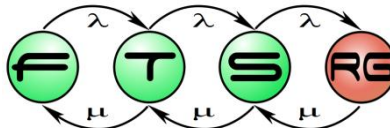


Terhelés modellek, előrejelzés

Gönczy László
gonczy@mit.bme.hu



Motiváció

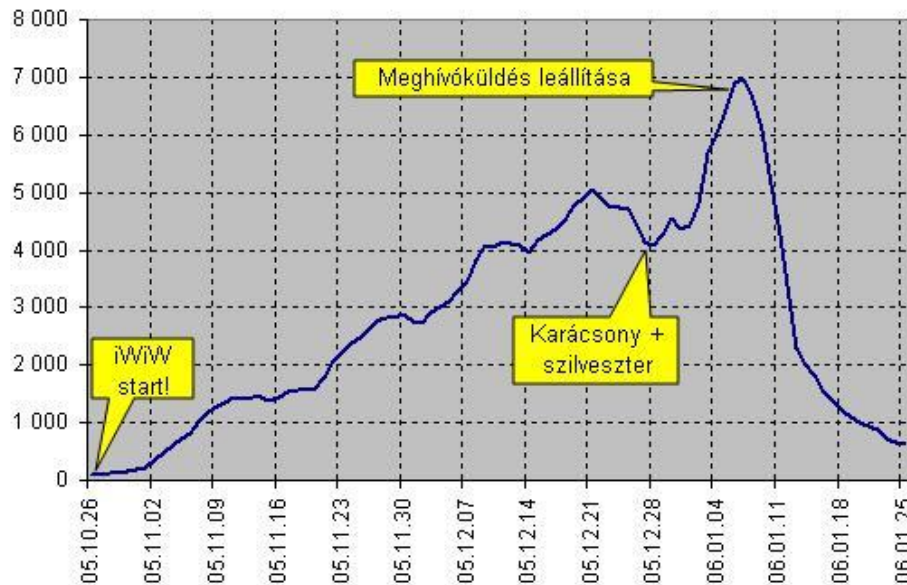
The collage consists of five overlapping screenshots of web browser windows:

- Top Left:** A window titled "SAP Developer Network (SDN): Downloads, Discussions, eLearning, and Documentation for Developer - Windows Internet Explorer". The page shows a message: "The SAP Developer Network is currently down. We are working to restore the site as quickly as possible. Current downtime is estimated at less than one hour. You may contact us with questions or concerns."
- Top Right:** A window titled "IEEE Xplore: Guest Home Page - Windows Internet Explorer". The page shows the IEEE Xplore logo and a message: "IEEE Xplore is temporarily unavailable. Please try again later."
- Middle Left:** A window titled "Server error! - Windows Internet Explorer". The page shows a "Server error!" message: "The server encountered an internal error and was unable to complete your request. Either the server is overloaded or there was an error in the application. If you think this is a server error, please contact the webmaster." Below this is an "Error 500" section with a link to "index.hu" and some technical details: "Wed Sep 12 16:36:17 2007 Apache/2.0.34 (Debian GNU/Linux) mod_fastcgi/2.4.2".
- Middle Right:** A window titled "easyJet.com - book cheap flights, hotels, car hire for low cost travel - Windows Internet Explorer". The page shows a message: "Sorry, but the easyJet sales system is temporarily unavailable [css01]". Below this is a message: "We regret that bookings cannot be made at present, while we undertake essential maintenance work on our systems. As this affects our central reservations computer, sales cannot currently be made via either our web site or our call centre. We anticipate that this work will be completed within 30 minutes. Thank you for your patience."
- Bottom Left:** A window titled "Gateway Timeout - In read". The page shows a message: "The proxy server did not receive a timely response from the upstream server. Reference #1.e50de1c1.1328609312.18d592a".

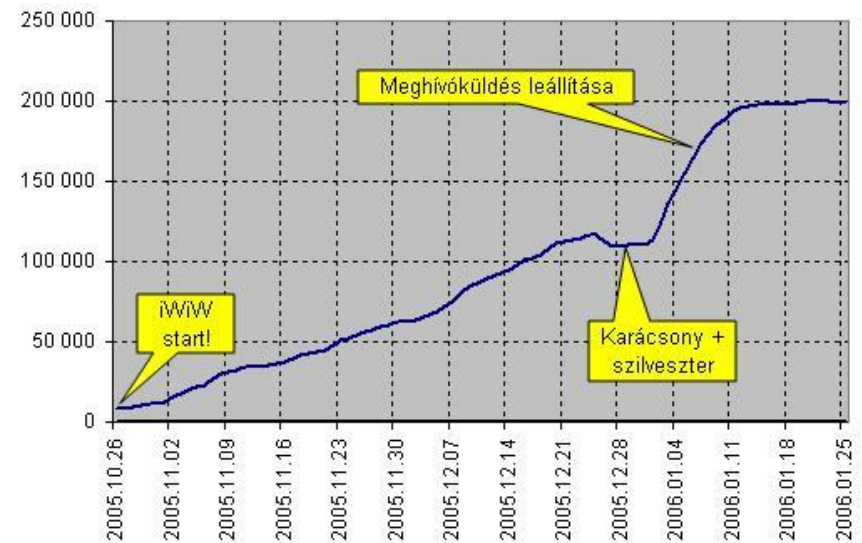
At the bottom right of the collage, there is a small Google logo and a message: "502. That's an error. The server encountered a temporary error and could not complete your request. Please try again in 30 seconds. That's all we know."

Valós terhelés példa (iwiw)

Napi regisztrációk (előző hét nap átlaga)

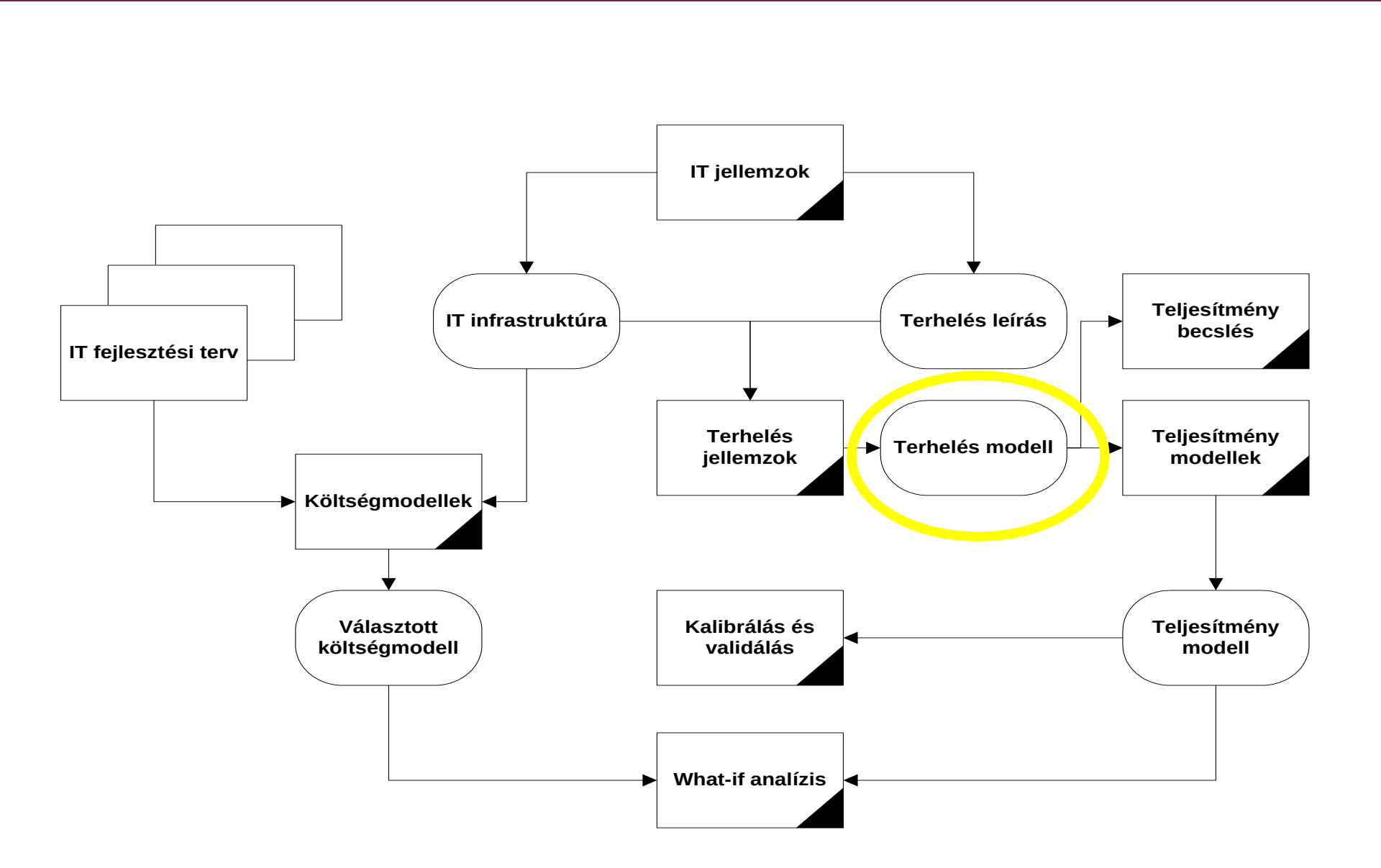


Napi egyedi látogatók (előző hét nap átlaga)



Forrás: <http://www.sg.hu/cikkek/42924/>

Terhelés modell



Terhelés modell meghatározása

- Adatgyűjtés
 - benchmarkok (részletesen Benchmark ea.)
 - ökölszabályok
 - „best practices”
 - mérések
- Monitorozás
 - belső (szerver)
 - külső (kliens, hálózat)
- Adatok rendszerezése
 - klaszter technikák

Terhelés modell minta

1 napi mérés adatai											
Érkezési ráta	20,000	kérés/óra	LAN 1	100	Mbps	LAN 2	100	Mbps	T3 Link	1.5	Mbps
Egy kérés szolgáltatás igénye											
E-business Funkció	Kérések száma	WS CPU (ms)	WS IO (ms)	AS CPU (ms)	AS IO (ms)	DB CPU (ms)	DB IO (ms)	LAN 1 (ms)	LAN 2 (ms)	T3 Link (ms)	Össz. szolg. igény
általános info	20,000	5.2	9.5	25.0	15.0	10.0	20.0	0.492	0.532	16.4	132.1
könyv keresés	18,900	4.8	8.5	18.0	14.0	13.0	40.0	0.328	0.352	12.0	91.0
könyv böngészés	14,120	4.9	8.2	13.0	12.0	13.0	40.0	0.287	0.328	12.0	83.7
bejelentkezés	8,020	5.1	8.4	12.0	10.0	13.0	20.0	0.295	0.492	11.5	80.8
regisztráció	892	32.0	15.0	16.0	30.0	15.0	20.0	0.655	0.000	32.8	126.4
bevásárlókosár	670	32.0	14.0	18.0	24.0	0.0	0.0	0.410	0.000	19.1	107.5
fizetés	584	31.0	15.0	35.0	90.0	30.0	80.0	0.819	0.901	43.7	326.4

Terhelés mérése ellenőrzött környezetben

Kliens gépen futó
tesztelő szkript

Dedikált szerver

Teljesítmény
monitorozás

Teljesítmény
monitorozás

DEDIKÁLT HÁLÓZAT

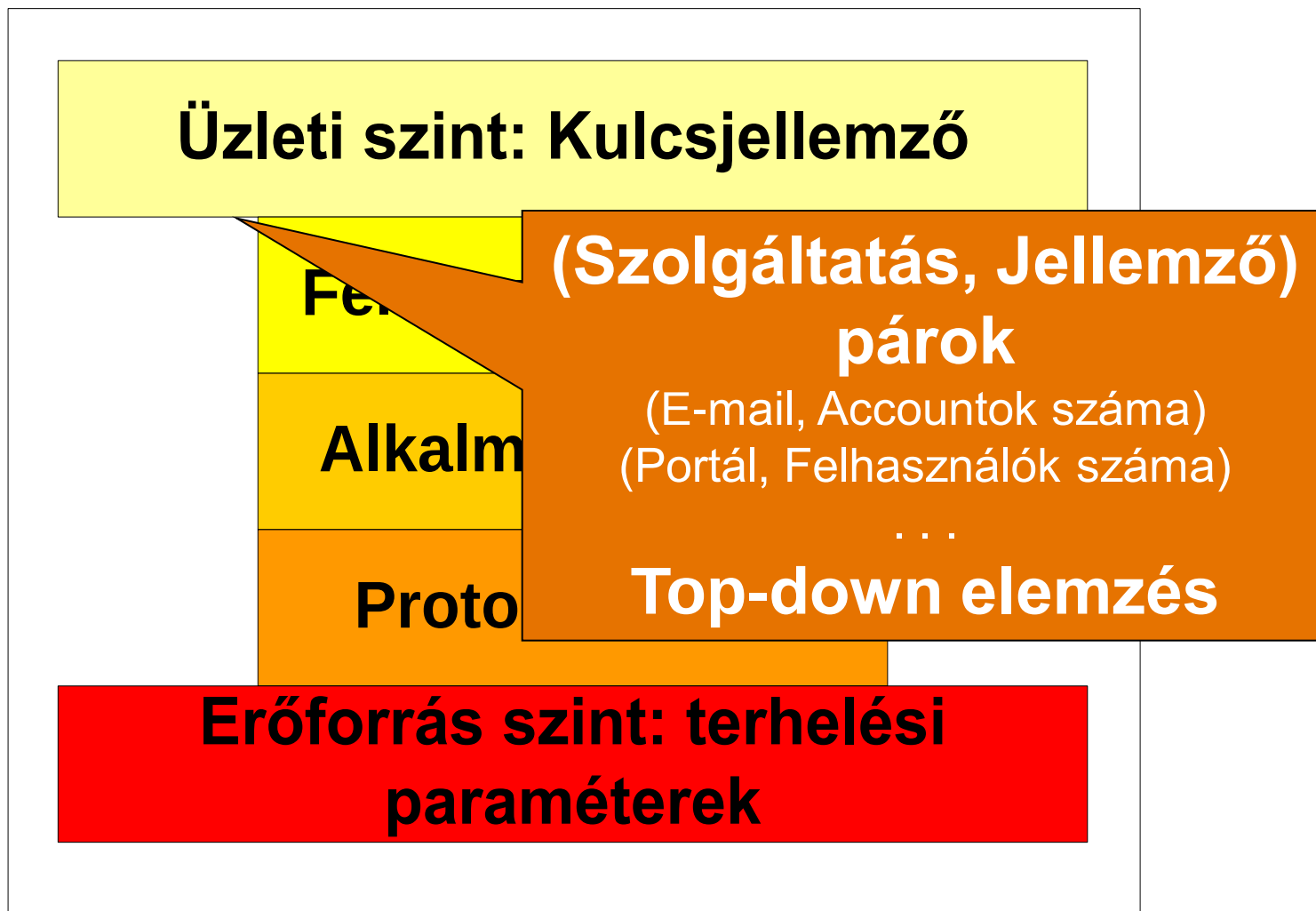
Terhelés előrejelzési stratégia



Hierarchikus előrejelzési modell

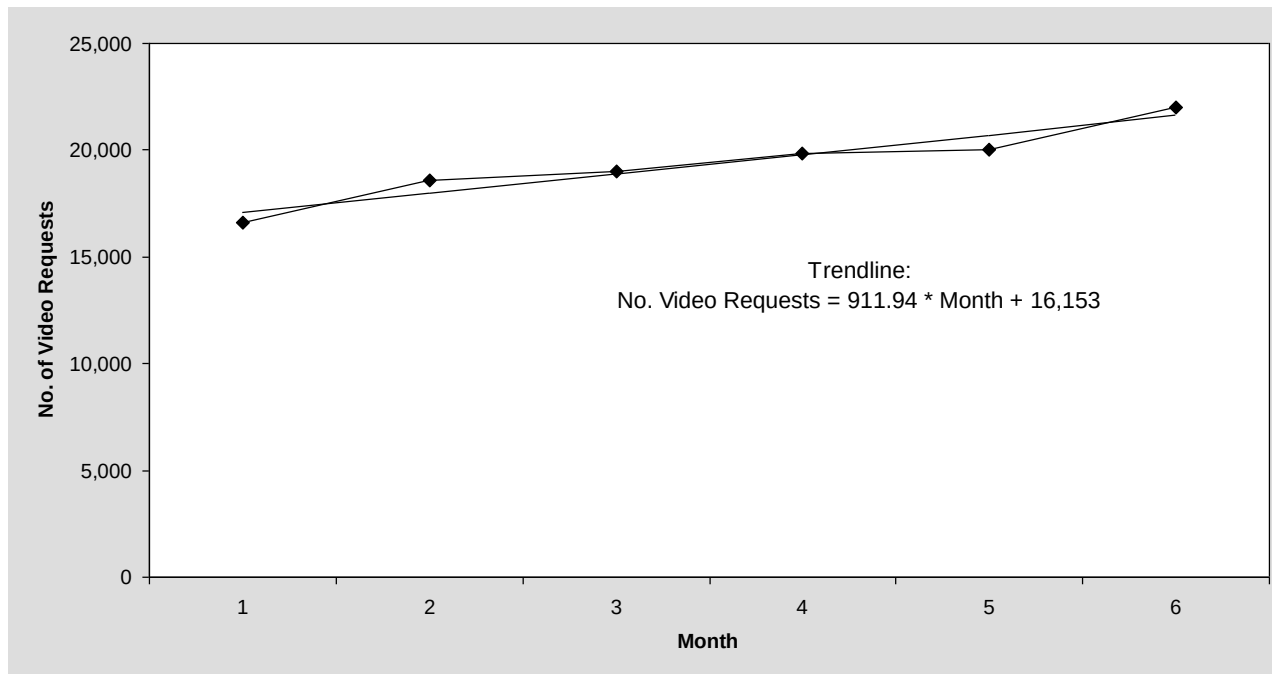


Hierarchikus előrejelzési modell



Terhelés előrejelzés

- A terhelési modell várható változása
- Többféle technika létezik
- Példa: lineáris regresszió



Terhelés típusai

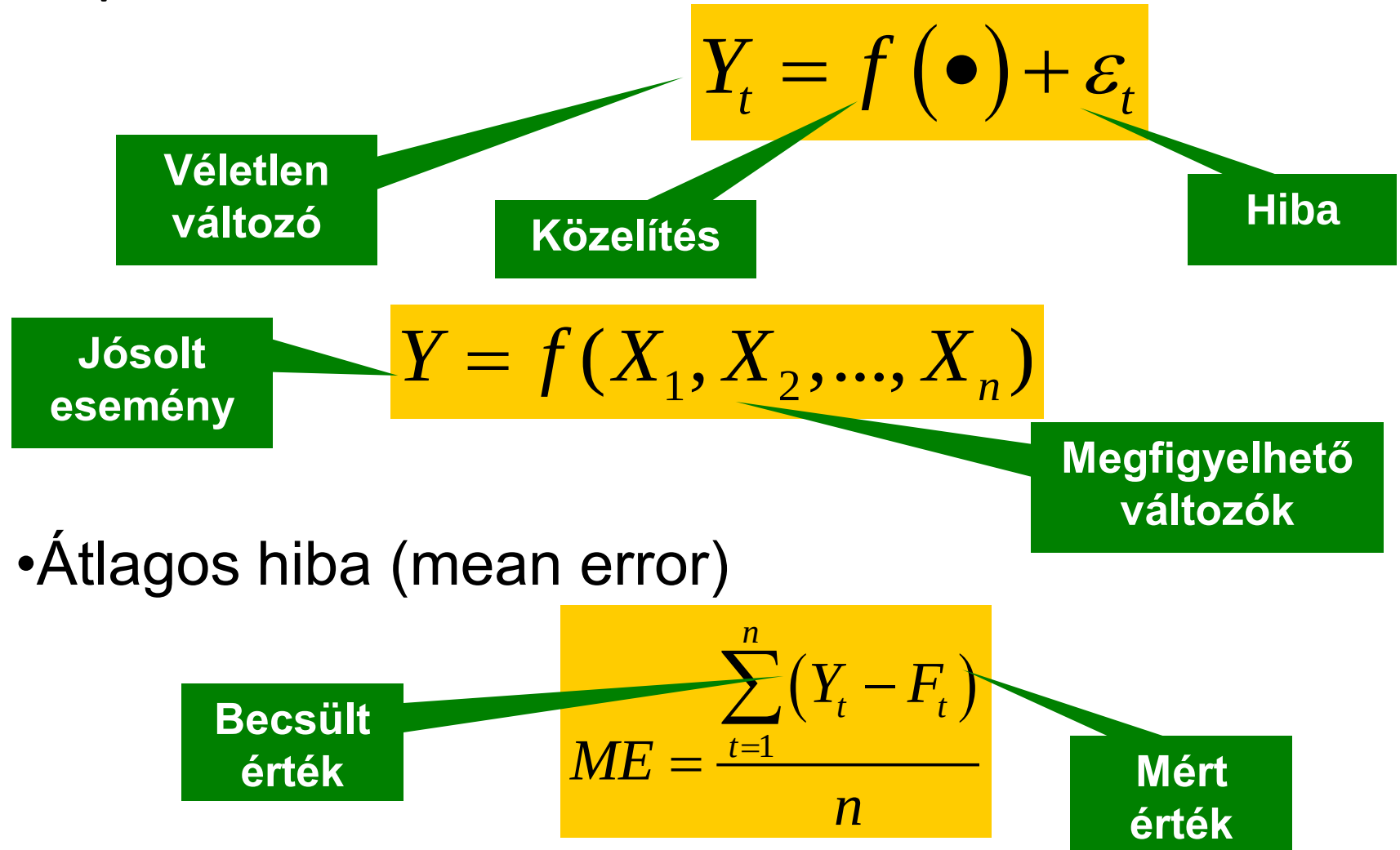
- Terhelési minta azonosítása
- Trend
 - Állandó jellegű változás, pl. folyamatos növekedés
- Időszakos
 - Visszatérő, periodikus hatások
- (Statikus)
- (Véletlenszerű)
- A mintához válasszunk előrejelzési módszer
 - Más területekről
 - Statisztika, meteorológia, tőzsde, stb.

Terhelés előrejelzési technikák

- Regresszió
 - Lineáris
 - Exponenciális
 - (Logaritmikus)
- Mozgó átlagok
- Exponenciális csúszóablak
- Könnyen automatizálhatók
 - Matlab
 - MS Excel
 - ...

Regressziós módszerek

- Alapelv:



Lineáris regresszió

- Egyszerű lin. függvény illesztése az adatokra
 - nem vár alapvető változást a rendszer viselkedésében

$$Y = a + bX$$

- Legkisebb négyzetek módszere
 - keressük azokat az a, b paramétereket, amelyekre

$$SSE = \sum_{t=1}^n \varepsilon_t^2 = \sum_{t=1}^n (Y_t - F_t)^2 \quad \text{minimális (Sum of Squared Errors)}$$

- cél:

$$\sum_{t=1}^n (Y_t - F_t)^2 = \sum_{t=1}^n [Y_t - (a + bX_t)]^2$$

Levezetés (parc. deriválás)

$$\frac{d \sum_{t=1}^n [Y_t - (a + bX_t)]^2}{da} = \sum_{t=1}^n (-2) [Y_t - (a + bX_t)] = 0$$

$$na = \sum_{t=1}^n (Y_t - bX_t)$$

$$a = \bar{Y} - b\bar{X}$$

**Xi, Yi a mért értékpárok
(pl. idő, terhelés)**

$$\frac{d \sum_{t=1}^n [Y_t - (a + bX_t)]^2}{db} = \sum_{t=1}^n X_t [Y_t - (a + bX_t)] = 0$$

$$\sum_{t=1}^n X_t \left[Y_t - \frac{1}{n} \sum_{t=1}^n (Y_t - bX_t) - bX_t \right] = \sum_{t=1}^n X_t Y_t - \frac{1}{n} \left(\sum_{t=1}^n X_t \right) \left(\sum_{t=1}^n Y_t \right) + \frac{1}{n} b \left(\sum_{t=1}^n X_t \right) \left(\sum_{t=1}^n X_t \right) - b \sum_{t=1}^n X_t^2 = 0$$

$$b = \frac{n \sum_{t=1}^n X_t Y_t - \left(\sum_{t=1}^n X_t \right) \left(\sum_{t=1}^n Y_t \right)}{n \sum_{t=1}^n X_t^2 - \left(\sum_{t=1}^n X_t \right)^2}$$

Lineáris regresszió (folyt.)

■ Korrelációs együttható (négyzete)

- a változó becsült és tényleges értékének kapcsolata
- 0 és 1 közti érték
- 0: nincs kapcsolat
- 1: függvényyszerű kapcsolat

$$R^2 = \frac{\sum_{t=1}^n (F_t - \bar{Y})^2}{\sum_{t=1}^n (Y_t - \bar{Y})^2}$$

■ Példa: E-mail szolgáltatás, 8 hétig mérjük a csúcsterhelést.

Hét	1	2	3	4	5	6	7	8
Max. terhelés (email/perc)	420	410	437	467	448	460	507	514

Hogyan közelíthető a terhelés változása?

Mekkora a korrelációs együttható?

Lin. regresszió példa

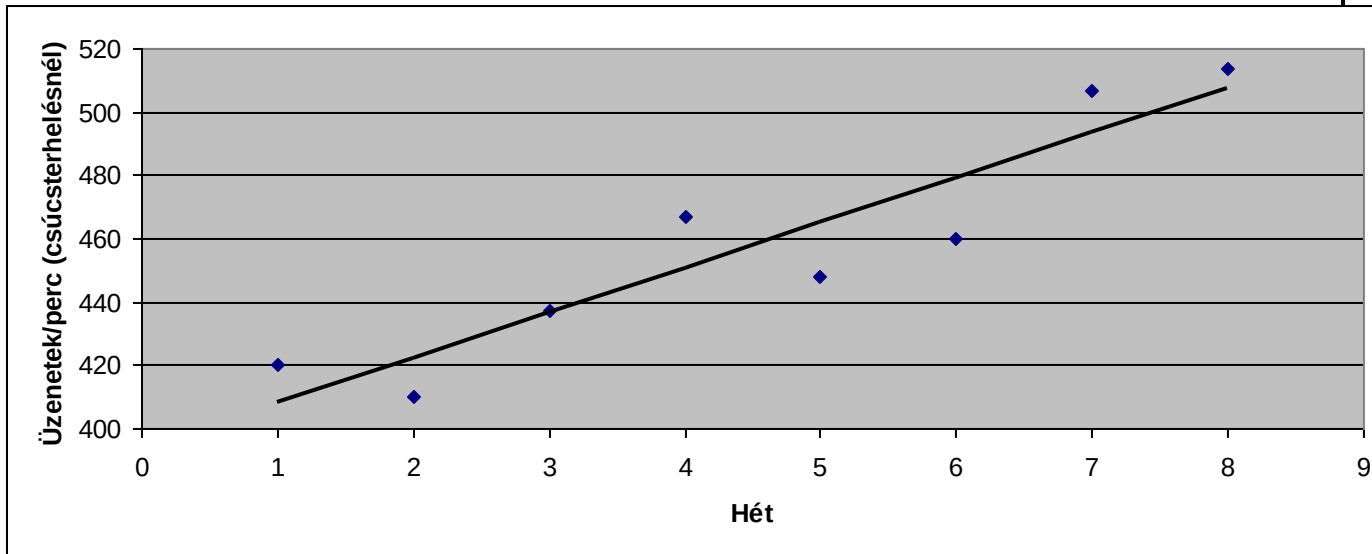
A legkisebb négyzetek módszerével

$$Y=393.98+14.20X$$

Korrelációs együttható:

$$R^2=0.855$$

Mért	Jósolt terhelés
420	408.18
410	422.38
437	436.58
467	450.78
448	464.98
460	479.18
507	493.38
514	507.58
	521.78



Két változó kapcsolatának vizsgálata

- Tfh. lineáris kapcsolat van az egyszerre bejelentkezett felhasználók száma és az elküldött emailek közt. (pl logok alapján)

Bejelentkezett felh. átlagos száma (1 óra alatt)	2450	2765	2241	2860	3011	2907	3209
Átl. terhelés (kimenő+bejövő emailek/óra)	19257	20488	18152	21450	21077	20639	22142

- Lineáris regressziós közelítés a legkisebb négyzetek módszerével:

ÜzenetekSzám = $f(\text{BejelentkezettFelhasználók})$

$Y=9480.48 + 3.95X$, $R^2=0.937$ erős kapcsolat

Nemlineáris módszerek

- Exponenciális közelítés:
 - jól illik a Web forgalom növekedéséhez

- Átalakítjuk a függvényt:

$$Y_t = a \times b^t$$

$$\log Y_t = \log a + t \log b$$

$$\log Y_t = Y', \log a = a', \log b = b'$$

$$Y' = a' + b't$$

- Legkisebb négyzetek módszere használható
- Pl. adottak a legnagyobb mért terhelés értékei

Mennyi a várható legnagyobb terhelés az év végén?

Hónap	1	2	3	4	5	6	7	8	9	10
Max. kérések/sec (Y_t)	1035	1100	1160	1250	1350	1555	1770	1950	2210	2630
$\ln(Y_t)$	6.942	7.003	7.056	7.13	7.207	7.349	7.478	7.575	7.7	7.874

Exp. terhelés példa

- Becslőfüggvény: $Y_t = a \times e^{bt}$
- Legkisebb négyzetek módszere a lineáris függvényre

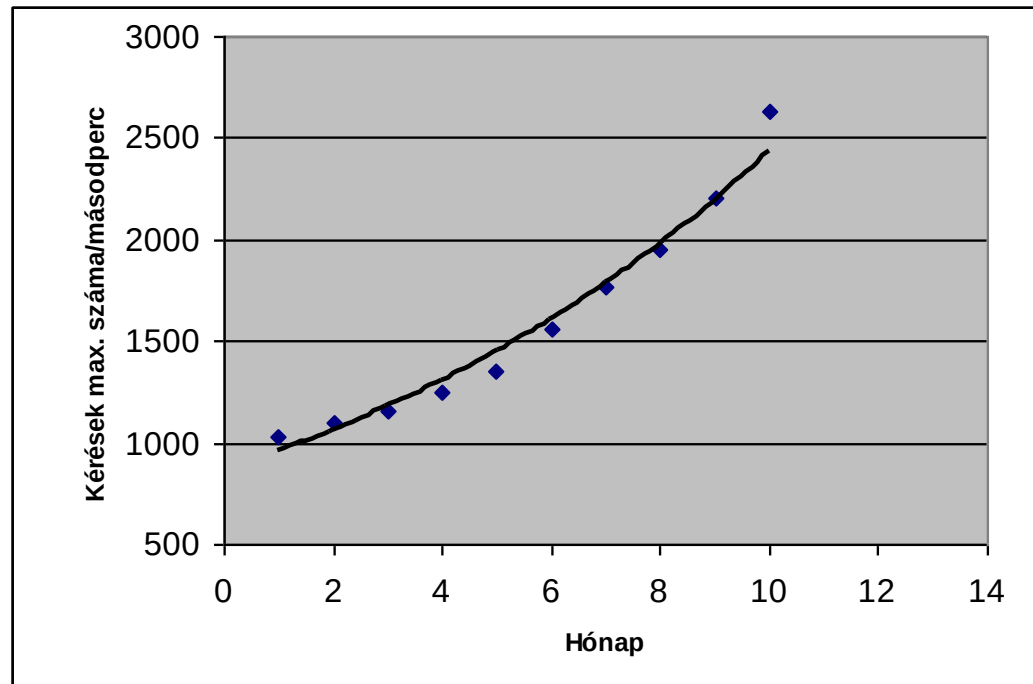
$$Y' = a' + b't, a' = 6.717, b' = 0.110, a = e^{a'}$$

- Eredmény:

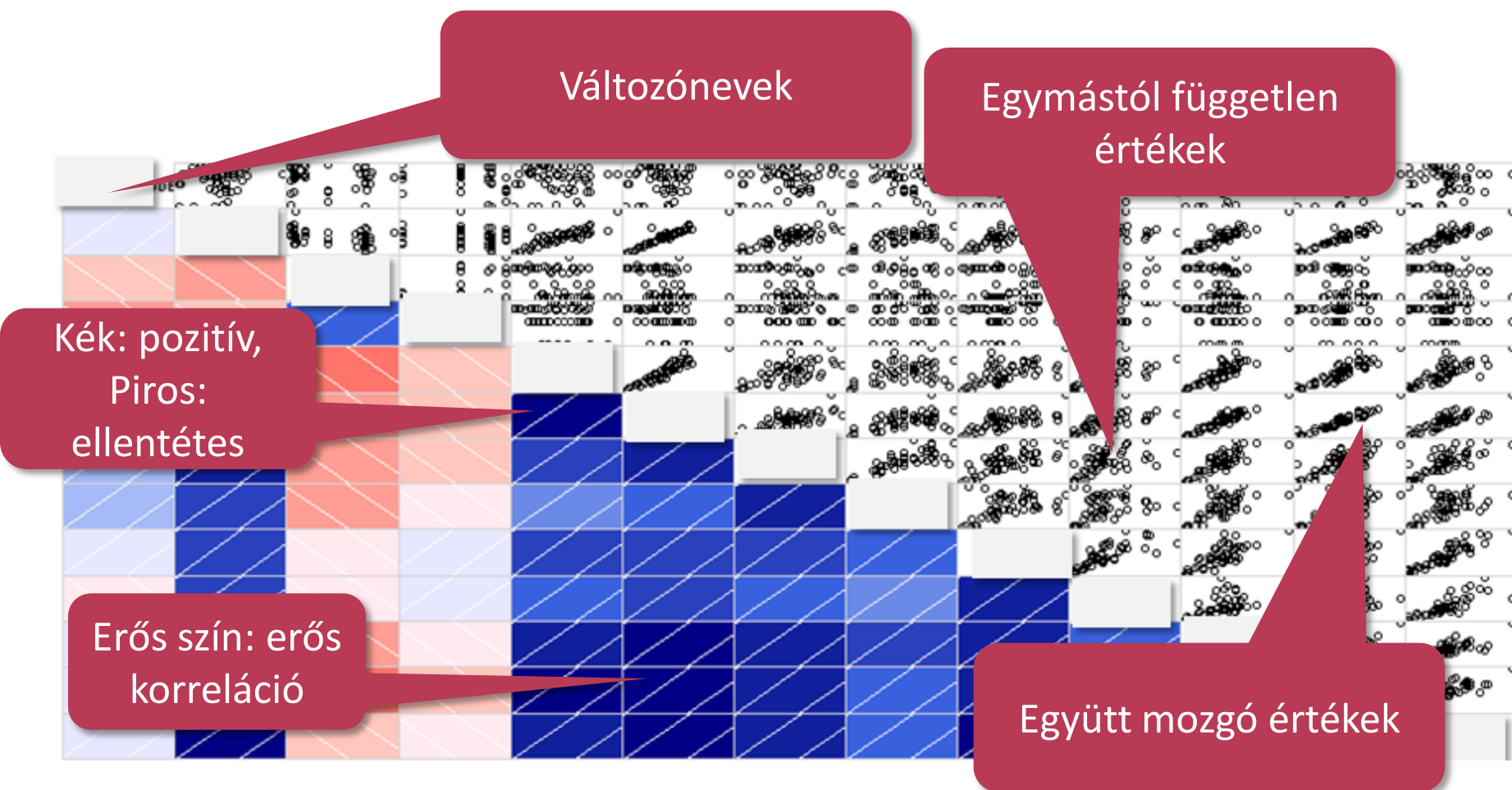
$$Y_t = 826.33 \times e^{0.11t}$$

- 12. hónap:

$$Y_t = 3093.3$$



Kitekintés: több érték páronkénti korrelációja



R statisztikai szoftver „corrgram” csomagjával előállítva.

Átló felett: scatterplot mátrix

Cél: együtt mozgó értékek kiszűrése, outlierok azonosítása.

→ Mik a terhelés/előrejelzés szempontjából lényeges változók?

Mozgó átlagok módszere

- Rövid távú előrejelzésre jó
- Egyszerre egy értéket ad meg
- A becsült érték az utolsó n érték átlaga

$$F_{t+1} = \frac{\sum_{i=t}^{t-n+1} Y_i}{n}$$

ahol Y_t a t . időpontban mért érték

F_{t+1} a becsült érték

n tipikusan 3 és 10 között van

(becslés hibája ne legyen túl nagy)

Exponenciális csúszóablak

- Egy értéket ad meg, az előző méréseket átlagolva
- Későbbi mérés nagyobb súllyal
- Súlyozza a mérési hibát
- Rövid távú előrejelzésre

$$F_{t+1} = F_t + \alpha(Y_t - F_t)$$

ahol F_t : a t. időpontra becsült érték

Y_t : t. időpontban mért érték

$Y_t - F_t$: mérési hiba a t. periódusban

α : súlyozás ($0 \leq \alpha \leq 1$, gyakorlatban $0,05 \leq \alpha \leq 0,3$)

A két módszer összehasonlítása

- Adott sávszélesség igények, a két módszerrel becsüljük a következő értéket

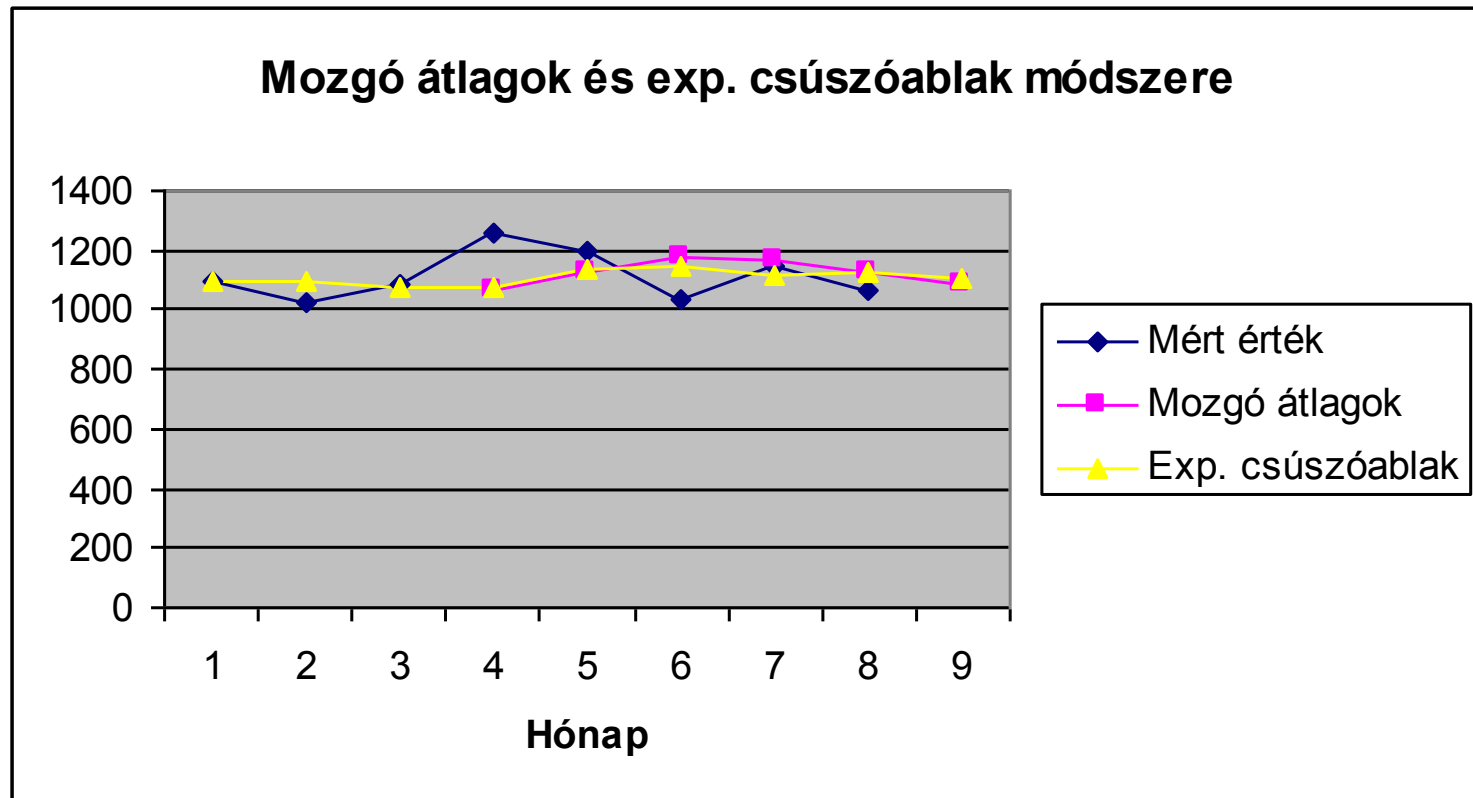
Hónap	Sávszélesség igény	Mozgó átlagok módszere (n=3)	Exp. csúszóablak ($\alpha = 0.3$)
1	1100		
2	1020		
3	1090		
4	1255		
5	1195		
6	1039		
7	1145		
8	1066		
9			

A két módszer összehasonlítása

- Adott sávszélesség igények, a két módszerrel becsüljük a következő értéket

Hónap	Sávszélesség igény	Mozgó átlagok módszere (n=3)	Exp. csúszóablak ($\alpha = 0.3$)
1	1100		1100.00
2	1020		1100.00
3	1090		1076.00
4	1255	1070.0000	1080.20
5	1195	1121.6667	1132.64
6	1039	1180.0000	1151.35
7	1145	1163.0000	1117.64
8	1066	1126.3333	1125.85
9		1083.3333	1107.90

A két módszer összehasonlítása



Terhelés karakterisztikái

- Webes terhelés alapjellemzői
 - Löketek (Burstiness)
 - Zipf törvénye teljesül
 - Kivétel: pl. file lekérések
 - „Heavy tailed” eloszlás (nem elhanyagolhatóak a ritka értékek)
- „Önhasonlóság” a löketeknél
 - 1 óra mintája $\sim$ 1 hét mintája
- Tervezésnél lényeges
 - Milyen gyakran vannak löketek?
 - Hogyan befolyásolja ez az áteresztőképességet?

Löketek karakterisztikája

λ_k : érkezési ráta a k . időszletben (n darab időszlet)

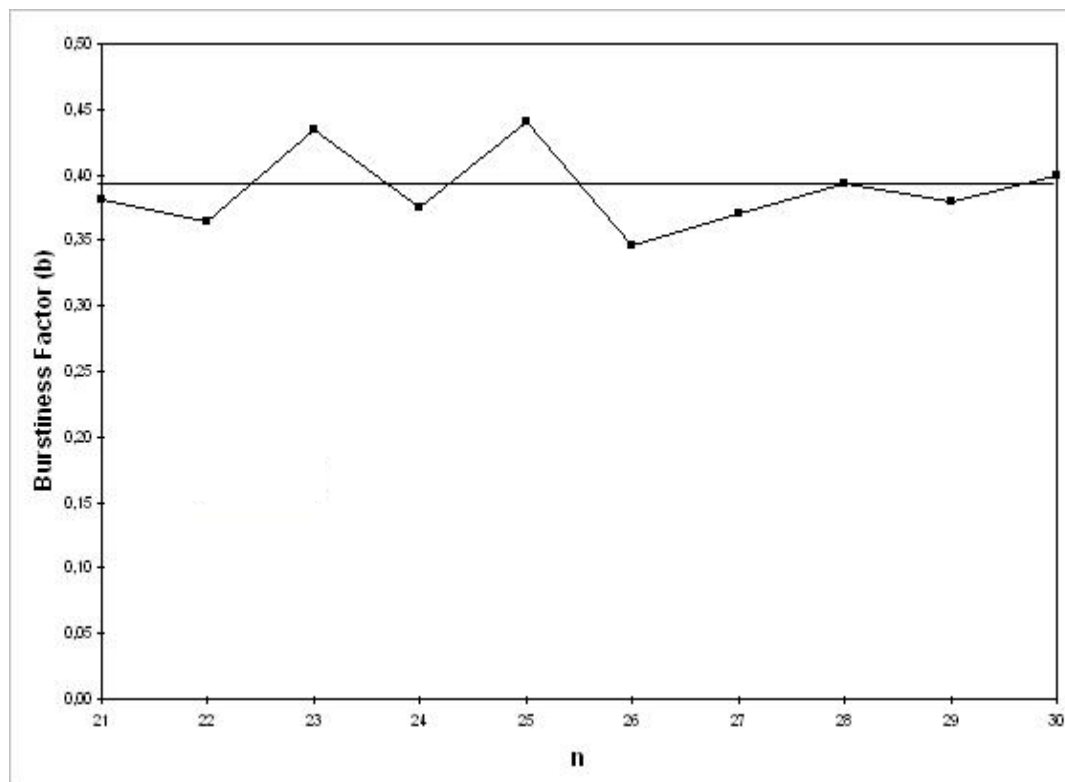
λ : érkezési ráta az egész időszakra

b : burstiness factor, a zsúfolt időszletek aránya

$$b = \frac{\sum_{i=1..n} \{0, \text{ ha } \lambda_i \leq \lambda; 1, \text{ ha } \lambda_i > \lambda\}}{n}$$

Burstiness factor

- Hogyan függ a burstiness factor a felbontás finomságától? (n)



A löketek hatása

- Hogyan függ az áteresztőképesség a burstiness factor-tól?

