MŰEGYETEM 1782

**Budapesti Műszaki és Gazdaságtudományi Egyetem**
Villamosmérnöki és Informatikai Kar

Kriston Melinda

# MAGYAR ÁRAMTŐZSDE ELEMZÉSE ÉS ELŐREJELZÉSE

KONZULENS

Dr. Pataki Béla

BUDAPEST, 2017

# HALLGATÓI NYILATKOZAT

Alulírott **Kriston Melinda**, szigorló hallgató kijelentem, hogy ezt a szakdolgozatot/ diplomatervet meg nem engedett segítség nélkül, saját magam készítettem, csak a megadott forrásokat (szakirodalom, eszközök stb.) használtam fel. Minden olyan részt, melyet szó szerint, vagy azonos értelemben, de átfogalmazva más forrásból átvettem, egyértelműen, a forrás megadásával megjelöltem.

Hozzájárulok, hogy a jelen munkám alapadatait (szerző(k), cím, angol és magyar nyelvű tartalmi kivonat, készítés éve, konzulens(ek) neve) a BME VIK nyilvánosan hozzáférhető elektronikus formában, a munka teljes szövegét pedig az egyetem belső hálózatán keresztül (vagy hitelesített felhasználók számára) közzétegye. Kijelentem, hogy a benyújtott munka és annak elektronikus verziója megegyezik. Dékáni engedéllyel titkosított diplomatervek esetén a dolgozat szövege csak 3 év eltelte után válik hozzáférhetővé.

Kelt: Budapest, 2017. 12. 14.

...…………………………………………….

Kriston Melinda

# Tartalomjegyzék

# Abstract

The topic of this thesis is to overview the implementation, comparison, and usability of Time-series Forecasting Methods, especially in the electricity market. A broad spectrum of methods has been used so far, for example, the nearest neighbour based models, ARMA, ARIMA models, artificial neural networks, regression models, moving averages, exponential refinement etc. The purpose of the thesis is to review them and selecting a group, then implementing and testing the method stack. The final goal is to estimate the next day price of the Hungarian Power Exchange.

Since the launch of the Hungarian Power Exchange (HUPX) in 2009, huge number of clearing benefits data have been stored, which are available to perform tests. In addition, MAVIR data publications are also available for the current state of the Hungarian electricity network.

The main problems, which were investigated:

- Processing and analyzing Hungarian stock exchange data.
- Examining the connection between collected data.
- The function of MAVIR data in price evolution.

In the last eight years, the Hungarian stock exchange has also operated in three different environments: independently, in the Czech-Slovak-Hungarian interconnection and also in the 4M Market Coupling

The European Union is developing a common electricity market. In a constantly changing environment, the most important task will be to better predict prices or to determine the evolution of consumption. Better predictions help better economic decisions.

In the thesis, the problem was investigated in several ways. I predicted the following day’s data using average and median of the previous period. I tried the ARIMA model and implemented the nearest neighbour algorithm that measures the proximity of the day with the absolute error. The highest hourly median calculating algorithm works, with the mean square error of 227,034 HUF.

# Összefoglaló

A szakdolgozat témája az idősor-előrejelzésre alkalmas módszerek áttekintése, implementálása, összehasonlítása, használhatóságuk feltérképezése. A módszerek igen széles spektrumát használták, javasolták már erre a célra, például a legközelebbi szomszéd alapú modellek, ARMA, ARIMA modellek, mesterséges neurális hálók, regressziós modellek, mozgó átlagolás, exponenciális simítás stb. ezek áttekintése és egy csoport kiválasztása, majd megvalósítása és tesztelése a feladat. A cél a magyar áramtőzsde következő napi árának minél pontosabb becslése.

A magyar áramtőzsde (HUPX) 2009-es indulása óta már több évnyi klíringeredmény áll rendelkezésre mélyrehatóbb elemző vizsgálatok elvégzéséhez. Ezen kívül a MAVIR adatpublikáció az elektromos rendszer pillanatnyi állapotáról is rendelkezésre állnak.

A vizsgált főbb problémák:

- A magyar tőzsdei adatok feldolgozása és elemzése
- Az összegyűjtött adatok közötti kapcsolat vizsgálata.
- A MAVIR adatok szerepe az ár alakulásában

Az elmúlt nyolc évben ráadásul háromféle környezetben is működött a magyar tőzsde: önállóan, a cseh-szlovák-magyar összekapcsolásban, valamint a 4M piac-összekapcsolásban. Az Európai Unió célja egy közös villamosenergia-piac. Az folyamatosan változó környezetben egyre fontosabb feladat lesz az árak minél pontosabb előrejelzése, a fogyasztás alakulásának meghatározása. A pontosabb előrejelzések segítik a jobb gazdasági döntéseket.

A dolgozatban megvizsgáltam többféle algoritmust. Átlagolással, mediánnal jeleztem előre következő napi adatot. Kipróbáltam az ARIMA modellt, illetve implementáltam egy legközelebbi szomszéd algoritmust, ami a napi abszolút hibával méri a napok közelségét. A legjobban az óránkénti mediánt számoló algoritmus működött, aminek az átlagos négyzetes hibája 227.034 forint lett.

# Összefoglaló fogalmai

**Klíringrendszer**: kölcsönös leszámítási rendszer, amelyben az elszámolási viszonyban levő pénzintézetek tartozásaikat kiegyenlítik. Készpénznélküli elszámolási rendszer, amikor készpénzben vagy valutában csak az egyenleget, vagyis a tartozások és követelések különbözetét, a klíringcsúcsot egyenlítik ki. Célja a pénzforgalom csökkentése. Belföldi viszonylatban, a bankügyletekben használatos. A külkereskedelemben a részt vevő kereskedők csak saját bankjukkal állnak kapcsolatban, velük hazai törvényes fizetési eszközben számolnak el. Nemzetközi szinten általában két ország jegybankja között szokásos vagy több ország közötti elszámolás központi rendezése. A klíringcsúcs az év végén keletkező egyenleg, amelynek kiegyenlítése történhet pótlólagos áruszállítással, arannyal vagy konvertibilis valutával, illetve jegybanki hitellel. [1]

**4M piac:** 4M piac-összekapcsolás – 4M Market Coupling. Az Európai Parlament és a Tanács 2009. július 13-i 714/2009/EK rendelete a villamos energia határokon keresztül történő kereskedelme esetén alkalmazandó hálózati hozzáférési feltételekről, és az 1228/2003/EK rendelet hatályon kívül helyezéséről szóló rendelettel összhangban a regionális szinten harmonizált menetrend-kezelési folyamat és a koordinált kapacitásallokáció került bevezetésre. Ennek megfelelően Magyarország 2012. szeptember 11-től a Cseh Köztársasággal és Szlovákiával volt piaci-összekapcsolásban. Miután Románia is belépett az EU tagállamok közé így ugyancsak kötelezett lett a vele határos Közép-Kelet Európai régió országai számára előírt követelmények és elvárások teljesítésében. 2013. augusztusában a nemzeti energiaszabályozó hatóságok jóváhagyták a cseh-szlovák-magyar áram piac-összekapcsolást Romániával. Ennek következtében 4M piac-összekapcsolás 2014. november 19-i kereskedési nappal sikeresen elindult.

A 4M MC projekt technikai és gazdasági célja, hogy a funkciók központosításával, a bonyolult, sok szereplő közötti többoldalú viszonyokat egy- az - egy típusú kapcsolattá alakítsa, ezáltal elősegítve a piac-összekapcsolás további piacok irányába történő kibővítését. A megvalósítás előtt a felek megállapodtak egy olyan megoldás kialakításában, amely a lehető legnagyobb mértékben kompatibilis a nyugat-európai régióval. Ezért a piac-összekapcsolás kiértékelési algoritmus bemeneti és kimeneti adatainak teljes kompatibilitását hozták létre ~~az~~ nyugat-európai régióval, lehetővé téve a két régió összekapcsolását, vagy a későbbiekben új felek csatlakozását. [2]

# 1 Magyar áramtőzsde

A pénzügyi idősorok előrejelzését nagyban nehezíti, hogy ezek általában zajosak, nem stacionáriusak, nemlineárisak és kaotikusak, továbbá gyakran fordul elő bennük strukturális törés is Ezen okok miatt a pénzügyi/tőzsdei idősorok előrejelzése az egyik legnagyobb kihívás a piaci szereplők számára.

**MAVIR**: Magyar Villamosenergia-ipari Átviteli Rendszerirányító Zártkörűen Működő Részvénytársaság. A MAVIR – mint az átviteli hálózat tulajdonosa – felelős az átviteli hálózat folyamatos működtetéséért és fejlesztéséért. Összehangolja a magyar villamosenergia-rendszer és a szomszédos hálózatok működését (export-import). [3] [4]

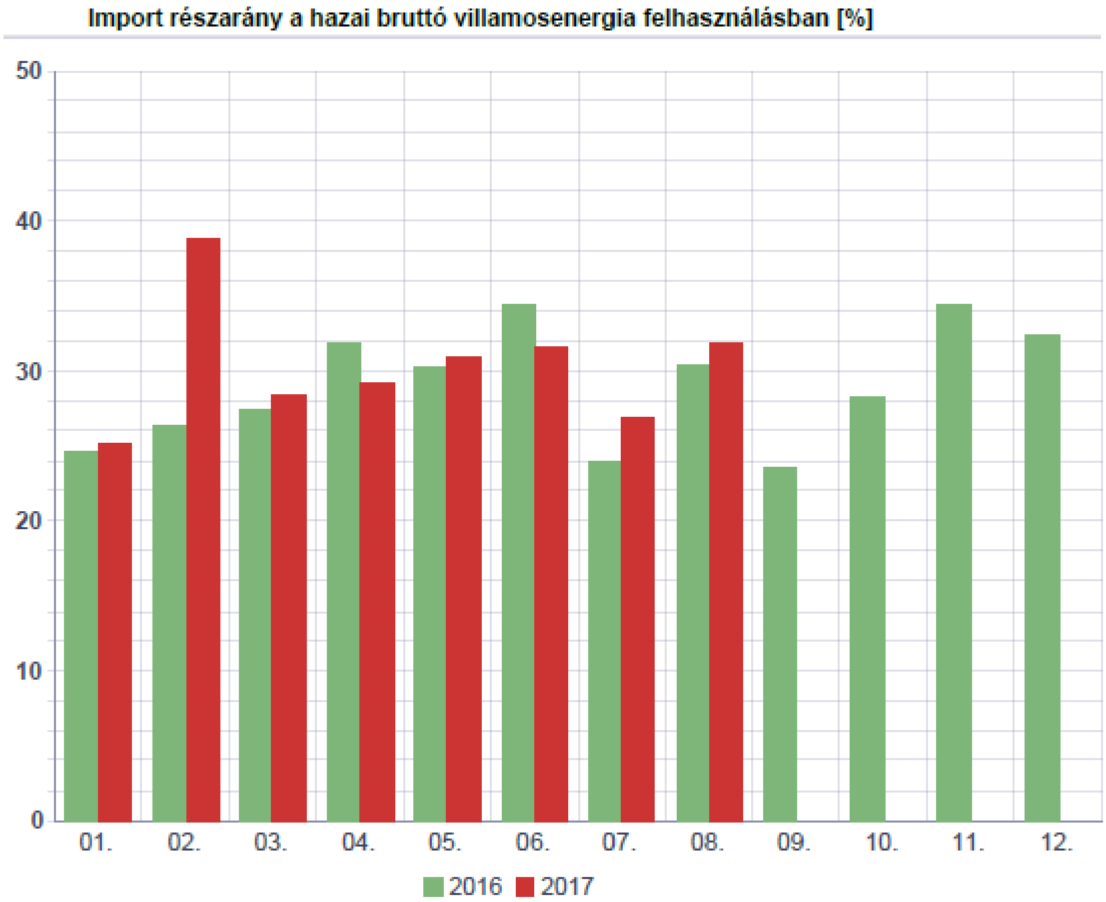
A rendszerirányítás feladata az országos energiarendszer teljesítmény-egyensúlyának fenntartása, a mérlegkörök tervektől való eltéréseinek kiegyenlítése. Ehhez meghatározza a szükséges tartalékokat, a szabályozás számára lekötött teljesítményeket. Elkészíti a hálózatfejlesztési stratégiát és javaslatot tesz az erőműpark fejlesztésére.

A MAVIR jelentős szerepet játszik a piacszervezés területén, folyamatosan nyomon követi és alkalmazza a jogszabályi környezet változását. A rendszerirányító alapította és működteti a magyar villamosenergia-tőzsdét (HUPX).

A MAVIR honlapról sok fajta adatot lehet beszerezni órás és 15 perces bontásban. Végül a terv és tény rendszerterhelést, az import-export adatokat, a határmetszéki áramlásokat és a széltermelést töltöttem le a teljes időszakra. Ezeken kívül a honlapon az egyes erőművek termelése és a frekvencia volt még elérhető, amiket nem töltöttem le, mivel úgy gondoltam, hogy a frekvencia nem lehet befolyásoló tényező az árak kialakulására, míg az erőműveknek csak az össztermelésére lehet szükségem, ami számítható a rendszerterhelésből és az importból. Ezeken kívül is használtam a honlapról diagramokat, adatokat szemléltetés okán.

A MAVIR honlapról beszerzett dokumentumok és diagramok tanulmányozása során kiderült, hogy a hazai termelés nem elegendő a hazai kereslet kielégítésére. Az igényt átlagban 30% import révén tudják biztosítani. [5] Az alábbi diagramon (1. ábra) jól látszik az import aránya a hazai villamosenergia-felhasználásban.

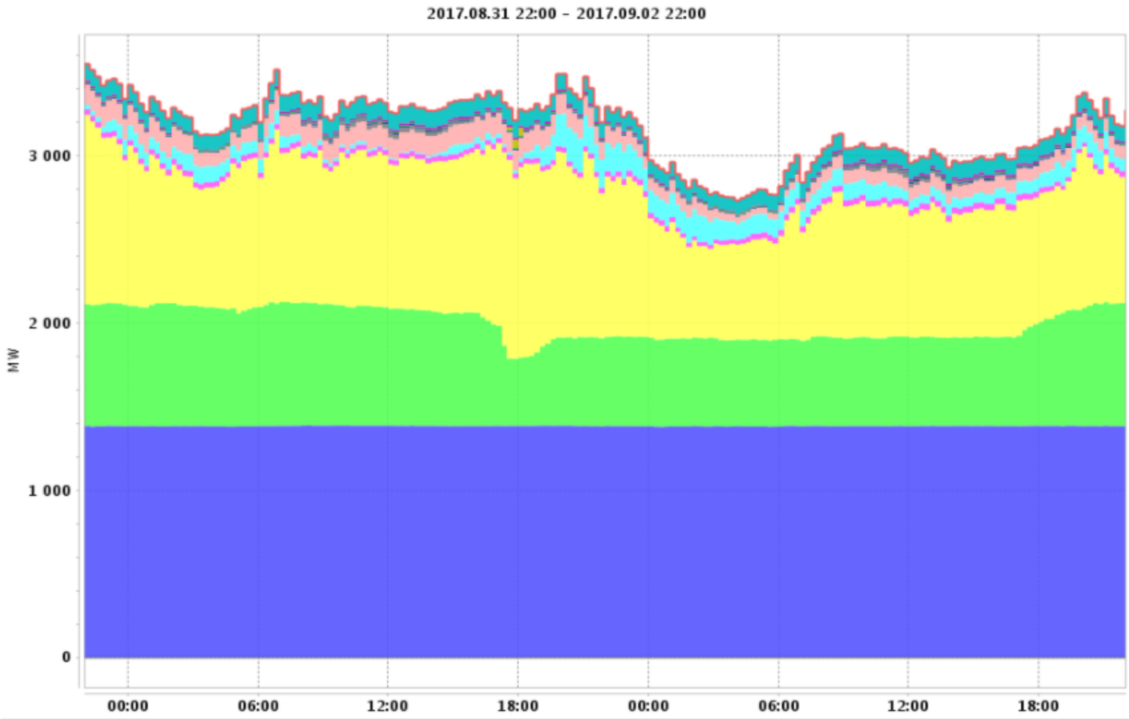
A piacnyitás óta a tényleges export-import szaldó nagysága a piaci szereplők – akár napi – üzleti döntésein, a piaci kínálaton, illetve a nemzetközi vezetékek aktuális átviteli kapacitásán múlik, tehát bizonytalan.




*1. ábra 2016, 2017 évi import százalékos aránya a teljes rendszerterhelésben hónapos bontásban*

Érdekes hogy július és szeptember hónapban hírtelen csökken az import mennyisége. Illetve 2017. év. februárja nagyon kiugrott.

A maradék 70%-ot belső termelés adja. Többfajta erőmű szolgáltatja az áramot Magyarországon, az alábbi diagramon (2. ábra) 48 órás időszak látható. Itt jól látszik, hogy a paksi atomerőmű áramtermelése a legmeghatározóbb, ezek után a kőszén-, gázerőművek adják a legtöbb energiát és egy elenyésző rész a zöld energia, a víz, szél és egyéb megújuló energiák.



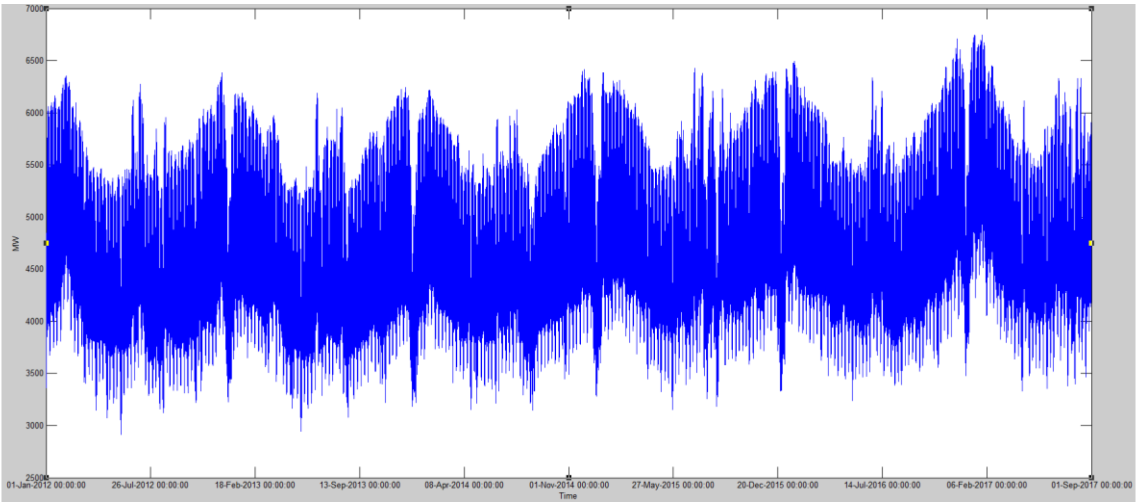

*2. ábra*

*A magyarországi erőművek termelése, fajtánként órás bontásban, alább az egyes színek jelentése*

Erőművi szumma term. tip. net.mérés (15p) · Nukleáris erőművek net.mérés (15p) · Barnakőszén-lignit erőművek net.mérés (15p) · Gáz (fosszilis) erőművek net.mérés (15p) · Feketekőszén erőművek net.mérés (15p) · Szárazföldi szélerőművek net.mérés (15p) · Biomassza erőművek net.mérés (15p) · Naperőművek net.mérés (15p) · Szemétégető erőművek net.mérés (15p) · Folyóvizes erőmvek net.mérés (15p) · Víztározós vízerőművek net.mérés (15p) · Olaj (fosszilis) erőművek net.mérés (15p) · Egyéb megújuló erőművek net.mérés (15p) · Egyéb erőművek net.mérés (15p)

2012. január 1. és 2017. szeptember 1. közötti adatokat gyűjtöttem be órás bontásban, ami 49680 mérést jelent. Meglepően jól látszik (3. ábra) az éves ciklus. Január, február, december több áramfogyasztás, nyáron leesik, bár az évközepén is vannak, kilengések gondolom ezek a hőségrekordos napok, illetve a decemberi hosszú ünnep karácsony és a szilveszter idején is jelentős változás figyelhető meg, de ebben az esetben visszaesés formájában, nem dolgozik senki. A 3. ábra a hitelesített tény rendszer terhelés mutatja, ennyi áramra volt igény az adott órában. Illetve az a trend is jól látszik, hogy évről évre több áramot fogyaszt a magyar nép.

Sajnos nem teljesen hiánytalanok az adatok. Általában mindegyikből van egy terv és egy tényleges adat. A tényleges adatok sokszor üresek ilyenkor a tervezet adatokkal helyettesítem őket.




*3. ábra*

*Magyarországi rendszerterhelés órásbontásban 2012 január 1 és 2017 szeptember 1 között*

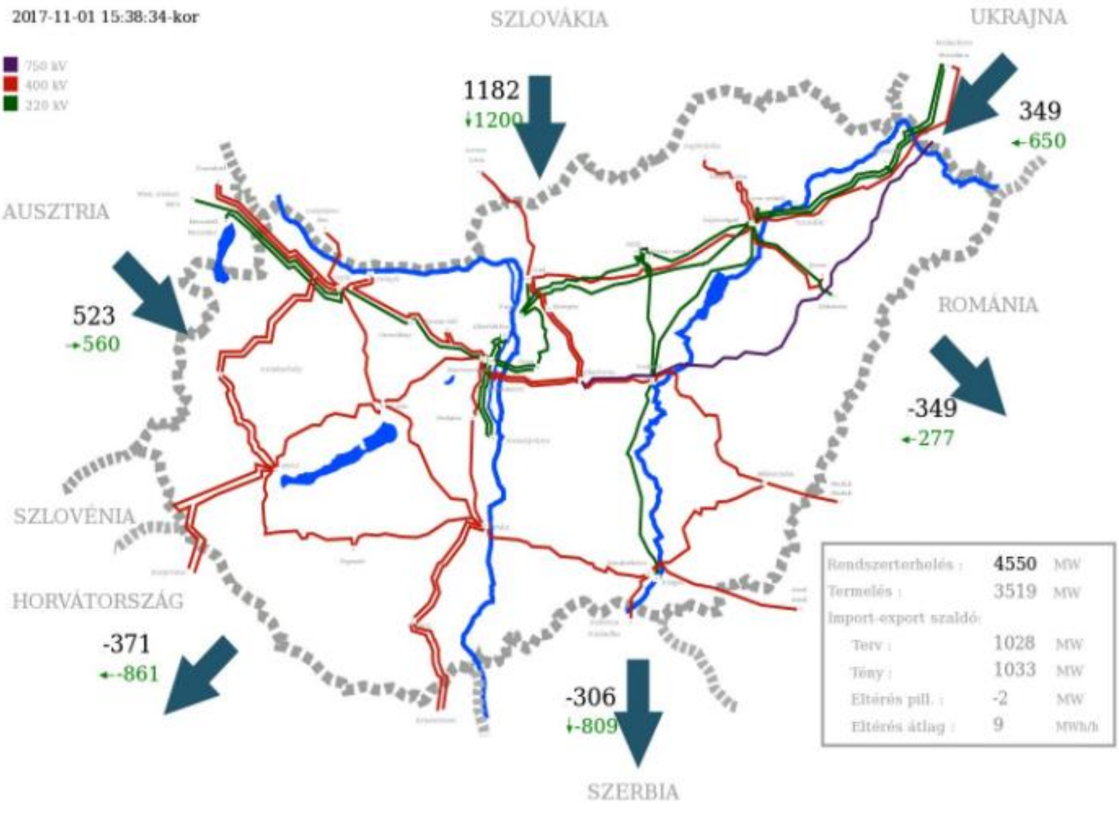
A másik nagy adatcsoport, amit le lehet tölteni a MAVIR-ról az a határmetszéki áramlások.

**Határmetszék**: a magyar és a szomszédos villamosenergia-rendszer közötti egyrendszerű vagy kétrendszerű távvezetékek, amelyek szinkron üzemmódban összekapcsolják a MAVIR és a szomszédos rendszerirányító szabályozási területeit. [6] A 4M piaci EU-s törvény kötelezi az országot, hogy minden határos Európai Uniós tagállam felé rendelkezzen hálózat hozzáféréssel. A 4. ábra szemlélteti, hogy honnan mekkora áram folyik, folyhat be. Jól látszik, hogy:

- Magyarország – Ausztria HUAT
- Magyarország – Horvátország HUHR
- Magyarország – Románia HURO
- Magyarország – Szerbia HUSR
- Magyarország – Szlovákia HUSK
- Magyarország – Ukrajna HUUK

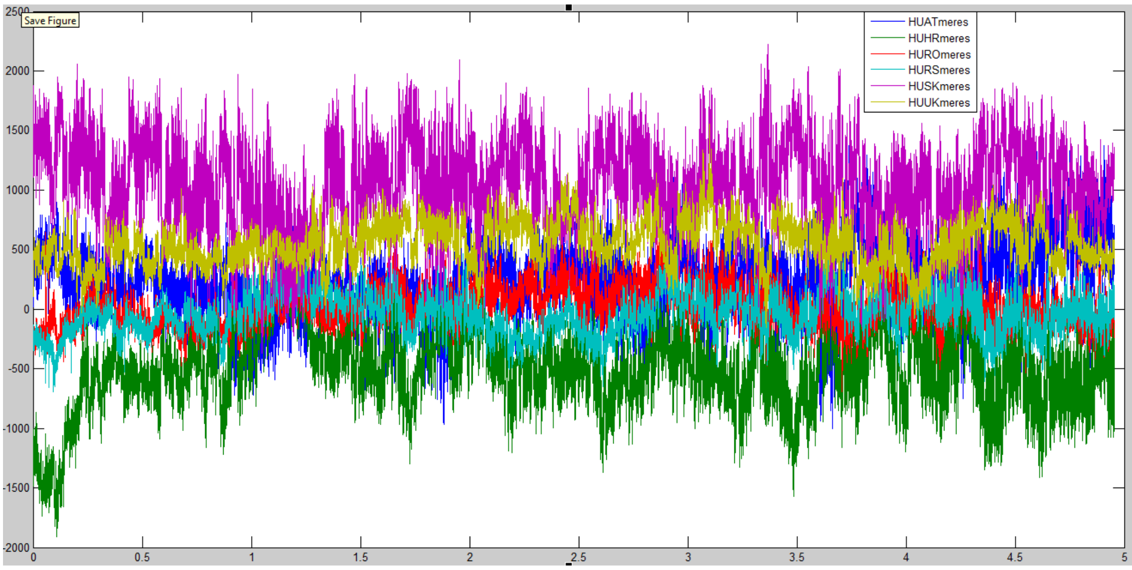
között van határmetszéki áramlás. Későbbiekben a 4 betűs rövidítéseiket fogom használni mikor rájuk, hivatkozom. Jelentős határmetszék az HUUK mert egy 750 kV egy 400 kV és 2 db 220 kV távvezetéken kapcsolódunk Ukrajnához, illetve jelentős még az ausztriai határmetszék és a szlovákiai, ezeken a pontokon több 400 kV-os illetve 220 kV-os távvezetéken kapcsolódunk. A következtetésemet alátámasztják még MAVIR honlapról beszerzett adatok, amelyek azt mutatják, hogy ezekből az országokból folyik be a legtöbb áram Magyarországra.

A 4. ábra egy olyan Magyarország térkép, ami az aktuális pillanatban - jelen esetben 2017. november 1-e 15:38-kor készült pillanatkép - mutatja a MAVIR honlapján a fennálló import-exportot és a rendszerterhelést. Éppen akkor, amikor készült a kép akkor Ausztria felöl 523 MW áram érkezett ténylegesen, míg a terv, ami zöld színnel van, 560 MW volt. Jól látszik, hogy az adott időpontban Szerbia felé áramot exportáltunk 306 MW, míg azt várta a MAVIR, hogy 809 MW lesz szüksége Szerbiának. A jobb alsó sarokban lévő táblázat összesíti a térképen látottakat még a rendszerterhelést és az erőművek össztermelését is feltünteti, így le tudjuk olvasni, hogy november 1-én 15:38-kor 4550 MW a rendszer terhelés, ami 3519 MW termeléssel és 1033 MW importtalimportál biztosítottak. Ami még látszik, hogy ebben a pillanatban egy nagyon jól becsült össz export-import szaldó volt.




*4. ábra*
*Határmetszéki áramlások pillanatnyi állapota 2017 November 1én 15:38:34kor*

Nagyon fontos megjegyezni, hogy nem Magyarország használja fel az összes import áramot, Horvátország, Szerbia és Románia rajtunk keresztül részesednek a Közép-Európában termelt áramból. [21] Az 5. ábra kaotikus színes világa mutatja be az egyes metszékekre jellemző exportot, importot.



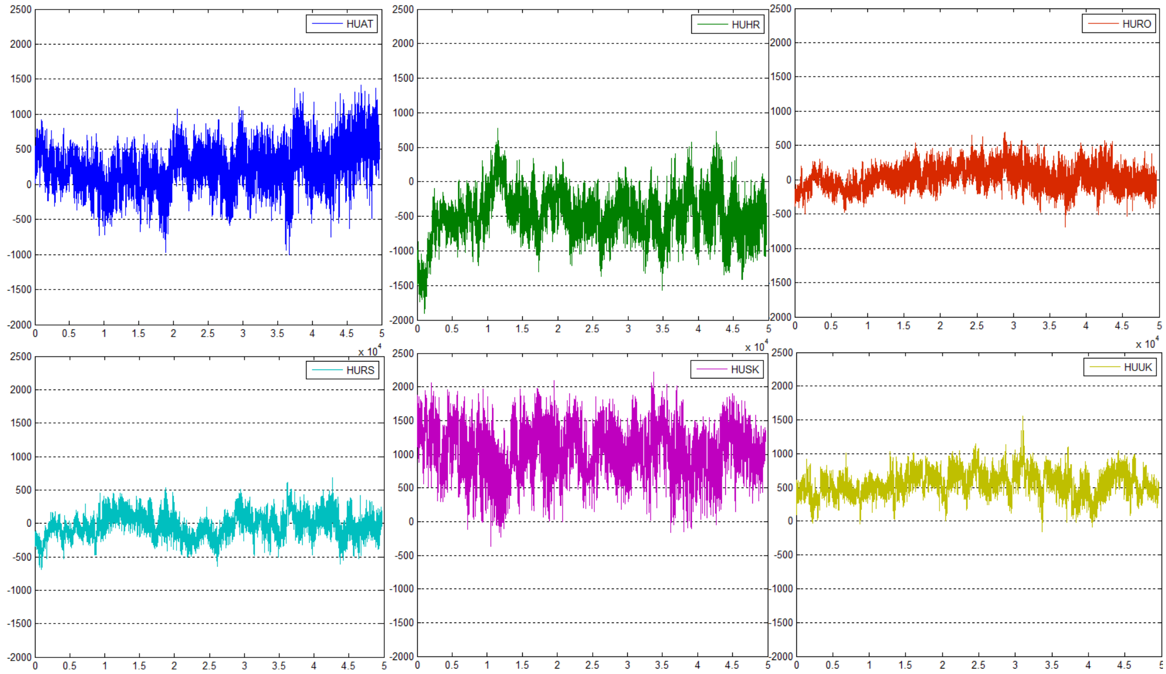

*5. ábra*

*Határmetszéki áramlások országos órás bontásban 2012 január 1 és 2017 szeptember 1 között*

Az 5. ábra egymáshoz viszonyítva mutatja be az 6 db határmetszék export-importját. Miért csak 6 db határmetszékről beszélek? Hiszen Magyarországnak 7 szomszédja van (Szlovénia a hetedik), ez nagyon egyszerű a 4. ábra szerint nincs köztünk távvezeték, amit a MAVIR csupa nullás adatok is alátámasztottak.

Már az 5. ábra is mutatja, hogy a zöld színű HUHR többnyire negatív, ami azt jelenti, hogy Horvátországba főleg exportálunk áramot. A magenta színű HUSK az esetek többségében pozitív az esetek döntőtöbbségében (eddig 88 alkalommal volt negatív) és kb. kétszerese a többi határmetszéknek, és az átlaga 1064,7 MW és a medián 1084,8 MW, minimuma -369MW, maximuma 2220,5 MW volt az elmúlt közel 6 évben.

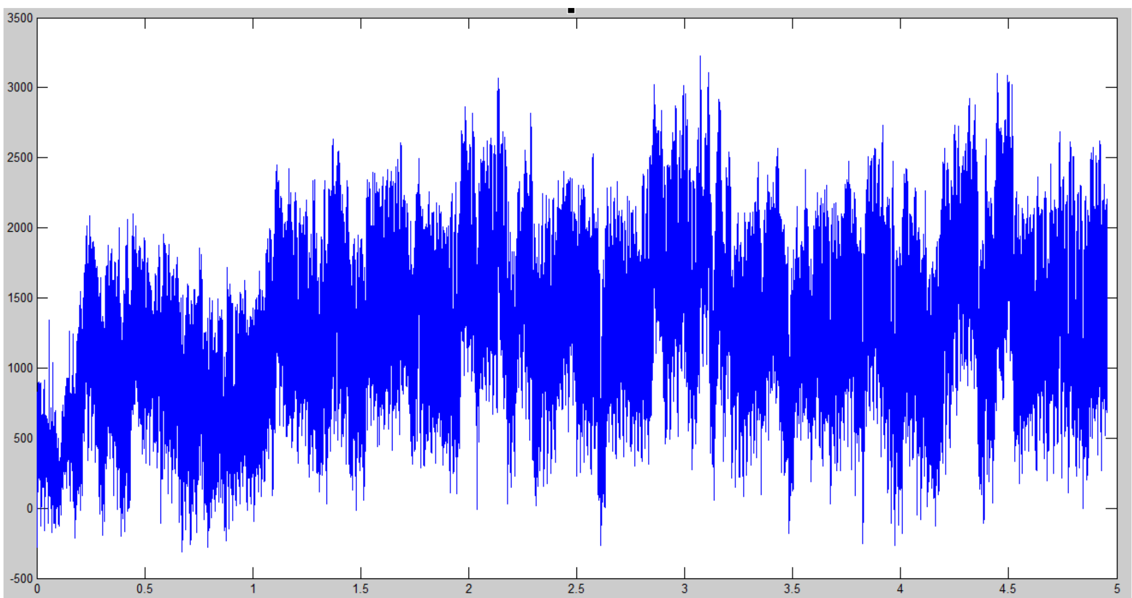
6. ábra külön mutatja bontva a határmetszékeket a jobb láthatóság kedvéért, azonosra állítottam az összes diagramon a léptékeket. Itt még jobban látszik, hogy HUSK-ból, HUUK-ból érkezik az import többsége. A HUUK határmetszéken átlagban 588,7 MW import érkezik, viszont nem mondható el, hogy csak import lenne ezen a határmetszéken. A megszerzett adatokban 55 alkalommal volt olyan óra, amikor exportált Magyarország Ukrajnába. Az utóbbi években már HUAT-ból is főleg importáltuk pedig volt olyan év mikor inkább ők vettek tőlünk. HUAT átlaga a teljes adatsorra 255,3 MW, viszont a maximuma 1419,7 MW, míg a minimuma -1005,7 MW. 9430 alkalommal exportáltunk Ausztriába áramot. HURO átlaga 37 MW míg HURS -32 MW nagyon kicsi aktivitás van ezen a két határmetszéken.




*6. ábra*

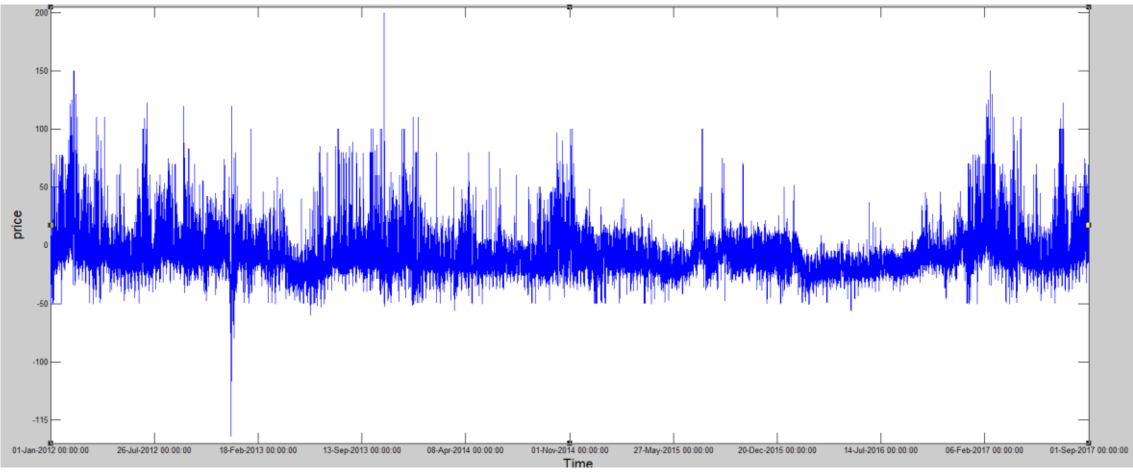
*Határmetszéki áramlások országos bontása egymás mellett azonos skálával 2012 január 1 és 2017 szeptember 1 között*

Majd összegeztem a határmetszéki adatokat és a 7. ábra lett az eredmény, aminek maximuma 3224 MW lett, ami megdöbbentő annak a fényében, hogy a tény rendszerterhelés átlagosan 4930 MW. Nagyon szabálytalannak tűnik a rendszerterheléshez képest az export-import.



*7. ábra*

*Az országos adatok alapján számított export-import szaldó*

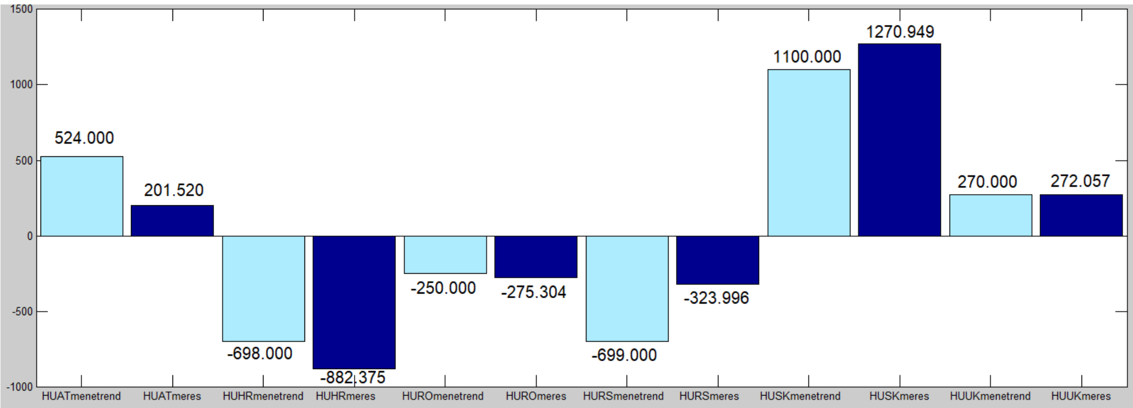



*8. ábra*

*Árak órásbontásban 2012 január 1 és 2017 szeptember 1 között*

A 8. ábra az árak alakulását mutatja be 2012. január 1. és 2017. szeptember 1 között órás bontásban, a hupx.hu-ról szereztem be az adatokat. Ami igazán érdekessé teszi, hogy nincsenek egyértelmű trendek, mint a rendszerterhelésnél. Illetve vannak nagyon kiugró értékek is mind pozitív mind negatív irányba.

A legkisebb ár -113,67 Ft volt 2012. december 26-án 5 órakor. A rendszerterhelés 3229,3 MW volt, ami kisebb, mint az átlagos rendszerterhelés, az import-export 263,5 MW volt. Ezek alapján arra tudok következtetni, hogy többet termeltünk, mint amire szükség volt, így gyorsan megpróbálták eladni az áram felesleget. Ez kétféleképpen jöhetett létre, az egyik, hogy nem jól számolták ki az ünnepek miatti keresletet, így eleve több termelés volt betervezve, mint szükséges volt, vagy az egyik nagy fogyasztó hirtelen kiesése okozhatta.




*9. ábra*

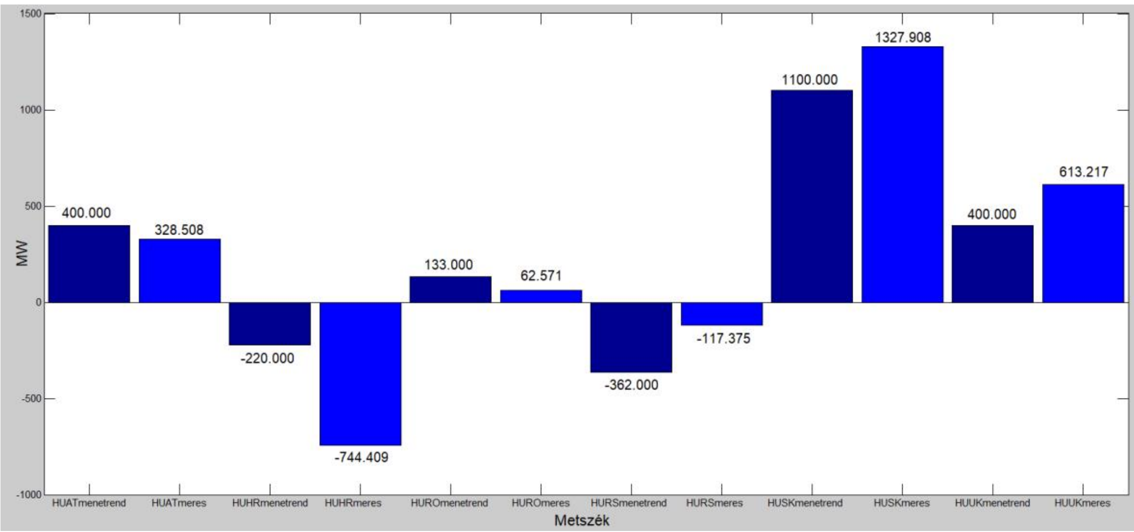
*Magyarország tervezett és tényleges export importja 2012 december 26-án 5 órakor*

Az itthoni túltermelésnél meg kell azt az esetet vizsgálni, hogy azért termeltünk-e többet mert valamelyik ország számára exportálni akartuk, de végül nem kellett neki. Ezt a 9. ábra szemlélteti határmetszékenként külön-külön. Kivonva a tervből a tényleges határmetszék forgalmat kiderült az, hogy a határmetszékek miatt összesen 15,8 MW többlet áram maradt Magyarországon, ez elhanyagolhatóan kicsi.

Már csak az az eshetőség maradt, hogy valamelyik nagy fogyasztó kiesett. Sajnos nem találtam olyan cikket a magyar online sajtóba, ami utalna erre a nagyon alacsony árra.

Sajnos sem a HUPX, sem a MAVIR honlapján nem találtam az okot, hogy miért volt ez a nagy ár változás.

A legnagyobb ár eddig 250 Ft volt 2013. október 27-én 17 órakor. Összes határmetszék eltérés -19,42 MW ennyivel többet importáltunk, ami szerintem nem sok. A 10. ábra szemlélteti határmetszékekre lebontva az akkori terv és tény import-export szaldót.




*10. ábra*
*Határmetszékek tény és terv áramlása 2013 október 27én 17kor*

Ekkor a rendszerterhelés 4545,8 MW volt, ami elég átlagosnak mondható, így a többlet fogyasztás nem okozhatta. Mint az előbb láttuk többet exportáltak, mint amennyit eredetileg terveztek.

Erre a napra is rákerestem a HUPX, MAVIR, GOOGLE keresőkben és semmi cikk, megjegyzés. Így szomorúan tudomásul vettem, hogy nem tudom a pontos okát a két nagy kilengésnek.

# 2 Előrejelzési és statisztikai módszerek

Abból indultam ki, hogy az áramtőzsde hasonlóan működik, mint a normál tőzsde, a kereslet és a kínálat határozza meg az árat, ezért tőzsdei algoritmusokat kerestem elsőnek, illetve megnéztem, hogy a matematikai tankönyvek, kiadott tananyagok milyen idősor elemzési elméleteket tartalmaznak.

A tőzsdebarát [7] honlapon van egy lista a különböző elven működő, különböző módon használható előrejelzésekről. A leírás főleg azt tartalmazza, hogy ezen elvek használata mellett mikor érdemes adni-venni. De mivel én itt most nem fogok eladni, így pontosan, úgy ahogy le vannak írva nem használhatók számomra, ezért megkerestem az elvek pontos matematikai háttérét, erre főkképpen a Számítástudományi és Információelméleti Tanszék által tartott Nagyméretű adathalmazok kezelése szeminárium anyagait használtam.

## 2.1 Idősorok

Az idősor megfigyelések egy sorozata, tipikusan időközönkénti mérés pl. naponta, óránként… A vizsgált problémában óránkénti méréseket használnak. Számít a sorrend, időbélyeg tartozik az egyes elemekhez. Ki tudjuk használni, hogy az egymást követő megfigyelések erősen korrelálnak egymással. Az idősorokat megkülönböztetjük a változók száma szerint, lehet egyváltozós pl. a hőmérséklet mérés, vagy lehet több változós pl. az időjárási adatok, páratartalom, hőmérséklet, légnyomás alapján következtetünk a várható időjárásra. Az idősor lehet stacionárius és nem stacionárius is. A nem stacionáriusra jellemzők hogy a variancia, átlag és a többi statisztikai paraméter változik az idő előre haladása során, vannak benne trendek, ciklusok. A stacionárius esetben nincsenek trendek, ciklusok, a statisztikai paraméterek állandók, az időtől függetlenek. Természetesen a stacionaritás nem ilyen fekete és fehér, létezik erős és gyenge stacionaritás is. Idősor előrejelzést nagyon sok helyen alkalmaznak. Időjárást, autóforgalmat, népességet, tőzsdét, fogyasztást is jeleznek előre. A feladatban jósolni kívánt adat az ár, ami tőzsdeszerű záróárra hasonlít.

## 2.2 Korreláció

A **korreláció** jelzi két tetszőleges érték közötti lineáris kapcsolat nagyságát és irányát (avagy ezek egymáshoz való viszonyát). Az általános statisztikai használat során a korreláció jelzi azt, hogy két tetszőleges érték statisztikai szempontból nem független egymástól. Az ilyen széleskörű használat során számos együttható, érték jellemzi a korrelációt, alkalmazkodva az adatok fajtájához. [8]

A korrelációs együttható előjele a kapcsolat irányát mutatja meg, a nagysága (0-1 közötti szám) pedig azt hogy mennyire hasonlítanak, korrelálnak egymással az adatok, az összefüggés erejét mutatja. A korreláció nem függ az adatok nagyságától, de érzékeny a mintavételezésre. [8]

Ha X és Y korrelálatlan, akkor r(X,Y) = 0, ha X és Y korrelálatlan, akkor nem feltétlenül függetlenek, de biztos, hogy nincs köztük lineáris típusú összefüggés. [8]

Az idősorok elemzésében és a jelfeldolgozásban gyakran alkalmazzák a korrelációt az összehasonlításokban.

Ha kiszámítjuk két adatsor értékkészletének korrelációját, akkor keresztkorrelációt kapunk.

Ha egy adatsort és egy időbeli eltoltjának korrelációját számoljuk így ki, akkor autokorrelációról beszélünk. A keresztkorreláció segít a két adatsor közötti összefüggés megtalálásában. Ha az egyik adatsort eltoljuk, akkor késleltetett hatások is felfedezhetők. Az autokorrelációval periódusok is kimutathatók ki az adatsorban. [8]

### 2.2.1 Autokorrelációs függvény

Definiáljuk a következőképpen az x(t) mért jel autokorrelációs függvényét: [9]

$$ACF(\tau) = \int_{-\infty}^{\infty} x(t)\, x(t-\tau)\, dt$$

Ez a függvény azt mutatja meg, hogy a jel, mennyire hasonlít önmagára (auto) ha $\tau$ idővel eltoljuk. Mivel a függvénynek sok pontja van, minden pontra megállapítjuk a hasonlóságot (szorzással) és ezt összegezzük (integrálással). Elég könnyű belátni, hogy ez a függvény akkor lesz maximális, ha az eltolás mértéke zérus, azaz $\tau = 0$. Ez következik az értelmezésből is, de abból is, hogy ebben az esetben a jel négyzetének integrálját kapjuk [9]

$$ACF(0) = \int_{-\infty}^{\infty} x(t)\, x(t)\, dt = \int_{-\infty}^{\infty} x^2(t)\, dt$$

A fenti autokorrelációs definíciót ezért általában annak normált változatával szokás helyettesíteni. Az eredetileg fent definiált függvény inkább autokovariancia függvénynek nevezik, és a zérus eltolással normált (vagy ha úgy tetszik a jel teljesítményére normált) változatát nevezik, autokorrelációs függvénynek Leginkább arról lehet megállapítani, hogy melyik függvénnyel van dolgunk, hogy a zérus eltolásra egyet mutat-e a függvény, vagy valami más értéket. Ha egyet, akkor a módosított, normált függvénnyel van dolgunk, ha eltér az egytől, vagy az eredeti első (kovariancia jellegű) definíciót használták. Mindkét esetben azonban biztos, hogy az ACF maximuma a zérus eltolásban van, és az is, hogy szimmetrikus a jobb és baloldalra (szimmetrikus függvény). Ez következik a fenti meghatározásból, hiszen elég csak felcserélni a két tényezőt és eltolni a tengelyt és ezzel máris bizonyítottuk a szimmetriát [9]

# 2.3 Regresszió

A statisztikában a regressziószámítás vagy regresszióanalízis során két vagy több véletlen változó között fennálló kapcsolatot modellezzük. A változók közötti kapcsolatok vizsgálata során, a változók kapcsolatának szorosságát (intenzitását) a korrelációval mérjük, míg a kapcsolat törvényszerűségét a regresszió írja le. A regressziós modell tulajdonságai alapján megkülönböztethetünk lineáris és nemlineáris regressziót, az adataink alapján pedig idősor, keresztmetszeti, és panel regresszió analízist. A regressziót leíró függvények sokfélék lehetnek, amelyeket a változók száma szerint és a kapcsolat milyensége alapján osztályozhatunk.[11]

## 2.3.1 Lineáris regresszió

A lineáris regresszió egyike a leggyakrabban alkalmazott statisztikai eljárásoknak, a trendelemzés alapja. Ha több folytonos változó lineáris kapcsolatban van egymással, akkor az egyik csoport segítségével (magyarázó változók) előre jelezhetjük a másik csoport értékét (eredmény változók). Szükségünk van a függő és független változó kiválasztására, de ez nem jelent oksági kapcsolatot! [10]

Az összefüggés segítheti a kapcsolat megértését és legfőképp releváns előrejelzéseink lehetnek. [12]

Nagyon jó kiindulási alap, hiszen általa meg tudjuk határozni, hogy a változók függetlenek vagy összefüggnek-e, ha igen menyire vannak hatással egymásra, milyen a kapcsolat erőssége, predikció. Ezen okokból nagyon gyakran lehet vele többváltozós tőzsdei előrejelzésekben találkozni. [11] A trend meghatározás gyakori módszere.

**Kétváltozós eset**:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Y az eredmény változó

X a magyarázó változó

$\beta_0$ és $\beta_1$ a regressziós együtthatók, míg az $\varepsilon$ véletlen változó.

A lehető legkisebb hibájú becslés a cél. A hibáról feltételezzük, hogy független a magyarázó változótól, és átlaga nulla. A becslés csak tökéletes kapcsolat esetén lenne hibamentes.[12]

**Elaszticitás:** A rugalmasság mérőszáma. Azt fejezi ki, hogy az X magyarázó változó 1% változása hány %-os változást okoz az eredmény változóban. [13]

$$E(Y;X) = \beta_1 \frac{X}{\beta_0 + \beta_1 X}$$

## 2.3.2 Legkisebb négyzetek módszere

$$\sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2$$

A **legkisebb négyzetek módszere** a mérések matematikai feldolgozásában használt eljárás. Nevét arról kapta, hogy az eltérések négyzetösszegét igyekszik minimalizálni. [15][14]

A módszer érzékeny a nagyon kilógó adatokra. Egy kilógó adat az egész eljárás eredményét megváltoztathatja, hamis képet adva az adatsorról. Különböző statisztikai tesztekkel szűrik az adatsort, hogy ne maradjanak benne mérési hibák. A kilógó adatokat elhagyják, vagy a kívülállókra kevésbé érzékeny módszerekkel alternatív becsléseket végeznek. Ilyen például a súlyozott regresszió, amiben a kívülálló adatok súlyát, és ezzel befolyását is csökkentik. [15]

## 2.4 Mozgóátlagolás

A mozgóátlagolás elméleti összefoglalója [16] és [17] honlap alapján készült. A mozgóátlagok számítása az idősorok hosszabb távú elemzésének egyik legegyszerűbb módja. Csak annyit tűzünk ki célul, hogy átlagolással kiszűrjük a durva, egészen rövid távú ingadozásokat. Végrehajtása matematikailag igen egyszerű, jól alkalmazkodik az idősor jellegéhez. Hátránya hogy az idősor megrövidül, és jól kell, megválasztani az átlagolandó tagok számát különben torzít. Simítja az idősort, de az extrém értékek erősen befolyásolják. Többnyire páratlan számú elemet átlagolunk, mert az adott tagnak az átlagát úgy számítjuk, hogy vesszük az adott tagot és előtte egy utána ugyan annyi számú további tagot. Ebből pedig már egyértelműen következik, hogy a mozgó átlagoknak van egy igen lényeges tulajdonsága, mégpedig az, hogy nem lehet minden egyes elemhez mozgóátlagot számítani, így a megfigyelt idősor eleje és vége elvész. Gyakran használják negyedéves trendek kiszámítására a devizapiacon vagy a tőzsdén. Egyszerű, nem igényel nagy gazdasági vagy matematikai tudást az alkalmazása. Készítenek egy rövid távú mozgóátlagot és egy hosszú távút, ha rövid távú felülről lefelé halad át a hosszú távún, akkor várhatón csökkenő trend fog következni, így érdemes eladni, ha a rövid távú alulról metszi a hosszú távút, akkor emelkedésre számítunk, így veszünk.

A mozgóátlagnak a 3 legismertebb fajtája a következő: egyszerű (vagy aritmetikai), exponenciális, és súlyozott. Módszerek közötti különbség a súlyozásban van, az egyszerű mozgóátlag minden egyes elemet egyforma súllyal vesz figyelembe, az exponenciális és a súlyozott pedig a frissebb adatokat nagyobb súllyal értékeli.

### 2.4.1 Egyszerű mozgóátlag

Az adatsor egyszerű számtani átlaga, ahol azonban - amint azt már fentebb említettünk nem szabad elfelejteni, hogy az egyszerű mozgóátlag késve követi a folyamatokat, így a trend megváltozását is késve jelzi. Mivel miden adatot egyforma súllyal vesz figyelembe, nem számol azzal a ténnyel, hogy a frissebb adatok jelentősége nagyobb.

Az intervallum megválasztását illetően annyit érdemes még megjegyezni, hogy „kisebb ablak" megválasztása esetén a mozgóátlag gyorsabban és érzékenyebben reagál a változásokra

A tőzsdei gyakorlatban azonban a mozgóátlagokat nem úgy számítják, hogy a kiválasztott adat környezetében végzik az átlagolást, hanem a kiválasztott adat előtti adatokra. Az átlag számítása tehát nem centrikus, hanem visszatekintő.

Következtetések:

- Amennyiben a vizsgált adat árfolyama átlépi a mozgóátlag értékét, vételi jelet kapunk. Ha az árfolyam a mozgóátlag alá süllyed, úgy eladási jelről beszélünk.
- Technicista ökölszabály szerint, amikor valamilyen "jól megválasztott" rövidebb intervallumra számolt mozgóátlag egy hosszabbat alulról metsz, vásárolni kell - lefelé történő metszés esetén pedig eladni.

### 2.4.2 Exponenciális simítás

Egyszerű modell. Nem feltételezünk az idősoros adatokban sem trend, sem szezonális hatást. Ez is egyfajta mozgó átlagolás azzal a különbséggel, hogy egy adott pont exponenciális simításának értékéhez elegendő a múltban közvetlenül előtte levő értékeket ismerni. Minden korábbi pont visszafelé haladva egyre kisebb súllyal számít. A súly értéke 0 és 1 között lehet. Az 1-hez közeli értékek nagy súlyt kapnak az aktuális pont kiszámításában, kevésbé simítanak, míg 0 közeli súlyok erős simítást végeznek.

#### 2.4.2.1 Egyszeres exponenciális simítás

Az exponenciális simítás módszer alapváltozata. Feltételezi, hogy a megfigyelt érték egy állandó körül ingadozik. A simítást $\alpha$ -val jelöljük. Az alfa azt mondja ki, hogy mennyire szeretnénk simítani, azaz mennyire vesszük figyelembe az elkövetett hibát. Az alfa helyes megválasztása kulcsfontosságú.

Az egyszeres exponenciális simítás egyenlete: $\bar{y}_{t+1} = \bar{y}_t + \alpha\,(y_t - \bar{y}_t) = (1 - \alpha)\bar{y}_t + \alpha y_t$ ahol:

$$0 \leq \alpha \geq 1$$

$\bar{y}_t$ a t időszakra vonatkozó trend szerinti érték

$y_t$ a t időszakra vonatkozó tényleges megfigyelés [20]

Ha az $\alpha = 1$ akkor az eljárás egyáltalán nem simít, ha közel 0 akkor jobban simít.

Az egyszeres exponenciális simítás csak egy időszakra ad érdemi előrejelzést. Feltétele hogy az idősorban ne legyen tartós tendencia. Nagyobb alfa esetén jobban követi a tényleges adatokat, míg kisebb alfával hosszabb távú ingadozást lehet meghatározni. Ezért mikor egy idősor feltehetőleg lineáris trendet követ, nem célszerű ezt a módszert használni. [18]

### 2.4.2.2 Kettős exponenciális simítás

A kétszeres simítás nem más, mint az egyszeresen simított sor ismételt egyszeres simítása, így figyelembe tudjuk venni a trend értékét is. A fentebbi képletben a simított és az előre jelzett értékek megegyeztek, most viszont a kétszeres simítás miatt nem ugyan azok. S lesz a simított érték és felső indexbe teszem a simítás számosságát. Ennek megfelelően az egyszeres simítás képlete: $S_t^{(1)} = (1 - \alpha)S_{t-1}^{(1)} + \alpha y_t$

A kétszeres simítás egyenlete pedig: $S_t^{(2)} = (1 - \alpha)S_{t-1}^{(2)} + \alpha S_t^{(1)}$ [18] [20]

Az így készített előrejelzés szintén torzított lesz ugyan, de a torzítás már lényegesen kisebb, mint az előzőekben, nagy α-k esetén pedig elenyésző lehet. Kisα-k esetén ezúttal is erős, nagyobbak esetén gyengébb a simítás. Ezzel a módszerrel tetszőleges hosszú és számú időszakra tudunk előrejelzést készíteni. Már kevés adattal is használható. [18]

## 2.5 Autoregresszív folyamatok

Az autoregresszív folyamatok és az ARMA illetve az ARIMA modell fejezet az Üzleti prognózisok idősoros modelljei Prof. Dr. Besenyei Lajos, Domán Csaba könyve [18] alapján és Balogh Péter és Nagy Lajos Ökonometria című könyvét [11] felhasználva készült, amelyek elérhetők a tankönyvtárban és az autoregresszíve moving average model című wikipedia [23] cikket dolgoztam még fel.

Az AR folyamatokkal általában azokat az idősorokat modellezhetjük, amelyekről feltételezhetjük, hogy jelen idejű értékeik alakulásában a közvetlen múlton kívül a véletlen hiba is beleszól.

**P rendű autoregresszív modell**

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

Ahol $c$ egy konstans és $\varepsilon_t$ a hibatag, $a_i$-k pedig a modell paraméterei.

A folyamat azért kapta a „regresszív” elnevezést, mert a fenti kifejezés nagyon hasonlít a többváltozós regresszióra. Az autoregresszív elnevezés pedig onnan származik, hogy ebben az esetben egy olyan regresszióról van szó, amelyben $X_t$-t a saját késleltetett értékeivel magyarázzuk.

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

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

Általánosabban:

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

Ennek segítségével az autoregresszív modell másik felírása a következőképp alakul:

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

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

Ahol:

$$a(B) = 1 - \sum_{i=1}^{p} a_i B^i$$

# 2.6 ARMA modell

A sztochasztikus módszerek a véletlennek jelentős hatást tulajdonítanak, ez a modellezésben fontos szerepet játszik. Ezek története Yule autoregresszív (1927), illetve Slutsky mozgóátlagolású modelljéig (1937) nyúlik vissza. Wold már 1954 alkalmazta a mozgóátlagolású modellt valós adatokra, illetve ő dolgozta ki a vegyes ARMA-modellek használatát (1954). [11]

A sztochasztikus idősor elemzés legegyszerűbb, és a gazdasági gyakorlatban leginkább elterjedt ágát jelentik, melyeket az AR- és az MA-modellek egyesítéseként állították elő. Az ARMA folyamatok jelentősége az utóbbi évtizedekben megnőtt, s a tapasztalatoknak köszönhetően matematikailag jól kezelhetőek és általánosíthatóak. Az ARMA folyamatok paramétereinek meghatározását, vagyis az illesztést általában empirikus idősorok alapján végezzük, figyelembe véve, hogy a létrehozandó modell a valóságnak csak megközelítését adja, törekednünk kell arra, hogy ez az illesztés minél pontosabb legyen. [18]

Fentebb már definiáltam az AR részét is. A MA-ról is volt már szó, viszont most szeretnék egy másik definíciót adni a MA-ra.

MA modell q-rendű képlete: [23]

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

Ahol θ modell paraméterei, ε itt is fehérzajnak felel meg, a μ pedig az átlag, amit gyakran tekintenek 0-nak. Az autoregresszív modellnél már használt backshift operátor segítségével így néz ki a mozgó átlag modell: [23]

$$x_t = \mu + \varepsilon_t + \sum_{i=1}^{q} \theta_i \; B^i \, \varepsilon_t$$

Ezek után már csak az ARMA(p,q) modell definícióját kell ismertetnem. A fentebb bemutatott autoregresszív és mozgó átlag modell segítségével a következőképen írható fel a két modell alap egyenletéből: [23]

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

Az autoregresszív jelző arra utal, hogy $X_t$ részben saját véges múltjára vonatkozó lineáris regressziójaként írható fel. A mozgóátlag jelző pedig azt fejezi ki, hogy a lineáris regresszió „hibatagja” az $\varepsilon_t$ fehérzaj mozgó átlaga, azaz a jelen és a véges múlt lineáris kombinációja. [18]

A modellezés első lépésében azt kell meghatároznunk, hogy megfigyeléseink milyen p és q rendű folyamatból származhatnak. Ezt a fázist a kiinduló modell felírásának, azonosításnak, identifikációnak nevezzük. Ennek módszere az, hogy meghatározzuk a ténylegesen megfigyelt idősor jellemzőit és összehasonlítjuk azokat az elméleti idősorok megfelelő jellemzőivel, majd megkeressük, hogy melyik elméleti modellel mutat idősorunk leginkább hasonlatosságot. Második lépésként a kiválasztott modell $\vartheta$ és $\alpha$ paramétereit becsüljük. A becslési folyamatok bonyolultabbak a korábbi fejezetekben tárgyalt alapeseteknél, a becslőfüggvények ritkán oldhatók meg explicite a paraméterekre, ezért többnyire iterációs eljárásokra van szükség. [18]

Az idősorokra vonatkozó egyik legárnyaltabb, legösszetettebb elemzés a Box és Jenkins által kidolgozott ARIMA modellekkel lehetséges. A számítások nehézkessége miatt az ARMA-modelleket csak nagyon kevesen használták egészen a számítógépek széles körű elterjedéséig, illetve amíg Box és Jenkins meg nem fogalmazta azokat a kritériumokat, amelyekkel minden idősorra meghatározható egy konkrét típusú és fokú ARIMA modell. Ez a modellezés elsősorban a sűrű megfigyeléssel rendelkező változók (pl. árfolyamok) nehezen megragadható szabálytalan ingadozásait, időbeli lefutásait próbálja leírni, alapesetben csupán saját múltbeli értékeik és a véletlen törvényszerűségei alapján. [11]

Az idősorok elméletében és alkalmazásában az autoregresszív és mozgóátlag-(ARMA) folyamatok jelentősége az utóbbi évtizedekben rendkívül megnőtt. Ez annak köszönhető, hogy az ARMA sztochasztikus folyamatok matematikai szempontból jól kezelhetők, és a folyamatok egy elég általános osztályát képviselik. Emiatt a gyakorlatban előforduló, stacionárius viselkedést mutató véletlen folyamatok nagy része jól közelíthető az ARMA folyamatokkal. [11]

# 2.7 ARIMA

Az ARIMA(p,d,q) modell az ARMA(p,q) modellt egy integráló, I(d) résszel egészíti ki. A d paraméter lehet nem egész szám is. Így az ARIMA modellnek 3 fő része van. AR(p), I(d), MA(q). A p, d, és a q nem negatív egész számok. ha valamelyik nulla, akkor az a rész kiesek, pl. így lehet belőle ARMA. Nem stacionárius adatokra használják. Legáltalánosabb, megengedi a stacionárius transzformációkat (differenciálás, logaritmizálás). Ebben a modellben p= autoregresszió rendje d= differenciák száma (nem szezonális különbségek) q= mozgóátlag rendje

Ismertebb ARIMA felparaméterezések

ARIMA (0,1,0)=véletlen bolyongás

ARIMA (1,1,0)=módosított elsőrendű autoregresszívmodell

ARIMA (0,1,1) nem állandó=egyszerű exponenciális simítás

ARIMA (0,1,1)=állandó egyszerű exponenciális simítás a növekedés

ARIMA (0,2,1) és (0,2,2) nem állandó=lineáris exponenciális simítás

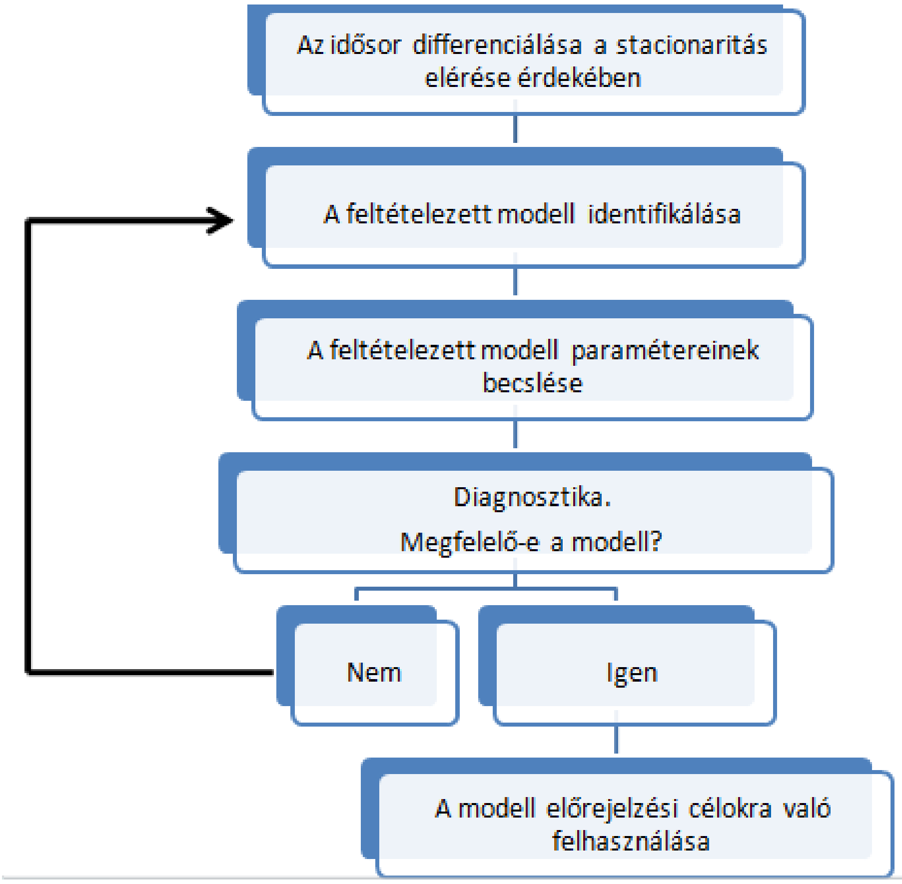
A „vegyes” modell -ARIMA (1,1,1)

Ha d=0, akkor az ARMA modellt kapjuk, ha d-szer deriválunk egy ARIMA(p,d,q) modellt, akkor is az ARMA modellhez jutunk. Amennyiben d=0, akkor az idősor stacionárius, amennyiben d=1, akkor nem stacionárius.

Vannak olyan stacionárius idősorok, amelyek esetében az autokorrelációs függvény lassan cseng le, és két távoli megfigyelés között is összefüggés mutatkozik. Ilyenkor két eset lehetséges. Az idősor egységgyököt tartalmaz, de mivel nagyon közel van az egyhez, ezért az egységgyök teszt téves eredményt mutat. A másik lehetőség, hogy az idősorban nincs egységgyök valóban, de hosszútávú korrelációkat tartalmaz, erre nem illeszkedik jól a szokványos ARIMA modell. Amennyiben újra differenciálnánk az idősort, az sem lenne megoldás, mert túl differenciált lenne az idősor. Granger és Joyeux (1980), valamint Hosking (1981) javasolta ennek a problémának az áthidalására, hogy a d differenciálási paraméter legyen nem egész érték.

Ha d értéke 0 és 0,5 közé esik, akkor az idősor hosszú távú függőségeket tartalmaz. Ha d értéke nagyobb, mint 0,5, akkor az idősor nem stacionárius, ha d=0, akkor az idősor egy fehérzaj folyamat.

Az idősor elemzéshez mindenképpen szükség van bizonyos előfeltevésekre, modellezésre, mivel itt nincs mód több független mintát venni, mint ahogyan a statisztika más területein; itt csak egy idősorunk van. Az ARIMA-modellezés lényege, hogy az idősorok leírására kidolgozott autoregressziós és mozgóátlagoláson alapuló eljárásokat (amelyek pedig azt mutatják, hogyan függ a megfigyelés mostani értéke az előző időszakok véletlen tényezőitől) egy közös modellbe építjük be. A Box és Jenkins által ajánlott általános módszer, ARIMA modellek alkalmazása idősorelemzésre, prognosztizálásra és ellenőrzésre az idősor elemzés Box-Jenkins módszertanaként lett ismert.




*11. ábra*

*ARIMA modellezés Box-jenkins féle módszerrel [11]*

Az ARIMA modellezés kiindulópontja annak megállapítása, hogy a vizsgálni kívánt idősorunk stacionárius-e, illetve, ha nem, akkor az, hogy alkalmas transzformációval stacionáriussá tehető-e. Ezzel eldöntöttük azt, hogy az adott idősorhoz illeszthető-e ARIMA modell, ha igen milyen (d) dimenzióval (fokkal) rendelkezik. A következő kérdés annak megválaszolása, hogy milyen típusú ARMA modell illesztésével próbálkozzunk, illetve, milyen legyen az autoregresszivitás (p) és/vagy, a mozgóátlagolás (q) rendje. Erre a kérdésre a választ a tapasztalati, vagy a transzformált idősor ACF és PACF értékei (autokorrelációs- és parciális autokorrelációs együtthatók) alapján adjuk meg. A modellezés e fázisát, modellazonosításnak (identifikációnak) nevezi a szakirodalom. Ezután a modellezés lépései alapvetően megfelelnek a már ismert lineáris regressziós modellezésnek. A választott modell paraméterbecslése után a modell ellenőrzése következik.

### 2.7.1 Stacionaritás biztosítása

Az idősor grafikonja, az autokorrelációs és parciális autokorrelációs függvények grafikonjai alapján valószínűsíteni lehet, hogy milyen rendű és fokú ARIMA folyamat illesztése vezethet eredményre.

A Box-Jenkins modellezés első lépésében az ARMA(p, q) folyamat paramétereit, vagyis q -t és p -t kell meghatározni. A fázis lényege tehát megtalálni a tapasztalati idősort legjobban leíró elméleti idősort. A munkában nagy segítségünkre lehet, ha a megfigyelt adatokat az idő függvényében ábrázoljuk. Ekkor szembesülhetünk azzal a ténnyel, hogy az idősorunkban milyen trend van. Amennyiben lineáris trenddel van dolgunk, akkor elegendő az adatsorunkat differenciálni. A differenciált adatokból készített ábránk már remélhetőleg nem mutat további trendet. Ám amennyiben mégis, ismételt differenciálásra van szükség. Mivel a gazdasági idősorok általában tartalmaznak trendet, így igen valószínű, hogy szükség lesz a differenciálásra. A tapasztalatok alapján azonban kétszeri differenciálással a trend problémája általában megszüntethető.

A vizsgált adatok időbeni ábrázolásán kívül egy másik ábra segítségével is el lehet dönteni, hogy szükséges-e a differenciálás. Ez a korrelogram (autokorrelációs függvény, ACF). Az egymást követő megfigyelések között fennálló összefüggések megállapítása az idősorok korrelációs struktúrájának leírását jelenti, mely az autokorrelációs és a parciális autokorrelációs együtthatók számításával történik.

Az autokorrelációs együtthatók becsült értékei, az Y idősor k időegységgel késleltetett értékei közötti lineáris korrelációs kapcsolat szorosságát mérik. Ha autokorrelációs együtthatók értékei lassan csökkennek, vagy majdnem lineárisan, ez indokolja a differenciaképzést. A megfelelő fokú differenciák elérését az autokorrelációs együtthatók gyors csökkenése jelzi. Ha az autokorrelációs együtthatók értékei a szezonális komponens hatásának megfelelően hullámoznak, akkor a szezonhatást először ki kell szűrni. A differenciálás elvégzése után, elkészítve a következő korrelogrammot, ismét csak a csökkenés mértékét kell vizsgálni.

Az idősorok stacionaritásának a vizsgálatára szokás használni a Dickey-Fuller tesztet. A Dickey-Fuller teszt úgynevezett „unit root" jelenlétét vizsgálja egy autoregresszív modellben. Unit root akkor van jelen a modellben, ha $\rho = 1$, ez a nem stacionárius eset.

## 2.7.2 Modell beazonosítása és együtthatók becslése

Az autokorrelációs függvény felrajzolása nem csak abban segít, hogy az idősorunkat stacionáriussá tudjuk tenni, hanem abban is, hogy az mozgóátlagolású (MA) tag q -fokára egy kezdeti becslést tudjunk adni. Ehhez a korrelogram alakját kell csak megvizsgálni. Ha a korrelogram q -nál kisebb értékeknél nem mutat semmilyen határozott alakot, míg q -tól nagyobb értékekre nulla, akkor a késleltetéseknek q -t kell választani. Vagyis pl. elsőrendű mozgóátlag (MA(1)) folyamat esetén kizárólag az első érték nem nulla, az összes többi az.

Az autoregresszív (AR) tag p kezdeti értékének eldöntésében a korrelogram helyett egy másik függvényt használunk, ez a parciális autokorreláció függvény (PACF). A PACF a magasabb rendű autokorrelációk hatását megtisztítja az alacsonyabb rendű autokorrelációk hatásaitól. A parciális korrelogram értéke egy bizonyos késleltetés után nulla körül fog mozogni. Ez a késleltetés lesz a p kezdeti értéke. Azaz egy elsőrendű autokorrelációs (AR(1)) folyamatnál a parciális korrelogram első eleme nem nulla, a többi mind nulla közelében marad.

A részleges autokorreláció függvény (PACF) az autokorreláció függvényből számítható ki. Az AR együtthatókat határozza meg, így a szignifikáns értékei alapján becsülhető az illesztendő modell AR tagjainak száma. Kiszámítása a Yule-Walker egyenletek megoldásával történik.

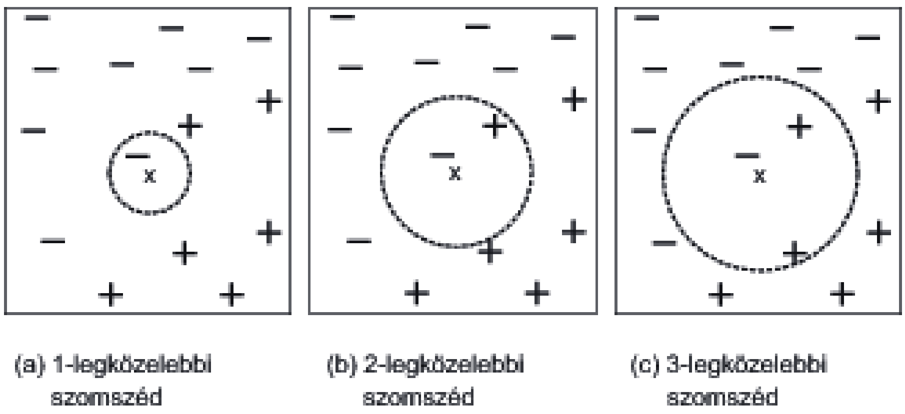
### 2.7.3 Diagnosztikai ellenőrzés

Ebben a fázisban ellenőriznünk kell, hogy megfelelően illeszkedik-e a modellünk az adatokhoz, vagyis a modellünk helyességét. A legfontosabb diagnosztikák a következők: a reziduumokra, illetve azok négyzetére vonatkozó Ljung-Box és Box-Pierce statisztikák, a normalitásra, a csúcsosságra és a ferdeségre vonatkozó tesztek.

A tesztek azt mérik, hogy a megbecsült ARIMA-modell mennyire illeszkedik az idősorra. Azt a hipotézist teszteljük, hogy a becsült reziduumok vajon tényleg normális fehérzajként viselkednek-e.

## 2.8 Legközelebbi szomszéd

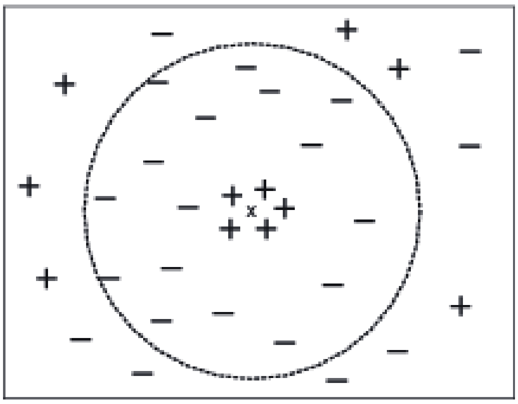
A legközelebbi szomszédok használatának indoklását legjobban a következő mondás szemlélteti: " *Ha valami úgy totyog, mint egy kacsa, úgy hápog, mint egy kacsa és úgy néz ki, mint egy kacsa, akkor az valószínűleg egy kacsa*." A legközelebbi szomszéd osztályozó minden egyes esetet egy adatpontként reprezentál egy $d$ -dimenziós térben, ahol $d$ az attribútumok száma. Egy adott teszteset esetén a meghatározott szomszédsági mértékek valamelyikével kiszámítjuk annak közelségét a tanulóhalmaz összes többi adatpontjához. Egy adott $z$ eset $k$ -legközelebbi szomszédja azt a $k$ pontot jelenti, amelyek a legközelebb vannak $z$ -hez.




*12. ábra*

Az 12. ábra a körök középpontjában lévő adatpont 1-, 2- és 3- legközelebbi szomszédját szemlélteti. Egy adatpontot a szomszédjainak osztálycímkéje alapján osztályozunk. Abban az esetben, ha a szomszédoknak egynél több címkéje van, az adatpontot a legközelebbi szomszédok többségi osztályához rendeljük hozzá. Az ábrán az adatpont 1-legközelebbi szomszédja egy negatív eset. Ezért az adatpontot a negatív osztályhoz rendeljük hozzá. Ha három legközelebbi szomszéd van az ábrán látható módon, akkor a szomszédság két pozitív és egy negatív esetet tartalmaz. A többségi szavazási sémával az adatpontot a pozitív osztályhoz rendeljük hozzá. Holtverseny esetén az adatpont osztályozásához véletlenszerűen választhatjuk valamelyik osztályt.

A fent leírtak $k$ helyes megválasztásának fontosságát hangsúlyozzák. Ha $k$ túl kicsi, akkor a legközelebbi szomszéd osztályozó a tanulóadatokban jelenlevő zaj miatt hajlamos lehet a túlillesztésre. Másrészt viszont, ha $k$ túl nagy, akkor a legközelebbi szomszéd osztályozó rosszul osztályozhatja a tesztpéldányt, mivel a legközelebbi szomszédok listája a szomszédságtól messzi adatpontokat is tartalmazhat 13. ábra.



*13. ábra Nagy k esetén*

## 2.9 Előrejelzés kiértékelése

Miután megválasztottuk az előrejelzési módszert meg kell róla győződnünk, hogy elfogadhatóan jó módszert választottunk. Ehhez érdemes az adathalmazt egy tanító és egy teszt részre külön bontani. A tanító halmaz alapján kiválasztjuk az előrejelzési modellt, majd előrejelzést adunk a teszt halmaz értékeire. Már csak a modell pontosságának kiértékelése van hátra, ha számunkra még nem elfogadható az adott feladathoz, akkor finomítunk a modellen és a paraméterezésén, és újra kiértékeljük.

A tanított modell pontosságának a mértéke a hiba mértéke. Minden kiértékelésnél azt mondjuk meg, hogy az adott paraméterekhez milyen hiba tartozik. Így megteremtődik a különböző megoldások összehasonlíthatósága és a visszacsatolás a javításhoz. A cél a hiba minimalizálása. Több fajta hibaszámítási módszert ismerünk. [11] [22]

### Abszolút százalékos hiba

APE (Absolute Percentage Error – abszolút százalékos hiba)

$$APE_t = 100 * \frac{|\, y_t - \hat{y}_t \,|}{y_t}$$

Ezt a mutatót minden egyes t értékre külön kell kiszámítani, ezért csak az egyes becslések százalékos hibáját adja meg. [11]

### Átlagos százalékos abszolút hiba

MAPE (Mean Absolute Percentage Error – átlagos százalékos abszolút hiba)

$$MAPE = \frac{\sum_{t=1}^{n} APE_t}{n}$$

Az APE átlagolásával számítjuk ki.

### Átlagos négyzetes hiba

MSE (Mean Squared Error – átlagos négyzetes hiba)

$$MSE = \frac{\sum_{t=1}^{n} (\hat{y}_t - y_t)^2}{n-k}$$

ahol $\hat{y}_t$ az eredményváltozó becsült (előrejelzett), $y_t$ pedig a tényleges értéket jelenti, k a becsléshez felhasznált paraméterek száma, n a megfigyelések számát jelöli.

RMSE (Root Mean Squared Error): a MSE négyzetgyökét kell venni

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} (\hat{y}_t - y_t)^2}{n-k}}$$

## 2.10Modellek összehasonlítása:

| Módszer | Főbb jellemzői | Előnyei | Hátrányai | Milyen időtávra |
|---|---|---|---|---|
| Lineáris regresszió | | Egyszerű | Pontatlan, ha sztochasztikus az adat | |
| Mozgó átlagolás | Az idősor dinamikus átlagát állítja elő | Egyszerű<br>Kevés adattal is működik | Pontatlan | rövidtáv, középtáv |
| Exponenciális átlagolás | Különböző súllyal veszi a vizsgált időszak résztendenciáit | Rövidtávon megbízható | Nehéz meghatározni a súlyokat | Rövidtáv, középtáv |
| Legközelebbi szomszéd | Hasonlóság vizsgálat | Már 2 nap után tud működni | Hasonlóság mértékét nehéz definiálni, a jó működéshez sok adat kell | |
| ARIMA | ARIMA(p,d,q)<br>p= autoregresszió,<br>d= differenciák száma,<br>q= mozgóátlag rendje | Finoman paraméterezhető, jó paraméterezés esetén jól jelez előre | Nem stacionárius folyamatokra, sok adat szükséges | Csak rövid távú előrejelzésre |

# 3 Különböző algoritmusok vizsgálata

A fentebb leírt algoritmusokat, statisztikai eljárásokat kipróbáltam a beszerzett adatokon. Ezek segítségével szeretném meghatározni, hogy melyik adattól mennyire, függ az ár, hogy ha csak az árat használom fel, akkor milyen pontossággal jósolható az ár, hogy a többi adatsort mennyire tudom beleépíteni az egyes módszerekbe, egyáltalán megéri-e belevenni őket. A magyar áramtőzsdét taglaló fejezetben már elkezdtem vizsgálni egy-egy kiugró adatot. Az egyik említett kiugró adat december 26-ára esik így úgy gondoltam, hogy érdemes lenne megvizsgálni úgy is az adatokat, hogy az ünnepeket kivesszük vagy csak a hétköznapokat vesszük.

Így hát ezen ötletek próbálkozások leírása lesz a következő fejezet.

## 3.1 Korreláció

Az adatok és az ár korrelációi a MATLAB corrcoef függvény segítségével:

*corrcoef (arak_without_nan (25:49680,1), arak_without_nan(1:49680-24,1)) = 0.6601*

Az árak 24 órával eltolt keresztkorrelációja

*corrcoef (arak_without_nan, HUATmeres_without_nan ) = 0.2457*

*corrcoef (arak_without_nan, HUHRmeres_without_nan) = 0.0216*

*corrcoef (arak_without_nan, HUROmeres_without_nan) = 0.1315*

*corrcoef (arak_without_nan, HURSmeres_without_nan) = -0.0149*

*corrcoef (arak_without_nan, HUSKmeres_without_nan) = 0.2308*

*corrcoef (arak_without_nan, HUUKmeres_without_nan) = 0.2190*

*corrcoef (arak_without_nan, ImportExportTeny_without_Nan) = 0.3662*

*corrcoef (arak_without_nan, RendszerTerheles) = 0.6202*

A matlab korrelációs mátrixokat ad vissza. [8] N valószínűségi változó (X1, ..., Xn), korrelációja egy n × n-es mátrix, amiben az i,j-edik elem corr(Xi, Xj). A korrelációmátrix szimmetrikus, mert Xi és Xj korrelációja megegyezik Xj és Xi korrelációjával, a mátrix átlójában 1-es állnak, mert a változók önnön korrelációi itt találhatók. Pl. egy 2 x 2 es mátrix $\begin{matrix} corr(x_i x_i) & corr(x_j x_i) \\ corr(x_i x_j) & corr(x_j x_j) \end{matrix}$

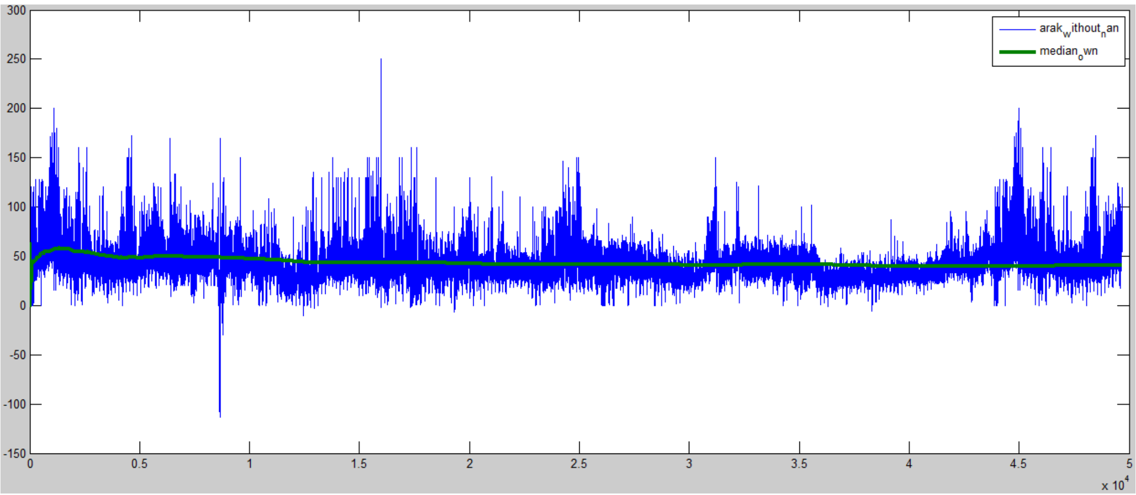
Így a matlab két bemenet esetén két 1-est és két egyforma értéket ad válaszként, úgy gondoltam felesleges információ a teljes mátrixot megtartanom.

A magyar áramtőzsde fejezetben már taglaltam, hogy kb. 30% importból tudja Magyarország fedezni az áramszükségletét, és érdekes módon az árak és az import korrelációja 36%-os, míg a rendszerterhelés 62% korrelál. Ez nem azt jelenti, hogy a rendszerterhelés 62%-ban határozza meg az árat. A korreláció az adatok közti összefüggések feltárására való, ez azt is alátámasztja, hogy míg az import-export korrelációja egyben 36% addig az egyes határmetszékek korrelációja összeadva biztos, hogy nem ennyi. Ez a korrelációs elemzés arra is nagyon jó volt, hogy lássam, hogy a HUHR és a HURS határmetszék hatása annyira jelentéktelen, hogy nem igazán korrelál az árral, és hogy a HUSK a legnagyobb forgalmú határmetszék nem a legkorrelálóbb.

## 3.2 Statisztikai algoritmusok

Készítettem egy pár előrejelzést statisztikai alapon, hogy a későbbiekben legyen mivel összehasonlítanom az eredményeimet, legyen valami, aminek a viszonyában jobbak vagy rosszabbak az aktuális algoritmusaim. Ezek az időjárás jós algoritmus alapján születettek, amilyen idő volt ma olyan lesz holnap is.

Első ilyen algoritmusom egy egyszerű medián kereséses algoritmus volt. Megkapta az első napot, mint adatbázist és utána a következő napra az addigi napok mediánját javasolta, mind a 24 órára. 14. ábra lett az eredmény, aminek meglepően jó MSE értéke lett, pontosan 444.5427 Ft hiba keletkezett azt az egy napot leszámítva a teljes adathalmazon. Jól látszik, hogy az elején mikor még kevesebb az adat követi az adatokat majd beáll egy viszonylag stabil értékre.



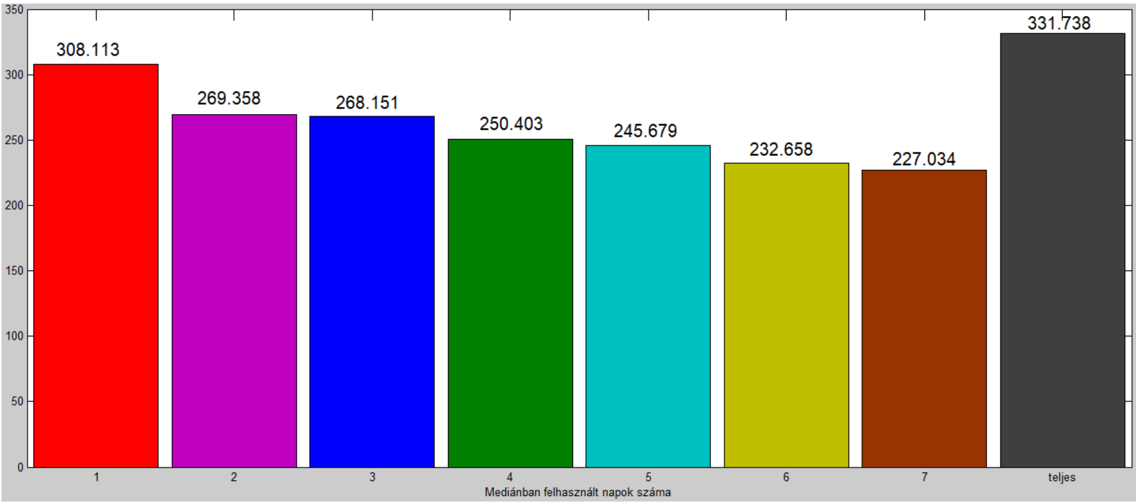

*14. ábra Mediánnal jósolt árak*

Ebből kiindulva kísérleteztem azzal mi van akkor, ha több napot kap meg alap adatnak és utána próbál előre jelezni. Ha 30 napos alapadattal dolgozik, akkor az MSE=438.0842 lett, ami jobb, mint ha csak egy napból indulnánk ki.

Ennek a medián képzési eljárásnak az a problémája, hogy nem követi az árak trendjét, így átalakítottam úgy hogy az előző napnak vegye a mediánját, így az MSE = 415.7456

Ezek után kipróbáltam mi van akkor, ha több napnak vesszük a mediánját. Kipróbáltam 2-től 28 napig tartó intervallumon a medián algoritmust és azt jött ki, hogy a legjobb eset az, ha 8 napnak veszem a mediánját, amikor is MSE = 360.3312. Az összes többi eset 363.4675 és 396.0762 között teljesített.

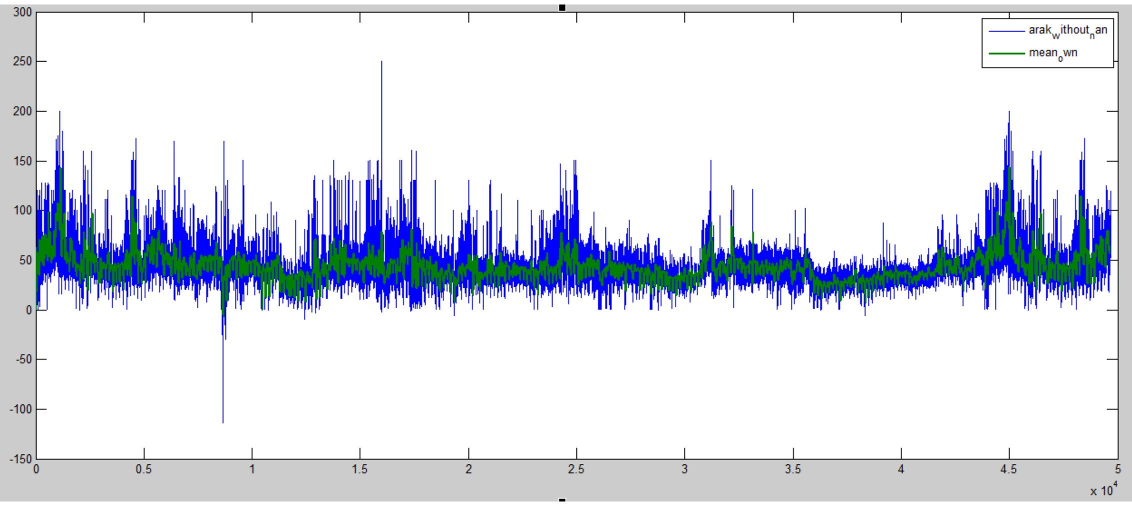
Egy kicsit módosítottam még az algoritmusom kialakítottam úgy hogy mindig az adat órai elemek mediánját vegye, így egyszer nem azt a lineáris vonalat kapom, mint eddig a napok, illetve ezt is lehet intervallummal módosítani. Nagyon jól teljesített, hiszen ha a teljes meglévő adathosszon néztem az aktuális óra mediánját, akkor az előre jelzett értékek négyzetes hibája 331.7381 lett. Ezek után még megnéztem, hogy milyen lesz az előrejelzés, ha 1 napból vagy 7 napból veszem. Jó eredménnyel zárultak ezek a tesztek ezért kiterjesztettem a kísérletet, a 14. ábra szemlélteti az eredményt.




*15. ábra Óránkénti medián előrejelzés MSE eredményei*

Az hogy óránként veszem a mediánt nem pedig egy árat adok meg nagyon sokat javított az algoritmuson.

A másik nagyon egyszerű statisztikai módszer, amit kipróbáltam az az átlagolás. Mindig az utolsó ismert napi átlagot jósoltam a következő napra. Míg nincs hirtelen változás addig nagy hibát nem vét, illetve 1 nap késéssel követi trendet is, ha van. Az átlagos négyzetes hibája MSE = 440.9133 lett (16. ábra).




*16. ábra Előző nap átlagával jósolunk előre.*

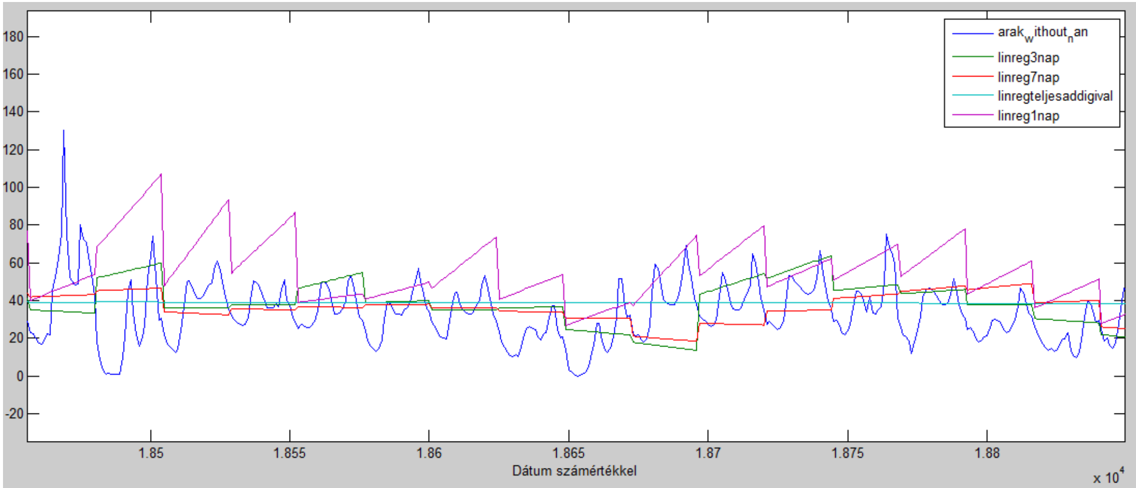
És ennek a párját a teljes adatsor átlagát jósolom a következő napra algoritmust is kipróbáltam, aminek az MSE = 454.6605 lett. Ez rosszabb mint eredmény, mint ha csak az előző napból jósolnék. Ha mindig a teljes adatsor átlagát veszem, egy idő után már nem követi a trend változását, vagy nagyon lassan, és így le lesz maradva az algoritmus, míg végül beáll egy értékre.

Ebből úgy tűnik, hogy rövidebb távot érdemesebb felhasználni a statisztikai előrejelzéseknél, így kipróbáltam az átlagolós algoritmust úgy hogy különböző hosszúságú átlagok segítségével jósoltam a következő napot. Ha 7 nap átlagát vettem (ilyenkor minimum 7 nap kell adatbázis szinten az algoritmusomnak) akkor az MSE = 357.9306 lett, ami hírtelen javulás az eddigi eredményekhez képest. 3 nap átlagolásával az eredmény egy kicsit rosszabb az MSE = 376.0104 lett. Kipróbáltam 8-tól 13-ig mindegyik átlagolási hosszal és rosszabb lett mind, mint a 7 napos átlagolás. A 14 napnak lett ezek után a legjobb eredménye MSE = 368.0844 majd tovább próbáltam egyesével növelve az átlagolt napok számát, és egy romlást tapasztaltam egészen a 21 napig, amikor egy kicsivel jobb lett az MSE = 375.8215 viszont ez rosszabb, mint a 14 vagy a 21 napos átlag viszont jobb, mint a 3 napos átlag.

Úgy tűnik, hogy van egy 7 napos ciklus az árakban.

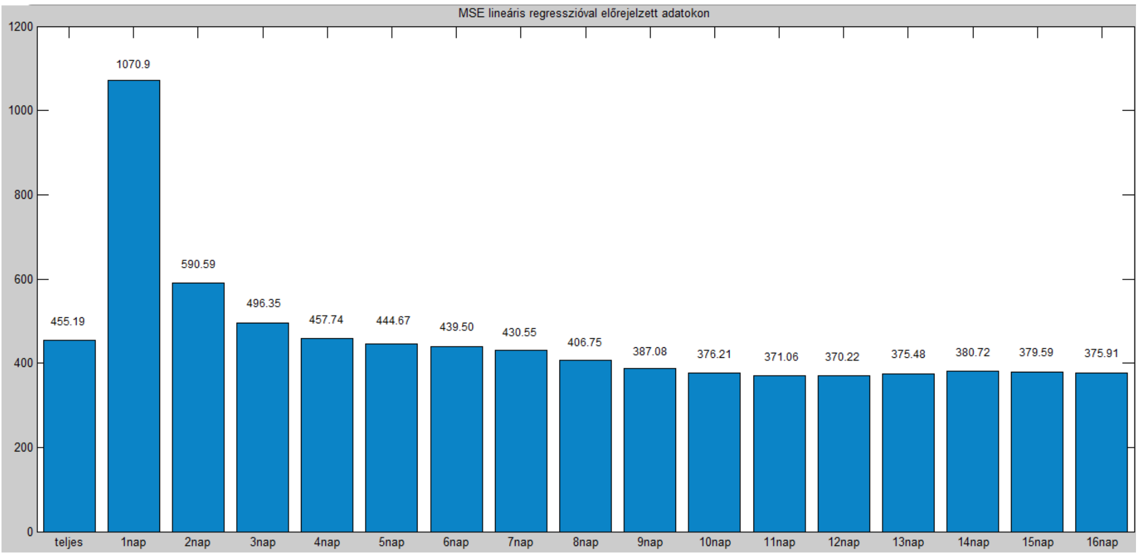
Az átlagolásnál is kipróbáltam az óránkénti átlagolást. Már az, hogy mindig a teljes adatsor átlagával jósolunk egy órát az is jobb, mint egy konstans meghatározása, az egésznapra. Ekkor a MSE = 337.8545. Itt is variáltam az átlagolandó ablak hosszával. 1 napos átlagolással a MSE = 305.2335, ami csökken egészen 8 napig amikor az MSE = 213,7337 a 7 napos ablak is hasonló értéket produkált ami pontosan MSE = 213.9501 volt majd növekedés következett.

Az első statisztikai fogalom, amit leírtam az a lineáris regresszió volt. A matlabnak erre is van kész megoldása, már csak jól kellett paramétereznem. Készítettem egy olyan kis scriptet, aminek meg kell adni az adatot, amit felszeretnénk dolgozni, illetve az idő tengelyt, azt hogy mennyi adatot szeretnénk alapból megadni az algoritmusnak, illetve hogy mennyi nap regressziójából számolja a következő napot.




*17. ábra Különböző regressziók, egy random kiemelt 16 nap hosszú szakaszon*

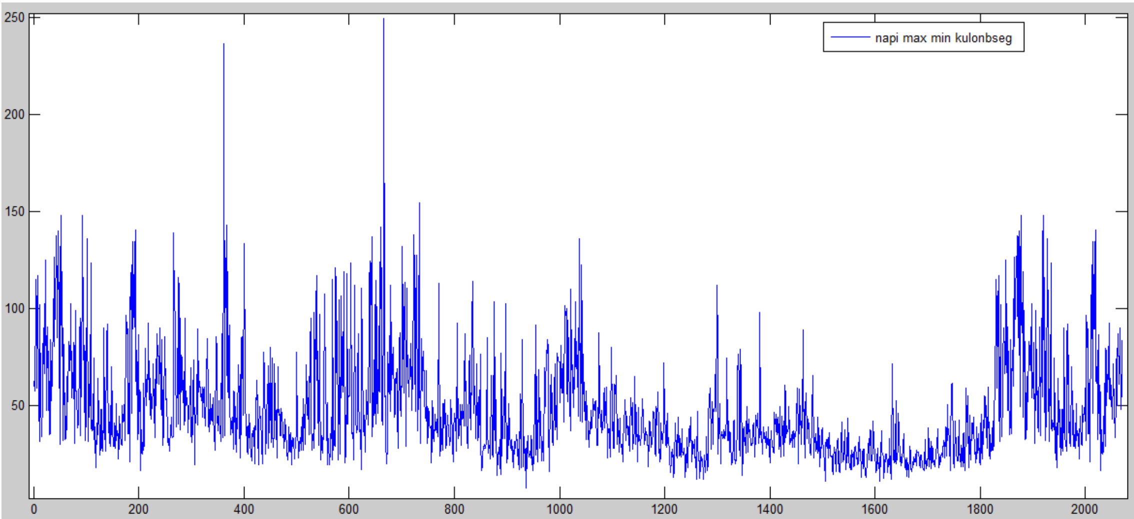
Végigpróbáltam 1 nap hosszú regressziótól a 16 hosszú regresszióig az összes lehetőséget, illetve azt, hogy ha az előtte lévő összes napot bent hagyom, akkor milyen eredményt kapok. A legérdekesebb az 1, 3 és a teljes regresszió lett, illetve még a 7 napot választottam ki bemutatás céljából (17. ábra).




*18. ábra MSE a lineáris regresszióval számolt eseteken*

A 18. ábra mutatja be egymáshoz képest az egyes lineáris regresszió hosszok használó algoritmusok milyen négyzetes hibával (MSE) jelzik előre az adatsorunkat.

Meglepően rossz lett, ha csak az utolsó napnak vettem a lineáris regresszióját és azt jósoltam előre a következő napra. Alapból itt azt vártam, hogy kb. az előző nap átlagát fogja jósolni egy kicsit módosított értékekkel, mivel jól látható, hogy az árnak van egy napi ingása és hajnalban a legolcsóbb, míg késő délután magasabb lesz az áram ára. Többször is megnéztem az 1 illetve 2 napos intervallumot és az volt a fura hogy a *polyfit* által jósolt meredekség és konstans sokkal nagyobb az 1 napos változatnál, ezt azzal tudom kapcsolatba hozni, hogy 1 napon belül túl nagy az árban az ingás (19. ábra).



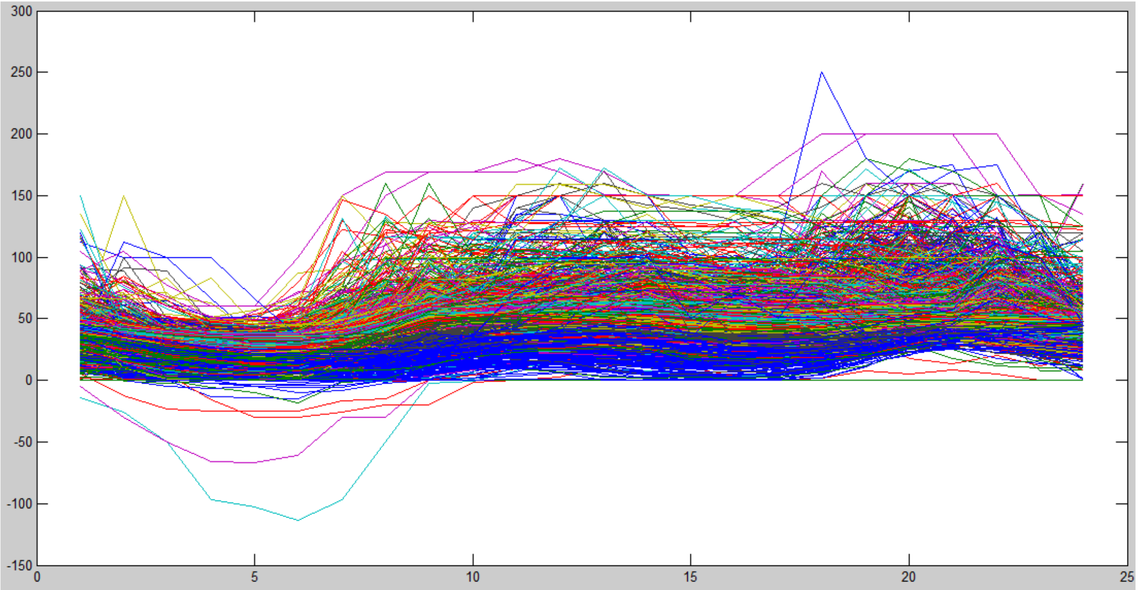

*19. ábra Napi maximum és minimum árak különbsége*

Viszont így jól látszik (17. ábra) az 1 napos regressziós előrejelzésnél hogy, a napi árnak van egy emelkedő tendenciája a késő délután, esti órákra. A 3 illetve 7 napos regressziónál is látszik, még hogy az esti órák drágábbak, nekik is van még egy szemmel is jó látható meredekségük, viszont már nem esnek bele abba a hibába, hogy a hírtelen kiugrásokat lekövessék.

Én azt vártam a lineáris regressziótól, hogy jobb lesz, mint a konstans áras átlagolós algoritmusom, hiszen az nem veszi figyelembe a napon belüli ár ingást, viszont azt kellett tapasztalnom, hogy rosszabbak lettek a regressziók. Míg az átlagolásnál a teljes adathosszú és az előző napból származó átlag volt rosszabb csak 400-nál, itt csak a 9 nap hosszú átlagolástól esett be 400 alá a hiba.

Eddig úgy tűnik, hogy x nap óránkénti mediánja vagy átlaga verhetetlen kicsi hibát produkál.

Kíváncsi lettem, hogy tényleg ennyire hasonlítanak egymásra az egyes órák, az átlagolás miért rosszabb a mediánnál, ezért kirajzoltattam úgy az adatokat, úgyhogy a napokat „egymásra” rajzoltam. 20. ábra lett az eredmény. 2070 nap van az adatbázisban és jól látszik, hogy 5 órakor a napok értéke 50 és 0 Ft között alakult a leggyakrabban, viszont vannak elvétve olyan esetek, amik nagyon kilógnak. Ezek a nagy kilógások rontják el az átlagolást, viszont a mediánra nem hatnak mert elszórva találhatóak az adatsoron.




*20. ábra Napok egymásra rajzolva*

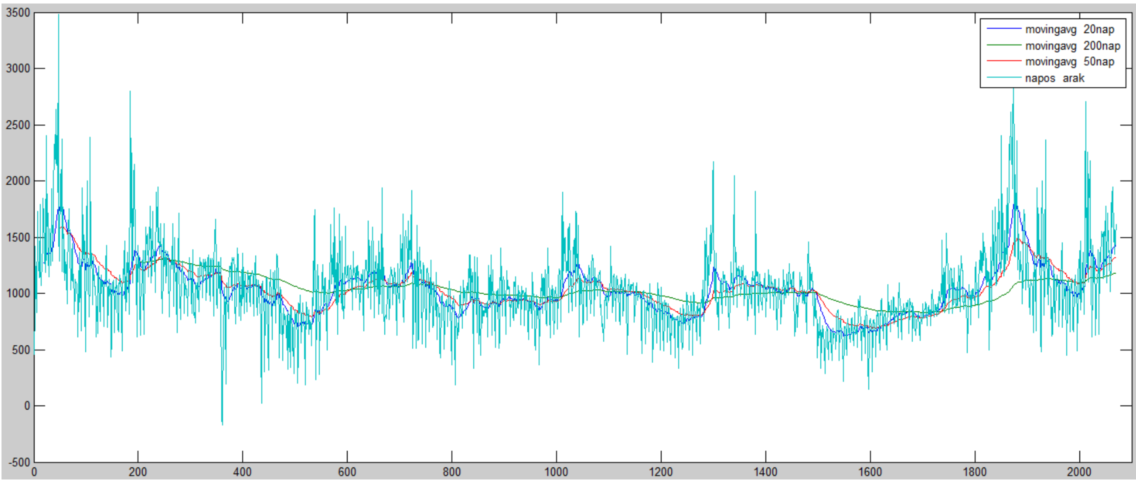
## 3.3 Mozgóátlagolás (MA)

A fentebbi algoritmusoknak a problémája hogy nem tudnak mit kezdeni a kiugró adatokkal, nem képesek előrejelezni, hogy hírtelen változás fog beállni az árakban.

A tőzsdén a mozgóátlagot is használnak vételi és eladási jelzésként. Az elektro tőzsdén valójában nincs a szó szoros értelmében vett adás és vétel, én ezen elméleteket a hírtelen változások előrejelzésére szeretném felhasználni. A tőzsdén 3 féle mozgó átlag ablak hosszat szoktak használni. 20, 50 és 200 periódusra vonatkozó egyszerű mozgóátlaggal figyelik az árfolyamot. Minél hosszabb periódusú mozgóátlag tör át az árfolyamon annál jelentősebb a jelzés.

Az első próbálkozás után kiderült, hogy így órás bontásban lévő árakra nem ad semmilyen értelmes ábrát, ha több napos átlagolást engedünk rájuk, kb. végig a napok felénél vannak az átlag vonalak ebben az esetben. Vagy órás felbontásban kell az átlagolást is csinálni vagy napos felbontásra kell hozni az árakat.

A napos felbontást próbáltam meg elsőnek, de így sem lett jobb sajnos túlságosan is mozgékony az ár. (21. ábra) Nagyon kevés olyan eset van, ahol napokon keresztül tudna úgy menni legalább a 200 napos átlag, hogy ne metssze az árakat.

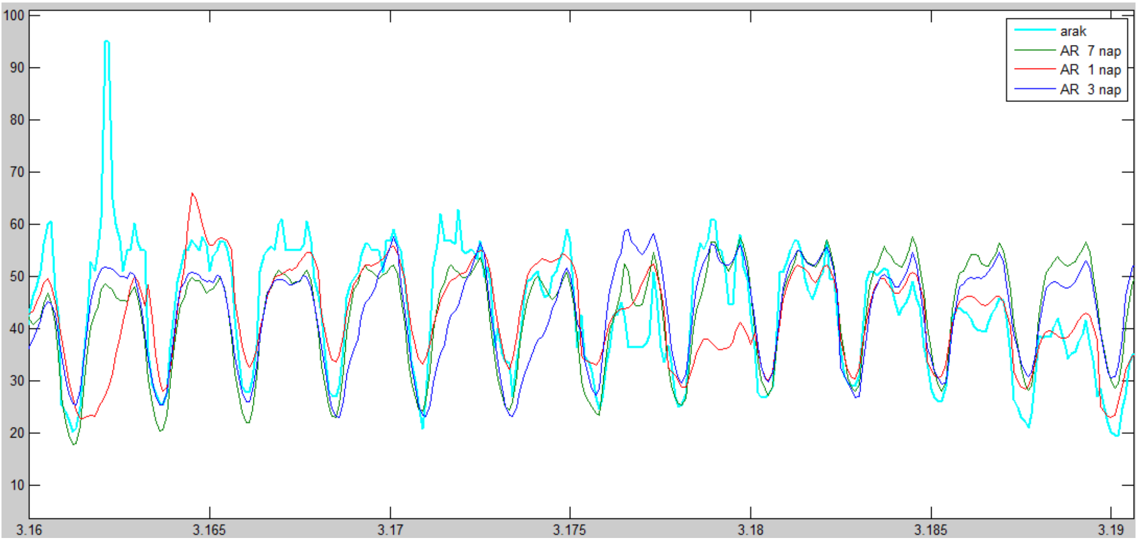



*21. ábra Mozgóátlagok a napos summázott adatokon*

Azért van, egy pár eset ahol szembetűnően elmászik az áraktól a 200 napos átlag, és mikor felülről metszi, utána tényleg nagyobb ár következik, mikor alulról akkor tényleg csökken az ár. Így ezen árak halmazán úgy tudom ezt az alkalmazást elképzelni, hogy a metszéshez még egy kikötést kell tennünk, hogy milyen idő intervallumban előtte nem volt metszés. Ezt úgy értem, hogy csak akkor vesszük figyelembe az aktuális előrejelzést, ha már legalább 20 napja nem metszette az adatsor, egyébként figyelmen kívül hagyjuk.

# 3.4 Autoregresszív folyamatok (AR)

A matlabnak van egy AR függvénye, ami az autóregresszív modellt definiálja. Mivel jól látható az árakon egy 24 órás ciklus. ezt kihasználva kísérleteztem az AR algoritmussal, illetve a medián, átlag algoritmusokban szerzett következtetéseket próbáltam még ki miszerint egy 7 napos ciklus van az árakban.




*22. ábra Autoregresszív folyamattal különböző előrejelzések eredményei*

A lineáris regressziónál is használt 1, 3, 7 napos esetek használtam itt is.

AR (1 nap) mse = 286.8185

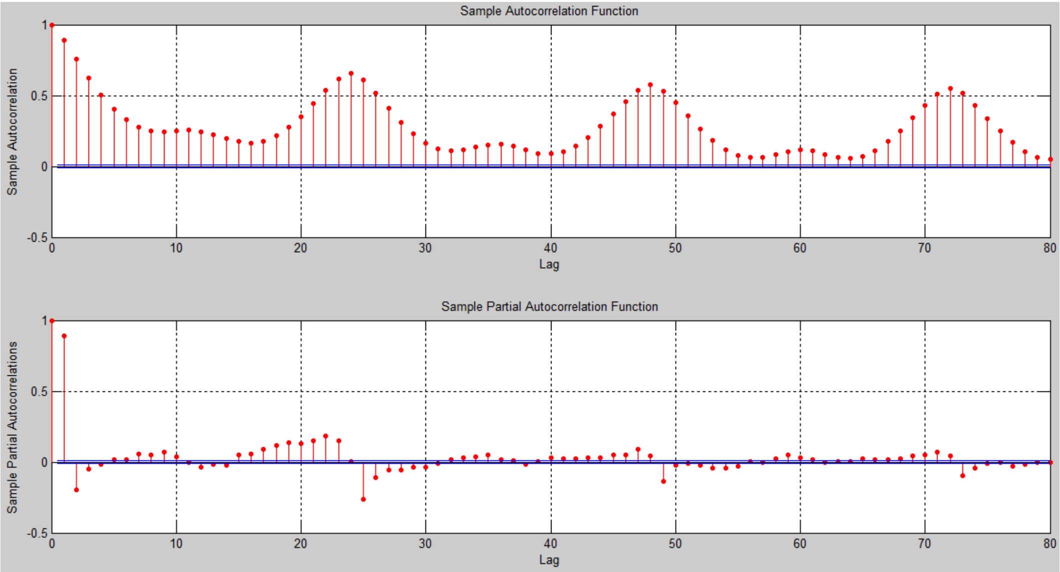
AR (3 nap) mse = 285.785

AR (7 nap) mse = 230.3533

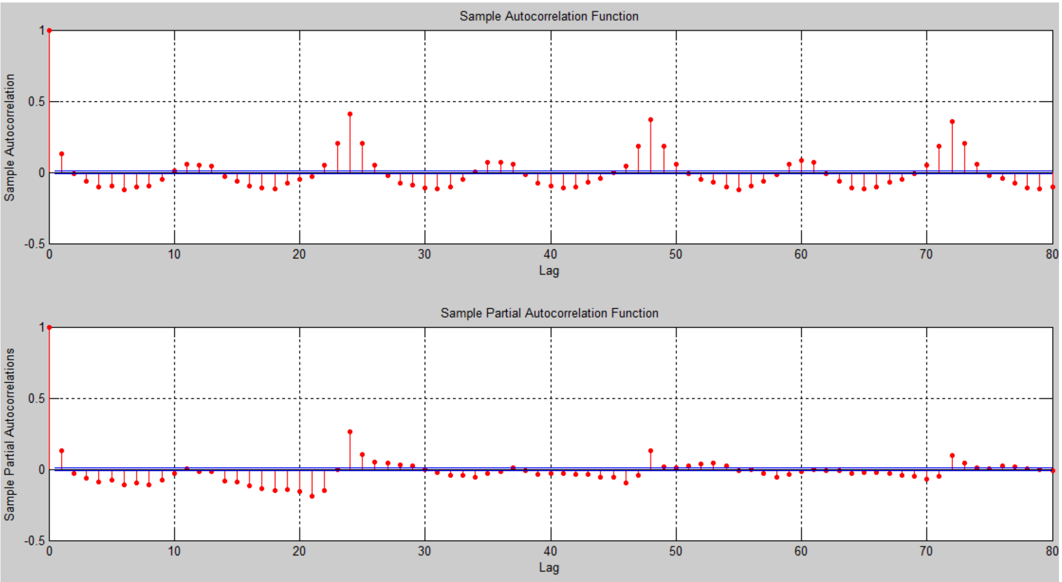
Azt lehet látni (22. ábra) hogy a 7 nap AR-je lett a legjobb, viszont ha ránézünk, az ábrára látszik, hogy a 7 napos változat nagyon lassan követi a trendet, míg az 1 napos azonnal leköveti. Viszont az 1 napos változat amiatt, hogy gyorsan követi a változásokat, egy kiugró adat után a következő napot nagyon rosszul jósolja meg, ahogy a 22. ábra 2 napja is mutatja, előtte az ár (világos kék) kimagaslóan nagy volt ezért az (piros) 1 napos regresszió magasabb értéket jósolt mint a többiek. Sajnos a 3 illetve 7 napos változatnak is van problémája, főleg a trendtől való lemaradás, ahogy a 22. ábra utolsó 3 napja is mutatja, míg az 1 napos regresszió már kisebb árakat jósol következőnek, közelít a valósághoz, addig a 3 illetve 7 napos változat még ugyan azt a magas árat jósolja, mint előtte lévő napokon.

## 3.5 ARIMA/ARMA

Elsőnek el kellett döntenem, hogy stacionárius-e az adathalmazom. A magyar áramtőzsde fejezetben a 7. ábra mutatja be az árakat. Valamiféle éves trendet lehet benne felfedezni és szemre növekszik is az ár. Mivel szemre nem tudtam eldönteni, hogy kell-e differenciálni az adatsort ezért megnéztem az eredeti adatokon az ACF és PACF értékeket 23. ábra ezt a két diagramot mutatja be.




*23. ábra ACF és PACF az eredeti adatokon*




*24. ábra 1 differenciált árakon a ACF és a PACF*

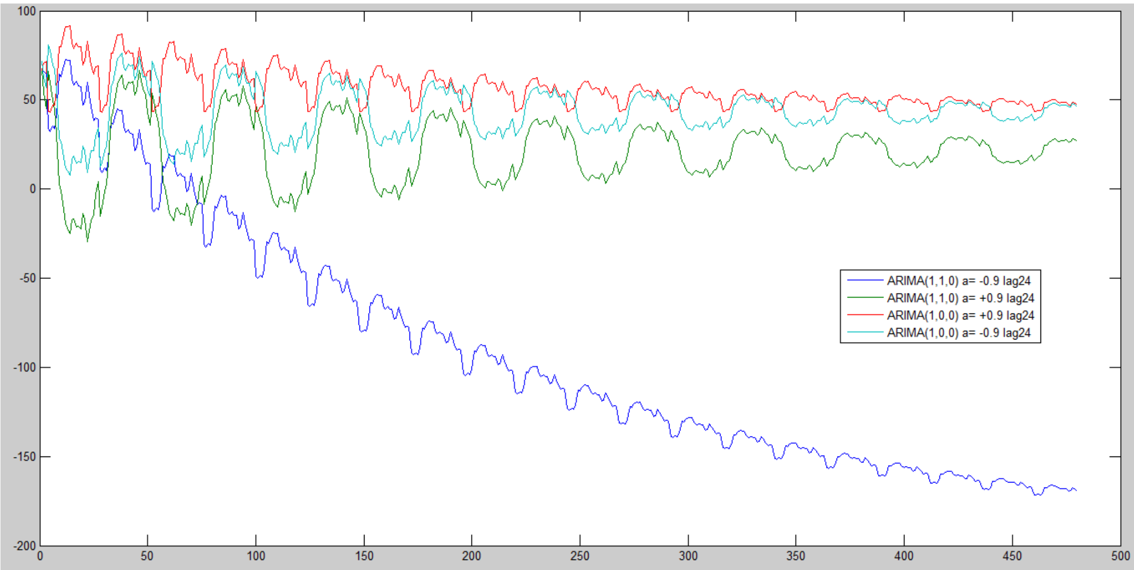
Differenciáltam az árakat, majd újra megnéztem az ACF és a PACF az újonnan kialakult adatsorra, 24. ábra.

A matlabnak nagyon jól összeszedett jól paraméterezhető tesztjei vannak a stacionaritás eldöntésére, mint az *adftest* ami a Dickey-Fuller tesztet használja vagy a *pptest* ami Phillips-Perron unit tesztje, és van *i10test* ami ezt a kettőt egyszerre teszteli. Ezek alapján az jött ki, hogy az alap adathalmazunk nem stacionárius, Így az I(d) = 0.

Mivel az ACF első diagram nem tartalmaz éles levágást, így nincs meghatározható MA tag. A PACF diagramon van éles levágás így AR domináns lesz az adatsorunk. Ökölszabályként megfogalmazható, hogy éles PACF levágás esetén olyan értékű AR(p), míg éles ACF levágás esetén olyan értékű MA(q) tagot kell választani, amelyik időbeli távolságnál ez a levágás megtörtént. Fontos megjegyezni, hogy általában a legjobb modellek vagy csak AR tagot, vagy csak MA tagot használnak. Persze lehetséges vegyes modell is, de ekkor a különböző részek kiolthatják egymást.

Ha megnézzük jobban a (23. ábra 24. ábra) PACF és ACF ábrákat akkor jól látható, hogy mindkettőn van egy 24 óránként ismétlődő kiugrás, illetve ezt már korábban is láttam, használtam a korábbi előrejelzésekben, itt is bele fogom venni a paraméterezésbe.

Két felé szedtem a meglévő árakat, első adathalmaz a „tanító” adathalmaz 2050 nappal és 20 maradt, amit előre fognak majd jelezni a különbözően felparaméterezett algoritmusok.



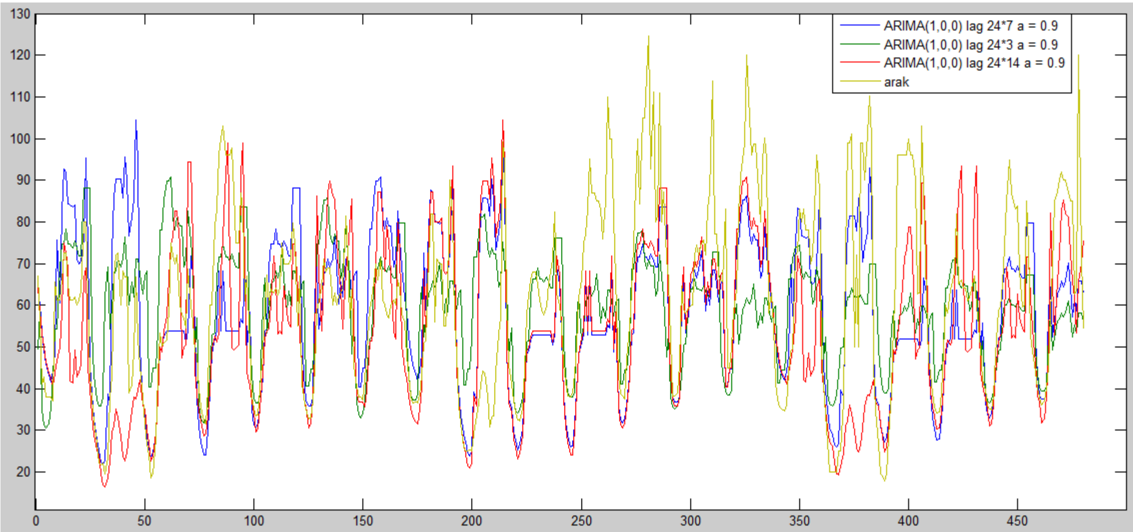

*25. ábra ARIMA modellek 24es regressziós késleltetéssel és 0.9es vagy -0.9es α-val*

A 25. ábra 4 db AR alapú modell, amik úgy voltak paraméterezve, hogy az értéke pont ellentétes, az az egyszer 0.9 másszor -0.9, ami azt jelenti, hogy az egyik esetben az előző értékek 0.9-eséből a másik esetben a -0.9-eséből indulunk ki, majd a hozzáadtuk a zajt. Jól látszik, hogy ennek hatására a világos kék és a piros pont ellentétes ciklusú lesz.

A sötét zöld és a sötét kék a differenciált párok, jól látszik, hogy nagyon elszaladnak a négyzetes hibájuk is 2000 felett van, a teszteken is jól látszik, hogy nem kell differenciálni az adatainkat.

Az α = 0.9-es ARIMA modell MSE = 516.2010, míg az α = -0.9-es változat a 968.5370 hibával zárt. Ezek alapján megfogom tartani a pozitív a-t.

MSE (24*3lag) = 376.3099, MSE (24*7lag) = 338.5795, MSE(24*14lag) = 392.1361 ezek, mind 0.9-es a-val lettek előre jelezve, úgy tűnik, hogy a 14 nap túl hosszú idő intervallum most is a 7 nap körüli érték lesz a megfelelő. 26. ábra




*26. ábra 24-es regressziós ciklusú ARIMA (1,0,0) modellek*

Ha az α már csak 0.5, akkor jobb lesz az előrejelzés ezeken az adatokon, már csak MSE = 331.4473, α = 0.2-es már kevésnek bizonyult MSE = 332.7707

Megtartottam az előző kísérletekből a 7 napos késleltetési paramétert, és az AR rész p nagyságával és az α értékeinek megválasztásával játszottam tovább. Kipróbáltam az ARIMA (2,0,0) modelltől az ARIMA (5,0,0)-ig számos α variációt. A kipróbált teszt esete közül az az ARIMA (3,0,0) modell lett a legjobb ahol az α értékei {0.6 0.3 -0.1}. Ennek az esetnek az átlagos négyzetes hibája MSE = 330.9795 lett.

Úgy látom, hogy az ARIMA modell nagy problémája hogy időközönként újra kell tanítani a modellt, egy frissebb adathalmazon, mert tényleg csak nagyon rövidtávon képes jól előre jelezni, illetve hogy egyes paraméterek finom hangolása miatt sok modellt kell tanítani mire megtaláljuk az igazit.

## 3.6 Legközelebbi szomszéd módszerek

A legközelebbi szomszéd az a nap lesz, ami a legkisebb abszolút hibát produkálja az utolsó ismert nappal. Így a kimeneten a legközelebbi szomszéd utáni napot fogom megadni. Ami még fontos, hogy nem csak az árak hasonlóságát vettem alapul, hanem a rendszerterhelést és a határmetszék adatait is.

Minden próbálkozáshoz készítettem egy mátrixot, ami az összes jósolt napot tartalmazz. Majd megnéztem, hogy ez a mátrix mekkora abszolút átlagos hibát produkál a valós adatokkal, hogy összehasonlítható legyen az egyes algoritmusok jósága.

Ez gyakorlatilag úgy nézett ki, hogy az utolsó ismert napot kivontam az összes napból, ami már az adatbázisban volt kivéve saját magát. Majd vettem az abszolút értékét az eredmények, ez után summáztam őket így minden napra létrejött egy abszolút relatív hiba. Ezek között megkerestem a legkisebbet, majd a legkisebb utáni napot adtam, meg mint előrejelzési nap.

Matlab kódként így néz ki:

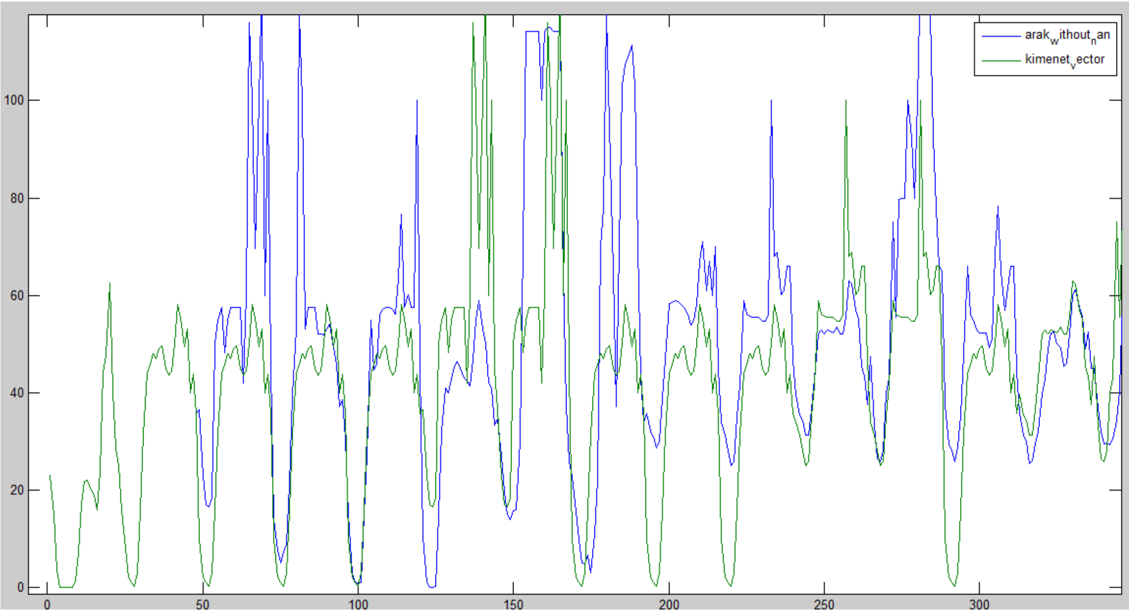
```
tmpAr = abs(Ar - ones(nn-1,1)*keresettNap);
[Arminimum, Arhely] = min(sum(tmpAr,2));
```

A fentebbi kód részlet az árakban keres hasonlóságot, de a változók kicserélésével, ugyan így össze lehet hasonlítani az összes határmetszéket, a rendszerterhelését vagy az export-import szaldot. A matlabban nagyon hatékonyak a mátrix műveletek, illetve rendelkezésre állt minden függvény, amire szükségem van ehhez a feladathoz.

Mivel azért tetszett nagyon ez az algoritmus, mert ha már van 2 egységnyi adatunk, akkor van minek a hasonlóságát vizsgálni így már 2 egység után tudok valamit előre jelezni. Így hát elsőnek úgy próbáltam ki az algoritmust, hogy 2 napot megkapott és mindig feltöltöttem az előrejelzés után a valós adattal. (27. ábra)

Átlagos százalékos abszolút hiba (MAPE) = 29.9160

Átlagos négyzetes hiba (MSE) = 442.3010




*27. ábra Első algoritmus 1 szomszéd 2 nap alap adatbázis*

Megpróbáltam nagyobb kezdeti adathalmazzal is. 14 nap abszolút százalékos hibája 29.8727 volt, ami alig különbözik attól, hogy ha kettő napot veszek, majd 30 napot kapott kiindulási alapnak, ahol 29.7640 lett a hiba. Ahogy várható hogy több kezdeti adatból jobb eredményeket tud találni viszont az különböző méretű bemeneti adatok száma nem befolyásolta eléggé az eredményt csak pár tizedesnyire.

Itt most az összes határmetszék és a bruttó rendszerterhelés is nézve volt, ha azoknak kisebb volt az abszolút napi hibájuk, akkor azon a napon lévő árat jósoltam. Így hát elkezdtem gondolkodni, hogy az egyik határmetszék nagyon hasonlít saját magára így folyton ő határozza meg az adott árat. Bővítettem az algoritmust és hozzá vettem egy statisztikát arról, hogy melyik adat alapján döntöttem az adott ár mellett.

Újra 2 napot kapott és lefutattam az algoritmust.

Ár = 664

Bruttó rendszer terhelés = 0

HUSK határmetszék = 2

HUUK határmetszék = 12

HUAT határmetszék = 222

HURO határmetszék = 724

HURS határmetszék = 424

HUHR határmetszék = 20

A korreláció vizsgálatnál azt mondtam, hogy a HURS és HUHR határmetszék nem korrelálnak az árral, viszont nagyon sokszor határozzák meg az árat az algoritmusban, míg a bruttó rendszerterhelés, ami jól korrelált az árral nullaszor így kiszedtem ezt a két határmetszéket az algoritmusból és újra futtattam.

Egyből javult a helyzet

MAPE = 29.3532

MSE = 428.1754

Megnéztem 30 nappal is és MAPE = 29.1928 lett, ami még mindig nem nagy változás. Így kidobom a következő kis korrelációjú határmetszéket is ami HURO volt. MAPE = 27.8299 lett és a MSE = 380.0701 is leesett, ezt is megnéztem 30 nappal. Egy egésszel jobb lett az eredmény a 30 napos alap adatbázissal.

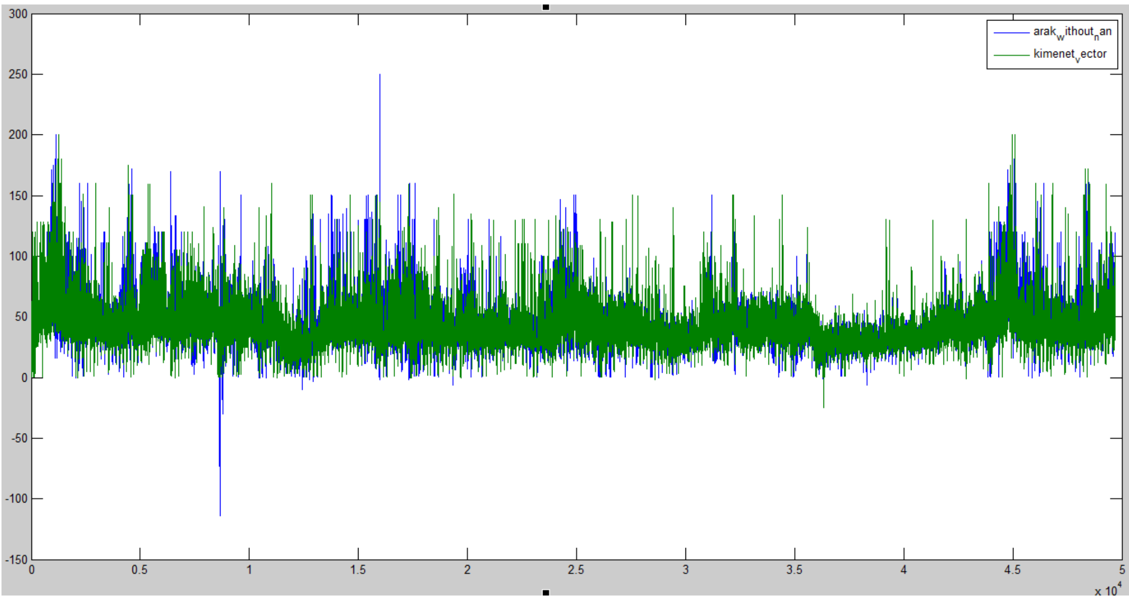
Kidobtam a HUUK-ot is, most romlott az eredmény az előző állapothoz képest a MAPE = 27.9164 lett. Ebben a felállásban volt az ár mellett bruttó rendszerterhelés, HUSK és HUAT határmetszék.

Ha csak HUAT és bruttó rendszerterhelés játszik, még akkor a MAPE = 27.9393.

Csak a bruttó rendszerterhelés és az árak voltak az algoritmusban akkor a MAPE = 27.2663, és végül csak az árat használtam, már amikor a MAPE = 27.2559

Végül az jött ki, ami a korrelációban is, hogy az ár önmagától függ a legjobban és utána a bruttó rendszerterheléstől. A HUUK, HUAT, HUSK határmetszékek hasonló mértékben határozzák meg, és nem hasonlítanak saját magukra annyira, hogy elvigyék nagyon rossz irányba a kiértékelést.

Egy 30 napos alap adatbázissal csak az árak hasonlítgatásával a MAPE = 27.0840 és a MSE = 349.4104. Az a 30 napos adathalmazzal rendelkező előrejelzést mutatja be. Azért itt jól látszik, hogy a hirtelen kiugrásokat nem tudja lekövetni az algoritmus, viszont a növekvős és csökkenő trendeket késéssel ugyan, viszont leköveti. (28. ábra)




*28. ábra 1 szomszéd, csak az árak figyelembevételével, 30 napos adatbázis*

Az előbbi paraméterezésekből azt sikerült leszűrnöm, hogy egy magához nagyon hasonlító adathalmaz teljesen el tudja rontani az egész számítást, így hogy ezen túl a hasonlóságokat finomítsam, több szomszéddal fogok, kísérletezni. Illetve megnézem az összes határmetszék 24 órával eltolt autokorrelációját.

*HUAT 24 órás autokorrelációja = 0.7655*

*HUHR 24 órás autokorrelációja = 0.8741*

*HURO 24 órás autokorrelációja = 0.7514*

*HURS 24 órás autokorrelációja = 0.8170*

*HUSK 24 órás autokorrelációja = 0.7470*

*HUUK 24 órás autokorrelációja = 0.8339*

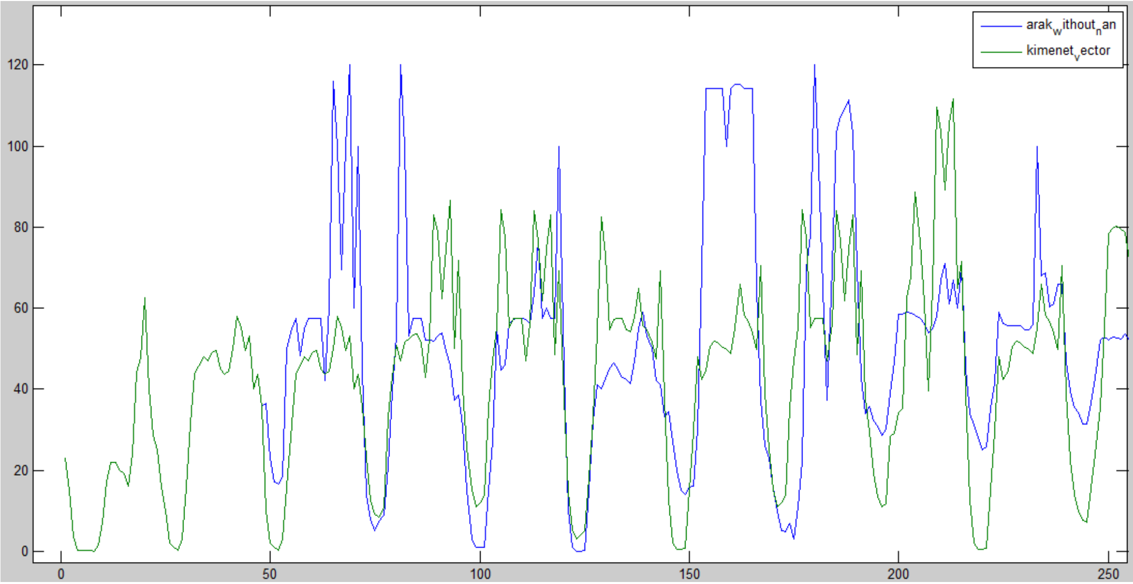
*Összesített Import-Export 24 órás autokorrelációja = 0.8888*

*RendszerTerheles 24 órás autokorrelációja = 0.8603*

Sokkal jobban korrelálnak magukkal, mint az ár saját magával. Így elkezdtem kísérletezni több szomszéddal vagy, csak néhány adatot használok, vagy csak az árat. Szerettem volna ha határmetszékeket valahogy felhasználni vagy legalább szűrni, hogy pontosan melyikek határozzák meg az árat, de úgy látszik, hogy a határmetszékek menetrendje miatt túlságosan is hasonlítanak a 24 órás eltoltjukra.

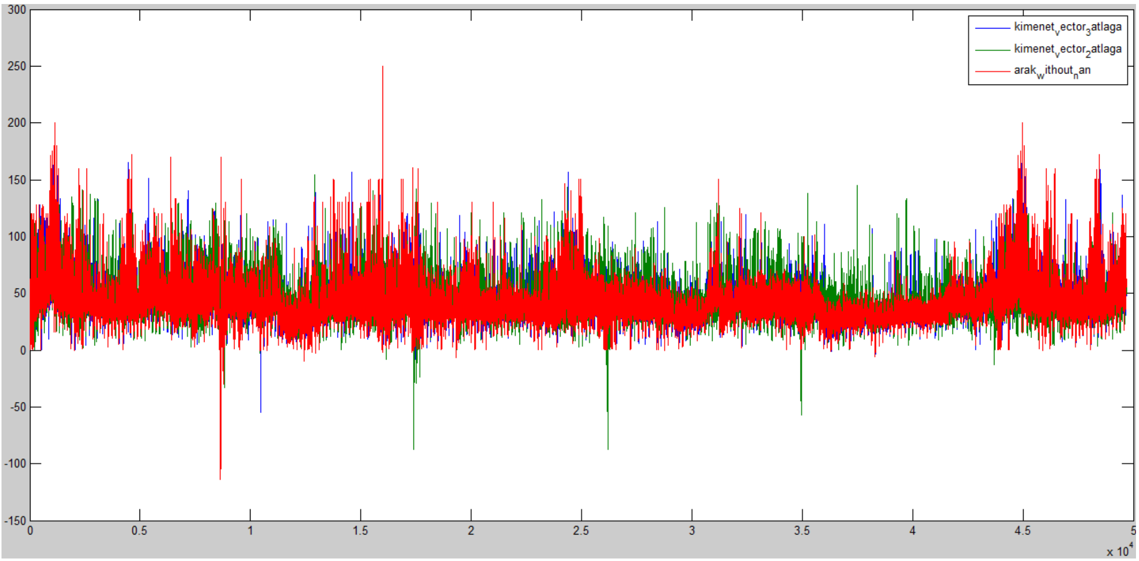
Elsőnek a két legközelebbi szomszéd átlagát néztem, és csak a bruttó rendszerterhelés és az ár volt benne a vizsgált adatokban. MSE = 345.9908 és a MAPE = 27.6153

29. ábra a két legközelebbi szomszéd átlagolását mutatja be, és abból is az első pár napot. Ezen jól látszik, hogy nincs olyan, mint az ábrán volt, hogy a második napi adat sokáig ismétlődik, illetve hogy az átlagolás miatt kevesebb az egyforma jósolt nap az elején.

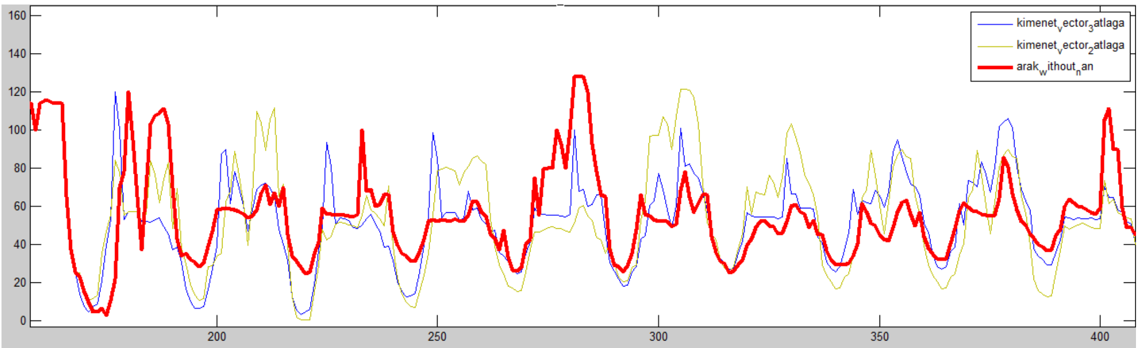



*29. ábra 2 legközelebbi szomszéd átlagával jósolok*

A három legközelebbi szomszéd átlagolása 7 nap alap adatbázissal, ami csak az ár és a bruttó rendszerterhelést tartalmazza. A MAPE = 27.3946 és a MSE = 263.9607 (30. ábra és 31. ábra). Jobb lett, de nem jelentősen. Itt volt egy másik átlagolási ötletem is. Ez, ami ilyen jó eredménnyel zárult úgy készült, hogy vettem az adatokon a hasonlóságokat, a leghasonlóbbnak megfogtam az árát és beletettem egy tömbbe, majd kiszedtem a már megtalált elem sorszámú adatot az összes adathalmazból és ezt újra megcsináltam a maradék adatokon. Ezzel az a probléma, hogy elcsúsznak az adatok olyan szempontból, hogy egy-egy elem kivétele után már lesznek olyan elemek az adathalmazban, amik normál esetben nem lennének szomszédok, viszont ezért cserébe minden választott elemet a tényleg leghasonlóbb adat alapján választjuk ki.

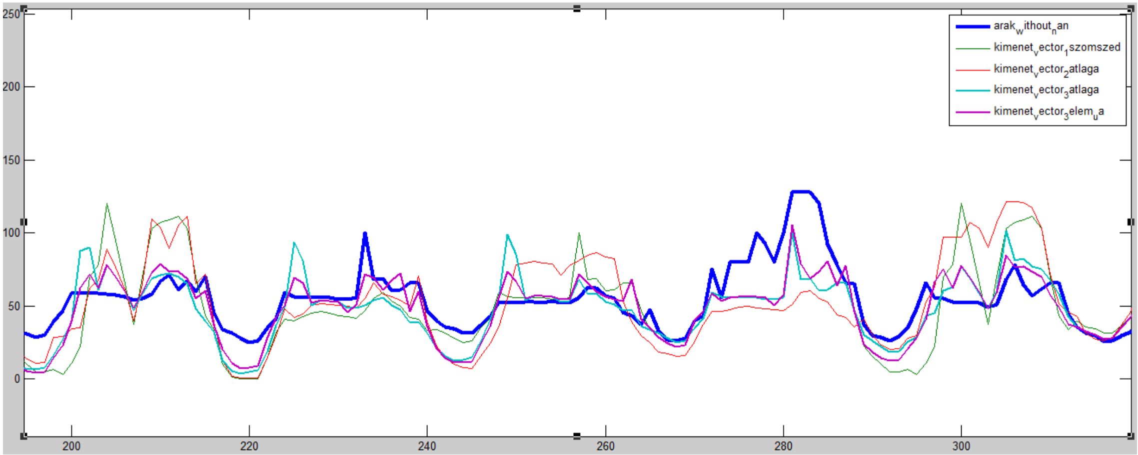



*30. ábra 2 illetve 3 legközelebbi szomszéd*




*31. ábra 2, 3 szomszéd közelebbről*

A másik ötlet az volt hogy megnézzük a hasonlóságokat majd megfogjuk az egyik adathalmaz leghasonlóbb x darab elemét és átlagoljuk őket ez fogja megadni hogy az adott halmaz mennyire hasonlít a keresett napra, ezt megcsináljuk az összes adathalmazzal. Ezzel az a baj, hogy egy adathalmaz dönti el csak hogy melyik napokat átlagoljuk, így sajnos elvesztettem azt a lehetőséget, hogy minden szomszédot különböző adathalmazok alapján válasszak ki, mivel végül is csak a bruttó rendszerterhelés és az árak vannak jelenleg használva így úgy éreztem, hogy ez nem akkora ár azért, hogy így nem csúsznak el a tényleges szomszédok. A kettő szomszédos átlagolás amiről ábrák vannak fentebb is ezen a módon készült.




*32. ábra Különböző szomszéd algoritmusok összehasonlítása*

A 32. ábra találomra kiemelt 5 nap. A kék az ár, amit szerettünk volna megjósolni. A lila az, amikor úgy átlagoltam, hogy minden adatsorból kiszedtem a 3 legjobbat és azokat átlagoltam. A világos kék pedig az az eset, amikor a 3at úgy választottam ki, hogy mindig kivettem egyet valamelyik adatsor alapján és a végén átlagoltam. A piros is ezen alapszik csak kettő elemmel, a zöld pedig csak egy szomszéd.

MSE (kimenet_vector_2elem_ua) = 344.9141

MSE (kimenet_vector_3kulonbozok_atlaga) = 263.9607

MSE (kimenet_vector_3elem_ua) = 238.6426

MSE (kimenet_vector_1szomszed) = 353.9069

Bár a fentebbi esetekben főleg csak ár és rendszerterhelés volt, nem mondtam le teljesen a határmetszékekről sem. Sajnos külön-külön használva a határmetszékeket, nem voltak valami sikeresek a tesztjeim, ezért megpróbálkoztam egyben az import-export szaldóval is. Ár, rendszerterhelés, import-export 1 szomszéd esetén az MSE = 361.5325 ez jobb mintha az összes metszéket külön-külön nézném, viszont rosszabb, mint ha csak az árak alapján jósolnék előre. Különböző adatok alapján vett 3 szomszéd átlagával előre jelzett MSE = 263.9607, itt úgy tűnik, hogy mindegy hogy van-e import-export vagy nincs, vagy nem használja, vagy pont ugyan azt találja meg így is. Felokosítottam egy kis statisztikával és sajnos úgy tűnik, hogy nem használja.

Sajnos azt kellett tapasztalnom, hogy az ár és a rendszerterhelés együttesét a legjobb választás adathalmaznak, a többi adat félre viszi az algoritmust. Nagyon úgy tűnik, hogy feleslegesen töltöttem le azt a sok adatot, mert a jelen algoritmusok egyike se tudja őket értelmesen kihasználni.

# 4 Eredmények, jövőbeli tervek

Összegyűjtöttem az idősor-előrejelzésre alkalmas módszerek egy csoportját majd, implementáltam és összehasonlítottam őket. A kipróbált algoritmusok a legközelebbi szomszéd, ARMA, ARIMA modellek, regresszió, egyszerű statisztikai módszerek, mint az átlagolás vagy a medián.

A MAVIR honlapjáról és a HUPX honlapról beszereztem a rendelkezésre álló adatokat.

Az adatok feldolgozás és implementálása matlabban történt, így megismerkedtem a matlab Econometrics Toolbox –ával, Financial Toolbox –ával és a Time Series Analysis modulal.

Jelenleg csak a matlabban futnak a scriptek és nem egy gombnyomásra, fel kell őket paraméterezni, ezt egyszerűsíteni kellene, felhasználó baráttá tenni, hogy bárki jósolhasson velük árat

Az egyszerű statisztikai módszerek, mint a medián és az átlag meglepően jól teljesíttek. Scripteket írtam hozzájuk, amiket fel lehet paraméterezni. Ezek a scriptek úgy lettek kialakítva, hogy 1 napot jelezzenek mindig előre, a beadott adatot és a kezdeti időpont között és visszatérjenek egy vektorral ezen előre jelzett értékek összefűzése után. Két féle medián és átlag algoritmus van, az egyik egy zsinórárat jósolunk, a másik viszont kihasználja a napon belüli ingást és az aktuális óra átlagát vagy mediánját veszi x napra visszamenőleg. A legjobban az a medián teljesített, ami 7 napot kapott feldolgozási ablaknak és az óráként adta meg a medián, ez csupán 227.034 átlagos négyzetes hibát vétett a több mint 5 évnyi adaton. Ezt érdemes lenne tovább vizsgálni vagy felhasználni más algoritmusokban is.

A legközelebbi szomszéd is figyelemre méltó eredményekkel zárt viszont az még kíván, egy kis utó munkát ki kellene mérni, hogy pontosan hány szomszéddal érdemes, melyik adatok felhasználása mellett, vagy akár ki lehetne a hasonlóságot mérő hiba algoritmust cserélni egy átlagos négyzetes hiba számítóra, vagy egy átlagos százalékos abszolút hibára és úgy megnézni, hogy az eredményeket.

A legjobb legközelebbi szomszéd algoritmus MSE = 263.9607

Miután ezek az egyszerűbb algoritmusok már jól eredménnyel zártak az ARIMA, ARMA, AR vagy MA modellek erőforrás igényesek és sok historikus adatot igényelnek ezért azokra nem fektettem végül akkora hangsúlyt. Tovább kellene velük kísérletezni. Viszont e rövid idő alatt is látszik, hogy sok lehetőség van még benne, sok miden finom hangolható még. Jelenleg a legjobb ARIMA modellem átlagos négyzetes hibája MSE = 330.9795 lett, ami bár jóval elmarad a medián algoritmus mögött azért nem olyan rossz.

Összegezve a jövőbeli terveket a legfontosabb egy felhasználó barát kialakítás az egyik már működő algoritmussal hogy addig is lehessen használni. Valahogy ki kellene utána úgy bővíteni, hogy kicserélhető legyen alatta az adatbázis bármilyen órás bontású adatsorra, mert akkor akár más típusú adatok előrejelzését is meg lehetne vele próbálni.

Majd vagy az ARIMA vagy a legközelebbi szomszéd algoritmusba kellene még több időt belefektetni. A legközelebbi szomszédnál ki lehetne cserélni egy másik hibaszámító algoritmusra a jelenlegi közelséget számító algoritmust. Az ARIMA-nak pedig több idő, és talán egy jobb hardver kellene, hogy ki lehessen pontosan kísérletezni, hogy milyen paraméterekkel működik a legjobban

# 5 Irodalom jegyzék:


[1] http://www.bibl.u-szeged.hu/ha/gazd/gazdasag/kl.html

[2] https://www.hupx.hu/hu/Piacosszekapcsolas/piacosszekapcsolastort/Lapok/default.aspx

[3] https://www.MAVIR.hu

[4] https://www.mvmpartner.hu/hu-HU/Szolgaltatasok/Villamos-Energia/Tudastar/FogalomTar

[5] https://www.MAVIR.hu/documents/10258/107815/Sz%C3%A9lkihaszn%C3%A1lts%C3%A1g+tanulm%C3%A1ny+2010.pdf/153d2d78-1c6f-4d54-858e-5bc46f56c352

[6] http://www.MAVIR.hu/c/document_library/get_file?uuid=81fd9f45-12cf-44e3-ad74-9492504a42ef&groupId=10258

[7] http://tozsdebarat.oldalunk.hu/site.php?sd=tozsdebarat&page=oeoxtYiFA0

[8] https://hu.wikipedia.org/wiki/Korrel%C3%A1ci%C3%B3

[9] Dr. Pór Gábor Méréstechinka 2013 http://www.tankonyvtar.hu/hu/tartalom/tamop412A/2010-0013_merestechnika/8_5_autokorrelacios_fuggveny.html

[10] https://www.mateking.hu/statisztika-2/regresszioszamitas/

[11] Balogh Péter és Nagy Lajos Ökonometria /Elméleti jegyzet/ (2013) http://www.tankonyvtar.hu/hu/tartalom/tamop412A/2011-0029_de_okonometria_elmelet/ch11.html

[12] Dr. Balázs Katalin Statisztika 1 oktatási segéd anyag http://psycho.unideb.hu/munkatarsak/balazs_katalin/stat1/stat1ora4.pdf

[13] Matematikai statisztikai elemzések 6. Regressziószámítás: kétváltozós lineáris és nemlineáris regresszió, többváltozós regresszió, Prof. Dr. Závoti József http://www.tankonyvtar.hu/en/tartalom/tamop425/0027_MSTE6/ch01s02.html

[14] Numerikus módszerek, Mészáros Józsefné http://www.tankonyvtar.hu/hu/tartalom/tamop425/0033_SCORM_GEMAK6841B/adatok.html

[15] https://hu.wikipedia.org/wiki/Legkisebb_n%C3%A9gyzetek_m%C3%B3dszere

[16] http://www.portfolio.hu/vallalatok/technikai-elemzes/technikai-elemzes-a-mozgoatlag-hasznalata.17774.html

[17] https://www.elemzeskozpont.hu/mozgoatlagok

[18] Üzleti prognózisok idősoros modelljei Prof. Dr. Besenyei Lajos, Domán Csaba (2011) http://www.tankonyvtar.hu/hu/tartalom/tamop425/0049_09_uzleti_prognozisok_idosoros_modelljei/989/index.html

[19] http://www.cs.bme.hu/nagyadat/

[20] http://elib.kkf.hu/okt_publ/szf_11_02.pdf

[21] A predictive pan-European economic and production dispatch model for the energy transition in the electricity sector, PowerTech, 2017 IEEE Manchester, 18-22 June, Laurent Pagnier, Philippe Jacquod http://ieeexplore.ieee.org/document/7980982/

[22] ANN approach for predicting economic trends based on electric energy consumption during natural disaster period, Knowledge, Information and Creativity Support Systems (KICSS), 2016 11th International Conference on, 10-12 Nov, Akanit Kwangkaew, Virach Sornlertlamvanich, Itsuo Kumazawa http://ieeexplore.ieee.org/document/7951405/

[23] https://en.wikipedia.org/wiki/Autoregressive%E2%80%93moving-average_model