

TDK dolgozat

László Zsolt

Debrecen
2014

Debreceni Egyetem
Informatikai Kar
Komputergrafika és Képfeldolgozás Tanszék

RSS hírek klaszteranalízise és FOREX árfolyamokra gyakorolt hatásának vizsgálata

Külső témavezető:

Kovács György
CEO, Analytical Minds Ltd.

Készítette:

László Zsolt
gazdaságinformatikus hallgató
4. évfolyam

Debrecen
2014

Tartalomjegyzék

1. Absztrakt	4
2. Bevezetés	5
2.1. RSS-hírcsatornák	7
2.1.1. Metaadatok	7
2.1.2. Az RSS technológia	8
2.1.3. Példa	8
2.2. FOREX kereskedés	9
2.2.1. A devizakereskedésről általánosan	9
3. Kapcsolódó modellek	11
3.1. Kendall és a bolyongó folyamatok	11
3.2. A Microsoft-részvények árfolyamának alakulása	12
3.3. Eugene Fama munkássága	13
3.3.1. Fama tapasztalatainak összefoglalása	14
3.3.2. A hatékony tőkepiacokról - másodszor	15
3.4. A piaci hatékonyságra irányuló további vizsgálatok	16
3.5. Összefoglalás	17
4. Az alkalmazott módszerek	18
4.1. Hierarchikus, agglomeratív klaszterezés	18
4.2. A használt különbözőségi függvények	19
4.2.1. A spektrum és bináris spektrum kernelek	20
4.3. Regressziószámítás	21
4.4. Logaritmikus hozamszámítás	23
4.4.1. Általános értelmezés	23
4.4.2. Ismeretlen hozam	23
4.5. Statisztikai tesztek	23
5. Eredmények	25
5.1. A felhasznált hírcsatornák	25
5.2. A vizsgált devizapárok	26
5.3. A kapott klaszterek	26
5.3.1. Példa	27
5.4. A nagy számosságú klaszterek vizsgálata	27

6. Összefoglalás	37
7. Munkamegosztás	39

Ábrák jegyzéke

2.1.	Forrás: MetaTrader 4, © 2000-2014, MetaQuotes Software Corp.	5
2.2.	Forrás: MetaTrader 4, © 2000-2014, MetaQuotes Software Corp.	6
3.1.	Az ábra pontjai a Microsoft-részvények két egymást követő napi hozamait mutatják 1990. március és 2001. július között. A bekarikázott pontok például annak felelnek meg, hogy valamely napon 1 százalékkal nőtt a részvény ára, a következő napon viszont 1 százalékos csökkenés következett be. A pontdiagram nem jelez szignifikáns kapcsolatot az egymást követő napok árfolyamváltozásai között.	13
5.1.	Magas klaszter-számosság magas egyezési aránnyal	32
5.2.	Ugyanazon, nagy számosságú klaszter vizsgálata más regressziós beállításokkal	33
5.3.	EUR-JPY devizapáros, 30-1 paraméterezés, 0,67-es egyezési arány, logaritmikus hozam-számítással produkálva	36

1. fejezet

Absztrakt

A devizakereskedelem a gazdasági szféra egyik legfontosabb és legaktívabb területe, rendkívül sok gazdasági szereplővel és befolyásoló tényezővel. A legtöbb résztvevő elsődleges célja, hogy az említett piacon az árfolyamok ingadozásából adódó lehetőségeket kihasználja, és ezáltal a lehető legnagyobb profitra tegyen szert.

A sikeres kereskedéshez mindenkinek szüksége van valamilyen prognózisra, számításokra, előzetes tapasztalatra, vagy megfigyelésekre, tehát egyfajta gondolkodásmódra, amely segítségével megpróbálja megbecsülni a jövőbeli árfolyammozgásokat.

Tudományos dolgozatunk célja, hogy fényt derítsünk arra, vajon az RSS hírcsatornákon felbukkanó hírek, információk kimutatható hatással vannak-e az árfolyamok alakulására, megfigyelhetők-e rendszerű, hasonló jellegű, gyakori változások, melyek a hírek megjelenésének időpontjában, vagy azt követően mutatkoznak az egyes devizapárok árfolyamaiban.

Amennyiben a vizsgálat alapjául szolgáló területet, forrásokat távolabbról és általánosabban tekintjük, egy, már ismert és sokak által analizált kérdéskörhöz, nevezetesen a hatékony piacok elméletéhez jutunk. Az általunk alkalmazott modell mellett a piaci hatékonyság fogalmát, illetve néhány, a már erre irányuló tesztet, modellt is megvizsgálunk, majd megpróbáljuk elhelyezni ezekhez képest saját szemléletünket és módszereinket.

Az elemzésekhez különböző hírforrásokból több tízezer hírt gyűjtöttünk össze 2013. októberétől 2014. októberéig, ezen eseményeket különböző klaszterezési eljárásokkal csoportosítottuk, majd az árfolyamok alakulásának tendenciáját és mértékét lineáris regresszió, illetve logaritmikus hozamszámítás segítségével igyekeztük megbecsülni, továbbá próbáltunk kapcsolatot teremteni a hírcsatornákon keresztül beáramló információk, és a kalkulált árfolyam-ingadozások között. Az eredményekből arra következtethetünk, hogy a hatékony piaci törvények érvényesülése mellett az árfolyamokban az RSS hírfolyamok hatása felfedezhető, azonban pontos, kvantitatív jellemzéshez további, részletesebb vizsgálatokra van szükség.

2. fejezet

Bevezetés

2014. március 18-án, 13:40-kor az alábbi hír jelent meg a BBC News Business hírcsatornán:

„Az angol központi bank két új alelnököt nevezett ki: Ben Broadbent került a monetáris politika élére, illetve Nemat Shafik lett a felelős a piaci és banki ügyekért.”

Az 2.1 ábrán a GBP-USD árfolyam alakulása látható a hír megérkezésének pillanatában, illetve az azt követő néhány órában. Jól látható, hogy a cikk megjelenése utáni percekben az angol font értéke gyengülni kezdett az amerikai dollárral szemben, majd viszonylag kis ingadozással, egy újabb szinten stabilizálódott, amit 120 perc elteltével ismét zuhanás követett.

Természetesen nem csupán az előző devizapárosra, hanem sok másikra is hatással volt az angol jegybank szervezetében bekövetkezett változás. Az 2.2 ábrán az angol font és az euró közötti kapcsolat látható, ugyanazon időintervallumban, mint az előző ábra esetén. A kép bizonyítja, hogy a bejelentés hatására az árfolyam hatalmas volatilitással mozgott az elkövetkező órákban.

Ha ismerjük a megválasztott pénzügyi vezető, avagy politikus nézeteit, korábbi tevékenységeit, az adott területet érintő kérdésekben támasztott álláspontját, mekkora eséllyel tudunk helyesen következtetni arra, hogy hivatala során milyen intézkedéseket fog eszközölni, és ezek hogyan fognak hatni különböző pénzügyi folyamatokra?



2.1. ábra. Forrás: MetaTrader 4, © 2000-2014, MetaQuotes Software Corp.



2.2. ábra. Forrás: MetaTrader 4, © 2000-2014, MetaQuotes Software Corp.

Az Internet elterjedésével a devizakereskedelem bárki számára elérhetővé vált. A nyomtatott és sugárzott média nyilvánvaló háttérbeszorulásával a kereskedők számára elsődleges hírforrásként a hírportálok szolgálnak, s a kereskedés tárgyához kapcsolódó hírek alapján próbálnak következtetni az árfolyamváltozások jellegére. Mindezt jól alátámasztja az alábbi, Várkúti Géza *FOREX-trading* című alapművéből[1] származó idézet:

„Ha elfogadjuk, hogy az árat a tréderek mozgatják, vagyis azok az emberek, akik trédeket indítanak, akkor már látjuk: az árat nem a gazdasági helyzet, hanem a kereskedelem mozgatja. Ha pl. az Európai Központi Bank emeli vagy csökkenti az alapkamatot, a tréderek reagálnak erre a hírre attól függően, hogy ők mit gondolnak az adott intézkedésről.

Ezt nagyon fontos, hogy megértse - a pénzpiacot az emberek emóciói mozgatják. A háttér lehet egy hír egy adott ország gazdaságáról, de attól még nem mozdul meg a kurzus. Először akkor moccan, amikor a tréderek a hírre reagálnak.”

Vajon a kiváltott hatás jellege és iránya a témájában és alapadataiban (a hír középpontjában álló személy, intézmény, esemény, folyamat, ország, terület stb.) megegyező hírek esetén hasonló? A bekövetkezett trendek, ingadozások újra és újra felbukkannak a megfelelő pillanatokban? Melyek azok a hírek, hírcsoportok, amelyek esetén a publikálást követő szisztéma állandó, vagy közel annak tekinthető?

A kérdésekre adott válaszok igen érdekes és új megvilágításba helyezhetik a devizapiacok jelenségeit, s elősegíthetik egy-egy nem várt fordulat mélyebb megértését, szükség esetén hatékonyabb kezelését. A hírek devizapiacokra gyakorolt hatásának automatikus vizsgálata azonban igen nehéz, interdiszciplináris feladat. Célkitűzéseit tekintve a Big Data területéhez kapcsolódik, hiszen másodlagos információkat szeretnénk kinyerni eredendően különböző célú, egymástól technikailag független adatokból. A hírek tartalmának megfelelő értelmezése a természetes nyelvek feldolgozásához (*Natural Language Processing* - NLP) kapcsolódó, meglehetősen komplikált feladat. Az elemzéseket nagyban segíthetik a szemantikus web már elérhető elemei, amelyek a híreket azok tartalmának értelmezését elősegítő metaadatokkal egészítik ki. Az árfolyamok elemzése alapvetően az idősoranalízis témaköréhez kapcsolódik, ugyanakkor a nagy mennyiségű adat feldolgozása miatt elengedhetetlen a hatékony, lehetőleg párhuzamos implementációk készítése.

Az árfolyamok gyűjtése és rögzítése könnyű feladat, hiszen több weboldalon is elérhetőek az időbélyegekkel ellátott árfolyam adatok. A hírek szisztematikus gyűjtéséhez a nagy hírszolgáltatók gazdasági

témájú RSS hírcsatornáit választottuk forrásul.

Az egyik legnehezebb kérdés a „hasznos tartalmú” hírek definiálása. Jóllehet az emberi olvasó számára kézenfekvő megítélni két hír vagy mondat tartalmi hasonlóságát, ennek automatizálása igen nehéz feladat, a mai napig nem léteznek olyan eszközök, melyek a természetes nyelven kifejezett mondatok, állítások tartalmát hibátlanul megértik. Sok esetben még egy emberi olvasó is csak megfelelő háttérismeretekkel, sok éves tapasztalattal értelmezhet többé-kevésbé helyesen, de bizonyosan nem teljes mélységében egy gazdasági hírt. A hasznos tartalmú hírek összerendezésére, azaz megkeresésére szintaktikus és statisztikai eszközöket vetünk be, nevezetesen megfelelő, sztringek között értelmezett távolságfüggvények/különbözőségi függvények segítségével próbáljuk a hasznos híreket csoportokba, klaszterekbe sorolni. Mivel sztringek között számos különböző módon definiálhatunk távolság-, illetve különbözőségi függvényt, több, széles körben elterjedt függvénnyel is elvégezzük a klaszterezést, s az eredményként előálló csoportok átmérőjének vizsgálatával megpróbáljuk kiemelni azt a módszert, amely a legtöbb szempontból a lehető legkisebb átmérőjű csoportokat, klasztereket állítja elő. A legmegfelelőbb klaszterező eljárás kiválasztását követően az ennek segítségével előálló csoportokban lévő híreket vizsgáljuk, s azt, hogy a hírek megjelenését követően valamely devizapár árfolyama mutatott-e azonos irányú mozgásokat. Amennyiben sikerül kimutatni összefüggéseket az RSS-hírek és a FOREX árfolyamok között, az hozzájárulhat a piacok mélyebb megértéséhez; következtetéseket vonhatunk le arra vonatkozóan, hogy milyen események lehetnek hatással a devizapiacokra.

Legjobb tudomásunk szerint az RSS-hírcsatornák és FOREX-árfolyamok közötti összefüggések tudományos igényességű vizsgálatával kapcsolatban nem születtek még általánosan elfogadott módszerek és modellek. Jelen munkát a 2014. áprilisában bemutatott dolgozattal[2] összehasonlítva a legfontosabb különbség az, hogy az RSS csatornákon megjelenő híreket klaszterezéssel csoportosítjuk, s az árfolyamváltozásokat az azonos csoportban szereplő hírek megjelenését követő időpillanatokban vizsgáljuk.

A dolgozat felépítése a következő. A bevezetés hátralévő részében röviden bemutatjuk az RSS technológia alapjait, továbbá vázoljuk a FOREX kereskedés működését. Futólag áttekintjük szerepét a gazdaságban, az elmúlt évtizedekre jellemző fejlődési tendenciáit és a kereskedés manapság legelterjedtebb formáit. A harmadik fejezetben a hatékony piaci elméletekkel foglalkozunk, sajnos viszonylag rövid formában, mindenesetre nagy hangsúlyt fektetve Eugene Fama tanulmányaira. A negyedik fejezetben az általunk alkalmazott klaszterezési eljárásokat és az egyes csoportok vizsgálatához felhasznált módszertant ismertetjük, míg az ötödik fejezetben értelmezzük a hasznos hírek vizsgálata során kapott eredményeket. A hatodik fejezetben az összefoglalással és a további fejlesztések bemutatásával zárjuk a dolgozatot.

2.1. RSS-hírcsatornák

2.1.1. Metaadatok

Mielőtt az RSS működésével és céljaival megismerkednénk, bevezetjük a metaadat fogalmát. Legegyyszerűbben talán úgy definiálhatjuk, hogy a metaadat adat az adatról. Például a könyvtári nyilvántartó kártya adatokat tartalmaz a könyv írójáról, címről, műfajáról, tárolási helyéről, a kölcsönzésekről, stb.. A nyilvántartó kártya tehát különböző aspektusokból írja le a könyvet, mind tartalmilag, mind magát a fizikai könyv objektumot. A kártyán szereplő adatok metaadatok, melyek célja a könyvek általános jellemzése, a több szempontú kategorizálás lehetővé tétele.

A metaadatok az Internethez kapcsolódó alkalmazások számára különösen lényegessé váltak, mivel napjainkban hatalmas információömegekből kell kiszűrniünk a számunkra hasznos információkat.

2.1.2. Az RSS technológia

A web 2.0 egyik legfontosabb eredménye a tartalomhoz való hozzáférés technológiai megújítása, amely lehetővé teszi, hogy ne kelljen „utánajárni” híreknek, az újonnan publikált tartalmakról automatikusan értesüljünk. Az első hírgyűjtő/hírelosztó szoftvereket (2004. körül) csak a nagyobb hírügynökségeken alkalmazták, mint például a Reuters, de ahogyan az átlagos felhasználók által publikált tartalmak, hálózati naplók (blogok) sokasodni kezdtek, rohamosan nőtt erre az igény a mindennapi életben, az internetező nagyközönség, az átlagos felhasználók köreiből is.

Az *RDF Site Summary* (RSS) technológia egy XML alapú jelölőnyelv szabvány. Legelterjedtebb alkalmazásaiban az RSS-dokumentumok általános értelemben tekintett híreket, illetve azokhoz kapcsolódó metaadatokat kódol(hat)nak. Az RSS állományok automatikus továbbítása és értelmezése képi az előző bekezdésben leírt hírszolgáltatások alapját. A technológia nevét újabban a *Really Simple Syndication*, azaz „nagyon egyszerű információterjesztés” kifejezés rövidítésének is tekintik. Napjainkra számtalan entitás használja az RSS-technológiát, ha híreket szeretne megosztani az RSS-csatornára feliratkozott olvasókkal, például hírportálok, blogok, szoftverfrissítési listák.

2.1.3. Példa

A szemléletesség kedvéért az alábbiakban egy, a BBC News csatornáról származó RSS-állományt mutatunk be.

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet title="XSL_formatting" type="text/xsl"href="/shared/bsp/xsl/rss/nolsol.xsl"?>

<rss xmlns:media="http://search.yahoo.com/mrss/"
    xmlns:atom="http://www.w3.org/2005/Atom" version="2.0">
  <channel>
    <title>BBC News - Business</title>
    <link>http://www.bbc.co.uk/news/business/#sa-ns_mchannel=
      rss&ns_source=PublicRSS20-sa</link>
    <description>The latest stories from the Business section
      of the BBC News web site.</description>
    <language>en-gb</language>
    <lastBuildDate>Fri, 21 Mar 2014 11:42:34 GMT</lastBuildDate>
    <copyright>Copyright: (C) British Broadcasting Corporation,
      see http://news.bbc.co.uk/2/hi/help/rss/4498287.stm
      for terms and conditions of reuse.</copyright>
    <image>
      <url>http://news.bbcimg.co.uk/nol/shared/img/bbc_news_120x60.gif</url>
      <title>BBC News - Business</title>
      <link>http://www.bbc.co.uk/news/business/
        #sa-ns_mchannel=rss&ns_source=PublicRSS20-sa</link>
      <width>120</width>
      <height>60</height>
    </image>
    <ttl>15</ttl>
    <atom:link href="http://feeds.bbci.co.uk/news/business/rss.xml"
      rel="self" type="application/rss+xml"/>
    <item>
      <title>Russian shares fall over sanctions</title>
      <description>Russian stock markets fell sharply on Friday
        as investors weighed the impact of western sanctions
```

```

over Ukraine.</description>
<link>http://www.bbc.co.uk/news/business-26678145#sa-
ns_mchannel=rss&ns_source=PublicRSS20-sa</link>
<guid isPermaLink="false">http://www.bbc.co.uk/
news/business-26678145</guid>
<pubDate>Fri, 21 Mar 2014 11:02:28 GMT</pubDate>
<media:thumbnail width="66" height="49" url=
"http://news.bbcimq.co.uk/media/imaqes/73725000/jpq/
_73725547_putin.yakunin.r.jpg"/>
<media:thumbnail width="144" height="81" url=
"http://news.bbcimq.co.uk/media/imaqes/73725000/jpq/
_73725548_putin.yakunin.r.jpg"/>
</item>
</channel>
</rss>

```

Az RSS-állomány az XML-dokumentumok általános felépítési sémáját követi, az RSS leírásában definiált jelölésekkel, ún. tag-ekkel jelölik magát a hírt, illetve a hozzá kapcsolódó metaadatokat, amelyek a tartalom értelmezését könnyít(het)ik meg. Az állomány első része magát a hírszatornát írja le, a második része pedig egy hírbejegyzést kódol. A hír egy rövid, tömör címmel rendelkezik, ami a <title> tag-ek között olvasható, majd egy rövid, de a címtől bővebb leírás olvasható róla a <description> tag-ek között. Metaadatként megjelenik a hír publikálásának dátuma, valamint a link, amelyen a teljes cikk található[3].

2.2. FOREX kereskedés

2.2.1. A devizakereskedésről általánosan

A nemzetközi devizapiac[4] (angolul: *foreign exchange market*, FOREX) egy globális, decentralizált piac, mely a devizákkal való kereskedést hivatott szolgálni. Ezen piac legfőbb résztvevői a nagyobb, nemzetközi bankok, továbbá multinacionális vállalatok, kormányzatok, azonban nem elhanyagolható a másik tábor szerepe sem, azaz a magánszemélyek rendkívül nagy csoportja. A világ számos táján funkcionálnak pénzügyi központok közvetítőként különböző típusú vevők és eladók széles skálája között (a hétvégéket leszámítva), így a devizatőzsdék is, amelyek lehetővé teszik ezen csoportok számára a nagymértékű, gyors és modern kereskedést. Ezeknek a devizapiacoknak a működése határozza meg a különböző devizák relatív árát (azaz a többi devizában kifejezett értékét). Jelentősége miatt kiemelhetünk két fontos, bankközi FX kereskedő platformot: az EBS-t (*Electronic Broking Services* - Elektronikus Bróker Szolgáltatások) és a Reuters 3000 Xtra-t.

A devizapiacok különböző pénzügyi intézetek segítségével, több szinten működnek. A színpalak mögött a legnagyobb kereskedési tételek általában abból adódnak, hogy a bankok olyan pénzügyi cégekhez fordulnak, akiket úgymond „dealer”-eknek, azaz elosztóknak hívunk, ugyanis ők aktívan részt vesznek a devizakereskedelemben. Sok esetben ezek a pénzügyi intézetek szintén bankok, ezért néha ezt a „színpalak mögötti” piacot bankközi piacnak is nevezzük, habár egyéb, például biztosító, vagy más funkciót betöltő társaságok is szerepet kapnak. Az ilyen, fontosabb elosztó szerepet betöltő intézetek között a devizakereskedés hatalmas méreteket ölthet, akár több száz millió dolláros pénzmozgások is előfordulnak.

A devizapiac a nemzetközi kereskedelmet és beruházást élénkíti és motiválja azáltal, hogy megvalósítja a devizák közötti átváltást, konverziót.

A modern devizapiac kialakulása az 1970-es években kezdődött meg, a devizapiaci tranzakciókra vonatkozó kormányzati megszorításokat követően 3 évtizeddel (a II. világháború után a Bretton Woods-i

pénzügyi rendszer stabilizálta a kereskedelmi és finanszírozási szabályokat a legjelentősebb iparral rendelkező országok között), amikor is az országok fokozatosan átváltottak az ingadozó árfolyamrátaakra a megelőző átváltási rendszert követően (az említett Bretton Woods-i rendszerben az árfolyamokat egy meghatározott szinten rögzítették).

A devizapiacot több tekintetben is egyedinek tekinthetjük:

1. először is, a nagy kereskedési volumenek miatt ez a szféra reprezentálja a legnagyobb eszközosztályt (olyan gazdasági elemek csoportját, amelyek vagyoneszközként funkcionálnak) a világban, ami hozzájárul a leglikvidebb pénzeszközök magas és könnyed forgalmához;
2. a földrajzi szétszórtsága miatt;
3. a folyamatos működése miatt, a nap mind a 24 órájában - a kereskedés megállás nélkül zajlik Greenwichi középídő szerint vasárnap este 22:00 órától egészen pénteken este 22:00 óráig;
4. a rengeteg gazdasági tényező (mint például nemzetek monetáris és fiskális politikája, gazdasági helyzete, jelentős vállalatok részvényárfolyamai, különböző tőzsdék, országok háborús időszakai, stb.) miatt, amik képesek befolyásolni a különböző árfolyamokat;
5. a relatíve (más piacok fix, magasabb nagyságrendű jövedelmezőségéhez viszonyítva) alacsony nyereségek miatt.

A felsoroltakat tekintve a devizapiacok állnak a legközelebb a tökéletes piac elméletéhez (a jegybankok esetleges árfolyam-stabilizálási beavatkozásai ellenére).

A Nemzetközi Fizetések Bankja (*Bank for International Settlements* - BIS) a világ legrégebbi, jelenleg is működő nemzetközi pénzügyi szervezete. A nemzetközi bankok közötti monetáris együttműködés legfőbb szervezője, a központi bankok központi bankja.

Kimutatások (a BIS előzetes, globális eredményei a 2013-as *Triennial Central Bank Survey of Foreign Exchange*-ből) szerint a devizakereskedelem 2013. áprilisában napi szinten, átlagosan 5,3 trillió dollár körül mozgott, azaz a 2010. áprilisi 4 trillióról, illetve a 2007. áprilisi 3,3 trillióról ez elég erős emelkedést jelent, amely szintén a devizakereskedelem jelentőségét acélozza.

3. fejezet

Kapcsolódó modellek

Jelen fejezet célja, hogy nagyon röviden ismertesse a már létező pénzügyi és gazdasági modelleket, amelyek alkalmasak az adott, illetve hasonló területek vizsgálatára.

Mindenképpen érdemesnek tartjuk, hogy a dolgozatban felvázolt vizsgálati módszerünket bizonyos aspektusok szerint viszonyítsuk az ún. „Hatékony piacok elméletéhez” (*efficient-market hypothesis - EMH*), amely a pénzügytan legfontosabb hipotézisei közé tartozik.

Habár az alábbiakban ismertetett fogalmak, elméletek, illetve tesztek elsősorban nem a devizapiacokra irányulnak (hanem leginkább a tőkepiacokra), mégis, eredendően hasonló területeken helyezhetjük el a dolgozat témáját és a fejezet további szakaszainak tartalmát. Ezen leírásokhoz forrásként és iránymutatóként egyaránt, Brealey és Myers[5] alapművét használtuk.

3.1. Kendall és a bolyongó folyamatok

Maurice Kendall brit statisztikus 1953-ban egy, a részvények és árujóságok széles skálájának árfolyamairól, illetve azok viselkedéséről szóló, meglehetősen ellentmondásos cikket nyújtott be a *Royal Statistical Society*-hez[6]. Kendall visszatérő, szabályos árciklusokra számított, de meglepetésére azt találta, hogy ilyenek nem léteznek, mindegyik sorozat „bolyongónak” tűnt¹. Kendall állítása szerint a különböző részvényárfolyamok véletlenszerűen követik egymást, amelyen azt értette, hogy az árfolyamok változásai egymástól függetlenek.

Ha a Kendall által tapasztaltakat összefoglaljuk, akkor a következő pontokról beszélhetünk:

1. Egymáshoz meglehetősen közel lévő árfolyamértékek sorozatát figyelembe véve a sorozat egyes elemei közötti véletlen változások olyan nagyok, hogy megnehezítik a jelenre is kiható, szisztematikus hatások esetleges felismerését.
2. Az előzőekből következik, hogy statisztikai módszerek alkalmazásával nehéz különbséget tenni valódi, eredendően bolyongó és olyan folyamatok között, amelyekben a szisztematikus, vagyis rendszerszerű hatásokat kiváltó elemek gyengék.
3. Ameddig nem sikerül megkerülnünk az ilyen komplikációkat, addig bármilyen trend, vagy modell illesztése az adott folyamatokra igencsak kockázatos vállalkozás. Egy közgazdász részéről elképzel-

¹Kendall felfedezése nem egészen előzmények nélküli. 53 évvel korábban a francia származású Louis Bachelier már említette az adott témát doktori disszertációjában. Bachelier-nek a véletlen folyamatok matematikai elméletével kapcsolatos eredményei öt évvel előzték meg Einsteinnek az egymással ütköző gázmolekulák véletlen mozgásáról szóló munkáját.[7][8]

hető, hogy különböző tesztekkel megvizsgálja az adatok és az előzetes elemzések alapján felállított hipotézisek illeszkedését, azonban közel lehetetlen, hogy döntsön olyan hipotézisek között, amelyek nagy mértékben eltérnek egymástól, mégis mind illeszkednek az adatokra.

4. Tapasztalati bizonyítékok és elméleti hátterek is erősítik azt a hitet, miszerint az aggregált indexszámok² sokkal inkább rendszerszerűbben viselkednek, mint azok komponensei. Ezen jelenség lehetséges okai a véletlen hatások redukálódása az átlagolásból adódóan és az összetevők konzekvens felbukkanása; ugyanakkor lehet éppen a véletlen műve is az egész. Amennyiben ez így van, akkor az aggregált sorozatokban hamis időbeli korrelációk lesznek és az index-számok használata a közgazdasági elemzésekben széleskörű újragondolást igényel.
5. A részvényárfolyamokon végzett elemzések kapcsán kiderült, hogy a sorozatokon belüli és a sorozatok közötti korrelációk egyaránt alacsonyak. Hacsak néhány egyedi részvény nem viselkedik az átlagosaktól eltérően, nem sok remény marad a tőzsdei árfolyammozgások előrejelzésére rendkívüli információk nélkül.

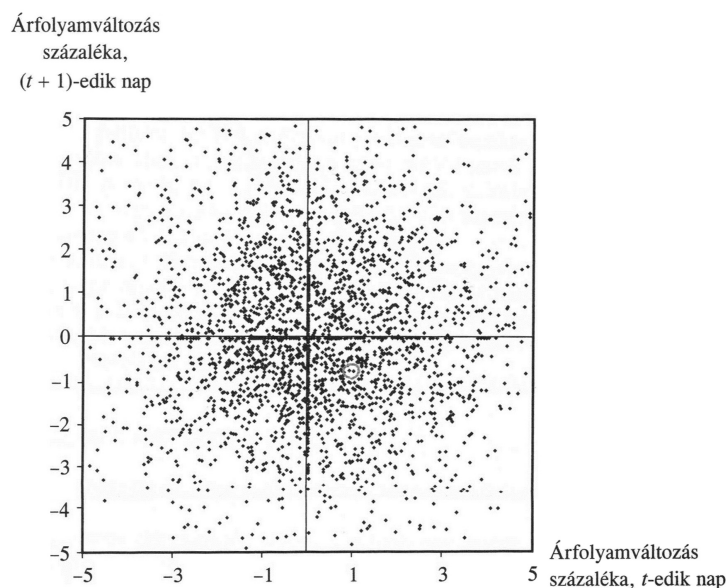
3.2. A Microsoft-részvények árfolyamának alakulása

Kendall gondolatainak szemléltetéséhez tekintsük a 3.1 ábrát. Mindegyik pont a Microsoft-részvények árfolyamának változását mutatja az egymást követő napokon. A grafikon délkeleti negyedében bekarikázott pontok egy olyan nap-párt jelölnek, amikor 1 százalékos növekedést 1 százalékos csökkenés követett. Ha a növekedést követő csökkenés szisztematikus lett volna, akkor sok pont lenne a negyedik síknegyedben és kevés az elsőben. Elég nyilvánvaló az ábra alapján is, hogy ilyen jellegzetességek nem igen léteznek, de ezt az állítást pontosabban is ellenőrizhetjük, ha kiszámítjuk minden egymást követő nap árfolyamváltozásainak korrelációs együtthatóját. Ha egy bizonyos típusú árfolyamváltozásból következtetni lehetne arra, hogy a következő napi változás is hasonló jellegű lesz, akkor a korrelációs együttható pozitív lenne. Ha az árfolyamváltozások között nincs összefüggés, akkor a korrelációs együttható értéke nulla. A fenti példában a korrelációs együttható értéke +0,022, vagyis elhanyagolható annak a valószínűsége, hogy egy adott napi árfolyam-növekedést a következő napon újabb árfolyam-növekedés követ.

Az eredményekből levonható következtetés szerint a Microsoft-részvény árának változásai gyakorlatilag korrelálatlanok, tehát a mai napi árfolyamváltozás semmilyen utalást nem ad arra nézve, hogy holnap milyen változás várható. Amennyiben ezt meglepőnek találjuk, gondoljunk arra az esetre, amikor nem ez lenne a helyzet, és a Microsoft-részvények árában bekövetkező változások előrejelzéseink szerint több hónapig tartósan fennmaradnának. Mi történne ilyenkor, ha a befektetők észrevennék ezt a profitszerzési lehetőséget? A felismerés lerombolná saját magát, önpusztító hatású lenne. Mivel a Microsoft-részvény jó befektetési lehetőséggel kecsegtet, a befektetők rohanni fognak, hogy vásárolhassanak belőle. Csak akkor fejeződnenek be a vásárlások, ha a részvény már csak a normális hozamot kínálja. Vagyis amint a befektetők felismerik a ciklust, a vásárlásukkal rögtön meg is szüntetik azt.

Vegyük észre, hogy mind Kendall kutatásai, mind az előzőekben bemutatott Microsoft-részvényekkel kapcsolatos példa, kizárólag magukból az idősorokból próbálták meg előrejelezni az árfolyamokat. A következő alfejezetben láthatjuk, hogy ez, tehát az idősorokon kívüli információk figyelmen kívül hagyása, a problémához való hozzáállásnak csak az első fokozata. Saját kutatásainkat már bonyolultabb és érde-

² Aggregált indexszámok alatt olyan mutatószámokat értünk, amelyeket valamilyen szempont alapján összetartozó, de különböző mutatószámokból származtatunk, például: egy egész iparágra jellemző jövedelmezőségi mutató, vagy a piaci portfólió várható hozama, stb..



3.1. ábra. Az ábra pontjai a Microsoft-részvények két egymást követő napi hozamait mutatják 1990. március és 2001. július között. A bekarikázott pontok például annak felelnek meg, hogy valamely napon 1 százalékkal nőtt a részvény ára, a következő napon viszont 1 százalékos csökkenés következett be. A pontdiagram nem jelez szignifikáns kapcsolatot az egymást követő napok árfolyamváltozásai között.

kesebb alapokra helyeztük, így nem csupán idősorok segítségével, hanem egy azoktól teljesen független forrásból származó információhalmazzal próbáljuk megérteni az árfolyammozgások miertjét.

3.3. Eugene Fama munkássága

Az előző alfejezetben bemutatottak alapján láthattuk, hogy a versenyző piacon az árfolyamoknak miért kell véletlen bolyongást követniük. Ha a múltbeli árváltozások alapján meg lehetne jósolni a jövőbeli árváltozásokat, a befektetők könnyedén, és nagy profitra tehetnének szert. A versenyző piacokon azonban ilyen lehetőségek nem, vagy csak igen ritkán adódnak, és ez utóbbi esetben is elég gyorsnak kell lenni, hogy élhessünk velük. Amint a befektetők, kereskedők megpróbálnak a múltbeli árakban rejtőző információk alapján előnyt szerezni, az árfolyamok is rögtön megváltoznak, egészen addig a pontig, ahol már nem lehet a múltbeli áralakulások vizsgálata révén extraprofithoz jutni. Ennek eredményeképpen, a múltbeli árakban megtalálható összes információ már a mai napi részvényárfolyamban tükröződni fog, úgy, hogy egy adott időszak árváltozásai függetlenek maradnak az előző időszak árváltozásaitól. Röviden, az árfolyamok véletlen bolyongást fognak követni.

A versenyző piacokon lényegében a részvények mai árának tükröznie kell a múltbeli árakban található információkat. Azonban, ha egy szinttel magasabbra lépünk, feltehetjük a kérdést, miszerint nem kellene-e a mai részvényárfolyamoknak a versenyző piacokon a befektetők rendelkezésére álló összes információt tükröznie? Ha az előbbi kérdésre a válasz igen, akkor a részvények ára éppen helyes lenne, az egyes hozamok pedig előrejelezhetetlenek lennének, bármilyen információt alapul véve is.

Eugene Fama, a hatékony piacokról alkotott művében[9], a piaci hatékonyságnak három nevezetes szintjét emelte ki, melyek, attól függően, hogy mennyi információ tükröződik a részvények árfolyama-

iban, rendre a következők: az első szinten az árfolyamok tartalmazzák a múltbéli árakban lévő összes információt. A közgazdaságtanban ezt nevezik a piaci hatékonyság gyenge (*weak*) formájának. Ha a piacok gyengén hatékonyak, akkor nem lehet tartósan extraprofitot elérni a múltbéli hozamok tanulmányozása révén. Az árak véletlen bolyongást fognak követni.

A hatékonyság második szintje azt követeli meg, hogy az árfolyamok ne csak az elmúlt időszaki árfolyamokat, hanem minden más közzétett, publikus információt is tükrözzenek; például azokat, amelyeknek a pénzügyi újságok, folyóiratok olvasása során juthatunk a birtokába. Ez a szint a piaci hatékonyság közepes (*semistrong*) formájaként ismert. Ha a piacok ebben az értelemben hatékonyak, akkor az árak azonnal reagálni fognak az olyan nyilvánosan közzétett információkra, mint például az utolsó negyedéves nyereséggel kapcsolatos bejelentések, új részvények kibocsátása, két vállalat összeolvadási terve, részvényfelaprózások, és hasonló.

Utolsó sorban felvázolhatjuk a hatékonyság erős (*strong*) formáját is; itt az árak már minden olyan információt tükröznek, amelyre a vállalat és az egész gazdaság alapos elemzése révén szert tehetünk. Továbbá, az erős formához szokott társulni a hatékonyság azon értelmezése is, amely, Fama-t idézve, az olyan befektetők, vagy csoportok feltételezését is magába foglalja, akik úgymond „monopolisztikus” hozzáféréssel rendelkeznek bizonyos információkhoz (lényegében bennfentesek).

3.3.1. Fama tapasztalatainak összefoglalása

Általános megfogalmazás[9] szerint a hatékony piacok elmélete lényegében azt a kérdést teszi fel, hogy tetszőleges időpontban az árak teljes mértékben tükrözik-e a rendelkezésre álló információkat? Fama elsőként a tőkepiaci eszközök elvárt hozamán alapuló elméleteket ismerteti és részletezi, majd újabb modelleket (martingál és véletlen bolyongás, illetve utóbbi kapcsán a fejezet elején bemutatott, Kendall-féle eredményekre is visszautal) származtat az elvárt hozamra építő feltételezésekből. Hely hiányában ezen szakaszok kifejtésétől most eltekintünk.

A cikkben leírtak összefoglalják, hogy az összes elérhető szakirodalom a témában implicit, vagy explicit módon, de azokon a feltételezéseken alapszik, hogy a piaci egyensúly kifejezhető az elvárt hozamok segítségével. Ez az alapja az úgynevezett elvárt hozamokra épülő, vagy más néven „korrekt játék” jellegű hatékony piaci modelleknek.

A fentiekben már bemutattuk a hatékonyság három szintjét. A gyenge formára vonatkozó tesztek a legterjedelmesebbek, és kijelenthetjük, hogy az ilyen jellegű tesztek eredményei meglehetősen támogatják a hipotéziseket. Habár találhattunk néhány, statisztikailag szignifikáns bizonyítékot arra nézve, hogy a folyamatosan következő árfolyam-, illetve hozamértékek függenek egymástól, ezek egy része konzisztens a „korrekt játék” modellekkel, a maradék pedig nem tűnik elegendőnek ahhoz, hogy definiáljuk a nem-hatékony piacok fogalmát. Továbbá, az egy napos, vagy hosszabb időintervallumokra kivetített ár- és hozamváltozások sem jelentenek elég bizonyítékot a „korrekt játékok” még „törekvőbb” utódjára, a véletlen bolyongásra.

Konzisztens bizonyítékok támasztják alá az átlagos részvényárfolyamok, illetve átlagos részvényhozamok napi változásainak pozitív függőségét, amely függőség felhasználható a korlátozott mértékben nyereséget ígérő kereskedési szabályok alapjaként. A Fama által korábban publikált adatokban[10] a függőség egy sorozatos korreláció formájában lelhető fel, amely konzisztensen pozitív értékeket jelent, azonban ezek egyúttal alacsonyak, nullához közeli; továbbá, csekély tendenciát mutattak a pozitív és negatív előjelű változások futamai egy véletlen bolyongású folyamatból elvárthoz képest. Ami még fontosabb, hogy a függőség Alexander[11] szűrő jellegű tesztjeiben is felbukkant, illetve Fama és Blume[12] is tapaszt-

talta a jelenséget a „megvásárol-és-megtart” jellegű, többletprofitot eredményező kereskedési stratégiák kapcsán (nagyon kis szűrőket alkalmazva). Azonban, bármilyen rendszerről beszéljünk is, amelyik ezt a szemléletmódot (a rövidtávú függések profitná való konvertálását) kísérli meg alkalmazni a gyakorlatban, az szükségszerűen azt is magával vonzza, hogy az elvárt profitot még a minimális tranzakciós díjak is elnyelik, amelyet az egyszerű kereskedőknek kell fizetnie a tőzsdéken. Így, a hatékony piacok egy ilyen, nem teljesen szigorú értelmezésében ez a pozitív függőség nem elég jelentős a hipotézisek elvetése melletti döntéshez.

A „korrekt játék” jellegű hatékony piaci modellek cáfolásához (egyetlen napnál hosszabb időtávot alapul véve) szükséges bizonyítékokat még nehezebben találhatunk. Cootner[8] és Moore[13] döntően negatív jellegű, de szintén alacsony mértékű sorozaton belüli korrelációt riportált az átlagos részvények heti árfolyamainak, hozamainak alakulására nézve, amely eredmények Fama[10] négy napos hozamokra vonatkozó elemzéseiben is felbukkantak. A Fama által publikált eredményekben[10] lévő futamtesztek alapján nem tűnik úgy, hogy bármi jel utalna a pozitív függőségre, illetve semmilyen más függőségre alapján véve sem. Lényegében, bármilyen autokorrelációt véltek is felfedezni a heti hozamok között, azok nem igazán használhatóak profitképes kereskedési szabályok kialakítására.

A piaci hatékonyság közepes formájára irányuló tesztek, amelyek kapcsán feltételezzük, hogy az összes nyilvánvalóan elérhető, publikált információ tökéletesen tükröződik az árfolyamokban, szintén alátámasztják a hatékony piaci hipotéziseket. Ha például Fama, Fisher, Jensen és Roll[14] munkáját tekintjük, észrevehetjük, hogy a részvényfelaprózásokkal kapcsolatos bejelentések, mint új információk a befektetők számára, azonnal beolvadnak az árfolyamokba, így azok teljes mértékben tükrözik ezeket az információkat (ebből következően a jövőbeli osztalékfizetésekre, és hasonló műveletekre is kihat).

A hatékonyság erős szintjének vizsgálatakor Niederhoffer és Osborne[15] kimutatta, hogy fontosabb tőzsdei értékpapírok (és egyéb termékek) esetén néhány specialistának egyfajta „monopolisztikus” hozzáférése van bizonyos információkhoz (származzon ez az adottság akár bennfentes létükből, akár mélyrehatóbb elemzésekből), és ezáltal a hatékonyság úgymond veszít erejéből. Végezetül, Scholes[16] is tapasztalta, hogy nem meglepő módon a bennfentesek saját cégükről többletinformációval rendelkeznek, ami ismét kisebb hatékonyságot eredményez.

3.3.2. A hatékony tőkepiacokról - másodszor

Eugene Fama 1991. decemberében megjelent tanulmányában[17] több, mint 20 évnyi tapasztalat után a hatékony tőkepiacokról újabb, részletes áttekintést adott, melyben a megelőző, fentiekben kifejtett írása óta felhalmozódott, hatalmas terjedelmű szakirodalmat is felhasználta. Fama a piaci hatékonyság három szintjét bizonyos mértékben módosította, egészen konkrétan a gyenge forma alatt egy általánosabb területet ért, tehát most már egyéb változók, mint például az osztalékhozamok és kamatráták, segítségével előrejelzett hozamokon alapuló munkákat is figyelembe vesz; a hatékonyság másik két szintje esetén túl nagy változások nem történtek, inkább csak azok címét módosította: a közepes forma felel meg ezentúl az úgynevezett esettanulmányoknak (*event studies*), illetve a hatékonyság erős formáját pedig a privát információkra irányuló teszteknek nevezi (*tests for private information*). Ahogyan Fama is megemlíti művében, a kapcsolódó szakirodalom végtelenül terjedelmessé vált az elmúlt (számára két) négy évtizedben, így jelen dolgozatban sajnos nem tudunk kitérni eme érdekes részletekre és modellekre.

3.4. A piaci hatékonyságra irányuló további vizsgálatok

A fent bemutatott, Microsoft-részvényekhez hasonló vizsgálatokat a különböző tőzsdeindexeken is elvégezhetjük, még hozzá teljesen analóg módon, azzal a minimális eltéréssel, hogy jelen esetben nem az egymást követő napi, hanem az egymást követő heti árakat tekintjük a vizsgálat alapjául. Ezen tesztek eredményeire nézve elmondhatnánk, hogy a heti hozamok alakulásában szerte a világon csak nagyon kevés alakzatra bukkanhatunk.

A piaci hatékonyság közepes formájának elemzéséhez a kutatók azt is megvizsgálták, hogy milyen gyorsan reagálnak a részvényárfolyamok az olyan különböző hírekre, mint például a vállalatok nyereségével, vagy osztalékával kapcsolatos bejelentések, a felvásárlásokról szóló hírek, vagy a makrogazdasági adatok közzététele. Ezen ponton érdemes megemlíteni, hogy jelen dolgozat lefedi ezen vizsgálni kívánt lehetőségeket is a devizaárfolyamokban tapasztalható ingadozásokra nézve.

A legtöbb vállalatfelvásárlás esetében a vevő a célvállalat aktuális piaci árán felül jókora prémium megfizetésére is hajlandó; ebből következően amikor egy vállalat felvásárlási kísérlet célpontjává válik, részvényeinek árfolyama a várható felvásárlási prémium következtében megemelkedik. Keown és Pinkerton[18] tanulmányából és eredményeiből láthatjuk, hogy amikor a nyilvánosság tudomást szerez a felvásárlási kísérletről, az átlagos célvállalat részvényeinek árfolyama jelentősen megugrik. A részvényárfolyam igazodása azonnali: a bejelentés napján bekövetkezett nagy ugrást követően az árváltozási várakozás megszűnik, és a továbbiakban nem figyelhető meg sem felfelé, sem pedig lefelé irányuló szisztematikus elmozdulás (*drift*) az árfolyamban. A tanulmány azt is kimutatta, hogy a felvásárlási bejelentést megelőző napokon a részvény árfolyama folyamatosan növekszik, ami a felvásárlásról szóló hírek kiszivárgására utal (gondoljunk csak a hatékonyság erős formájára, illetve az információkhoz való „monopolisztikus” hozzáférésre). Néhány befektető már a bejelentést megelőzően elkezd vásárolni a célvállalat részvényeit. A hatékony piacok elméletének megfelelően, amint az információ publikussá válik, ez azonnal és teljes mértékben tükröződik az árfolyamokban.

Patell és Wolfson egy tanulmányából[19] az is kiderül, hogy mennyire gyorsan reagálnak a részvényárfolyamok az új információk hatására. Azt találták, hogy amikor egy vállalat közlésezi esedékes gyorsjelentését, vagy osztalékpolitikai változást jelent be, az árfolyam változásának legnagyobb része a bejelentést követő 5-10 percen megy végbe.

A hatékonyság erős formájának tesztelése során megvizsgálták különböző hivatásos értékpapír-piaci elemzők javaslatait, valamint olyan befektetési, illetve nyugdíjalapokat kerestek, amelyek előreláthatóan képesek a piac teljesítményét felülmúlni. Néhány kutató talált ugyan csekély mértékű, tartós teljesítménybeli többletet bizonyos alapoknál; legalább ugyanennyien viszont arra a következtetésre jutottak, hogy a professzionálisan irányított befektetési alapok nem hozzák be a menedzsment költségeit. Mark Carhart[20] tanulmánya nagyjából 1500 amerikai befektetési alap átlagos hozamát elemzi. A kimutatásban észrevehetünk éveket, amikor a befektetési alapok felülmúlták a piacot, azonban legalább ilyen gyakran fordul elő a másik eset is. A Carhart által alkalmazott módszer csak elég durva összehasonlításra alkalmas, mivel a befektetési alapok hajlamosak különböző piaci szektorokra specializálódni, azaz csak alacsony kockázatú, vagy csak nagyvállalati részvényekkel foglalkoznak; ezek hozama pedig lehetett éppen a piaci hozam alatt. Az ilyen különbségek elkerülésére minden alapot egy olyan portfólióval kell összehasonlítani, amelyet hasonló értékpapírokból állítottak össze. Carhart tanulmánya ugyan kiterjedt erre is, a következtetés azonban változatlan maradt: az alapok a költségek levonása előtt nagyjából ugyanakkora, a költségek levonása után viszont alacsonyabb hozamot biztosítottak, mint a hasonló értékpapírokból összeállított referenciaportfóliók.

A nyereségbejelentési reakciók vizsgálatára épít Bernard és Thomas[21] kutatása, amely kapcsán elemzéseikkel (a részvényárfolyamoknak a meglepetésszerűen jó, vagy rossz nyereségi, 1974 és 1986 közötti években érvényes adatok bejelentését követő teljesítménye alapján) arra a következtetésre jutottak, hogy a vállalatok legjobb jövedelmezőségi hírekkel rendelkező felső 10 százalékának részvényei több, mint 4 százalékponttal múlták felül a legrosszabb vállalatokét a bejelentést követő két hónapban. Ezek alapján úgy tűnik, hogy a befektetők alul reagálják a nyereség-bejelentéseket, és azok jelentőségét csak a későbbi információk fényében tudják megfelelően értékelni.

3.5. Összefoglalás

A fejezetben bemutatottak alapján láthattuk, hogy a hatékony piaci hipotézisek rendkívül tág témakörrel rendelkeznek, az elmúlt fél évszázadban igen terjedelmes szakirodalom halmozódott fel ezen területeken. Sajnálatos módon, a dolgozat keretei miatt csak röviden tudtuk ezen munkáknak egy töredékét megemlíteni.

Jelen tudományos munka szemszögéből tekintve a legfontosabb tanulság a már létező szakirodalomra nézve, hogy amennyiben ragaszkodunk a piaci hatékonyság Fama által definiált három fokozatához, úgy az RSS hírek klaszteranalízise, majd az ezt követő kapcsolatok feltárása az egyes hírcsoportok és a devizaárfolyamok ingadozásai között leginkább a második szintnek felel meg. Lényegében, Fama szavaival élve, egyfajta esettanulmányokat folytatunk arra nézve, hogy a publikusan hozzáférhető és a modern kor eszközeivel közzétett információk hogyan és milyen gyorsasággal hatnak a gazdasági folyamatokra.

A következő fejezetben a vizsgálatok első lépését, a hírek klaszterekbe sorolásának folyamatát és ezen módszerek technikai hátterét, az azt követő fejezetben pedig az alkalmazott matematikai eljárásokat mutatjuk be, amelyeket a kapcsolatok megvilágítása során használtunk. Egyrészt alkalmazzuk a folytonos hozamszámítással kapcsolatos ismereteket, amelyekre a korai vizsgálatok is építettek (lásd a Fama által említett, elvárt hozamon alapuló modellekről szóló szakaszt, és a hozzá kötődő szakirodalmat), illetve maradunk, a dolgozat korábbi formájában taglalt lineáris regressziószámításnál.

4. fejezet

Az alkalmazott módszerek

Jelen fejezetben az adatbázisunkban szereplő hírek csoportosítása, valamint az előálló csoportokhoz kapcsolódó árfolyam szekvenciák elemzése során alkalmazott módszereket ismertetjük.

4.1. Hierarchikus, agglomeratív klaszterezés

Általánosan fogalmazva egy klaszterező algoritmus célja az, hogy bizonyos, nem feltétlenül matematikai objektumokat úgy soroljon csoportokba, hogy az egy csoporton belüli elemek valamilyen tekintetben hasonlóbba legyenek egymáshoz, mint a más csoportokban lévő elemekhez. Számos klaszterező eljárással találkozhatunk az irodalomban, és ezek közül kell kiválasztanunk azt, amely hírek csoportosítására a legmegfelelőbb. A talán legelterjedtebbnek nevezhető módszer a k -közép klaszterezés, mely a klasztereket, csoportokat azok középpontjával azonosítja. A k -közép klaszterezés alkalmazása során elengedhetetlen, hogy az egy csoportba tartozó elemek számtani, esetleg súlyozott átlagát számítsuk. Ez hírek esetén nyilvánvalóan nem kivitelezhető, hiszen hírek összeadását, hírek átlagát nem tudjuk értelmezni. A k -medián klaszterezés során elkerülhetjük a hírekkel végzett algebrai műveleteket, azonban a k -közép klaszterezéshez hasonlóan ez a módszer is paraméterként várja a csoportok becsült számát, a k -t. Mivel ilyen becsléssel nem rendelkezünk és a hírek nagy száma miatt sok k -ra végrehajtani a klaszterezést igen számításigényes, a k -medián klaszterezést sem használhatjuk hatékonyan. Más megközelítés a sűrűségbecslés, nevezetesen az adott adathalmazra valamilyen valószínűségi sűrűségfüggvény keverék illesztése. Ezen megközelítés legjellemzőbb példája a *Gaussian Mixture Model* illesztése, ahol az adathalmaz eloszlását súlyozott, szükség esetén többdimenziós Gauss-sűrűségfüggvények keverékével próbáljuk modellezni. Az említett megközelítésnél szükséges, hogy egy adott pontban ki tudjuk értékelni az illesztett sűrűségfüggvényeket, ami hírek esetén megint nehezen értelmezhető. Választásunk ennek megfelelően a szintén igen elterjedt hierarchikus klaszterezésre esett, ahol egy, a hírek között értelmezett különbözőségi függvényt (*dissimilarity function*) és egy küszöbszámot kell csak specifikálnunk, mely az egy klaszterbe eső elemek megengedett legnagyobb különbözőségéhez kapcsolódik a megadott különbözőségi függvény tekintetében.

Az egyszerűség kedvéért feltehetjük, hogy a hírek az angol abc kisbetűiből és szóköz karakterekből állnak, azaz az s és t sztringeket formálisan $s, t \in W = \cup_{n=0}^{\infty} \Sigma^n$ módon jelölhetjük, ahol Σ egy 27 elemű abc szimbólumait tartalmazza. Egy $d : W \times W \rightarrow \mathbb{R}$ különbözőségi függvénytől az alábbi tulajdonságok teljesülését várjuk el. Bármely $s, t \in W$ esetén

1. $d(s, t) \geq 0$,

2. $d(s, s) = 0, \forall s \in W,$
3. $d(s, t) = d(t, s), \forall s, t \in W.$

Megjegyezzük, hogy a különbözőségi függvény tulajdonságai hasonlóak a *metrika* tulajdonságaihoz, azonban a háromszög egyenlőtlenség, valamint a (2)-es tulajdonságnál a szigorú $d(s, t) = 0 \leftrightarrow s = t$ előírás teljesülése nem szükséges.

Az általunk alkalmazott hierarchikus, agglomeratív klaszterezés algoritmusát az alábbi pontokban foglaljuk össze:

-
1. Bemenet: hírek halmaza $H \subset W$; ϵ küszöbérték; $d(s, t) : W \times W \rightarrow \mathbb{R}$ különbözőségi függvény.
 2. Kimenet: csoportosított hírek: $\mathcal{C} \subset 2^H$.
 3. Inicializálásként minden hír egy különálló klaszterhez tartozik, azaz $\mathcal{C} = \{\{h\} | h \in H\}$.
 4. Megkeressük azon két klasztert, amelyre a klaszterek között értelmezett

$$e(C_i, C_j) = \frac{1}{|C_i| \cdot |C_j|} d(c_i, c_j), \quad (4.1)$$

különbözőségi függvény értéke minimális, ahol $c_i \in C_i, c_j \in C_j, C_i, C_j \in \mathcal{C}$, azaz

$$\{min_i, min_j\} = argmin_{i,j,i \neq j} e(C_i, C_j), \quad (4.2)$$

ahol $C_i, C_j \in \mathcal{C}$. Könnyen látható, hogy a klaszterek között értelmezett $e : \mathcal{C} \times \mathcal{C} \rightarrow \mathbb{R}$ különbözőségi függvény két klaszter elemeinek átlagos különbözőségét adja.

5. Ha $e(C_{min_i}, C_{min_j}) > \epsilon$, akkor a klaszterezés befejeződött, nincsenek már összevonható klaszterek, a legközelebbi klaszterek különbözősége nagyobb, mint ϵ . Az előállt klaszterek a $\mathcal{C}$ halmaz elemei. Ellenkező esetben legyen $\mathcal{C} = (\mathcal{C} / \{C_{min_i}\} \cup \mathcal{C} / \{C_{min_j}\}) \cup \{C_{min_i} \cup C_{min_j}\}$, és térjünk vissza a negyedik pontra.
-

A hierarchikus, agglomeratív klaszterezés működése könnyen értelmezhető. Nem definiáltuk még azonban a különbözőségi függvényt, amelyet a klaszterezés során használni fogunk.

4.2. A használt különbözőségi függvények

Célunk, hogy híreket csoportosítsunk, s ehhez az informatika nyelvén sztringek, azaz egy véges abc szavai között kell egy olyan függvényt találnunk, amely ezek különbözőségét méri. Nyilvánvalóan a távolság függvények, azaz metrikák teljesítik a különbözőségi függvény tulajdonságait, így elsőként tekinthetjük az elterjedt sztring távolság függvényeket, például a szerkesztési vagy Levenshtein-távolságot, amely két sztring távolságát az azokon alkalmazott „csere”, „beszúrás” és „törlés” műveletek legkevesebb számaként definiálja, melyekkel egyik sztring a másikká alakítható. Bár a Levenshtein-távolság függvény igen elterjedt sztringek hasonlóságának mérésére, hierarchikus klaszterezésre történő alkalmazása két szempontból jelenthet problémát: egyrészt értékkészlete az $\mathbb{R} \cup 0$ halmaz, így a klaszterezéshez használt $\epsilon \in \mathbb{R}$ megválasztása igen nehéz feladat; másrészt a hírek hossza nagyban befolyásolja azok távolságát. Célszerű lenne egy olyan különbözőségi függvényt választani, amely abszolút abban az értelemben, hogy értékkészlete egy korlátos $[a, b]$ tartományba esik, és értékét kevésbé befolyásolja az összehasonlított sztringek hossza.

Sztringek hasonlóságának mérésére széleskörűen alkalmazott úgynevezett sztring kernelek léteznek. Ezen sztring kernelek $k(s, t) : W \times W \rightarrow \mathbb{R} \cup 0$ típusú függvények, amelyek működésüket tekintve le-

írhatók $k(s, t) = \langle \phi(s), \phi(t) \rangle$ módon, ahol a $\phi(s) : W \rightarrow \mathbb{R}^N$ egy sokdimenziós sajátstígtérbe képező függvény és N jellemzően nagy szám, akár végtelen is lehet. A kernel függvények tehát a belsőszorzat egyfajta általánosításának tekinthetők, melyekben a tényleges belsőszorzat kiszámítását egy nagydimenziós térbe történő transzformáció előzi meg. Ideális esetben a $k(s, t)$ függvény értéke kiszámítható anélkül, hogy a $\phi(s)$ függvényt ismernénk. A kernel-függvények használatának nagyszerűsége abban rejlik, hogy azok a belső szorzat tulajdonságaival rendelkezve bárhol felhasználhatóak egy matematikai módszerben, ahol belsőszorzat-számítás történik, kibővítve ezzel az algoritmus lehetőségeit. Alkalmazásuk alapvető a statisztikai tanulóalgoritmusok területén, ahol lineáris osztályozó módszerekből nem-lineáris osztályozó kapható az algoritmusban szereplő belsőszorzatok megfelelő kernel függvényre történő cseréjével.

Célunk egy különbözőségi függvény konstruálása, mely abszolút abban az értelemben, hogy egy korlátos intervallumba képez. Ha egyszerű $x, y \in \mathbb{R}^N$ számvektorokat szeretnénk klaszterezni, felhasználhatjuk a korrelációból származtatott

$$d(x, y) = \frac{1}{2} \left(1 - \frac{\langle x, y \rangle}{\sqrt{\langle x, x \rangle \langle y, y \rangle}} \right), \quad (4.3)$$

függvényt, mely éppen a kívánt tulajdonságot teljesíti. Analóg módon definiálhatunk számunkra kedvező tulajdonságokkal rendelkező különbözőségi függvényt sztringekre a sztringeken értelmezett kernelek felhasználásával:

$$d(s, t) = \frac{1}{2} \left(1 - \frac{k(s, t)}{\sqrt{k(s, s)k(t, t)}} \right). \quad (4.4)$$

A különbözőségi függvény tehát adott, már csak a sztringek kezelésére alkalmas $k(s, t)$ kernel-függvényeket kell meghatároznunk. A dolgozat hátralévő részében két különböző kernelt használunk klaszterezésre, a spektrum és bináris spektrum kerneleket[22]. Ezeket részletesen a következő szakaszban mutatjuk be.

4.2.1. A spektrum és bináris spektrum kernelek

A spektrum-kernel a sztringeket alkotó l hosszúságú részsstringek gyakoriságának hasonlósága alapján állapítja meg a vizsgált sztringek hasonlóságát, illetve különbözőségét. Sajnos a kernel-trükk a számítások lerövidítése érdekében itt nem alkalmazható, így a kernel függvényt az abban megjelenő ϕ sajátstígtérbe történő transzformációval definiáljuk. Legyen $\phi_l(s) : W \rightarrow \mathbb{R}^{\Sigma^l}$, és

$$\phi_l(s) = (\psi_a(s))_{a \in \Sigma^l}, \quad (4.5)$$

ahol $\psi_a(s)$ azt adja meg, hogy az a -val jelölt l hosszúságú sztring hányszor fordul elő s -ben. Ezt követően a spektrum kernel

$$k_l^{\text{spektrum}}(s, t) = \langle \phi_l(s), \phi_l(t) \rangle \quad (4.6)$$

módon definiálható, s (4.4) alapján behelyettesítéssel kapható a $d_l^{\text{spektrum}}(s, t)$ különbözőségi függvény. Megjegyezzük, hogy a spektrum kernelnek van egy l paramétere, s különböző l értékek esetén különböző eredményeket kaphatunk. Természetesen a klaszterezés műveletigénye is függvénye l -nek, hiszen az angol abc kisbetűit és a szóköz karaktert használva $l = 2$ esetén csak 27^2 , azonban $l = 5$ esetén már 27^5 sajátstígtérre kell kiszámolni minden s sztringre a $\phi_l(s)$ függvény értékének meghatározásához.

A spektrum-kernel egy változatának tekinthető a bináris spektrum kernel, melynek leírása analóg a spektrum-kernelével, azzal a különbséggel, hogy a $\phi(s) : W \rightarrow \{0, 1\}^{\Sigma^l}$, és $\psi_a(s)$ azt adja meg, hogy az a -val jelölt l hosszúságú sztring előfordul-e a vizsgált s sztringben: értéke 1, ha előfordul, különben 0. A spektrum kernellel analóg módon (4.4) alapján bevezethetjük a $d_l^{\text{bináris}}(s, t)$ különbözőségi függvényt.

A különbség tehát a spektrum és bináris spektrum kernelek között az, hogy utóbbi az egy sztringben előforduló azonos l hosszúságú részsstringek előfordulásainak számát nem veszi figyelembe, csak az előfordulás tényét. Megjegyezzük, hogy a bináris spektrum kernel is paraméterezhető a vizsgált részsstringek hosszát specifikáló l paraméterrel, így különböző l értékek esetén különböző klaszterezési eredményekre számíthatunk.

Két hír tartalmi hasonlóságának méréséhez a hírek értelmezésére lenne szükség. A gazdasági hírek gazdasági folyamatokat befolyásoló pozitív vagy negatív hatásának eldöntése még a szakemberek számára is nehéz feladat. Ugyanakkor, a sztring-kernelek használatán alapuló osztályozás és klaszterezés jelenleg *state-of-the-art* technikának számít[23] az automatizált szövegfeldolgozás területén, szerteágazó alkalmazásokkal a protein-osztályozástól[24] a szöveg alapú adatbányászatig[25].

4.3. Regressziószámítás

Az árfolyamok időbeli változását ideális esetben egy $c(t) : \mathbb{R} \rightarrow \mathbb{R}$ típusú, időben értelmezett függvénnyel írhatjuk le. Jellemzően a $c(t)$ függvény nem áll rendelkezésünkre, annak csak egy mintavételezett változatával, a $c^T(m) : \mathbb{N} \rightarrow \mathbb{R}$ függvénnyel dolgozunk, ahol T az árfolyam mintavételezési ideje, és $c^T(m) = c(mT)$ teljesül. Például $T = 5s$ esetén az árfolyam értékét csak minden ötödik másodpercben ismerjük, így $c^5(1)$ az adatgyűjtés kezdetét követő ötödik másodpercben, $c^5(2)$ az adatgyűjtés kezdetét követő tizedik másodpercben adja meg a vizsgált árfolyamot.

Legyen a mintavételezés kezdete a T_0 időpillanat. Jelöljön C egy N elemszámú klasztert és $T^C = \{t_1, t_2, \dots, t_N\}$ a klaszterben szereplő hírek megjelenési időpontjainak halmazát. Hogy a hírek megjelenésének időpillanatait kompatibilissé tegyük a mintavételezett $c^T(m)$ függvénnyel, bevezetjük a megjelenések időpontjainak jellemzésére az $M^C = \{m_1, m_2, \dots, m_N\}$ halmazt, ahol $m_i = \left\lfloor \frac{t_i - T_0}{T} \right\rfloor$ jelöli azon árfolyamminta sorszámát, melyet követően a klaszter i -edik eleme hírként megjelent. Az időpontok kerekítése nyilvánvaló hibát eredményez. Ezen hibáktól két okból tekintünk el:

1. A tapasztalat azt mutatja, hogy az árfolyamok változását leíró függvény többnyire nem sima. Ennek megfelelően nem tudhatjuk, hogy az árfolyam az m_i és $m_{i\pm 1}$ időpontok között hogyan változik, így nem garantált, hogy egy esetleges interpoláció csökkenti a hibát.
2. Mivel a későbbiekben a kerekített időintervallumoktól nagyságrendekkel nagyobb időtartományokkal dolgozunk és az árfolyamfüggvények jellemzően korlátos változásúak, a kerekítés során megjelenő hibát elhanyagolhatónak tekinthetjük.

Vizsgálatunk alapját a C klaszter elmeinek megjelenését követő árfolyammozgások becslése képezi. Célunk tehát meghatározni, hogy valamely $m_i \in M^C$ időpontot követően kT , $k = 1, \dots$, idővel milyen irányba, illetve milyen mértékben mozdul az árfolyam. Az árfolyammozgások becslésére egyszerű lineáris regressziót alkalmazunk, azaz az árfolyamfüggvényt a vizsgált időpillanatban egy lineáris függvénnyel közelítjük, s az árfolyam változását az illesztett lineáris függvény meredekségével jellemezzük. Jelölje L a hírek megjelenését követően vizsgálni kívánt időpillanatok számát és D azok időbeli távolságát. Például $L = 25$ és $D = 10s$ választással a hír megjelenését követő $LD = 250s$ időtartamban vizsgáljuk az árfolyamváltozásokat.

A meredekség meghatározása során ki kell jelölnünk, hogy a lineáris közelítést milyen skálán szeretnénk alkalmazni: ha az egyenest sok mintapontra illesztjük, annak meredeksége a hosszútávú trendeket jellemzi, ha viszonylag kevés, egymást követő mintára, akkor a rövidtávú változásokat írhatjuk le vele. Jelölje K

a regressziószámítás során felhasznált mintapontok számát. Az általánosság megszorítása nélkül, az egyszerűbb formalizmus kedvéért feltéve, hogy K páratlan, meghatározzuk az $r^{m_i} \in \mathbb{R}^L$ vektort, ahol

$$r_j^{m_i} = \frac{\sum_{k=-K/2}^{K/2} \left(c^T(m_i + jD + k) - \frac{1}{K} \sum_{k=-K/2}^{K/2} c^T(m_i + jD + k) \right) k}{\sum_{k=-K/2}^{K/2} \left(c^T(m_i + jD + k) - \frac{1}{|K|} \sum_{k=-K/2}^{K/2} c^T(m_i + jD + k) \right)^2}, \quad (4.7)$$

ahol $j = 0, \dots, L-1$. A fenti eljárással egy C klaszter vizsgálatakor minden $m_i \in M^C$ hír megjelenéshez előállítunk egy r^{m_i} vektort, mely j -edik eleme a jD időpont körüli K mintapontra illesztett egyenes meredekségeként azt adja meg, hogy a hír megjelenését követő jD időpillanatban milyen irányba és milyen mértékben mozgult az árfolyam. (Természetesen az illeszkedések jóságát is megvizsgáltuk, méghozzá az ismert statisztikai eljárás, azaz a determinációs együttható meghatározásával, minden illesztett egyenes esetén.)

Feltevésünk szerint ha a hasonló hírek tartalma hatással van az árfolyamokra, akkor bizonyos C klaszterekhez tartozó r^{m_i} vektorok esetén lennie kell olyan időtartamnak, amely eltelte után az árfolyamok többnyire ugyanazon irányba változnak.

Ezt egyszerűen vizsgálhatjuk, ha az r^m , $m \in M^C$ vektorokat egy olyan $R^C \in [-1, 1]^{L \times |M^C|}$ mátrixba szervezzük, amely m -edik sora az r^m vektort tartalmazza. A mátrix sorai a C klaszterben lévő hírek előfordulásaihoz tartoznak, a mátrix j -edik oszlopából pedig azt olvashatjuk ki, hogy hogyan változtak az árfolyamok a hírek megjelenését követő jD időpillanatban.

Ha egy R^C mátrix j -edik oszlopa sok azonos előjelű értéket tartalmaz, az arra utal, hogy a C klaszterhez tartozó hírek megjelenését követő jD időpillanatban sokszor mozdul ugyanazon irányba az árfolyam. Az értékek nagyságából az árfolyamváltozások nagyságának jellegére vonatkozó információkat nyerhetünk ki.

Az egyszerűség kedvéért modellünkben csak az árfolyamváltozások irányával foglalkozunk, azaz az R^C mátrix elemeinek előjeleit vizsgáljuk. Megjegyezzük, hogy ezen megközelítést több irodalmi forrás is alkalmazza az árfolyamváltozások korreláltságának vizsgálatára (a 3. fejezetben már bemutattuk, hogy Fama munkájában[9] is megtalálhatóak hasonló eljárások, amelyeket Niederhoffer és Osborne[15] tanulmányaiból meríti leginkább). A különböző C klaszterek kvantitatív jellemzése érdekében bevezetünk egy, az R^C mátrixból származó mérőszámot, az ún. *egyezési rátát*. Legyen a $q^C \in \mathbb{R}^{|M^C|}$ vektor az R^C mátrix azon oszlopa, amelyben a legtöbb egyező előjelű érték található és legyen $\#q^C$ az egyező előjelű értékek száma a q^C vektorban. Ekkor a $p^C \in \mathbb{R}$ egyezési rátát $p^C = \frac{\#q^C}{|M^C|}$ módon definiáljuk. Könnyen látható, hogy minden esetben $p^C \in [0.5, 1]$ teljesül, és $p^C = 1$ esetén a C klaszterbe sorolt hírek megjelenését követően a vizsgált időpillanatok közül legalább egy esetében az árfolyamok mindig ugyanazon irányba mozdulnak.

A p^C egyezési rátát felhasználhatjuk arra, hogy a C klaszterben szereplő hírek árfolyamokra gyakorolt hatását jellemezzük.

A bevezetett p^C értékek felhasználásával igen sok szempont alapján vizsgálhatjuk az összegyűjtött adathalmazt, s a vizsgálatot számos szubjektív aspektus figyelembevételével árnyalhatjuk. Ezekről részletesen az eredményeket bemutató fejezetben számolunk be.

Előrevetítjük ugyanakkor, hogy a bemutatott p^C mérőszám a hosszútávú folyamatok, trendek hatásként jelentkező hibákkal terhelt. Ezen hibákat két módon csökkenthetjük:

1. A modellezést pontosítva, további becsléseket végezhetünk a hosszútávú trendek mértékére vonatkozóan, s az általunk vizsgált rövidtávú trendeket korrigálva, a hosszútávú trendek figyelembevételével alakítjuk ki a p^C mérőszámot. Ezen megközelítéssel a későbbiekben kívánunk foglalkozni.
2. A másik lehetőség, hogy statisztikai tesztek végzünk, arra vonatkozóan, hogy egy adott, viszonylag magas p^C értékkel jellemzett C klaszter esetén a p^C -t előállító időpillanatokban megjelenő meredekségek időben független viselkedést mutatnak-e vagy köztük kapcsolat állapítható meg.

4.4. Logaritmikus hozamszámítás

4.4.1. Általános értelmezés

A logaritmikus hozam másik neve folyamatos kamatozás, ugyanis a kamatfizetés technikailag minden időpillanatban történik. Számítási módjának megértéséhez először vezessük be az r hozamú, éven belül m alkalommal történő kamatos kamatozás képletét:

$$V_1 = V_0 \cdot \left(1 + \frac{r}{m}\right)^m, \quad (4.8)$$

ahol V_0 az alaptőke, V_1 pedig a tőke a futamidő végén.

Ha m tart a végtelenhez (azaz a bank minden pillanatban jóváírja a kamatot), akkor határértéke e^r , ahol e az Euler-féle szám:

$$\lim_{m \rightarrow \infty} \left(1 + \frac{r}{m}\right)^m = e^r. \quad (4.9)$$

Így a folytonos kamatozás képlete:

$$V_1 = V_0 \cdot e^{t \cdot r}. \quad (4.10)$$

4.4.2. Ismeretlen hozam

Tegyük fel, hogy a fenti (4.10) egyenletben az ismeretlen az r hozam, amelyet az egyes devizaárfolyamok esetén a t és $t + D$ időpillanatok között keresünk (a képletet most már a dolgozat témájára vetítve), ahol D a két időpillanat között eltelt idő hossza, percben megadva.

Így, az egyenletet átrendezve az egyes r hozamokat a szoftveres implementációban a következőképpen határoztuk meg:

$$r = \ln \left(\frac{S_{t+D}}{S_t} \right), \quad (4.11)$$

ahol S_{t+D} , illetve S_t a megfelelő időpontokhoz tartozó devizaárfolyamok.

4.5. Statisztikai tesztek

Legyen C híreknek egy klasztere, amelyek megjelenését követően az árfolyammozgások az előző szakaszban bevezetett mérőszám szerint $p^C = 0,8$ mértékű egyezést mutatnak. Legyenek az R^C mátrix sorai a kapcsolódó hírek megjelenési idejének megfelelően rendezve, és tegyük fel, hogy a mátrix p^C értékét

szolgáltató oszlopában előforduló előjelek a

$$q^{C,1} = (-1, -1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, -1, -1) \quad (4.12)$$

vektorral jellemezhetők, ahol $q_i^{C,1}$ azt adja meg, hogy a C klaszterben szereplő, időrendben i -edik hír megjelenését követően a legmagasabb p^C értéket eredményező időpillanatban csökkent vagy növekedett az árfolyam. Mivel a pozitív irányú elmozdulások egy tömbben helyezkednek el, érezhető, hogy a p^C által jelzett hatás nem felel meg a valóságnak, feltételezhető, hogy a tisztán pozitív elmozdulással járó időszakban valamilyen nem modellezett tényező hatása tükröződik. Ugyanakkor, ha a $p^C = 0,8$ érték a

$$q^{C,2} = (1, 1, -1, 1, 1, 1, -1, 1, -1, 1, 1, 1, 1, 1, -1, 1, 1, 1) \quad (4.13)$$

árfolyamváltozások eredményeként áll elő, akkor feltételezhetjük, hogy sokkal inkább a C klaszterbe sorolt, a korábban definiált tekintetben hasonló hírek és a vizsgált devizapár kapcsolatát jellemző tulajdonságról, semmint egy külső hatás eredményéről van szó. Nagyobb mintaszám esetén mindez fokozottan érvényesül.

Ezen a ponton ismét egyszerűsítéssel élünk, és eltekintünk attól, hogy az egy klaszterben szereplő hírek között időben nagyon különböző időtartamok is eltelhetnek. A hírek megjelenésének pontos időpontjaitól eltekintve, csak azok egymásutánosságára építve próbáljuk meghatározni, hogy mennyire karakterisztikus tulajdonsága a C klaszterbe sorolt híreknek a bevezetett p^C egyezési ráta egy adott devizapárra nézve. Az előző példából intuitívan is érzékelhető, hogy az azonos irányú elmozdulások összefüggő blokkjainak mérete lehet a meghatározó. Valóban, az alább bemutatásra kerülő tesztek is ezt, vagy ezekből származtatott mutatókat vizsgálnak.

A C klaszterben szereplő hírek megjelenéseit követő jD időpontban az árfolyamváltozások előjeleit egy Bernoulli-folyamat realizációjának tekintjük. Meghatározzuk a mintában az azonos előjelű változásokat tartalmazó futamok számát (például az $q^{C,1}$ vektor esetén $N_r = 3$, a $q^{C,2}$ vektor esetén $N_r = 9$), valamint a mintában szereplő különböző előjelű elemek teljes számát (N_+ és N_-). Null-hipotézisként feltesszük, hogy az egyes változások egymástól függetlenek, s a hipotézisről a nem-paraméteres Wald-Wolfowitz futampróba [26] alapján döntünk.

A null-hipotézis megtartása esetén kijelenthetjük, hogy a C klaszterben szereplő hírek megjelenését követően a p^C leíró eredményező időpontban az árfolyamok a kijelölt szignifikancia szinten függetlenül változnak, azaz a változások előjele egy C -re jellemző Bernoulli-eloszlást követ. Az eloszlás paramétere éppen p^C , így az empirikusan bevezetett p^C mutatónak statisztikai tartalmat adhatunk. A null-hipotézis elvetése esetén megállapíthatjuk, hogy az árfolyamváltozások előjeleinek sorozata nem független, valószínűsíthetően olyan külső hatások érvényesülnek, amelyeket a modellalkotáskor elhanyagoltunk.

Megjegyezzük, hogy a null-hipotézis elvetése esetén továbbra is él a lehetőség, hogy a C klaszterbe sorolt hírek egy másik, $j' \neq j$ indexszel jellemezhető időpillanatban $p^{C'} < p^C$ mértékű, de érzékelhető hatást gyakorolnak az árfolyamokra. Ezt azért tartjuk fontosnak kiemelni, mert a nagy mennyiségű adat miatt tett egyszerűsítések következtében, az egyes klasztereket csak a legerősebb leíróval jellemeztük. Ennek megfelelően a statisztikai eredmények az RSS-hírek árfolyamokra gyakorolt hatásának, mint feltételezésnek a megerősítést teszik csak lehetővé, cáfolatát nem.

5. fejezet

Eredmények

5.1. A felhasznált hírcsatornák

Mivel az RSS-csatornákról érkező hírek könnyen gyűjthetőek, a hírek leírása tömör, metaadatként elérhető a hír forrására mutató link, valamint a hír megjelenésének a pontos időpontja, az RSS-csatornákon érkező hírek kiváló alapot biztosítanak ahhoz, hogy azokat összevessük az érintett árfolyamok változásaival. Kiemeljük továbbá, hogy a jelenleg is fejlődő szemantikus web tovább bővítheti az RSS technológia lehetőségeit: többek között a Reuters hírügynökség támogatásával már fejlesztés alatt állnak olyan RSS technológiák, amelyek a hírek automatikus értelmezését lehetővé téve további metaadatokkal bővítik ki a jelenlegi szabványt, ezzel megkönnyítve a világhálón megjelenő információk és azok hatásainak vizsgálatát.

Az RSS-csatornákról gyűjtött híreket tartalmazó adatbázisunk pontosan 1 évet ölel fel: 2013. október 1-től 2014. október 1-ig tartalmazza a lejjebb felsorolt hírforrásokból begyűjtött adatokat.

Természetesen ahhoz, hogy a vizsgálatokat minél nagyobb hatékonysággal végezzük, sok alapadatra, forrásra volt szükség: mind a hírek, mind az árfolyamok szemszögéből.

Igyekeztünk azon hírforrásokra, hírcsatornákra feliratkozni, amelyeken az árfolyamváltozások szempontjából a legrelevánsabb hírek jelennek meg:

1. BBC News - Asia,
2. BBC News - Business,
3. BBC News - Europe,
4. BBC News - Middle East,
5. BBC News - Politics,
6. BBC News - Technology,
7. BBC News - US & Canada,
8. CNN.com - Business,
9. CNN.com - U.S.

A kilenc csatornáról összesen 39 110 darab hír gyűlt össze.

5.2. A vizsgált devizapárok

Távlati célkitűzésünk a jelenleg forgalomban lévő legfontosabb négy deviza egymáshoz viszonyított változásainak vizsgálata:

USD	United States Dollar	Amerikai dollár
GBP	British Pound	Angol font
EUR	Euro	Euro
JPY	Japanese Yen	Japán yen

A felsorolt 4 devizát érintő árfolyam táblázatok a következő forrásból származnak:
<http://www.histdata.com/>.

A honlapon található, letölthető változatok közül az 1 perces, azaz a legkisebb és legpontosabb bontásban lévőkön végeztük a vizsgálatokat. Ez egy devizapáros esetén, 1 hónapos időtartamot tekintve átlagosan körülbelül $30\,000 \cdot 4$, azaz $120\,000$ értéket jelent. A négyszeres szorzó onnan származik, hogy egy adatsor 4 arányt tartalmaz minden percben: nyitó, legmagasabb, legalacsonyabb és záró értéket. A teljes adatbázist tekintve durván 9 millió árfolyamráta-val rendelkezünk, ami jól illusztrálja az elemzések nagy számításigényét.

5.3. A kapott klaszterek

Az előző fejezetben leírt modell sok egyszerűsítést tartalmaz és sok olyan faktor figyelembevételét elhanyagolja, amelyek fontosságát nehéz megítélni, a bemutatott módszer könnyen értelmezhető, az eredmények reprodukálhatóak, és itt emeljük ki, hogy a számításigény is kezelhető. Feltett célunk a modell további finomítása a már korábban is említett szempontok szerint, azonban a fenti egyszerű modellel végzett elsődleges tesztek is igen érdekesnek ítélt eredményeket produkáltak, amelyeket szeretnénk megosztani a nyilvánossággal. Jelen fejezetben az előállt eredményekből emelünk ki néhányat.

Az alábbi táblázatban láthatjuk az egyes különbözőségi függvényekkel, $\epsilon = 0,4$ választással kapott klaszterek tulajdonságait: a táblázat sorai a kernel-függvények különböző paraméterezései szerint tagolják az eredményeket, míg az oszlopokban a különböző számosságú klaszterek gyakoriságát különítettük el.

Különbözőségi függvény	Klaszterek számossága						Összesen
	2	3	4	5	6-9	~50	
d_2^{binary}	834	82	17	4	2	1	940
d_3^{binary}	540	26	2	1	0	1	570
$d_2^{spectrum}$	1098	126	28	8	4	1	1265
$d_3^{spectrum}$	521	20	3	2	0	1	547
$d_4^{spectrum}$	469	14	3	1	0	1	488

Jóllehet különböző paraméterekkel különböző klaszterek álltak elő, némi hasonlóság mégis felfedezhető: a kis számosságú klaszterek dominálnak, és minden esetben előáll egy nagy, 50 körüli elemszámmal rendelkező klaszter. Ahhoz, hogy helytálló következtetéseket vonhassunk le, a nagy számosságú klasztereket vizsgáljuk részletesen a fejezet hátralévő részében. A kis számosságú klaszterek nagy száma a vizsgált hírek rövidegével indokolható. Mivel az RSS-folyamokon jellemzően tömör, néhány mondatot tartalmazó hírek jelennek meg, a hasonló tartalom azonosítása nehéz feladat. További vizsgálatokban ezt

alapvetően két módon javíthatjuk: egyrészt a hírekhez kapcsolódó linkek követésével azok bővebb leírását felhasználva; másrészt bizonyos kifejezések (például IMF) megjelenéséhez nagyobb súlyt rendelve.

Összességében elmondható, hogy az új módszereknek köszönhetően kapott, felhasználható eredmények mennyisége viszonylag alacsony, de a fejlemények annál érdekesebbek. A klaszterezési eljárások tartalmilag az elvárásainknak megfelelően csoportosították a felhalmozott híreket. Az automatizált metódusok által képzett csoportokból szeretnénk az alábbiakban néhány kiemelkedő példát kiragadni, amelyekkel, habár nem 100%-os bizonyossággal és csak rövid időintervallumokat alapul véve, de bizonyos tekintetben esetleges ellenpéldákkal szolgálhatunk a közepesen hatékony piacokra vonatkozó feltételezésekre, illetve alkalmas alapot képezhetünk a további analízisekhez.

5.3.1. Példa

A klaszterezés működésének illusztrálására bemutatunk egy öt elemű klasztert, amely elemei valóban hasonló tartalmú hírek:

- 2014.06.27. 18:51:00 - *Summit backs Juncker in blow to UK*
EU leaders nominate Jean-Claude Juncker as next European Commission president, despite strong UK opposition.
- 2014.07.02. 4:09:00 - *Juncker to be face of EU*
Jean-Claude Juncker is a step closer to becoming the next president of the European Commission after EU leaders approved his nomination, despite strong opposition from Britain.
- 2014.07.15. 11:47:00 - *MEPs to approve Juncker nomination*
MEPs are expected to approve the nomination of Jean-Claude Juncker as head of the European Commission, despite UK opposition.
- 2014.07.15. 14:28:00 - *Juncker voted in as EC president*
MEPs elect Jean-Claude Juncker as head of the European Commission, despite opposition from some UK members.
- 2014.07.15. 15:23:00 - *Juncker elected to top EU job*
MEPs elect Jean-Claude Juncker - a veteran champion of EU integration - as head of the European Commission despite UK opposition.

5.4. A nagy számosságú klaszterek vizsgálata

A különböző paraméterű különbözőségi függvényekkel végzett klaszterezések eredményeként előálló legnagyobb, nagyságrendileg 50 elemű klaszterek nagyon hasonlóak voltak, néhány elemtől eltekintve ugyanazon híreket sorolta egy klaszterbe a hierarchikus klaszterezés.

Az ide tartozó hírek, események a különböző hírportálokon viszonylag sűrűn jelennek meg, és lényegében különböző, a világ bármely táján elhelyezkedő országok, államok állapotát írják le, mindezt gazdasági, politikai, kulturális szempontokból, amelyek közvetetten, de jó kiindulási alapot nyújthatnak ahhoz, hogy a különböző devizaárfolyamok alakulását tanulmányozzuk, és az azokban rejlő ingadozásokat magyarázzuk.

A 2-es paraméterezésű, bináris spektrum kernelt felhasználva felépített, magas elemszámú klaszterbe sorolt hírek a következők voltak:

- 2014.04.02. 11:24:00 - *Niue profile*
Provides an overview of Niue, including key events and facts
- 2014.04.02. 12:35:00 - *Macedonia country profile*
Provides an overview of Macedonia, including key events and facts about this European country
- 2014.04.02. 15:22:00 - *Belarus country profile*
Provides an overview of Belarus, including key events and facts about this east European country
- 2014.04.02. 16:23:00 - *Poland country profile*
Provides an overview of Poland, including key events and facts about this European country
- 2014.04.08. 11:39:00 - *Indonesia country profile*
Provides an overview of Indonesia, including key events and facts about the world's fourth most populous nation
- 2014.04.10. 13:34:00 - *Hungary country profile*
Provides an overview of Hungary, including key events and facts about this eastern European country
- 2014.04.10. 16:56:00 - *Latvia profile*
Provides an overview of Latvia, including key events and facts about this country which lies on the Baltic Sea
- 2014.04.10. 18:50:00 - *United Kingdom country profile*
Provides an overview of the United Kingdom, including key events and facts about this European country
- 2014.04.23. 11:19:00 - *Canada country profile*
Provides an overview of Canada, including key events and facts about the world's second-largest country
- 2014.04.29. 12:42:00 - *Australia country profile*
Provides an overview of Australia, including key facts and events
- 2014.04.29. 14:54:00 - *Papua New Guinea profile*
Provides an overview of Papua New Guinea, including key events and facts
- 2014.04.29. 17:10:00 - *Philippines profile*
Provides an overview of Philippines, including key events and facts.
- 2014.05.01. 17:37:00 - *Iceland country profile*
Provides an overview of Iceland, including key events and facts
- 2014.05.01. 17:53:00 - *Andorra country profile*
Provides an overview of Andorra, including key events and facts.
- 2014.05.01. 18:09:00 - *Denmark country profile*
Provides an overview of Denmark, including key events and facts about this European country
- 2014.05.05. 13:48:00 - *Finland country profile*
Provides an overview of Finland, including key events and facts about this northern European country.

- 2014.05.05. 18:30:00 - *Georgia country profile*
Provides an overview of Georgia, including key events and facts
- 2014.05.08. 15:15:00 - *Croatia country profile*
Provides an overview of Croatia, including key events and facts about this European country which used to be part of Yugoslavia
- 2014.05.08. 17:46:00 - *Sweden country profile*
Provides an overview of Sweden, including key events and facts about this European country.
- 2014.05.14. 12:16:00 - *Monaco country profile*
Provides an overview of Monaco, including key events and facts about tiny kingdom on the Mediterranean Sea
- 2014.05.14. 13:50:00 - *San Marino country profile*
Provides an overview of San Marino, including key events and facts about this tiny European country
- 2014.05.14. 17:31:00 - *Lithuania country profile*
Provides an overview of Lithuania, including key events and facts about this small country on the Baltic Sea
- 2014.05.15. 12:44:00 - *Vietnam country profile*
Provides an overview of Vietnam, including key events and facts about this east Asian country
- 2014.05.20. 16:14:00 - *Belgium country profile*
Provides an overview of Belgium, including key events and facts about this European country
- 2014.05.21. 13:12:00 - *Bosnia-Herzegovina*
Provides an overview of Bosnia, including key events and facts about this European country which used to form part of Yugoslavia.
- 2014.05.21. 16:03:00 - *Vanuatu country profile*
Provides an overview of Vanuatu, including key events and facts
- 2014.05.21. 16:38:00 - *Palau country profile*
Provides an overview of Palau , including key events and facts about this tropical island nation in the northern Pacific.
- 2014.05.22. 11:15:00 - *Laos profile*
Provides an overview of Laos, including key events and facts
- 2014.05.22. 12:22:00 - *China profile*
Provides an overview of China, including key events and facts.
- 2014.05.22. 13:03:00 - *Xinjiang profile*
Provides an overview of Xinjiang, including key events and facts.
- 2014.05.22. 13:10:00 - *Germany country profile*
Provides an overview of Germany, including key events and facts about this European country
- 2014.05.22. 13:33:00 - *Tokelau profile*
Provides an overview of Tokelau, including key events and facts

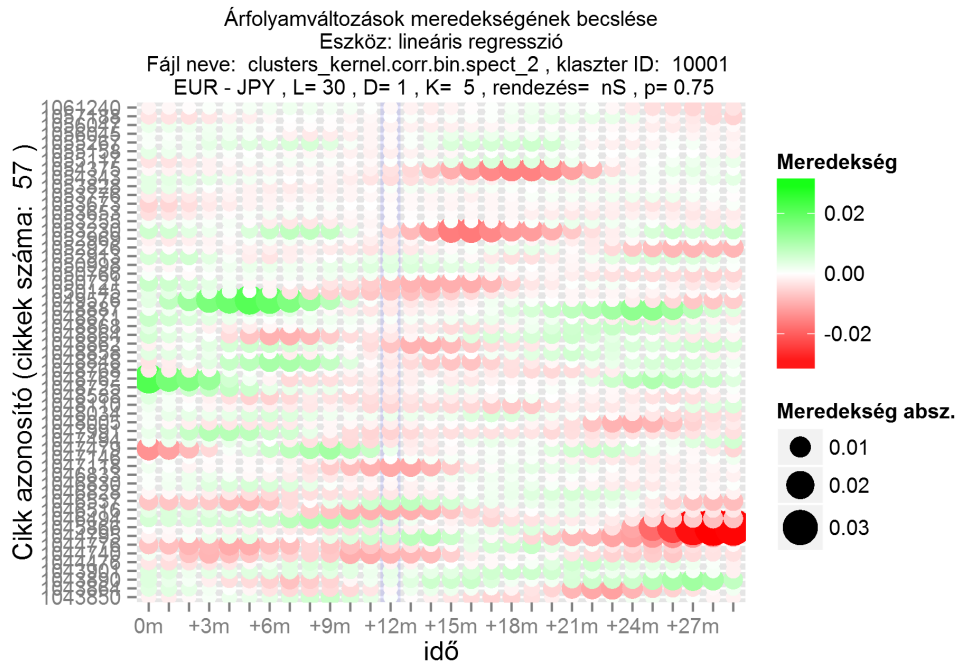
- 2014.05.22. 13:57:00 - *Mongolia profile*
Summary: *Provides an overview of Mongolia, including key events and facts.*
- 2014.05.22. 15:47:00 - *Singapore profile*
Provides an overview of Singapore, including key events and facts.
- 2014.05.29. 14:26:00 - *Kazakhstan profile*
Provides an overview of Kazakhstan, including key events and facts about this vast, mineral-rich Central Asian country.
- 2014.06.04. 17:59:00 - *Estonia country profile*
Provides an overview of Estonia, including key events and facts about this small country on the Baltic sea
- 2014.06.10. 14:22:00 - *Netherlands profile*
Provides an overview of the Netherlands, including key events and facts about this European country
- 2014.06.10. 18:45:00 - *Greece country profile*
Provides an overview of Greece, including key events and facts about this European country
- 2014.06.12. 12:48:00 - *Albania profile*
Provides an overview of Albania, including key events and facts about this European country
- 2014.07.01. 11:47:00 - *Jordan profile*
Provides an overview of Jordan, including key events and facts about this small country in the Middle East.
- 2014.07.01. 13:19:00 - *Oman profile*
Provides an overview of Oman, including key events and facts about the oldest independent state in the Arab world
- 2014.07.01. 17:18:00 - *United Arab Emirates profile*
Provides an overview of the United Arab Emirates - known as the UAE - including key events and facts about this federation of seven states.
- 2014.07.03. 14:20:00 - *Iraqi Kurdistan profile - Overview*
Provides an overview of Iraqi Kurdistan, including key events and facts about this autonomous region
- 2014.07.03. 16:30:00 - *France country profile*
Provides an overview of France, including key events and facts about this European country.
- 2014.07.08. 13:42:00 - *Iran country profile*
Provides an overview of Iran, including key events and facts about this country whose supreme leader is an Islamic cleric
- 2014.07.08. 16:59:00 - *United States country profile*
Provides an overview of the United States, including key events and facts about this superpower

- 2014.07.09. 12:44:00 - *Indonesia country profile*
Provides an overview of Indonesia, including key events and facts about the world's fourth most populous nation
- 2014.07.09. 18:46:00 - *Cambodia profile*
Provides an overview of Cambodia, including key events and facts.
- 2014.07.15. 11:51:00 - *Cook Islands profile*
Provides an overview of the Cook Islands, including key events and facts.
- 2014.07.15. 15:32:00 - *Slovenia country profile*
Provides an overview of Slovenia, including key events and facts about this European country which used to be part of Yugoslavia
- 2014.07.23. 12:29:00 - *Hong Kong profile*
Provides an overview of Hong Kong, including key events and facts.
- 2014.07.23. 18:03:00 - *Marshall Islands profile*
Provides an overview of the Marshall Islands, including key events and facts.
- 2014.07.24. 14:16:00 - *Bulgaria country profile*
Provides an overview of Bulgaria, including key events and facts about this eastern European country
- 2014.08.14. 16:00:00 - *Czech Republic profile*
Provides an overview of the Czech Republic, including key events and facts about this European country which was part of the former Czechoslovakia.
- 2014.09.26. 19:42:00 - *Fiji country profile*
Provides an overview of Fiji, including key events and facts about this tropical island and ethnically diverse former British colony in the Pacific.

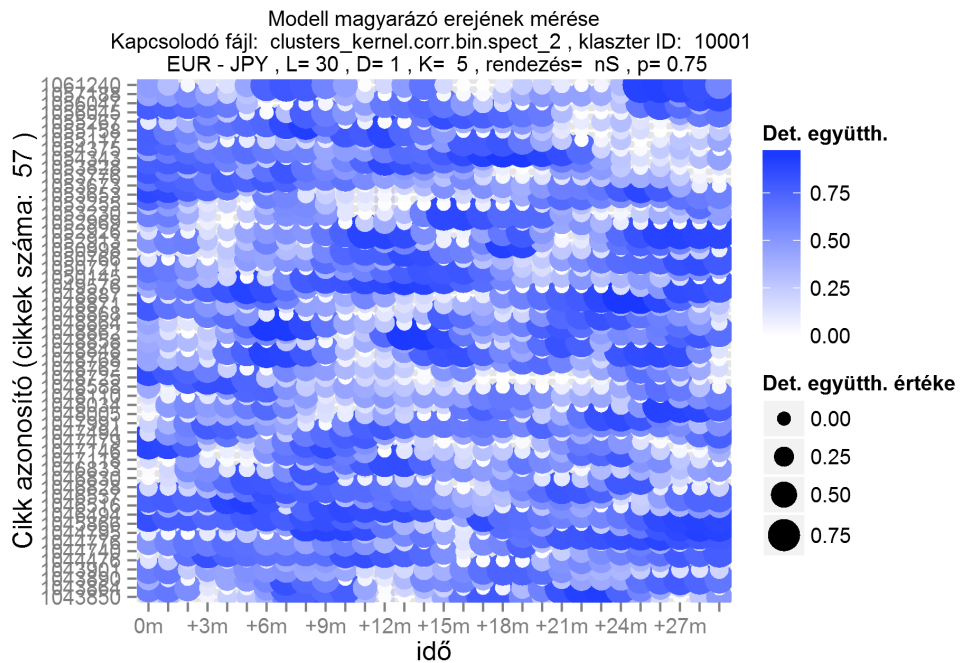
A hírek szemmel láthatóan különböző országok politikai berendezkedését, gazdasági tulajdonságait mutatják be. Több devizapár esetén is megvizsgáltuk mind lineáris regressziószámítással, mind logaritmikus hozamszámítással, hogy a klaszterbe tartozó hírek megjelenését követően van-e olyan időpont, amikor az árfolyam hasonló irányú elmozdulásokat mutat.

Választásunk a megvizsgált devizapárosok közül az EUR-JPY-ra esett, ugyanis ezen devizák esetén tapasztaltuk a legmagasabb egyezőséget az egy időpontokhoz tartozó árfolyam-ingadozások esetén (továbbá vegyük észre, hogy többségükben az európai és ázsiai államokról van szó a kapcsolódó hírekben, amely szintén ezen páros választását indokolja), amelyeket az alábbi grafikonokkal szemléltetünk.

Az 5.1 ábrán láthatjuk az EUR-JPY devizapáros árfolyamának alakulását az országokat bemutató hírek megjelenését követően, az árfolyamok változását lineáris regresszióval közelítve, a következő paraméterezéssel: a vizsgált időpontok száma 30, a közöttük lévő időtartam hossza 1 perc, míg az egyenes illesztéséhez felhasznált árfolyamértékek száma 11. Az elemzésre alkalmas klaszter elemszáma 57, amelyekből az esetek 75%-ában a hírek megjelenését követő 12. percben az árfolyam ugyanazon irányba mozdult: csökkent. Az illesztett regressziós egyenesek determinációs együtthatóit (R^2) az 5.1(b) ábrán vizualizáltuk. A legjobb p értéket produkáló, 12-edik perchez tartozó oszlopban az 57-ből csupán 8 esetben jelenik meg 0,25 alatti determinációs együttható, ami a regressziós eljárás alkalmazhatóságát indokolja.

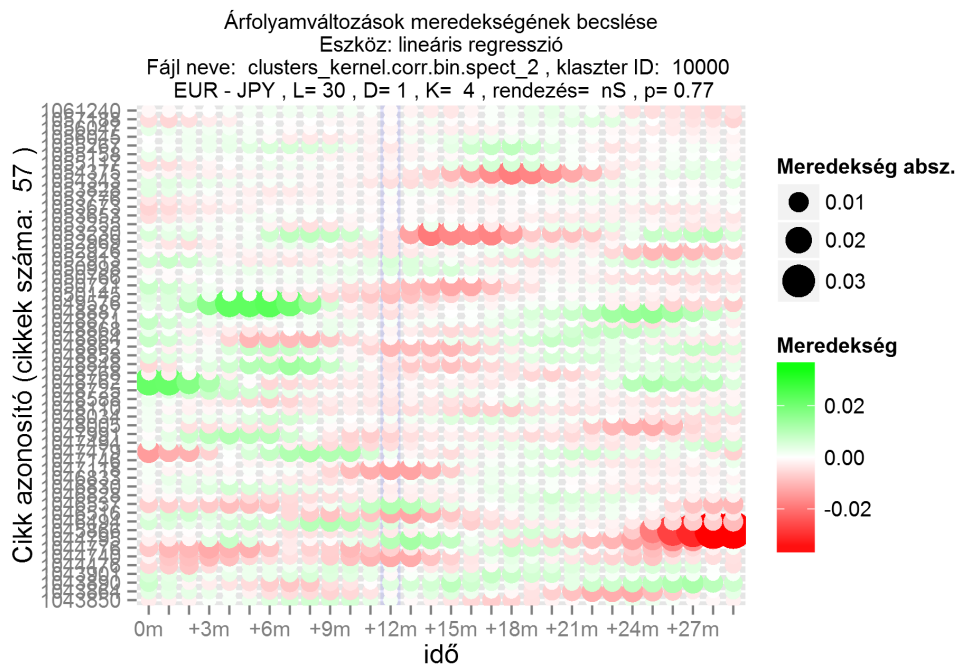


(a) EUR-JPY devizapáros, 30-1-5 paraméterezés, 0,75-ös egyezési arány, lineáris regresszió segítségével produkálva

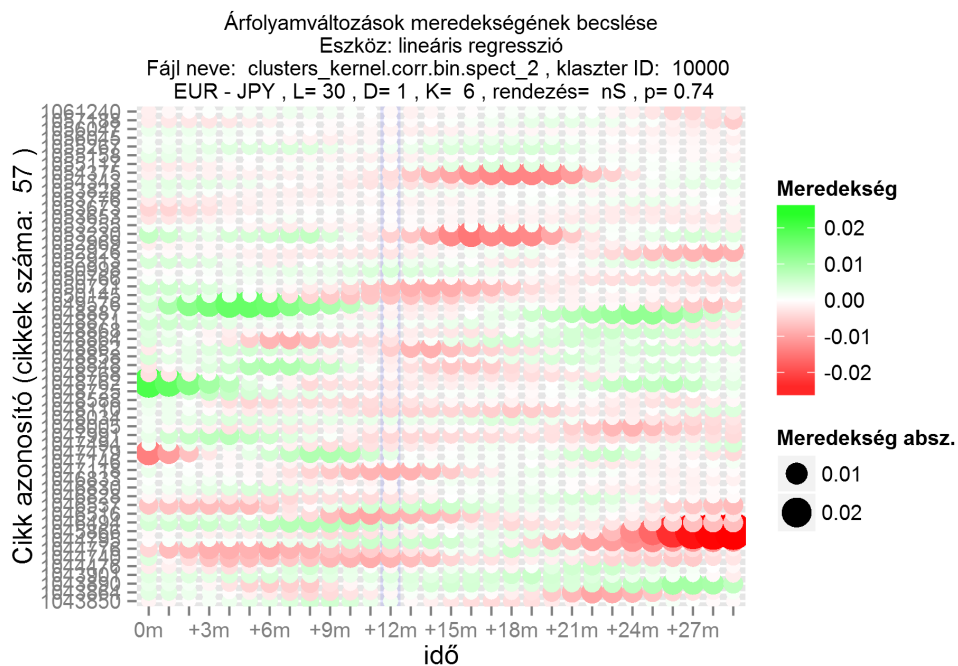


(b) A 5.1(a) kimenethez tartozó, regressziós „jóságot” ellenőrző táblázat

5.1. ábra. Magas klaszter-számosság magas egyezési aránnyal



(a) EUR-JPY devizapáros, 30-1-4 paraméterezés, 0,77-es egyezési arány, lineáris regresszió segítségével produkálva



(b) EUR-JPY devizapáros, 30-1-6 paraméterezés, 0,74-es egyezési arány, lineáris regresszió segítségével produkálva

5.2. ábra. Ugyanazon, nagy számosságú klaszter vizsgálata más regressziós beállításokkal

Ezen eredmény tehát azt sugallja, hogy a klaszterbe sorolt, hasonló tartalmú hírek hatással vannak az árfolyam alakulására. Jóllehet ezt teljes bizonyossággal nem jelenthetjük ki, további teszteket végeztünk, hogy az állítás valóságtartalmát megerősítsük.

Elképzelhető, hogy az EUR-JPY árfolyam a hírek által átölelt időszakban hosszú távú növekvő trendeket mutatott. Ha ezen időszakokból több hír származik, az összefüggő, azonos előjelű árfolyamváltozások sorozataként kell megjelenjen az árfolyamváltozásokat bemutató mátrix egy oszlopában. Ezen külső behatás vizsgálható a Wald-Wolfowitz teszt segítségével. Null-hipotézisként feltettük, hogy az árfolyamváltozások iránya független, azaz a legjobb egyezést produkáló, 12-edik percnél jelentkező oszlopban az előjelekből álló +1, -1 sorozat elemei egymástól függetlenek. A teszt segítségével arról hozunk döntést adott szignifikancia szint mellett, hogy el kell-e vetnünk a null-hipotézist. Ha a null-hipotézist elvetjük, az azt jelenti, hogy hosszabb +1, -1 sorozatok jelennek meg, mint ami a független sorozat esetén elfogadható. A tesztek során azt kaptuk, hogy 0,05 szignifikancia szinten a null-hipotézist elfogadhatjuk, azaz a fent leírt modell értelmében külső hatás nem érvényesül.

Egy másik megközelítésben néhány egyszerűsítéssel élve arra adunk durva becslést, hogy mekkora a valószínűsége annak, hogy 57 árfolyamváltozást megvizsgálva az esetek 75%-ában a változások iránya megegyezik. Számos széleskörűen elfogadott modell és elmélet, valamint a tapasztalat azt támasztja alá, hogy az árfolyamok hosszabb távon korrelálatlanok. Például néhány órával korábbi árfolyammozgások egyáltalán nem, vagy nagyon gyenge korrelációt mutatnak egy adott pillanat árfolyamváltozásaiival. Mivel az 57 hír 26 különböző napon jelent meg, és 3 esettől eltekintve a hírek megjelenése közötti időbeli különbség több, mint egy óra, feltehetjük, hogy a hírek megjelenésének időpontjaiban az árfolyamváltozások iránya egymástól független. Ezen feltevést az előző teszt is alátámasztja. Ha a hírek megjelenésekor egymástól független az árfolyamváltozások iránya, akkor a hírek megjelenéseit követő 12-edik percekben is függetlennek tekinthető. Feltéve, hogy a vizsgált klaszterben szereplő hírek nincsenek hatással az árfolyamokra, kiszámíthatjuk, mekkora a valószínűsége annak, hogy 57 független árfolyamváltozás vizsgálata során az esetek 75%-ában egyező előjelet kapunk. Könnyen belátható, hogy -1, +1 számokból álló n elemű sorozat, melyben rendre k és $n - k$ szimbólum egyezik meg, éppen $2 \binom{n}{k}$ darab van. Az n elemű, előjeleket tartalmazó sorozatok száma 2^n . Annak valószínűsége, hogy egy n elemű, egyenlő valószínűséggel megjelenő -1, +1 számokat tartalmazó sorozatban legalább l érték megegyezzen (feltéve, hogy $l \geq \frac{n}{2}$):

$$P = \frac{2}{2^n} \sum_{k=l}^n \binom{n}{k}. \quad (5.1)$$

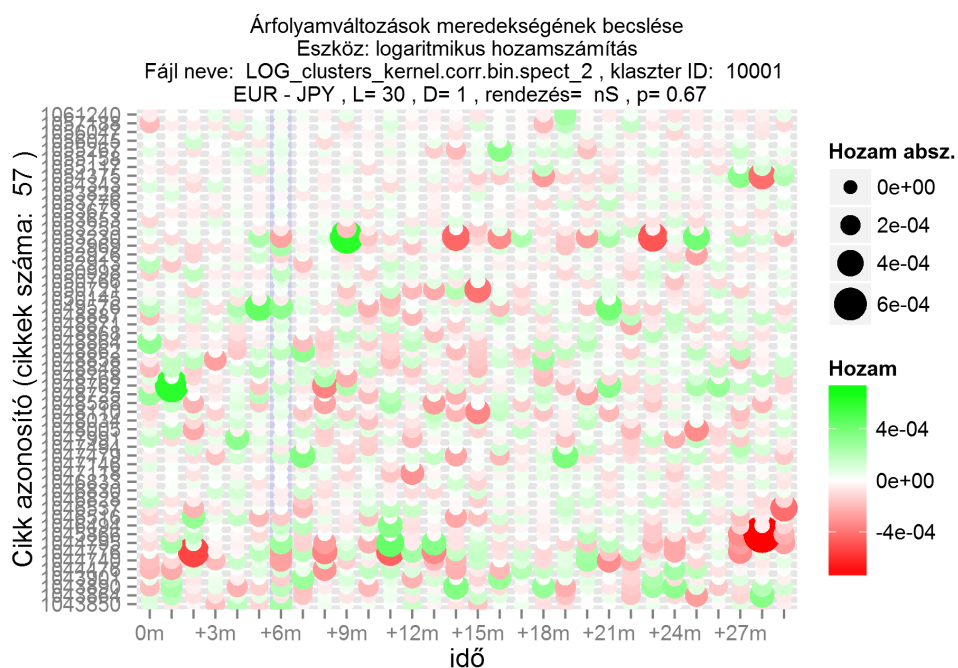
Az $n = 57$, $k = 43$ választással $P = 2,79 \cdot 10^{-6}$ adódik. Mivel 30 időpillanatot vizsgálunk (eltekintve az árfolyamnak ezek egymásutániságából adódó esetleges korreláltságtól), annak valószínűsége, hogy az előálló klaszterhez tartozó mátrix valamely oszlopában legalább 43 elem megegyezzen, $30 \cdot P = 0,000083$. Ha figyelembe vesszük, hogy 6 devizapárt vizsgáltunk, és a devizapárok együttes mozgásai szintén nagyon gyengén korreláltak, legalább 43 minta egyezésének valószínűsége még mindig csak 0,0006 körüli. Ezen valószínűség becslés (az egyszerűsítések mellett) akkor érvényes, ha az árfolyamokból véletlenszerűen választunk egymástól kellően távoli elemeket. A különbség a teljesen véletlenített mintaválasztás és a vizsgált eset között az, hogy mi az egy klaszterbe eső hírek megjelenését használtuk referenciaként a vizsgált minták kiválasztására. Feltételezhetjük tehát, hogy a referenciapontok megválasztása vezetett oda, hogy egy kis valószínűségű esemény bekövetkezett, azaz a hasonló tartalmú hírek hatása az árfolyamban megjelenik.

Bemutattuk ugyan a lineáris regresszió determinációs együtthatóink mátrixát, a vizsgálati módszer és az eredmények stabilitását tovább vizsgáljuk a kiértékelési módszer paramétereit (a regresszióhoz használt mintapontok számát) variálva. Arra a kérdésre keressük a választ, hogy a kapott kvalitatív eredmények megváltoznak-e a regresszióhoz használt mintapontok számának variálásával. Az 5.2(a) és 5.2(b) ábrákon a rendre $K = 4$ és $K = 6$ választással a lineáris regresszióhoz használt pontok számát 9-re, illetve 13-ra változtattuk. A legjobb eredményeket továbbra is a 12-edik percnél találjuk, rendre az esetek 77%-ában, illetve 74%-ában egyezik a változások iránya. Mindkét esetben megvizsgáltuk a sorozatok függetlenségét a Wald-Wolfowitz futamhossz teszttel, s a korábbihoz hasonlóan azt kaptuk, hogy az előjelek sorozatai 0,05 szignifikancia szinten függetlennek tekinthetők, külső hatás nem érvényesül. A korábbihoz hasonló egyszerűsített becsléssel hasonló eredményeket kapunk arra vonatkozóan, hogy mekkora az esélye annak, hogy a kapott egyezés a véletlen műve. Ahogy az az eredményekből látható, a regresszió paramétereinek adjusztálásával az eredmény és a következtetések nem változnak.

Megvizsgáltuk továbbá a klaszter elemeit, illetve az EUR-JPY árfolyam változásait úgy, hogy lineáris regresszió helyett logaritmikus hozamszámítással becsültük az árfolyamváltozások irányát. Megjegyezzük, hogy ha csak az előjeleket vizsgáljuk, ez az eset a lineáris regresszió speciális esetének tekinthető akkor, amikor 2 pontot használunk az egyenes illesztésére. Az eredmények az 5.3 ábrán láthatóak. A hatodik percben az árfolyamváltozások iránya az esetek 67%-ában egyezett. A lineáris regresszióhoz képest az egyezési ráta csökkent. Futamhossz tesztekkel kimutattuk, hogy a korábbi értelemben tekintett külső hatás nem érvényesül az egymást követő előjelek sorozatában, ugyanakkor megjegyezzük, hogy annak a valószínűsége, hogy ez az egyezés a véletlen műve, jelentősen növekszik: $P = 0,016$ valószínűséggel egyezik egy 57 elemű, két lehetséges elemet egyenlő valószínűséggel tartalmazó véletlen sorozatban 38 elem; figyelembe véve, hogy 30 egymást követő időpontot vizsgálunk, a véletlen egyezés valószínűsége 0,48-ra növekszik. Ha azt is figyelembe vesszük, hogy több árfolyampárt vizsgáltunk, nagy valószínűséggel az egyezés a véletlen műve. A lineáris regresszióval és logaritmikus hozamszámítással kapott eredmények látszólagos ellentmondása azzal oldható fel, hogy a két adatponton alapuló logaritmikus hozamszámítás jóval érzékenyebb a zajokra és véletlen változásokra, mint a több mintaponton alapuló lineáris regressziós számítások.

Végezetül álljon itt néhány szó arról, hogy a fentiekben bemutatott eredményeket hogyan is magyarázhatjuk, mivel indokolhatjuk az országok, városok, régiók gazdasági állapotát prezentáló híresemények és az árfolyam-ingadozások között feltételezett kapcsolatot. Alapvetően a devizák az egyes, földrajzilag elkülöníthető gazdasági szektorokhoz kötődnek, és azok legfontosabb gazdasági és pénzügyi mutatószámainak tekinthetők, ha az egymáshoz viszonyított értékeket tekintjük. Ebből kifolyólag, az ilyen típusú, régiókat leíró hírek talán a legmegfelelőbb kiindulási alapnak titulálhatóak az olyan kereskedők számára, akik a tőzsdei terminológiában használt, széles körben ismert, úgynevezett fundamentális elemzés (*fundamental analysis*) hívei. Ezen összefoglalók az előbb említett, mélyreható analízisekhez rendkívül sok információt hordozhatnak magukban (a klaszterszámosság mellett ez is egy fontos oka volt annak, hogy bemutatni kívánt példának választottuk meg őket): megemlítik a legfontosabb, aktuális gazdasági eseményeket, közlik a politikai vezetők kinevezéseit, tevékenységeiket, különböző gazdasági és a társadalmat egészében leíró, demográfiai mutatószámok alakulását, mint például a bruttó nemzeti jövedelem, vagy a munkaerőpiac jellemzői, illetve számtalan más, hasznos adatot.

A kvantitatív eredmények és a vizsgált klasztert alkotó hírek gazdasághoz, politikához erősen köthető tartalmának fényében az EUR-JPY árfolyamban kis valószínűsége ellenére is megjelenő nagy egyezési ráta például magyarázatul szolgálhat olyan esetekre, amikor az RSS hírek árfolyamokban jelentkező hatása kimutatható. Erősebb állítások megfogalmazásához további vizsgálatok szükségesek, melyekhez megfelelő



5.3. ábra. EUR-JPY devizapáros, 30-1 paraméterezés, 0,67-es egyezési arány, logaritmusos hozamszámítással produkálva

kiindulási állapot biztosíthatnak a vizsgálataink során azonosított klaszterek, illetve az elemzési módszer.

6. fejezet

Összefoglalás

A hírek és árfolyamok közötti kapcsolatok jelenlétét nyilvánvaló tényként kezelhetjük, azonban a kapcsolatok és összefüggések kvantitatív leírása és jellemzése még nem valósult meg. A kereskedés utóbbi években tapasztalt átalakulásának köszönhetően mind a kereskedési platform, mind a kereskedőket befolyásoló hírek elsődleges forrása az Internet. Ennek megfelelően a világ különböző pontjain kereskedő felhasználók ugyanazon híreket olvashatják, miközben ugyanazon devizapárokkal kereskednek, így a hírek árfolyamokra gyakorolt hatása is fokozottan jelentkezik, mint a korábbi években. Az összefüggések feltárása a devizapiacok működésének mélyebb megértését teszi lehetővé, és magyarázatot adhat olyan jelenségekre, amelyek a kereskedők kollektív viselkedésének eredményeként állnak elő egy-egy gazdasági hír megjelenését követően.

A dolgozat fő célkitűzése az RSS-hírfolyamok és FOREX-árfolyamok kapcsolatának vizsgálata volt. Ahogy azt a bevezetésben részleteztük, ez egy igen nagy és komplex feladat, így általános érvényű következtetéseket messze nem vonhatunk le, azonban az első lépéseket megtehetjük ezen ígéretes kutatási területen.

Az előzőekben bemutatott, kiragadott eredményeket, vagy akár a teljes, releváns eredményhalmaz bármely elemét tekintve is, rájöhettünk arra, hogy amennyire a közepesen hatékony piacokra vonatkozó hipotéziseket nem utasíthatjuk el, legalább ugyanannyira könnyelműen nem is fogadhatjuk el őket. A bemutatott klaszterekből kiindulva láthattuk, hogy van egy bizonyos, úgynevezett átfutási ideje a nyilvánosan közzétett információknak, mielőtt azok teljes mértékben beépülnének az árfolyamokba - és szerencsénkre, az általunk implementált, szoftveres megoldás alkalmas az ezen rövid, de jelentős időintervallumok vizsgálatára, illetve a beépülés folyamatának figyelemmel kíséretére.

A létező, széleskörűen elfogadott elméletek és módszerek tanulmányozása, az azokban részletezett tanulságok figyelembe vétele mellett kidolgoztunk egy modellt, amely kezelhető számításigénnyel teszi lehetővé a hír- és árfolyamok egyszerű összefüggéseinek vizsgálatát. A klaszterezési eljárásokkal készített csoportok jellemzésére bevezetett mérőszámaink könnyen interpretálhatóak és számíthatóak.

A kidolgozott modellt implementáltuk, s alkalmaztuk az előző fejezet elején felsorolt devizákból kombinálható párosokhoz kapcsolódó adatsorokra. Megállapíthatjuk, hogy az előálló eredmények sok esetben jól értelmezhetőek, a vizsgált klaszterekbe tartozó események gazdasági és történeti hátterét, a hozzájuk kapcsolódó asszociációkat megvizsgálva és értelmezve az árfolyamokra gyakorolt hatás elfogadható.

Mivel az árfolyamváltozásoknak nincs etalonnak tekinthető magyarázata, a modell és az eredmények validálása igen nehéz feladat. Az elsődleges eredmények *proof-of-concept* módon arra utalnak, hogy a kidolgozott modell működőképes, ugyanakkor általános érvényű kijelentések megfogalmazásához további

munkára és tesztekre van szükség:

1. Figyelemmel kell követnünk, hogy az előállt eredmények alapján megfogalmazott elvárások összhangban lesznek-e a jövőbeli árfolyamváltozásokkal, amikor a hírcsatornákon a már létező klaszterekbe illeszthető hírek jelennek meg.
2. A vizsgálati módszer kidolgozása során számos tényezőt elhanyagoltunk. Ezen tényezők és aspektusok figyelembevételével a modellt tovább kell finomítanunk, hogy pontosabb, könnyebben értelmezhető eredményekhez jussunk. A modell finomítása során az idősoranalízis és sztochasztikus folyamatok területéről további eszközöket és eljárásokat kell adaptálnunk a problémához.
3. Igen előnyös lenne kidolgozni egy eljárást, amely a hírek tartalmához egy árfolyam szempontjából értelmezhető jelzőt, címkét rendel. Ez a feladat egy újszerű kutatási terület eredményeinek alkalmazását jelenti, nevezetesen a természetes szöveg tartalmának visszaadását abban a speciális esetben, amikor egy szöveget egyetlen szóval szeretnénk jellemezni.
4. Az egyes klaszterekhez tartozó egyezési ráták időbeli változásokat mutathatnak, melyeknek figyelembevétele további fontos aspektusa lehet az eredmények helyes értelmezésének.
5. A jövőbeli elemzéseket és azok hitelességét még két fontos lehetséges lépés segíthetné elő: a bemutatottakon kívüli, más különbözőségi függvényeket is bevezethetnénk, valamint bizonyos kifejezésekhez nagyobb súlyt rendelve segíthetnénk elő a klaszterek kialakulását.
6. Végezetül, további elemzéseket nem csak a rövid RSS hírek felhasználásával, hanem az RSS csatornában lévő linkeken keresztül elérhető információk felhasználásával még hatékonyabban végezhetnénk.

7. fejezet

Munkamegosztás

Kovács György tevékenysége:

- szoftverfejlesztés: "keretbefoglalás" (CMake alkalmazása), melynek köszönhetően a szoftver funkciói szeparáltan, kerek egységet alkotnak;
- kiértékelt eredmények kiegészítése, kisimítása.

Saját tevékenységem:

- a kapcsolódó modellek tanulmányozása, vizsgálata;
- hír- és árfolyam-adatbázis felépítése, az adatok rendezése, és a feldolgozhatóság megvalósítása;
- szoftverfejlesztés: tényleges funkciók implementálása (C++ kódok megvalósítása), melyek lehetővé tették az elemzést;
- a szoftver kötegelt futtatása, eredmények generálása az elkészített klaszterek alapján;
- a különböző statisztikai adatok alapjául szolgáló adatállományok létrehozása, és feltöltése;
- megfelelő eredmények kiválasztása, kiértékelése, ok-okozati összefüggések feltárása;
- az elemzésekhez szükséges források beszerzése;
- SPSS tesztek elvégzése.

Közösen megoldott problémakörök:

- a szoftver funkciónak részletes megtervezése;
- R script-ek létrehozása a grafikus ábrázoláshoz, illetve azok további finomítása, pontosítása, bővítése;
- feldolgozási módszerek megtervezése;
- az eredmények szerkezetének és tartalmának megalkotása;
- a kimenetek (mind a köztes, szöveges állományok, mind a grafikus, végleges outputok) formai jegyeinek megalkotása;

- az eredmények értékelésére megalkotott mutatószámok megtervezése;
- a kapott eredmények részletes elemzése;
- az eredmények, statisztikai próbák és egyéb produktumok értelmezése.

Irodalomjegyzék

- [1] Várkúti, G.: *FOREX-trading - Az alapoktól a mesterfogásokig*, Fekete Sas Kiadó, Budapest, 2011
- [2] László, Zs. - Kovács, Gy.: *RSS hírcsatornák FOREX árfolyamokra gyakorolt hatásának vizsgálata*. Tudományos Diákköri Dolgozat, 2014. április, Debreceni Egyetem - Informatikai Kar, helyi forduló
- [3] Bradley, N.: *Az XML kézikönyv, Új, átdolgozott kiadás*, SZAK Kiadó, Bicske, 2005
- [4] http://en.wikipedia.org/wiki/Foreign_exchange_market
- [5] Brealy, Richard A. - Myers, Stewart C., *Modern vállalati pénzügyek*, Panem Könyvkiadó, Budapest, 2011
- [6] Kendall, M. G., *The Analysis of Economic Time-Series. Part I. Prices*. Journal of the Royal Statistical Society, 96. 1953. 11-25. old.
- [7] Bachelier, L.: *Theorie de la Speculation*. Gauthier-Villars, Párizs, 1900
- [8] Cootner, P. H.: *The Random Character of Stock Market Prices*, M.I.T. Press, Cambridge, Mass., 1964. 17-78. old.
- [9] Fama, E. F.: *Efficient Capital Markets: A Review of Theory and Empirical Work*. Journal of Finance, 25. 1970. május, 383-417. old.
- [10] Fama, E. F.: *The Behaviour of Stock Market Prices*. Journal of Business, 38. 1965. január, 34-105. old.
- [11] Alexander, S. S.: *Price Movements in Speculative Markets: Trends or Random Walks*. Industrial Management Review, 2. 1961. május 7-26. old. (újrányomtatva: [8], 199-218. old.)
- [12] Fama, E. F. - Blume, M.: *Filter Rules and Stock Market Trading Profits*. Journal of Business, 39. 1966. január, 226-241. old.
- [13] Moore, A.: *A Statistical Analysis of Common Stock Prices*. Kiadatlan Ph.D. tézis, Graduate School of Business, University of Chicago, 1962.
- [14] Fama, E. F. - Fisher, L. - Jensen, M. - Roll, R.: *The Adjustment of Stock Prices to New Information*. International Economic Review, X. 1969. február, 1-21. old.
- [15] Niederhoffer, V. - Osborne, M. F. M.: *Market Making and Reversal on the Stock Exchange*. Journal of the American Statistical Association, 61. 1966. december, 897-916.

- [16] Scholes, M.: *A Test of the Competitive Hypothesis: The Market for New Issues and Secondary Offsprings*. Kiadatlan Ph.D. t  zis, Graduate School of Business, University of Chicago, 1969.
- [17] Fama, E. F.: *Efficient Capital Markets: II*. Journal of Finance, 46. 1991. december, 1575-1617. old.
- [18] Keown, A. - Pinkerton, J.: *Merger Announcements and Insider Trading Activity*. Journal of Finance, 36. 1981. szeptember, 855-869. old.
- [19] Patell, J. M. - Wolfson, M. A.: *The Intraday Speed of Adjustment of Stock Prices to Earnings and Dividend Announcements*. Journal of Financial Economics, 13. 1984. j  nius, 223-252. old.
- [20] Carhart, M. M.: *On Persistence in Mutual Fund Performance*. Kiadatlan k  zirat, University of Chicago, 1994. december
- [21] Bernard, V. L. - Thomas, J. K.: *Post-Earnings Announcement Drift: Delayed Price Response or Risk Premium?* Journal of Accounting Research, 27. 1989. mell  klet, 1-36. old.
- [22] G  rtner, T.: *A survey of kernels for structured data*. SIGKDD Explor. Newsl., 5.(1.), 2003. j  lius, 49-58. old.
- [23] Lodhi, H. - Saunders, C. - Shawe-Taylor, J. - Cristianini, N. - Watkins, C.: *Text Classification using String Kernels*. Journal of Machine Learning Research, 2. 2002. febru  r, 419-444. old.
- [24] Leslie, C. - Eskin, E. - Noble, W. S.: *The spectrum kernel: a string kernel for SVM protein classification*. Pac. Symp. Biocomput., 2002., 564-575. old.
- [25] Karatzoglou, A. - Feinerer, I.: *Kernel-based machine learning for fast text mining in R*. Computational Statistics and Data Analysis, 54.(2.), 2010. febru  r, 290-297. old.
- [26] Wald, A. - Wolfowitz, J., *On a test whether two samples are from the same population*, Ann. Math. Statist., 11. 1940., 147-162. old.