Rendszertelen idősorok modellezése spline-interpolációval*

Rappai Gábor,
a Pécsi Tudományegyetem Közgazdaságtudományi Karának intézetigazgató egyetemi tanára
E-mail: rappai@ktk.pte.hu

Az interpolációs módszerek speciális osztályáról, a spline-interpolációról kapunk áttekintést a tanulmány segítségével. Amennyiben a nemekvidisztáns idősort kiegészítjük (átlaggal, esetleg az előző értékkel feltöltjük, vagy interpolációval pótoljuk), és az így keletkező ekvidisztánst modellezzük, gyakran fals eredményre jutunk: ugyanis nem ritka, hogy a kiegészített idősor más tulajdonságokkal rendelkezik, mint az eredetit generáló folyamat. A szerző célja annak bemutatása, hogy a rendszertelen idősorok kiegészítése nem történhet mechanikusan. Dolgozatában néhány rendszertelen empirikus idősoron demonstrálja a bemutatott eljárásokat, majd néhány általános konklúziót fogalmaz meg.

TÁRGYSZÓ:
Idősorelemzés.
Interpoláció.

* A szerző ezúton mond köszönetet a TÁMOP-4.2.2.C-11/KONV-2012-0005 számú „Jól-lét az információs társadalomban” című pályázatnak kutatás támogatásáért. Köszönet illeti a pályázatban is közreműködő kollégáimat, illetve a *Statisztikai Szemle* ismeretlen lektorát értékes tanácsaikért.

Az idősorelemzés szakirodalma túlnyomó többségben olyan jelenségekkel foglalkozik, amelyekben a megfigyelések időpontjai egymástól azonos távolságra vannak, vagyis két megfigyelt időpont között egyenlő hosszúságú idő telik el. Az ilyen idősorokat tartalmazó adatállományok esetében tulajdonképpen nincs szükség a megfigyelés dátumának (időpontjának) feljegyzésére, teljes körű információt kapunk akkor is, ha csak a kezdő, illetve végső időpontot, valamint a megfigyelések gyakoriságát (az idősor frekvenciáját) tüntetjük fel.

Ezeket az idősorokat ekvidisztánsnak[1] nevezzük, és modellezésük során gyakran élünk (beláthatóan információvesztés nélkül) azzal az egyszerűsítéssel, hogy a tényleges dátum helyett az időpontokat fiktív, számtani sorozatot alkotó egész számokkal jelöljük. (A leggyakrabban alkalmazott időmegjelölés a szokásos $t = 1, 2, \ldots, T$.) Az idősor-elemzési technikák ugyanakkor nem szűkíthetők le az egyenletes (rendszeres) idősorok modellezésére, ugyanis számos olyan esettel találjuk szembe magunkat, amikor a megfigyelések nem egyenlő időközönként követik egymást.

A gazdaságtudományok területén szemléletes példa a pénzügyi piacokon mért árfolyam-idősorok esete, ahol – elsősorban a hétvégi, ritkábban az ünnep- vagy tőzsdeszünnapok következtében – az idősorokban „lyukak" vannak, ezáltal még az amúgy egymástól rendszeresen 24 óra távolságra levő záró árfolyamadatok sem ismétlődnek szabályosan. Az egy napnál nagyobb gyakoriságú pénzügyi idősorok esetén pedig a rendszertelenség tekinthető általánosnak, hiszen a valamennyi üzletkötést tartalmazó árfolyamidősoroknál az azonos távolság kritériuma úgysem teljesül, hiszen semmi sem garantálja, hogy a brókerek 10 vagy 20 másodpercenként kötnek üzletet. Szintén gyakran találkozunk rendszertelen idősorokkal a kereslet mikroszintű modellezésében, amikor az egyébként nem megfigyelhető keresletet a fogyasztással (azaz a rendszertelen időközökben megvalósuló vásárlásokkal) helyettesítjük.

A nem egyenletes időközökben keletkező megfigyelések modellezésének széles tárházát használják a tengerbiológiában, az asztrofizikában, a meteorológiában. Önmagában megérne egy hosszabb fejtegetést, hogy mi okból keletkeznek ilyen rendszertelenül megfigyelt idősorok, mennyiben lenne javítható a helyzet a mérési módszer tökéletesítésével, ám ezzel itt nem foglalkozunk.[2] A rendszertelenséget (időbeli

[1] Az ekvidisztáns helyett az angol nyelvű szakirodalom gyakran használja az evenly-spaced vagy equally-spaced, a német nyelvű az regelmäßige kifejezést is. Ezek a megjelölések viszonylag ritkán kerülnek elő, ugyanis „hírértéke" az egyenletesség nem teljesülésének van.

[2] Érdemes belegondolni, hogy a rendszertelen idősoroknak tulajdonképpen három altípusa különböztethető meg: a rendszeres, de nem mindig azonos időközönként keletkező; a rendszeres, de időnként kimaradó; illetve a teljesen rendszertelenül keletkező idősor. Ezek között a dolgozatban nem teszünk különbséget.

egyenetlenséget) adottságként fogjuk fel, ugyanakkor kijelenthető, hogy a modellezést nehezítő anomália, kellemetlen tulajdonság.

Az idők során különböző megoldások fejlődtek ki a változó időintervallumokat tartalmazó idősorok modellezésére:

1. Elsősorban pénzügyi idősorok esetén (de a csapadék modellezésében is) használatos, hogy a meglevő lyukakat 0-kal (bizonyos esetekben az utolsó megfigyelt tényleges értékkel[3]) töltjük fel, és az így kiegészített idősorokon végezzük el a modellezést. Intuitív módon is könnyen belátható, hogy ez a megoldás nagy mennyiségű kiegészítés alkalmazásakor teljesen tévútra vihet, ezért óvatosan kezelendő.

2. Nyilvánvalóbb megoldásnak tűnik, hogy keressünk az idősornak egy olyan frekvenciát, amelyre minden megfigyelés ráilleszthető, majd interpoláljunk a hiányzó időpillanatokhoz kvázi megfigyeléseket. Lineáris interpoláció esetén a megoldás mindenképpen gyors és kényelmes, ugyanakkor a nemlinearitásra vonatkozó tesztek ilyenkor kevéssé lesznek hatásosak, meglepően gyakran előfordul, hogy olyankor is nemlinearitást mutatnak, amikor az eredeti adatok között ez nem volt tapasztalható (*Schmitz* [2000]).

3. Lényegesen komplexebb (a dolgozatban nem tárgyalt megoldás), ha az idősor kovariancia-struktúrájából (autokovariancia-függvényéből) kiindulva képzünk becslőfüggvényt, amivel a hiányzó helyeket ki tudjuk egészíteni. Amennyiben az adathiányos szakaszt olyan valószínűségi változók jellemzik, amelyek valószínűség-eloszlása megegyezik az ismert adatok eloszlásával, akkor alkalmazható a Lomb–Scargle-algoritmus (eredeti leírását lásd *Lomb* [1976], jó áttekintést ad róla *Schmitz* [2000]), amely a rendszertelen adatokra szolgáltat periodogramot, és amelynek fontos jellemzője, hogy nem szükséges feltételezésekkel élni az adathiányos szakaszra vonatkozóan.

4. Amennyiben az adatsorunk ténylegesen rendszertelenül keletkezik, vagyis az ekvidisztáns idősorból hiányzik egy-egy megfigyelés, célszerű lehet bevezetni a folytonos időt feltételező modelleket (continuous-time model). A probléma már viszonylag korán megjelent a modellezési szakirodalomban, érdemben *Jones* [1985], *Bergstrom* [1985] és *Hansen–Sargent* [1991] foglalkozott a kérdéssel. A dolgozatnak nem célja a folytonos időt feltételező modellekre vonatkozó eredmények bemutatása, az érdeklődőknek ajánljuk *Brockwell* [2001]

[3] Az árfolyam-modellezésben a leggyakoribb feladat a hozam előrejelzése, ebben az esetben a hiányzó hozamadatok 0-val történő feltöltése ekvivalens a hiányzó árfolyamadat legutolsó tényleges adattal való helyettesítésével (forward-flat interpolation).

vagy *Cochrane* [2012] kitűnő összefoglalóját. Az ilyen modellek paraméterbecslésének általánosan használt, állapottér modellen alapuló megoldását kimerítően tárgyalja *Wang* [2013]. Ez a modellosztály leginkább előrejelzési célra használatos, ugyanakkor gyenge pontja, hogy az előrejelzés megint csak ekvidisztáns idősort feltételezve készül.

Ebben a tanulmányban az interpolációs módszerekről, pontosabban egy speciális osztályukról a spline-interpolációról[4] kapunk áttekintést. Amennyiben a nemekvidisztáns idősort kiegészítjük (átlaggal, esetleg az előző értékkel feltöltjük, vagy interpolációval pótoljuk), és az így keletkező ekvidisztáns idősort modellezzük, gyakran meglehetősen fals eredményre jutunk, ugyanis nem ritka, hogy a kiegészített idősor más tulajdonságokkal rendelkezik, mint az eredetit generáló folyamat. Célunk annak bemutatása, hogy a rendszertelen idősorok kiegészítése annak ellenére sem történhet mechanikusan, hogy a statisztikai-ökonometriai programcsomagok a lehetőséget „tálcán kínálják".

A továbbiakban először áttekintetjük a gyakrabban alkalmazott interpolációs technikákat, viszonylag részletesen tárgyalva a spline-interpoláció alapvető tulajdonságait, illetve sajátosságait. Ezt követően szimulált (fiktív) idősorokon mutatjuk meg, hogy milyen torzulásokat eredményezhet az adatgeneráló-folyamat(ok) felismerése során, ha a rendszertelen idősorokat előbb feltöltjük, majd a kiegészített idősor(ok)ra végezzük el a szokásos teszteket. A tanulmány végén néhány rendszertelen empirikus idősoron demonstráljuk a bemutatott eljárásokat, legvégül néhány általános konklúziót fogalmazunk meg.

1. A probléma kezelése hiányzó adatok feltételezésével

A rendszertelen idősorok kezelésének megszokott útja, ha azzal a feltételezéssel élünk, hogy létezik egy „eredeti" idősor, ami tulajdonképpen ekvidisztáns, csak nem ismerünk belőle néhány megfigyelt értéket. Ilyenkor a hiányzó adatok kezelésének leggyakrabban alkalmazott módszere az idősori interpoláció.

Az interpoláció általánosabban használatos eljárás, vagyis nem csak akkor alkalmazható, ha hiányzó vagy vélt hiányzó adatot akarunk pótolni. Minden olyan becslést így nevezünk, amelyben az idősor „közepén" (értsd nem a megfigyelési időszа-

[4] A spline kifejezésnek mindeddig nem honosodott meg magyar megfelelője. A szó eredetileg a hajógyártásból származik, a hosszú, rugalmasan hajlítható, a hajótest formáját jól követő lécekre (dongákra) használatos. A spline-okra vonatkozó első matematikai hivatkozás *Schoenberg* [1946] cikkében olvasható.

kon túl!) található időponthoz rendelünk hozzá egy ex post becslést. Jelen dolgozatban az interpoláció két típusát mutatjuk be:

– a *lineáris* (illetve az ezzel gondolatvilágában azonos log-lineáris) és
– a *spline*

közelítést. Mindkét eljárás ugyanazzal a lépéssel indul: meg kell határoznunk az idősorra jellemző gyakoriságot, vagyis azt a frekvenciát, aminek alkalmazásával kijelöljük a hiányzó (interpolálandó) adatok helyét. Bizonyos esetekben a kérdés triviális, hiszen adott egy „természetes" gyakoriság, csak valamely okból nem keletkezik minden elvárt időpillanatban adat. (Gondoljunk a már említett napi záróárfolyam-idősorban szombatonként és vasárnaponként keletkező lyukakra. Ekkor a természetes megfigyelési gyakoriság a naponkénti, ugyanakkor minden 6. és 7. érték hiányzik.) Más esetekben nincs ilyen kézenfekvő megoldás, hiszen például a világcsúcsok egy sportágban vagy a kormánykoalíció erejét mutató mandátumarány változása elméletileg sem ugyanolyan időközönként következik be. A már többször említett tőzsdei példában hasonló problémák keletkeznek akkor, amikor a különböző tőkepiacokon (tőzsdéken) a nemzeti sajátosságok következtében eltérő időpontokban megjelenő ünnepnapok okoznak rendszertelenséget.

Általánosan javasolt eljárás, hogy a feltételezett gyakoriság legyen a tényleges megfigyelések között előforduló *legkisebb* távolság. Erről ugyanakkor könnyen belátható, hogy nem feltétlenül eredményez olyan frekvenciát, amelyre valamennyi tényleges érték illeszkedik. Két megjegyzést fűznénk a feltételezett idősori frekvencia megállapításához:

– egyrészről kívánatos, hogy minél több (lehetőleg az összes) eredeti megfigyelés megfeleltethető legyen a feltételezett idősor egy konkrét pontjával, ami – könnyen átláthatóan – a minél sűrűbb megfigyelési gyakoriság melletti érv;
– másrészről el kell(ene) kerülni, hogy az interpolált értékek száma meghaladja (egyes érvelések szerint megközelítse) a tényleges (valós) adatok számát, mindez a túl sűrű feltételezett megfigyelési gyakoriság ellen szól.[5]

A feltételezett gyakoriság bevezetésével már egyenletessé tett, ám hiányzó adatokat tartalmazó idősor felírását követően az interpoláció azt jelenti, hogy meg kell becsülnünk a folyamat lefutását minden ismert két empirikus érték között.

[5] A szakirodalom az elmúlt mintegy két évtizedben sokat foglalkozott azzal a problémával, hogy az extrém nagy sűrűségű idősorok (ultra-high frequency data) modellezése esetén az említett kritériumok nehezen teljesíthetők. Az ilyen, tipikusan tőzsdei üzletkötéseket tartalmazó idősorok modellezési lehetőségeinek kitűnő összefoglalása olvasható *Engle* [1996] munkaanyagában.

Az interpoláció eredményeként keletkező kiegészített idősorral kapcsolatban két követelményt támaszthatunk:

– ahol ilyen létezik, ott a megfigyelt idősori értékeket adja vissza,
– legyen viszonylag sima, azaz diszpreferálja a töréseket.

Vezessük be a következő jelöléseket! Legyen a megfigyelt (empirikus) idősorunk

$$y_{t_1}, y_{t_2}, \ldots, y_{t_k}, \ldots, y_{t_T},$$

ahol $t_2 - t_1$ nem feltétlenül egyezik meg $t_3 - t_2$ távolsággal. Legyen Δ az a legnagyobb távolság, amelyre igaz, hogy valamennyi $t_k - t_{k-1}$ megegyezik Δ-val vagy annak egész számú többszörösével.[6]

Ekkor képezhető a következő hiányos idősor:

$$y_{t_1}, y_{t_1+\Delta}, y_{t_1+2\Delta}, \ldots, y_{t_1+j\times\Delta}, \ldots, y_{t_T},$$

ahol $y_{t_1+j\times\Delta}$ eredetileg nem megfigyelt, vagyis interpolációval előállítandó adat akkor, ha $t_1 + j \times \Delta$ nem esik egybe egyetlen eredeti t_k-val sem.

Az interpoláció során a feladatunk tehát az, hogy valamilyen eljárással becsüljük azokat az értékeket, melyek olyan időpontokhoz tartoznak, amelyből eredetileg nem származik empirikus adatunk. Triviális, ám a korábban felállított kritériumrendszernek nem teljesen megfelelő eljárás a lineáris interpoláció. Ekkor ha

$$y_{t_{k-1}} = y_{t_1+(j-1)\times\Delta} < y_{t_1+j\times\Delta} < y_{t_k} = y_{t_1+(j+1)\times\Delta},$$

ahol $y_{t_1+j\times\Delta}$ eredetileg hiányzik, ugyanakkor két „szomszédja” ismert, akkor az interpolációval pótolt érték

$$\hat{y}_{t_1+j\times\Delta} = y_{t_{k-1}} + \frac{y_{t_k} - y_{t_{k-1}}}{t_k - t_{k-1}} = y_{t_{k-1}} + \frac{y_{t_k} - y_{t_{k-1}}}{2\Delta}.$$

Az utóbbi felírás alapján könnyen belátható, hogy amennyiben a két empirikus érték között egynél több hiányzó adat található, akkor az interpoláció a nevező értelemszerű módosításával egyszerűen elvégezhető. Általánosságban a lineáris interpoláció felírható formulája:

$$\hat{y}^{LIN} = (1-\lambda)\, y_{t_{k-1}} + \lambda y_{t_k}, \qquad /1/$$

[6] Ez gyakran, de nem feltétlenül, megegyezik a két tényleges megfigyelés közötti minimális távolsággal.

ahol $y_{t_{k-1}}$ az utolsó nem hiányzó adat, y_{t_k} a következő nem hiányzó adat és λ megmutatja a hiányzó adat relatív pozícióját a két ismert, empirikus érték között. (Látható, hogy amennyiben egy érték hiányzik, akkor felezni kell az ismert különbséget; ha kettő hiányzó adat van, akkor harmadolni és így tovább.)

Könnyen átlátható, hogy az így képzett egyszerű lineáris interpoláció meglehetősen „töredezett" folyamatot szolgáltat, az eljárással keletkező becsült függvény meredeksége gyakran és ugrásszerűen váltakozik. Ennek a töredezettségnek a tompítására szokták alkalmazni a log-lineáris interpolációt, ahol a becsült érték a

$$\hat{y}^{LOGLIN} = e^{(1-\lambda)ln\, y_{t_{k-1}} + \lambda ln\, y_{t_k}} \qquad /2/$$

képlettel keletkezik.

Miközben a logaritmálás varianciastabilizáló jellegénél fogva az eljárás némiképpen simább megoldást szolgáltat, újabb problémaként merül fel az esetleges negatív értékek kezelésének nehézsége, így – noha kiszámítása meglehetősen egyszerű – a bemutatott lineáris, illetve log-lineáris interpoláció inkább csak durva tájékozódásra használatos.[7]

A sima függvények megtalálására fejlesztették ki az ún. spline-interpolációt. Az eljárás eredeti definíciója szerint szakaszonként adjuk meg az $S(t)$ interpoláló függvényt, úgy, hogy az kielégítsen bizonyos speciális feltételeket. Amennyiben – mint eddig is – a megfigyelések helyét $t_1 < t_2 < \ldots < t_T$ pontok jelölik, és a megfigyelt értékekről feltételezzük, hogy ezek az idő függvényében alakulnak, vagyis $y_{t_k} = f(t_k)$, akkor olyan $S(t)$ függvényt keresünk, amely teljesíti a következő feltételeket:[8]

$$S(t) = S_{t_k}(t) \quad t \in [t_1, t_T], \qquad /F1/$$

$$S(t_k) = y_{t_k}, \qquad /F2/$$

$$S_{t_k}(t_{k+1}) = S_{t_{k+1}}(t_{k+1}). \qquad /F3/$$

E feltételek tulajdonképpen a következőket jelentik: az interpoláció szakaszokból áll, és akár minden szakaszra különböző függvényt definiálhatunk; az interpoláló

[7] Az eljárások értelemszerűen továbbfejleszthetők: amennyiben nem csak a hiányzó adatot közvetlenül megelőző, illetve követő ismert értéket használjuk fel, akkor az interpoláció simasága javítható (ilyen például az EViews programban használatos cardinal spline módszer).

[8] Ezeket a feltételeket F1, F2 stb. számozással jelöljük.

függvény a tényleges megfigyeléseket képes reprodukálni; valamint az interpoláció eredményképpen kapott görbe folytonos (hiszen a közbülső megfigyelések két szakaszhoz is tartoznak, de ott /F3/ értelmében mindkét szakasz egyenlő értékkel bír). Az előbbi három feltétel a spline-interpoláció általános definíciója.

Annak függvényében, hogy milyen típusú $S(t)$ függvényeket használunk, más-más spline-eljárásokról beszélhetünk.[9] A leggyakoribb megoldás, hogy $S(t)$ függvényeket a polinomok közül választjuk, mégpedig úgy, hogy magasabb fokszámú polinomok esetében a közbülső pontokban csatlakozó szakaszoknál a deriváltak (meredekség) egyezőségét is megköveteljük. Általánosságban egy spline *p*-ed fokú és *m*-ed rendű, ha szakaszonként legfeljebb *p*-ed fokú polinomokból áll, és a közbülső pontokban a találkozó szakaszok deriváltjai *m*-ed rendig megegyeznek.[10]

A továbbiakban két – a gyakorlatban viszonylag elterjedt – spline-interpolációt mutatunk be:

– inkább csak didaktikai okból a lineáris spline-t és a
– harmadfokú, másodrendű spline-t.

A lineáris spline bemutatása során elsőként fókuszáljunk mindössze egy szakaszra: legyen a vizsgált intervallum $[t_{k-1}, t_k]$, melynek – feltevésünk szerint – két végpontján ismert érték helyezkedik el, így ha meg tudjuk határozni a két empirikus érték között lefutó, interpolált görbét (a sztochasztikus folyamat alakulását), akkor a szakaszon található hiányzó adatokat csak le kell olvasnunk erről a függvényről.

Definíció szerint a spline-ra igaz, hogy

$$S_{t_{k-1}}(t_{k-1}) = \alpha_{t_{k-1}} + \beta_{t_{k-1}} t_{k-1} = y_{t_{k-1}},$$

$$S_{t_{k-1}}(t_k) = \alpha_{t_{k-1}} + \beta_{t_{k-1}} t_k = y_{t_k}.$$

Ebből a kétismeretlenes, kétegyenletes rendszerből az ismeretlen paraméterek ($\alpha_{t_{k-1}}, \beta_{t_{k-1}}$) rendre meghatározhatók, vagyis a spline felírható.[11]

[9] Noha az általános definíció megengedi, hogy akár minden szakaszon más-más függvénytípust használjunk, általában azonos függvényosztályból származtatjuk az interpoláló függvényeket.

[10] Nyilvánvalóan erős megszorítást jelent az a feltételezés, miszerint egy folyamat adott intervallumon folytonosan differenciálható függvény szerint fut le. (Gondoljunk például az árfolyam-modellezésben kitüntetett szerepet játszó Brown-mozgásra, ahol a differenciálhatóság sehol sem teljesül!) Ezért itt is szükséges hangsúlyozni, hogy az interpolációs technikák nem „csodaszerek", hanem körültekintően és óvatosan alkalmazandó „sebtapaszok".

[11] Az egyenletrendszer megoldhatósága szemmel látható, hiszen az együtthatómátrix determinánsára $t_k - t_{k-1} > 0$ definíciószerűen teljesül.

A megoldás egyébiránt azonos a már bemutatott lineáris közelítéssel, vagyis:

$$S^{LIN}(t) = y_t = y_{t_{k-1}} + \frac{y_{t_k} - y_{t_{k-1}}}{t_k - t_{k-1}}(t - t_{k-1}) \quad t \in [t_{k-1}, t_k]. \qquad /3/$$

A felírásból jól látható, hogy a spline paraméterei minden $[t_{k-1}, t_k]$ intervallumban változnak, illetve változhatnak.

A gyakorlatban – mivel ésszerű számolásigénnyel megfelelő rugalmasságot biztosít – általában harmadfokú, másodrendű spline-interpolációt[12] alkalmazunk. Harmadfokú spline esetén a korábban tárgyalt /F1/–/F3/ feltételek újabb hárommal[13] egészülnek ki:

$$S'_{t_{k-1}}(t_k) = S'_{t_k}(t_k), \qquad /F4/$$

$$S''_{t_{k-1}}(t_k) = S''_{t_k}(t_k), \qquad /F5/$$

$$S''(t_1) = 0 \qquad S''(t_T) = 0, \qquad /F6/$$

Az interpolációhoz szükséges görbék meghatározása során tehát keressük a

$$\begin{aligned} S^{CUB}(t) &= S_{t_{k-1}}(t_{k-1}) = y_{t_{k-1}} = \\ &= \alpha_{t_{k-1}} + \beta_{t_{k-1}}(t - t_{k-1}) + \gamma_{t_{k-1}}(t - t_{k-1})^2 + \delta_{t_{k-1}}(t - t_{k-1})^3 \quad t \in [t_{k-1}, t_k] \end{aligned} \qquad /4/$$

kifejezéshez tartozó paramétereket. Belátható, hogy összesen $4(T-1)$ darab ismeretlen paraméterhez a /F2/–/F6/ feltételek pontosan ugyanennyi egyenletet határoznak meg, így a feladat megoldható.[14]

Az előbbiekben bemutatott, általánosan definiált harmadfokú, másodrendű spline-interpoláció helyett gyakran alkalmazzák az ún. *Catmull–Rom-spline*-okat (első leírását lásd *Catmull–Rom* [1974], a továbbiakban *CRS*). Az eljárás akkor alkalmazható, ha feltételezhetjük, hogy a rendszertelen idősor tulajdonképpen nem más, mint egy ekvidisztáns idősor, melyből hiányoznak megfigyelések.

[12] Amikor a harmadfokú, másodrendű spline-interpolációról esik szó, általában egyszerűen harmadfokú spline-ról (cubic spline) beszélünk.

[13] A felírás az ún. természetes spline-ra vonatkozik, elvben nem kizárt, hogy a második derivált a kezdő, illetve a végső megfigyelésnél nem 0.

[14] A bizonyítást lásd *Mészárosné* [2011]. Az ismeretlenek és feltételek számának megegyezése természetesen csak szükséges, ám nem elégséges feltétele az egyenletrendszer egyértelmű megoldhatóságának. A megoldás egzisztenciája és unicitása megkívánja az együtthatómátrix nem szinguláris voltát is.

Vezessük be a

$$y_{t_0+j\times\Delta} = y_j$$

jelölést, így az idősor első, biztosan megfigyelt értéke y_0, a második értéke y_1, és így tovább. Az eljárás lényege, hogy feltesszük, minden (megfigyelt, vagy éppen hiányzó) pont egy harmadfokú polinomon fekszik, melynek az adott helyen nemcsak az értékét, de a deriváltját is ismerjük.

$$y(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3$$

Nézzük az első két pont esetén mindez mit jelent:

$$\begin{aligned} y(0) &= \alpha_0, \\ y(1) &= \alpha_0 + \alpha_1 + \alpha_2 + \alpha_3, \\ y'(0) &= \alpha_1, \\ y'(1) &= \alpha_1 + 2\alpha_2 + 3\alpha_3. \end{aligned}$$

Oldjuk meg az egyenletrendszert az ismeretlen paraméterekre:

$$\begin{aligned} \alpha_0 &= y(0), \\ \alpha_1 &= y'(0), \\ \alpha_2 &= 3\left[y(1) - y(0)\right] - 2y'(0) - y'(1), \\ \alpha_3 &= 2\left[y(0) - y(1)\right] + y'(0) + y'(1). \end{aligned}$$

Mindezt visszahelyettesítve az eredeti polinomba, és elvégezve a szükséges egyszerűsítéseket kapjuk a következő harmadfokú polinomot:

$$y(t) = \left(1 - 3t^2 + 2t^3\right) y(0) + \left(3t^2 - 2t^3\right) y(1) + \left(t - 2t^2 + t^3\right) y'(0) + \left(-t^2 + t^3\right) y'(1). \quad /5/$$

Az /5/ egyenlet megoldása során a nehézséget az okozza, hogy a különböző megfigyelt értékeknél nehezen adható meg az illesztett (illesztendő) görbe deriváltja (meredeksége).

A CRS-eljárás során feltesszük, hogy az előbbi deriváltak a megfigyelt értékekből egyszerűen meghatározhatók. Keressük a spline-t az $\left[y_j, y_{j+1}\right]$ szakaszon! Legyenek a keresett meredekségek a következők:

$$y'(j) = \frac{y_{j+1} - y_{j-1}}{2},$$
$$y'(j+1) = \frac{y_{j+2} - y_j}{2}.$$

Így a korábban bemutatott harmadfokú polinom felírható mátrix alakban a következőképpen:

$$y(t) = \begin{bmatrix} 1 & t & t^2 & t^3 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -3 & 3 & -2 & -1 \\ 2 & -2 & 1 & 1 \end{bmatrix} \begin{bmatrix} y_j \\ y_{j+1} \\ \dfrac{y_{j+1} - y_{j-1}}{2} \\ \dfrac{y_{j+2} - y_j}{2} \end{bmatrix}.$$

Mindez minimálisan átalakítva:

$$y(t) = \begin{bmatrix} 1 & t & t^2 & t^3 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -3 & 3 & -2 & -1 \\ 2 & -2 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -\frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & -\frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} y_{j-1} \\ y_j \\ y_{j+1} \\ y_{j+2} \end{bmatrix},$$

majd a két belső mátrixot összeszorozva

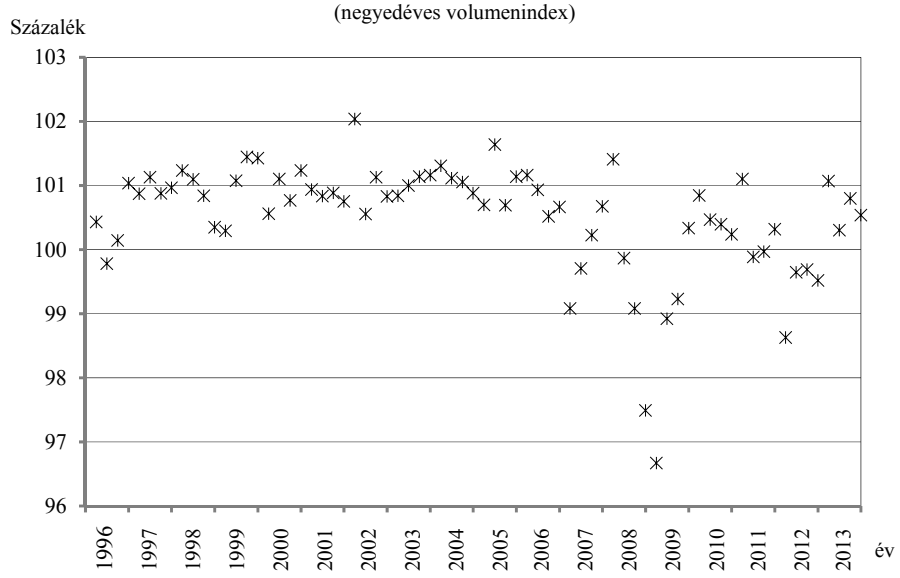
$$y^{CRS}(t) = \frac{1}{2} \begin{bmatrix} 1 & t & t^2 & t^3 \end{bmatrix} \begin{bmatrix} 0 & 2 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 2 & -5 & 4 & -1 \\ -1 & 3 & -3 & 1 \end{bmatrix} \begin{bmatrix} y_{j-1} \\ y_j \\ y_{j+1} \\ y_{j+2} \end{bmatrix}. \qquad /6/$$

Az előbbi egyenlettel meghatározott, viszonylag könnyen számszerűsíthető görbe reprezentálja az idősor alakulását két kijelölt pont között. (Minden különösebb ma-

gyarázat nélkül látható, hogy az interpolációval keletkezett görbék minden szakaszon változhatnak.)

Tekintsük a következő rendkívül egyszerű példát! A magyar reál GDP negyedéves változását jellemző volumenindexek 1996 és 2013 között az 1. ábrán látható módon alakultak:

1. ábra. A magyar GDP alakulása 1996 és 2013 között
(negyedéves volumenindex)



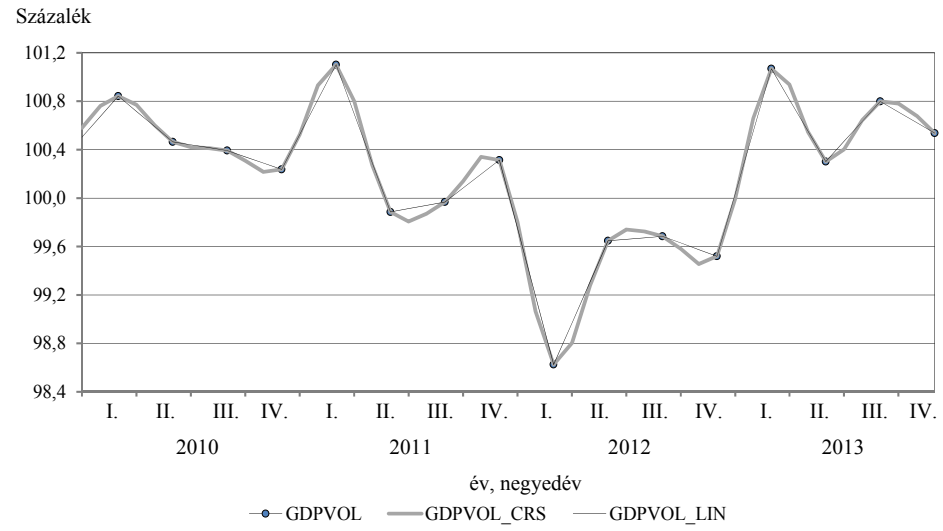
Forrás: KSH (http://www.ksh.hu/docs/hun/xstadat/xstadat_evkozi/e_qpt001b.html).

A pontdiagramon, miután minden negyedévhez egy értéket rendeltünk (vagyis összességében 72 elemű idősorunk van), viszonylag nehezen követhetők a tendenciák, ezért is használunk idősorok esetén általában – némiképp félrevezető módon – vonaldiagramot.[15] Amennyiben a negyedéves tényadatokat összekötjük, tulajdonképpen – ki nem mondva – értékeket interpolálunk az empirikus értékek közé. A korábban elmondottaknak megfelelően többféle módon is elvégezhetjük az interpolációt, a 2. ábrán lineáris és Catmull–Rom-spline-nal végzett interpoláció segítségével becsült havi bontású fiktív idősorok láthatók.[16] Annak érdekében, hogy az ábrán jobban elkülönüljenek a tényadatok (GDPVOL), és a lineáris, illetve CRS interpolációval nyert értékek (GDPVOL_LIN, GDPVOL_CRS), ezért csak az utolsó négy év adatait szerepeltettük.

[15] A grafikus ábrázolásra vonatkozó szabályok, elvek tekintetében lásd például *Hunyadi* [2002].

[16] A Központi Statisztikai Hivatal is elvégzi a negyedéves GDP-adatok havi bontásra sűrítését (igaz nem a volumenindexek, hanem az értékadatok tekintetében), de az egy teljesen más gondolatmeneten alapuló, ezért értelemszerűen teljesen eltérő eredményre vezető eljárás.

2. ábra. A magyar GDP alakulása 2010 és 2013 között
(negyedéves volumenindex interpolálásával nyert havi adatok)



A 2. ábráról leolvasható a kétféle interpoláció eredményeképpen keletkező, egymástól esetenként jelentősen eltérő becsült idősor. Érdemes felfigyelni arra, hogy olykor a polinomiális spline segítségével becsült értékek „túlfutnak" a lineáris interpoláció által sugallt folyamatokon (tipikusan így van ez trendfordulók környezetében, például 2011 vagy 2012 közepén!). Pontosan az ilyen, a nehezen megmagyarázható túlfutások miatt merül fel a gondolat, hogy az interpolációs eljárásokat óvatosan kell kezelni.

2. Az eredeti adatgeneráló folyamat torzulása interpolációval kiegészített idősorok esetén

Ebben a fejezetben, az adatgeneráló folyamat (data generating process – DGP) torzulásának szemléltetése érdekében szimulációt alkalmaztunk.[17] Törekedtünk arra, hogy az alkalmazott modellek összehasonlíthatók legyenek, ennek érdekében a szimuláció során felhasznált konstansok (paraméterek) a különböző jellegű folyama-

[17] Az idősorok szimulációját az EViews 8.0 programcsomaggal végeztük.

toknál azonosak, ahol ez nem lehetséges, hasonlók legyenek. Az elemzés logikája mindvégig ugyanaz, tehát

1. alkalmasan választott modellel 1 000 elemű idősorokat generálunk;

2. a keletkezett fiktív (szimulált) idősorokból véletlenszerűen kihagyunk „megfigyeléseket" (az elemzés során előbb az eredeti idősor 10, 20, és így tovább, végül 90 százalékát hagytuk el);

3. az így létrejött rendszertelen idősorokban a hiányzó értékeket
– először az adott folyamat várható értékével feltöltjük,
– másodszor köbös spline-interpolációval kiegészítjük;

4. végezetül (1 000 független kísérlet alapján) megvizsgáljuk, hogy a feltöltött, illetve kiegészített idősor legfontosabb tulajdonságai menynyiben térnek el az eredetileg generált idősor alapvető jellemzőitől.

Három, az empirikus idősorok esetén nagy gyakorisággal előforduló adatgeneráló folyamatot elemeztünk, melyek

– elsőrendű vektor-autoregresszív, azaz VAR(1) modellel meghatározott;
– sztochasztikus trendet tartalmazó (véletlen bolyongást követő);
– első rendben integrált, egymással tökéletes kointegrációs kapcsolatban álló

idősorokat eredményeztek. Valamennyi szimulált idősorra érvényes, hogy az első „megfigyelést" megelőző elem (y_0) értéke 0, a felhasznált véletlen változók normális eloszlású, 0 várható értékű, 1 szórású fehérzaj-folyamatok (ezek jelölése egy folyamat esetén ε_t , két folyamat esetén $\varepsilon_{1t}, \varepsilon_{2t}$).

Elsőként, annak érdekében, hogy az idősorok között kimutatható ok-okozati összefüggések torzulását elemezni tudjuk, az általánosan használt Granger-próba logikáján alapuló vektor-autoregresszív modellből származó idősorokat generáltunk, a következő modell szerint:

$$y_{1t} = 0,9y_{1,t-1} + 0,4y_{2,t-1} + \varepsilon_{1t},$$
$$y_{2t} = 0,9y_{2,t-1} - 0,4y_{1,t-1} + \varepsilon_{2t}.$$

Mindez mátrix alakban így írható fel:

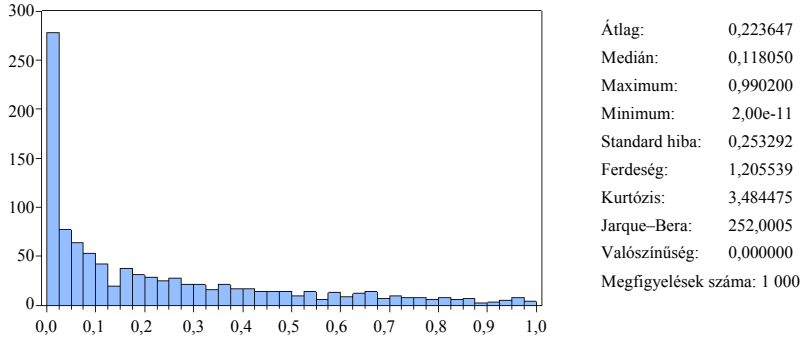
$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} 0,9 & 0,4 \\ -0,4 & 0,9 \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}.$$

Közismert, hogy a VAR-modellel felírható folyamatok akkor stacionáriusak, ha az együtthatómátrixának valamennyi sajátértéke az egységkörön belül van, valamint a paramétermátrixban a főátlón kívüli elemek különböznek 0-tól, mivel esetünkben mindkét feltétel teljesül, így a modellben szereplő változók Granger-okságban vannak egymással. A szimulációval azt vizsgáljuk, hogy előfordulhat-e, hogy a rendszertelen idősorok feltöltését vagy kiegészítését követően az okság „elveszik".

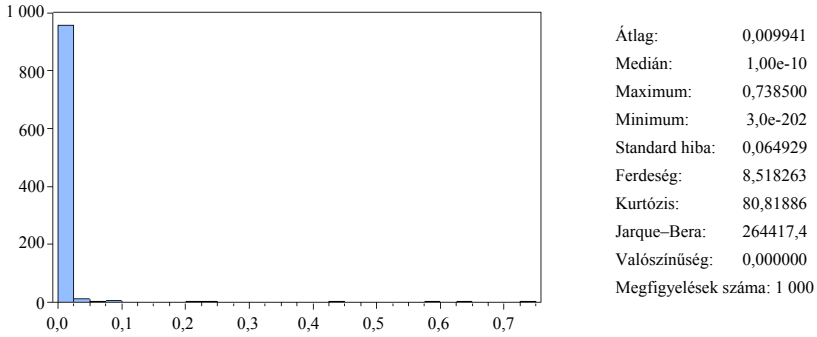
A VAR-modellekre vonatkozó szimulációs eredmények érzékeltetéséhez tekintsük a 3. *a*) és 3. *b*) ábrákat.

3. ábra. VAR adatgeneráló folyamatból származó változók közötti Granger-okság tesztjeinek p-értékei

*a) A hiányzó értékek a várható értékkel feltöltve**



*b) A rendszertelen idősor pótlása spline-interpolációval**



Megjegyzés. Amennyiben az eredeti megfigyelések 90 százaléka hiányzik.

Az ábrákból leolvasható, hogy amikor a hiányzó adatokat a várható értékkel pótoltuk, a Granger-okságot tesztelő Wald-próba[18] 5 százalékos szinten mindössze 355

[18] A sztochasztikus folyamatok tulajdonságainak vizsgálata során általánosan használt próbák leírása megtalálható például *Hunyadi* [1994] cikkében, illetve *Rappai* [2013] könyvében.

esetben veti el a nullhipotézist (vagyis talál ok-okozati összefüggést), és az eredeti adatgeneráló folyamatokhoz illeszkedő döntések száma 10 százalékos szignifikanciaszinten is csak 472. Ugyanakkor a spline-interpolációval kiegészített idősoroknál a helyes döntések száma 1 000 esetből – 5, illetve 10 százalékos szinten – rendre 970, illetve 979. Mindebből arra következtethetünk, hogy a másodrendű, harmadfokú spline-interpoláció alkalmazásakor kisebb annak a veszélye, hogy viszonylag sok hiányzó adat esetén is helytelenül ismerjük fel az adatgeneráló folyamatot, azaz a változók közötti ok-okozati összefüggést.

Az idősoros alapvetésekben mindig kiemelt figyelmet fordítunk a véletlen bolyongás folyamatra, melynek jelentőségét két dologgal is magyarázhatjuk: egyrészt a random walk az egységgyök-tesztekben a nullhipotézis alatti modellspecifikációt jelenti, másrészt az eltolásos véletlen bolyongás a sztochasztikus trend alapesete. Ennek megfelelően két random walk folyamatot szimuláltunk:

– véletlen bolyongás eltolás nélkül

$$y_t = y_{t-1} + \varepsilon_t ,$$

– véletlen bolyongás eltolással

$$y_t = 0{,}01 + y_{t-1} + \varepsilon_t .$$

Megvizsgáltuk, hogy az értékek elhagyását, majd kiegészítését követően elképzelhető-e, hogy az egységgyököt tartalmazó folyamat stacionáriusnak tűnik, az egységgyök létezésének tesztelésére kiterjesztett Dickey–Fuller-próbát alkalmaztunk. Ezután közös trendet tartalmazó idősorokat szimuláltunk. Végtelenségig leegyszerűsített modellünkben a Granger által javasolt specifikációt követtük (*Granger* [1988]). A két együttmozgó folyamat:

$$y_{1t} = x_t + \varepsilon_{1t},$$
$$y_{2t} = 2x_t + \varepsilon_{2t},$$

ahol

$$x_t = x_{t-1} + \varepsilon_t .$$

A kointegráltság tesztjére az Engle–Granger kétlépcsős tesztet (EG-teszt) használtuk, és azt vizsgáltuk, hogy az elméletben együttmozgó (közös trendet tartalmazó) idősorok esetében hányszor fordul elő, hogy a teszt a kointegráció hiányát mutatja.

A korábbiakban bemutatott szimulációk legfontosabb eredményeit az 1. táblázatban foglaljuk össze.

1. táblázat

A hibásan felismert adatgeneráló folyamatok száma 1 000 szimulált idősor esetén, 5 százalékos szignifikanciaszint mellett

Kihagyott megfigyelések aránya (százalék)	Feltöltés módja	Folyamat(ok)			
		VAR	RW ($\mu = 0$)	RW ($\mu = 0,01$)	ECM
10	feltölt	0	232	249	0
	kiegészít	0	57	52	0
20	feltölt	0	441	416	1
	kiegészít	0	47	48	0
30	feltölt	0	591	567	8
	kiegészít	0	52	56	0
40	feltölt	0	713	723	33
	kiegészít	0	36	60	0
50	feltölt	0	823	815	38
	kiegészít	0	65	57	0
60	feltölt	0	906	893	34
	kiegészít	0	67	70	0
70	feltölt	9	960	955	44
	kiegészít	0	79	43	0
80	feltölt	151	979	980	20
	kiegészít	1	73	67	0
90	feltölt	645	998	999	9
	kiegészít	30	123	105	1

Megjegyzés. A táblázatban „feltölt" jelöli, ha a hiányzó adatokat a várható értékkel pótoltuk, illetve „kiegészít" jelöli, ha a hiányzó adatokat másodrendű, harmadfokú spline-interpolációval helyettesítjük. A fejlécben a VAR a vektor-autoregresszív modellt, az RW a véletlen bolyongást, az ECM pedig a kointegrált rendszer (mivel ez hibakorrekciós mechanizmussal is felírható) jelöli.

Az 1. táblázat adatai jól mutatják, hogy

– ok-okozati kapcsolat feltételezése esetén, amennyiben a rendszertelen idősorok viszonylag nagy arányban tartalmaznak adathiányt, egyre gyakrabban kerülhetünk abba a szituációba, hogy az adatgeneráló-folyamatok szintjén meglevő Granger-okságot a hiányzó adatok kiegészítésével elfedjük, a szimuláció azt támasztja alá, hogy a spline-

interpoláció jobb tulajdonságokkal bír, mint a várható értékkel történő pótlás;

– véletlen bolyongásból származó, hiányzó adatokat tartalmazó idősoroknál a várható értékkel történő feltöltés egyértelműen hibás megoldás, ugyanakkor a spline-interpoláció alkalmazása csak jelentős arányban hiányzó érték mellett okozhatja az eredeti adatgeneráló folyamat félrespecifikálását (ne feledjük, hogy a kiterjesztett Dickey–Fuller-próba 5 százalékos szinten, 1 000 eredeti ekvidisztáns idősor esetén önmagában is mintegy 50 esetben hibás döntést sugall!);

– közös trendet tartalmazó idősoroknál szintén azt tapasztaltuk, hogy a spline-interpolációval történő adatkiegészítés kevesebb (szimulációnkban szinte semmilyen) félrespecifikálást eredményez, ezért egyértelműen ajánlható.

Szimulációs eredményeink alapján bátran kijelenthetjük, hogy amennyiben az idősor nemekvidisztáns, akkor a spline-interpolációval operáló adatkiegészítés kevesebb veszéllyel jár, mint a hagyományos módszerek.

3. Két illusztratív példa az interpolációval keletkező érdekes eredményre

Ebben a fejezetben az előzőkben bemutatott spline-interpolációt illusztráljuk két empirikus adatállományon.[19] A futtatások eredménye hangsúlyozottan illusztráció, így a becslési eredményeket nem kívánjuk sport-, illetve pénzügy-szakmai újdonságok megalapozására felhasználni.

Első példánkban két ismert úszó, az olimpiai és világbajnok *Gyurta Dániel*, illetve nagy ellenfele *Michael Jamieson* (Nagy-Britannia) által az elmúlt évek világversenyein 200 méteres mellúszásban, 50 méteres medencében elért eredményeit vizsgáljuk. Az összehasonlítandó eredmények az 2. táblázatban olvashatók:[20]

Láthatjuk, hogy mindkét versenyző eredményei rendszertelenül keletkeznek (természetesen más lenne a helyzet, ha valamennyi versenyük, illetve edzésük eredményét feljegyeznénk, de ezzel itt nem foglalkozunk), ráadásul a nem egyenletes időkö-

[19] További érdekes példa olvasható a hozamgörbe spline alapú becslésére *Kopányi* [2010] disszertációjában.

[20] Az adatok forrása a Nemzetközi Úszószövetség honlapja, ahol az időszakos világranglistákból kigyűjthetők az egyéni eredmények. (http://www.fina.org/H2O/index.php?option=com_wrapper&view=wrapper&Itemid=804). Amennyiben az adott napon a versenyző többször is rajthoz állt, a legjobb eredményét szerepeltetjük.

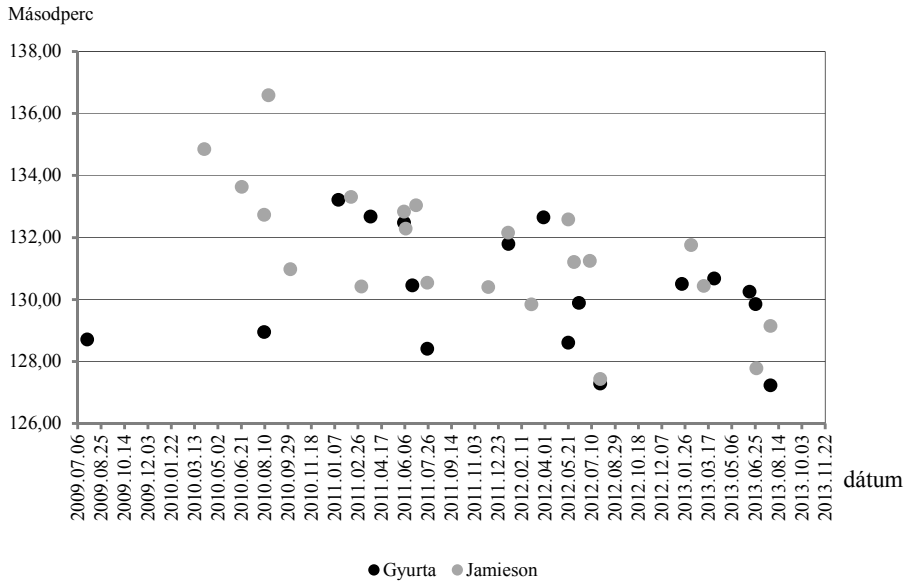
zökből származó adatok keletkezési időpontjai sem feltétlenül esnek egybe (nyilvánvalóan csak akkor, ha ugyanazon a versenyen indultak). Az eredményeket a 4. ábra szemlélteti.

2. táblázat

Úszóeredmények 2009 és 2013 között
(perc:másodperc.századmásodperc)

Dátum	Esemény	Gyurta	Jamieson
2009. 07. 26.	Világbajnokság	2:08.71	
2010. 04. 03.	British Gas Bajnokság		2:14.85
2010. 06. 22.	British Gas Nyílt Nemzetközi Bajnokság		2:13.63
2010. 08. 09.	Európa-bajnokság	2:08.95	2:12.73
2011. 01. 15.	Flanders Swimming Cup	2:13.21	2:16.59
2010. 10. 04.	Brit Nemzetközösségi Játékok		2:10.97
2011. 02. 11.	BUCS LC Bajnokság		2:13.31
2011. 03. 05.	British Gas Bajnokság		2:10.42
2011. 03. 25.	Budapest Open	2:12.67	
2011. 06. 04.	Barcelona Mare Nostrum	2:12.48	2:12.83
2011. 06. 08	Di Canet Mare Nostrum		2:12.28
2011. 06. 22.	Magyar Bajnokság	2:10.45	
2011. 06. 30.	Scottish Gas Nyílt Nemzetközi Bajnokság		2:13.04
2011. 07. 24.	Világbajnokság	2:08.41	2:10.54
2011. 12. 02.	Dán Nyílt Bajnokság		2:10.40
2012. 01. 13.	Viktoria Emlékverseny		2:12.15
2012. 01. 14.	Flanders Swimming Cup	2:11.79	
2012. 03. 03.	British Gas Bajnokság		2:09.84
2012. 03. 29.	Nyílt Nemzeti Bajnokság	2:12.65	
2012. 05. 21.	Európa-bajnokság	2:08.60	2:12.58
2012. 06. 02.	Mare Nostrum		2:11.21
2012. 06. 13.	Budapest Open	2:09.89	
2012. 07. 06.	6. EDF Nyílt Úszóbajnokság		2:11.24
2012. 07. 28.	Londoni Olimpia	2:07.28	2:07.43
2013. 01. 19.	Flanders Speedo Cup	2:10.50	
2013. 02. 08.	Derventio eXcel February Festival		2:11.75
2013. 03. 07.	British Gas Nyílt Nemzetközi Bajnokság		2:10.43
2013. 03. 29.	Budapest Open	2:10.68	
2013. 06. 13.	Sette Colli Trophy	2:10.25	
2013. 06. 26.	Magyar Bajnokság	2:09.85	
2013. 06. 28.	British Gas Bajnokság		2:07.78
2013. 07. 28.	Világbajnokság	2:07.23	2:09.14

4. ábra. Gyurta Dániel és Michael Jamieson versenyeredményei



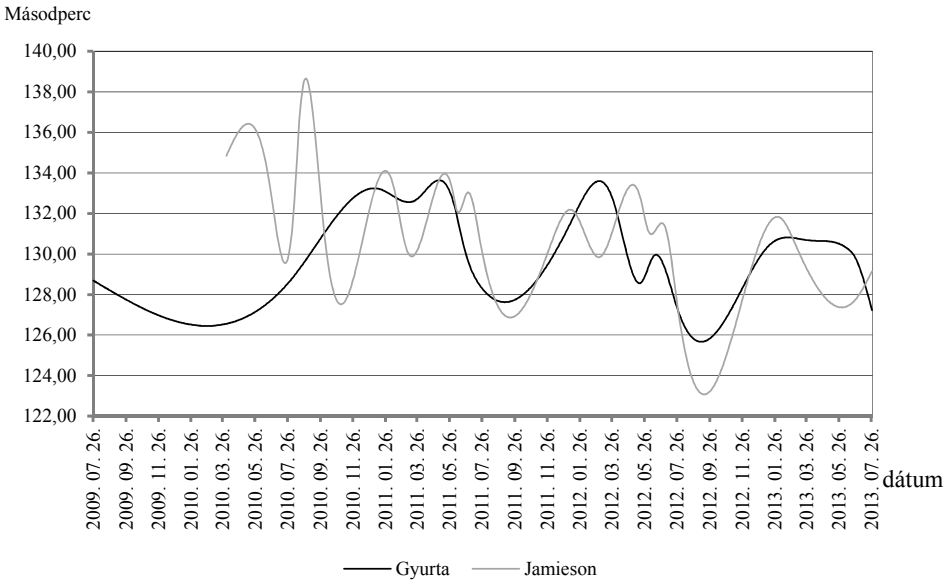
A 4. ábra – valljuk be – nem túlságosan informatív: az eredmények nehezen azonosíthatók, és főképpen nehezen hasonlíthatók össze. A vizsgált mintegy 5 évben 32 különböző időpontból származnak eredmények, ezek közül mindössze hét olyan alkalom volt, amikor mindketten indultak, ezáltal rendelkeznek eredménnyel. Az látható, hogy például a londoni olimpián vagy a 2013-as világbajnokságon Gyurta megelőzte ellenfelét (ügyeljünk arra, hogy a kevesebb idő jelenti a jobb eredményt!), de a teljes vizsgált időszakban nehéz összehasonlítani a teljesítményeket.

Érdekes lenne összevetni a két klasszis teljesítményét a teljes időhorizonton, például úgy, hogy két verseny közötti időszakra interpoláljuk a várható eredményeket. Ismét hangsúlyozandó, hogy semmilyen sportszakmai kérdést nem vizsgáltunk, tehát nem kívánjuk megítélni, hogy van-e létjogosultsága különböző felkészülési fázisokban (edzőtábor előtt, után, közben stb.) levő versenyzőket, mindössze azt illusztráljuk, hogy elvben lehetséges különböző, egymástól nem azonos távolságban levő időpontok adataiból interpolációval becsülni az eredmény változását. Mivel a megfigyelt versenyidőpontokról nem tételezhető fel, hogy eredetileg ekvidisztáns idősorból[21] származnak, csak néhány adat hiányzik, ezért harmadfokú, másodrendű spline-interpolációt alkalmaztunk.

[21] Hiszen nem arról van szó, hogy minden hónap meghatározott napján rendeznek versenyeket, csak Gyurta vagy Jamieson nem indult mindegyiken, hanem hosszabb kihagyások és sűrűbb „versenyidények" váltogatják egymást.

Az interpolációval meghatározott teljesítményértékek láthatók az 5. ábrán.

5. ábra. Gyurta Dániel és Michael Jamieson spline-módszerrel interpolált eredményei



Valószínűleg az úszáshoz kevésbé értők is látják, hogy az interpoláció eredményeképpen létrejött fiktív értékek nem feltétlenül reálisak. Az ábrából például azt lehet leolvasni, hogy a skót fiú a londoni olimpiára olyan mértékben fejlődött, hogy noha az olimpiai döntőt elveszítette, de utána „benne volt" egy sokkal jobb eredmény, akár a világcsúcs is. Majd a 2012-es idény elmúltával ismét gyengébb eredményei voltak, amelyek gyorsan javulni kezdtek, ám a világbajnokságra már túljutott a legjobb eredményén. Ezzel szemben Gyurta mindvégig kiegyensúlyozottabb, kevésbé szóródó eredményeket ért el, melyeknek éves minimuma mindig az év fő versenyén jelentkeztek. Ha mindezt a statisztikai modellezés során oly fontos, ám sokszor elfeledett verifikációként fogjuk fel, akkor láthatjuk, hogy a spline-interpoláció mechanikus alkalmazását óvatosan kell kezelnünk.

Tekintsünk egy másik, a dolgozat elején elméletben már többször hivatkozott példát! Közismert (*Bélyácz* [2009] 77. old.), hogy a piaci modell logikája alapján egy adott részvény hozama felírható így:

$$r_i = \alpha + \beta r_M \ ,$$

ahol r_i az *i*-edik (tőzsdei) befektetés hozama, r_M a piaci portfolió hozama, α és β a modell becsülendő paraméterei, melyek közül az utóbbinak kitüntetett szerepe van, ugyanis gyakran használják az adott befektetés kockázatosságának proxy-jaként.

A konkrét paraméterbecslési eljárásban a kiválasztott befektetésre (leggyakrabban tőzsdén forgó részvényre) vonatkozóan meghatározzuk a hozamot a *t*-edik időpontra

$$r_{it} = \frac{p_{it} - p_{i,t-1}}{p_{i,t-1}} \approx \Delta log\, p_{it} \ ,$$

ahol p_{it} az adott részvény záróárfolyama a *t*-edik napon.

Hasonló logikával a tőzsdeindex alakulásból kiindulva elvégezhető a piaci portfolió hozamának közelítése, ezáltal a modellben szereplő mindkét változó idősora rendelkezésünkre áll. Mivel a befektetések hozama szinte mindig stacionárius folyamatból származó idősor, így a paraméterbecslés OLS-sel általában hatékonyan megoldható.[22]

Tanulmányunk témája szempontjából nagy jelentősége van annak, hogy az említett részvény-záróárfolyamok, illetve a tőzsdeindex napi záró értékei elvben napi rendszerességű ekvidisztáns idősort alkotnának, ám csak hétköznapokon keletkeznek, vagyis rendszertelen idősorainkból hiányoznak a szombat-vasárnapi értékek, illetve a tőzsdeszünnapok. A β-becslés során bevett gyakorlat, hogy a hétvégeken, illetve tőzsdei szünnapokon (valamint a ritkán előforduló kereskedési felfüggesztések esetén) az adott részvény, illetve részvényindex hiányzó záróárfolyamát (-értékét) helyettesítjük az utolsó tényadattal, ami praktikusan azt jelenti, hogy a hozamok idősorát nullákkal töltjük fel. Előző fejtegetésünkből kitűnt, hogy a 0 értékekkel való pótlás komoly veszélyekkel jár, ugyanis például az is előfordulhat, hogy a valójában egymással ok-okozati viszonyban álló változók között nem lesz kimutatható a kapcsolat.

Tekintsük a 2014-es első négy hónapját, és vizsgáljuk meg, hogy milyen különbséget okoz, ha a Budapesti Értéktőzsdén forgó legfontosabb részvényekre vonatkozó β-becslést különböző kiegészítésekkel végezzük el! A 6. ábra a Budapesti Értéktőzsde indexének (BUX) alakulását mutatja a vizsgált időszakban.

Az idősor „szakadozottsága” a hétvégék, illetve tőzsdei kereskedési szünnapok (például a húsvéti ünnepek) miatt keletkezik, hasonlót találnánk, ha az elemzésünkbe vont MOL, MTELEKOM, OTP, RICHTER árfolyamait vagy az azokból számítható hozamokat ábrázolnánk.

Elvégeztük a korábban tárgyalt kockázati proxy, vagyis a β-együttható becslését három módon:

– csak a hétköznapok (tőzsdei munkanapok) figyelembe vételével, vagyis feltételezve, hogy péntek és hétfő között a hozam éppen úgy

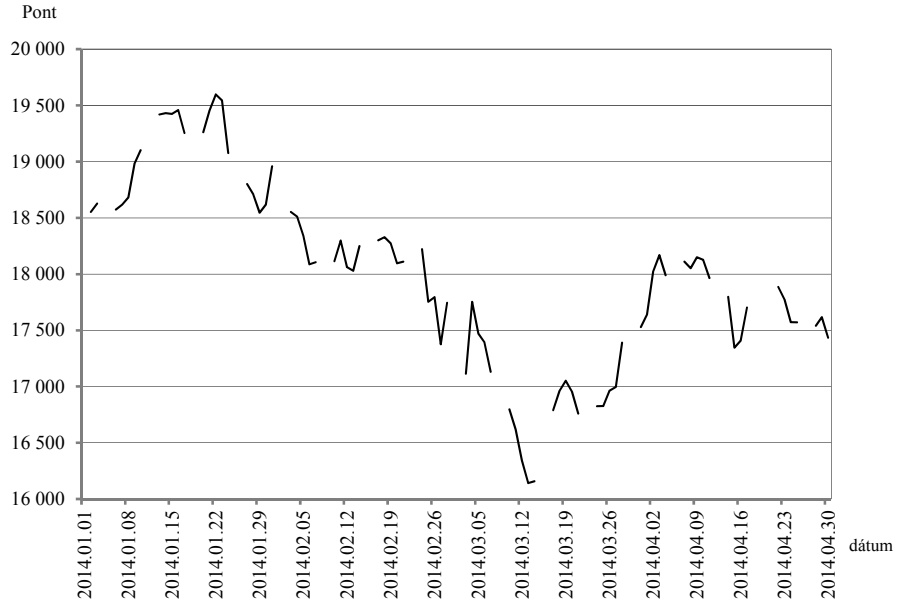
[22] Egy korábbi tanulmányunkban (lásd *Varga–Rappai* [2002]) bemutattuk, hogy a paraméterbecslés GARCH-specifikációt végezve korrektebb eredményekre vezet, de ezzel itt nem kívánunk foglalkozni.

képződik, mint hétfő és kedd között (ebben az esetben 83 hozamadatunk lett mind a BUX, mind a konkrét részvények idősorában);

– az év első négy hónapjának valamennyi napjához rendeltünk hozamértéket, mégpedig úgy, hogy a tőzsdei szünnapokon 0 hozamot tételeztünk fel, a további napokon a hozam úgy keletkezett, mint az előző pontban (ekkor 120 elemű idősoraink lettek);

– a tényleges hozamok közötti hiányzó adatokat harmadfokú, másodrendű spline-interpolációval pótoltuk, és az így kiegészített idősorok alapján becsültük az együtthatót (így 118 adatunk lett valamennyi idősorban, ugyanis az év első két napjához – mivel előtte nem volt megfigyelt értékünk – nem tartozik interpolált érték).

6. ábra. A BUX alakulása 2014 első négy hónapjában



Az eredmények (lásd a 3. táblázatot) önmagukért beszélnek: a hozamok 0-val történő feltöltését követően a becsült együtthatók alig térnek el – megítélésünk szerint hibásan – csak a hétköznapok figyelembe vételével, így a hozam keletkezésének időtartamával nem kalkuláló becslésektől. Ugyanakkor a spline-interpolációval kiegészített adatsorok alapján becsült β-együtthatók – különösen a nagyobb kockázatosságot jelző, 1-nél nagyobb abszolút értékű esetekben – jelentősen meghaladják az eredetileg becsült értékeket, felhívva a figyelmet arra, hogy amennyiben egy befektetés árfolyama a (hosszú) hétvégék, azaz kihagyások után eltérően változik, mint a

piac egészét reprezentáló tőzsdeindex értéke, akkor ez a befektetés kockázatosságának növekedését mutatja.

3. táblázat

β-együttható becslése különböző idősorok alapján

Részvény	Idősor jellege (megfigyelt értékek száma)		
	csak hétköznap ($T = 83$)	hétvégén 0 hozam ($T = 120$)	spline interpolációval kiegészített idősor ($T = 118$)
MOL	0,9109	0,9176	0,9040
MTELEKOM	0,4704	0,4695	0,5017
OTP	1,2715	1,2658	1,6214
RICHTER	1,0825	1,0771	1,2744

4. Konklúziók

Az információs társadalom egyik legszembeötlőbb jellemzője, hogy a gazdasági-társadalmi jelenségek elemzői óriási adatáradattal találkoznak kutatásaik, illetve munkájuk során. Miközben a gazdaság modellezése szempontjából szerencsés, hogy az adatállományok egyre nagyobbak, az idősorok pedig hosszabbak, aközben nem feledkezhetünk meg arról, hogy az információáradatnak negatív oldala is van: az adatok minősége mindinkább romlik. Az adatminőség kategóriája rendkívül összetett fogalom, esetünkben nem a hivatalos statisztikában szokásos jelentéstartalommal használjuk a kifejezést, hanem a modellező szemszögéből értelmezzük azt. Modellezési szempontból egyik kifejezetten kedvezőtlen, viszonylag alaposan körüljárt tulajdonsága a nagy adathalmazoknak, hosszú idősoroknak a volatilitás (változékonyság, operacionalizálva a szórás) növekedése. Emellett számos további, a modellépítés szempontjából kényelmetlen jelenséggel is szembe találhatjuk magunkat: outlierek jelennek meg, strukturális törések alakulnak ki stb. Ebben a tanulmányban is olyan tulajdonságot vizsgáltunk, amely az információs társadalomban vált mindennapivá.

A rendszertelen idősorok kialakulásának egyik legfontosabb oka, hogy már nem csak egy – korábbi vélekedésünk szerint megfellebbezhetetlen – adatszolgáltató, jól tervezhető adatközléseire támaszkodunk, így gyakran szembesülhetünk nemekvidisztáns idősorokkal. A rendszertelenül megfigyelt jelenségek modellezése során a

modellező alapdilemmája a következő: a megfigyelt értékek egy részének kihagyásával, információt veszítve ugyan, de a hagyományos eszközökkel modellezhető idősort vizsgáljunk-e, vagy a rendszertelenséget valamilyen adekvát eszközzel kezelve valamennyi empirikus megfigyelés alapján készítsünk modellbecslést? Ebben a dolgozatban azt mutattuk be, hogy az adatpótlás módszerének megválasztása fontos feladat a rendszertelen idősorok modellezése során, ugyanis a nem szerencsés adatkiegészítések akár az idősor alaptulajdonságait is megváltoztathatják.

A bemutatott spline-interpolációs eljárás viszonylag egyszerűen átlátható, ráadásul a standard programcsomagok jelentős része támogatja, ezért ajánlható a modellezőknek, ha az adatvesztést is el kívánják kerülni, de arra is vigyázni akarnak, hogy az idősor alaptulajdonságai ne torzuljanak. Minden korábban bemutatott pozitív tulajdonság ellenére érdemes óvatosságra is inteni, ugyanis az adatpótlás (még ha körültekintően történik is!) azt a veszélyt hordozza, hogy nem csak az empirikus adatokra támaszkodunk következtetéseinkben. Milyen mértékben engedhető meg az, hogy a modellező „tisztított" adatokra épített konstrukcióra alapozza döntéseit? A kérdés már a statisztikai etika témakörébe tartozik, és a válasz messze meghaladja jelen tanulmányunk kereteit!

Irodalom

BÉLYÁCZ I. [2009]: *Befektetési döntések megalapozása*. Aula Kiadó. Budapest.

BERGSTROM, A. R. [1985]: The Estimation of Parameters in Non-Stationary Higher-Order Continuous-Time Dynamic Models. *Econometric Theory*. Vol. 1. No. 2. pp. 369–385.

BROCKWELL, P. J. [2001]: Continuous-Time ARMA Processes. In: *Shanbhag, D. N. – Rao, C. R.* (eds): *Handbook of Statistics 19*; *Stochastic Processes: Theory and Methods*. Elsevier. Amsterdam. pp. 249–276.

CATMULL, E. – ROM, R. [1974]: *A Class of Local Interpolating Splines*. In: *Barnhill, R. E. – Reisnfeld, R. F.* (eds.): *Computer Aided Geometric Design*. Academic Press. New York. pp. 317–356.

COCHRANE, J. H. [2012]: *Continuous-Time Linear Models.* National Bureau of Economic Research Working Paper. No. 5807. Cambridge.

DOOB, J. L. [1953]: *Stochastic Processes*. Wiley. New York.

ENGLE, R. F. [1996]: *The Econometrics of Ultra-High Frequency Data*. National Bureau of Economic Research Working Paper. No. 5816. Cambridge.

ENGLE, R. F. – GRANGER, C. W. J. [1987]: Co-integration and Error Correction: Representation, Estimation and Testing. *Econometrica.* Vol. 55. No. 2. pp. 251–276.

GRANGER, C. W. J. [1988]: Some Recent Developments in a Concept of Causality. *Journal of Econometrics*. Vol. 39. No. 2. pp. 199–211.

HANSEN, L. P. – SARGENT, T. J. [1991]: Prediction Formulas for Continuous-Time Linear Rational Expectations Models. In: *Hansen, L. P. – Sargent, T. J.* (eds.): *Rational Expectations Econometrics.* Westview Press. Boulder. pp. 209–218.

HUNYADI L. [1994]: Egységgyökök és tesztjeik. *Szigma*. 25. évf. 3. sz. 135–164. old.

HUNYADI L. [2002]: Grafikus ábrázolás a statisztikában. *Statisztikai Szemle*. 80. évf. 1. sz. 22–52. old.

JONES, R. E. [1985]: Time Series Analysis with Unequally Spaced Data. In: *Hannan, E. J. – Krishnaiah, P. R. – Rao, M. M.* (eds.): *Handbook of Statistics. Vol. 5. Time Series in the Time Domain*. Elsevier. North-Holland, Amsterdam. pp. 157–177.

KOPÁNYI SZ. [2010]: *A hozamgörbe dinamikus becslése*. Phd-értekezés. Budapesti Corvinus Egyetem. Budapest.

LOMB, R. [1976]: Least-Squares Frequency Analysis of Unequally Spaced Data. *Astrophysics and Space Science*. Vol. 39. No. 2. pp. 447–462.

MÉSZÁROS J. [2011]: *Numerikus módszerek*. Digitális Tankönyvtár. Miskolci Egyetem. Miskolc.

RAPPAI G. [2013]: *Bevezető pénzügyi ökonometria*. Pearson Education. Harlow.

SCHMITZ, A. [2000]: *Erkennung von Nichtlinearitäten und wechselseitigen Abhängigkeiten in Zeitreihen.* WUB-DIS-2000-11. Wuppertal. http://juser.fz-juelich.de/record/154233/files/FZJ-2014-03612.pdf

SCHOENBERG, I. J. [1946]: Contributions to the Problem of Approximation of Equidistant Data by Analytic Functions. *Quarterly Applied Mathematics. Parts A and B.* Vol. 4. No. 1. pp. 45–99. pp. 112–141.

VARGA, J. – RAPPAI, G. [2002]: Heteroskedasticity and Efficient Estimates of BETA. *Hungarian Statistical Review*. Special Number 7. pp. 127–137.

WANG, Z. [2013]: cts: An R Package for Continuous Time Autoregressive Models via Kalman Filter. *Journal of Statistical Software*. Vol. 53. No. 5. http://www.jstatsoft.org/v53/i05/paper

Summary

The study examines spline-interpolation, a special class of interpolation techniques. When any unequally-spaced time series is refilled with its mean or a lagged or interpolated value, the new, evenly-spaced time series often has a different data generating process compared to the original one. The paper places focus on the fact that irregular time series cannot be supplemented mechanically. It also presents the procedures for irregular empirical time series and formulates some general conclusions.