Idősorok elemzése

Spektrálelemzés, DF és ADF tesztek

2017. november 14.

SPEKTRÁL-ELEMZÉS

Példa - BKV villamosenergia-terhelési görbéje

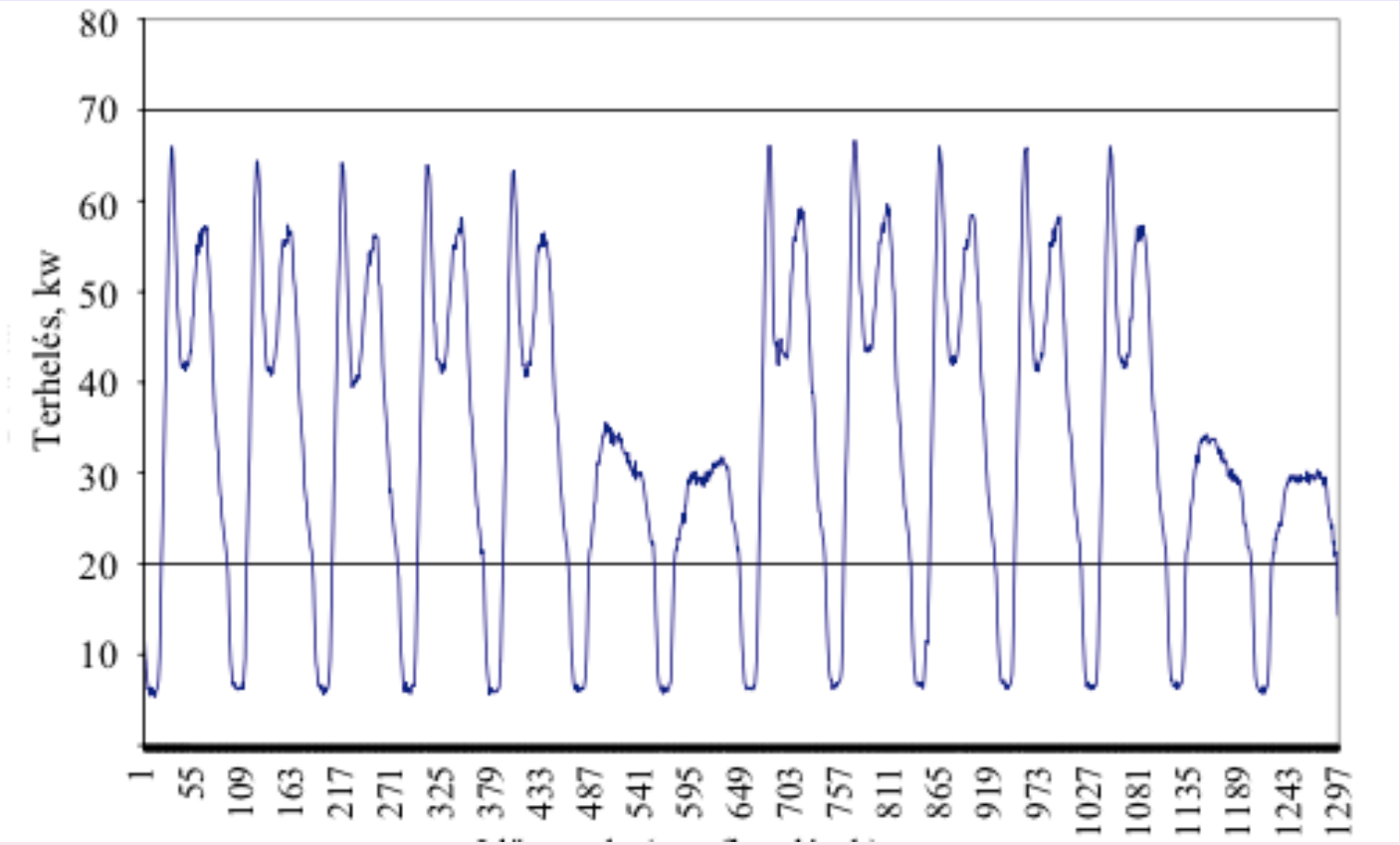


Figure: BKV villamosenergia-terhelési görbéje, negyedóránkénti mérések (2 hét adatai, ezer kw-ban mérve)

Az idősor alakja sugallja a napi periodicitás jelenlétét.

Példa - Fogyasztói árindexek

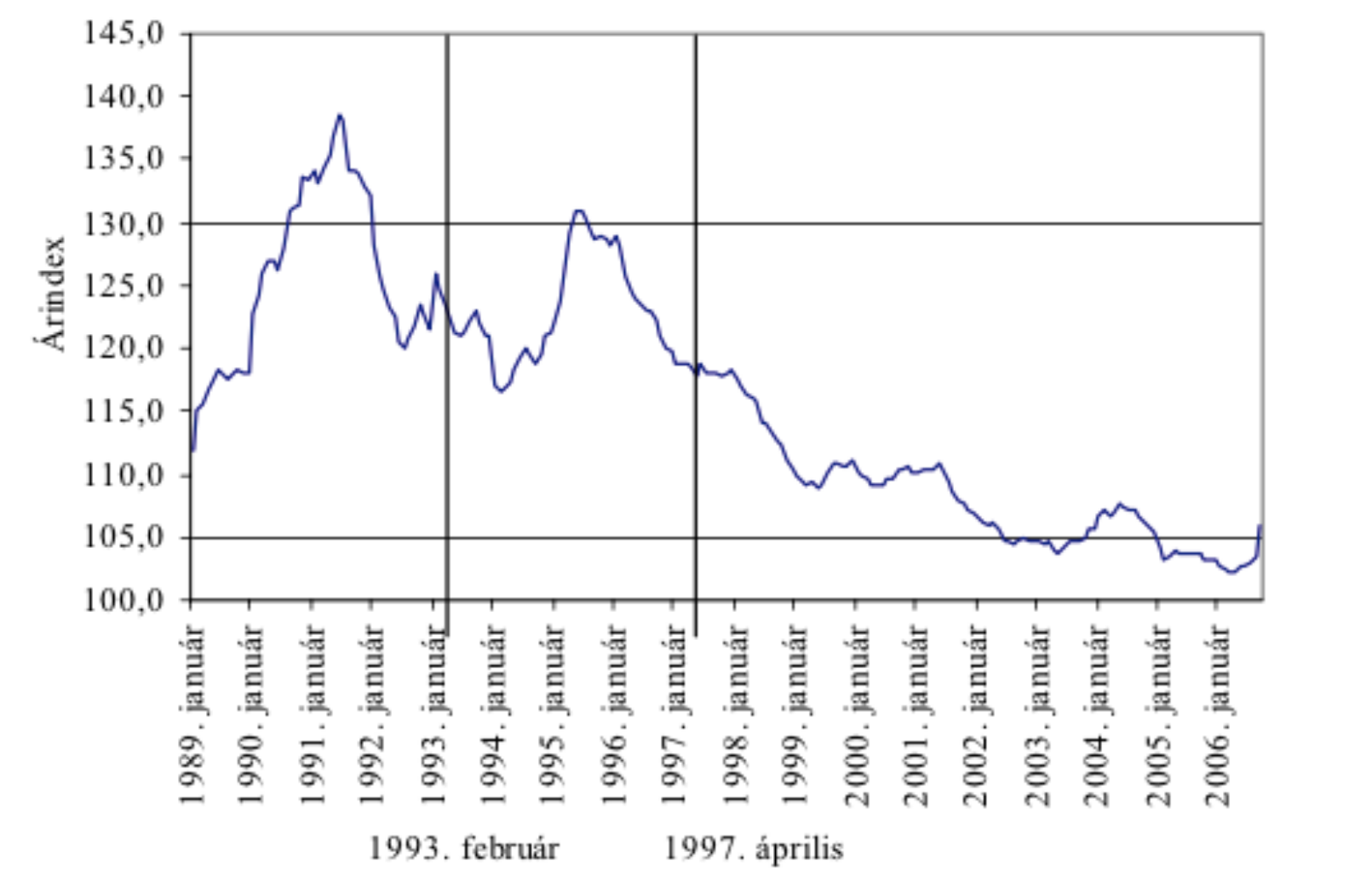


Figure: Fogyasztói árindexek

Spektrál-elemzés

A Wold-tételben meghatároztuk annak feltételét, hogy egy adott $(c_0, \ldots, c_q)$ sorozat mikor lehet egy q-adrendű MA folyamat autokovariancia függvénye. Ezt most megpróbáljuk általánosítani tetszőleges stacionárius folyamatra.

Tétel 1 (Herglotz tétel).

Legyen (c_k) egy valós számsorozat. Ekkor létezik olyan (x_t) stacionárius folyamat, melyre $c_k = \mathrm{E}(x_t x_{t+k})$ pontosan akkor, ha c_k előáll

$$c_k = \int_{-\pi}^{\pi} e^{ik\lambda} F(d\lambda)$$

alakban, ahol $i = \sqrt{-1}$ és $F(d\lambda)$ szimmetrikus mérték a $[-\pi, \pi]$ intervallumon.

A tételben szereplő F mértéket spektrális mértéknek nevezzük.

Spektrál-elemzés

Az $F(d\lambda)$ szimmetrikus mérték azt jelenti, hogy minden $A \subset [-\pi, \pi]$ mérhető halmazhoz tartozik egy véges $F(A) \geq 0$ szám, melyre

$$F(A \cup B) = F(A) + F(B), \quad \text{ha } A \cap B \neq \emptyset,$$
$$F(A) = F(-A).$$

Ekkor az integrált közelítő összegekkel definiáljuk az

$$\int_{-\pi}^{\pi} e^{ik\lambda} F(d\lambda) = \lim_{\substack{n \to \infty \\ \max(\delta_j) \to 0}} \sum_{j=1}^{n} e^{ik\lambda_j} F([\lambda_{j-1}, \lambda_j])$$

formában, ahol $-\pi = \lambda_0 < \lambda_1 < \ldots < \lambda_n = \pi$ az intervallum felosztása, $\delta_j = \lambda_j - \lambda_{j-1}$ pedig a j-dik részintervallum nagysága.

1. eset: F abszolút folytonos a Lebesgue mértékre.

Ekkor tetszőleges $A \subset [-\pi, \pi]$ mérhető halmaz esetén létezik egy $f(\lambda)$ integrálható függvény olyan, melyre

$$F(A) = \int_A f(\lambda) d\lambda.$$

Ezt hívjuk spektrális sűrűségfüggvénynek. Ekkor

$$c_k = \int_{-\pi}^{\pi} e^{ik\lambda} f(\lambda) d\lambda,$$

ami éppen f Fourier-transzformáltja (konstanstól eltekintve). Ha egy idősor autokovarianciái a fenti alakban előállíthatók, akkor azt mondjuk, hogy folytonos spektruma van.

Spektrál-elemzés

Visszafelé elvégezve a transzformációt adódik, hogy

$$f(\lambda) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} c_k e^{-ik\lambda},$$

ahonnan a $c_k = c_{-k}$ és $e^{ik\lambda} = \cos(k\lambda) + i\sin(k\lambda)$ tulajdonságok alapján

$$f(\lambda) = \frac{1}{2\pi} \left(c_0 + 2 \sum_{k=1}^{\infty} c_k \cos(k\lambda) \right).$$

Példák:

- ha (ε_t) fehérzaj, akkor $f(\lambda) = \frac{\sigma^2}{2\pi}$
- ha (x_t) MA(1) folyamat, akkor $f(\lambda) = \frac{1}{2\pi}|\beta_0 + \beta_1 e^{-i\lambda}|^2$
- ha (x_t) AR(1) folyamat, akkor $f(\lambda) = \frac{c_0}{2\pi}\left|\frac{1}{1-\alpha e^{-i\lambda}}\right|^2$

2. eset: F csak néhány pontra koncentrált mérték.

Legyenek ezek a pontok a $\lambda_j \in [-\pi, \pi]$, $\lambda_j = \lambda_{-j}$, $j \in \mathbb{N}$ pontok, a σ_j^2 súlyokkal. Ekkor

$$F(A) = \sum_{\lambda_j \in A} \sigma_j^2.$$

Ekkor azt mondjuk, hogy az idősornak diszkrét spektruma van. Ezt felhasználva a (c_k) sorozatra azt kapjuk, hogy

$$c_k = \sum_{j=-\infty}^{\infty} \sigma_j^2 e^{ik\lambda_j} = \sigma_0^2 + 2\sum_{j=1}^{\infty} \sigma_j^2 \cos(k\lambda_j)$$

az F mérték szimmetria tulajdonsága miatt.

Spektrál-elemzés

Természetesen felmerül a kérdés, hogy van-e olyan folyamat, melynek a fenti alakú diszkrét spektruma van.

Tétel 2.

Ha idősorunk előáll az

$$x_t = A_0 + \sum_{j=1}^{\infty} (A_j \cos(\lambda_j t) + B_j \sin(\lambda_j t))$$

alakban, ahol (A_j, B_j) *független,* 0 *várható értékű és* $\sigma_j\sqrt{2}$ *szórású v.v-k, akkor* (x_t) *autokovariancia függvénye éppen a fenti alakú.*

Azaz a diszkrét spektrum azt jelenti, hogy idősorunk véletlen periódusok összegeként áll elő.

Spektrál-elemzés

Egyszerűsítsük tovább a modellt: tegyük, fel hogy

$$x_t = a_0 + \sum_{j=1}^{m} (a_j \cos(\lambda_j t) + b_j \sin(\lambda_j t)) + e_t$$

alakú, ahol a_k, b_k, λ_k konstansok, (e_t) pedig fehérzaj. Ezek az ún. rejtett periódusú idősorok $\lambda_1, \ldots, \lambda_m$ rejtett periódusokkal.

Legyen a fenti idősor egy véges megfigyelése $(x_1, \ldots, x_N)$, és tegyük fel, hogy m és $\lambda_1, \ldots, \lambda_m$ ismertek. Feladatunk az ismeretlen (a_k, b_k) együtthatók becslése a minta alapján.
LS-becslés:

$$\min_{a_k, b_k} \left(\sum_{t=1}^{N} (x_t - a_k \cos(\lambda_k t) - b_k \sin(\lambda_k t))^2 \right) \quad k = 1, \ldots, m.$$

Spektrál-elemzés

Egyszerű számolással adódik, hogy

$$\hat{a}_j = \frac{2}{N}\sum_{t=1}^{N} x_t \cos(\lambda_j t) \quad \text{és} \quad \hat{b}_j = \frac{2}{N}\sum_{t=1}^{N} x_t \sin(\lambda_j t)$$

Probléma: a fenti λ_j frekvenciák általában nem ismertek!

Definíció 1.

Minden $\lambda \in [-\pi, \pi]$ értékre legyen az $I_N(\lambda) = (a(\lambda))^2 + (b(\lambda))^2$ függvény az ún. periodogram függvény, ahol

$$a(\lambda) = \sqrt{\frac{2}{N}}\sum_{t=1}^{N} x_t \cos(\lambda t) \text{ és } b(\lambda) = \sqrt{\frac{2}{N}}\sum_{t=1}^{N} x_t \sin(\lambda t).$$ *Azaz*

$$I_N(\lambda) = \frac{2}{N}\left|\sum_{t=1}^{N} x_t e^{-it\lambda}\right|.$$

Tétel 3.

Tetszőleges $\lambda \in [-\pi, \pi]$ esetén teljesül, hogy

$$I_N(\lambda) = 2 \sum_{k=0}^{N-1} \hat{c}_k \cos(\lambda k),$$

ahol

$$\hat{c}_k = \frac{1}{N} \sum_{t=1}^{N-k} x_t x_{t+k}.$$

Ha a periodogramot éppen a rejtett periódushoz tartozó pontban számoljuk ki, akkor

$$A(\lambda_k) = \sqrt{\frac{N}{2}}\, \hat{a}_k, \quad \text{és} \quad B(\lambda_k) = \sqrt{\frac{N}{2}}\, \hat{b}_k$$

adódik.

Tétel 4.

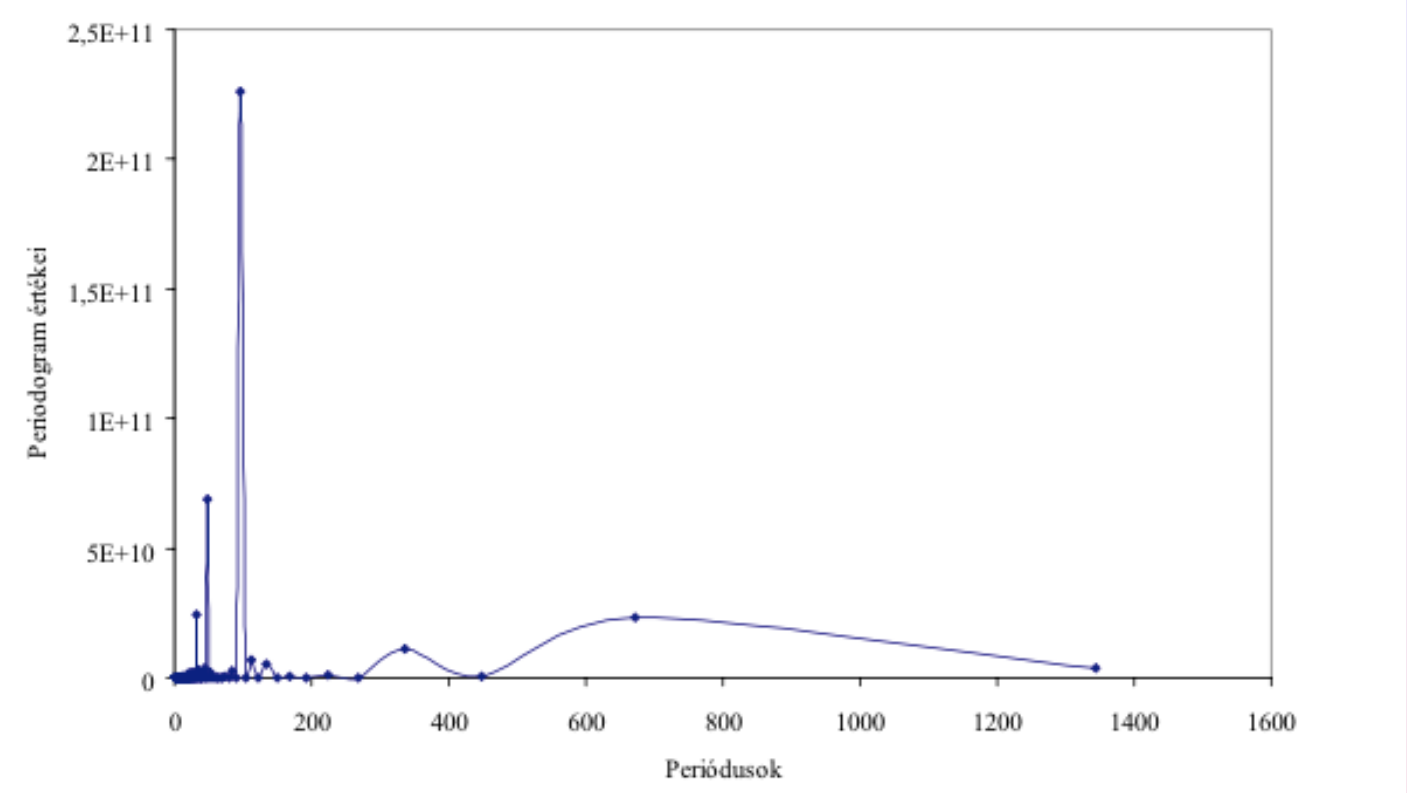
Tegyük fel, hogy a mintabeli megfigyelések száma végtelen tart. Ekkor a periodogram határértékére

$$\lim_{N\to\infty} I_N(\lambda) = \begin{cases} 0, & \lambda \neq \lambda_k \ k = 1, \dots, m \\ +\infty, & \lambda = \lambda_k \ \textit{valamilyen } k\textit{-ra} \end{cases}$$

adódik. A fenti konvergencia sztochasztikus értelemben értendő.

Numerikus számításoknál a periodogramot a $\lambda = 0, \frac{2\pi}{N}, \frac{4\pi}{N}, \dots$ ún. Fourier-frekvenciákon értékelik ki. Nagy mintaelemszám esetén a periodogram kiugró értékei nagy valószínűséggel rejtett frekvenciákat jelentenek. Ezek tesztelesére számos próba áll rendelkezésre (pl. Fisher-féle g-próba, Whittle-féle módosított g-próba).

Példa - BKV adatok spektruma



A numerikus adatok alapján megállapítható, hogy a becsült sűrűségfüggvény a 96-os periódusnál (periódus$= 2\pi/\lambda_k$) veszi fel a maximumát, ami a negyedórás megfigyelések miatt egy napos ciklusnak felel meg ($24 \cdot 4 = 96$).

Példa - BKV villamosenergia-terhelési görbéje

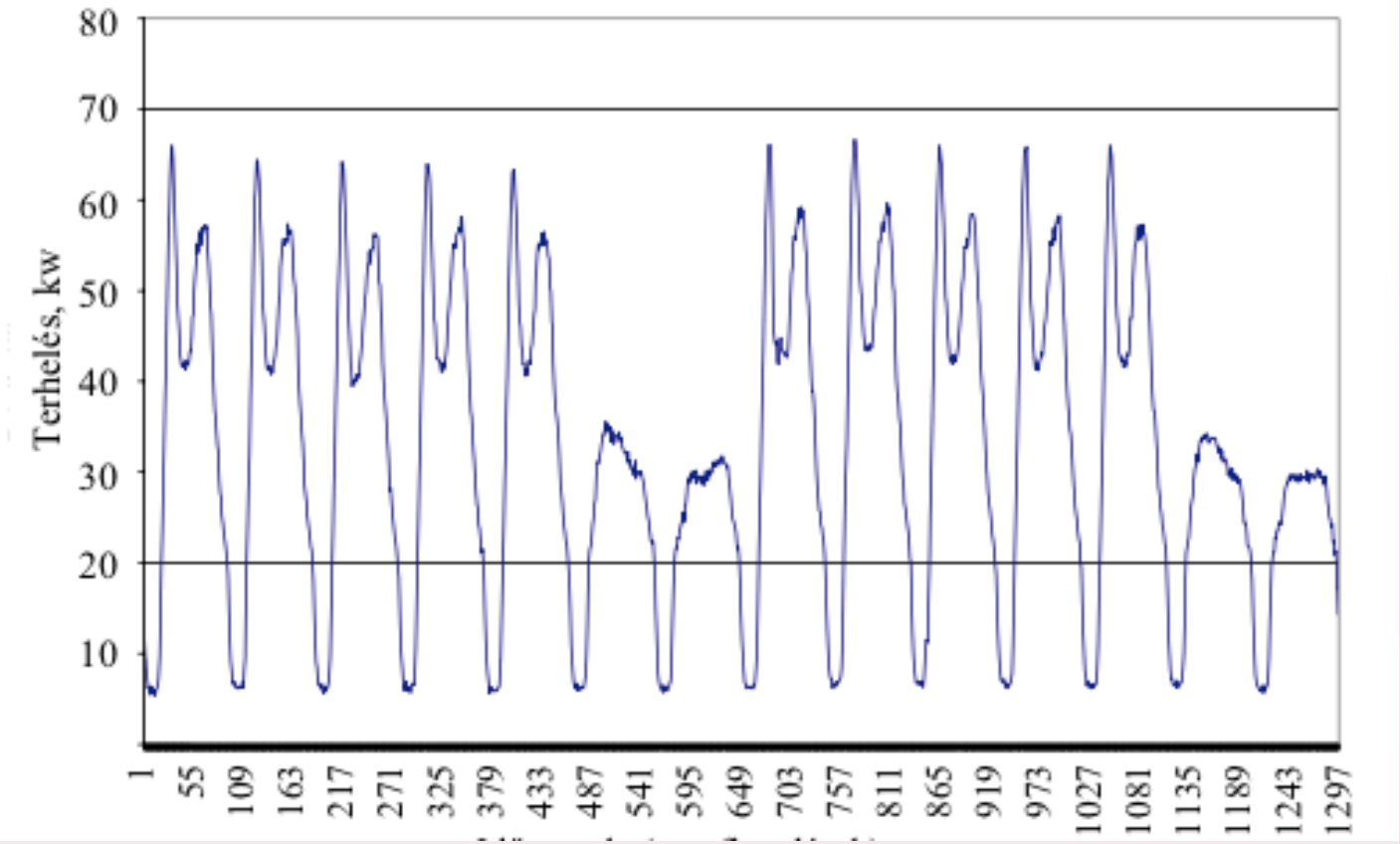
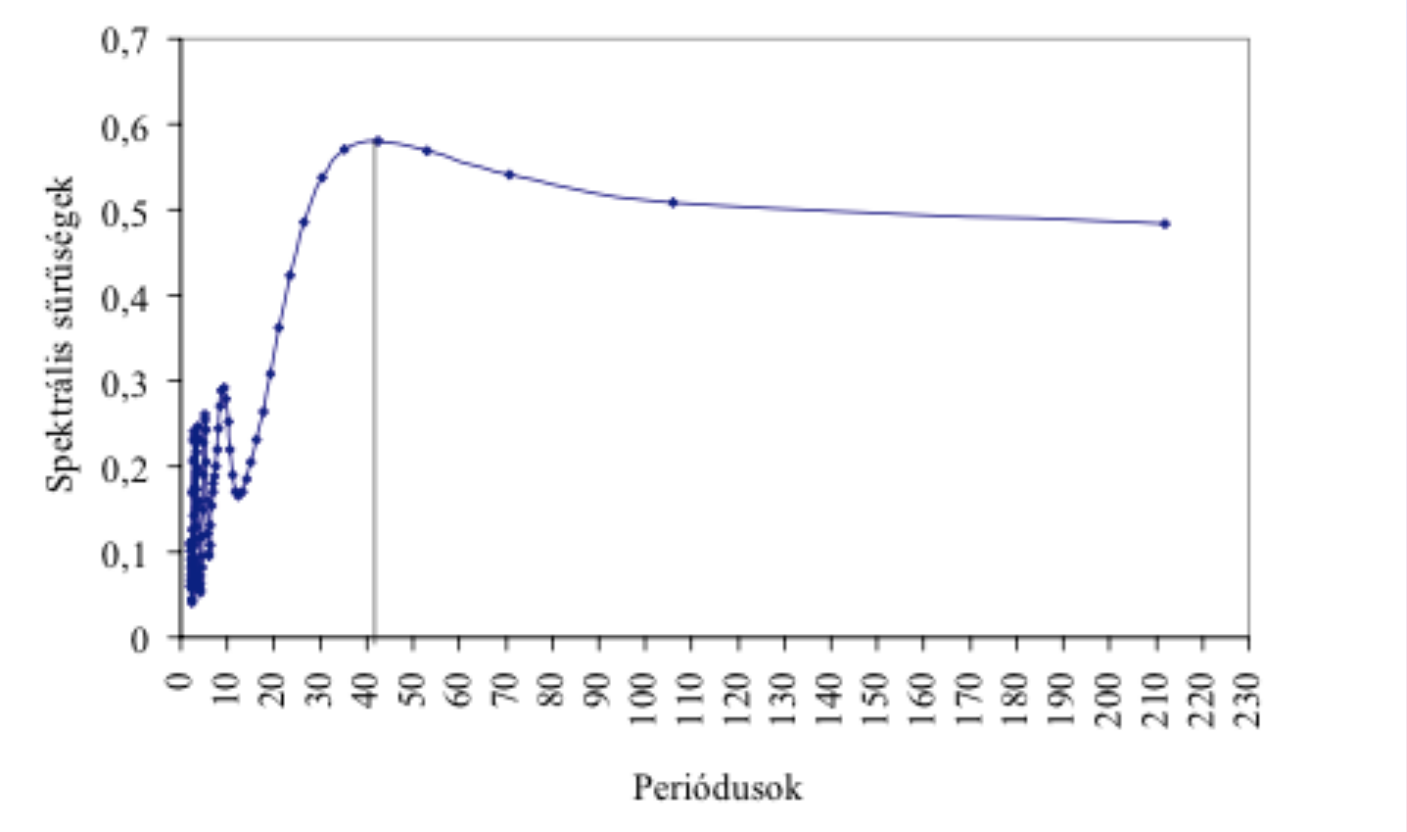


Figure: BKV villamosenergia-terhelési görbéje, negyedóránkénti mérések (2 hét adatai, ezer kw-ban mérve)

Példa - fogyasztói árindexek spektruma



A maximumot a $42,5$ hosszú perióduson veszi fel, ami kb. $3,5$ éves ciklusnak felel meg.

Példa - Fogyasztói árindexek

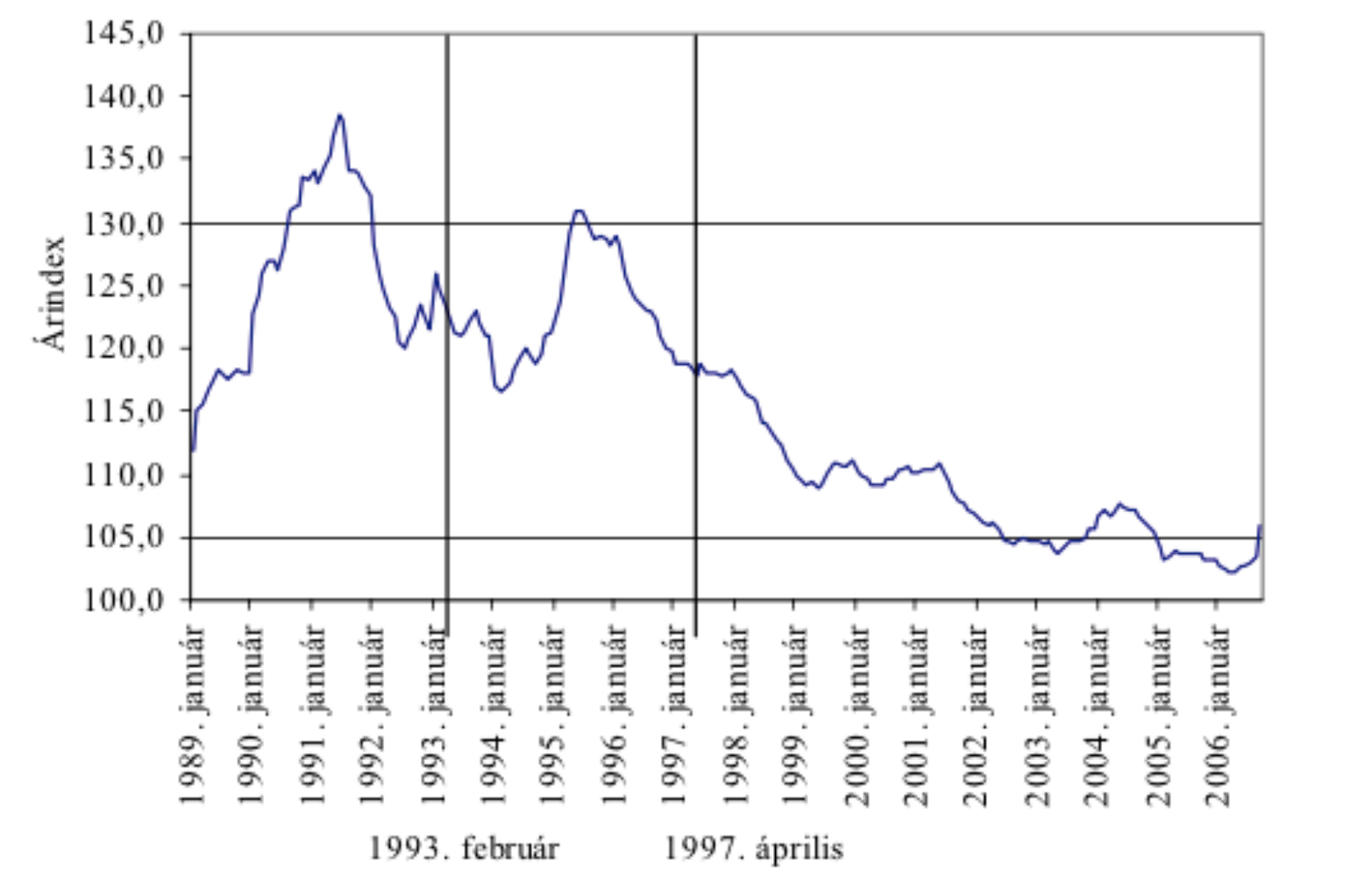


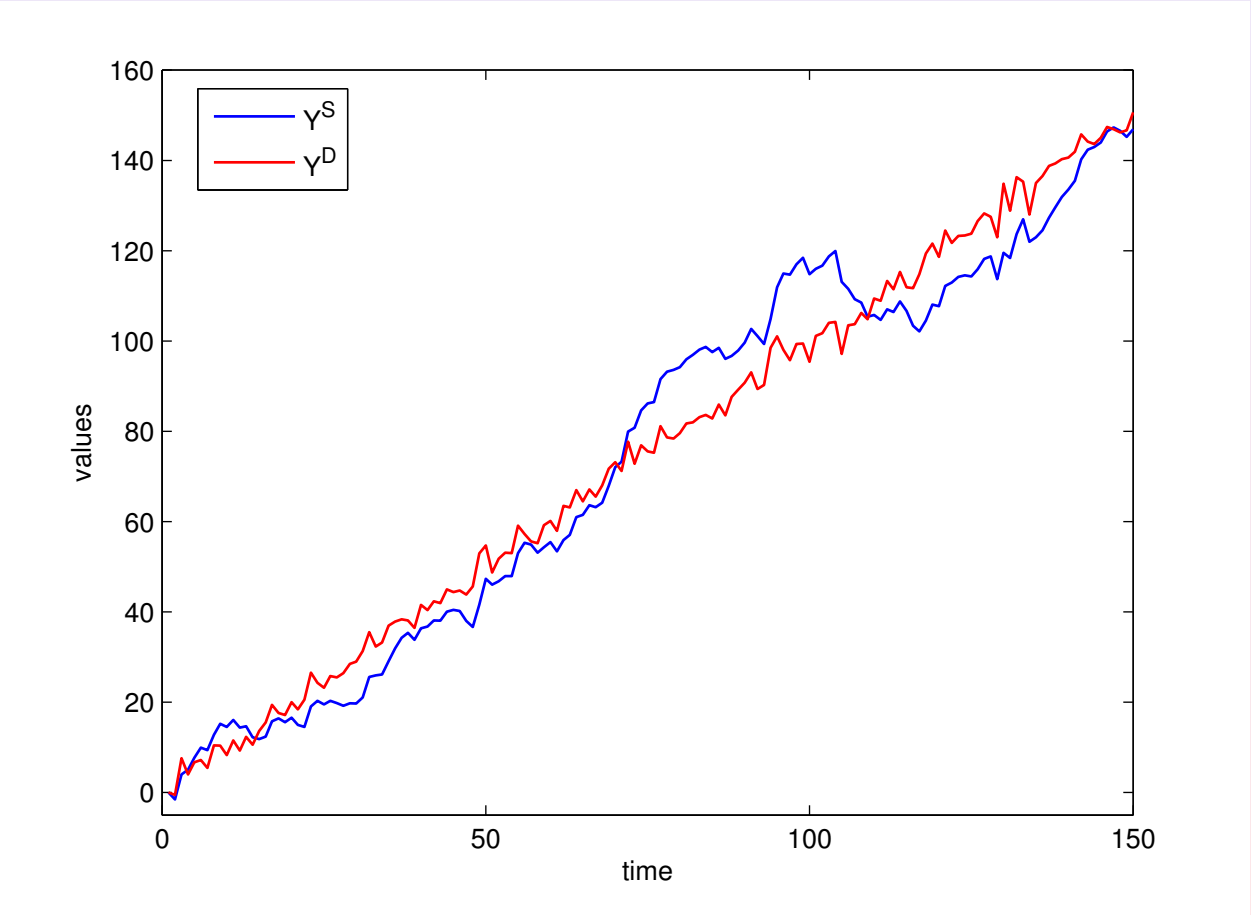
Figure: Fogyasztói árindexek

STACIONARITÁSI TESZTEK

Stacionaritási tesztek

Nem stacionárius folyamatok két alapesete:

- determinisztikus trendet tartalmazó idősorok (trend-stac.)
- egységgyök (unit root) folyamatok (differencia-stacionárius)



Példa

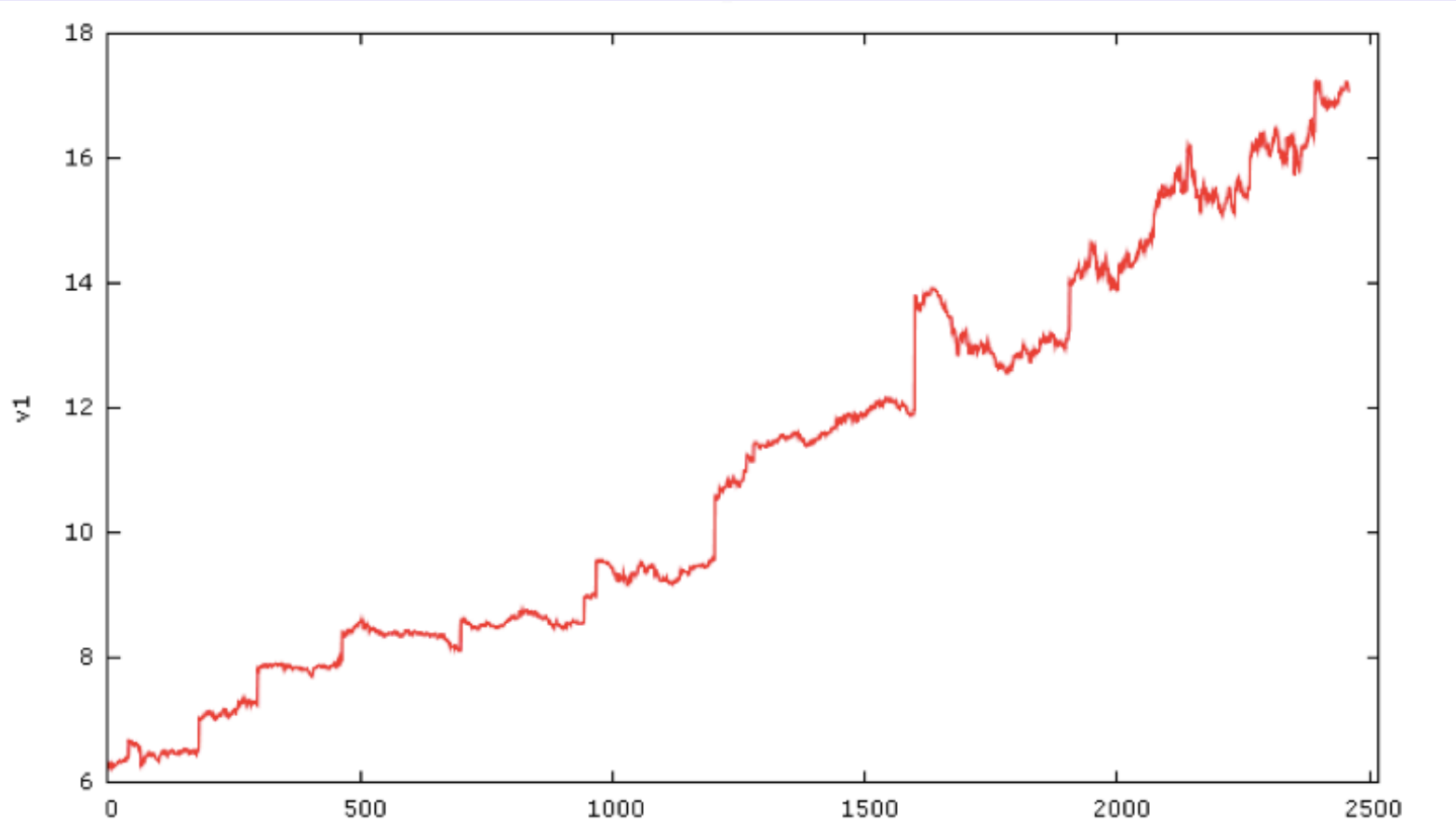


Figure: Francia frank napi árfolyama 1986 és 1993 között, 2459 megfigyelés

Stacionaritási tesztek

Tekintsük az alábbi modellt:

$$y_t = \alpha + \delta t + u_t,$$

ahol u_t ARMA folyamat, azaz

$$\underbrace{(1 - \alpha_1 z - \ldots - \alpha_p z^p)}_{A(Z)} u_t = \underbrace{(1 + \beta_1 z + \ldots + \beta_q z^q)}_{B(z)} \varepsilon_t.$$

Tegyük fel, hogy B invertálható, és legyen

$$A(z) = (1 - \lambda_1 z)(1 - \lambda_2 z) \ldots (1 - \lambda_p z).$$

Ekkor, ha $\lambda_1, \ldots, \lambda_p$ az egységkörön belül helyezkednek el, akkor

$$u_t = \frac{B(z)}{A(z)} \varepsilon_t = \Psi(z) \varepsilon_t$$

jóldefiniált, és $\Psi(z)$ gyökei az egységkörön kívülre esnek. Azaz y_t trend-stacionárius folyamat lesz.

Stacionaritási tesztek

Tegyük fel, hogy $\lambda_1 = 1$. Ekkor

$$A(z)u_t = (1-z)(1-\lambda_2 z)\dots(1-\lambda_p z)u_t = B(z)\varepsilon_t,$$

azaz

$$(1-z)u_t = \frac{B(z)}{(1-\lambda_2 z)\dots(1-\lambda_p z)}\varepsilon_t = \Psi^*(z)\varepsilon_t,$$

ahol Ψ^* gyökei az egységkörön kívülre esnek. Ekkor

$$\begin{aligned} y_t - y_{t-1} &= (1-z)y_t = (1-z)\alpha + \delta t - \delta(t-1) + (1-z)u_t = \\ &= 0 + \delta + \Psi^*(z)\varepsilon_t \end{aligned}$$

ún. egységgyök folyamatot kapunk.

Stacionaritási tesztek

Tehát

- ha $A(z)$-nek van 1 darab egységgyöke, a többi gyök pedig az egységkörön kívül van, akkor egységgyök folyamatunk lesz,
- ha minden gyöke az egységkörön kívülre esik, akkor trend-stacionárius folyamatunk van.

Több egységgyök esetén a fenti gondolatmenetet ismételni kell az egységgyökök multiplicitásának megfelelően.

Ezek lesznek az ún. ARIMA(p, d, q) folyamatok, ahol d az egységgyökök számát jelöli. Az ilyen folyamatok d-edrendbeli differenciázása vezet stacionárius folyamathoz.

Stacionaritási tesztek

Hogyan kell ezeknél a folyamatoknál a trendet eltávolítani?

- Trend-stacionárius esetben egy egyszerű kivonás (osztás) megoldja a problémát
- Egységgyök folyamat esetén ez nem jó megoldás (ld. például az eltolásos véletlen bolyongást), mert ekkor a kivonás/osztás nem szünteti meg a reziduumok szórásának időbeli változását. Emiatt használjuk ezekben az esetekben a differenciázást.
- Trend-stacionárius esetben viszont a differenciázás nem jó megoldás, mert ekkor egységgyök kerülne az MA részbe!

Stacionaritási tesztek

Tehát minden egyes idősornál alapvetően fontos kérdés, hogy vajon a vizsgált folyamat tartalmaz-e egységgyököt vagy sem.

Viszont, az egységgyök tesztek tárgyalása előtt figyelembe kell vennünk az alábbiakat:

- véges minták állnak csak rendelkezésünkre, így nem lehet megállapítani, hogy a folyamat tényleg a fenti, precízen definiált, általános egységgyök folyamat, vagy pedig egy stacionárius folyamat;
- csak azt lehet tesztelni, hogy előre meghatározott rendű autoregresszív folyamatok tartalmaznak-e egységgyököt, vagy sem;
- azaz tesztelésre alkalmas hipotézis akkor fogalmazható meg, ha előre korlátozott rendű autoregresszív folyamatok alkotják a vizsgálat tárgyát.

Stacionaritási tesztek

Dickey-Fuller teszt: autokorrelálatlan reziduumok esetén alkalmazható csak. A becsült regresszió

$$y_t = \alpha + \rho y_{t-1} + \delta t + \varepsilon_t$$

alakú, ahol $\varepsilon_t \sim N(0, \sigma)$ i.i.d. folyamat. A kérdés az, hogy ρ értéke 1 lesz-e, azaz egységgyök folyamatunk van-e, vagy sem. Tehát

$$H_0: \ \rho = 1.$$

α és δ értékének függvényében időtrendet tartalmazó ill. nem tartalmazó, eltolásos vagy eltolás nélküli véletlen bolyongás folyamatokat kapunk H_0 teljesülése esetén.

Stacionaritási tesztek

Autokorrelált reziduumok esetén: a Phillips-Perron (DF teszt korrigált változata) teszt, valamint a kibővített DF-teszt (ADF) alkalmazható, de most eltolás nélküli vagy eltolásos egységgyök folyamatokat keresünk úgy, hogy y_t késleltett értékei is megjelennek a modellben, de ezek már csak stacionárius formában.

KPSS teszt: trend-stacionaritás vs. stacionaritás tesztelésére

LMC teszt: AR(p) vs. ARIMA($p, 1, 1$) tesztelésére

Példa

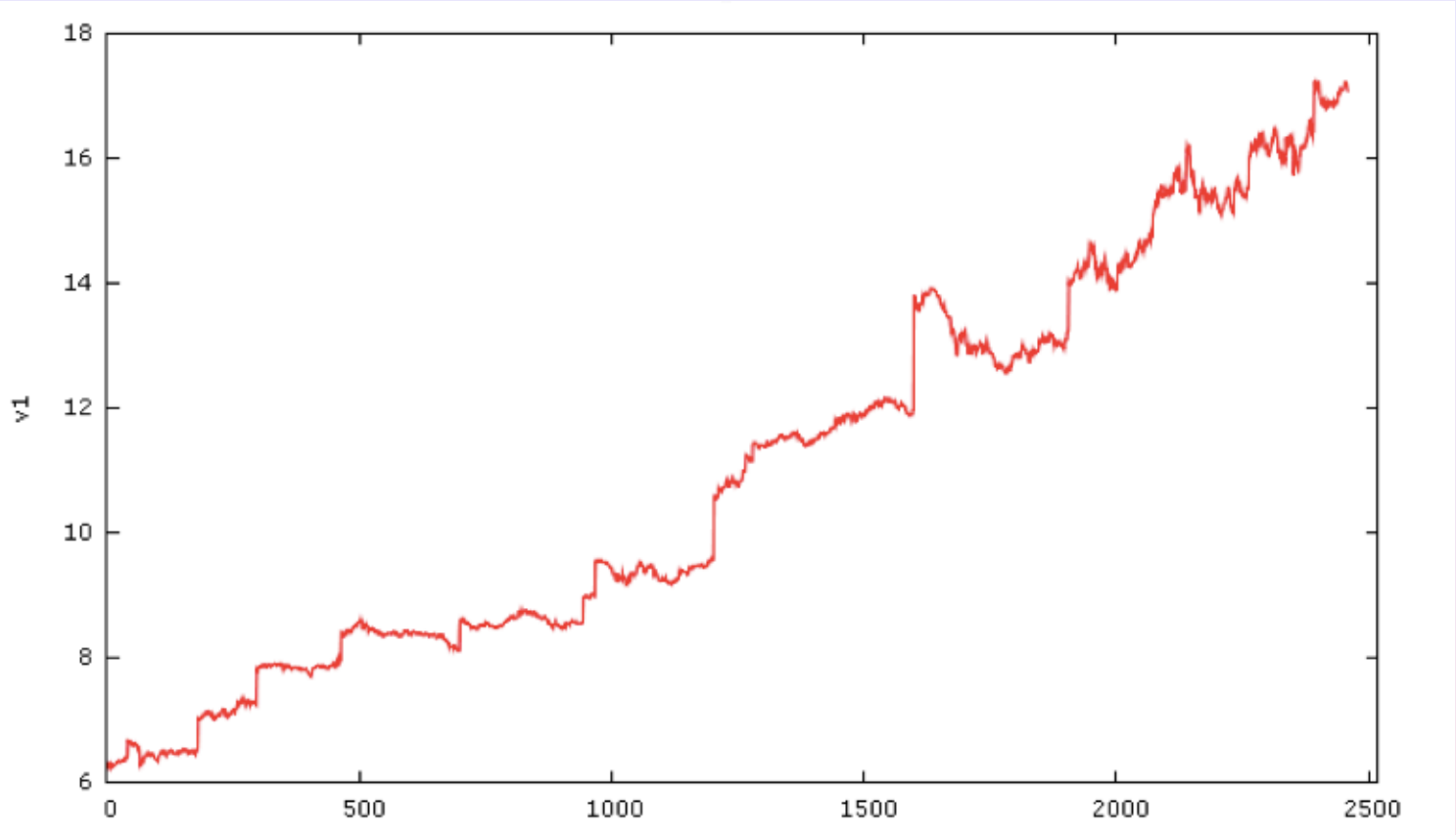
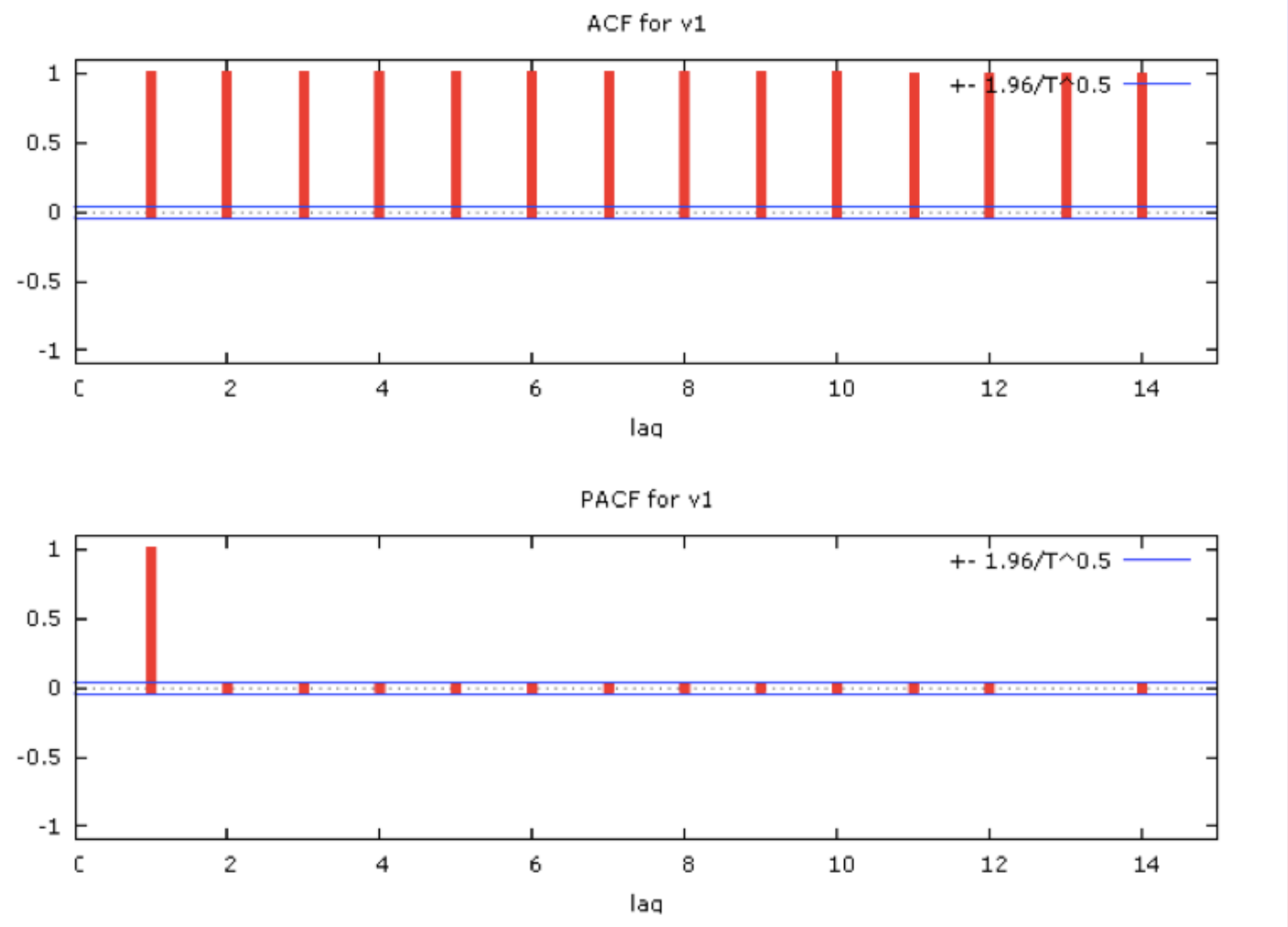


Figure: Francia frank napi árfolyama 1986 és 1993 között, 2459 megfigyelés

Biztos, hogy nem stacionárius az idősor, hiszen egyértelmű lineáris trendet tartalmaz.

Példa



A nem-stacionárius viselkedés innen is egyértelműen leolvasható.

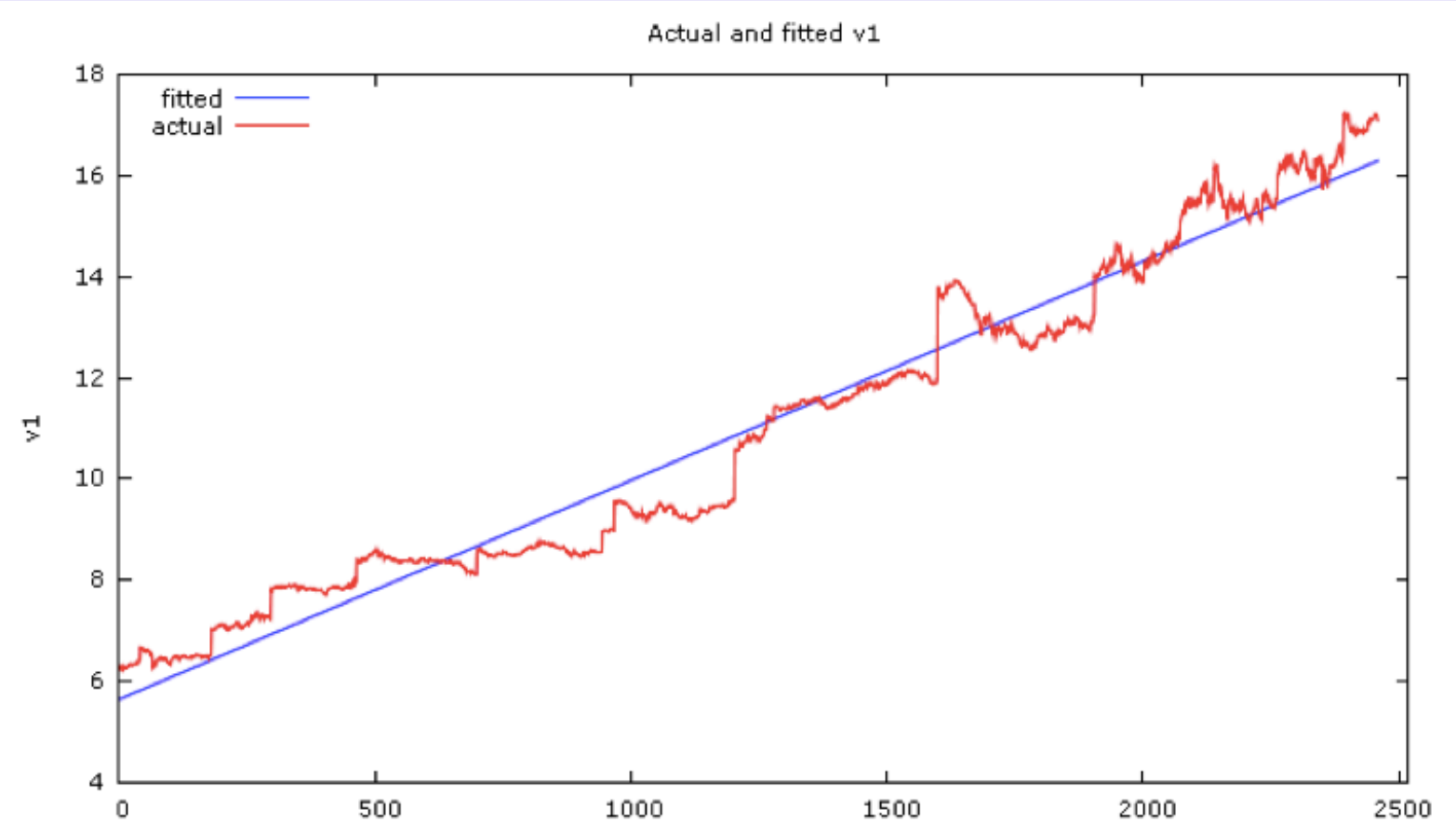
Példa

Autocorrelation function for v1

LAG	ACF		PACF		Q-stat.	[p-value]
1	0.9985	***	0.9985	***	2454.7787	[0.000]
2	0.9971	***	-0.0052		4903.2766	[0.000]
3	0.9955	***	-0.0098		7345.3686	[0.000]
4	0.9940	***	0.0010		9781.0852	[0.000]
5	0.9925	***	-0.0008		12210.4307	[0.000]
6	0.9910	***	-0.0055		14633.3434	[0.000]
7	0.9895	***	0.0013		17049.8580	[0.000]
8	0.9880	***	0.0005		19459.9980	[0.000]
9	0.9865	***	0.0016		21863.8040	[0.000]
10	0.9850	***	0.0025		24261.3287	[0.000]
11	0.9835	***	0.0045		26652.6529	[0.000]
12	0.9821	***	-0.0006		29037.7866	[0.000]
13	0.9806	***	-0.0002		31416.7438	[0.000]
14	0.9791	***	-0.0038		33789.4867	[0.000]

Kivétel nélkül minden késleltetésen szignifikáns az autokorreláció és alig csökkennek az értékek, azaz idősorunk egyértelműen nem stacionáriusnak tűnik.

Példa



Lineáris trendet illesztve az adatokra jól látszik, hogy idősorunkat a lineáris trendtől megtisztítva kellene stacionarizálni. De hogyan? Determinisztikus avagy sztochasztikus ez a trend?

Példa

A fenti kérdés eldöntésére a kibővített Dickey-Fuller (ADF) teszt ad választ.

```
Augmented Dickey-Fuller tests, order 1, for v1
sample size 2457
unit-root null hypothesis: a = 1

   test with constant
   model: (1 - L)y = b0 + (a-1)*y(-1) + ... + e
   1st-order autocorrelation coeff. for e: 0.000
   estimated value of (a - 1): 8.58636e-005
   test statistic: tau_c(1) = 0.19528
   asymptotic p-value 0.9724

   with constant and trend
   model: (1 - L)y = b0 + b1*t + (a-1)*y(-1) + ... + e
   1st-order autocorrelation coeff. for e: 0.000
   estimated value of (a - 1): -0.00641982
   test statistic: tau_ct(1) = -2.76327
   asymptotic p-value 0.2111
```

Példa

Ha az ereseti idősor csak konstanssal integrált, de a konstans és trend körül stacioner, akkor determinisztikus a trend:

```
test with constant
model: (1 - L)y = b0 + (a-1)*y(-1) + ... + e
1st-order autocorrelation coeff. for e: 0.000
estimated value of (a - 1): 8.58636e-005
test statistic: tau_c(1) = 0.19528
asymptotic p-value 0.9724
```

Ha az idősor konstans és trend körül sem stacioner, akkor differenciázni kell:

```
with constant and trend
model: (1 - L)y = b0 + b1*t + (a-1)*y(-1) + ... + e
1st-order autocorrelation coeff. for e: 0.000
estimated value of (a - 1): -0.00641982
test statistic: tau_ct(1) = -2.76327
asymptotic p-value 0.2111
```

Mindkét esetben tartható a nullhipotézis, azaz nem stacioner a folyamat. Ezért differenciázni fogunk!

Példa

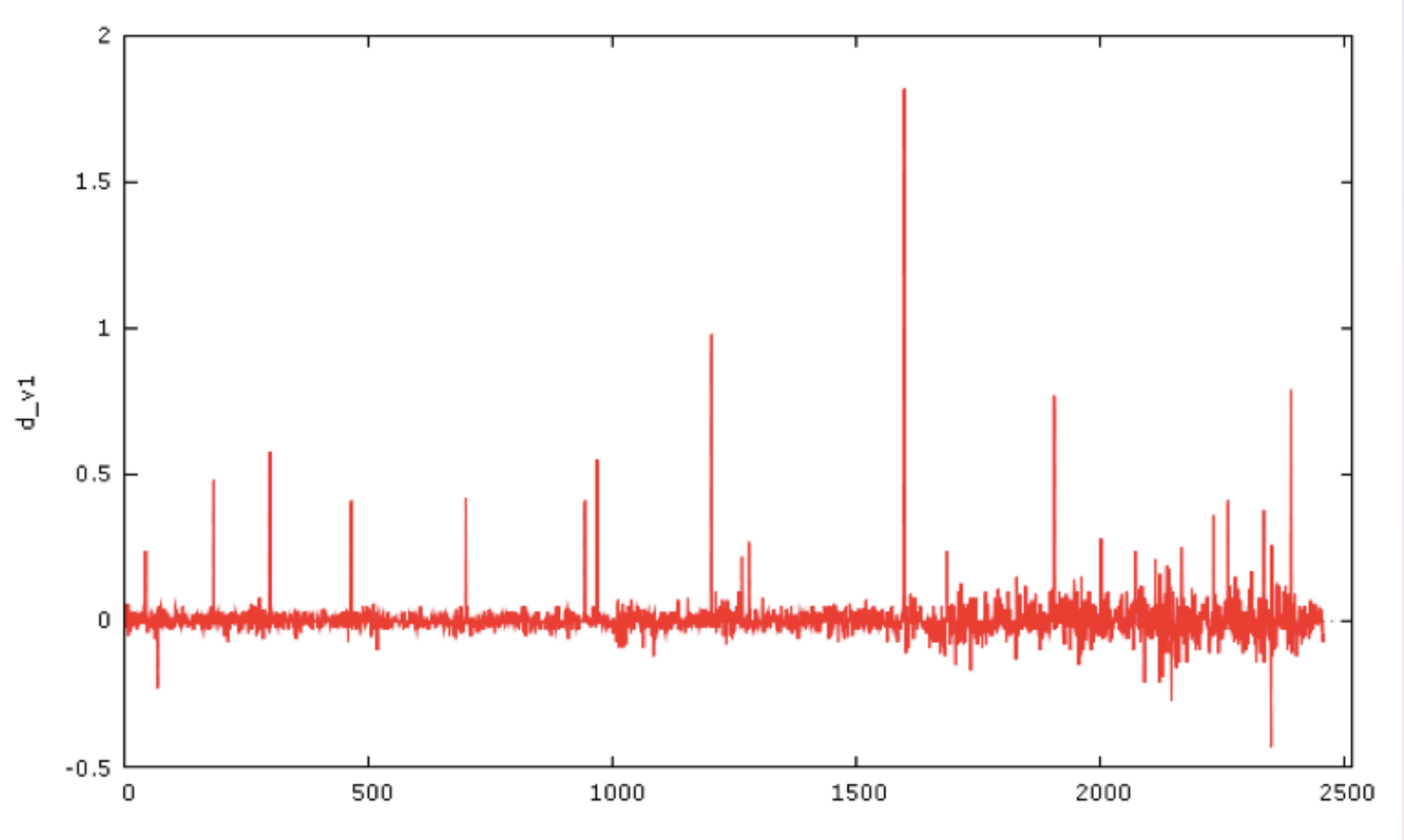


Figure: Az idősor első differenciázottja

Példa

Autocorrelation function for d_v1

LAG	ACF	PACF	Q-stat.	[p-value]
1	-0.0113	-0.0113	0.3134	[0.576]
2	0.0064	0.0062	0.4132	[0.813]
3	0.0045	0.0047	0.4640	[0.927]
4	-0.0076	-0.0075	0.6064	[0.962]
5	-0.0010	-0.0012	0.6087	[0.988]
6	0.0107	0.0107	0.8899	[0.989]
7	-0.0069	-0.0066	1.0066	[0.995]
8	-0.0084	-0.0087	1.1805	[0.997]
9	-0.0084	-0.0087	1.3561	[0.998]
10	-0.0191	-0.0190	2.2565	[0.994]
11	-0.0089	-0.0092	2.4511	[0.996]
12	-0.0132	-0.0134	2.8831	[0.996]
13	0.0135	0.0135	3.3365	[0.996]
14	-0.0194	-0.0190	4.2634	[0.994]

Most már a korrelogram-on is látszik a stacionaritás, a Ljung-Box Q-statisztikák sem szignifikánsak.

Példa

Ennek teljesen formális igazolásához újra segítségül hívjuk az ADF-tesztet.

```
Augmented Dickey-Fuller tests, order 1, for d_v1
sample size 2456
unit-root null hypothesis: a = 1

   test with constant
   model: (1 - L)y = b0 + (a-1)*y(-1) + ... + e
   1st-order autocorrelation coeff. for e: 0.000
   estimated value of (a - 1): -1.00503
   test statistic: tau_c(1) = -34.9874
   asymptotic p-value 5.278e-037

   with constant and trend
   model: (1 - L)y = b0 + b1*t + (a-1)*y(-1) + ... + e
   1st-order autocorrelation coeff. for e: 0.000
   estimated value of (a - 1): -1.00548
   test statistic: tau_ct(1) = -34.9911
   asymptotic p-value 6.389e-058
```

A nullhipotéziseket elvetjük, azaz a transzformált idősor már stacionárius, nem tartalmaz egységgyököt.