Önállólaboratórium

**Kriston Melinda**

**Idősor elemzés, Idősor előrejelzés**

Konzulens : Pataki Béla

Tartalomjegyzék

[Feladat 2](#_Toc452908000)

[Adatgyűjtés 2](#_Toc452908001)

[Adatkiválasztás 2](#_Toc452908002)

[Adattisztítás 5](#_Toc452908003)

[Legközelebbi szomszéd 5](#_Toc452908004)

[Definíciója 5](#_Toc452908005)

[Megvalósítás 6](#_Toc452908006)

[Első algoritmusom 7](#_Toc452908007)

[Aránnyal módosított 7](#_Toc452908008)

[Több szomszéd 8](#_Toc452908009)

[Csak hétköznap 10](#_Toc452908010)

[Irodalomjegyzék 11](#_Toc452908011)

[Használt rövidítések, fogalmak 11](#_Toc452908012)

Feladat

Idősor-előrejelzésre alkalmas módszerek áttekintése, implementálása, összehasonlítása, használhatóságuk feltérképezése. A módszerek igen széles spektrumát használták, javasolták már erre a célra (, legközelebbi szomszéd alapú modellek, ARMA, ARIMA modellek, mesterséges neurális hálók, regressziós modellek stb.), ezek áttekintése és egy csoport kiválasztása, majd megvalósítása és tesztelése a feladat. Részben szimulált, részben valós, mért adatokon célszerű tesztelni a megvalósított módszereket.

A magyar áramtőzsde (HUPX) 2009-es indulása óta már több évnyi klíringeredmény áll rendelkezésre mélyrehatóbb elemző vizsgálatok elvégzéséhez. Az elmúlt 6 évben ráadásul három féle környezetben is működött a magyar tőzsde: önállóan, a cseh-szlovák-magyar összekapcsolásban, valamint a 4M piac-összekapcsolásban.

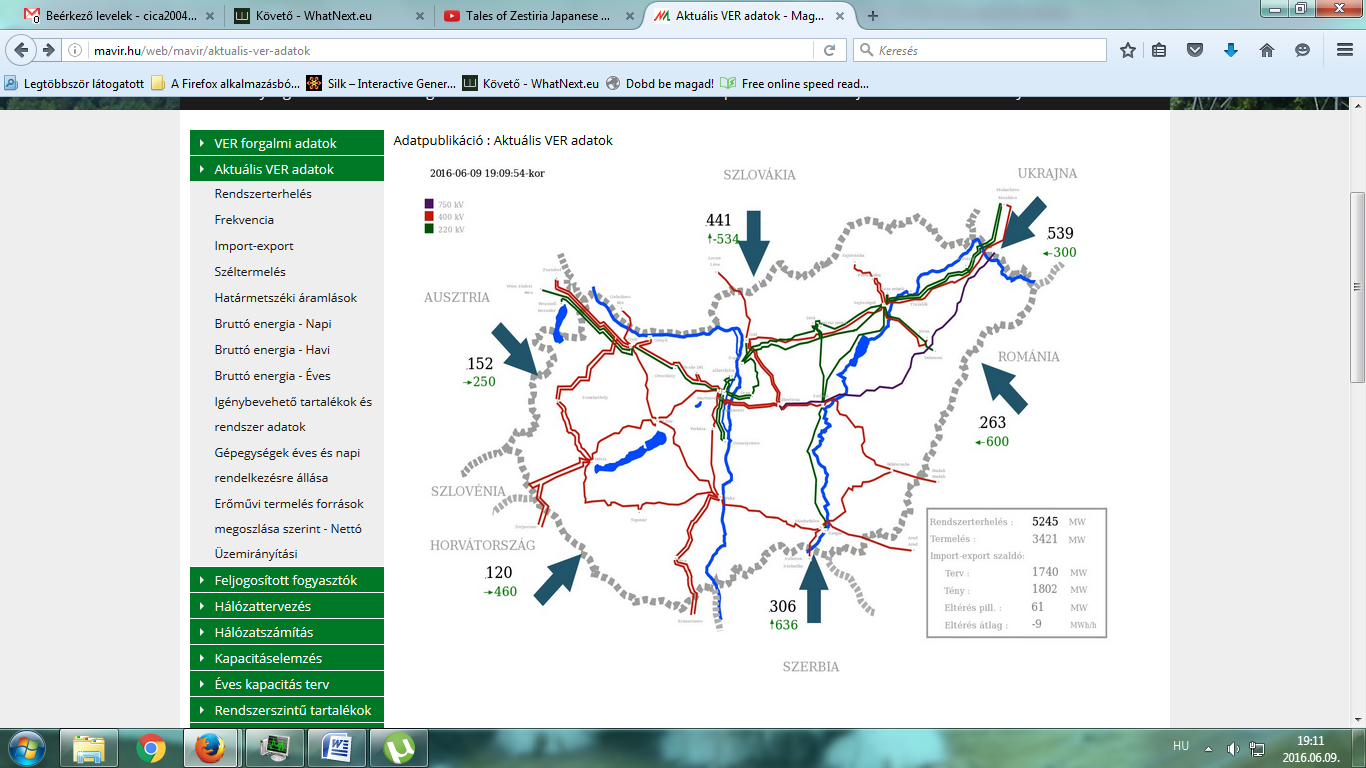
Adatgyűjtés

A mavir.hu-ról letöltöttem órás bontásba mindennapra az elérhető adatokat. Nagyon sok fajta adatot tudtam így megszerezni, mint például rendszerterhelés, import-export, magyarországi széltermelés, határmetszéki áramlások. Ezek között az adatok között sok olyan volt ami függött valamelyik másiktól vagy számomra nem hordozott semmilyen plusz felhasználható tényt. Szerettem volna még időjárás adatokat vagy a zöldenergia termelésről adatokat gyűjteni, de sajnos elég hiányos volt az ami elérhető volt. Ezen kívül rendelkezésre álltak még a VET tanszéktől az áramtőzsde árak a 2012 évtől 2015ig.

Adatkiválasztás

Elsőnek is a beszerzett adatok nem voltak teljesek hogy korrelációt tudjak nézni ezért az üres órákat költöttem -10000. Így a későbbiekben egyszerű volt ezen fals adatok kiszűrése.

Megnéztem minden adatra amit letöltöttem hogy mennyire korrelál az árral. Mint kiderült a Bruttó hazai rendszerterheléstől függ a legjobban az ár.



Ár és a bruttó tény rendszerterhelés = 0.6956

Ár és HUAT (Magyarország - Ausztria) határmetszéki áramlás menetrend = 0.4995

Ár és HUAT (Magyarország - Ausztria) határmetszéki áramlás mérés = 0.0090

Ár és HUHR (Magyarország - Horvátország) határmetszéki áramlás menetrend = 0.2559

Ár és HUHR (Magyarország - Horvátország) határmetszéki áramlás mérés = 0.0235

Ár és HURO (Magyarország - Románia) határmetszéki áramlás menetrend = 0.1018

Ár és HURO (Magyarország - Románia) határmetszéki áramlás mérés = 0.0085

Ár és HUSR (Magyarország - Szerbia) határmetszéki áramlás menetrend = 0.0926

Ár és HUSR (Magyarország - Szerbia) határmetszéki áramlás mérés = 0.0165

Ár és HUSK (Magyarország - Szlovákia) határmetszéki áramlás menetrend = 0.2040

Ár és HUSK (Magyarország - Szlovákia) határmetszéki áramlás mérés = 0.0116

Ár és HUUK (Magyarország - Ukrajna) határmetszéki áramlás menetrend = 0.0878

Ár és HUUK (Magyarország - Ukrajna) határmetszéki áramlás mérés = 0.0057

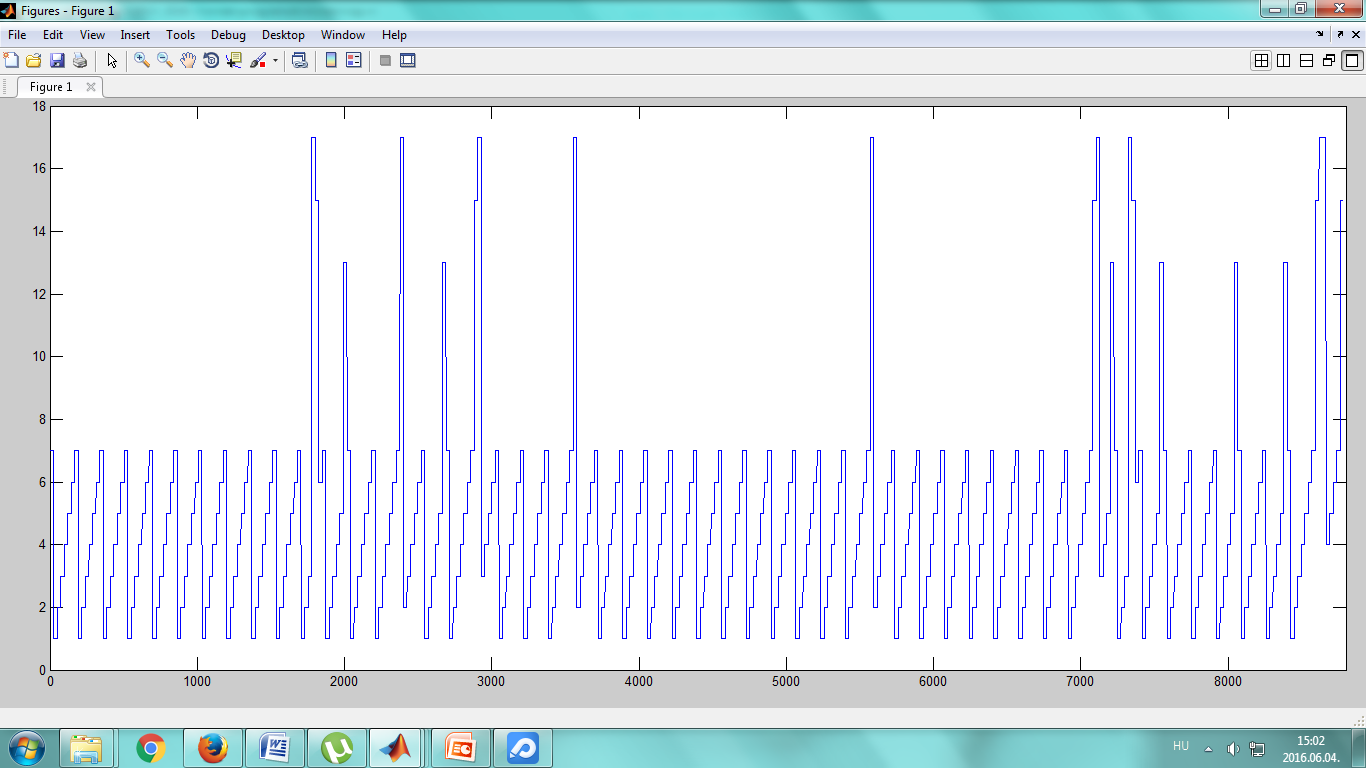
Ár és magyarországi szélerőművek termelése = -0.0160

Ár és a nettó tény rendszerterhelés = 0.6983

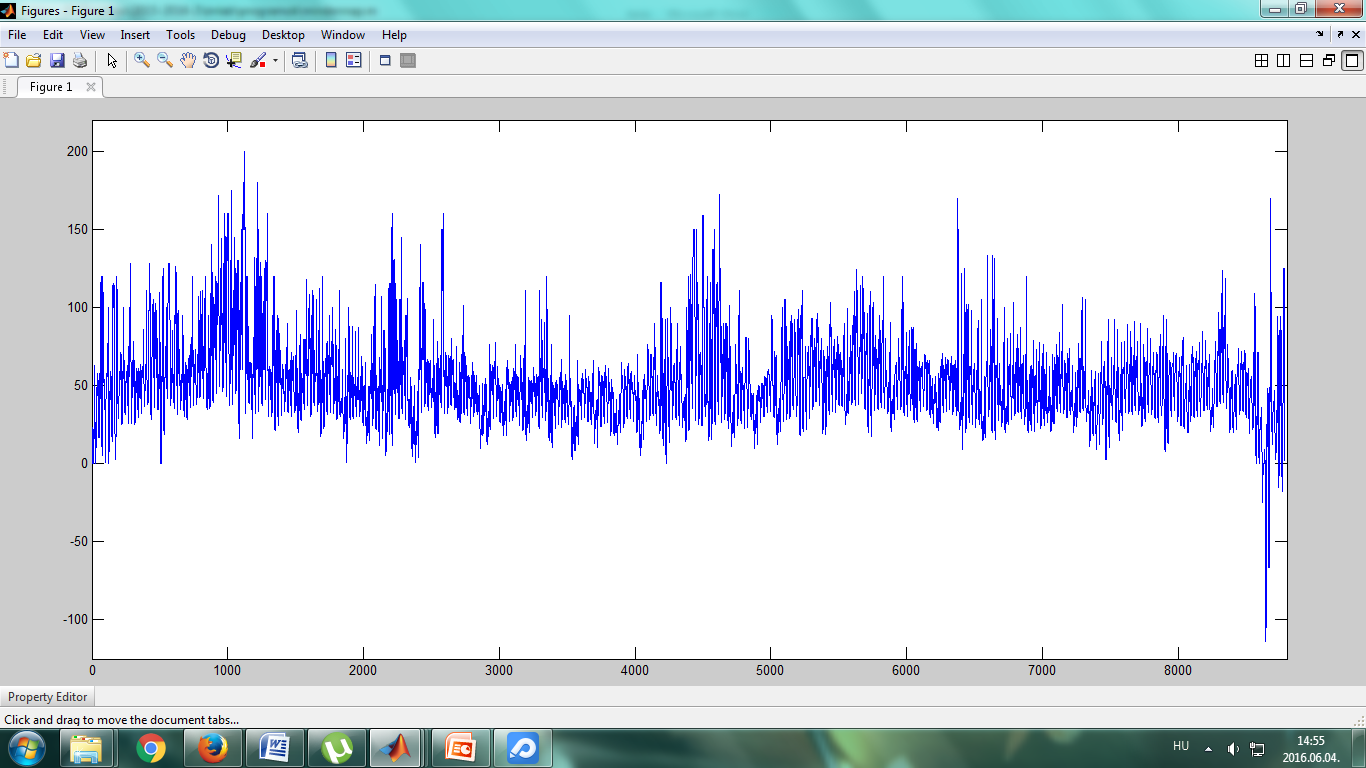
Még a SK-HU Szlovákia-Magyarország határmetszéki áramlást vettem bele mert szerettem volna valahogy még is csak figyelembe venni Németország termelését.

Ezen kívül készítettem egy napok vektort a 2012es amiben a napok fajtáját jelölöm. Minden hétfő 1es értéket kapott, a keddek 2esek, és így tovább a vasárnap 7es lett. A speciális napok, mint ünnep = 17 , ünnep miatti szabadság = 15, áthelyezett munkanap = 13. Ez azért kellett mert később kiderült hogy ha mondjuk pénteket találok a leghasonlóbbnak akkor egy szombati nap fog kerülni a kimenetre ami sajnos nem jó, mert a hétvégék és az ünnepnapok kiugró adatokat szolgálnak. Így végül a csak a 2012 évvel foglalkoztam.

Napok vektor :

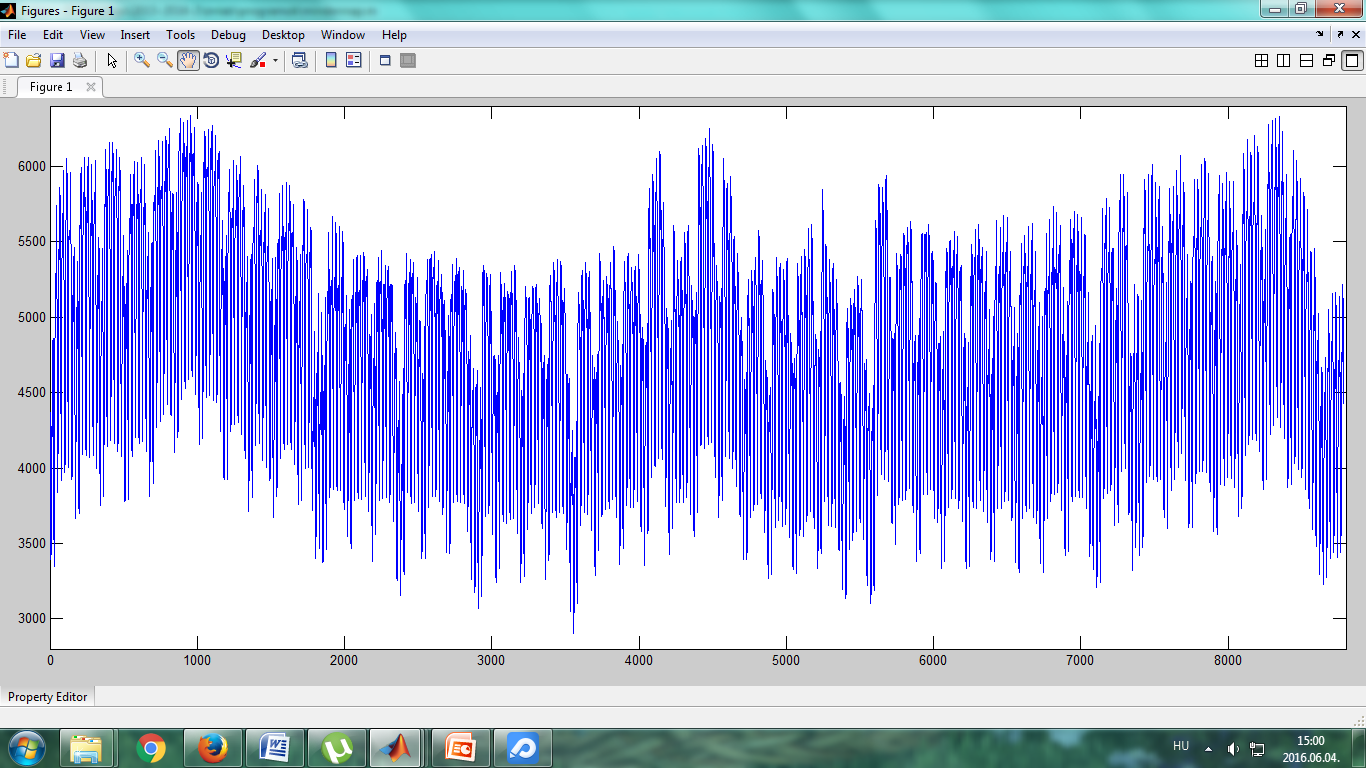


2012 évi árak alakulása :

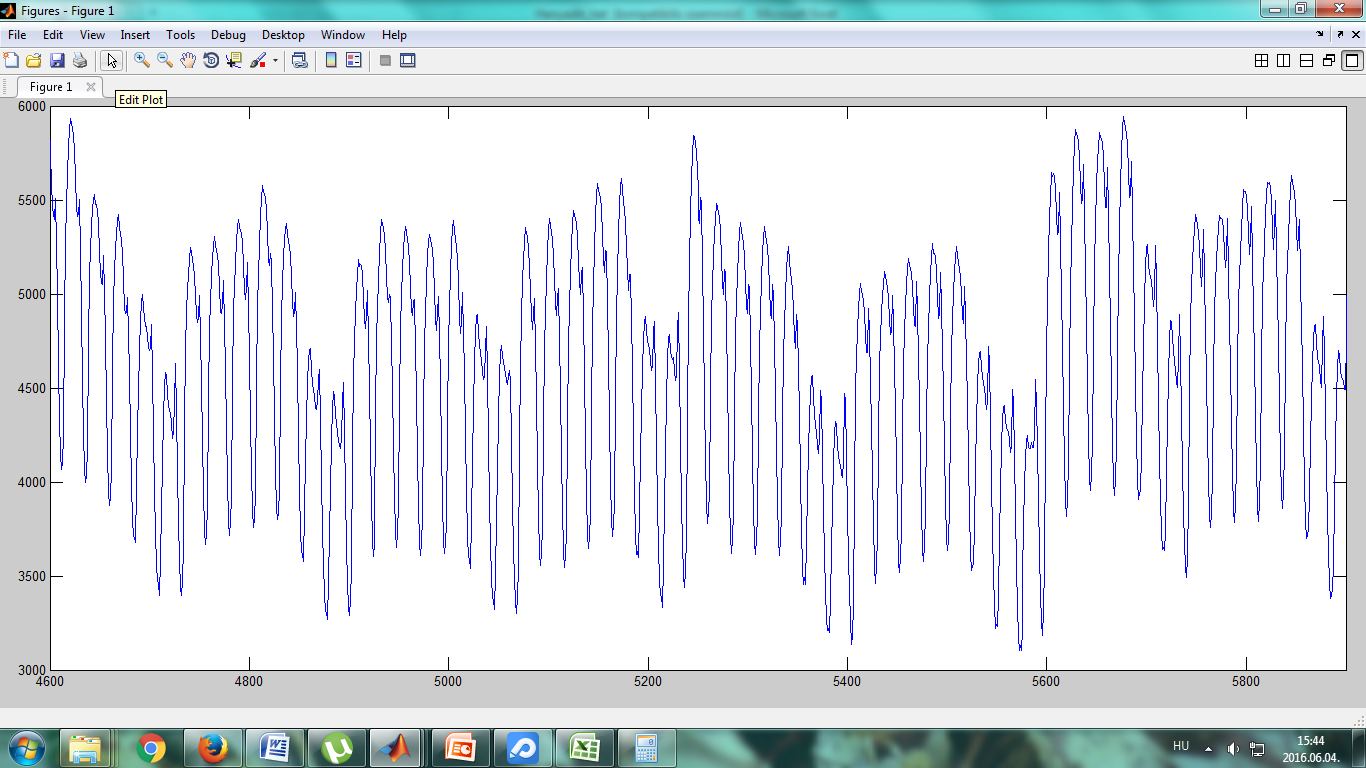


Itt az látszik hogy vannak napok amik kb az ünnepnapokra hétvégre esnek amik nagyon eltérnek a többitől. A legjobban ez évvégén látszik a karácsonyi szünetben amikor is mínuszba ment az ár. Sikerült sokkal többet termelni mit ami szükséges volt, pedig azért karácsonykor sok égőt és egyéb világító giccset használnak az lakossági felhasználók, itt hogy így helyzet alakult ki az azért lehetet mert vagy SK felöl sok áram érkezett vagy nem számoltak vele hogy a céges felhasználóknak nem kellesz áram.

2012 évi bruttó hazai rendszer terhelés :



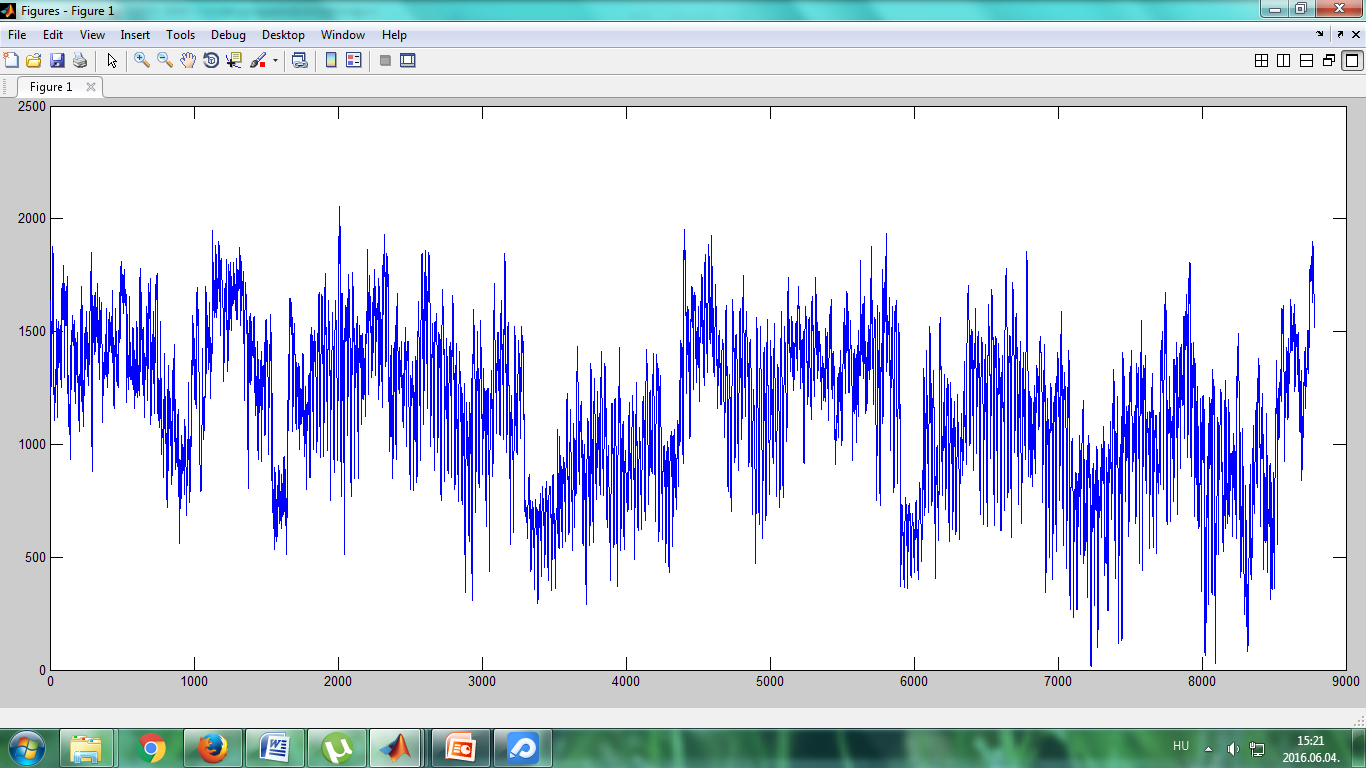
Karácsony környékén jól látszik hogy a rendszerterhelés alaposan lecsökkent. Ebből arra következtettem, hogy az ipari felhasználók amik igazán meghatározóak a rendszerterhelés szempontjából. Ezért is lehet hogy 5000 környékén hirtelen nagyon lecsökken a rendszerterhelés ekkor vannak a nyári szabadságolások. 233 nap volt az évben augusztus 20 ami a diagramon 5592 - 5616 ig van. Augusztus 20án volt olyan óra amikor 3104 MWig lement a rendszerterhelés.



Az említett nyári szabadságolási időszak 4600 = július 9, 5900 = szeptember 1 éjfél.

Összességében jól látszik hogy az ár jól korrelál a bruttó hazai termeléssel.

(SK - HU) Szlovákia-Magyarország határmetszéki áramlás 2012 :



Nagyon változó a folyó áram. Ő lesz az igazi probléma. Sajnos itt nem nagyon látszik elsőnek semmilyen összefüggés közte és az árak között illetve a között hogy milyen típusú nap van.

Adattisztítás

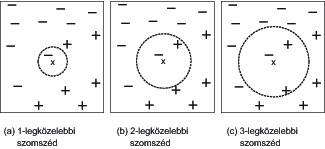
Sajnos a mavirról letöltött adatok nem voltak mindig teljesek. Voltak esetek amikor órák, sajnos néha napok is kimaradtak. Az órákat lineáris illesztéssel jósoltam meg, míg ha nap maradt ki akkor azt mondta hogy ugyan olyan mint tegnap volt.

Legközelebbi szomszéd

Mivel nem tudtam összegyűjteni elég adatot, ezért egy olyan algoritmus kerestem ami akár kevés elemmel is elég helyesen működik. Végül a legközelebbi szomszédra eset a választás, mert könnyen alakítható és már 1 ismert nap után úgy mód képes jósolni. És minél több adatunk van annál pontosabb lesz az algoritmusunk. Ráadásul egyszerű mátrix műveletekkel lekódolható matalban így nagyon gyorsan lefutott.

Definíciója

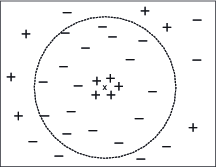
A legközelebbi szomszédok használatának indoklását legjobban a következő mondás szemlélteti: *" Ha valami úgy totyog, mint egy kacsa, úgy hápog, mint egy kacsa és úgy néz ki, mint egy kacsa, akkor az valószínűleg egy kacsa.''* A legközelebbi szomszéd osztályozó minden egyes esetet egy adatpontként reprezentál egy http://www.tankonyvtar.hu/hu/tartalom/tamop425/0046_adatbanyaszat/math/eq_1435.png -dimenziós térben, ahol http://www.tankonyvtar.hu/hu/tartalom/tamop425/0046_adatbanyaszat/math/eq_1436.png az attribútumok száma. Egy adott teszteset esetén a meghatározott szomszédsági mértékek valamelyikével kiszámítjuk annak közelségét a tanulóhalmaz összes többi adatpontjához. Egy adott http://www.tankonyvtar.hu/hu/tartalom/tamop425/0046_adatbanyaszat/math/eq_1437.png eset http://www.tankonyvtar.hu/hu/tartalom/tamop425/0046_adatbanyaszat/math/eq_1438.png -legközelebbi szomszédja azt a http://www.tankonyvtar.hu/hu/tartalom/tamop425/0046_adatbanyaszat/math/eq_1439.png pontot jelenti, amelyek a legközelebb vannak http://www.tankonyvtar.hu/hu/tartalom/tamop425/0046_adatbanyaszat/math/eq_1440.png -hez.



Az ábra a körök középpontjában lévő adatpont 1-, 2- és 3- legközelebbi szomszédját szemlélteti. Egy adatpontot a szomszédainak 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 http://www.tankonyvtar.hu/hu/tartalom/tamop425/0046_adatbanyaszat/math/eq_1441.png helyes megválasztásának fontosságát hangsúlyozzák. Ha http://www.tankonyvtar.hu/hu/tartalom/tamop425/0046_adatbanyaszat/math/eq_1442.png túl kicsi, akkor a legközelebbi szomszéd osztályozó a tanulóadatokban jelenlevő zaj miatt hajlamos lehet a túlillesztésre. Másrészt viszont, ha http://www.tankonyvtar.hu/hu/tartalom/tamop425/0046_adatbanyaszat/math/eq_1443.pngtúl nagy, akkor a legközelebbi szomszéd osztályozó rosszul osztályozhatja a tesztpéldányt, mivel a legközelebbi szomszédok listája a szomszédságtól messzi adatpontokat is tartalmazhat

k -legközelebbi szomszéd osztályozás nagy k esetén



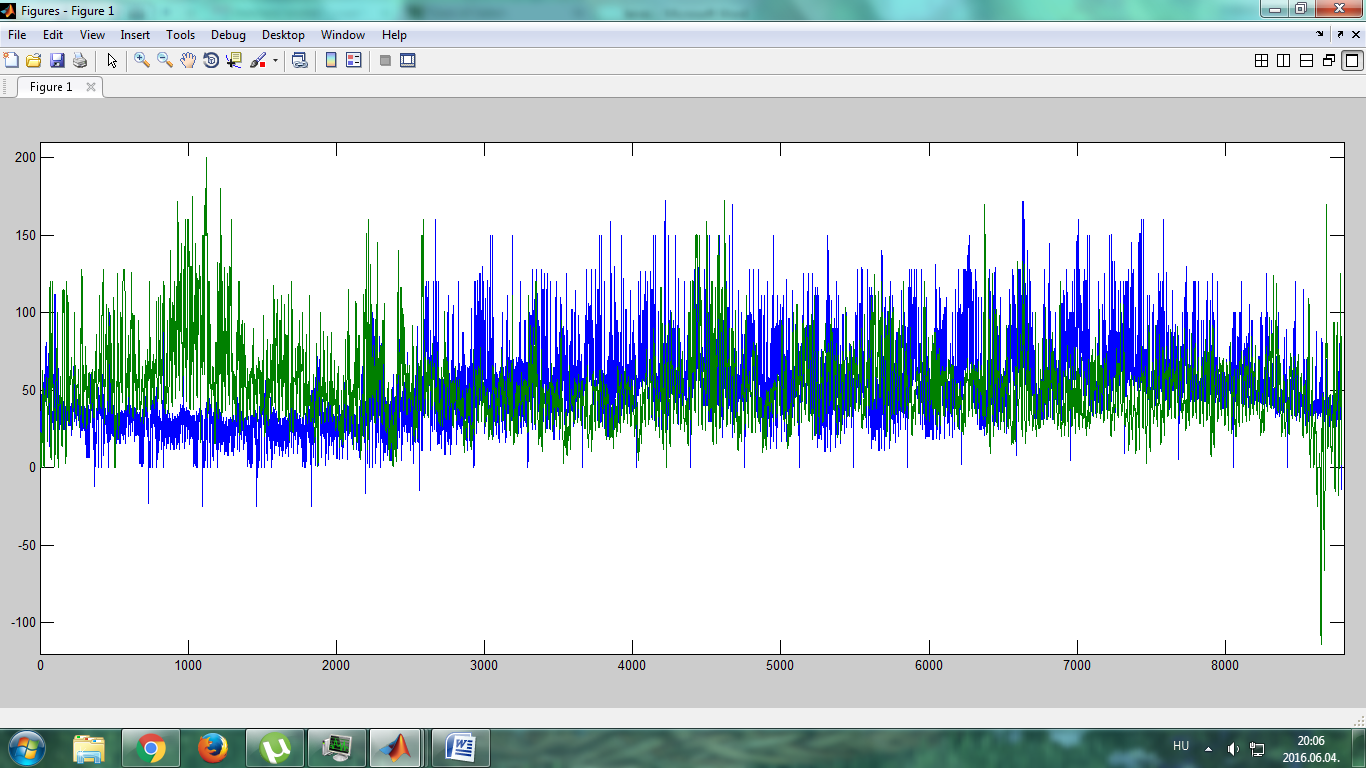
Megvalósítás

A legközelebbi szomszéd az a nap lesz ami a legkisebb abszolút hibát produkálja az utolsó ismert nappal. Így a kimeneten a legközelebbi szomszéd után napot fogom megadni.

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

Első algoritmusom

Mindennaphoz megkerestem a legközelebbi szomszédot mind a bruttó haza rendszer terhelésben (innentől kezdve csak röviden Br) mind a SK-HU határmetszéki áramlásban (a továbbiakban SK) és természetesen az árban is. Majd azt a helyet vettem ami a legkisebb hibát okozza a három mátrixban. Így végül első neki futásra 16.1649 hibát kaptam.



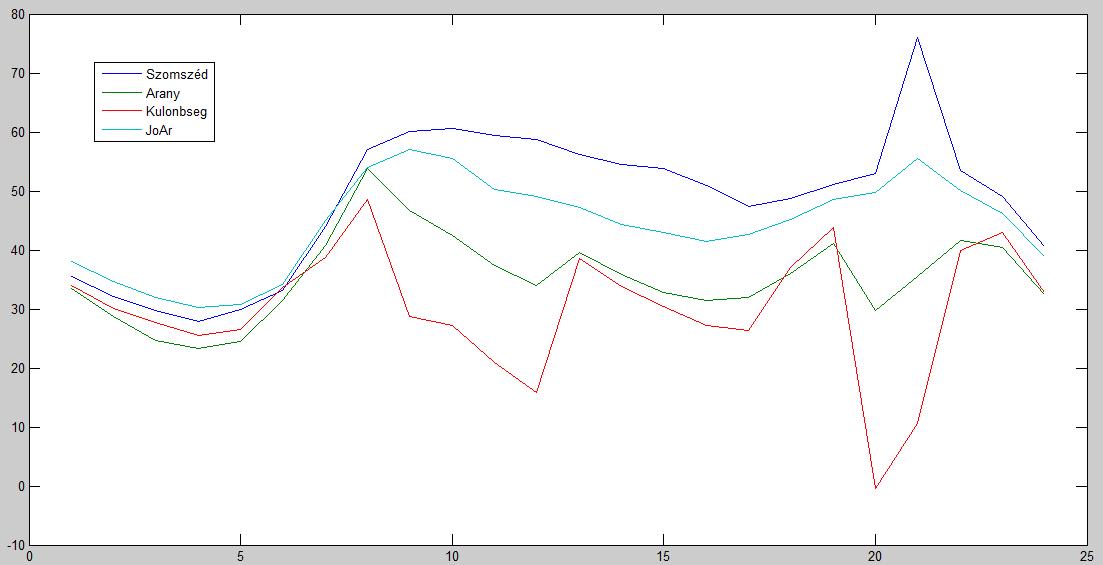
Ezen a rajzon jól látszik hogy még nagyon nem az igazi.

Ezek után megpróbáltam finomítani a dolgon különböző módosításokkal.

Aránnyal módosított

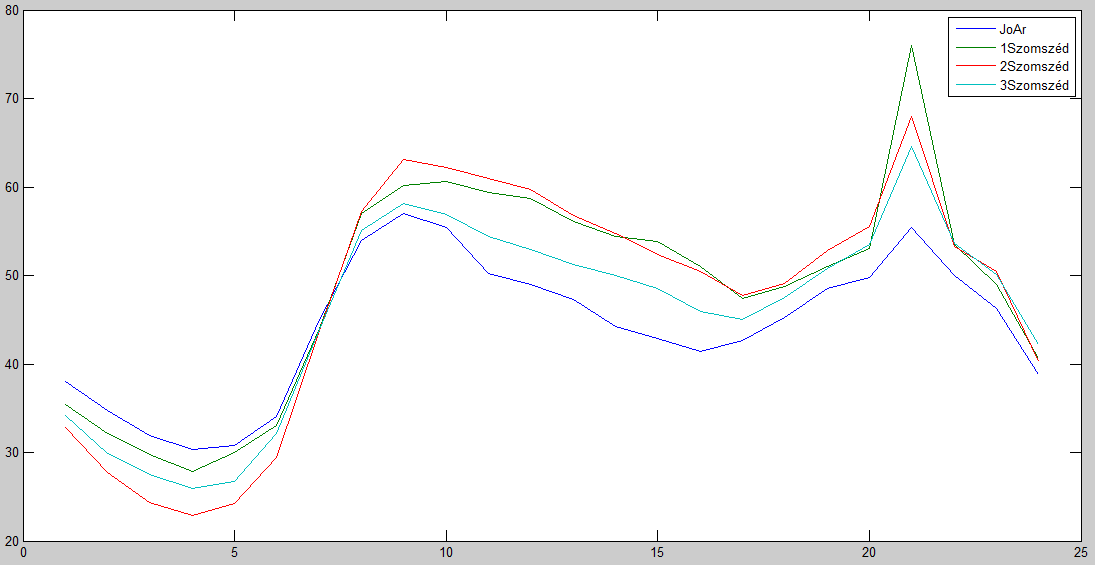
Elsőnek egy aránnyal módosítottam a kimeneti értékeket. Ez úgy történt hogy megnéztem hogy az utolsó ismert nap hogy korrelál az éppen kiválasztott nappal. Megnéztem a korrációt a Br, SK és az ár esetén is és ezekből korrelációs értékekből képeztem egy-egy számot úgy hogy az összegük 1 legyen. Majd megnéztem hogy mennyire tér el egymástól a kiválasztott és az utolsó ismert nap, vettem a különbségüket. Majd ezt a különbséget megszoroztam az aránnyal és hozzáadtam a kimenethez. Ez sajnos első neki futásra rosszabb eredményt adott mint csak a sima algoritmus. A hiba 16.7906 lett.

Kíváncsiságból csak a különbséggel módosított kimenet is kipróbáltam hát ha az jobb és az arány létrehozása nem volt jó. De sajnos ez még rosszabb eredmény hozott. 18.0478 lett



Több szomszéd

A legközelebbi szomszédok keresésénél általában nem csak egyet szoktak megkeresni ha nem annyit amennyi beleesik egy bizonyos tartományba. Ezt az ötletet felhasználva megkerestem még két szomszédot és átlagoltam a hozzájuk tartozó kimeneteket.



Ez már sokkal elfoghatóbb eredményeket adott

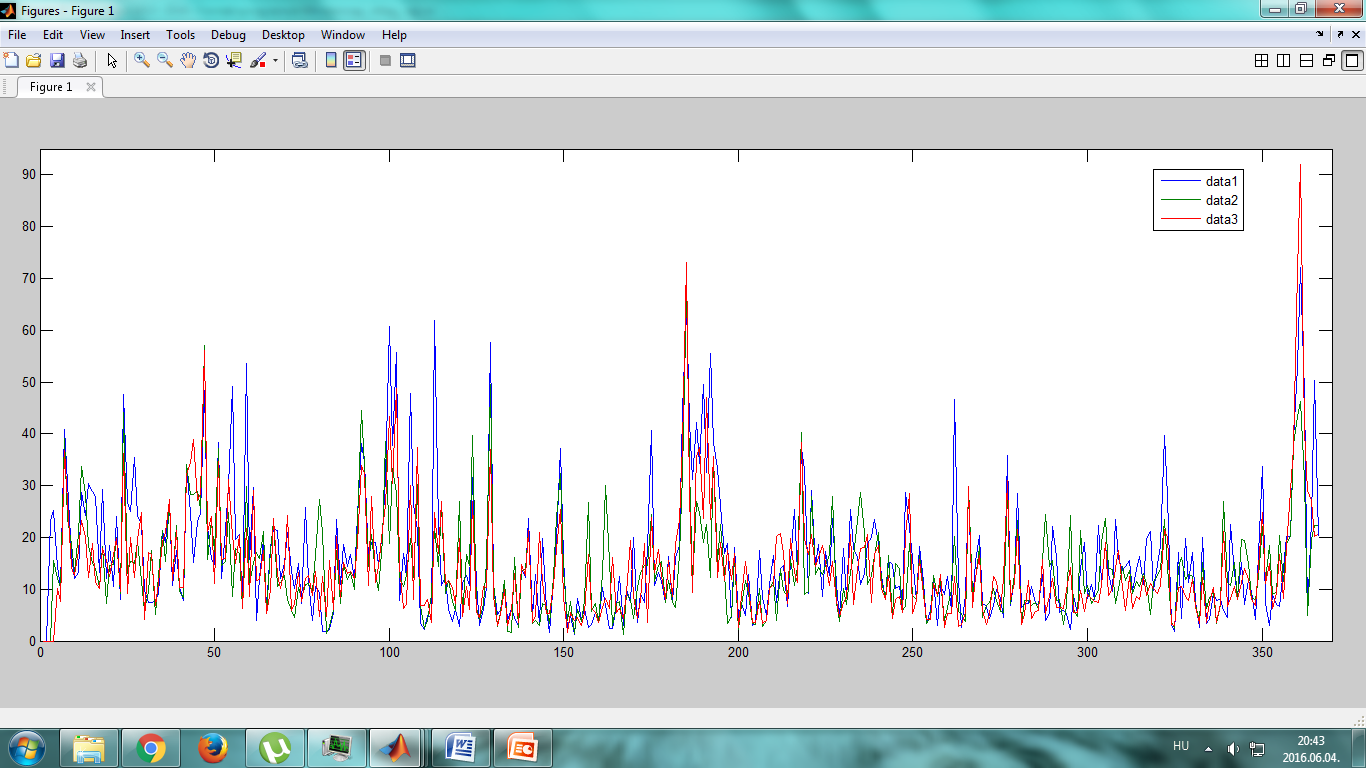
1 szomszéd = 16.1649

2 szomszéd = 14.1596

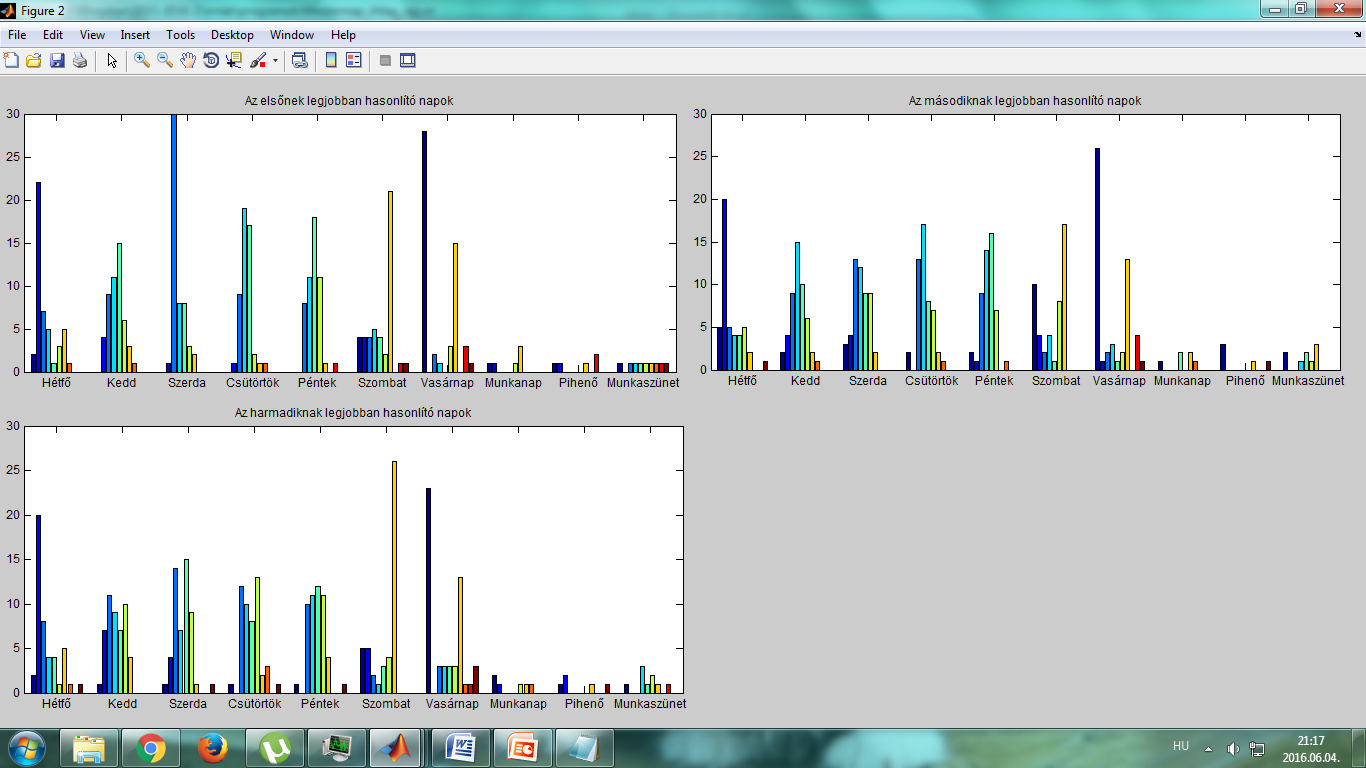
3 szomszéd = 14.2196

Sajnos rosszabb lett az eredmény ha már 3 szomszédot vettem figyelembe. Ezért megnéztem hogy mi lett a probléma.

A hibák alakulása az egyes napokhoz :



Itt sajnos jó látszik hogy van olyan eset amikor a három szomszéd rosszabb hibát ad mit a kettő vagy van olyan eset is amikor csak 1 megtalálása az legjobb döntés. Így kiegészítettem úgy az algoritmust számítsa ki mind a kettő és mind a három szomszéd esetén az talált napok hibájának átlagát és a legkisebb hibájú esetet adja hozzá a kimeneti mátrixhoz. Sajnos ez 14.7339 hibával működött. Ami nem lett jobb mint sima algoritmusok így felmerült a kérdés hogy mi lehet a baj. Vagy lehetne úgy keresni hogy ha hétfő van akkor csak a hétfők között keressünk. Ezért megnéztem hogy melyik naphoz melyiket találja meg.



Azt vettem észre hogy ha vasárnap volt az utolsó ismert nap akkor hétfőt találta meg legnagyobb valószínűséggel ami nem jó, hiszem a vasárnap után akár ünnep is lehet. Egyébként nagyon összevissza vannak a találta napok és keresett napok.

Ezekből az okokból kifolyólag készítettem egy olyan felállást ahol csak hétköznapokat vettem figyelembe.

Csak hétköznap

A nap mátrixnak innentől van igazán nagy értelme mert az alapján elsőnek létrehoztam egy új mátrixot amiben csak a hétköznapok vannak (1-5 sorszámig). Azok a hétköznapok amik ünnepek voltak vagy egyéb speciális napok azok is kikerültek a mátrixból. Így végül egy 246 elemű tömböt kaptam.

Ezek az adatokra is kipróbáltam az aránnyal módosított algoritmust és a többszomszédságot is.

1 szomszéd esetén = 14.1965

Kimenet aránnyal módosítva = 18.3529

Kimenet különbséggel módosítva = 21.9497

Sajnos itt sem voltak jók csak a sima aránnyal módosított esetek.

Majd jöttek a több szomszédos algoritmusok. Ezek már lényegesebb jobb arányt értek el. Bár sajnos úgy tűnik hogy a 246 hétköznapi munka nap kevés adatnak bizonyul három jó szomszéd megtalálásához.

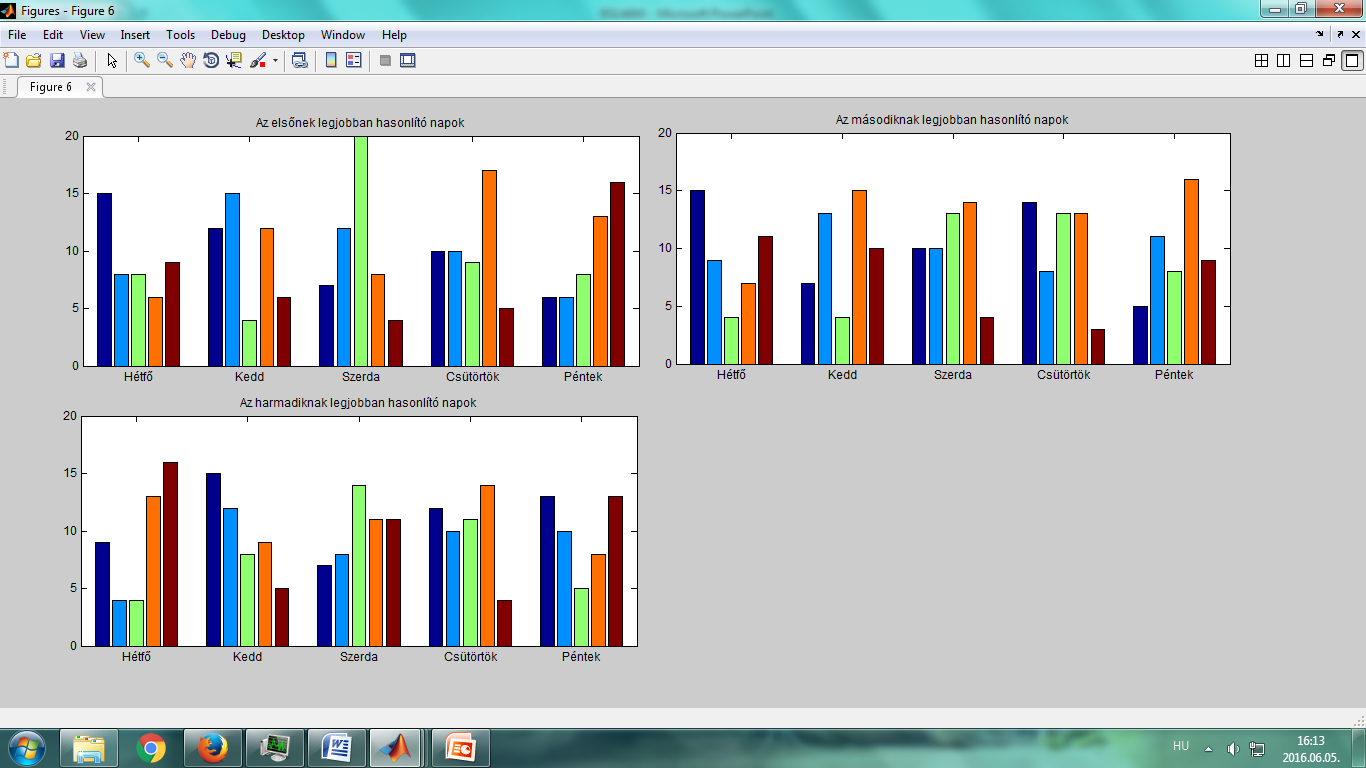
2 szomszéd átlaga = 12.3722

3 szomszéd átlaga = 14.3752

Változó átlag = 14.6856

Végül itt is rosszabb lett az a eset ha hol 3 hol 2 szomszédot veszek. Finomítanom kellesz rajta.

Itt is megnéztem hogy melyik naphoz melyik napot találja meg:



Itt már sokkal jobb aránnyal talál ugyan olyan típusú napot mint az előbb. De még mindig nehéz lenne azt kijelenteni hogy a hétfők a hétfőkre hasonlítanak.

Összességében jó ötlet volt ki hagyni a speciális napokat a vizsgálatból

Irodalomjegyzék

<http://www.tankonyvtar.hu/hu/tartalom/tamop425/0046_adatbanyaszat/ch05s02.html>

Használt rövidítések, fogalmak

Br : Bruttó hazai rendszerterhelés

SK : SK-HU határmetszéki áramlás

Legjobb szomszéd : legkisebb abszolút hibát okozó elem az eddigi ismert napokból.