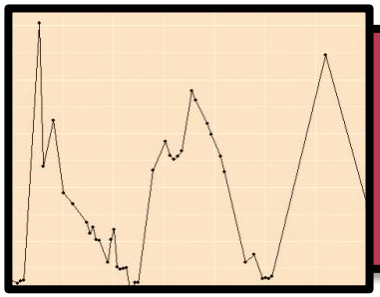
Idősorok elemzése

Salánki Ágnes

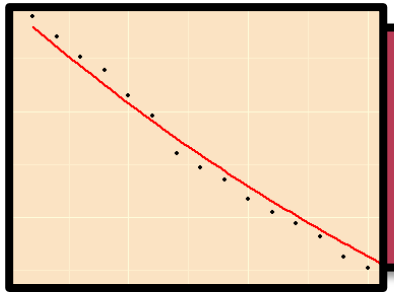
salanki.agnes@gmail.com

2012.04.13.

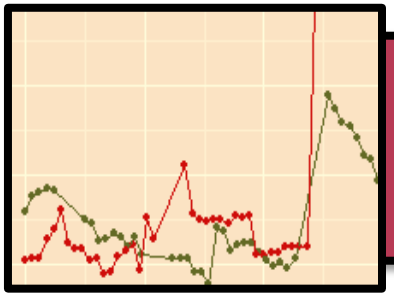
Idősorok analízise



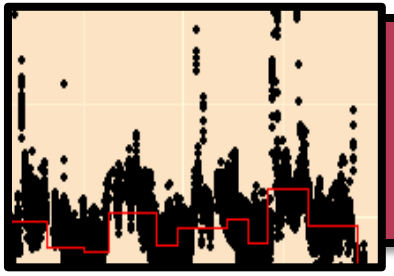
Alapfogalmak



Komponenselemzés



Összehasonlítás



Tárolás/Indexelés

Alapfogalmak

- Eddig: rekordok egy adatbázisban (valami struktúrával), a sorrend nem számít
  - Pl. vásárló kosara (pelenka → sör)

- Sorrend is: szekvenciaelemzés
  - Pl. időben különböző vásárlások (fényképezőgép → fényképnyomtató *egy hónapon belül*)

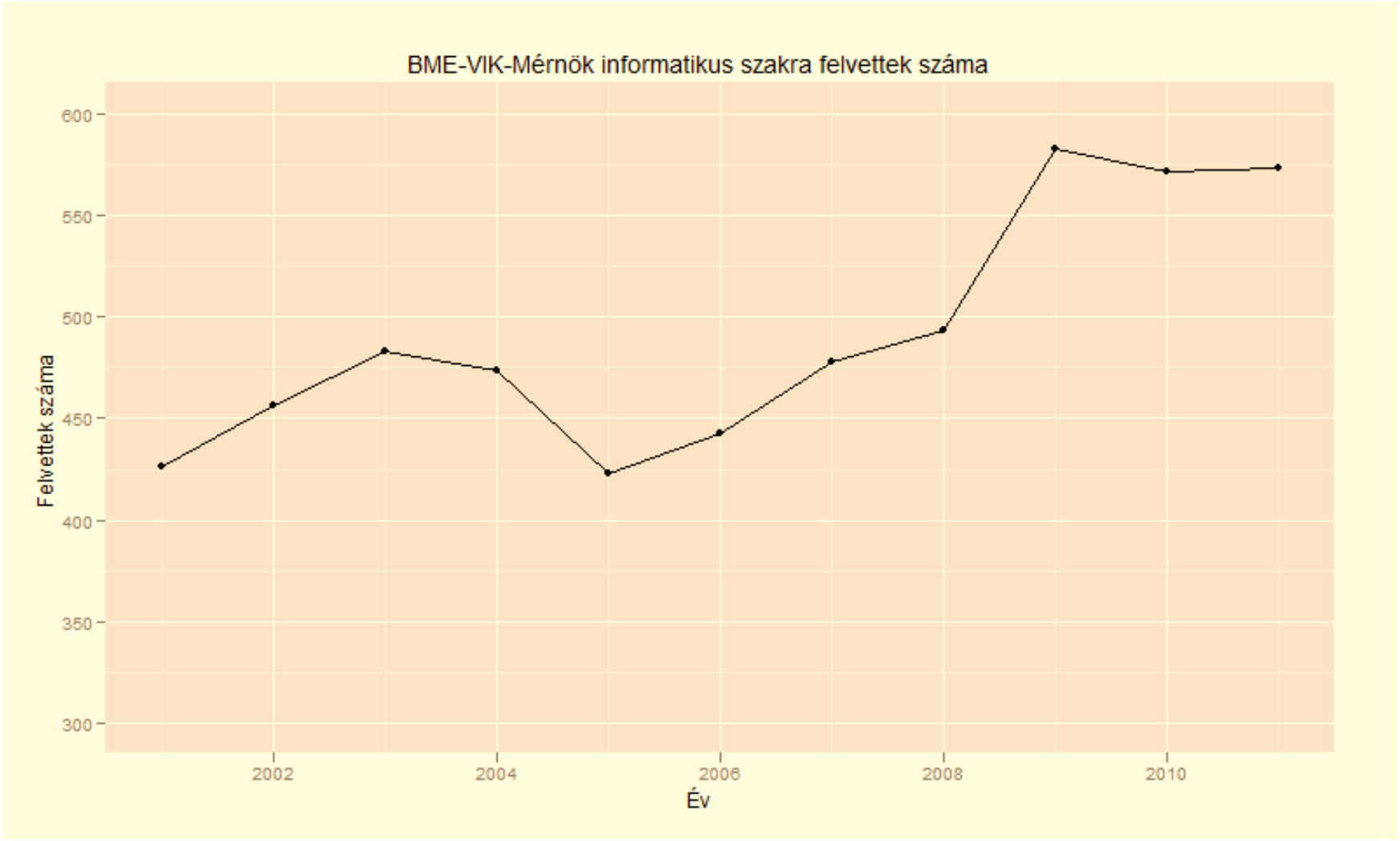
- Időbélyeg is: idősorelemzés

Alapfogalmak

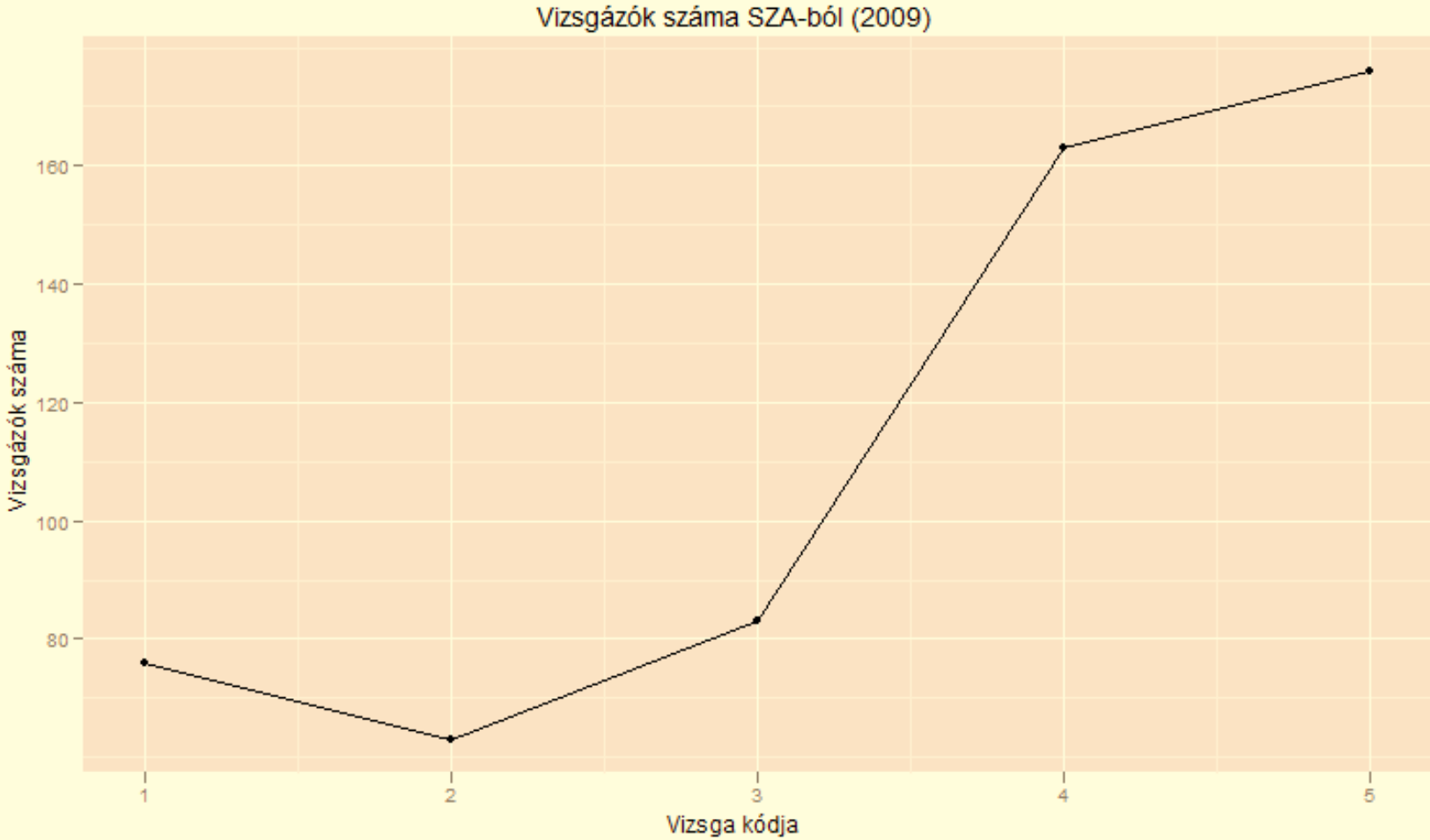
- Idősor-adatbázis – megadott időközönként rögzített érték- vagy esemény szekvenciák
  - Megj.: tehát minden idősor egy szekvencia is

- Alkalmazások
  - Természeti jelenségek adatai: hőmérséklet, légnyomás
  - Tőzsdeelemzés és –jóslás
  - Folyamatirányítási rendszerek

Idősor


BME-VIK-Mérnök informatikus szakra felvettek száma
Felvettek száma
600
550
500
450
400
350
300
2002
2004
2006
2008
2010
Év

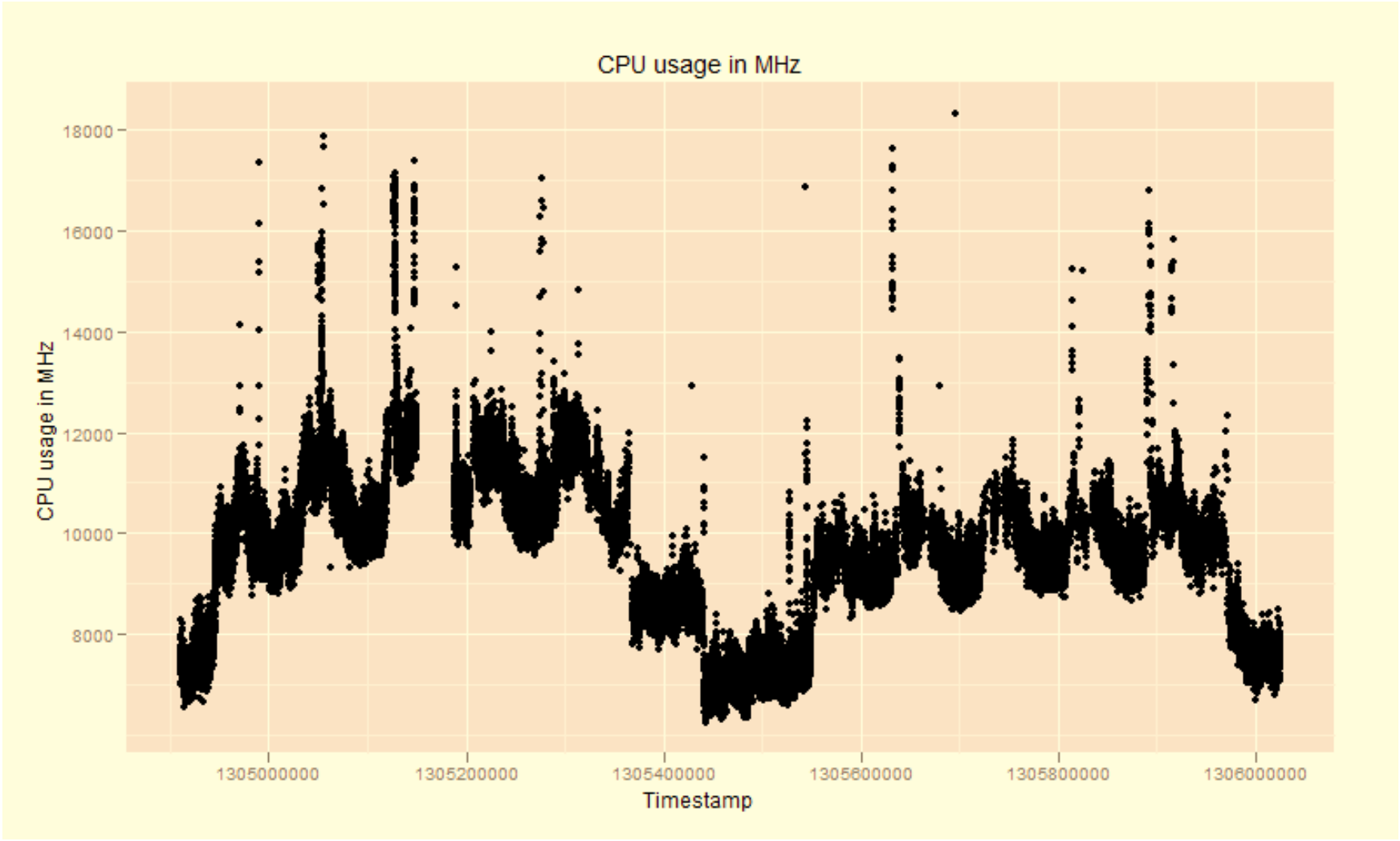
Idősor


Vizsgázók száma SZA-ból (2009)
Vizsgázók száma
160
140
120
100
80
1
2
3
4
5
Vizsga kódja

Idősor


Átmentek aránya SZA vizsgákon (az összes aznap vizsgázóhoz képest, 2009)
Átmentek aránya
0.7
0.6
0.5
0.4
0.3
1
2
3
4
5
Vizsga kódja

Idősor


CPU usage in MHz
18000
16000
14000
12000
10000
8000
CPU usage in MHz
1305000000
1305200000
1305400000
1305600000
1305800000
1306000000
Timestamp

Komponensek

$$Y = T + S + C + I$$

- **T**rendmozgás
- **S**zezonális mozgás
- **C**iklikus mozgás
- **I**rreguláris mozgás

Komponensek

$$Y = T + S + C + I$$

- **T**rendmozgás
- **S**zezonális mozgás
- **C**iklikus mozgás
- **I**rreguláris mozgás

általános irány egy hosszú időszakon belül

Komponensek

$$Y = T + S + C + I$$

- **T**rendmozgás
- **S**zezonális mozgás
- **C**iklikus mozgás
- **I**rreguláris mozgás

időszakok szerint rendszeres mozgás; pl. Valentin napi virágeladás

Komponensek

$$Y = T + S + C + I$$

- **T**rendmozgás
- **S**zezonális mozgás
- **C**iklikus mozgás
- **I**rreguláris mozgás

hosszú távú változások a trend körül;
megj.: nem feltétlenül periodikus

Komponensek

$$Y = T + S + C + I$$

- **T**rendmozgás
- **S**zezonális mozgás
- **C**iklikus mozgás
- **I**rreguláris mozgás

véletlenszerű, előre semmilyen módon nem jósolható

Idősorok elemzése

- Egyedi idősorok jellemzése
  - dekompozíció
  - becslések a jövőre nézve

- Idősorok viszonya egymáshoz
  - távolságok meghatározása
  - összehasonlítás

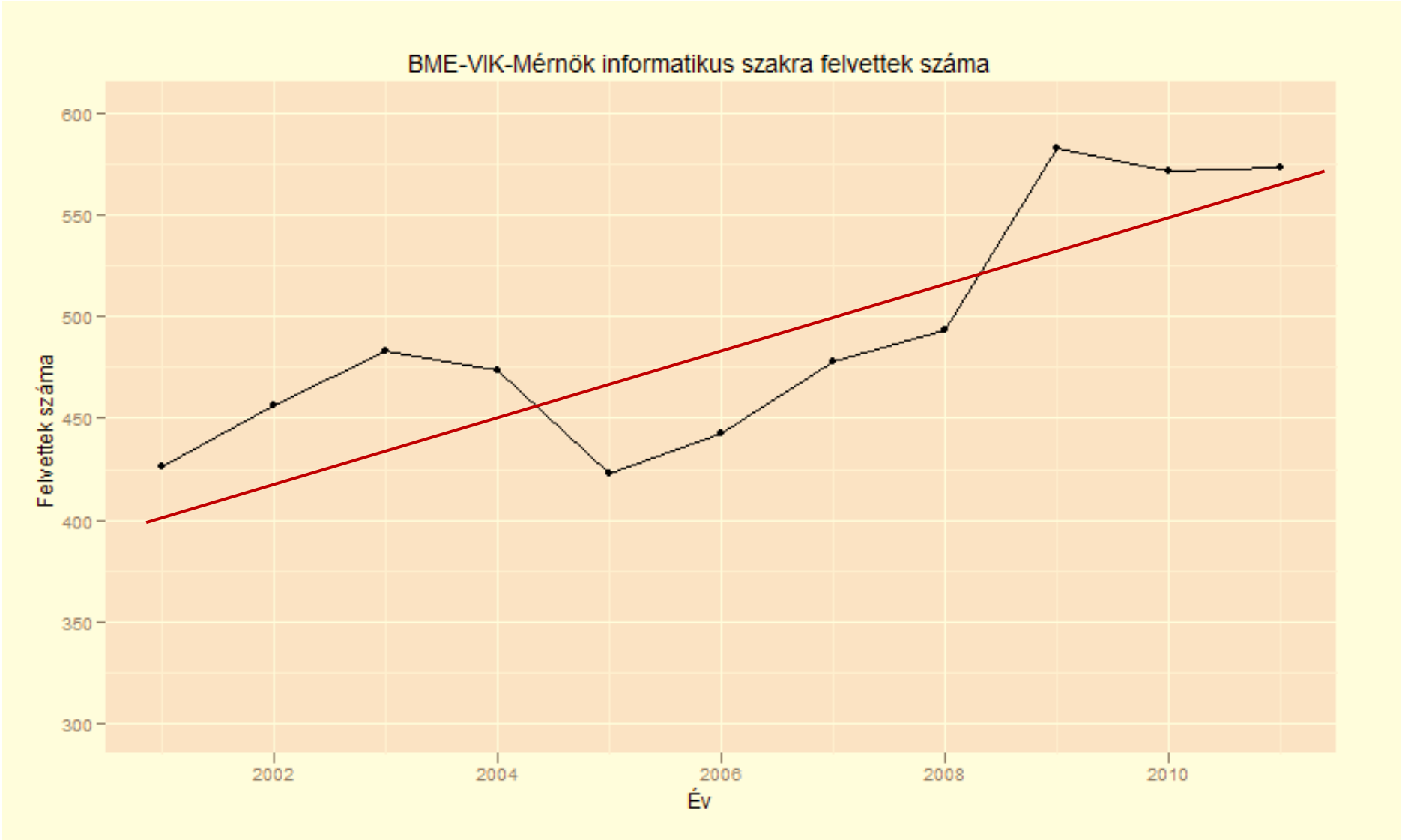
Trendelemzés

Dekompozíció, becslések

Trendelemzés

- Szabad kéz módszere ☺

Szabad kéz módszere



Trendelemzés

- Szabad kéz módszere ☺
- Analitikus módszer
  - regresszió

Regresszió

- Adott: $\boldsymbol{x} = [x_1, x_2, \dots x_n]$ független és $\boldsymbol{y} = [y_1, y_2, \dots, y_m]$ függő változók.
- Keressük: $\boldsymbol{y} = f(\boldsymbol{x}, \beta)$ összefüggést és ebben a β paramétert

- Lineáris regresszió
  - a függő változó a függetlenek lineáris kombinációja
- Nemlineáris regresszió
  - paramétereiben lineáris függvények
  - interpoláció spline-okkal

Lineáris regresszió

- Sejtés: az y és az x között lineáris kapcsolat
- Legyen n a tanítóhalmaz mérete, illesszünk rá hipersíkot:

$$\tilde{y} = \beta_0 + \sum_{j=1}^{n} \beta_j x_j$$

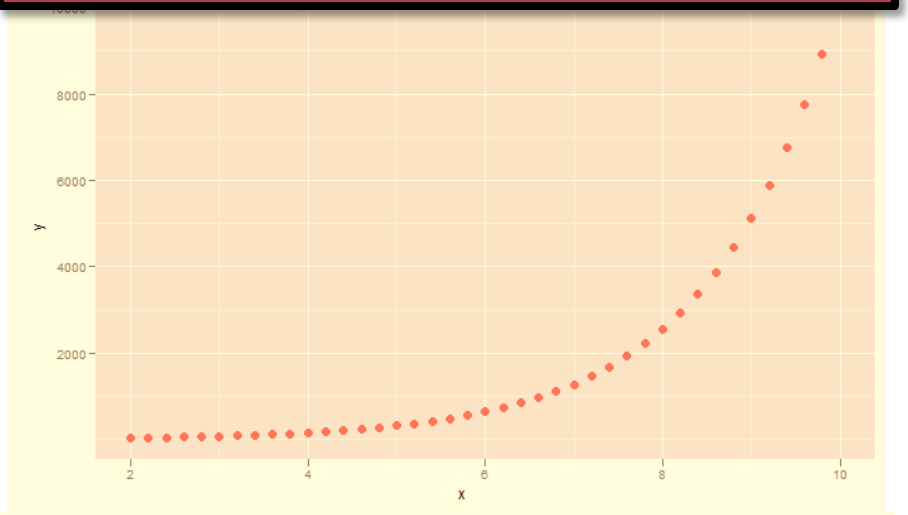
- A β paraméter meghatározása: legkisebb négyzetek módszere a teljes adathalmazra

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{n} x_{i,j}\, \beta_j)^2$$

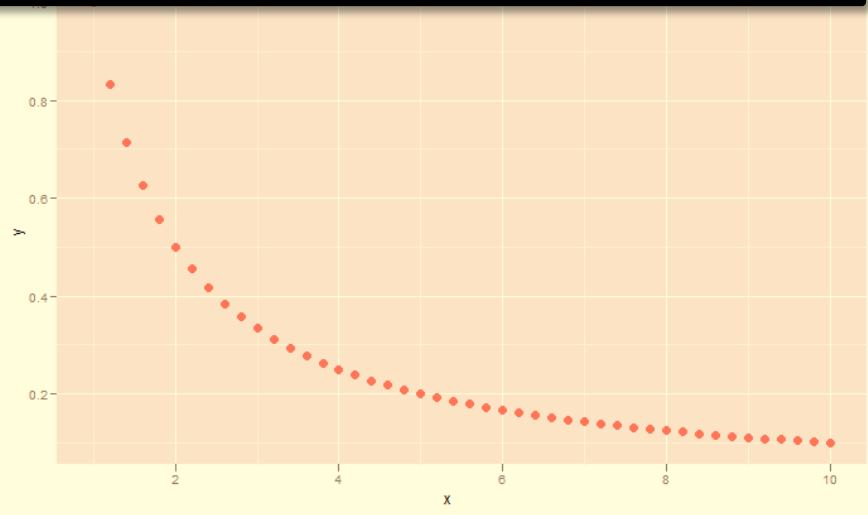
- Innen megvan az optimális β paramétervektor ☺
- DE: csak óvatosan…

Paramétereiben lineáris r.

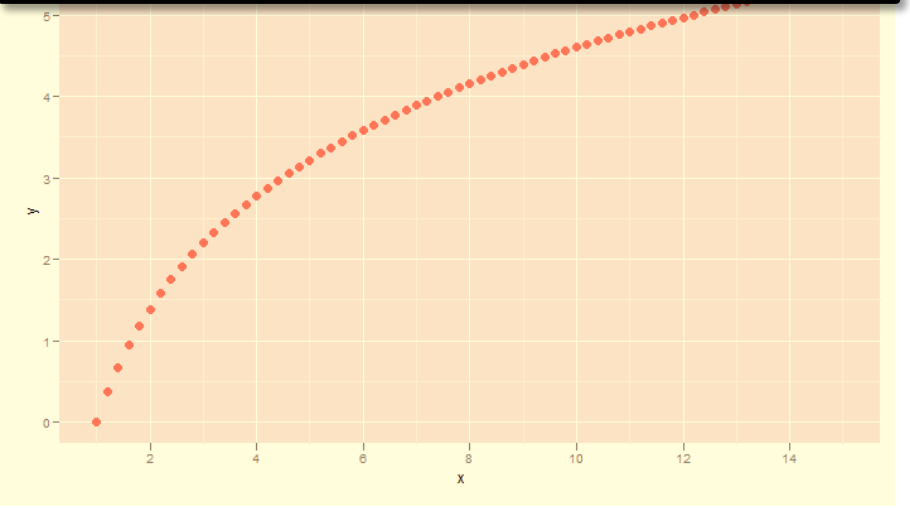
$y = a \cdot b^x$



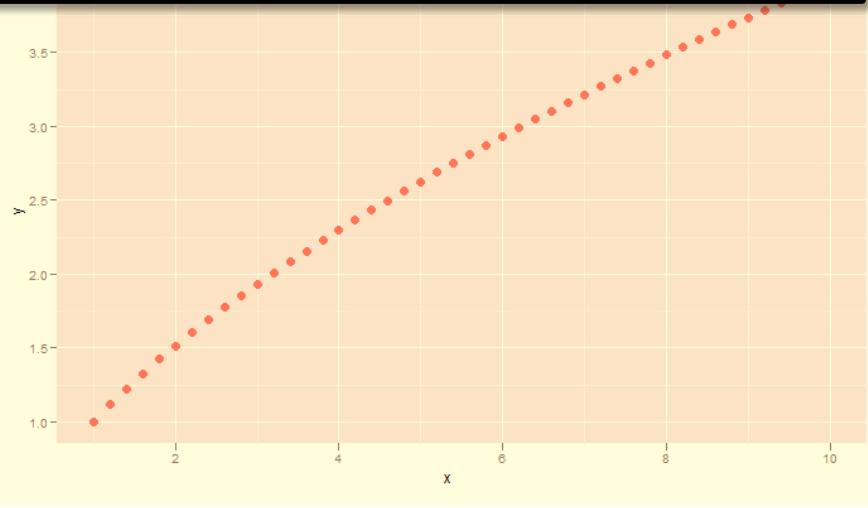
$y = a + b \cdot {}^1/_x$



$y = a + b \cdot \log x$

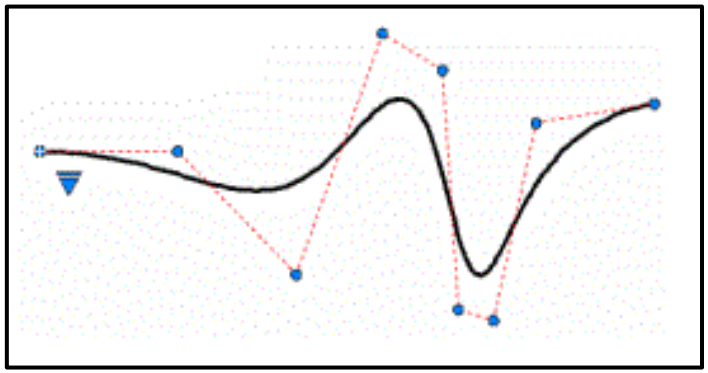


$y = a \cdot x^b$



Interpoláció spline-okkal

- Spline-ok



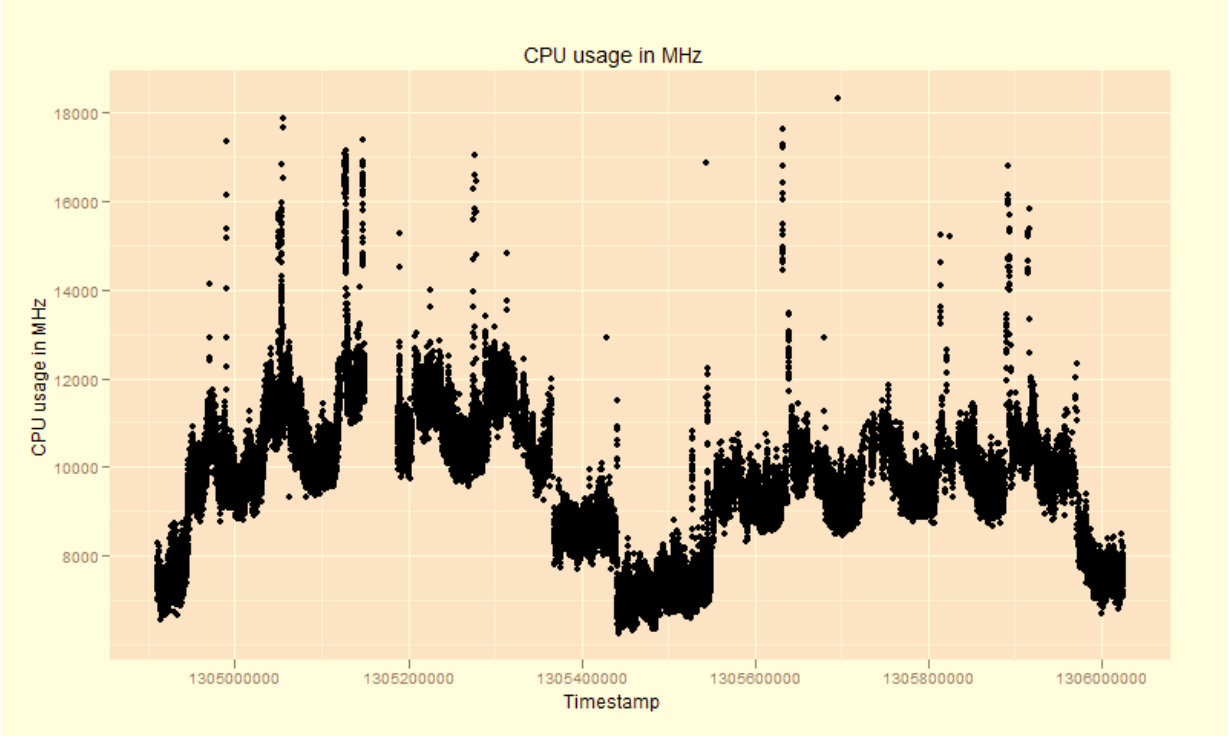
- Regressziós modellezés: általános modell
- Interpoláció: egy-egy pontra lokális függvény

- Alkalmazás:
  - Korlátozott számú mérőpont
  - Mérési hiba
  - Tehát előrejelzést nem tudunk adni!

Kép forrása: http://exchange.autodesk.com/autocadlt/hun/

Trendelemzés

- Szabad kéz módszere ☺
- Analitikus módszer
  - Regresszió
- Simítás?
  - Mozgó átlag!



Módszer

k-adrendű mozgóátlag

- k hosszú ablak, a benne szereplő elemek átlagát vesszük

Eredeti	3	5	1	0	8	10	6	2	1	3
$k = 3$	NA	3	2	3	6	8	6	3	2	NA
$k = 4$	NA	NA	2.9	4.1	5.4	6.3	5.6	4.4	NA	NA

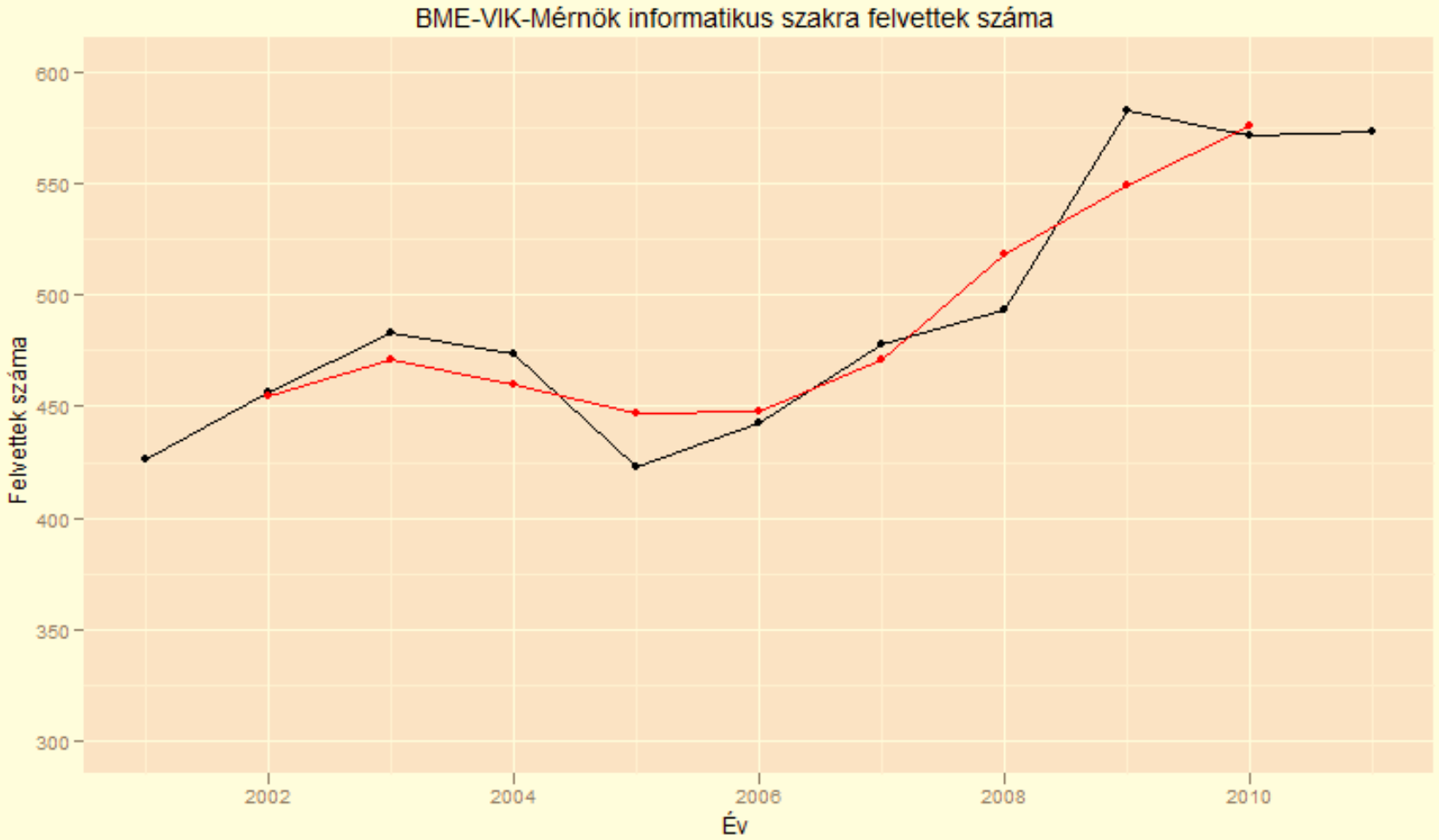
$$k_{2,3} = \frac{3 + 5 + 1}{3}$$

$$k_{7,4} = \frac{\mathbf{0.5} \cdot 8 + 10 + 6 + 2 + \mathbf{0.5} \cdot 1}{4}$$

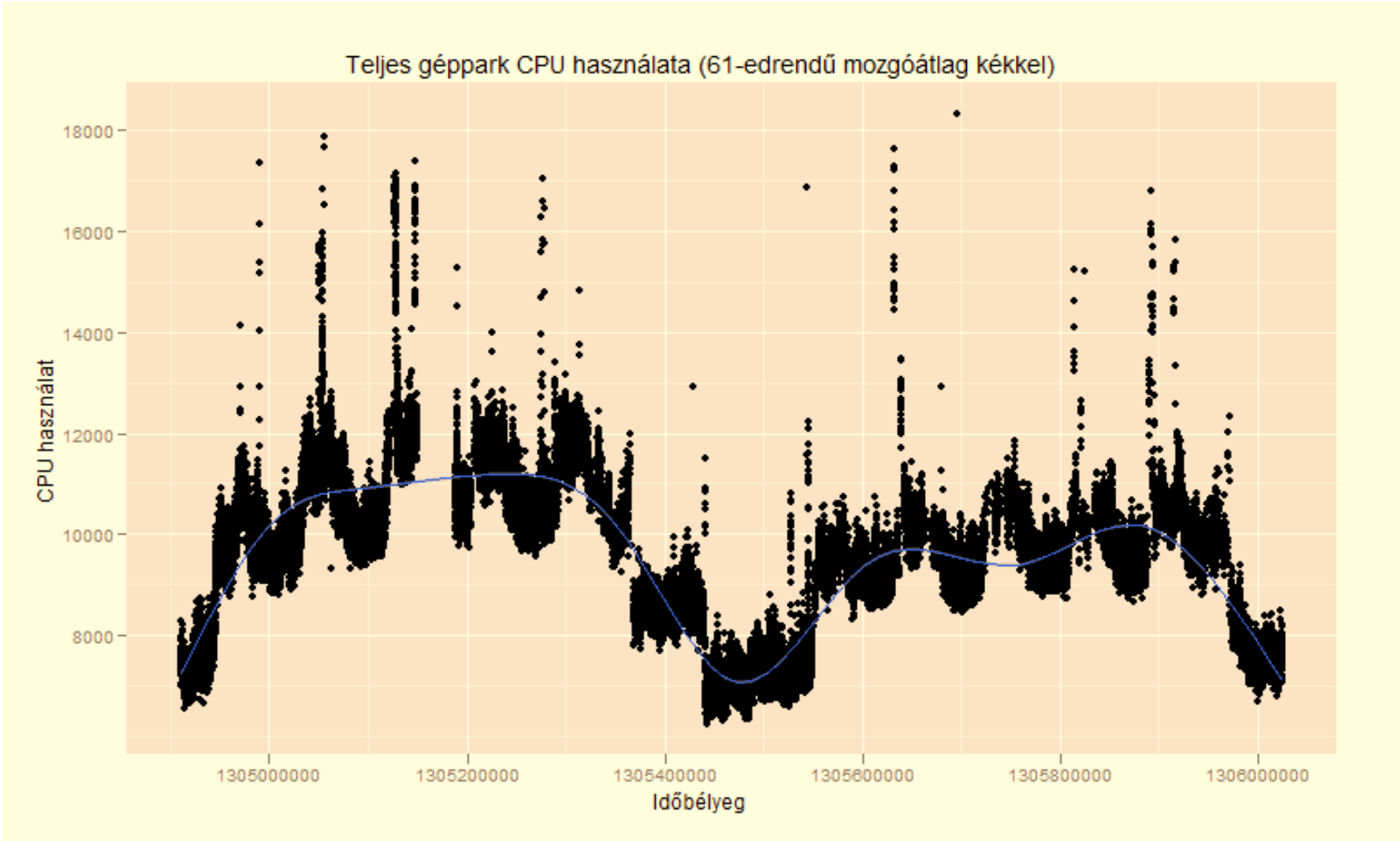
- Súlyozá

Eredeti	3									
$k = 3$	NA	3	2	3	6	8	6	3	2	NA
$k = 3$ ([1,3,1])	NA	3.8	1.6	1.8	6.8	8.8	6	2.6	1.8	NA

k-adrendű mozgóátlag



k-adrendű mozgóátlag



Szezonális mozgások

- Cél: a szezonális mozgások kiiktatása/becslése
  - Minden, ami naptári intervallumhoz köthető
  - Példa: pékség és virágüzlet január-március

- Szezonális index: havi átlagtól való eltérés %-ban
  - Ezekkel kell leosztani/ezeket kell kivonni

Szezonális mozgások

- Alapfeladat: kezdeményezett mobilhívások (millió)

Év	I. negyedév	II. negyedév	IIII. negyedév	IV. negyedév
2001	777	975	1044	984
2002	963	1115	1159	1162
2003	1068	1207	1206	1219
2004	1138	1307	1347	1332

Szezonális mozgások

- Alapfeladat: kezdeményezett mobilhívások (millió)

Év	I. negyedév	II. negyedév	IIII. negyedév	IV. negyedév
2001	777	975	1044	984
2002	963	1115	1159	1162
2003	1068	1207	1206	1219
2004	1138	1307	1347	1332

Lépések:

1. Trend kiszámítása (itt: negyedrendű mozgóátlag)

Szezonális mozgások

- Alapfeladat: kezdeményezett mobilhívások (millió)

Év	I. negyedév	II. negyedév	IIII. negyedév	IV. negyedév
2001			968.3	1009
2002	1040.9	1077.5	1112.9	1137.5
2003	1154.9	1167.9	1183.8	1205
2004	1235.1	1266.9		

Lépések:

1. Trend kiszámítása (itt: negyedrendű mozgóátlag)

Szezonális mozgások

- Alapfeladat: kezdeményezett mobilhívások (millió)

Év	I. negyedév	II. negyedév	IIII. negyedév	IV. negyedév
2001			968.3	1009
2002	1040.9	1077.5	1112.9	1137.5
2003	1154.9	1167.9	1183.8	1205
2004	1235.1	1266.9		

Lépések:

1. Trend kiszámítása (itt: negyedrendű mozgóátlag)
2. Szezonális index kiszámítása

Szezonális mozgások

- Alapfeladat: kezdeményezett mobilhívások (millió)

Év	I. negyedév	II. negyedév	IIII. negyedév	IV. negyedév
2001			1.08	0.98
2002	0.89	1.04	1.04	1.02
2003	0.93	1.03	1.02	1.01
2004	0.92	1.03		

Lépések:

1. Trend kiszámítása (itt: negyedrendű mozgóátlag)
2. Szezonális index kiszámítása

Szezonális mozgások

- Alapfeladat: kezdeményezett mobilhívások (millió)

Év	I. negyedév	II. negyedév	IIII. negyedév	IV. negyedév
2001			1.08	0.98
2002	0.89	1.04	1.04	1.02
2003	0.93	1.03	1.02	1.01
2004	0.92	1.03		

Lépések:
1. Trend kiszámítása (itt: negyedrendű mozgóátlag)
2. Szezonális index kiszámítása
3. Átlagolás

Szezonális mozgások

- Alapfeladat: kezdeményezett mobilhívások (millió)

Év	I. negyedév	II. negyedév	IIII. negyedév	IV. negyedév
2001			1.08	0.98
2002	0.89	1.04	1.04	1.02
2003	0.93	1.03	1.02	1.01
2004	0.92	1.03		
Átlag	**0.91**	**1.03**	**1.05**	**1.00**

Lépések:
1. Trend kiszámítása (itt: negyedrendű mozgóátlag)
2. Szezonális index kiszámítása
3. Átlagolás

Ciklikus mozgások

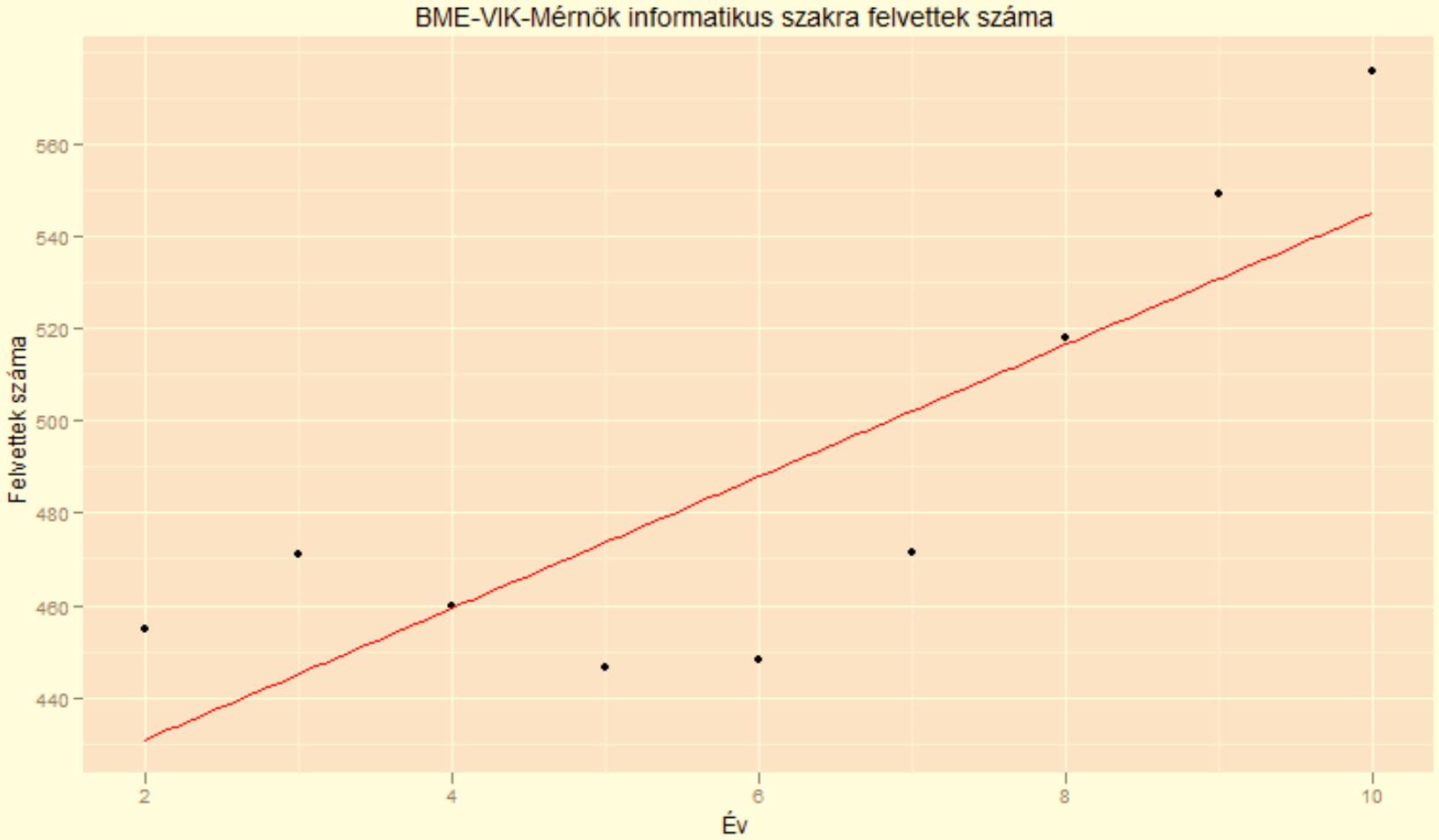
- Cél: ciklikus komponens meghatározása

- Szezonmentesített adatokon:

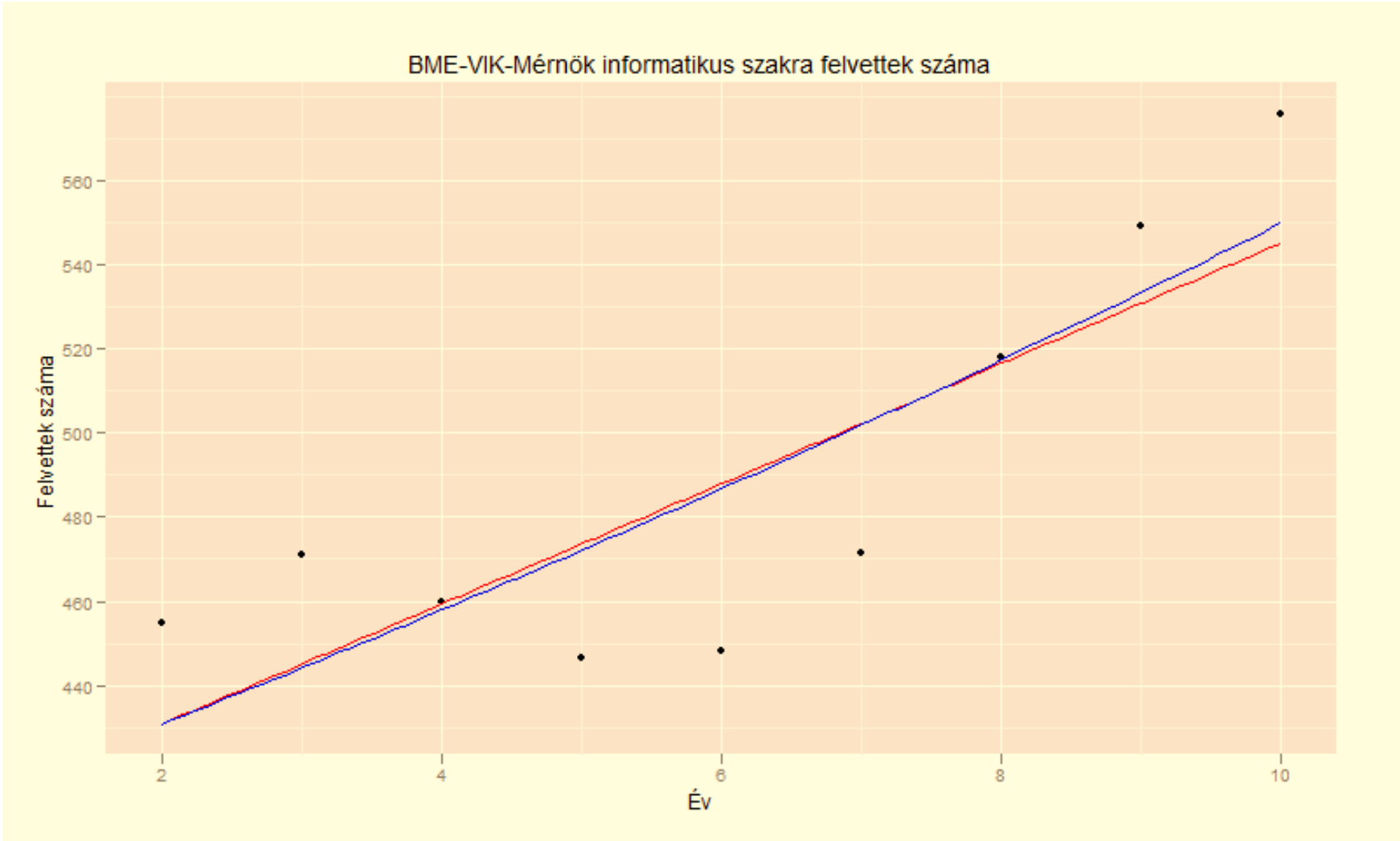
Ciklikus = mozgóátlag- trendfüggvény

Lépések:
1. Trend kiszámítása (mozgóátlag)
2. A mozgóátlagra trendfüggvényt illesztünk
3. Ciklikus = mozgóátlag - trendfüggvény

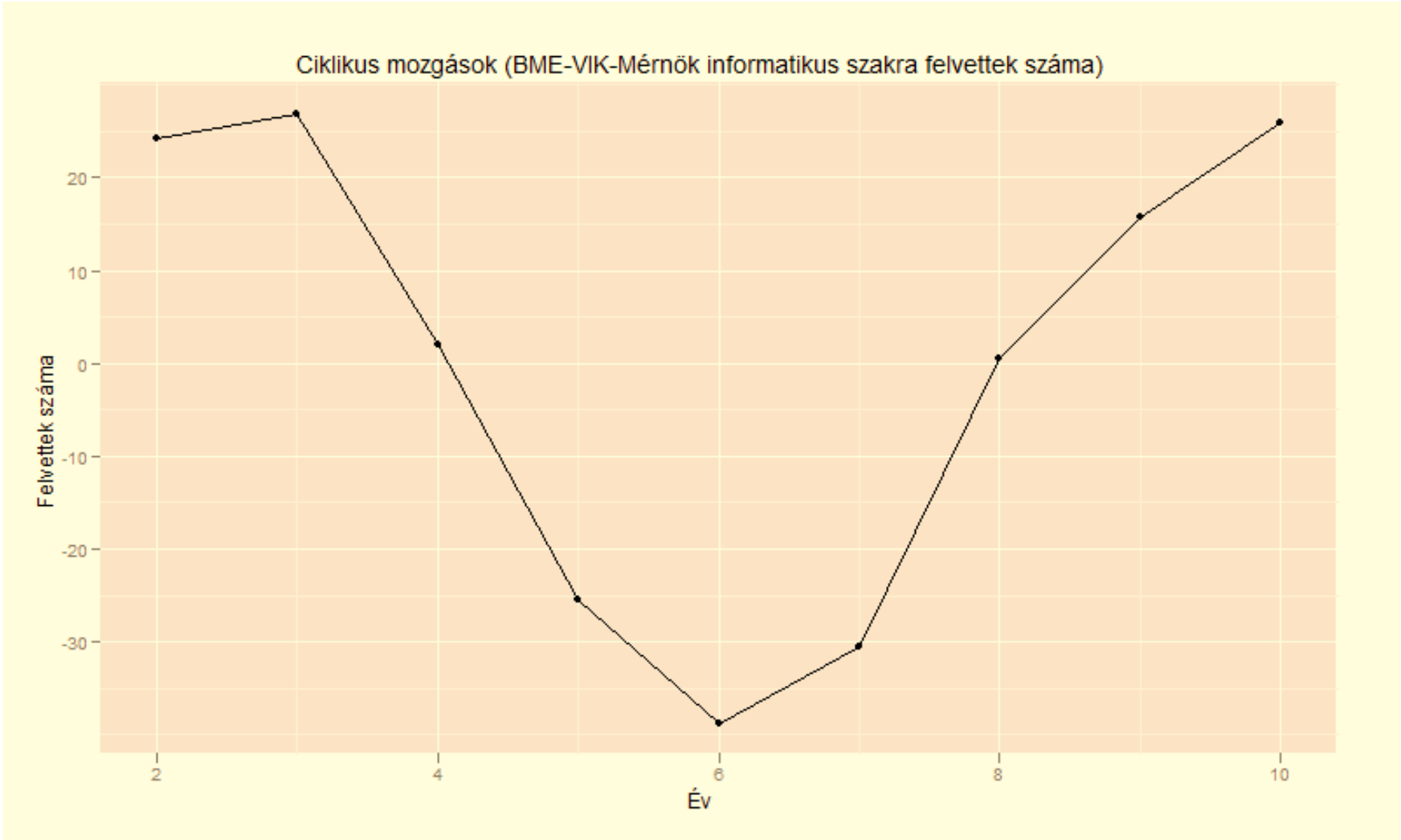
Ciklikus mozgások 2.



Ciklikus mozgások 2.



Ciklikus mozgások 2.



Összehasonlítás

Távolságfüggvények, dinamikus idővetemítés

Összehasonlítás motivációk

- Teljes egyezés (*whole sequence matching*)
  - egy szekvencia-halmazban talál hasonlóakat egymáshoz
  - Pl. olyan termékek, amiknek az eladási mutatói hasonlóak

- Rész-szekvencia keresése (*subsequence matching*)
  - egy előre definiált rész-szekvenciát keresünk az idősorokban
  - Pl. EKG-ban adott rendellenesség keresése

Minkowski távolság

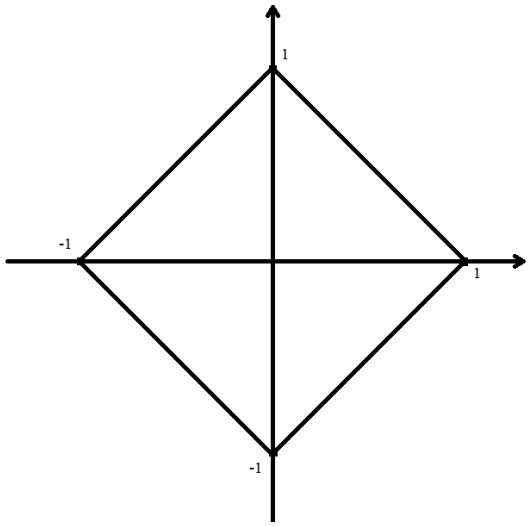
- Két n dimenziós adatvektor távolsága lehet például:

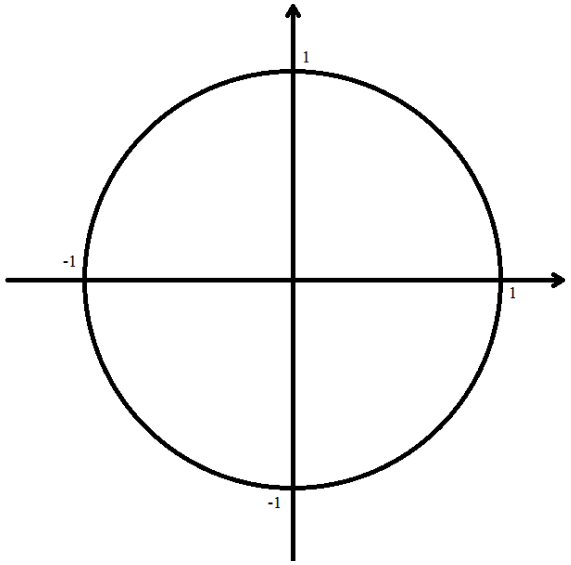
$$d_p(x,y) = \sqrt[p]{\sum_{k=1}^{n}(x_k - y_k)^p} = \|x - y\|_p$$

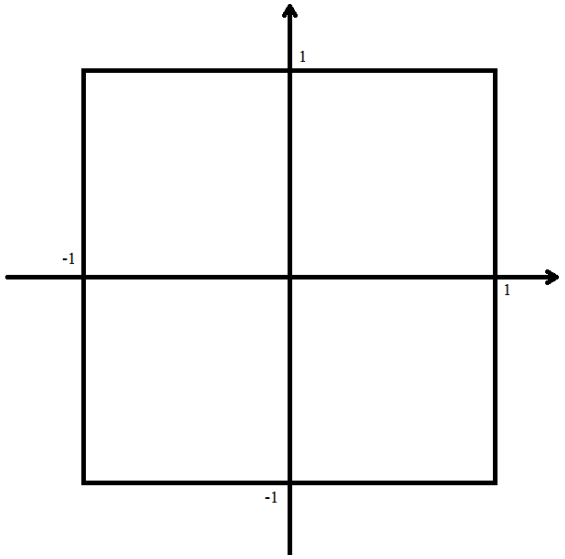
Speciális Minkowskik

- Manhattan távolság $p = 1$
  - $d_1(x, y) = \sum|x_k - y_k|$
- Euklideszi távolság $p = 2$
  - $d_2(x, y) = \sqrt[2]{(\sum(x_k - y_k)^2)}$
- Chebyshev távolság $p \rightarrow \infty$
  - gyakorlatilag a megegyező indexű elemek távolságai közül a legnagyobb

Speciális Minkowskik







A (0, 0) ponttól 1 egységre lévő pontok halmaza $p = 1, p = 2$ és $p = \infty$ esetén

Euklideszi problémák

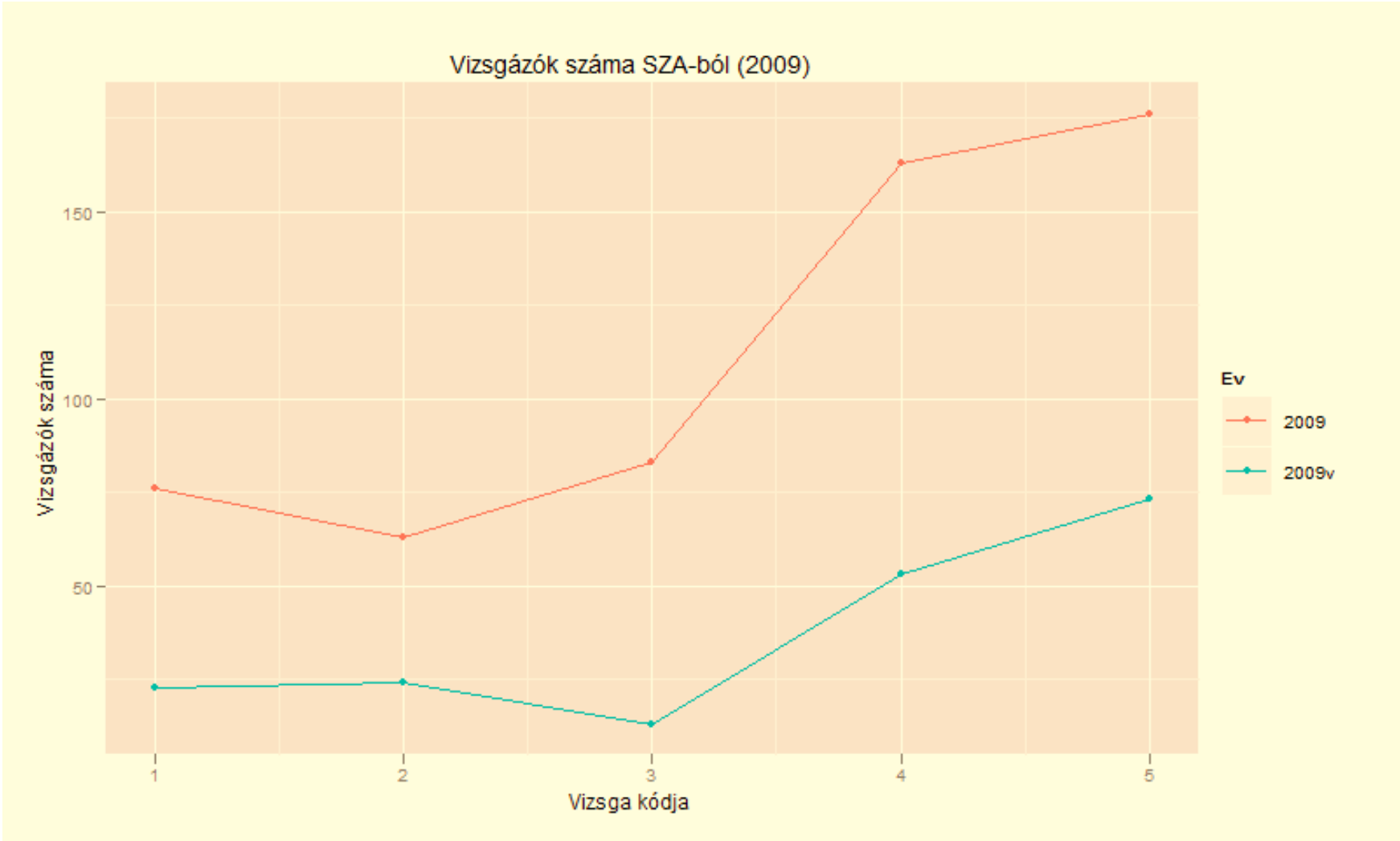
- Outlierek
  - Mozgóátlag
  - Zajszűrés

- Y tengely menti eltolás különböző
- Y tengely menti skálázás különböző

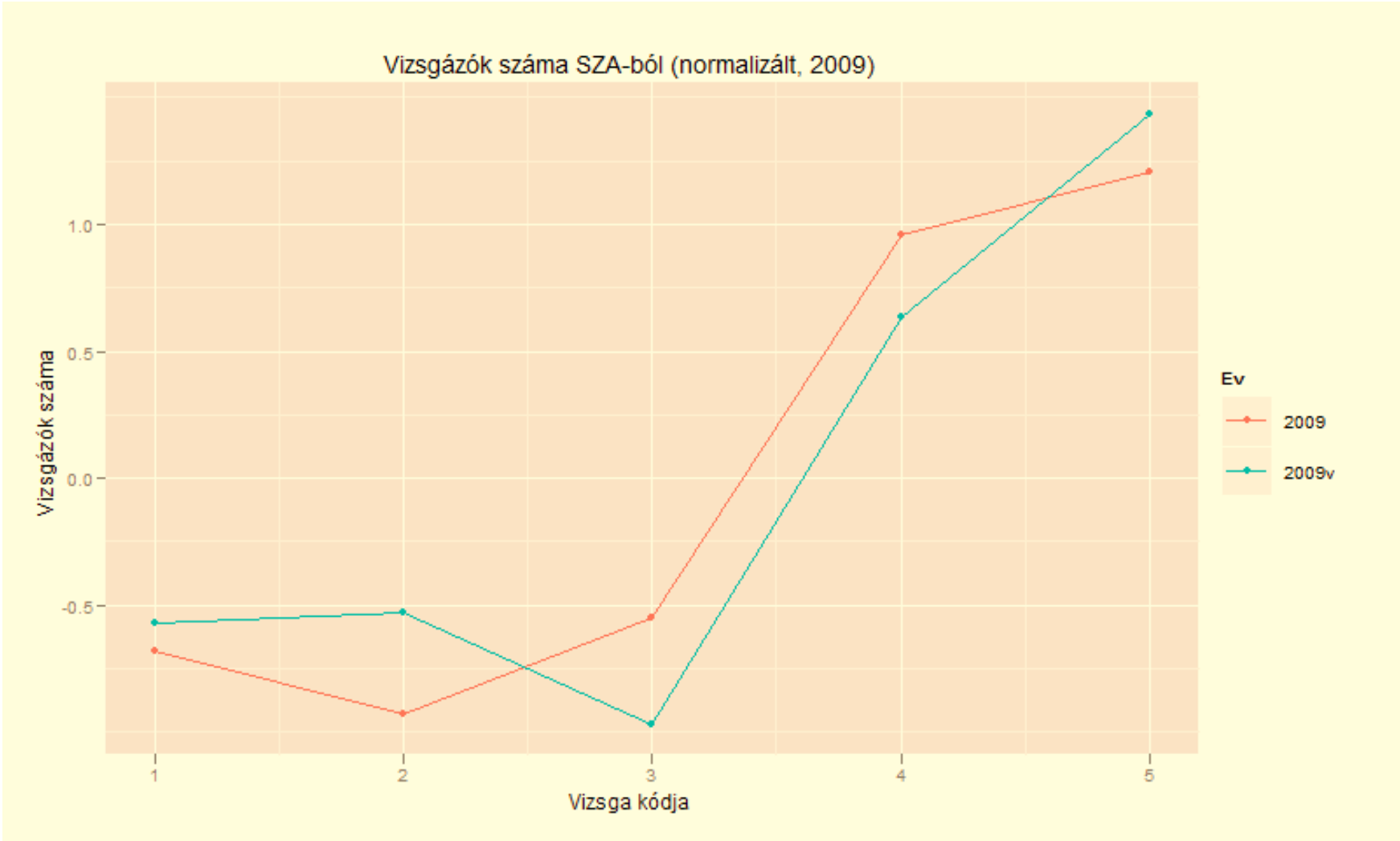
„normalizálás”:

$$y_i = \frac{y_i - Átlag(y)}{SZÓRÁS(y)}$$

Euklideszi problémák



Euklideszi problémák



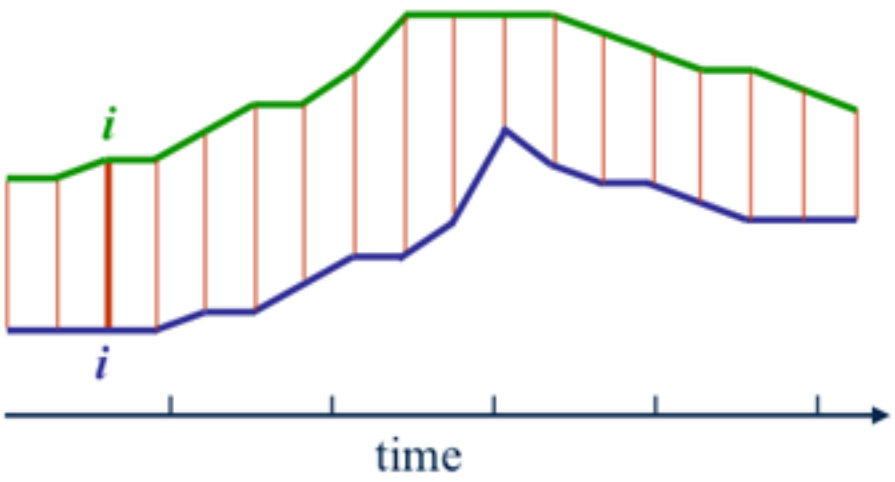
Euklideszi problémák 2

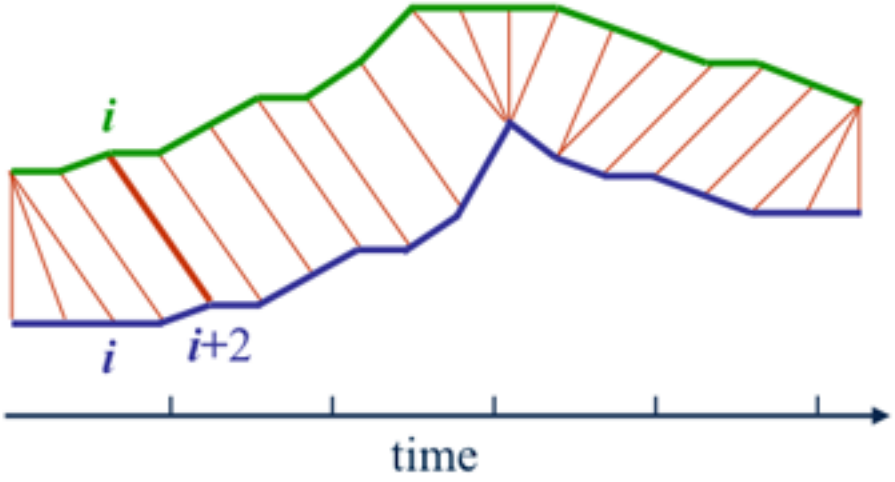
- X tengely menti eltolás?
- Összehasonlítás kiugró értékek alapján?
- Különböző hosszúságú idősorok?

Dinamikus idővetemítés

Dinamikus idővetemítés

- Dynamic Time Wraping
- Az idősorok pontjait nem indexenként hasonlítjuk össze
  - Motiváció pl. hangfelismerésnél





Kép forrása: http://homepages.inf.ed.ac.uk/group/sli_archive/slip0809_c/s0562005/img/DTWExplain.png

Dinamikus idővetemítés számítása

Lépések

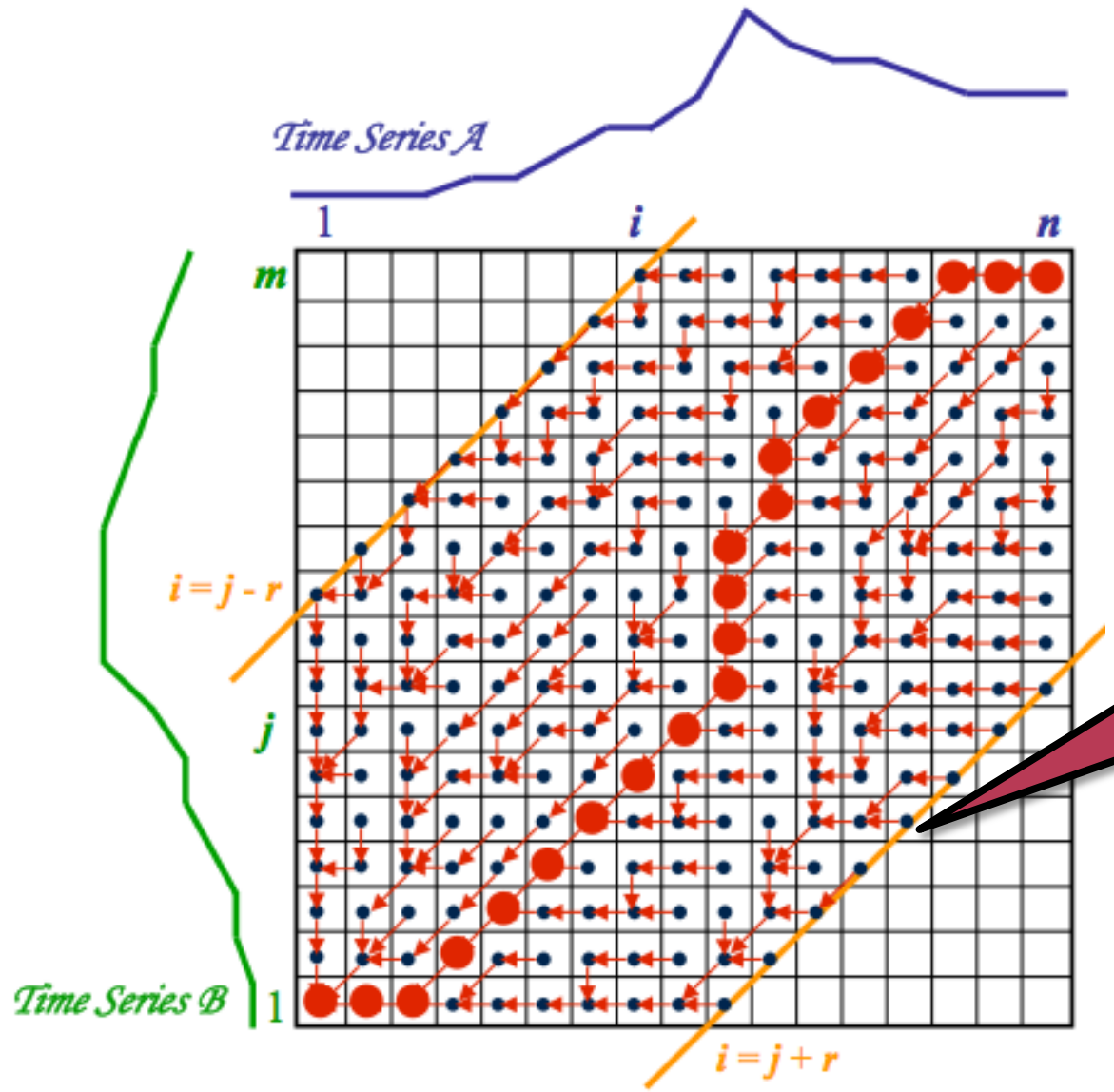
1. $n \times m$-es $\boldsymbol{D}$ mátrixban rögzítjük a sorok egymástól való távolságát
2. Lépkedünk a szomszédos mezőkön

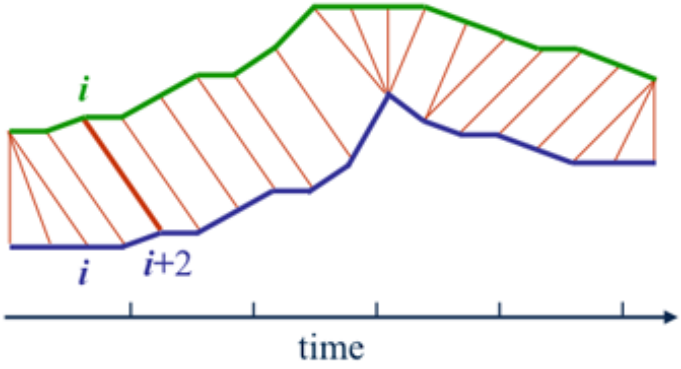
Kell: $p = [p_1, p_2, \dots p_k]$ útvonal a $D[1,1]$ és $D[n,m]$ között
Cél: minimális költségű út

Szabályok:

1. Minden lépésben előre haladunk (nem távolodhatunk, tehát $[i,j] \rightarrow [\tilde{\imath}, \tilde{\jmath}]$ esetén $\tilde{\imath} \geq i, \tilde{\jmath} \geq j$)
2. Az út folytonos, mindig csak szomszédos cellákra léphetünk

Dinamikus idővetemítés





Sakoe-Chiba sáv

(**Ikatura par.)

Kép forrása: http://homepages.inf.ed.ac.uk/group/sli_archive/slip0809_c/s0562005/theory.html

Más hasonlóságok

- Befoglaló négyszögek hasonlósága
  - Inkább a különbözőek szűrhetőek ki
  - Nem túl hatékony

- Burkolt szegmensek hasonlósága
  - Két konfigurációs paraméter
    - Szegmensek közötti megengedett eltérés: ε
    - Azonos szegmensek minimális értéke: min_supp
  - Gyakorlatban is alkalmazott

Tárolás/Indexelés

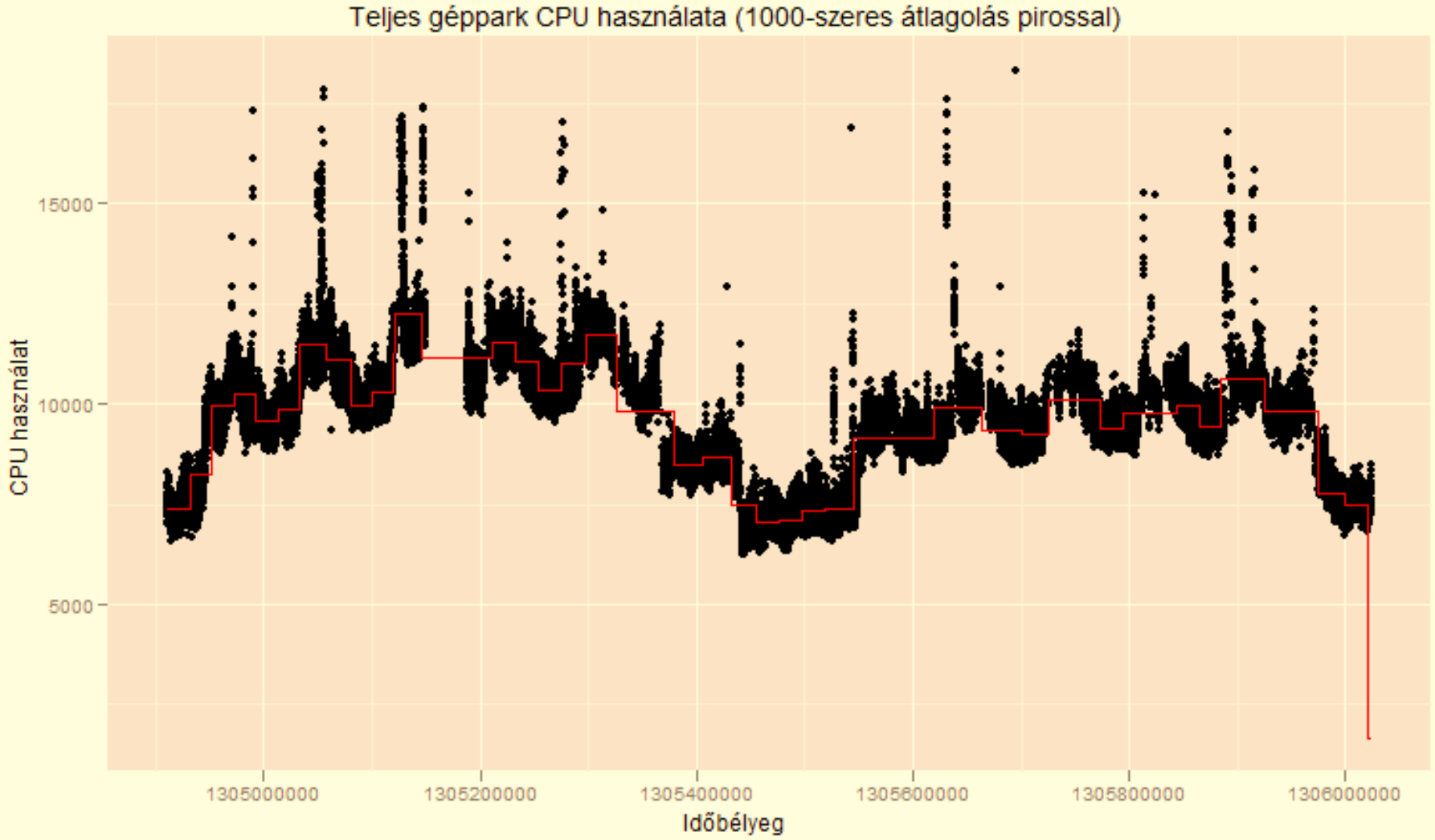
DFT, szegmentálás, fontos pontok

Idősorok hatékony tárolása

- k-adrendű átlagolás
- Szegmentálás
- „Fontos pontok”

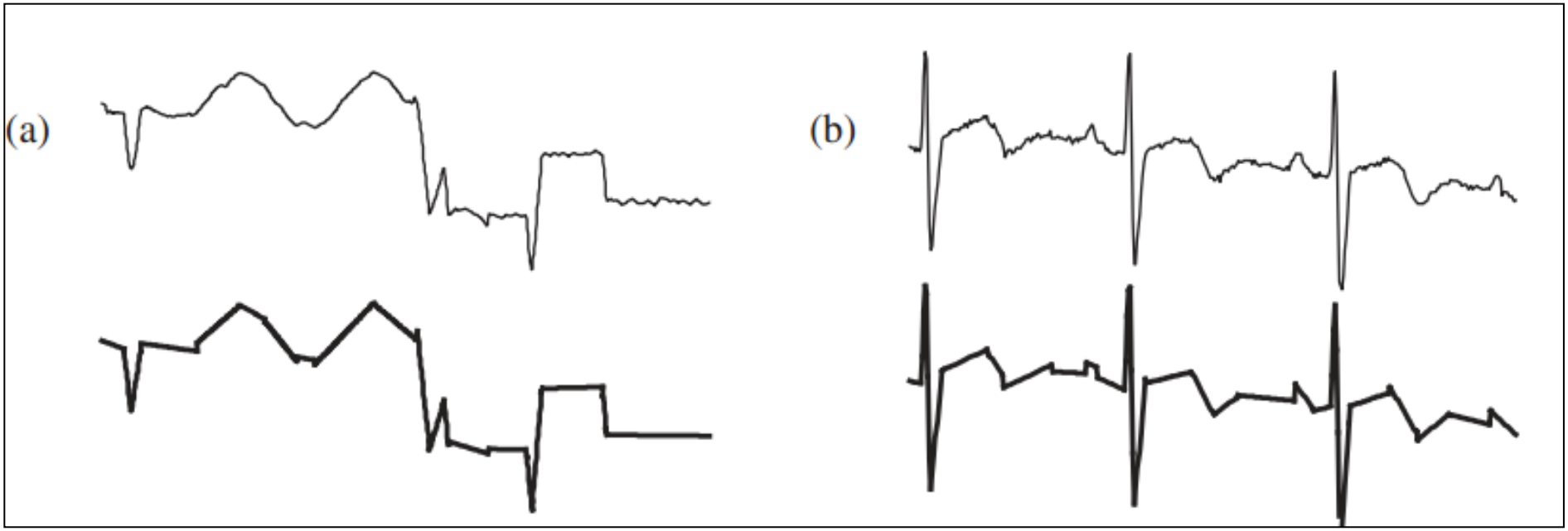
- Diszkrét Fourier-transzformáció

k-adrendű átlagolás



Szegmentálás

- Cél: az idősort egyenes szakaszokkal közelítsük
  - Minél kevesebb egyenes legyen VAGY
  - Minél pontosabban közelítsen



Kép forrása: Keogh, Chu, Hart, Pazzani. *Segmenting Time Series: A Novel Approach*

Szegmentálás 2

- Csúszóablakos algoritmus
  - Egyszerű, online, de ha nagyon „rezeg”, nem jó
  - pl. egészségügy
- Top Down
  - Minden lehetséges töréspontot megvizsgál
  - A legkisebb hibájúnál vág
- Bottom Up
  - A legfinomabb felbontással kezdjük
  - A legjobban illeszkedző i. és $i+1$. szakaszt összeolvasztjuk, majd frissítjük a távolságokat

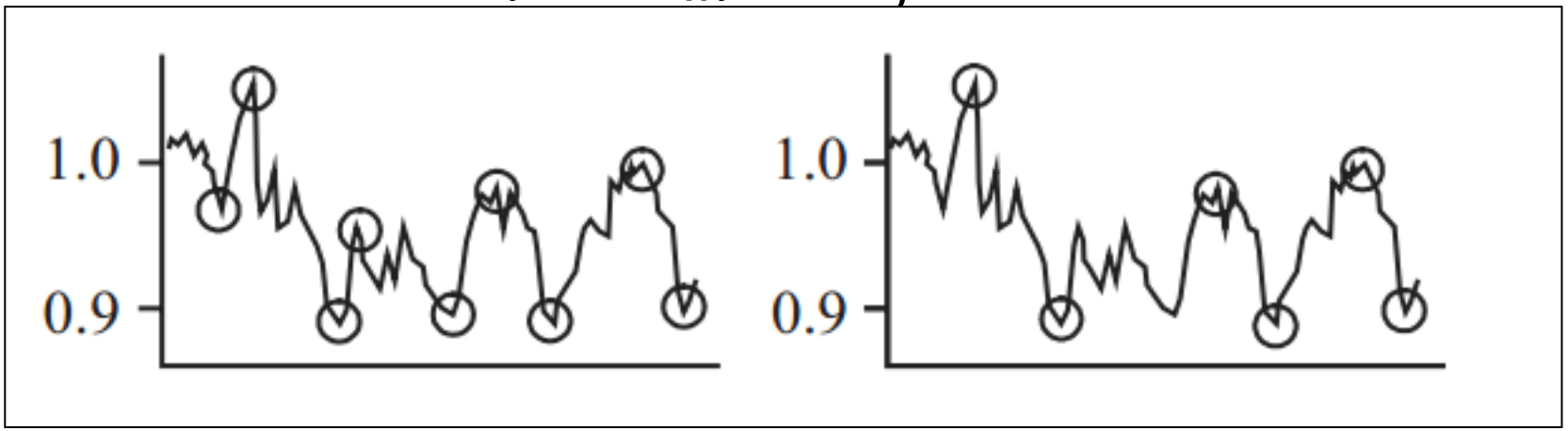
Összehasonlítás

- Magyarázat
  - E – maximális hiba egy adott szegmensre
  - ME – teljes közelítő hiba
  - K – szegmensek száma
  - L – a szegmens átlagos hossza

Algoritmus	Konfiguráció	Online	Bonyolultság
Csúszóablakos	E	IGEN	$O(Ln)$
Top-Down	E, ME, K	NEM	$O(Kn^2)$
Bottom-Up	E, ME, K	NEM	$O(Ln)$

Fontos pontok módszere

- Válasszuk ki a reprezentatív pontokat, a többieket hagyjuk el
- Összehasonlításnál a többi idősornak is csak a fontos pontjait vesszük figyelembe
- Miért lehet egy a_m pont fontos minimum?
  - Minimális az $a_i, \ldots, a_m, \ldots, a_j$ szekvenciában



Kép forrása: Fink, Pratt. Indexing of Compressed Time Series

Diszkrét Fourier-transzformáció

- Cél: időtartományból frekvenciatartományba transzformálni az adatokat

- Miért jó?
  - a zajszűrés könnyebb
  - a transzformálás lineáris
    $(af_1(t) + bf_2(t) = aF_1(\omega) + bF_2(\omega))$
  - a tengelyeken való eltolások megjelennek (kompenzálni tudunk a fr.t.-ban is)
  - azonos Euklideszi távolság az idősorok és tr.-ik között

Diszkrét Fourier-transzformáció 2

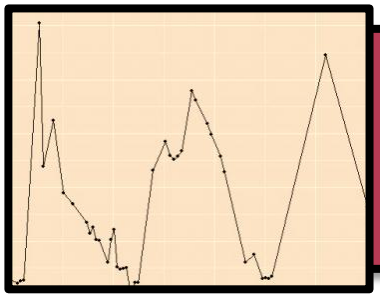
Képzése

$$F[k] = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} f(i) e^{\frac{-2\pi jki}{N}}$$

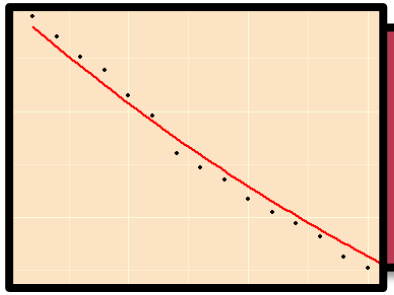
$$\forall\ 0 \ \leq k \ \leq N-1\text{-re}$$

1. Amit kapunk: Fourier-együtthatók
2. Tömörítés: az első néhány együttható alapján jellemzünk csak
3. Készítünk az együtthatók alapján egy keresőfát, amiben már könnyű lesz keresni

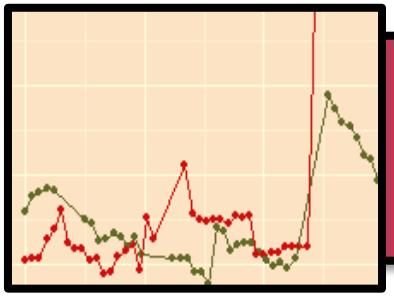
Idősorok analízise – Összefoglalás



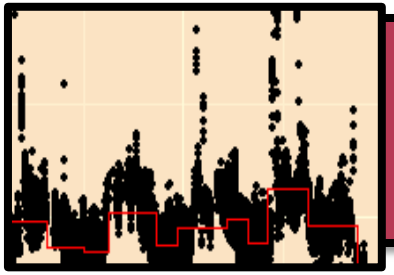
Alapfogalmak



Komponenselemzés



Összehasonlítás



Tárolás/Indexelés