Budapesti Műszaki és Gazdaságtudományi Egyetem
Villamosmérnöki és Informatikai Kar
Méréstechnika és Információs Rendszerek Tanszék

Bartók Ferenc

TŐZSDEI IDŐSOROK ELEMZÉSE ÉS ELŐREJELZÉSE

KONZULENS

Dr. Pataki Béla

BUDAPEST, 2014

Tartalomjegyzék

Jelölések

A dolgozat során gyakran használt jelölések összefoglalva a következők:

- t: időváltozó (diszkrét)
- X_t , Y_t: idősor t darab időponttal
- μ: átlag, várható érték
- σ, σ^2: szórás, szórásnégyzet
- ε_t: hibatag, független és azonos eloszlású véletlen változó normál eloszlásból mintavételezve, 0 átlaggal - $\varepsilon_t \sim N(0, \sigma^2)$
- φ: autoregresszív modell paramétere
- θ: mozgó átlag modell paramétere
- c: konstans
- B: „backward shift” operátor
- τ : módosított Thompson Tau módszernél használt érték
- RMSE: átlagos négyzetes hiba gyöke (Root Mean Square Error)
- MAE: átlagos abszolút hiba (Mean Absolute Error)
- AIC: Akaike Information Criterion
- ACF: autokorrelációs függvény (AutoCorrelation Function)
- ARCH, GARCH: Autoregressive Conditional Heteroskedasticity, Generalized ARCH

1 Bevezetés

„A pénzügyi előrejelzés, vagy kifejezetten a tőzsdei előrejelzés manapság az egyik legfelkapottabb kutatási terület.” [1] Ezt bizonyítja az is, hogy rengeteg új tanulmány található ebben a témakörben.

Szeretném kihangsúlyozni, hogy bár első ránézésre e témakörrel kapcsolatban sokaknak az jut eszébe, hogy „nem igen van értelme ilyesmivel foglalkozni”, vagy „úgy se lehet értelmes eredményt produkálni”, mégis érdemes ezt mélyebben átgondolni. Személyes tapasztalatom, hogy nem csak egyszerű egyéni próbálkozások vannak e téren, hanem komoly bankok is alkalmaznak komoly szakértőgárda által megalkotott (különböző) előrejelző rendszereket, amik segítik a működésüket. Ezek a rendszerek értelemszerűen nem publikusak. Továbbá a téma kutatását indokolja az is, amit az összefoglalóban is említettem: pontosabb előrejelzések segítségével jobb gazdasági döntések hozhatóak, melyek a források jobb allokációjához vezetnek, ami a fejlődést segíti, gyorsítja. Mindezek mellett azt is fontos megemlíteni, hogy az itt kifejlesztett módszerek más alkalmazási területekre is átültethetőek – főként olyan területekre, ahol hasonló jellegű idősorokról van szó.

Alapvetően 2 nézet terjedt el a pénzügyi, tőzsdei idősorok előrejelzésével kapcsolatban [2]. Az egyik az úgynevezett hatékony piac hipotézis (Efficient Market Hypothesis - EMH) elméletre épít, ami ugyanis kimondja, hogy az aktuális piaci árak teljes mértékben a valós árat tükrözik, azaz minden információ bele van „kódolva” az árba. Így tehát a befektetők nem tudják folyamatosan „legyőzni” a piacot, vagyis nem képesek az általagosnál jobb eredményt elérni. A hatékony piac hipotézisnek 3 alcsoportja van (gyenge, félerős, erős), amikről bővebben itt [2] olvashatunk. Ezzel szemben helyezkedik el a másik nézet, ami azt mondja ki, hogy különböző elemzési módszerekkel mégis csak „legyőzhető” a piac. Ez a kérdés már hosszú évek óta nyitott.

A tőzsdei idősorok tulajdonságai (kevés minta, magas zaj, nemstacionárius, nemlineáris, véletlenszerű, kaotikus…) rendkívül megnehezítik az előrejelzést. A dolgozatban egy eddig kevésbé vizsgált megközelítést alkalmazok a problémára. Az alapötlet az, hogy kilógó értékek és változási pontok detektálásával, majd ezen értékek külön kezelésével építem fel a modellt. A kutatásom elején nem találkoztam hasonló megközelítéssel, majd rátaláltam Yang, Huang, Chan, King és Lyu [3] munkájára, ahol

hasonló ötlettel próbálkoztak. A fő különbség az, hogy az imént hivatkozott cikkben a szerzők nem választották külön a változási pontot és a kilógó értéket.

Az munkámhoz az adatokat az ingyenesen elérhető Yahoo Finance weboldalról töltöttem le.

A bevezetés végén szeretném megjegyezni, hogy a dolgozatban megtartom az átvett ábrákon az angol nyelvű elnevezéseket, illetve néhol a szövegben is előfordulhat ilyen.

2 Elméleti háttér

2.1 Idősorokról általánosan

2.1.1 Idősor bevezetés

Idősor alatt megfigyelések egy sorozatát értjük [13]. Ezek a megfigyelések, mérések tipikusan azonos időközönként történnek, történtek [14]. Ilyen időköz lehet például naponta, óránként, percenként, stb.. Előfordulhat, hogy egy idősor nem csak azonos időközönkénti méréseket tartalmaz, de ez ritkának mondható. Az idősort az adatokhoz tartozó időbélyeg különbözteti meg a szekvenciális adatoktól, adatbázisoktól (utóbbinál időbélyeg nincs).

Az idősorokat többféleképpen is szokták reprezentálni. Ezek közül az egyik elemi reprezentáció egy *t* hosszú idősorra a következőképp néz ki:

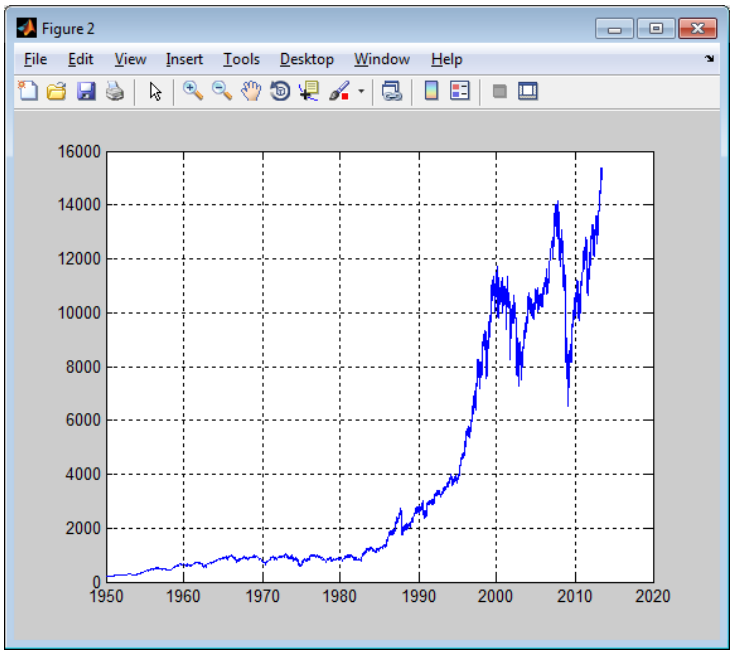
$$\{X_t : t = 1,2 \dots\}$$

Egy másik reprezentáció lehet például a következő:

$$X = F(t)$$

Azaz az X idősor a *t* függvénye.

Az idősorok ábrázolása egy mért érték esetén tipikusan úgy történik, hogy egy koordinátarendszerben a vízszintes tengelyen szerepel az idő (t), a függőleges tengelyen pedig a mért érték. Erre látható egy példa az alábbi ábrán:



1. ábra Dow Jones tőzsdeindex (naponkénti záró értékek 1950 és 2013 között)

Az idősorokat több szempont alapján is fel lehet osztani. Az egyik alapszempont az attribútumok száma szerinti felosztás. Ekkor beszélünk egyváltozós (univariate), és többváltozós (multivariate) idősorokról. Egyváltozós idősor lehet például a fenti képen látható idősor, ahol csak a Dow Jones index napi záró értékeit látjuk az idő függvényében. Másik példa lehet a hőmérséklet ábrázolása az idő függvényében. Többváltozós idősor esetén a tőzsde napi záró értéket kiegészíthetjük például a napi nyitó értékkel, napi kereskedett mennyiséggel, stb., így több megfigyelt változót kapunk. (Értelemszerűen többváltozós idősor lehetne például egy tőzsde napi záró értéke és a levegő napi átlaghőmérséklete is.)

Az idősorok felosztásánál egy másik fontos szempont az idősor stacionaritása szerinti vizsgálat. Ez alapján beszélünk stacionárius és nemstacionárius folyamatokról, illetve esetünkben idősorokról.

A stacionárius idősorok olyan idősorok, amelyek statisztikai tulajdonságai nem változnak az idő haladása során [17]. Ilyen tulajdonság például az átlag, variancia, autokorreláció. A nemstacionárius idősorok statisztikai tulajdonságai ezzel szemben változnak az idő haladása során (nem konstans például a variancia). Ezt okozhatják például trendek, ciklusok, véletlen bolyongások (random walk) [18]. A legtöbb statisztikai előrejelző módszer alapja az, hogy különböző matematikai transzformációkkal (ezek később bemutatásra kerülnek) közel stacionáriussá alakítják az idősort, majd maga az előrejelzés azt veszi alapul, hogy a statisztikai tulajdonságai

nem változnak meg a transzformált idősornak. Az előre jelzett eredményt utána visszatranszformáljuk az eredeti tartományba.

Az idősoroknak több fontos tulajdonsága is van. Ezek közül érdemesnek tartom kiemelni azt, hogy bizonyos idősorok esetén az egymást követő megfigyelések erősen korrelálnak egymással [13]. Erre egy jó példa a hőmérséklet mérése. Például a 10 óra 10 perckor mért érték és a 10 óra 11 perckor ugyanazon a napon mért érték erősen korrelál egymással. Ezzel hasonlóan tőzsdei idősorokra általánosságban igaz az, hogy az időben közel álló értékek erősen korrelálnak egymással (ezt tanúsítja a később bemutatott ACF [29][30] például a Dow Jones indexen).

Az idősoroknak rendkívül sok alkalmazási területe van. Az idősorok elemzése alapján elsősorban előrejelzést szokás elvégezni, de a múltbeli adatok vizsgálata is sok új és érdekes információt feltárhat. Tipikus alkalmazási területei a következők (a teljesség igénye nélkül): tőzsde, időjárás, közlekedés, természeti katasztrófák, víz-, gáz-, áramfogyasztás, népesség stb..

2.1.2 Modellezés, pénzügyi idősorok csoportosítása

A modellezés segítségével betekintést nyerhetünk azon mechanizmusokba, amelyek generálják az idősort [14]. Fontos kérdés a modell bonyolultsága, lehetőleg arra kell törekedni, hogy egyszerű modellt kapjunk (Occam borotvája elv).

Niels Bohr mondta a következőt: *„Prediction is very difficult, especially if it's about the future."* [17]. Azaz magyarul: Az előrejelzés nagyon nehéz, különösen, ha a jövőről szól. Ez egy fajta figyelmeztetésként szolgálhat, hogy könnyű olyan modellt találni, ami a múltbeli adatokra jól (vagy túl jól) illeszkedik, de az már jóval nehezebb, hogy megfelelő előrejelzést legyünk képesek végezni.

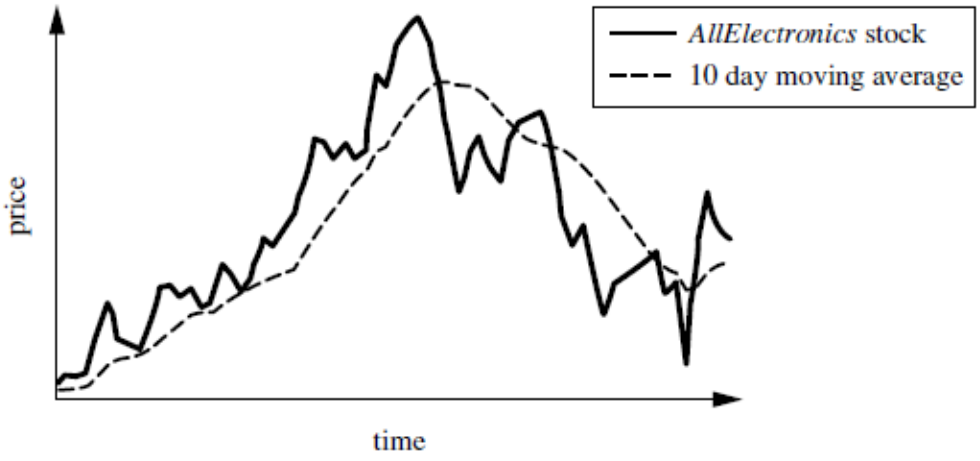
Idősorok modellezésére sokféle módszer létezik. Idősorok modellezése alatt legtöbbször az idősor dekomponálását értjük a trendanalízis 4 alapkomponensébe. Ezt a modellt így jelölik [14]:

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

(megjegyzés: szorzatjel helyett összeggel is jelölik)

A trendanalízis 4 alapkomponensének (T, C, S, I) a jelentése alább olvasható:

- T, azaz trendmozgás: Ez az általános irányt adja meg hosszabb időszakokra. A trendmozgást egy úgynevezett trend görbével, vagy trend egyenessel szokás megadni. A trend meghatározására szokták használni például a mozgó átlagot, ami az alábbi ábrán is látszódik.



2. ábra Trend [14]

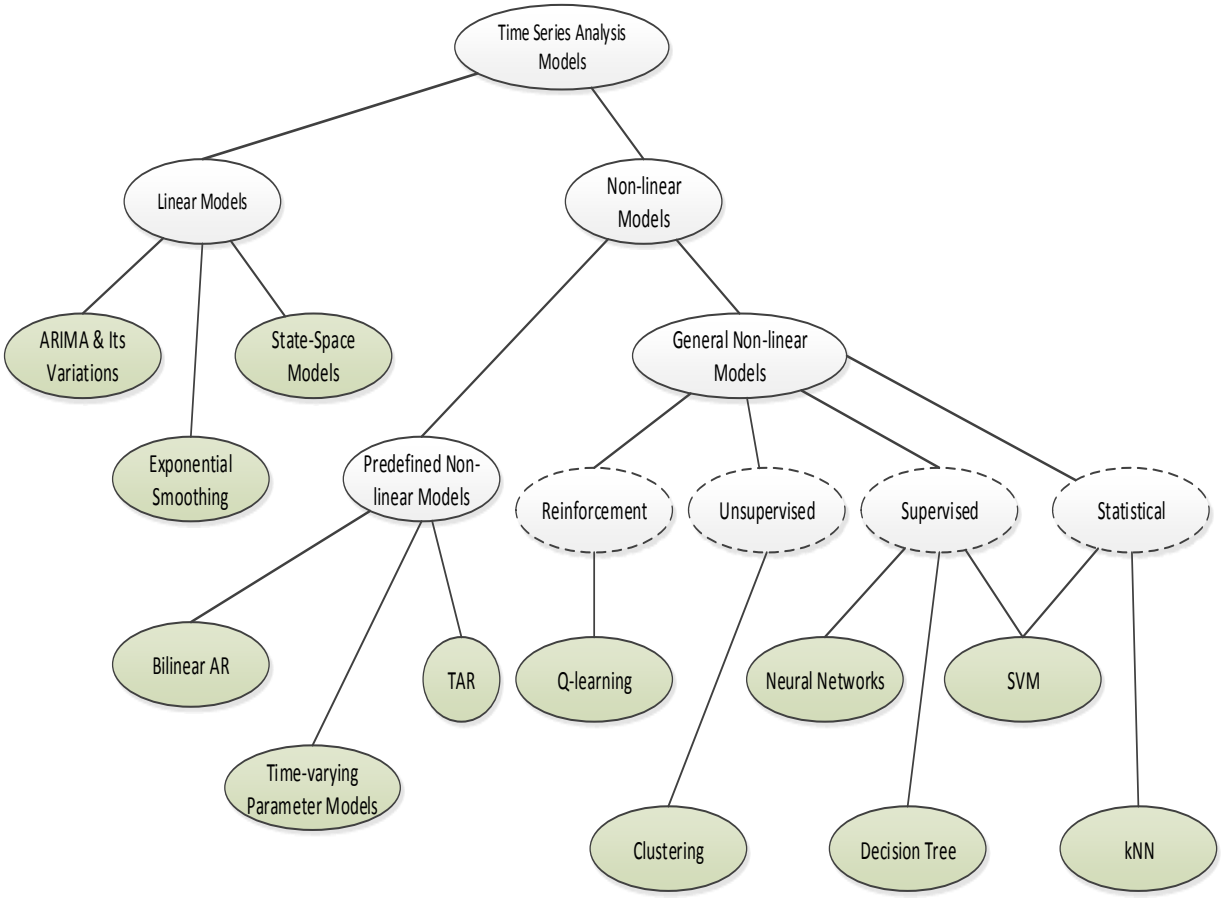
- C, azaz ciklikus mozgás: Ez a mozgás ciklusokat jelent az idősorban. Ezek a ciklusok hosszú távú változások a trend körül. Fontos tulajdonsága a ciklikus mozgásnak, hogy lehet periodikus, de nem periodikus is.
- S, azaz szezonális mozgás: Ezek rendszeres, vagy naptárhoz köthető mozgások. Hasonló a ciklikus mozgáshoz, de fő különbsége, hogy ez mindig periodikus és a periódus ideje nem lehet több 1 évnél.

  Példák: Karácsony előtti nagy vásárlások, vagy nőnapi virágeladás növekedése minden évben.
- I, azaz irreguláris mozgás: Ez a mozgás a véletlenszerű eseményeket jelenti.

  Példák: természeti katasztrófa, háború, tőzsde manipulálása, stb..

Ezen általános modell után kitérek a konkrétabb modellek egyfajta csoportosítására. Az idősor analízishez használt modelleket alapvetően 2 csoportra, **lineáris** és **nemlineáris** modellekre lehet osztani [4]. Az irodalomkutatásom során számos, idősorokhoz használt modellel találkoztam. Egy összefoglalót olvashatunk

különböző modellekről Yang, King, Chan és Huang cikkében [4]. A következőkben ezt az összefoglalót dolgozom fel az idősorokhoz használt modellek csoportosításának bemutatásához, mely csoportosítást az alábbi képen láthatunk:



3. ábra Modellek csoportosítása [4]

A lineáris modellek egyszerűek és könnyen alkalmazhatóak. Az ARIMA modell egy tipikus lineáris modell, melyet gyakran használnak összehasonlítási alapként. A valós életben gyakran előfordul azonban, hogy a vizsgált idősor nemlineáris tulajdonságokkal rendelkezik, emiatt a kutatók vizsgálódtak nemlineáris modellekkel is. Ilyenek például a bilineáris autoregresszív, küszöb autoregresszív (TAR) vagy egyéb időben változó paraméterű modellek. A szerzők ezeket a modelleket az úgynevezett előre definiált osztályba sorolják. Ezen csoport mellett beszélhetünk az általános nemlineáris modellekről (ezt a csoportot gépi tanulással létrehozott modelleknek is nevezik). Ennek a csoportnak több alosztálya is van (megerősítéses, nem felügyelt, felügyelt, statisztikai). Ezek közé tartozik például a Q-tanulás, klaszterezés, döntési fa, neurális hálózatok, szupport vektor gépek, vagy a k-legközelebbi szomszéd modell.

Érdemesnek tartom megjegyezni, hogy ezen modellek mellett gyakran találkoztam az úgynevezett ARCH, GARCH modellekkel főként gazdasági idősorokkal kapcsolatban.

A következőkben bemutatom a nemstacionárius idősorok típusait, melyek bizonyos mértékben magukban hordozzák az általános modellnél bemutatott változókat, mozgásokat. A **pénzügyi** idősorokat ilyen szempontból a következő típusokra szokás bontani [18]:

- Véletlen bolyongás (random walk):

$$X_t = X_{t-1} + \varepsilon_t$$

A véletlen bolyongás azt írja le, hogy a *t*-edik időpillanatban az idősor értéke egyenlő lesz az előző értékkel plusz egy 0 átlagú fehér zajjal (hibatag). Erre a folyamatra jellemző, hogy a varianciája változik az idő haladásával, illetve, hogy nem az átlaghoz visszatérő (mean reverting) folyamatokhoz tartozik.

- Véletlen bolyongás sodródással (random walk with drift):

$$X_t = c + X_{t-1} + \varepsilon_t$$

Az előzőhöz hasonló jellemzői vannak ennek a típusnak. Annyiban különbözik, hogy egy plusz konstans megjelenik az egyenletben.

- Determinisztikus trend:

$$X_t = c + \beta_t + \varepsilon_t$$

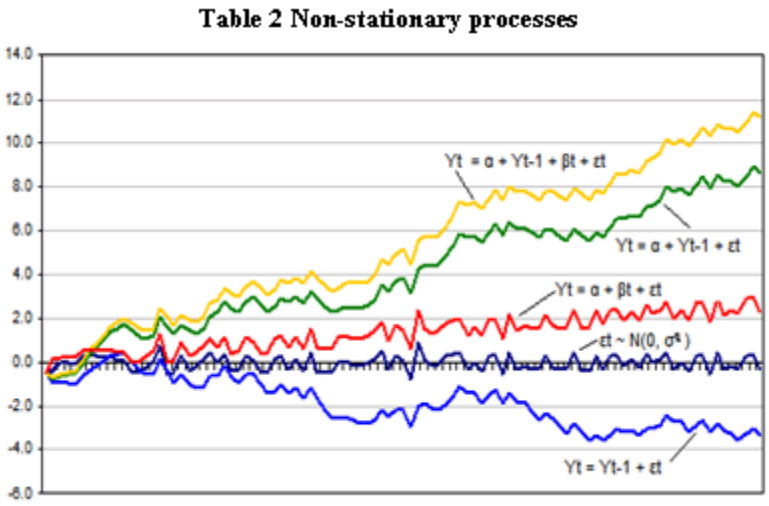
Az előző típustól abban tér el, hogy a *t*-edik időpillanatban nem az előző időpillanatbeli értéket, hanem egy időtől függő trendet használ fel. Ebben az esetben az átlag folyamatosan nő (vagy csökken) a trendnek megfelelően.

- Véletlen bolyongás sodródással és determinisztikus trenddel (random walk with drift and deterministic trend):

$$X_t = c + X_{t-1} + \beta_t + \varepsilon_t$$

Ez a típus kombinálja az előző 2 típust, mint ahogy az a képletből és a nevéből is kivehető.

A következő ábrán az imént tárgyalt nemstacionárius idősorok típusaira láthatunk példákat:



4. ábra Nemstacionárius idősorok [18]

2.1.3 Előrejelzés

Az idősorelemzés egyik legfontosabb alapfeladata az előrejelzés. Előrejelzéseket folyamatosan használnak az üzleti életben, pénzügyben, közgazdaságtanban, közigazgatásban, és sok egyéb területen is [15]. Előrejelzéseket azért készítünk, hogy segítsék a döntéseinket különböző területeken. Az előrejelzés szorosan összefügg a statisztikai modellek építésével. Ahhoz, hogy képesek legyünk egy változót előre jelezni először építenünk kell egy modellt és meg kell becsülnünk a modell paramétereit a megfigyelt múltbeli adatok alapján. Az előrejelzés azon alapul, hogy keresünk egy olyan matematikai formulát vagy eljárást, amely megközelítőleg generálja a múltbeli adatokat [14].

Diebold [15] 6 kérdéstípust fogalmaz meg, amelyek fontosak bármilyen előrejelzéssel kapcsolatban (az alábbi felsorolást és kérdéseket nagyrészt szó szerint vettem át):

1) (Döntési környezet és veszteség függvény) Milyen döntést fog segíteni az előrejelzés? Mik a következményei az előrejelzési modell használatának és értékelésének? Hogyan számszerűsítünk egy „jó” előrejelzést? Illetve hogyan számszerűsítjük az előrejelzési hiba okozta költséget vagy veszteséget? Hogyan számítunk ki optimális előrejelzést?
2) (Jósolandó objektum) Mit kell előre jelezni? Mekkora az adat mennyisége, milyen a minősége? Egy változót vagy több változót jóslunk? Vannak hiányzó megfigyelések?

3) (Előrejelzés megfogalmazása) Például idősorok esetében egy egyszerű legjobb becslés a fontos számunkra; vagy a lehetséges jövőbeli értékek egy ésszerű tartománya fontos, amely tükrözi az előrejelzési probléma bizonytalanságát; vagy a lehetséges jövőbeli értékek egy valószínűségi eloszlása a fontos?

4) (Előrejelzési határ) Mekkora távolságra jelezzünk előre és miért? A legjobb modellezési és előrejelzési módszer valószínűleg függ a távolságtól.

5) (Információs halmaz) Milyen információra alapozzuk az előrejelzést? Létezik elérhető adat az előrejelzendő idősorhoz?

6) (Módszerek és komplexitás) Melyik előrejelzési módszer a legjobb az adott problémára? Milyen komplex legyen a modell?

Az előrejelzések részletes ismertetése nem célja ezen fejezetnek. Az előző fejezetben a modellek áttekintése által példákat láttunk különböző előrejelzésre alkalmas modellekre.

2.2 Példák ARIMA alkalmazásokra a szakirodalomban

A munkám során az ARIMA modellel dolgoztam. Ebben a fejezetben néhány olyan alkalmazási példát mutatok be, ahol a kiugró érték és változási pont detektálása szerepel az alkalmazásban, vagy magának az idősornak az előrejelzéséről van szó. Továbbá néhol röviden kitérek az áttekintett munkákban szereplő kiugró értékek detektálásával kapcsolatos információkra is.

Watson, Tight, Clark és Redfern [6] áttekintik a főként szállítással, utazással, forgalommal kapcsolatos idősorokban használt módszereket kilógó érték detektálására és hiányzó érték kezelésére (ezen idősorok nem feltétlen rendelkeznek hasonló tulajdonságokkal, mint a gazdasági, tőzsdei idősorok). Megemlítik, hogy (addig) a legtöbb módszer az idősor Box-Jenkins ARIMA struktúrában való reprezentálását feltételezi, és ezeket a modelleket sokféle idősorra alkalmazzák.

Alább olvasható egy példa a cikkben feldolgozottak közül:

- 1966-1976-ig terjedő havi forgalmi adatokkal próbálták előre jelezni az 1977-es évet. Ezekben az adatsorokban kilógó értéket okozhat például baleset, időjárás, karbantartás, stb.. Végül egy ARIMA(12,1,7) modellt dolgoztak ki, és 12 hónappal előre jelezve 5% körüli hibákat kaptak. Megemlítik, hogy ez a modell elég magasrendűnek számít. A hibát kis mértékben tovább tudták csökkenteni komplexebb modellekkel.

Thalassinos és Pociovalisteanu [12] 1997-2007-ig terjedő időszakban a román tőzsde napi adatait (2507 megfigyelés) felhasználva vizsgálták az Európai Unió különböző intézkedéseinek és egyéb külső események hatását (például az új lej bevezetése). Először modellnek egy egyszerű regressziót alkalmaztak (az AIC, Lagrange paraméterekkel finomítottak a modellen). Továbbá elvégezték a Dickey-Fuller tesztet stacionaritás vizsgálata céljából. Különböző problémák miatt áttértek az ARIMA modellre, és végül az ARIMA(1,1,0) modellt választották. Ezzel a modellel mindig 17 nappal előre jelezve 1 hónapos időtartamban a legnagyobb hiba -4.23% volt. Ezek mellett elvégezték a CUSUM és CUSUMQ teszteket struktúra változás detektálásához. A struktúra változás időpontjának 2001 második félévét kapták eredményül, amit hivatkozással meg is magyaráztak. Látható az alábbi ábrán, hogy az idősor tulajdonságai megváltoztak ebben az időszakban.

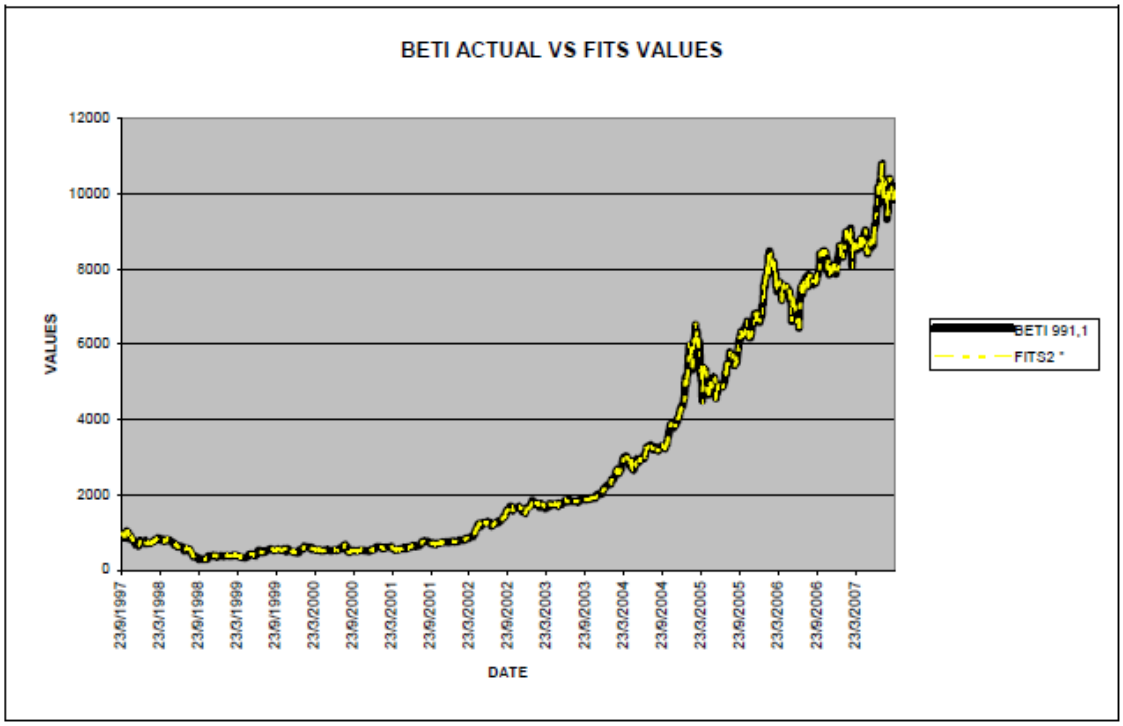


Diagram 7: ARIMA (1,1,0) Model, Actual vs Fits Values

5. ábra A román tőzsde 1997-2007 között ARIMA modell illesztéssel [12]

Tsay [10] 3 valós példát mutatott be, ahol ARIMA modell volt az alapja a kiugró érték és a változási pont detektálásának:

- Havi légiforgalmi megtett kilométerek az adatok 1960-tól 1977-ig (216 megfigyelés). Először logaritmizálták az idősort a variancia stabilizálása érdekében, majd egy ARIMA modellt javasoltak az idősorra. A kilógó érték és változási pont detektáló algoritmus 6 szintváltást és 4 additív kilógó értéket talált. Jelentős varianciaváltozást nem talált a másik algoritmus. A kapott időpontokat megvizsgálták és magyarázták.
- Ebben a példában 369 darab IBM záró tőzsdei értékét használták fel. Szintén a logaritmusát használták az adatnak. A használt modell ARIMA(0,1,1). Az előre becsülthöz közeli értéket kaptak a varianciaváltozás vizsgálatánál. Továbbá egy másik a cikkben bemutatott algoritmus 6 lehetséges kilógó értéket és 3 lehetséges változási pontot detektált.
- Ebben a példában tv-vel és rádióval kapcsolatos adatokkal (rendelésekkel) dolgoztak. 274 adatpontot használtak fel. Az algoritmus egy jelentős varianciaváltozást fedezett fel. Az idősort módosítva, a varianciaváltozást stabilnak minősítették, majd futtatták a kilógó érték és változási pont detektáló algoritmust, ami egy additív és két innovatív kilógó értéket talált. Amennyiben a talált varianciaváltozást figyelmen kívül hagyták, úgy 15 kilógó értéket és változási pontot detektált az algoritmus. Problémát okozhat, ha figyelmen kívül hagyjuk a varianciaváltozást.

2.3 Kiugró érték és változási pont detektálása

2.3.1 Definíciók

A kilógó vagy kiugró érték (angolul: outlier) az egy olyan megfigyelés, ami jelentősen eltér ugyanazon minta többi tagjától [5]. Maga a neve is jól mutatja, hogy „kilóg”, „kiugrik” a többi érték közül. A dolgozat során mindkét magyar elnevezést alkalmazom.

Watson, Tight, Clark és Redfern [6] a következőképp fogalmazta meg a kilógó értéket: „A kilógó érték definíciója lehet: egy olyan megfigyelés, ami nem reprezentatív, hamis vagy diszharmonikus. Egy olyan megfigyelésnek is tekinthető, amely nem a célpopulációhoz tartozik.”.

Hawkins úgy definiálja a kilógó értéket, mint „egy olyan megfigyelés, amely annyira eltér a többitől, hogy gyanús, hogy egy másik mechanizmus generálta” [7].

Johnson megfogalmazásában „a kiugró érték az egy olyan megfigyelés az adathalmazban, amely inkonzisztensnek tűnik az adathalmaz többi részével” [7].

Fontos megjegyezni, hogy az irodalomkutatásom során azt tapasztaltam, hogy különböző szerzők különböző elnevezéseket használnak a kilógó értékkel és a változási ponttal kapcsolatban. Előfordul az is, hogy a kilógó érték név alá veszik a változási pontot is, és a kilógó értékeket nem különböztetik meg. Összességében el lehet mondani, hogy miközben ugyanazon alaptípusokról beszélnek, igen sokféle elnevezés, besorolás terjedt el, nincs egy egységes elfogadott rendszer.

Fox a kilógó értékeket 2 típusra osztotta fel [8]. Ezeket nevezték el később additív és innovatív kilógó értéknek. Az additív kiugró érték (AO) egyetlen megfigyelésre van hatással [9] (ilyen lehet például a feljegyzési hiba [6]). Ezen zavar után az idősor visszatér a normális kerékvágásba, mintha semmi se történt volna. Az innovatív (IO) kilógó érték viszont egy adott ponttól kezdve az összes megfigyelésre kihat (ez a típus változási pontnak tekinthető).

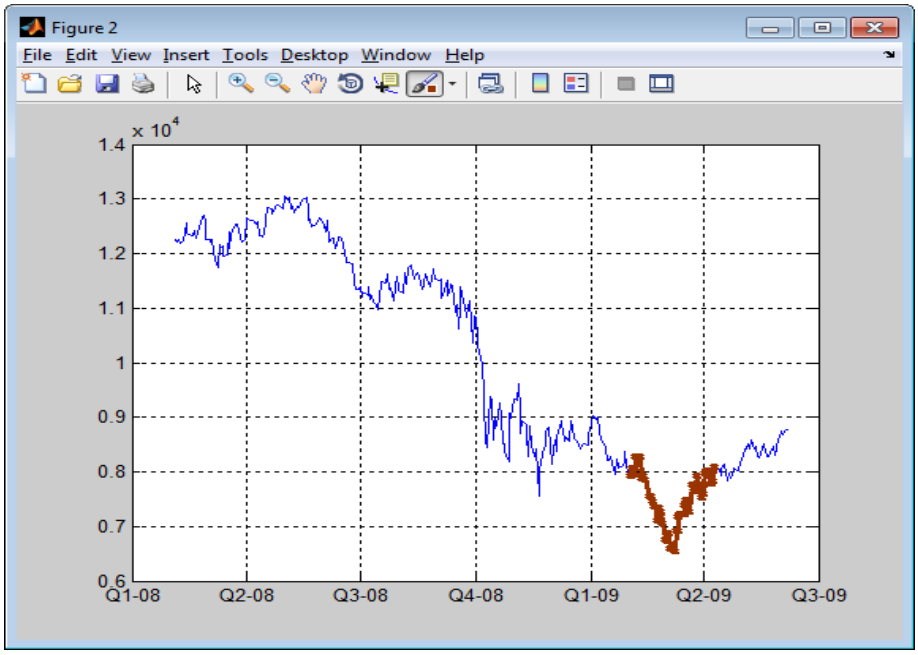
Yamanishi és Takeuchi [11] a változási pontot külön kezelik a kilógó értékektől. Leírásuk szerint a változási pont egy olyan változáshoz köthető, ahol megváltoztak a folyamat statisztikai jellemzői.

Tsay [10] struktúraváltozásokat és kilógó értékeket különböztet meg. A kilógó értékek mellett 2 további típust definiál a struktúra változás halmazán belül. Az egyik a szint változás (level shift - LS), amelynél egy mennyiségbeli szintváltás figyelhető meg egy időpontban, és ez a változás állandó. A másik a varianciaváltozás (variance change - VC), ami pedig a variancia megváltozását jelenti egy adott pontban. A szint változást további csoportokra is osztja a szerző. Ezen típusoknak a formális definíciója megtalálható a hivatkozott cikkben.

A dolgozatom során pontos formális definíciót nem használok annak eldöntésére, hogy egy adott pont változási pont vagy pedig kiugró érték. Általánosan

megfogalmazva, kilógó értéknek tekintem azon értékeket, amelyek feltűnően más értéket vesznek fel, mint a múltbeli környezetükből következne, majd viszonylag rövid idő elteltével az idősor visszatér a múltbeli környezetükből következő szinthez. Ezzel szemben a változási pont nem tér vissza a múltbeli környezetből következtethető szinthez (legalábbis nem rövid időn belül), a statisztikai jellemzők hosszabb távra megváltoznak. A változási pontok és kilógó értékek fogalmánál az „idő rövidségével" kapcsolatban sokáig lehetne vitatkozni, elmélkedni. Én ebbe mélyebben nem kívánok belemenni.

Az alábbi ábrán egy kilógó értékhalmazt láthatunk barna színnel kijelölve:



6. ábra Dow Jones index kilógó érték

Látható, hogy a kijelölt szakasz előtt Q4-08-tól kezdődik egy stacionáriusnak mondható rész, amiből ez a kijelölt szakasz kilóg, majd az idősor Q2-09-től visszatér a Q4-08a-t közvetlenül követően tapasztalható viselkedéshez. Ez nem mondható tipikus kilógó értéknek, de a dolgozatom során az ilyen jellegzetességű ponthalmazokat is annak tekintem. Az ábrán tovább látható Q3-08-nál és Q4-08-nál is egy változási pont.

2.3.2 Javasolt módszerek

A kiugró érték és változási pont detektáló módszereket alapvetően 2 csoportba lehet osztani [7]. Az egyik az egyváltozós, a másik a többváltozós módszerek. Az eddigi munkám során még nem használtam fel a napi záróértékeken kívül más adatot (például napi kereskedett mennyiség), ezért az egyváltozós módszerekre koncentráltam. Egy

másik csoportosítása a módszereknek a parametrikus és a nem parametrikus (modell nélküli) módszerek. Az előbbi megközelítés általában azon megfigyeléseket tekinti kilógó értékeknek, amelyek a leginkább eltérnek a modell feltételezéseiből következő értéktől. Nem parametrikus módszerek közé tartoznak a különböző távolság alapú, és klaszterezési módszerek.

Kilógó érték detektálásához viszonylag magától értetődő egyszerű módszernek tűnik az, amikor a naponkénti (vagy a legkisebb időközönkénti) legnagyobb eltéréseket keressük meg. Ez lényegében egy differenciálást jelent. Ez a módszer persze nem garantálja, hogy a legtöbb kilógó értéket megkapjuk. Vegyük például a 6. ábra Dow Jones index kilógó értékét, ahol nem várhatjuk el, hogy ez a módszer megtalálja a kijelölt kilógó értékhalmazt. Ezen módszer továbbfejlesztése lehet, ha nem csak a naponkénti legnagyobb eltéréseket figyeljük, hanem az eredeti idősor értékeit vonjuk ki egy adott időintervallumra vonatkozó mozgó átlagból.

A módosított Thompson Tau módszer, egy statisztikai módszer a kilógó értékek keresésére [33]. A módszer lépései a következők:

- kiszámoljuk az átlagot és a szórást
- kiszámoljuk minden pontra az abszolút eltérést az átlagtól
- kiszámoljuk a τ értékét a Student féle t eloszlás kritikus értéke alapján (részletes leírás [33])
- ezután pedig a τ szorozva a szórás értékéhez hasonlítjuk az adott megfigyelés abszolút eltérését
- amennyiben az eltérés nagyobb ennél az értéknél a megfigyelést kilógó értéknek jelöljük és töröljük
- majd ezeket a lépéseket ismételjük addig, amíg már nem találunk kilógó értéket

Habár ez a módszer nem idősorokra van specializálva, mert nem veszi figyelembe, hogy melyik adathoz milyen időbélyeg tartozik, a tesztelések során korrekt eredményeket kaptam vele (erről a későbbiekben lesz szó).

Yamanishi és Takeuchi [11] egy olyan módszert dolgoztak ki, amely első lépése az, hogy az algoritmus megtanul egy valószínűségi modellt az adatok alapján. Ezen tanuló algoritmus képes követni a változó adatokat, úgy, hogy a múltbeli információkat

fokozatosan elfelejti. Ez után pontozzák az összes adatot az alapján, hogy mennyire tér el a megtanult modell által jósolttól – ahol a nagyobb pontszám jelzi a kilógó érték nagyobb valószínűségét. A változási pontok keresését úgy végzik, hogy kilógó értéket keresnek azon idősorban, ami a kiszámított pontozás mozgó átlagából áll. A cikkben külön kezelik az idősorokat és az egyéb adatsorokat. Idősorok esetére egy autoregresszív modellt megtanuló algoritmust dolgoztak ki. A pontozási definíciók megtalálhatóak a tanulmányban.

Változási pont detektálására egy lehetséges másik módszer lehet a Kolmogorov-Smirnov teszt alkalmazása az idősorban jelen lévő megfigyelésekre, amellyel azt vizsgálhatjuk, hogy az idősor adott időpontja előtti és utáni szakasz mintái ugyanabból az eloszlásból származnak-e.

3 ARIMA modell és a Box-Jenkins módszer

Az ARIMA modell és a Box-Jenkins módszer igen gyakran előfordul az idősorokkal kapcsolatos tanulmányokban. Ezen ok miatt is döntöttem úgy, hogy megismerkedek ezzel a megközelítéssel.

3.1 ARIMA modell

Az ARIMA az angol AutoRegressive Integrated Moving Average kifejezésnek a rövidítése. Ennek a segítségével idősorokat szoktak modellezni főként előrejelzés céljából [19]. Az ARIMA modell az ARMA modell általánosítása, amely stacionárius sztochasztikus folyamatok leírására szolgál [20][27]. Az ARMA modell 2 fő részből áll: AR(p), MA(q), azaz az autoregresszív és a mozgó átlag modell részből. A p, q paraméterek jelentik a különböző részek fokát, rendjét. Ezek a paraméterek nem-negatív egész számok, és ha valamelyik értéke 0, akkor az a rész kiesik a modellből.

Az autoregresszív modell azt írja le, hogy a változó értéke (lineárisan) függ ugyanezen változó korábbi értékeitől. Egy p rendű autoregresszív modellt a következőképp definiálhatunk [20][21]:

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t$$

Ahol c egy konstans és ε_t a jelöléseknél bemutatott hibatag, φ_i-k pedig a modell paraméterei.

A „backward shift” vagy „backshift” (B) operátor segítségével is felírhatjuk a modellt. Az operátort a következőképp definiáljuk:

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

Általánosabban:

$$B^k X_t = X_{t-k}$$

Ennek segítségével az autoregresszív modell másik felírása a következőképp alakul:

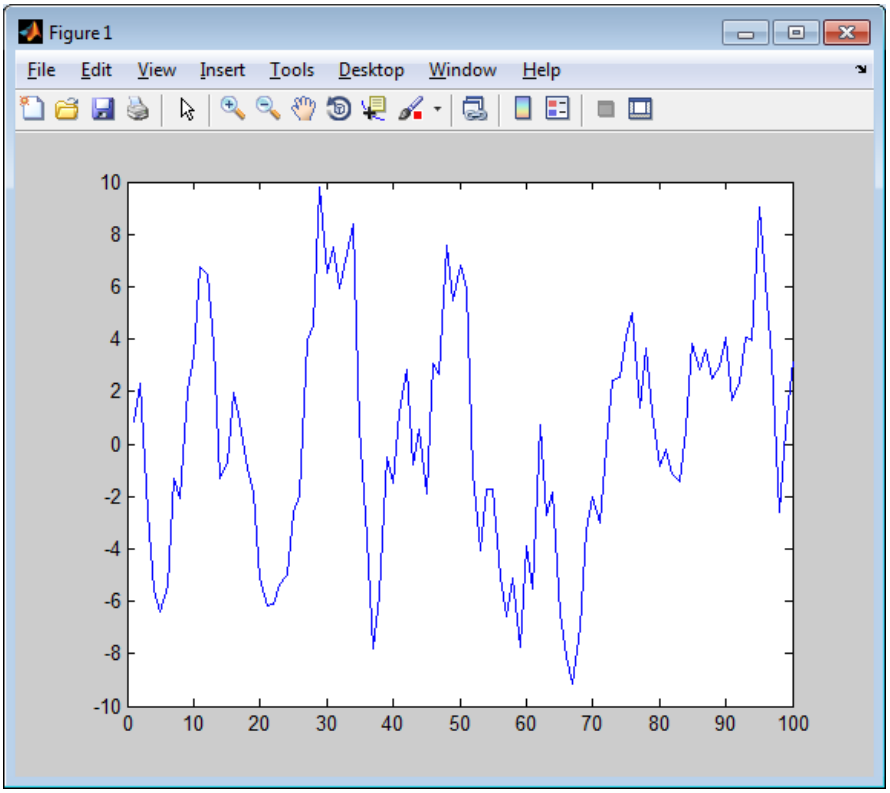
$$X_t = c + \sum_{i=1}^{p} \varphi_i B^i X_t + \varepsilon_t$$

$$\varphi(B) X_t = c + \varepsilon_t$$

Ahol:

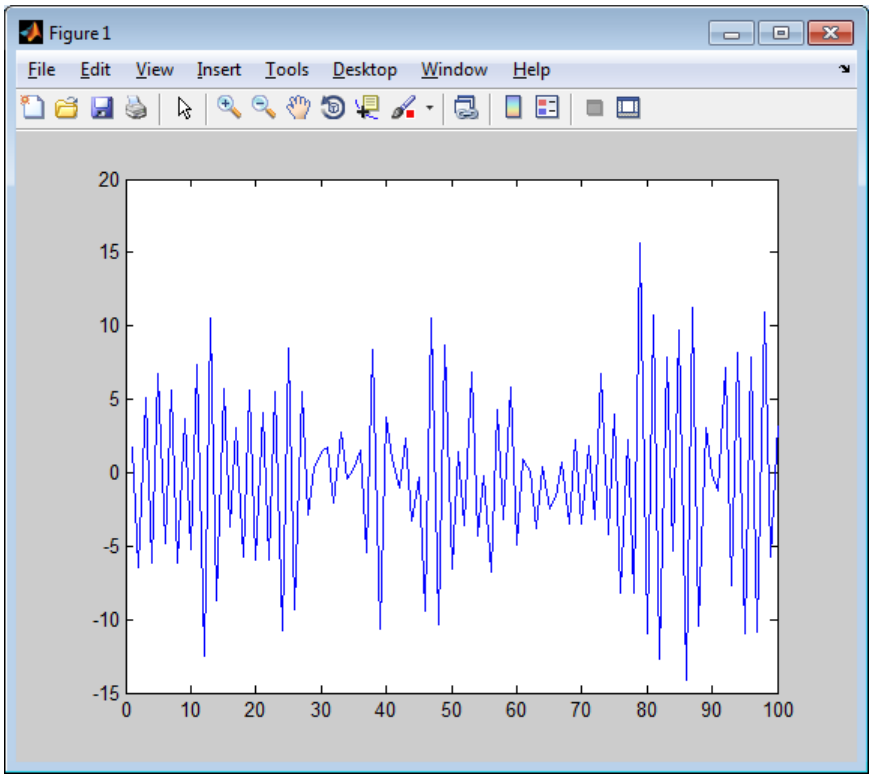
$$\varphi(B) = 1 - \sum_{i=1}^{p} \varphi_i B^i$$

Figyeljük meg, hogy miként viselkedik egy AR(1) folyamat különböző $\varphi_1$értékkel. Először a $\varphi_1$értéke legyen 0.9 (a konstans értéke 0, a variancia pedig 10). Ekkor szimulált adatokkal (100 darab megfigyelés) a következőt kapjuk:



7. ábra AR(1) $\varphi_1 = 0.9$

Látható, hogy ahogy elkezd egy irányba növekedni vagy csökkeni, akkor egy ideig ezt meg is tartja. Ez logikus, mivel az előző érték 0.9-szereséből indulunk ki mindig, amihez hozzáadjuk a zajt. Ezzel szemben egy AR(1) folyamat, ahol a $\varphi_1$értéke -0.9 hasonlóan szimulált adatokkal a következőképp néz ki:



8. ábra AR(1) $\varphi_1 = -0.9$

Látható, hogy itt már egy erős oszcillálás van jelen. Ez is érthető, mivel az aktuális érték mindig az előző közel ellentettjéből indul ki.

A mozgó átlag modell a hibatag korábbi értékeire épít. Egy q rendű mozgó átlag modellt a következőképp definiálhatunk:

$$X_t = \mu + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

Ahol μ az átlag (gyakran 0-nak tekintik), ε fehérzajnak felel meg, θ_i-k a modell paraméterei. A korábban bevezetett „backshift” operátor segítségével az előzőhöz hasonló elven a következőképp írható le a mozgó átlag modell:

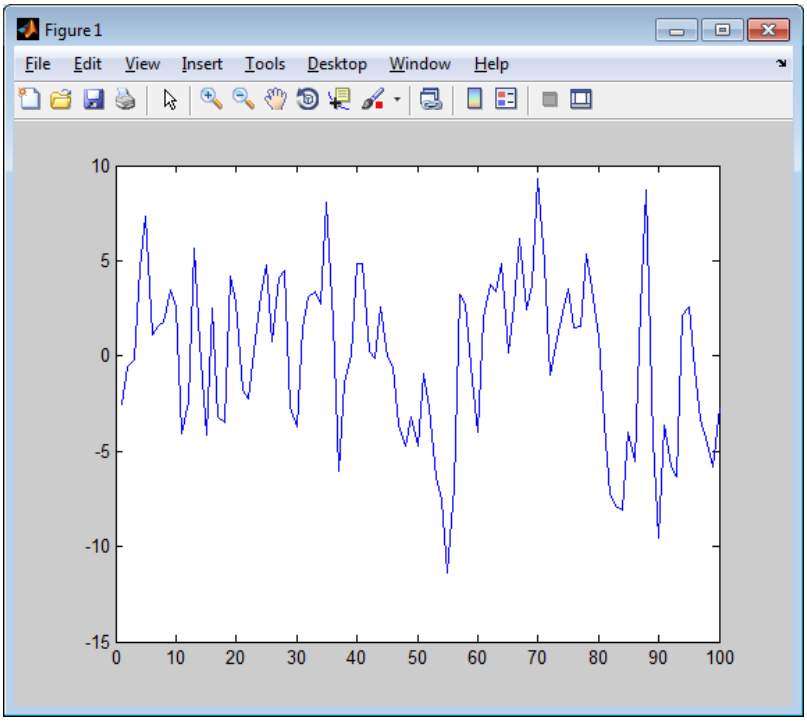
$$X_t = \mu + \varepsilon_t + \sum_{i=1}^{q} \theta_i B^i \varepsilon_t$$

$$X_t = \mu + \theta(B)\varepsilon_t$$

Ahol:

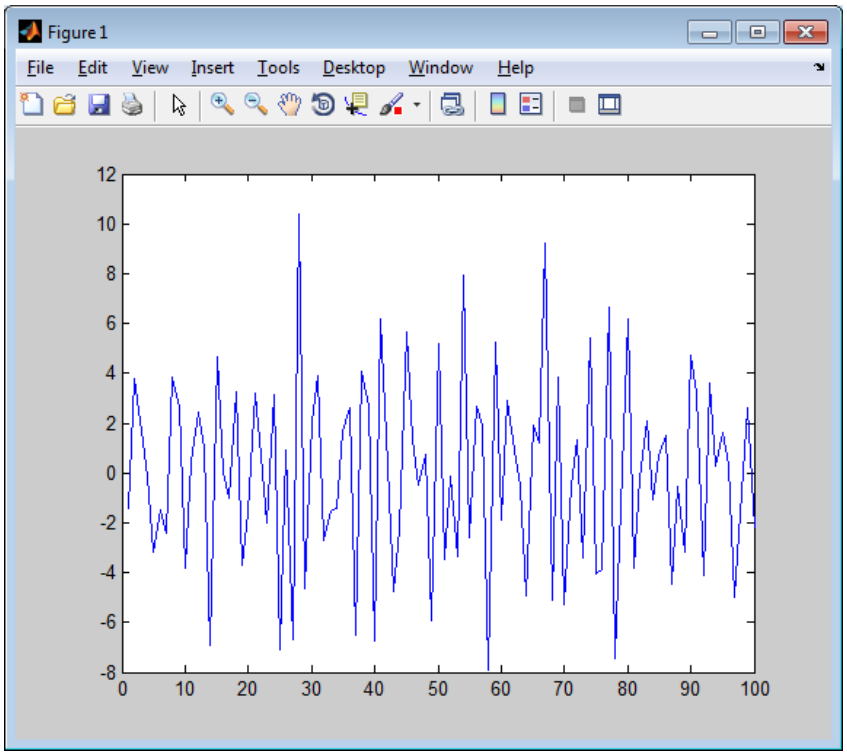
$$\theta(B) = 1 + \sum_{i=1}^{q} \theta_i B^i$$

Figyeljük meg, hogy miként viselkedik egy MA(1) folyamat különböző θ_1 értékkel. Először a $\theta_1$értéke legyen 0.9 (a konstans értéke 0, a variancia pedig 10). Ekkor szimulált adatokkal (100 darab megfigyelés) a következőt kapjuk:



9. ábra MA(1) $\theta_1 = 0.9$

Ez az ábra valamennyire hasonló a 7. ábrán látottakhoz, csak itt „gyorsabban” történik az irányváltás.



10. ábra MA(1) $\theta_1 = -0.9$

Ezen a grafikonon is inkább oszcillálás látható, mint az AR modellnél hasonló értékkel. Természetesen különbségek megfigyelhetőek, mint például a kevésbé sűrű irányváltoztatások.

Egy ARMA(p,q) modellt a következőképp írhatunk le az elsőként bemutatott autoregresszív és mozgó átlag modell ábrázolások segítségével:

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

A „backshift” operátorral leírt modellekből az ARMA(p,q) modell a következő formákban írható fel:

$$\left(1 - \sum_{i=1}^{p} \varphi_i B^i\right) X_t = \left(1 + \sum_{i=1}^{q} \theta_i B^i\right) \varepsilon_t$$

Vagy:

$$\varphi(B) X_t = \theta(B) \varepsilon_t$$

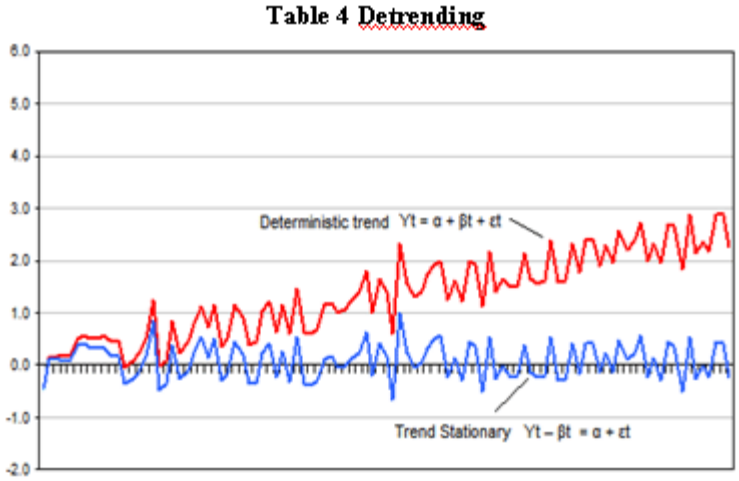
Megjegyzés: Ezekből az egyenletekből a c és a μ változókat kihagytuk.

AR modell esetén egy adott számú előrejelzés úgy történik, hogy az első ismeretlen értéket (legyen X_t) kiszámoljuk az autoregresszív modell képletéből, ahol a zajt 0-nak tekintjük (mivel ez a zaj várható értéke). Ezután a következő értéket (X_{t+1}) hasonlóan számoljuk ki, de az előtte levő ismeretlen érték helyére az elsőnek kiszámolt értéket (X_t) helyettesítjük. Majd ezt folytatjuk a megadott lépésszámig. Az ARMA modell előrejelzése ezen módszer általánosításából történik.

Az ARIMA(p,d,q) modell az ARMA(p,q) modellt egy integráló, I(d) résszel egészíti ki. A d paraméter lehet nem egész szám is [22]. Ez az I(d) rész az idősor (közel) stacionáriussá alakítására használatos. Pontosabban a differenciálás műveletének a rendjét adja meg a d paraméter. A következőkben bemutatom a stacionáriussá alakítás módszereit, majd a stacionaritás vizsgálatára használt tesztekről is írok.

A következő módszereket használják a nemstacionárius idősorok stacionáriussá, vagy közel stacionáriussá transzformálásához: trend eltávolítása (de-trending), különbségképzés (differencing). Fontos megjegyezni, hogy ezek a módszerek nem kezelik az összes fajta nemstacionaritást.

Ha a trendet eltávolítjuk, akkor beszélhetünk trend-stacionárius idősorról [17][18]. Tehát amennyiben az idősorban trendet vélünk felfedezni, érdemes ezzel megpróbálkozni a stacionárius idősor létrehozásának reményében. Ezt a transzformációt szemlélteti a következő ábra:



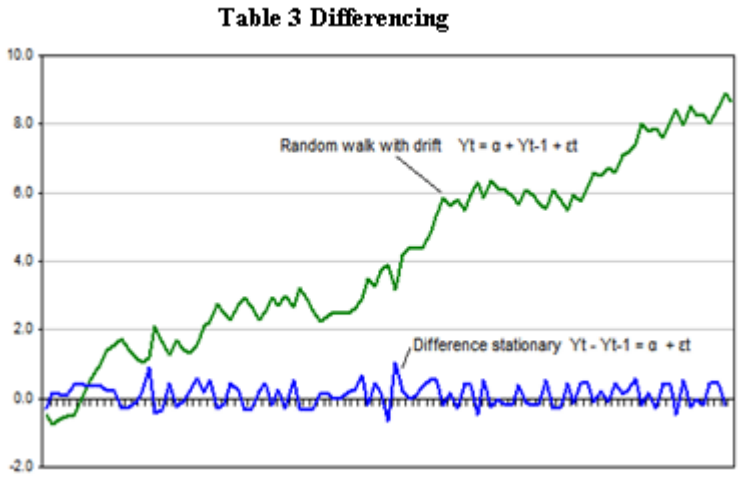
11. ábra Trend eltávolítása[18]

A képen látható, hogy az eredeti (piros) idősor egy determinisztikus trend típusú idősornak tekinthető. Ebből az idősorból a trendet felismerve, majd azt kivonva kapjuk a kék, trend-stacionárius idősort.

Az idősor stacionáriussá alakításban a másik használt módszer a különbségképzés (differencing). Az elsőrendű különbségképzés a következőnek felel meg:

$$DIFF = X_t - X_{t-1}$$

Azaz minden időponthoz tartozó értékből kivonjuk az előtte levő időpontbeli értéket. Értelemszerűen a differenciálás során egy adattal rövidebb idősort kapunk. A másodrendű különbségképzés 2 egymás utáni differenciálást jelent. Ahány különbségképzést végzünk az idősoron, annyi lesz az I(d) értéke az ARIMA(p,d,q) modellben. A következő képen egy differenciálást láthatunk:



12. ábra Differenciálás [18]

A korábban bemutatott „backward shift” operátort felhasználva bevezetésre kerül maga a különbségképző operátor [35]:

$$\nabla = 1 - B$$

Ekkor maga a különbségképzés a következőt jelenti:

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

Látható, hogy ez csak egy másik leírása magának a differenciálásnak.

Az idősorok stacionaritásának a vizsgálatára szokás használni a Dickey-Fuller, és az Augmented Dickey-Fuller tesztet. A Dickey-Fuller teszt úgynevezett „unit root”

jelenlétét vizsgálja egy autoregresszív modellben [24]. Ez a következőt jelenti: legyen egy AR(1) modellünk, ami így írható le:

$$x_t = \rho x_{t-1} + u_t$$

Ahol u_t a hibatag. Unit root akkor van jelen a modellben, ha $\rho = 1$, ez a nemstacionárius eset. Ha $|\rho| < 1$, akkor erős stacionaritásról van szó.

A unit rootnak a modell stacionaritására való hatása a következőképp mutatható be. Legyen

$$x_0 = 0$$

Ekkor unit root jelenlétét feltételezve a modell a következő:

$$x_t = x_{t-1} + u_t$$

Behelyettesítésekkel ezt kapjuk:

$$x_t = x_0 + \sum_{j=1}^{t} u_j$$

A varianciát kiszámolva látható, hogy függ t-től, azaz az időtől:

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

Például:

$$Var(x_1) = \sigma^2$$

$$Var(x_2) = 2\sigma^2$$

Tehát ez a korábban bemutatott nemstacionárius folyamatnak felel meg. A Dickey-Fuller teszt nullhipotézise az, hogy van unit root a modellben. A teszt kiértékelésének az eredménye 0, vagy 1, azaz elfogadja a nullhipotézist, vagy elutasítja. Ezzel az a probléma, hogy a teszt nem tudja megkülönböztetni unit root, és a közel unit root folyamatokat. Ezzel szemben a bővített Dickey-Fuller teszt egyfajta bizonyosságot is megad a nullhipotézis elutasításával kapcsolatban (minél nagyobb negatív szám az eredménye, annál inkább elutasítja a nullhipotézist).

3.2 Box-Jenkins módszer

Az ARIMA modell és a Box-Jenkins módszer szorosan összefügg [26]. Ez a módszer az alábbi 3 fő lépésből áll:

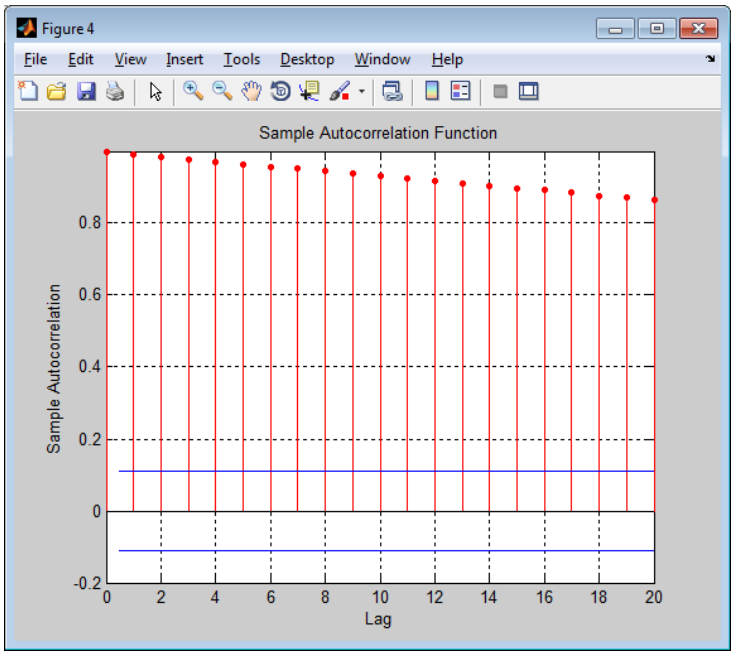
1. Modell azonosítása, választása

2. Modell paraméterek becslése

3. Modell ellenőrzése

Ebből a 3 lépésből az 1. lépést részletezem mivel az tekinthető a legproblematikusabbnak. A 2. lépésre, azaz a paraméterek becslésére például az általam használt MATLAB-ban is létezik implementált algoritmus. A 3. lépésnek, a modell ellenőrzésének pedig a legfontosabb része a modellel való előrejelzés eredményének összehasonlítása a valós adatokkal.

Az 1. lépés, azaz a modell kiválasztása röviden az ARIMA p, d, q értékeinek meghatározását takarja. Elsőként a d paraméter értékét kell meghatározni, mivel az ARMA modell feltételezi a stacionaritást [27][28], és a d paraméter ismeretében stacionáriussá transzformálhatjuk az idősort. Ennek a meghatározása a legtöbb esetben az autokorrelációs függvény értékeinek a vizsgálatával történik. Az ACF az idősor különböző időpillanatokban mért értékeinek a korrelációit adja meg [29][30]. Az ACF értékei numerikus értékek, amelyek megmutatják, hogy az időben megadott távolságokra (lag) az értékek mennyire erősen korrelálnak egymással. Például 1-es távolság azt vizsgálja, hogy az időben 1 távolságra lévő értékek mennyire korrelálnak egymással az egész idősort tekintve. Az értékkészlete a korrelációknak [-1,+1], ahol +1 erős pozitív korrelációt jelent, míg -1 az erős negatív korrelációt jelenti. Ezeket az értékeket az ACF függvény kirajzoltatásával szokták megvizsgálni, ami például a Dow Jones index egy részére így néz ki:

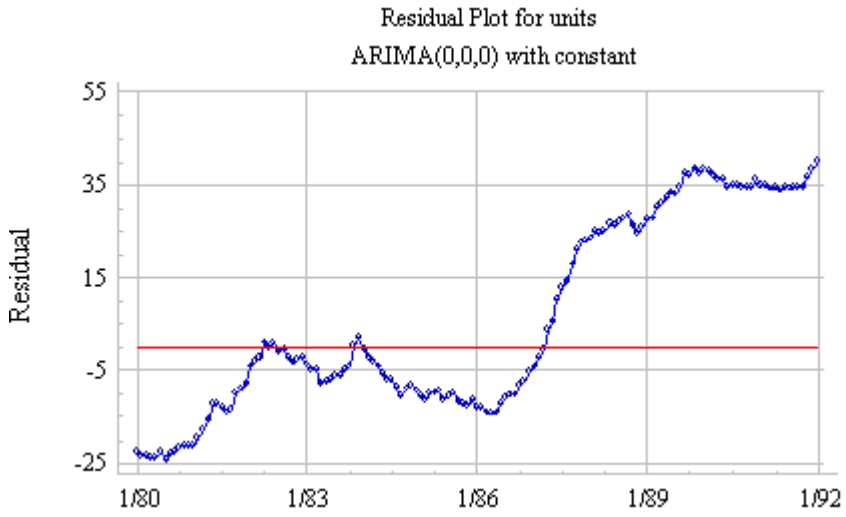


13. ábra DJIA 2008.02.-2009.06. ACF

Az ábrán a vízszintes tengelyen látható az időbeli távolság (lag) napokban mérve, a függőleges tengelyen pedig a korreláció értékek. Látható, hogy minél közelebb van 2 érték az időben, annál inkább korrelálnak egymással.

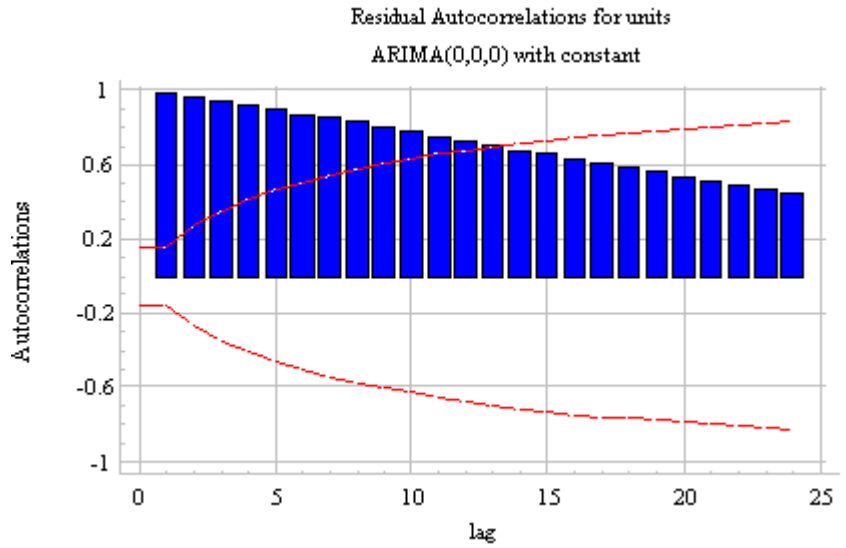
Alapvető esetben a különbségképzés rendje azon legkisebb értéknek felel meg, amelyiknél az idősor egy (jól) megadott konstans körül ingadozik, és aminek az ACF grafikonja gyorsan csökken a 0 felé. Amennyiben az idősor ACF grafikonján nagy időbeli távolságoknál is jelentősnek mondható korreláció fedezhető fel, akkor szükséges lehet további differenciálás. Sok esetben igaz az, hogy ha az ACF értékei fokozatosan csökkennek vagy nőnek, akkor még szükséges további különbségképzés [31]. Az is elmondható, hogy a differenciálás rendje gyakran akkor optimális, amikor a szórás a legkisebb. Van olyan szoftver, ami a szórás minimalizálásával becsli meg a *d* paramétert. Ezt úgy teszi, hogy különböző *d* értékű ARIMA modelleket illeszt az idősorra, de a modellekben nem használ együtthatókat csak konstanst. Ilyen együtthatók lehetnek egyébként például az autoregresszív, mozgó átlag, regressziós, szezonális, stb. együtthatók. Előfordulhat túldifferenciálás is, amire figyelmeztető jel lehet például az ACF 1-es időeltolásánál (1 hosszú időbeli távolságnál) lévő magas negatív érték (kb. -0.5). Fontos kiemelni, hogy „kőbe vésett”, minden esetben jól működő módszerek nincsenek a differenciálás rendjének meghatározása, inkább csak irányelvek léteznek. A következőben egy példán mutatom be a differenciálás és az ACF vizsgálatát.

Egy értékesítési idősorra egy ARIMA(0,0,0) modellt illesztettek az együtthatók kihagyásával, csak egy konstans taggal. Az alábbi ábrán láthatjuk az idősor átlagtól való eltérését:



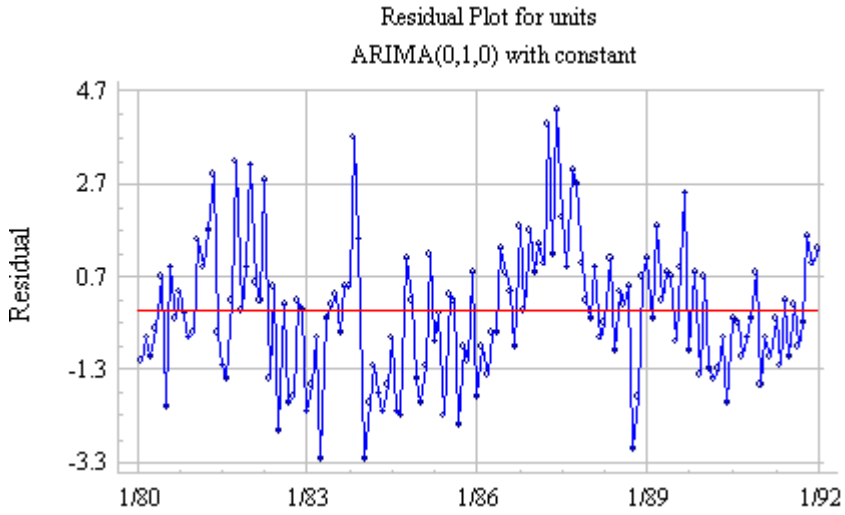
14. ábra Példa idősor I(d) vizsgálatához [28]

Ennek az idősornak az ACF kirajzolása lassú csökkenést mutat, ami a nemstacionárius idősorok jellemzője:



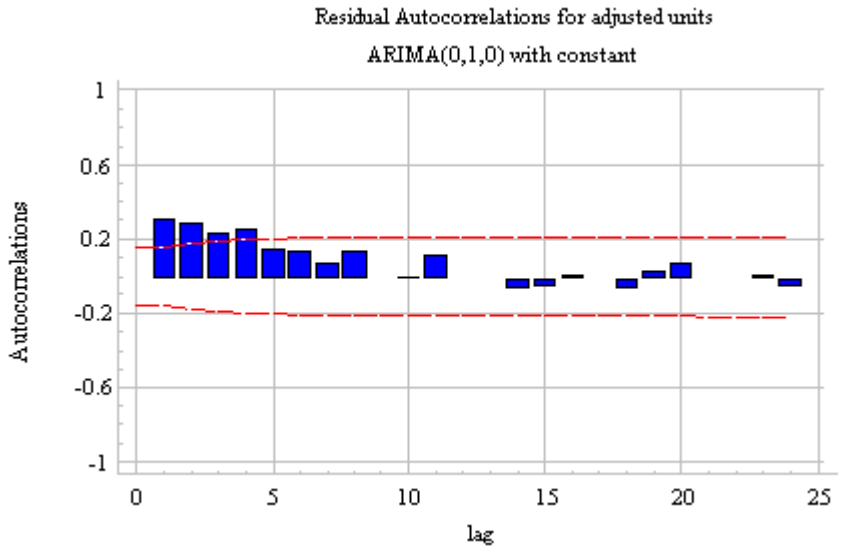
15. ábra Példa idősor I(d) vizsgálatához – ACF [28]

Egy differenciálás után az idősort kirajzolva a következőt kapjuk:



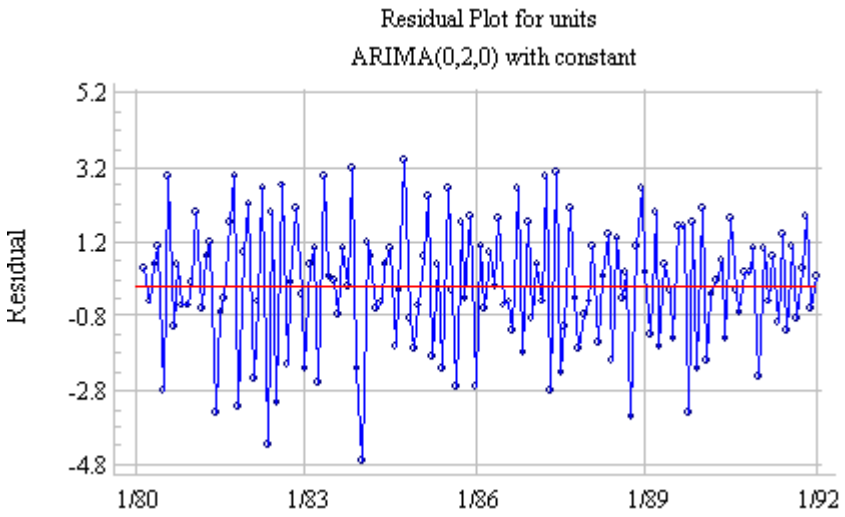
16. ábra Példa idősor I(d) vizsgálatához - I(1) [28]

Szemmel jól látható, hogy az eredeti idősor hosszú távú trendjét sikerült eltávolítani. A differenciált idősor ACF kirajzolása a következő:



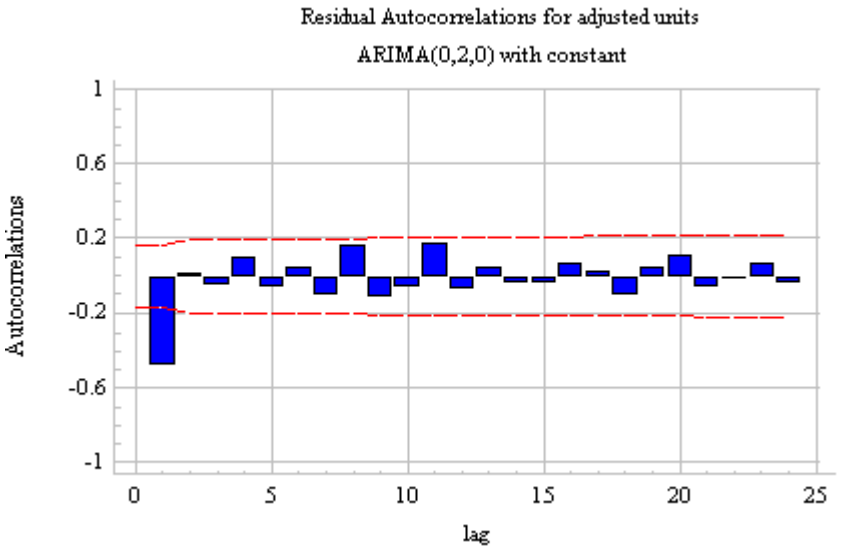
17. ábra Példa idősor I(d) vizsgálatához – ACF [28]

Látható egy enyhe pozitív korreláció. Továbbá a differenciálás elvégzésével a szórás jelentősen csökkent. Kérdés, hogy szükség van-e további differenciálásra. [28] szerint mivel a trendet eltávolítottuk és a megmaradt autokorreláció kicsinek mondható, ezért úgy tűnik, hogy nincs szükség még egy differenciálásra. Egy második differenciálás esetén a következő idősort és ACF grafikont kapjuk:



18. ábra Példa idősor I(d) vizsgálatához - I(2) [28]

Az idősoron túldifferenciálás jelei mutatkoznak meg, mivel szinte minden egymás utáni időpillanatban ellentétes előjele van az értéknek.



19. ábra Példa idősor I(d) vizsgálatához – ACF [28]

Az ACF 1. periódusánál látható magas negatív érték ezt megerősíti, ezért valószínűleg túldifferenciált idősort kaptunk.

Ezen példa alapján láthattuk néhány irányelv használatának a bemutatását a gyakorlatban is.

A I(d) meghatározása után az AR(p) és MA(q) meghatározása következik a Box-Jenkins módszer 1. lépésében. Ezen komponensek meghatározására is inkább irányelvek léteznek, mint sem konkrét, mindig jól alkalmazható módszerek. A legtöbb megközelítés az ACF és PACF függvényeket veszi alapul. A PACF az a parciális

autokorrelációs függvény (Partial ACF), amely röviden megfogalmazva abban különbözik az ACF-től, hogy csak az adott távolságban (lag) lévő értékek közti korrelációt vizsgálja, kiküszöbölve az ACF-nél fellépő korreláció továbbterjedését [32]. Pontosabb definíció itt megtalálható: [23]. Kirajzolása teljesen hasonló az ACF kirajzolásához.

Amennyiben a PACF grafikonján egy éles *levágás*, az ACF-nél pedig egy lassú csökkenés figyelhető meg, akkor ez AR jellemzőnek tekinthető (azaz az autokorreláció könnyebben magyarázható AR taggal, mint MA taggal). Amennyiben az ACF grafikonján egy éles levágás figyelhető meg, akkor ez MA jellemzőnek tekinthető. Ökölszabályként megfogalmazható, hogy éles PACF levágás esetén olyan értékű AR(p), míg éles ACF levágás esetén olyan értékű MA(q) tagot kell választani, amelyik időbeli távolságnál ez a levágás megtörtént. Fontos megjegyezni, hogy általában a legjobb modellek vagy csak AR tagot, vagy csak MA tagot használnak. Természetesen előfordulhat, hogy vegyes ARMA modell bizonyul jobbnak, de ilyen vegyes modellek esetén vigyázni kell azzal, hogy az AR és MA részek kiolthatják egymás hatásait. Ilyen esetben érdemes megpróbálkozni azzal, hogy csökkentjük 1-gyel mind az AR, mind az MA rendjét.

Két további tanács a tagok meghatározására:

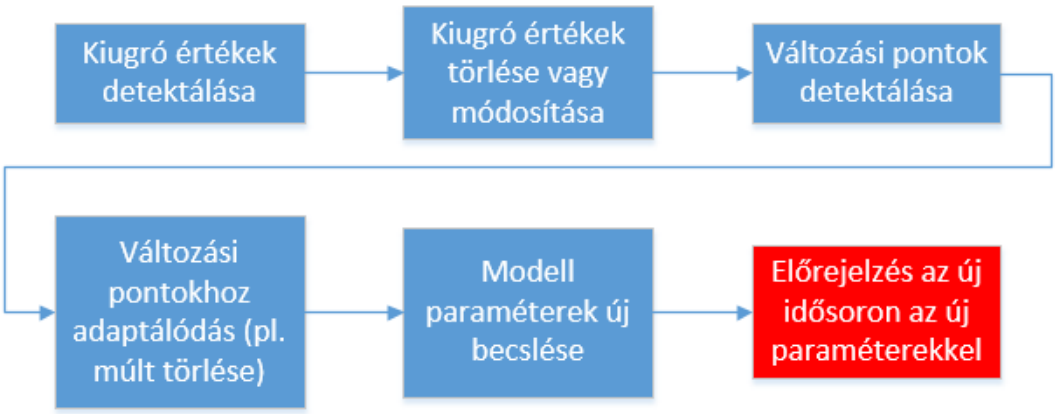
- Amennyiben egységgyököt fedezünk fel a modell AR részében (az AR koefficiensek összege közel 1), akkor érdemes 1-gyel csökkenteni az AR rendjét és növelni a különbségképzés rendjét 1-gyel.
- Amennyiben egységgyököt fedezünk fel a modell MA részében (az MA koefficiensek összege közel 1), akkor érdemes 1-gyel csökkenteni az MA rendjét és csökkenteni a különbségképzés rendjét 1-gyel.

4 Kombinált előrejelző rendszer megvalósítása és tesztelése

4.1 Kiugró értékek, változási pontok detektálása és adaptáció

Mint azt a bevezetőben már olvashattuk, az alapvető célja a kutatásomnak annak vizsgálata, hogy mennyiben javít az előrejelzésen, ha a megtalált kiugró értékek és változási pontok alapján a modell módosításra kerül, adaptálódik. Az adaptációt 2 részre érdemes bontani, a kiugró értékekkel kapcsolatos és a változási pontokkal kapcsolatos adaptációra. Egyelőre a legegyszerűbbnek nevezhető adaptációval teszteltem az előrejelzést mindkét esetben. A kiugró értékeknek minősített adatpontokat egyszerűen csak törlöm az idősorból. A változási pontoknál az utolsó megtalált változási pont előtti szakaszt törlöm az idősorból.

Az alábbi ábrán láthatjuk a kombinált előrejelző rendszer egyszerű vázlatát:

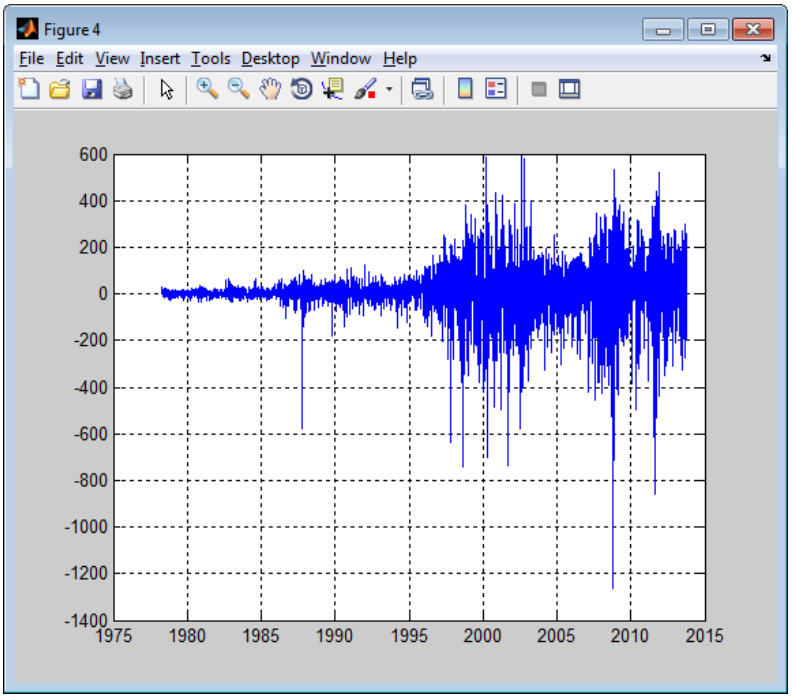


20. ábra Kombinált előrejelző rendszer vázlata

Fontos megemlíteni, hogy a későbbiekben egy fejlesztési lehetőségnek tűnik, ha a modell paraméterek új becslése előtt elvégezzük a modell teljesen új azonosítását, kidolgozását.

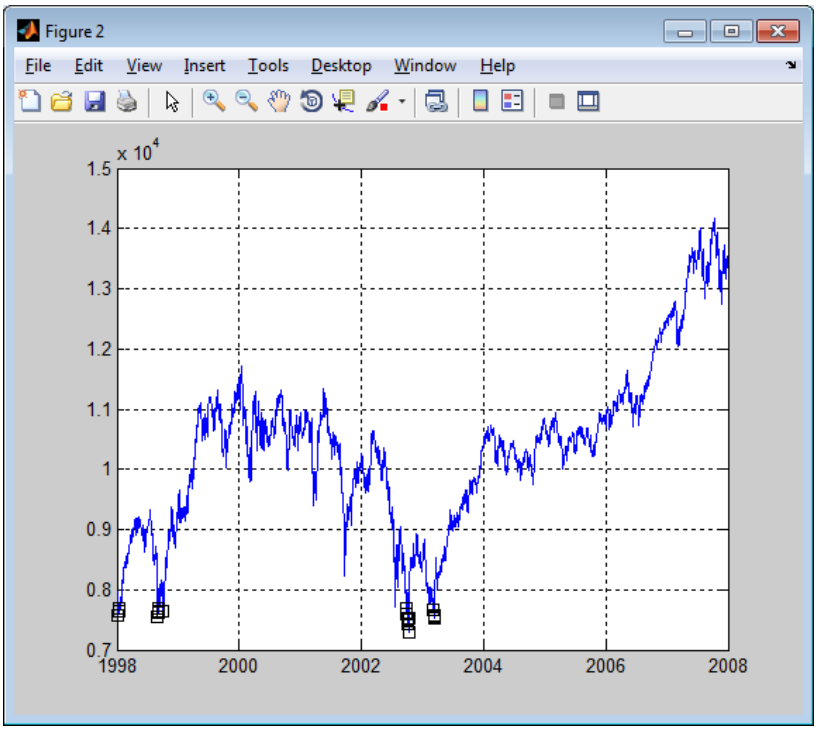
Kiugró érték detektálásával kapcsolatban először különböző hosszúságú mozgó átlagok és az eredeti idősor különbségének a módszerével próbálkoztam. Az így kapott idősorban a legnagyobb eltéréseket vizsgáltam. A Dow Jones indexre alkalmazva ezt a módszert a kapott időpontokat interneten visszakeresve elmondható, hogy leginkább a

legnagyobb napi árváltozások (nevezetes események) időpontját, illetve a legnagyobb százalékos változások időpontját kaptam eredményül. A különbség mellett vizsgáltam a relatív (százalékos) különbséget is, amely lényegében hasonló eredményeket hozott. Alább látható az eredeti Dow Jones idősor és a mozgó átlag különbsége:



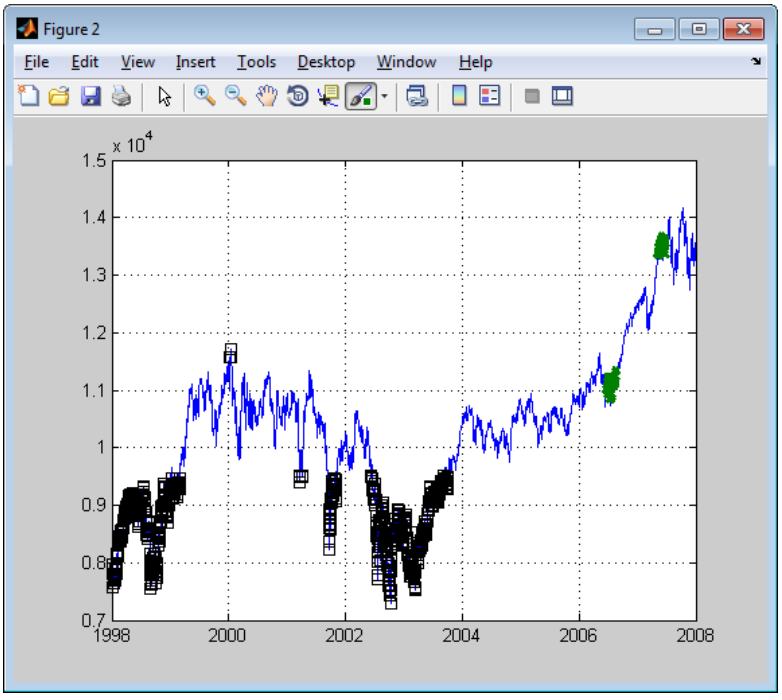
21. ábra Dow Jones index és mozgó átlag különbsége

A továbbiakban a módosított Thompson Tau módszert használtam különböző idősorokhoz. Ez a módszer publikusan elérhető a Matlab weboldaláról. Ahol a függvény paraméteréről írok, ott a módszer α értékét értem. Alább látható a Dow Jones index egy szakaszára a kapott eredmény, ahol is a módszert úgy alkalmaztam, hogy az időben legújabb megfigyeléseket nem vettem be a lehetséges kilógó értékek halmazába:



22. ábra Dow Jones index kilógó érték keresése

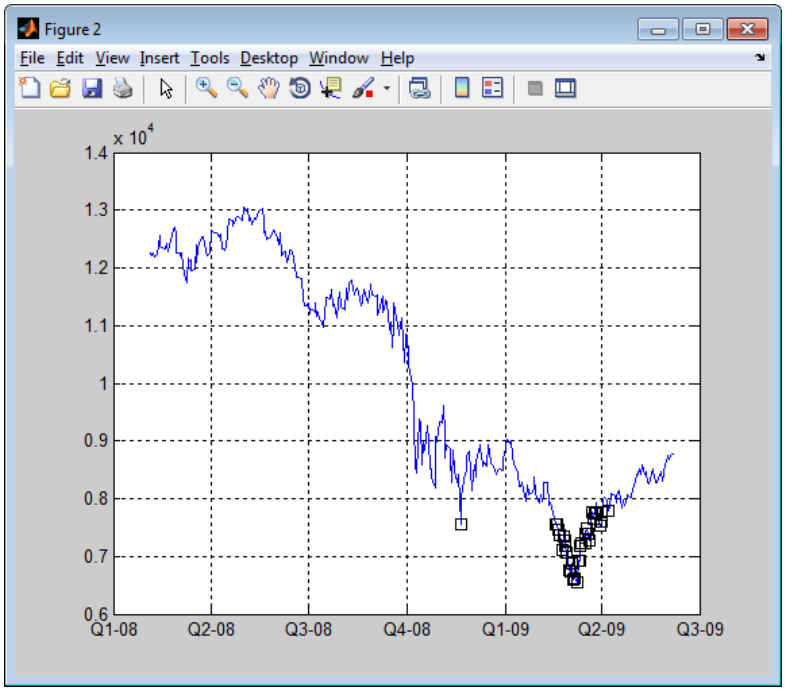
A képen a fekete négyzettel jelölt pontok jelentik a kilógó értékeket. Láthatjuk, hogy vannak olyan nem megjelölt adatpontok, amelyek kilógó értéknek minősülnének (mint például a 2002-es határ előtti negatív kicsúcsosodás). Más paraméterrel futtatva többet is megtalált az algoritmus, ami alább látható:



23. ábra Dow Jones index kilógó érték keresése 2

Ezzel kapcsolatban lehet vitatkozni, hogy mit tekinthetünk itt változási pontnak és mit kilógó értéknek. A 2002 és 2004 közötti eltérő szakaszt én 2 változási pont által körülvett szakasznak tekintem, mivel viszonylag hosszabb az a kiugrás. Továbbá az első fekete négyzettel megjelölt szakasz végén is egy változási pont található. Mindenesetre az elmondható, hogy a legtöbb jobban eltérő részt felismerte a módszer (de ez a módszer nem képes megkülönböztetni a kilógó értéket a változási ponttól). Az ábrán zölddel megjelölt részeken változási pontok találhatóak, amelyet ez a módszer nem talált meg. Akár úgy is fogalmazhatunk, hogy a két zöld jelölés közötti szakasz egy változási pont halmaznak tekinthető. Ezen a szakaszon egy kilógó érték is megfigyelhető, amit nem talált meg a módszer, mivel a legújabb megfigyeléseket nem vettem be a lehetséges kilógó értékek halmazába. Érdekes eredmény volt az, hogy az utóbbi ábrán fekete négyzettel jelölt adatpontokat törölve az idősorból rosszabb előrejelzést kaptam, mint az előtte levő ábrán megtalált kilógó értékek törlésével.

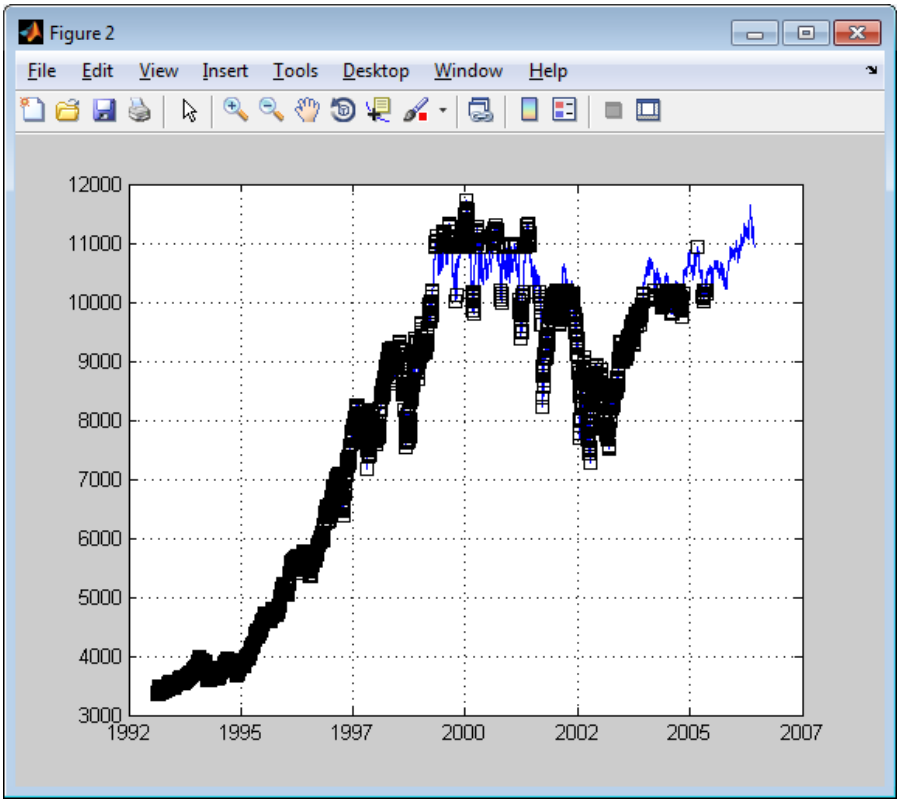
Az alábbi ábrán a dolgozatban már korábban szereplő Dow Jones index szakaszán látható a módosított Thompson Tau módszer eredménye:



24. ábra Dow Jones index kilógó érték keresése 3

Itt elmondható, hogy a korábban kilógó értéknek minősített szakaszt jól megtalálta a módszer.

A következő ábrán ugyanúgy a módosított Thomspon Tau módszer eredménye látszódik:



25. ábra Dow Jones index kilógó érték és változási pont keresés

Itt inkább a változási pont halmazait találta meg a módszer, néhány kilógó értékkel kiegészítve. Jól látszik, hogy a kéken megmaradt szakaszok előtt és között erősen más statisztikai jellemzőkkel bíró szakaszokat felismerte a módszer. Az idősor elején, 1995 körül van egy változási pont, amikor is egy trendet lehet felfedezni, ahol néhol előfordulnak kilógó értékek. Ezután az első kék szakasz kezdetén is van egy változási pont, majd a végén szintén van egy (kilógó értéknek túl hosszú lenne). A második kék szakasz elején szintén található egy változási pont. Ezek közül lényegében mindent kilógó értéknek jelölt a módszer, mivel a változási pontot nem tudja megkülönböztetni a kilógó értékektől

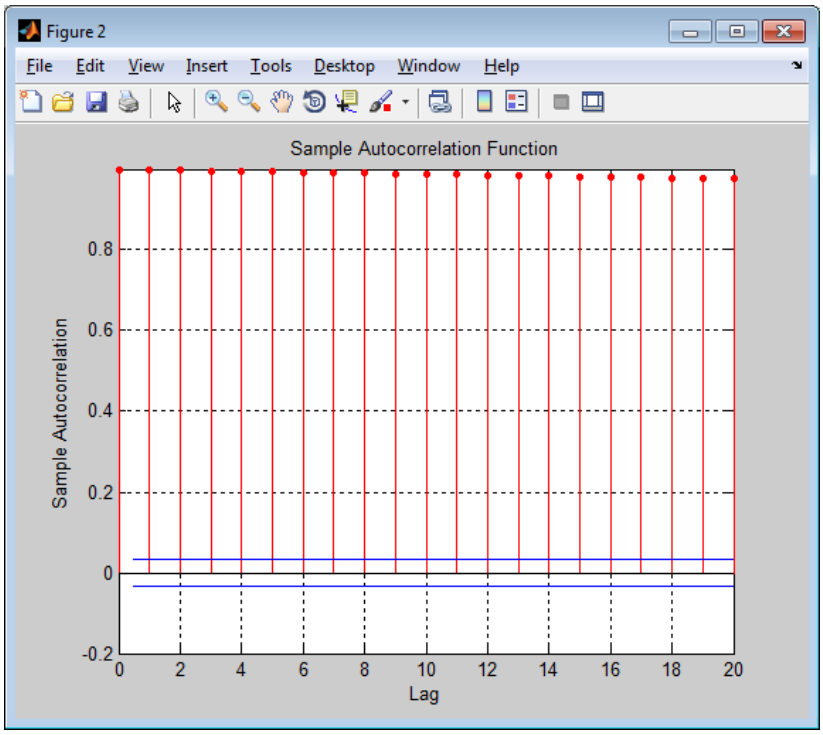
Változási pontok detektálásához a Kolmogorov-Smirnov tesztet elsőnek úgy alkalmaztam, hogy a módosított Thompson Tau módszer által megtalált időpontok által kettévágott szakaszokra vizsgáltam a nullhipotézis elfogadását. Mivel a Thompson Tau módszer nem a változási pontok keresésére alkalmas, ezért kipróbáltam az egész idősoron végighaladva a Kolmogorov-Smirnov tesztet (a trendet differenciálással eltávolítottam az idősorból). Ez a tesztelt idősoron nem adott túl jónak mondható

eredményt. Továbbá teszteltem úgy is a módszert, hogy az adott időpont által kettévágott idősor 2 szakaszát mozgó átlagoltam. Így már ígéretesebb eredményeket is kaptam, de ez a módszer még további tesztelést igényel.

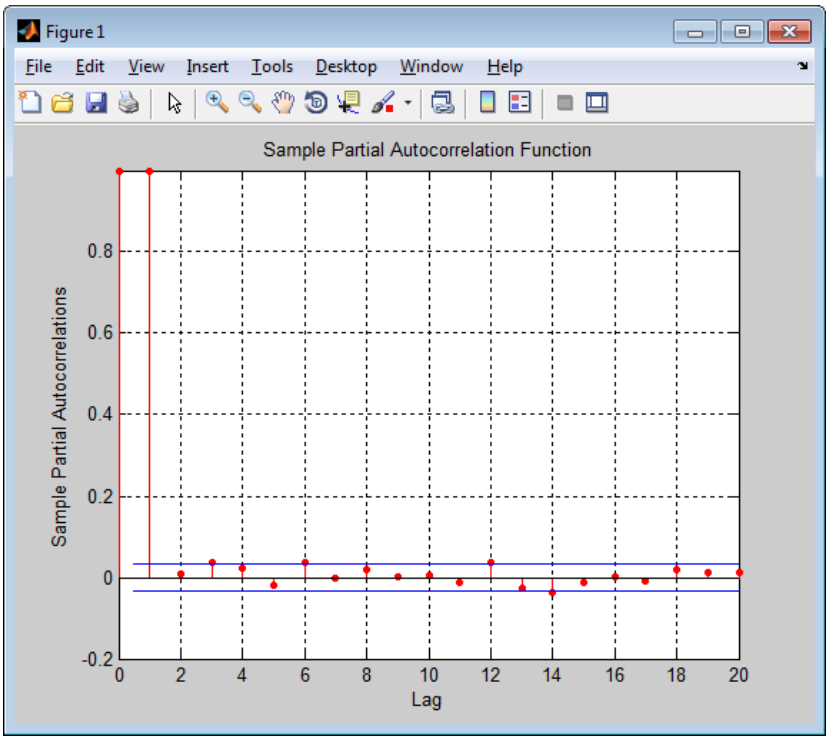
4.2 Eredmények bemutatása

A kombinált előrejelző rendszer eredményét először egy olyan (egyszerű) modellen mutatom be, amelyet a 3.2-ben leírtak szerint dolgoztam ki. A modell létrehozásáról csak röviden írok, a részletes magyarázatokra most nem térek ki. Az elsődleges cél az adaptáció utáni előrejelzés jóságának a bemutatása.

A kiválasztott idősor a Dow Jones index 1993.02.25.-2006.06.08. közötti szakasza. Különösebb megfontolás nem áll annak hátterében, hogy ezt a szakaszt választottam ki. Annyi kikötésem volna a szakasz kiválasztásakor, hogy még a 2008-as válság ideje ne kerüljön bele ebbe a vizsgálódásba. Erre megvizsgálva az ACF és PACF értékeket a következőt kaptam:



26. ábra Dow Jones index ACF



27. ábra Dow Jones index PACF

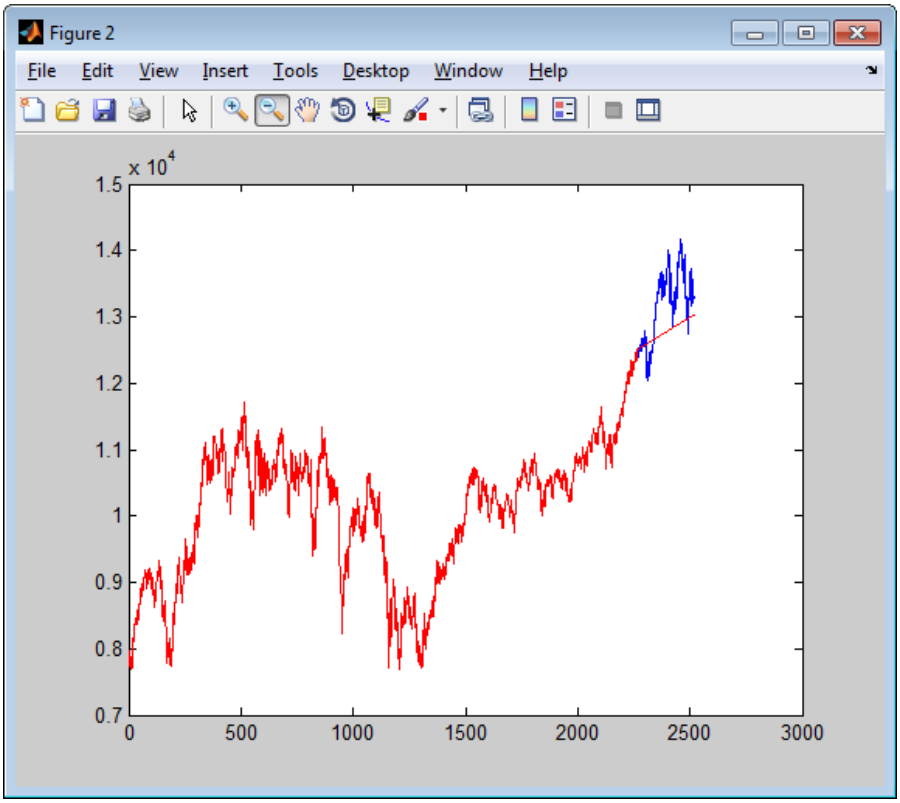
Az utóbbi ábrából látható, hogy az ACF esetén az 1 távolságra lévő korreláció terjedt tovább a többi távolságra (mivel a PACF a továbbterjedést szűri ki). Az ACF tipikus nemstacionárius viselkedést mutat, ezért differenciálás szükségesnek tűnik. A differenciálást elvégezve már olyan ACF, PACF grafikonokat kaptam, ami nem feltétlen indokolja további AR és MA tagok használatát. ARIMA(0,1,0) modellt alkalmazva 248 napot jeleztem előre. A 248 napnyi előrejelzést azért választottam, mert viszonylag hosszabb távra terveztem tesztelni a rendszert. Ekkor az RMSE értéke 212.37, a MAE értéke 161.96 lett. Miután alkalmaztam a módosított Thompson Tau módszert (25. ábra), és elvégeztem az adaptációt (töröltem a megjelölt értékeket), az előrejelzés eredménye jobb lett. Az RMSE értéke 182.5-re csökkent, a MAE érték pedig 151.86-ra csökkent.

A kombinált előrejelző rendszer képességeit több a szakirodalomban vizsgált idősorhoz és módszerhez is összehasonlítottam.

Fuchs és Sellner [34] 3 tőzsde indexet (osztrák, német, egyesült államokbeli) vizsgáltak meg és próbáltak előre jelezni. Az adatokat 1998 januárjától 2007. december végéig használták fel. A tanítóhalmaz 2006.12.31-ig tartalmaz adatokat, a teszthalmaz pedig az egész 2007-es év. A szerzők többféle modellt, módszert is kipróbáltak: Double Expnonential Smoothing, ARIMA, ARIMA-GARCH. Az ARIMA modell esetén a p, d,

q értékeket úgy határozták meg, hogy (1,1,1)-től végigiteráltak (8,1,8)-ig, és azt a modellt választották ki az adott idősorhoz, ahol a legkisebb AIC (Akaike Information Criterion) értéket kapták.

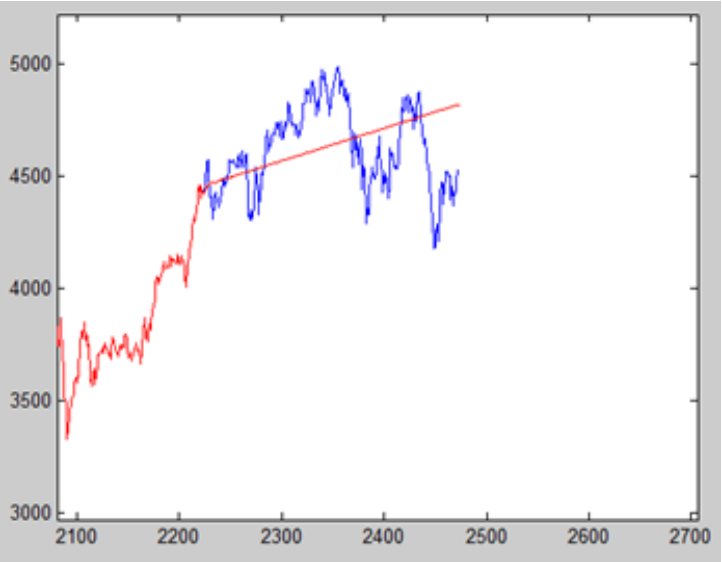
A Dow Jones indexhez ARIMA(8,1,4)-et alkalmaztak végül. Így az előrejelzésnél RMSE értéknek 550.21-et kaptak. Ugyanezt a modellt használva én 632.658-at kaptam hibának. Ennek az eltérésnek az okát az elvégzett vizsgálódás során nem sikerült megtalálni. Valószínűleg valamilyen egyéb beállítást is használtak a szerzők, amelyről a hivatkozott prezentációban nem tettek említést. Érdekesség továbbá az, hogy például az ARIMA(7,1,7)-es modellel előre jelezve 538.69-et kaptam hibának, tehát jobb lett az előrejelzés. A modell kiválasztásakor a szerzők nem a legkisebb RMSE értéket elérő modellt keresték, hanem a legkisebb AIC értéket produkálót, ezért fordulhatott ez elő. A továbbiakban az ARIMA(8,1,4)-es modellel vizsgálódtam. A módosított Thompson Tau módszert alkalmazva 0.022-es paraméterrel az adaptáció után az előrejelzés RMSE értéke 632.658-ról 600.55-re csökkent. Alább látható az adaptáció utáni előrejelzés ARIMA(8,1,4)-es modellel:



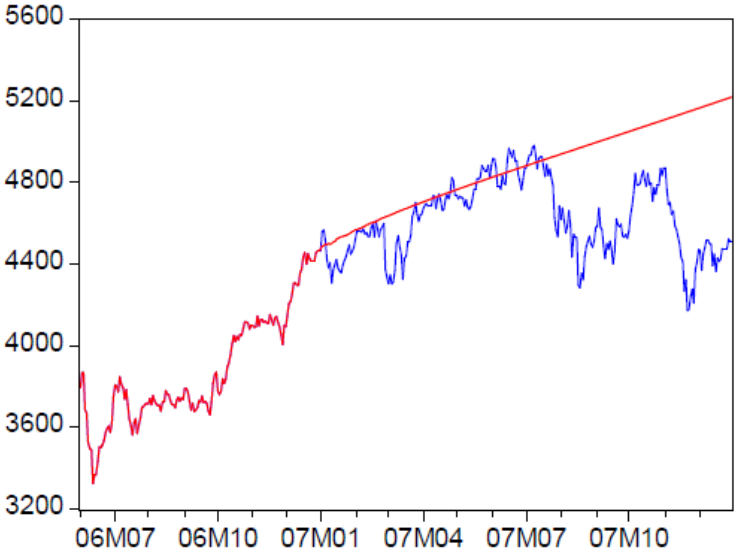
28. ábra Dow Jones index 1998-2008 előrejelzése adaptáció után

Az ábrán a piros szín jelzi az előrejelző rendszert, a kék pedig az eredeti idősort.

Az osztrák indexhez (ATX) ARIMA(7,1,7)-et alkalmaztak. Ugyanezzel a modellel szintén más RMSE-t kaptam, 212.74-et a 373.66 helyett. Másik ARIMA modellt alkalmazva, például (4,0,0)-át az RMSE értéke 201.6636, tehát jobb. A további vizsgálatokat az ARIMA(7,1,7) modellel végeztem. Ugyanezzel a modellel az előrejelzés különbségét az alábbi 2 képen láthatjuk:



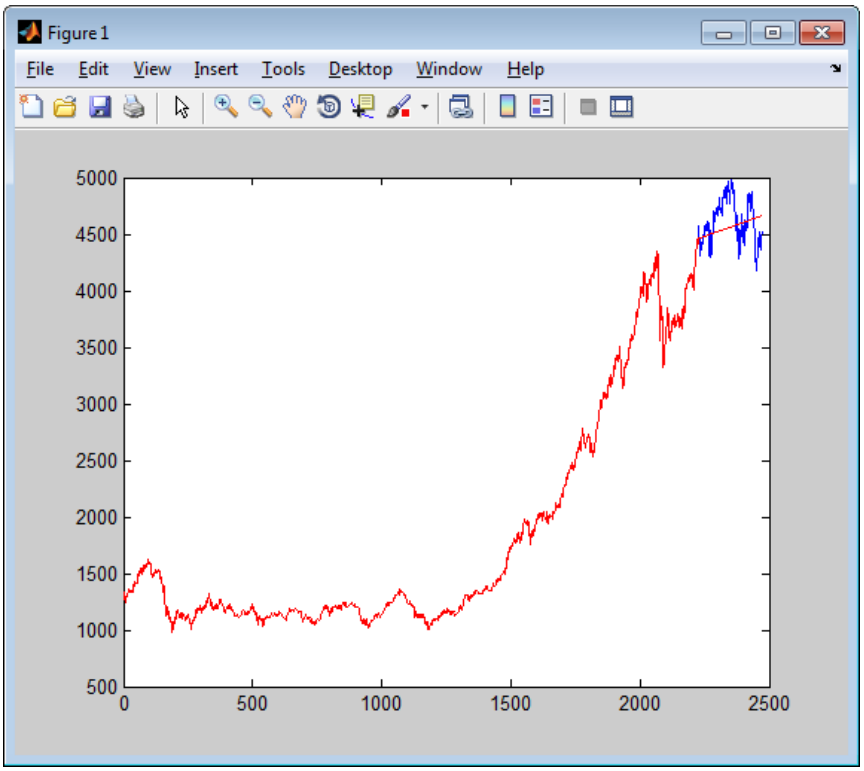
29. ábra ATX saját előrejelzés



30. ábra ATX előrejelzés [34]

Megjegyzés: A 2 képen a horizontális és vertikális nagyítás nem egyezik meg pontosan.

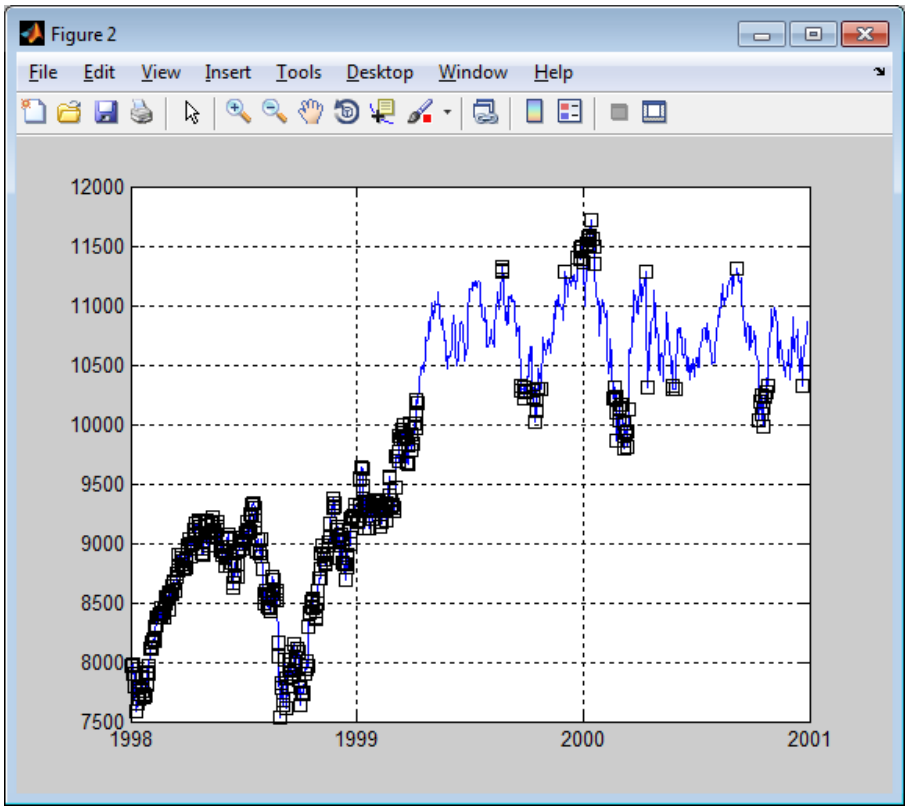
Ezen az idősoron a tesztelés során nem sikerült javítani az adaptálással. Az alábbi ábrán látható a teljes idősor:



31. ábra ATX idősor előrejelzéssel

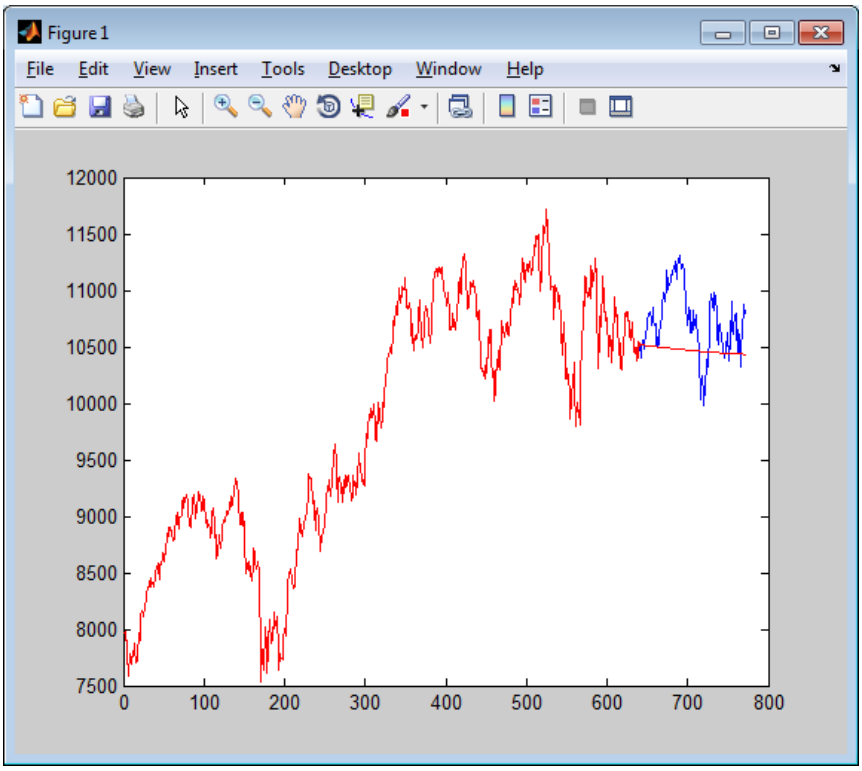
Egy jelentős változási pont figyelhető meg az *x* tengelyen 1500-as érték környékén. Ez a tény is közrejátszhat a javítás sikertelenségében, mivel az a változási pont detektáló algoritmus, amit alkalmazok, még további fejlesztésre szorul.

Yang, King, Chan és Huang [4] 2 idősort vizsgáltak meg, a Dow Jones indexet és a Hong Kong's Hang Seng indexet. A Dow Jones index esetében a modell tanításához felhasznált időszakasz 1998.01.02. - 2000.06.29-ig tart, a tesztelés pedig 2000.06.30.-2000.12.29-ig. A szerzők a kidolgozott SVM modellüket összehasonlították egy AR(4), illetve egy RBF hálózat előrejelzésével. Az előrejelzéshez itt RMSE érték helyett MAE, illetve UMAE, DMAE hibaértékeket vizsgáltak. AR(4)-es modellt alkalmazva 88.74-es MAE érték helyett 379.19-et kaptam. Ez már jelentős különbségnek mondható, de ennek a különbségnek okát ebben az esetben sem sikerült kideríteni. A módosított Thompson Tau módszerrel és az utána történő adaptálással 205.68-ra is lecsökkenthető a MAE érték (ekkor a függvény paraméterének 0.07-et adtam meg). Az alábbi ábrán látható a Thompson Tau módszerrel elvégzett kilógó értékek és változási pontok detektálása:

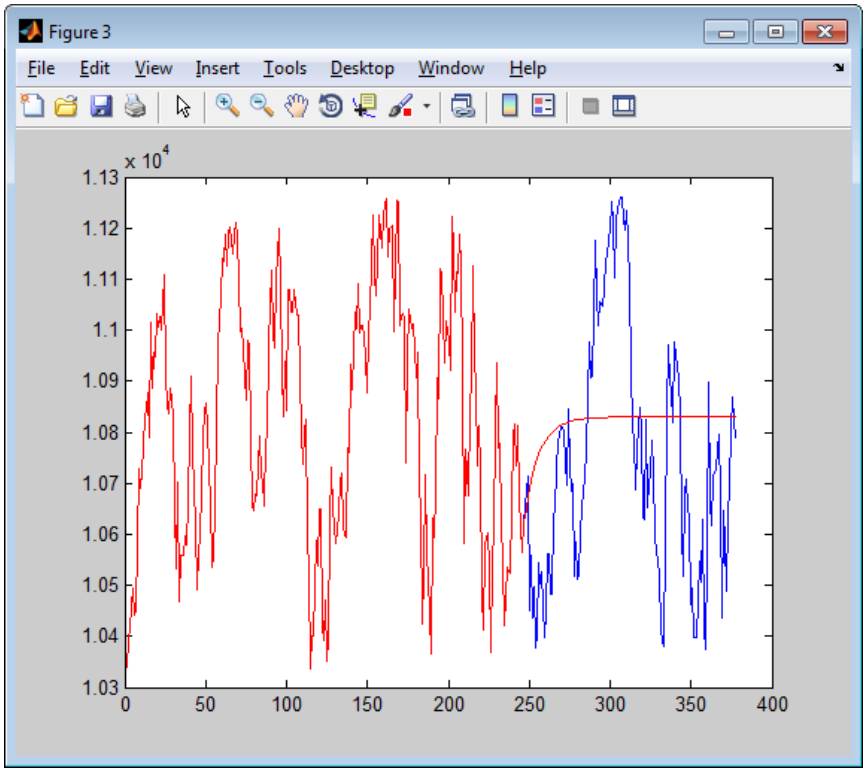


32. ábra Dow Jones index 1998-2001 detektált kilógó értékek és változási pontok

A módszer az 1999 körüli változási pontot sikeresen felismerte, illetve további kilógó értékeket is felismert. Az alábbi ábrán látható az adaptáció előtti, majd alatta az adaptáció utáni előrejelzés:



33. ábra Dow Jones index 1998-2001 adaptáció nélküli előrejelzés



34. ábra Dow Jones index 1998-2001 adaptáció utáni előrejelzés

5 Összefoglalás, jövőbeli tervek

A kutatásom során azt vizsgáltam, hogy a tőzsdei idősorok előrejelzését lehetséges-e javítani a kilógó értékek és változási pontok külön kezelésével, azaz ezen információk alapján történő adaptálással. Megismerkedtem az alap tőzsdei fogalmakkal, majd részletesebb betekintést nyertem az idősorok világába. Az irodalomkutatásom során összegyűjtöttem néhány gyakran használt modellezési módszert az idősorokkal kapcsolatban. Megismerkedtem részletesen az ARIMA modellel, és a Box-Jenkins módszerrel. Bemutattam a kilógó értékek és változási pontok jelentését, különböző értelmezéseit, majd ismertettem néhány módszert ezek detektálására. Bizonyos módszereket teszteltem valós adatokra, majd ezen eredményeket felhasználtam a kombinált előrejelzőben, ahol az adaptálás után sikerült javítani az előrejelzésen több idősor esetén is. Összességében elmondható, hogy érdemes ezen ötlet irányában további kutatásokat végezni, mivel ígéretesek az eredmények az eddigiekben alkalmazott viszonylag egyszerű módszerekkel is.

A jövőbeli tervek között a legfontosabb első lépés egy jól működő változási pont detektáló módszer kidolgozása, akár a jelenlegi továbbfejlesztésével, akár egy teljesen új módszer implementálásával. Továbbá rendkívül fontos magának az adaptációnak a fejlesztése is. Például a kilógó értékek egyszerű törlése helyett, a kilógó értékek módosítását is érdemes megvizsgálni. Másik példa az adaptáció fejlesztésére, hogy az idősor módosítása után a teljes modell újra kidolgozásra kerül. Ezek mellett nagy javítást eredményezhet egy jobb modell kiválasztása is. Mindenképpen fontos feladat további különböző adatsorokon tesztelni a módszert.

Irodalomjegyzék

[1] Ritanjali Majhi, G. Panda, G. Sahoo, P. K. Dash and D. P. Das: *Stock Market Prediction of S&P 500 and DJIA using Bacterial Foraging Optimization Technique*, 2007 IEEE Congress on Evolutionary Computation (CEC 2007)

[2] William Leigh, Cheryl J. Frohlich, Steven Hornik, Russell L. Purvis, and Tom L. Roberts: *Trading With a Stock Chart Heuristic*, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS, VOL. 38, NO. 1, JANUARY 2008

[3] Haiqin Yang, Kaizhu Huang, Laiwan Chan, Irwin King, and Michael R. Lyu: *Outliers Treatment in Support Vector Regression for Financial Time Series Prediction*

[4] Haiqin Yang, Irwin King, Laiwan Chan and Kaizhu Huang: *Financial Time Series Prediction Using Non-fixed and Asymmetrical Margin Setting with Momentum in Support Vector Regression*

[5] Grubbs, F. E. (February 1969): *Procedures for detecting outlying observations in samples*, Technometrics **11** (1): 1–21,

[6] Watson S. M., Tight M., Clark S. and Redfern E. (1991): *Detection of outliers in time series*, Institute of Transport Studies, University Of Leeds. Working Paper 362

[7] Ben-Gal I., *Outlier detection*, In: Maimon O. and Rockach L. (Eds.) *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*, Kluwer Academic Publishers, 2005, ISBN 0-387-24435-2. CHAPTER1

[8] Jussi Tolvi: *Outliers in time series: A review*

[9] Daniel Felix Ahelegbey: *Time Series Outliers* http://dfkahey.webs.com/presentations/Robust.pdf

[10] Ruey S. Tsay*: Outliers, Level Shifts, and Variance Changes in Time Series*, Journal of Forecasting, Vol. 7, 1-20 (1988)

[11] Takeuchi, J., Yamanishi (2006), K.*: A unifying framework for detecting outliers and change points from time series*, Knowledge and Data Engineering, IEEE Transactions on (Volume:18 , Issue: 4)

[12] E. Thalassinos, D.-M. Pociovalisteanu: *A time series model for the romanian stock market*

[13] Bodon Ferenc, Buza Krisztián: *Adatbányászat*

[14] Jiawei Han, Micheline Kamber: *Data Mining: Concepts and Techniques*, Second Edition

[15] Francis X. Diebold: *Elements of Forecasting*, Fourth Edition, August 2006, University of Pennsylvania

[16] Wikipedia: *Stationary process*, http://en.wikipedia.org/wiki/Stationary_process (megtekintve: 2014.10.15.)

[17] http://people.duke.edu/~rnau/411diff.htm (megtekintve: 2014.10.15.)

[18] http://www.investopedia.com/articles/trading/07/stationary.asp (megtekintve: 2014.10.15.)

[19] Wikipedia: *ARIMA*, http://en.wikipedia.org/wiki/ARIMA (megtekintve: 2014.10.15.)

[20] Wikipedia: *Autoregressive moving average*, http://en.wikipedia.org/wiki/Autoregressive_moving_average (megtekintve: 2014.10.15.)

[21] Wikipedia: *Autoregressive model*, http://en.wikipedia.org/wiki/Autoregressive_model (megtekintve: 2014.10.15.)

[22] http://www.math.elte.hu/probability/markus/AlkmatTS1/idosorok_1_3.pdf (megtekintve: 2014.10.15.)

[23] http://www.math.elte.hu/probability/markus/AlkmatTS1/idosorok_1_1.pdf (megtekintve: 2014.10.15.)

[24] Wikipedia: *Dickey-Fuller test*, http://en.wikipedia.org/wiki/Dickey-Fuller_test (megtekintve: 2014.10.15.)

[25] Wikipedia: *Augmented Dickey-Fuller test*, http://en.wikipedia.org/wiki/Augmented_Dickey-Fuller_test (megtekintve: 2014.10.15.)

[26] Wikipedia: *Box-Jenkins*, http://en.wikipedia.org/wiki/Box-Jenkins (megtekintve: 2014.10.15.)

[27] Terence C. Mills and Raphael N. Markellos: *The Econometric Modelling of Financial Time Series*, Third Edition, 2008, Cambridge University Press, ISBN-13 978-0-511-38103-4

[28] http://people.duke.edu/~rnau/411arim2.htm (megtekintve: 2014.10.15.)

[29] Wikipedia: *Autocorrelation*, http://en.wikipedia.org/wiki/Autocorrelation (megtekintve: 2014.10.15.)

[30] http://www.forecastingsolutions.com/arima.html (megtekintve: 2014.10.15.)

[31] http://sfb649.wiwi.hu-berlin.de/fedc_homepage/xplore/tutorials/xegbohtmlnode39.html (megtekintve: 2014.10.15.)

[32] http://people.duke.edu/~rnau/411arim3.htm (megtekintve: 2014.10.15.)

[33] John M. Cimbala: *Outliers*, 2011

[34] Regina Fuchs and Richard Sellner: *Forecasting Stock Market Time Series*

[35] Közgazdasági és pénzügyi idősorok BME tantárgy VISZM021, Telcs András által tartott előadás jegyzete