Tőzsdei idősorok előrejelzése adatbányászati módszerekkel*[1]

Badics Milán Csaba

Az elmúlt 20-25 évben rengeteg adatbányászati módszert kezdtek el alkalmazni tőzsdei idősorok előrejelzésére, egyre újabb és szofisztikáltabb modellek jelentek meg a szakirodalomban és a piaci alkalamazások során is. Kutatásomban ezért azt vizsgálom, hogy a különböző adatbányászati modellek mennyire használhatóak az aktív portfóliókezelésben, külön kitérve a zajszűrő és hibrid módszerek alkalmazhatóságára. Célom az volt, hogy olyan árfolyam-előrejelzésen alapuló kereskedési stratégiát mutassak be, amely tranzakciós költségek mellett is eredményes lehet. A különböző adatbányászati módszerek előrejelző képességét az OTP záróárfolyamának idősorán teszteltem.

Journal of Economic Literature (JEL) kódok: C45, G14, G17

Kulcsszavak: tőzsdei idősorok előrejelzése, kereskedési stratégia, neurális háló, független komponenselemzés, empirikus dekompozíció, adatbányászati modellek

Bevezetés

A tőzsdei idősorok alakulása már évtizedek óta a befektetők figyelmének középpontjában áll, és próbálják különféle módszerekkel előrejelezni azt. A nagy érdeklődésre való tekintettel akadémiai körökben is egyre több kutatás kezdett el foglalkozni az idősorok előrejelzésének lehetőségeivel. Először a közgazdaságtanban használatos, hagyományos statisztikai/ökonometriai modelleket kezdték alkalmazni, azonban az idősorok speciális jellegzetességei miatt – mint például a nemlinearitás, a nemstacioner tulajdonság, a magas zaj/jel arány – ezek kevésbé bizonyultak eredményesnek. Ekkor fordultak a műszaki életben gyakran alkalmazott nemparaméteres, kevesebb statisztikai megkötéssel rendelkező adatbányászati/gépi tanulási módszerek felé, és ezek eszköztára új perspektívát nyitott a pénzügyi idősorok hatékonyabb előrejelzése előtt.

* Jelen cikk a szerző nézeteit tartalmazza, és nem feltétlenül tükrözi a Magyar Nemzeti Bank hivatalos álláspontját. Badics Milán, a Magyar Nemzeti Bank PADS PHD ösztöndíjprogramjának hallgatója.

1 Ezúton szeretnék köszönetet mondani *Ferenczi Tamásnak,* aki segített a releváns szakirodalom feldolgozásában és megértésében. Szintén köszönettel tartozom *Hans Zoltánnak, Szoboszlai Mihálynak* és *Márkus Balázsnak,* akik hasznos tanácsokkal láttak el a kutatási eredményeim megfogalmazása során. A tanulmány eredeti, egy hosszabb változata a Budapesti Értéktőzsde által szervezett X. Kochmeister-díj pályázatán első díjat ért el 2014 májusában. A tanulmányban említett további releváns problémákat a Magyar Nemzeti Bank PADS PHD ösztöndíjprogramjának hallgatójaként kutatom tovább.

Az elmúlt 30 évben az adatbányászati módszerek egyre több változatát kezdték el alkalmazni tőzsdei adatok alakulásának vizsgálatára. Először az egyik legnépszerűbb, a neurális háló különböző fajtáit alkalmazták a statisztikai módszerekhez képest nagy haszonnal. Mivel az előrejelzési pontosság kicsi javítása is akár hatalmas többletprofitot eredményezhet, ezért mind a befektetői, mind az akadémiai körökben egyre népszerűbbé vált a különböző hálózatok közül a legoptimálisabb megkeresése, annak megfelelő parametrizálása. Idővel azonban a nagy siker és a széles körű alkalmazás miatt – mint minden előrejelzésre épített stratégia, ha sokan kezdik el használni egyszerre – az erre épített befektetési döntések átlagon felüli profitszerzési lehetősége is egyre csökkent.

Ugyanakkor ez nem jelentette azt, hogy a befektetési döntéshozók, illetve a kutatók ezután elfordultak volna ezektől a módszerektől. Éppen ellenkezőleg, egyre több energiát fektettek a műszaki élet egyéb területein már sikerrel használt módszerek idősor-előrejelzésre való átültetésébe. Többek között kipróbálták a többi adatbányászati módszer módosított változatait (SVR, Random Forest), illetve a zajszűrő (ICA, PCA) és dekompozíció alapú (EMD, wavelet) technikákat is. Emellett elterjedt a többlépcsős hibrid módszerek alkalmazása és az egyes előrejelzések kombinálása is. Mára már rengeteg módszert és modellt fejlesztettek ki, és a tőzsdei idősor-előrejelzésen alapuló stratégia a legnépszerűbbek közé tartozik. Ezek alkalmazása ugyanakkor komoly kihívás is, mivel a hatékony előrejelzéshez szükséges a különböző modellek előnyeinek és hátrányainak ismerete.

Kutatásomban ezért bemutatom a legismertebb, aktív portfóliókezelésre alkalmas adatbányászati módszereket, azok előnyeit és hátrányait, melyiket mikor és milyen formában érdemes alkalmazni; illetve kitérek arra is, hogy melyek a jelentősebb kutatási irányok napjainkban. Célom az volt, hogy a teljes folyamatot bemutassam az előre jelezni kívánt részvényárfolyam kiválasztásától (cikkemben az OTP napi záróárfolyamai) a szükséges inputváltozók és a használható adatbányászati módszerek definiálásán át egészen a kereskedés megvalósításáig, mintegy útikönyvet adva ezzel az olvasó kezébe az előrejelzésen alapuló aktív portfóliókezeléshez.

1. A tőzsdei idősorok jellemzői, előrejelzési módszerei, nehézségei

A pénzügyi idősorok előrejelzését nagyban nehezíti, hogy ezek általában zajosak, nemstacionáriusak, nemlineárisak és kaotikusak, továbbá gyakran fordul elő bennük strukturális törés is (*Hall*, 1994; *Li* et al., 2003; *Yaser és Atiya*, 1996; *Huang* et al., 2010; *Lu* et al., 2009, *Oh* és *Kim*, 2002; *Wang*, 2003). Ezen okok miatt a pénzügyi/tőzsdei idősorok előrejelzése az egyik legnagyobb kihívás a piaci szereplők számára.

Az előrejelző módszereket, amelyeket a tanulmányokban ismertetnek, két kategóriába lehet osztani: statisztikai/ökonometriai és adatbányászati/gépi tanulási módszerek. A tradicionális statisztikai módszerek közé tartozik a lineáris regresszió, a mozgóátlagolás, az exponenciális simítás, az ARIMA, a GARCH és a VAR. Ezek a módszerek akkor adnak jó előrejelzési eredményeket, ha a pénzügyi idősorok lineárisak vagy közel lineárisak, a való életben azonban nem ez a jellemző. Emellett a hagyományos statisztikai módszerek nagy mennyiségű historikus adatot követelnek, és a jó előrejelzési eredményhez emellett megkövetelik azt is, hogy ezek eloszlása normális legyen (*Cheng* és *Wei*, 2014).

Ezeket a feltételezéseket küszöbölik ki az adatbányászati módszerek, amelyek jobban képesek modellezni az idősorok nemlineáris struktúráját. Idesoroljuk a neurális hálók mellett a tartóvektorgépeket (Support Vector Machines – SVM) és a döntési fák különböző fajtáit is. Ezek az adatvezérelt és nemparametrikus módszerek ismeretlen kapcsolatokat is képesek feltárni és kezelni az empirikus adatok között, így hatékonyabban előre jelezhetik a bonyolult és nemlineáris tőzsdei adatok változását (*Chen* et al., 2003; *Chun* és *Kim,* 2004; *Thawornwong* és *Enke*, 2004; *Enke és Thawornwong,* 2005; *Hansen* és *Nelson*, 2002). Az elmúlt években megjelenő, egyre több adatbányászati cikk és alkalmazás is azt mutatja, hogy ezek az alkalmazások versenyképesek, és jelentős előnyöket mutatnak fel a hagyományos módszerekhez képest (*Lu* et al., 2009; *Duan* és *Stanley*, 2011; *Huang* et al. 2010; *Ni* és *Yin*, 2009).

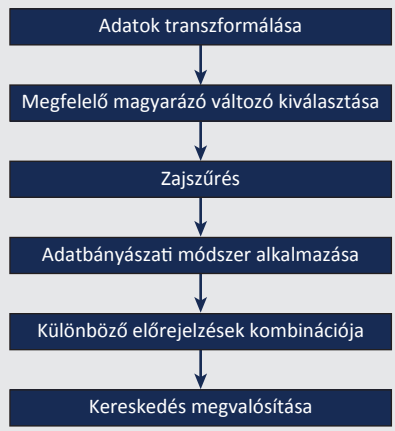
Mivel azonban minden adatbányászati módszernek vannak hátrányai, emiatt akadémiai körökben egyre népszerűbb az egyes adatbányászati technikák keresztezése (hibridálása). Az alapötlet az, hogy a hibrid módszerek kiküszöbölik az egyedi módszerek hátrányait, és szinergiát alkotva, javítják az előrejelzések pontosságát. A módszernek alapvetően három különböző fajtája van. Az első a „divide and conquer" (oszd meg és uralkodj) elven alapszik, aminek az a lényege, hogy komplex problémák esetén érdemes lehet több kisebb problémára felosztani a kérdést, majd külön-külön megoldani azokat. Ennek egyik legelterjedtebb alkalmazása a tőzsdei előrejelzésben az empirikus dekompozíció (Empirical Mode Decomposition – EMD; *Cheng és Wei,* 2014). A második esetben megpróbáljuk kiszűrni a modellek input változóiból a zajt, így elősegítve, hogy pontosabb eredményt kapjunk. Erre leggyakrabban a függetlenkomponens-analízist (Independent Component Analysis ICA) használják, amely azon elven alapu, hogy az input változókból független komponenseket létrehozva (IC-k) megállapítható, hogy melyik komponens tartalmazza a zajt, és azt eltávolítva, növelni tudjuk az előrejelzés pontosságát (*Lu* et al., 2009). A harmadik módszer pedig a különböző adatbányászati modellek előrejelzéseinek kombinálása az egyszerű aggregálástól kezdve a bayesi átlagoláson át a Lasso-regresszióig. A kombinálási módszerek azon alapulnak, hogy az egyes módszerek együttes figyelembevételével az előrejelzés varianciája csökkenthető (*Sermpinis* et al., 2012).

Egy tőzsdei előrejelzésen alapuló, aktív portfóliókezelő stratégia megalkotása során tehát a következő kihívásokkal, nehézségekkel kell szembenézni:

1. megfelelő magyarázó változók kiválasztása (*feature selection*),
2. zajszűrés, jelfeldolgozás (*financial signal processing*),
3. valamilyen adatbányászati módszer alapján előrejelzés a paraméterek optimalizálása mellett (*forecasting with data mining methods*),
4. különböző előrejelzések kombinálása (*combining data mining techniques*).

A teljes folyamatot, beleértve az adatok előkészítését, transzformálását és a kereskedést az *1. ábra* mutatja.

1. ábra
A kutatásban bemutatott modellkeret felépítése



Forrás: saját szerkesztés

2. Az adatbányászati modell felépítésének folyamata

2.1. Zajszűrés és hibrid módszerek

Ahhoz, hogy pontos előrejelzést tudjunk készíteni, szükséges, hogy a részvények árfolyammozgása mögötti látens változókat megtaláljuk, és felhasználjuk a modellezés során. Az ilyen problémák megoldására a mérnöki gyakorlatban már elterjedt módszer, a függetlenkomponens-elemzés alkalmazható. Ez az eljárás képes arra, hogy feltárja az adatsorok változását befolyásoló, rejtett komponenseket, és ezeket különválassza egymástól, méghozzá úgy, hogy azok a lehető legkevésbé függjenek egymástól, és lineáris kombinációjukból felírhatóak legyenek az eredeti adatsorok (*Kapelner* és *Madarász*, 2012).

A függetlenkomponens-elemzéssel lehetőség nyílik arra, hogy megtaláljuk és eltávolítsuk a zajkomponenst a modellezéshez használt adatokból, így javítva az előrejelzés pontosságát (*Lu,* 2010). A módszert gyakran használják a műszaki életben *jelfeldolgozásra* (*Beckmann* és *Smith*, 2004), *arcfelismerő rendszereknél zajszűrésre* (*Déniz* et al., 2003) és természetesen *tőzsdei idősorok előrejelzésére* is. *Oja* et al. (2000) függetlenkomponens-elemzést használtak, hogy csökkentsék a modell input adatainak zaj/jel arányát, majd autoregresszív modellel jelezték előre a devizaárfolyamokat.

Egy kicsit más szempontból közelíti meg a zajszűrést az EMD dekompozíciós eljárás, amely nem az input változókból próbálja kiszűrni a zajt, hanem magából az eredeti idősorból. Az empirikus dekompozíció lényege a korábban említett „divide and conquer" elv. Az eljárást *Huang* et al. (1998) fejlesztette ki, és a Hilbert–Huang-transzformáción alapszik. Ez az eredeti idősort véges számú IMF-ekre bontja fel, amelyek könnyebben kezelhetők és erősen korreláltak, így könnyebb egyesével előre jelezni őket, majd ezeket aggregálva, megkapni az eredeti idősor előrejelzését (*Cheng és Wei,* 2014). Ezt a módszert gyakran használják *földrengésjelek dekompozíciójára* (*Vincent* et al., 1999), *szélsebesség* (*Guo* et al., 2012) és akár *turizmus előrejelzésére* is (*Chen* et al., 2012). Kutatásomban emiatt az ICA mellett ezt a módszert kombináltam egy adatbányászati modellel.

2.2. Lehetséges adatbányászati módszerek

Az adatbányászati módszerek közül a pénzügyi idősorok előrejelzésére a legelterjedtebbnek és legnépszerűbbnek a különböző neurális hálózatok számítanak (*Cao* és *Parry*, 2009; *Chang* et al., 2009; *Chavarnakul és Enke,* 2008; *Enke ésThawornwong,* 2005). Ezek az adatvezérelt, nemparametrikus módszerek nem követelnek erős modellfeltevéseket, sem előzetes statisztikai feltételezéseket az input adatokról, továbbá bármilyen nemlineáris függvényt képesek modellezni (*Vellido* et al., 1999; *Zhang* et al., 1998). *Atsalakis* és *Valavanis* (2009) közel száz tanulmányt feldolgozó cikkében rámutat, hogy a különböző neurális hálózatok közül az előrecsatolt (feed forward neural network – FFNN) és a rekurrens (recurrent neural networks – RNN) hálókat alkalmazzák a leggyakrabban a kutatók a pénzügyi idősorok előrejelzésére. Előrecsatolt neurális hálók közül a hiba-visszaterjesztéses (back-propagation neural network – BPN), míg rekurrens hálók közül az Elman- és a Jordan-hálók a legnépszerűbbek.

További megoldás lehet pénzügyi idősorok előrejelzésére tartóvektorgépek, döntési fák, genetikus algoritmusok használta is. A módszerek nagy száma miatt nagyon időigényes lehet megtalálni, hogy egyes idősorok esetén melyik a leghatékonyabb megoldás; illetve, ahogy láttuk, mindegyiknek van előnye és hátránya is, ezért gyakran használnak többet a modellezés során, majd kombinálják ezek eredményeit. Mivel egy rejtett réteggel ren-

delkező, előrecsatolt neurális háló bármilyen komplex problémát tud modellezni (*Chauvin és Rumelhart*, 1995), ezért én is ezt használtam a kutatásomban.

2.3. Adatbányászati módszerek kombinálása

Az idősor-előrejelzés irodalmának egyik legérdekesebb kérdése az, hogyan kombináljunk különböző előrejelzési technikákat. Több kutató is rámutatott, hogy a különböző technikákat – főleg rövid távú előrejelzés esetén – érdemes kombinálni, ami azért előnyös, mert kiküszöböli az egyes módszerek hiányosságait (*Zhang és Wu*, 2009; *Armstrong*, 1989). Habár *Timmermann* (2006) tanulmányában rámutatott, hogy egy egyszerű átlagolás is felveheti a versenyt a szofisztikáltabb technikákkal, azonban vannak olyan esetek, amikor az egyik módszer jóval pontosabb, mint a többi, így az átlagolás nem elég hatékony. *Granger és Ramathan* (1984) a regressziós technikát ajánlotta biztató eredményekkel, míg *Swanson és Zeng* (2001) a bayesi átlagolást. Szinte minden szerző azt az állítást fogalmazta meg, hogy a különböző előrejelzési módszerek kombinálása szükséges; arról azonban nem született egyezség, hogy mikor melyiket érdemes használni, így elemzésemben többet is alkalmaztam.

3. A tanulmányban alkalmazott módszerek bemutatása

3.1. Függetlenkomponens-elemzés

Ha az adatbányászati modelleket úgy tanítjuk, hogy nem vesszük figyelembe azok lehetséges zajtartalmát, akkor az ronthatja az általánosítás képességét a teszthalmazon, illetve túltanuláshoz vezethet. Az input adatok zajszűrése ezért kiemelt feladat a modellezés során, amit én függetlenkomponens-elemzéssel fogok megoldani. Most ennek elméleti hátterét mutatom be.

Legyen $X=[x_1, x_2, \ldots, x_m]^T$ egy többdimenziós adatmátrix mxn-es mérettel, ahol $m \leq n$ és a megfigyelt kevert jelek $\mathbf{x}_i$ mérete 1xn i = 1, 2, ..., m. Az ICA-modell alkalmazása esetén ez az X mátrix felírható a következő alakban:

$$X=AS=\sum_{i=1}^{m} a_i s_i ,$$

ahol a_i az i-edik oszlopa az mxm méretű ismeretlen **A** keverőmátrixnak (mixing matrix) és $\mathbf{s}_i$ az i-edik sora az mxn méretű,m „source" S mátrixnak. Az s_i vektorok azok a látens adatok, amelyeket nem tudunk közvetlenül megfigyelni a kevert x_i adatokból, de utóbbiak ezen látens adatok lineáris kombinációjaként írhatóak fel (*Dai* et al., 2012). A függetlenkomponens-elemzés célja, hogy megtaláljuk azt az mxm méretű W mátrixot, (demixing matrix), amelyre teljesül, hogy

$$Y=WX,$$

ahol y_i az i-edik sora az Y mátrixnak, i = 1, 2, …, m, és ezek a vektorok statisztikailag függetlenek (független komponensek). Ha a W mátrix az A keverőmátrix inverze, $W=A^{-1}$, akkor a független komponenseket (y_i) tudjuk használni, hogy megbecsüljük az eredeti látens jeleket (s_i) (*Lu,* 2010).

Függetlenkomponens-elemzés során egy optimalizációs problémát oldunk meg úgy, hogy megválasztjuk a független komponensek statisztikai függetlenségének egy objektív függvényét, és optimalizációs eljárásokkal megkeressük a W mátrixot (*Lu* et al., 2009). Több ilyen kifejlesztett/kidolgozott eljárás létezik (*Bell és Sejnowski*, 1995; *David* és *Sanchez*, 2002; *Hyvärinen* et al., 2001), amelyek általában nem felügyelt tanítási algoritmusokat használnak, hogy maximalizálják az IC-k statisztikai függetlenségét. Az ICA egyik leggyakoribb megoldási módja a FastICA algoritmus (*Hyvärinen* et al., 2001), amelyet én is alkalmaztam a W mátrix definiálására.

3.2. Empirikus alapú dekompozíció (EMD)

Az empirikus alapú dekompozíció egy nemlineáris jeltranszformációs eljárás, amit *Huang* et al. (1998) fejlesztett ki nemlineáris és nemstacioner idősorok dekompozíciójára. Ez a módszer az eredeti idősort különböző időskálájú, oszcilláló IMF (Intrinsic Mode Function) komponensekre bontja fel (*Yu* et al., 2008). Minden egyes IMF-nek két feltételt kell kielégítenie: egyrészt a lokális minimumok és maximumok össz-számának és a függvény nullhelyei számának különbsége maximum egy lehet, másrészt a lokális átlagnak nullának kell lennie (*Cheng és Wei,* 2014). Ez az algoritmus a következő:

1. Határozzuk meg az összes lokális minimumát és maximumát x(t)-nek.
2. Határozzuk meg az alsó $x_u(t)$ és felső $x_l(t)$burkolóját x(t)-nek.
3. A felső és az alsó burkolót használva, adjuk meg az idősor átlagát: $m_1(t)=[x_u(t)+x_l(t)]/2$.
4. Számoljuk ki az eredeti idősor, x(t) és az előző lépésben kapott átlag, $m_1(t)$ idősor különbségét: $h_1(t)=x(t)–m_1(t)$, ami az első IMF-et – $h_1(t)$ – adja meg, ha kielégíti a fent említett két feltételt.

5. Miután megkaptuk az első IMF-et, ugyanezt az iterációs algoritmust folytatjuk addig, amíg meg nem kapjuk a végső idősort, a reziduális komponenst – r(t) –, ami egy monoton függvény, és azt jelzi, hogy le kell állítanunk az algoritmust (Huang et al., 1998).

Az eredeti idősort x(t) visszakaphatjuk az IMF komponensek és a reziduális összegeként:

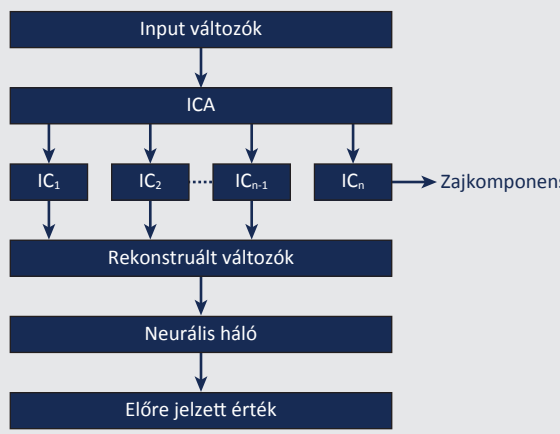
$$X(t)=\sum_{i=1}^{n} h_i(t)+r(t).$$

A kapott IMF-ek közel ortogonálisak egymásra, és nulla közeli az átlaguk (Yu et al., 2008). A reziduális az eredeti idősor trendkomponense, míg az IMF-ek csökkenő sorrendben egyre alacsonyabb frekvenciájúak (Cheng és Wei, 2014).

3.3. Az ICA–BPN és az EMD–BPN hibrid modellek rövid bemutatása

Az első hibrid modell, amelyet én is alkalmaztam, három lépésből épül fel: először ICA-módszer segítségével meghatározza az input változók független komponenseit (IC-ket), majd ezekből TnA (Testing and Acceptance) módszerrel *Cheung* és *Xu* (2001) kiválasztja a zajkomponenst és ezt kiszűri, végül BPN neurális háló segítségével előre jelzi az idősort. Ezt a folyamatot mutatja be a

2. ábra
Az ICA–BPN-modell folyamatábrája

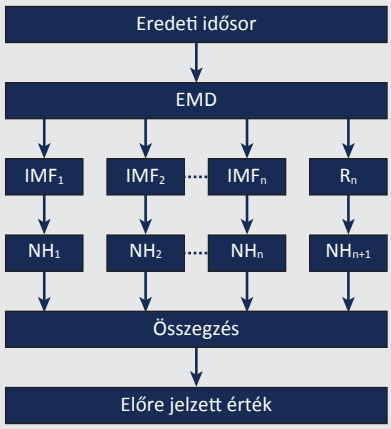


Forrás: saját szerkesztés

A cikkben használt másik hibrid modell is három lépésből épül fel: elsőként felbontjuk az eredeti idősort az EMD-módszer szerint az IMF-komponensekre és a reziduálisra, ezután minden egyes IMF esetén egy BPN-modell segítségével előre jelezzük a következő időszaki

értékeket, majd az eredeti idősor előre jelzett értékét ezek összegeként konstruáljuk. Ezt a hibrid módszert mutatja be a *3. ábra*:

3. ábra
Az EMD–BPN-modell folyamatábrája



Forrás: saját szerkesztés

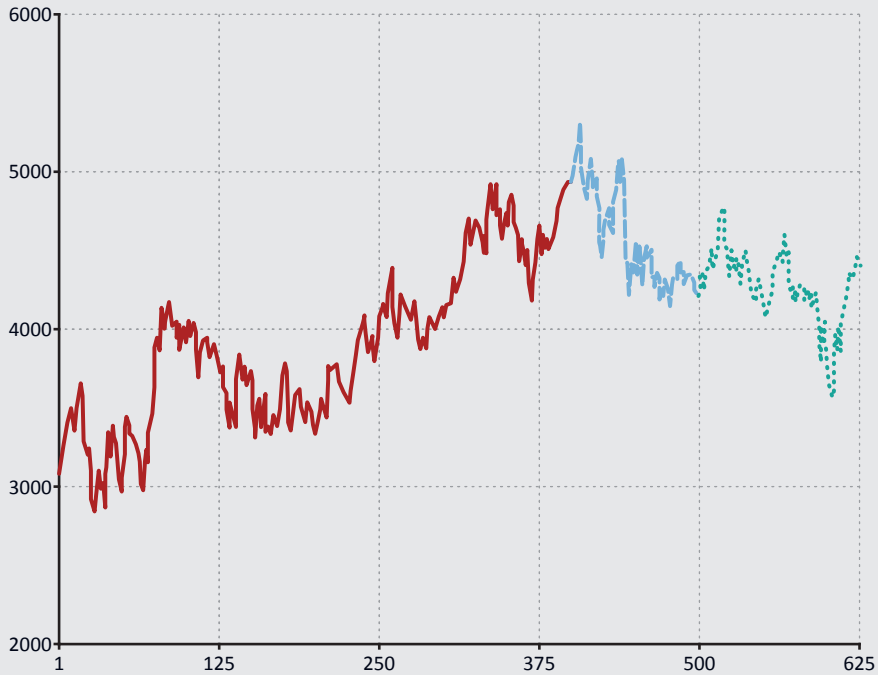
4. Empirikus elemzés

4.1. Adatok és teljesítménykritériumok

Kutatásomban a Budapesti Értéktőzsdén forgalmazott OTP-részvény záró árfolyamának előrejelzése alapján valósítottam meg kereskedési stratégiákat, és vizsgált időszaknak a 2011. 10. 03. és 2014. 04. 11. közötti intervallumot választottam. Az idősort tanuló, tesztelő és validáló adathalmazokra bontottam, ezek aránya 64%, 16% és 20% lett, így a két és fél éves idősor utolsó fél évén teszteltem az előrejelzési modelleket. Az árfolyam alakulását a *4. ábra* mutatja, ahol folytonos, szaggatott és pontozott vonallal jelöltem a különböző halmazokat.

A modellezéshez 8 technikai indikátort választottam, amelyeket széles körben alkalmaznak sok sikerrel, többek között *Kara* et al. (2011) is. Az indikátorok vizsgált időszakbeli statisztikai tulajdonságait az *1. táblázat* tartalmazza.

4. ábra
Az OTP árfolyamának alakulása a vizsgált időszakban



Forrás: saját szerkesztés

1. táblázat
A technikai indikátorok statisztikai jellemzői

	Max.	Min.	Átlag	Szórás
Súlyozott MA	5302	2835	4061,4	516
Momentum	789	-814	15,3	242,3
Stochastic K%	100	0	53,5	31,3
Stochastic D%	98,8	2,6	53,4	27,2
RSI	88,5	15,2	51,6	15,4
MACD	235,5	-231,4	3,4	76,3
LW R%	0	-100	-47	30,6
A/D Oszcillator	100	0	51,2	28,7

saját szerkesztés

4.2. A különböző módszerek előrejelzési eredményei

Kutatásom empirikus részében három adatbányászati modellt alkalmaztam, mivel azonban mindegyik esetén az algoritmusok gyors konvergenciájához szükséges, hogy az inputadatok normalizálva legyenek, ezért első lépésként ezzel kezdtem a modellezést. Minden egyes változó esetén a következő módszerrel transzformáltam az adatokat a [0,1] intervallumba:

$$x'_i = \frac{x_i - x_{min}}{x_{max} - x_{min}},$$

ahol x_{min} és x_{max} az egyes változók minimuma és maximuma az adott idősor esetén.

Első lépésben az egyik legnépszerűbb neurális hálót (backpropagation neural network – BPN) használtam a modellezés során. A megfelelő paraméterek (rejtett rétegben lévő neuronok száma, tanulási ráta) kiválasztáshoz a grid search eljárást használtam. A hálózat input rétege 8 neuronból állt (a magyarázó változók számának megfelelően), míg a köztes rétegben a 11, 12, 13, 14 neuronszámú hálózatokat teszteltem. A hálózatnak egy kimenete volt: a részvény napi hozama. *Lu* (2010) tanulmánya alapján alacsony tanulási ráták (0,01, 0,02, 0,03, 0,04, 0,05) mellett teszteltem a modelleket a tanulási folyamat alatt. Konvergenciakritériumként azt a szabályt alkalmaztam, hogy a tanulási folyamat leáll, ha az RMSE-mutató kisebb lesz, mint 0,0001, vagy eléri az 1000-dik iterációt. Azt a hálózati topológiát választottam optimálisnak, amely esetén a teszthalmazon a legkisebb az RMSE. A *2. táblázat* mutatja a neurális hálózat különböző paraméterei esetén a teszthalmazon mért teljesítményt, amely alapján a későbbiekben validációs halmazon történő modellezés során 8-12-1-es topológiával és 0,05 tanulási rátával rendelkező hálózatot használtam.

A validációs időszakban az eredeti és az előrejelzett árfolyamot, az abszolút hibát, illetve az előjeltalálatot mutatja be a*z 5. ábra.*

Mivel a pénzügyi idősorokra jellemző, hogy magas a zaj/jel arány, ezért második modellemben a BPN-háló használata előtt függetlenkomponens-elemzéssel kiszűrtem az inputváltozókból a zajt. Ehhez szükséges volt egyrészt a független komponensek (IC-k) előállítása, majd a TnA algoritmus segítségével a zajkomponens definiálása.

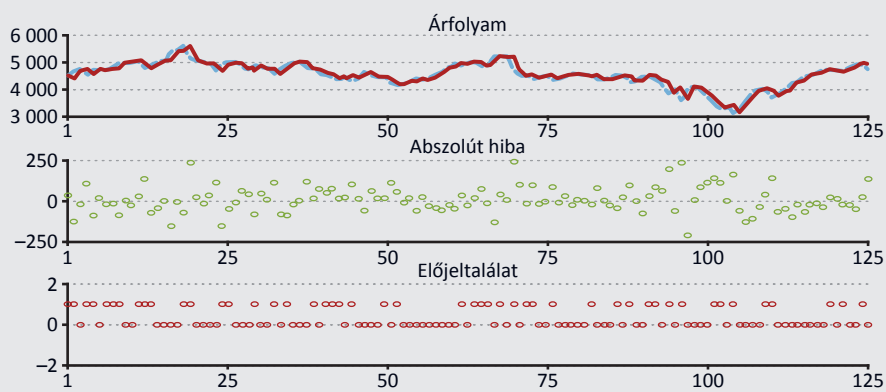
A TnA algoritmus során egyesével elhagytam az egyes IC-ket, majd ezután rekonstruáltam az input mátrixot, és megvizsgáltam hogy ez mennyire tér el az eredetitől. Az eltérést az RHD-mutatóval mértem. Mivel 8 input változót használtam, ezért 7-szer kell ezt a műveletet elvégeznem, hogy megtaláljam a zajkomponenst. Ezek RHD-értékeit mutatja a *3. táblázat*.

2. táblázat
Különböző paraméterű BPN-hálózatok hibája a teszthalmazon

Rejtett rétegben lévő neuronok száma	Tanulási ráta	Validációs RMSE
11	0,01	0,124111
	0,02	0,120873
	0,03	0,119689
	0,04	0,119021
	0,05	0,118578
12	0,01	0,120424
	0,02	0,117532
	0,03	0,116893
	0,04	0,116581
	0,05	0,116369
13	0,01	0,124840
	0,02	0,123034
	0,03	0,121980
	0,04	0,121219
	0,05	0,120619
14	0,01	0,124489
	0,02	0,120798
	0,03	0,119771
	0,04	0,119247
	0,05	0,118872

saját szerkesztés

5. ábra
A BPN-modell előrejelzési pontossága a validációs időszakon



Forrás: saját szerkesztés

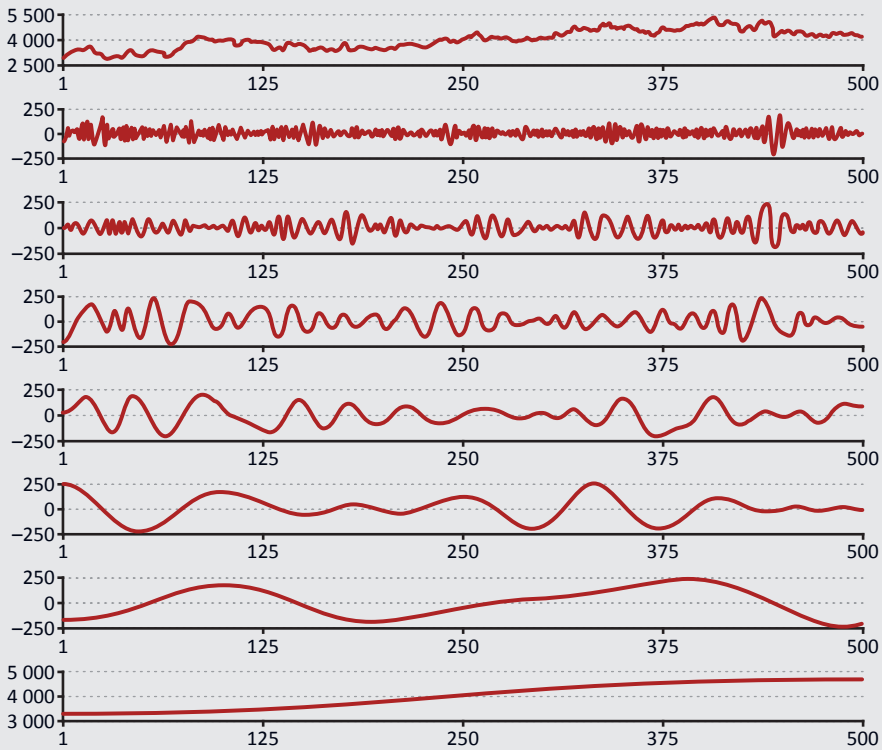
3. táblázat Különböző rekonstruált inputmátrixok RHD-értékei	
Főkomponensek	**RHD**
IC1, IC2, IC3, IC4, IC5, IC6, IC7	4,3674
IC1, IC2, IC3, IC4, IC5, IC6, IC8	3,6260
IC1, IC2, IC3, IC4, IC5, IC7, IC8	4,4830
IC1, IC2, IC3, IC4, IC6, IC7, IC8	2,4118
IC1, IC2, IC3, IC5, IC6, IC7, IC8	3,7873
IC1, IC2, IC4, IC5, IC6, IC7, IC8	3,9655
IC1, IC3, IC4, IC5, IC6, IC7, IC8	7,1473
IC2, IC3, IC4, IC5, IC6, IC7, IC8	7,7748
saját szerkesztés	

A táblázat alapján megállapítható, hogy az ötödik komponens a zaj. A modell második lépéseként a rekonstruált változók felhasználásával BPN-hálózatot építettem. Az optimális paraméter kiválasztása teljesen hasonlóan működött a korábban bemutatotthoz, az ICA–BPN-modell esetén is a 8-12-1 topológiájú hálózat lett az optimális.

A harmadik módszer esetén a tőzsdei idősorok komplex dinamikája miatt az eredeti idősort az EMD-módszer segítségével IMF-ekre bontottam fel, és ezeket külön-külön előre jelezve, majd összeadva kaptam meg az eredeti idősor előre jelzett értékét. Több tanulmányhoz hasonlóan (*Yu* et al., 2008; *Cheng és Wei,* 2014), ennél a módszernél én is az árfolyamokat jeleztem előre. A *6. ábra* mutatja az OTP árfolyamának empirikus alapú dekompozícióját.

A legfelső sorban jeleztem az eredeti idősort, majd alatta az egyre kisebb frekvenciájú IMF-eket (IMF1, IMF2, …, IMF8), és végül legutolsóként a trendnek megfeleltethető reziduumot. A módszer második lépésenként minden egyes IMF-et különböző paraméterű neurális hálókkal előre jeleztem, majd a kapott értékeket aggregálva kaptam meg az OTP következő napi záró árfolyamának értékét. Mivel ebben az esetben végső soron 8 idősort kellett előrejeleznem, illetve ezekhez meghatároznom az optimális inputok számát (hány késleltetést alkalmazzak a NAR modellben), ezért ez a korábbi két modellhez képest jóval komplexebb és időigényesebb folyamat volt. A probléma megoldhatósága érdekében a késleltetések számát minden egyes IMF esetén 10-ben határoztam meg *Mingming* és *Jinliang* (2012) alapján, így csak az optimális neuronszámot és tanulási rátát kellett megkeresnem. Ezeket a különböző IMF-ek esetén a *4. táblázat* mutatja.

6. ábra
Az OTP árfolyamainak empirikus alapú dekompoziciója



Forrás: saját szerkesztés

4. táblázat
Különböző IMF-ek optimális paraméterei

IMF	Neuronok száma	Tanulási ráta
1	12	0,05
2	12	0,05
3	12	0,05
4	12	0,025
5	12	0,025
6	12	0,025
7	13	0,025
8	13	0,025

saját szerkesztés

A három modell optimális paramétereinek megtalálása után a validációs időszakra való előrejelzéshez használtam őket *(5. táblázat)*.

5. táblázat
Különböző módszerek teljesítménye a validációs halmazon
(OTP)

Modell	RMSE	MAPE (%)	DA (%)
BPN	0,018864	113,38	61,6
ICA–BPN	0,018738	107,79	60,8
EMD–BPN	0,026672	292,47	56,8

saját szerkesztés

A táblázatok alapján látni, hogy a fejlettebb hibrid módszerek előjel-előrejelzési aránya nem jobb a sima BPN-modellnél, azonban később érdemes lesz azt is megnézni, hogy az elért profit szempontjából felülmúlják-e az első modellt.

Előtte azonban még megvizsgáltam, hogy a három módszert kombinálva javulnak-e az előrejelzési eredmények. Ahogy korábban említettem, a kombinálás segítségével ki tudjuk küszöbölni az egyes módszerek hátrányait, ezáltal jobb előrejelzést és magasabb profitot tudunk elérni. A három módszer (sima átlag, bayesi átlag, GRR) háromfajta kombinálásával kapott előrejelzés eredményeit foglalja össze a 6. táblázat.

6. táblázat
A három módszer kombinálásával kapott eredmények
(OTP)

Modell	RMSE	MAPE (%)	DA (%)
Átlag	0,018854	144,82	64,8
Bayes-i átlag	0,018733	107,87	60,8
GRR	0,019087	151,21	61,6

saját szerkesztés

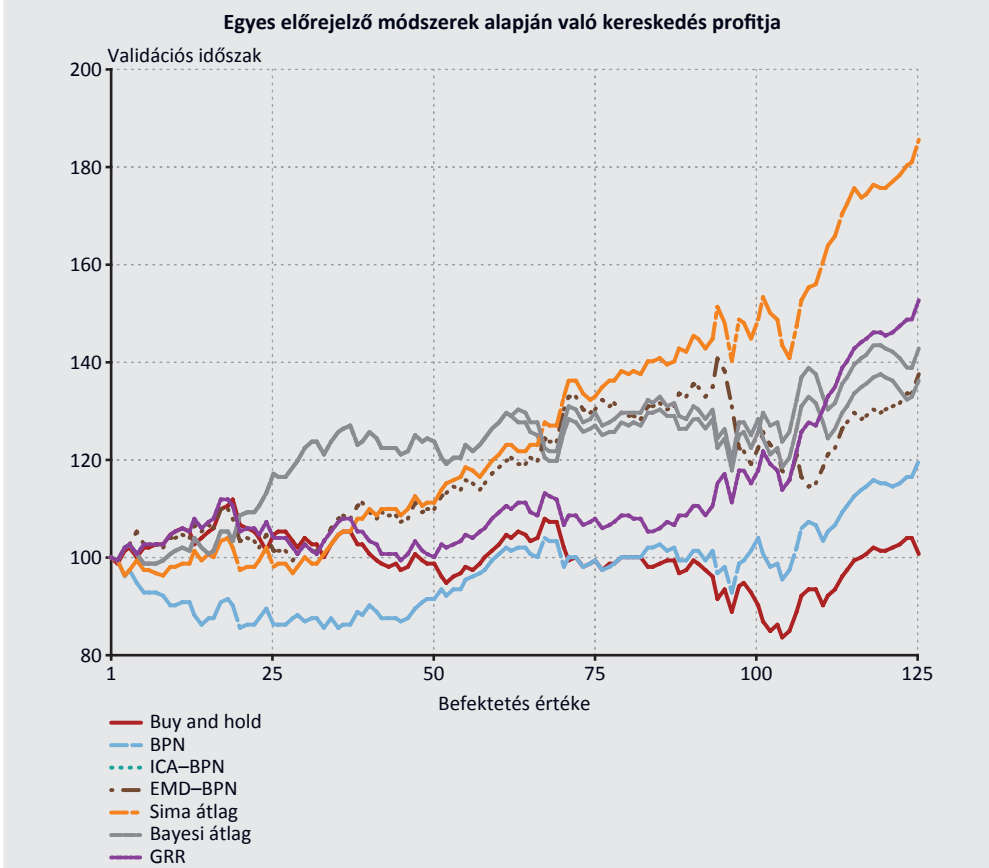
A három adatbányászati és a három kombinációs modell validációs halmazon elért profitjait pedig a *7. táblázat* és a *7. ábra* mutatja (0,1%-os tranzakciós költség figyelembevételével).

7. táblázat
A 6 modell által generált profit a validációs halmazon
(OTP)

	Buy and hold	BPN	ICA–BPN	ICA–EMD	Átlag	Bayes-i átlag.	GRR
Éves hozam	2,50%	36,46%	71,53%	64,01%	124,27%	62,71%	85,16%
Éves volatilitás	30,63%	30,22%	29,68%	29,92%	28,73%	29,81%	29,52%

saját szerkesztés

7. ábra
Az egyes adatbányászati és kombinációs modellek alkalmazásával elérhető profit a validációs halmazon
(OTP)



Forrás: saját szerkesztés

A táblázatot vizsgálva, egyrészt megfigyelhetjük, hogy habár az előjel-előrejelzési arány a két komplexebb adatbányászati modell esetén rosszabb volt, mint a sima neurális hálónál, profit szempontjából azonban messze felülmúlják azt. Másrészt mindhárom modell jobban teljesít, mint a „buy and hold". Emellett azt is feltűnő, hogy a három kombinációs módszerből profit szempontjából a sima átlagolás teljesít a legjobban. Ez elsőre furcsának tűnhet, mivel ez a legkevésbé szofisztikált átlagolási módszer. Ha azonban belegondolunk, hogy a másik kettő a tanuló- és teszthalmazon elért hibák alapján súlyozza a modelleket, és itt a profit szempontjából jobban teljesítő modellek (ICA–BPN és ICA–EMD) a sima neurális hálóhoz képest rosszabbul teljesítettek, így ezeket kisebb súllyal átlagolja, akkor már érthető, hogy emiatt alacsonyabb profitot eredményeznek ezek a kombinációk. Tehát a két módszer (ICA–BPN és EMD–BPN) a validációs időszakon közel ugyanolyan jól jelez előre, mint a teszt- és tanuló időszakon, míg a sima neurális háló rosszabbul; így, amikor a szofisztikáltabb modellek nagyobb súllyal szerepelnek a kombinációban (sima átlagolás), akkor az több profitot eredményez. Érdemes lenne a későbbi kutatások során megvizsgálni, hogy mi történne, ha nem az RMSE, hanem a profit alapján történne a másik két kombinációs módszer súlyainak megválasztása.

5. A módszer alkalmazásának kihívásai, további kutatási lehetőségek, konklúzió

Ahogy az előző fejezet eredményei alapján láttuk, az adatbányászati technikák segítségével megvalósuló, aktív portfóliókezelés képes felülmúlni a „buy and hold" stratégiát. A modellalkalmazás ugyanakkor nem egyszerű, megvannak a maga nehézségei és korlátai. Ezek közé tartozik az optimális módszerek, paraméterek kiválasztása, illetve bizonyos időszakonként (akár naponta is) a modellek újrakalibrálása, ami rendkívül időigényes folyamat, és nagy körültekintést követel meg.

A modellek folytonos fejlesztésére emiatt a következőkben pár továbblépési lehetőséget, kutatási irányt szeretnék bemutatni. Természetesen érdekes lehet megvizsgálni, hogy a különböző értékpapírokra ugyanazok a módszerek adják-e a legpontosabb előrejelzést, illetve ha nem, akkor egy adott részvénynek milyen jellemzője okozza az eltérést. Továbbá érdemes figyelembe venni és alkalmazni a cikkben kevésbé tárgyalt adatbányászati módszereket a genetikus algoritmusok használatától kezdve a döntési fákon át a szöveges adatbányászatig. Utóbbi az elmúlt 2-3 év leggyorsabban fejlődő területe: a piacon megjelenő híreket automatizálva elemzik, és megállapítják azok várható hatását az egyes értékpapírokra (*Hagenau* et al., 2013).

Kereskedés szempontjából fontos lehet vizsgálni, hogy milyen profitot tudunk elérni tőkeáttétel esetén, hiszen erre rendkívül sok piacon van lehetőség. *Sermpinis* et al. (2012) bemutatott erre egy módszert, ahol az árfolyamok mellett a papírok volatilitását is előrejelezte, és aszerint határozta meg a tőkeáttétel mértékét, hogy ez mekkora volt (magas tőkeáttétel alacsony volatilitás esetén, alacsony pedig magas volatilitás esetén). Ez azonban szinte egyedi eset, a kutatások nagy része nem foglalkozik ezzel. Érdemes lehet ezért alaposabban megvizsgálni egyrészt a volatilitást előrejelzésének lehetőségét, másrészt az ennek alapján felállított tőkeáttételi szabályok érvényességét.

A felsorolt indokok alapján láthatjuk, hogy az adatbányászattal ugyan jelentős többlethozamot lehet elérni a hagyományos befektetési stratégiákhoz képest, azonban a kivitelezése rendkívül komplex probléma, amely komoly erőforrásokat és szakértelmet követel meg.

Felhasznált irodalom

Armstrong, J. S. (1989): Combining forecasts: the end of the beginning or the beginning of the end? *International Journal of Forecasting*, 5, 585–588.

Atsalakis, G. S. – Valavanis, K. P. (2009): Surveying stock market forecasting techniques – Part II. Soft computing methods. *Expert Systems with Applications*, 36(3), 5932–5941.

Beckmann, C. F. – Smith, S. M. (2004): Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Transactions on Medical Imaging*, 23(2), 137–152.

Bell, A. J. – Sejnowski, T. J. (1995): An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7, 1129–1159.

Cao, Q. – Parry, M. E. (2009): Neural network earnings per share forecasting models: A comparison of backward propagation and the genetic algorithm. *Decision Support Systems*, 47(1), 32–41.

Chang, P-C. – Liu, C-H. – Lin, J-L. – Fan, C-Y. – Ng, C. S. P. (2009): A neural network with a case based dynamic window for stock trading prediction. *Expert Systems with Applications*, 36(3), 6889–6898.

Chauvin, Y. – Rumelhart, D. E. (1995): Backpropagation: Theory, architectures, and applications. New Jersey: Lawrence Erlbaum associates.

Chavarnakul, T. – Enke, D. (2008): Intelligent technical analysis-based equivolume charting for stock trading using neural networks. *Expert Systems with Applications*, 34(2), 1004–1017.

Chen, A-S. – Leung, M. T. – Daouk, H. (2003): Application of neural networks to an emerging financial market: Forecasting and trading the Taiwan Stock Index. *Computers – Operations Research*, 30(6), 901–923.

Chen, C. F. – Lai, M. C. – Yeh, C. C., (2012): Forecasting tourism demand based on empirical mode decomposition and neural network. *Knowledge-Based System* 26, 281–287.

Cheng, C-H. – Wei L-Y. (2014): A novel time-series model based on empirical mode decomposition for forecasting TAIEX. *Economic Modelling*, 36, 136–141.

Cheung, Y. M. – Xu, L. (2001): Independent component ordering in ICA time series analysis. *Neurocomputing*, 41(1–4), 145–152.

Chun S-H., Kim S. H. (2004): Data mining for financial prediction and trading: application to single and multiple markets. *Expert Systems with Applications*, 26 (2), 131–139.

Dai, W. – Wu, J-Y. – Lu, C-J. (2012): Combining nonlinear independent component analysis and neural network for the prediction of Asian stock market indexes. *Expert Systems with Applications*, 39(4), 4444–4452.

David, V. – Sanchez, A. (2002): Frontiers of research in BSS/ICA. *Neurocomputing*, 49(1), 7–23.

Déniz, O. – Castrillón, M. – Hernández, M. (2003): Face recognition using independent component analysis and support vector machines. Pattern Recognition Letters, 24(13), 2153–2157.

Duan, W-Q. – Stanley, H. E. (2011): Cross-correlation and the predictability of financial return series. *Physica A: Statistical Mechanics and its Applications*, 390(2), 290–296.

Enke, D. – Thawornwong, S. (2005): The use of data mining and neural networks for forecasting stock market returns. *Expert Systems with Applications*, 29(4), 927–940.

Granger, C.W.J. – Ramanathan, R. (1984): Improved methods of combining forecasts. *Journal of Forecasting*, 3(2), 197–204.

Guo, Z. – Zhao, W., Lu, H. – Wang, J. (2012): Multi-step forecasting for wind speed using a modified EMD-based artificial neural network model. *Renewable Energy*, 37(1), 241–249.

Hall, J. W. (1994): Adaptive selection of US stocks with neural nets. In Deboeck, J. G. (ed.): *Trading on the edge: Neural, genetic and fuzzy systems for chaotic financial markets*. Wiley Finance, 45–65.

Hansen, J. V. – Nelson, R. D. (2002): Data mining of time series using stacked generalizers. *Neurocomputing*, 43(1), 173–184.

Hagenau, M. – Liebmann, M. – Neumann, D. (2013): Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision Support Systems*, 55(3), 685–697.

Huang, N. E. – Shen, Z. – Long, S. R. – Wu, M. C. – Shih, H. H. – Zheng, Q. – Yen, N. C. – Tung, C. C. – Liu, H.H. (1998): The empirical mode decomposition and the Hilbert spectrum for nonlinear and nonstationary time series analysis. *Proceedings of the Royal Society of London A – Mathematical, Physical – Engineering Sciences*, Series A, 454, 903–995.

Huang, S-C. – Chuang, P-J. – Wu, C.F. – Lai, H-J. (2010): Chaos-based support vector regressions for exchange rate forecasting. *Expert Systems with Applications*, 37(12), 8590–8598.

Kara, Y. – Boyacioglu, M. A. – Baykan, Ö. K. (2011): Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of Istanbul Stock Exchange. *Expert Systems with Applications*, 38, 5311–5319.

Kapelner, T. – Madarász, L. V. (2012): Független komponens analízis és empirikus tesztjei kötvényhozamok felhasználásával. TDK-dolgozat.

Li, T. – Li, Q. – Zhu, S. – Ogihara, M. (2003): A survey on wavelet applications in data mining. *SIGKDD Explorations*, 4(2), 49–68.

Lu, C-J., Lee, T-S. – Chiu, C-C. (2009): Financial time series forecasting using independent component analysis and support vector regression. *Decision Support Systems* 47(2), 115–125.

Lu, C-J. (2010): Integrating independent compontnt analysis-based denoising scheme with neural network for stock price prediciton. *Expert Systems with Applications*, 37(10), 7056-7064.

Mingming, T. – Jinliang, Z. (2012): A multiple adaptive wavelet recurrent neural network model to analyse crude oil prices. *Journal of Economics and Business*, 64(4), 275–286.

Ni, H. – Yin, H., (2009): Exchange rate prediction using hybrid neural networks and trading indicators. *Neurocomputing*, 72(13–15), 2815–2823.

Oja, E. – Kiviluoto, K. – Malaroiu, S. (2000): Independent component analysis for financial time series. In Proceeding of the IEEE 2000 adaptive systems for signal processing, communications, and control symposium, Lake Louise, Canada. 111–116.

Oh, K. J. – Kim, K.-J. (2002): Analyzing stock market tick data using piecewise nonlinear model. *Expert System with Applications*, 22(3), 249–255.

Sermpinis, G. – Dunis, C. – Laws, J. – Stasinakis, C. (2012): Forecasting and trading the EUR/USD exchange rate with stochastic Neural Network combination and time-varying leverage. *Decision Support Systems*, 54(1), 316-329.

Swanson, N.R. – Zeng, T. (2001): Choosing among competing econometric forecasts: regression-based forecast combination using model selection. *Journal of Forecasting*, 20(6), 425–440.

Timmermann, A. (2006): Chapter 4: Forecast Combinations. In Elliott, G. – Timmermann, A. (eds.): *Handbook of Economic Forecasting* 1, Elsevier, 135–196.

Thawornwong, S. – Enke, D. (2004): The adaptive selection of financial and economic variables for use with artificial neural networks. *Neurocomputing*, 56, 205–232.

Vincent, H. T. – Hu, S-L. J. – Hou, Z. (1999): Damage detection using empirical mode decomposition method and a comparison with wavelet analysis. Proceedings of the Second International Workshop on Structural Health Monitoring, Stanford, 891–900.

Vellido, A. – Lisboa, P. J. G. – Vaughan, J. (1999): Neural networks in business: A survey of applications (1992–1998): *Expert Systems with Applications*, 17(1), 51–70.

Wang, Y-F. (2003): Mining stock prices using fuzzy rough set system. *Expert System with Applications*, 24(1), 13–23.

Yaser, S. A-M. – Atiya, A. F. (1996): Introduction to financial forecasting. *Applied Intelligence*, 6, 205–213.

Yu, L. – Wang, S. – Lai, K.K. (2008): Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm. *Energy Economics*, 30(5), 2623-2635.

Zhang, G. – Patuwo, B. E. – Hu, M. Y. (1998): Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, 14, 35–62.

Zhang, Y. D. – Wu, L. N. (2009): Stock market prediction of S–P 500 via combination of improved BCO approach and BP neural network. *Expert Systems with Applications*, 36, 8849–885.