BÁNKI KÖZLEMÉNYEK
ÉVFOLYAM 1. SZÁM

Nagylétszámú termelővállalatok állományi rendelkezésre állása karakterisztikájának leírása ARIMA modellel

Estimation of available staff of large production companies using mathematical ARIMA modelling

Fehér András István
Óbudai Egyetem, Biztonságtudományi Doktori Iskola, Budapest, Magyarország
feher.andras@uni-obuda.hu

***Összefoglalás* — Az iparvállalatok egyre növekvő problémája a rendelkezésre álló munkavállalói állomány bizonytalansága: hány emberrel számolhatnak az elkövetkezendő napokban, hetekben, hónapokban? [1] [2]**

Egyfelől az üzleti célok hatékonyságot ösztönző elvárásrendszere, másfelől a munkaerőpiac átalakulása készteti a nagyvállalatokat a mind pontosabb, és lehetőleg tudományos alapokra támaszkodó prediktív számításokra.

Feltételezésem szerint, megfelelő matematikai módszerrel, modellezhető egy adott állomány jövőbeli valószínű rendelkezésre állása, historikus adatok segítségével.

Jelen cikkben azt vizsgálom, hogy historikus és jogszerűen használható adatok ismeretében egyáltalán modellezhető-e egy adott munkavállalói állomány jövőbeli rendelkezésre állása, és meghatározható-e a teljes állomány beosztásának karakterisztikája.

Választásom a Box és Jenkins által kidolgozott ARIMA modellekre esett, a konkrét tapasztalati idősorra vonatkozó relevanciájuk, valamint matematikai és informatikai matematikai szempontból való jó kezelhetőségük miatt.

Kulcsszavak: időrelemzés, ARIMA modell, beosztástervezés, predikció

***Abstract* — The ever-increasing problem of industrial companies is the insecurity of the available staff: how many people can be counted on in the coming days, weeks or months? [1] [2]**

On one hand, the expectations of a business-to-business incentive (eg lean management) and on the other, the transformation of the labour market, force large companies to develop more precise and predictable calculations relying on scientific basis.I suppose a suitable mathematical algorithm could be the right method to use for modelling the probable availability of a given team of employees by means of historical data.

In the present article, I examine whether, with the knowledge of historical and legally usable data, the future availability of a given workforce can be modelled at all, and whether the characteristics of the headcount planning of the entire workforce can be determined.

I chose the ARIMA models developed by Box and Jenkins because of their relevance to the specific empirical time series and their good manageability from a mathematical and informatics point of view.

Keywords: time series, ARIMA model, headcount planning, predictivity

1. BEVEZETÉS

A témát bevezető cikkből [3] kiderült, hogy milyen hatása van a cégek üzleti folyamataira a megfelelően optimalizált beosztástervezésnek, illetve milyen jogi és üzleti keretek között kell (lehet) a feladatot elvégezni. Az állományi optimumkeresés célja, hogy mindenkor a megfelelő számú és kompetenciájú munkaerő álljon rendelkezésre. Amennyiben a kompetencia kérdéskörétől – amely nem tárgya e tanulmánynak – eltekintünk, marad a számosság kérdése.

A termelővállalatok, különböző paraméterek alapján, például megrendelési mennyiség, technológia, tervek stb., pontosan tudják, hogy mikor, hány embernek kell felvennie a munkát. A cél az, hogy ezt a számot a legjobban megközelítsük, és lehetőség szerint ne alulról. A megvalósulást befolyásolja, hogy a beosztott sokaság egy eddig tudományosan nem meghatározott része nem fog rendelkezésre állni az adott műszakkezdetkor. Kutatásom során arra keresem a választ, hogy lehet-e a korábban leírtaknál igazoltan pontosabb becslést adni a várhatóan megjelenők számáról oly módon, hogy a megoldás figyelembe vegye a vonatkozó adatvédelmi és egyéb előírásokat, és automatikus adatgyűjtés során rögzített adattömegből dolgozzon.

Jelen cikkben azt vizsgálom, hogy historikus adatok ismeretében egyáltalán modellezhető-e egy adott munkavállalói állomány jövőbeli rendelkezésre állása, és meghatározható-e a teljes állomány beosztásának karakterisztikája. E munka során, adatvédelmi okokból, nem dolgozhatok személyes adatokkal. Feltételezem azonban – és kutatásomat eleve így is kezdem –, hogy adott munkavállalói állomány jövőbeni rendelkezésre állása valószínűségi becsléséhez nincs szükség személyes adatok ismeretére, ezáltal a prediktív matematikai modell jogszerűen használható tetszőleges sokaságra.

2. MATEMATIKAI MODELL

Idősorelemzésre több módszert használhatunk, ám a különböző módszerek alkalmazhatóságának vannak előfeltételei (pl. stacionaritás, komponensre bonthatóság stb.), amelyek teljesülése határozza meg, hogy mely

módszer(ek) szerinti vizsgálattól várhatjuk az idősor karakterisztikájának megértését, és alkalmazható(k) megfelelően robusztus modellalkotásra.

Kutatásaim során azt tapasztaltam, hogy a különféle módszerek szerinti vizsgálatot sokszor akár az elemzést végző(k) tudásszintje, preferenciája, és egyéb, nem kizárólag tudományos szempontok befolyásolják. Adott idősor ugyanis többféle vizsgálat alapfeltételeit is teljesítheti, ám egyáltalán nem biztos, hogy a különféle modellek egyformán erősek lesznek. Az elvárható alapossághoz hozzátartozik, hogy megfelelő gondossággal körbejárjam a szóba jöhető modellek szerinti elemzéseket, és bizonyosságot szerezzek arról, hogy a megfelelő módszert alkalmazom.

Az egyes módszerek sok esetben akkor is alkalmazhatók (matematikailag levezethetők), ha az idősor nem teljesíti maradéktalanul az alapfeltételeket, azonban ezen vizsgálatok eredményeit fenntartásokkal kell fogadni. Egyszerű példával szemléltetve, egy nem működő mutatós óra is mutatja a pontos időt naponta kétszer, s ha épp akkor pillantunk rá, könnyen tekinthetjük megfelelően működőnek, holott nem az. Ily módon, egy, az adott módszer szerint nem megfelelően „teljesítő" idősorra alkalmazott modell is adhat fals pozitív eredményt.

Mielőtt rátérek a vizsgálatára, fontosnak tartom kitérni arra a kérdésre, hogy mi a helyzet a többi, leginkább már legújabbkori módszerek alkalmazhatóságával. Ilyen módszerek például a Neurális hálózat alapú becslések [4], a gépi tanulás (Machine Learning) [5], mélytanulás (Deep Learning) [6], vagy Bayesi becslés elméletek. [7] Ezek alkalmazhatóságának is megvannak. a kritériumai, a mi idősorunk év vége felé növekvő jellegéből adódó tulajdonsága nem teszi lehetővé ezek szabályos használatát [8]

Az idősorelemzés matematikájának nincs olyan általánosságban alkalmazható módszertana, melyet bármilyen idősorra alkalmazva azt mondhatnánk, hogy minden létező módszertant kipróbálva jutottunk az optimális eredményre. [9] Ezen a területen helye van a kutatói emberi tényezőnek, akinek feladata megtalálni a reálisan legjobb kompromisszumot. Hivatkozom a MOA elvre (Mission Oriented Application - feladatorientált alkalmazás, mely mint mutató arra vonatkozik, hogy az adott eszközt vagy megoldást milyen igényű feladatokra lehet alkalmazni), e témában is szem előtt kell tartani az elemzés célját.

Mindez még inkább alátámasztja, hogy a modell kiválasztásánál körültekintően kell eljárni, több szabályosan szóba jöhető modell esetén pedig érdemes az idősort mindazok szerint megvizsgálni (megjegyzés: a témával foglalkozó matematikusok sokszor „ránézésre" meg tudják mondani, merre érdemes indulni, ám ez hivatkozható tudományosságot nélkülöző megállapítás).

Jelen cikkben, a fenti szempontokat és módszereket a kutatásaim során megvizsgálva, a széles körben alkalmazott autoregresszív és mozgóátlag alapú modellalkotás lehetőségét elemzem. Feltételezésem szerint ugyanis a vizsgált idősorra alkalmazható valamely e tárgykörbe eső sztenderd eljárás.

3. AZ ADATHALMAZ

A vizsgálni kívánt adattömeg MySQL adatbázis-kezelő szoftverben áll rendelkezésre, amit a Login Autonom Kft. EASE++ Workhour és Holiday szoftverei táplálnak valós idejű, automatikus adatrögzítéssel. A folyamat első lépéseként az adatbázisból kinyertem (exportáltam) az anonimizált adatokat Microsoft Excel táblázatkezelő szoftverbe, amit a kezelhetőség miatt célszerű megtenni. Az így kinyert adathalmazt a cél és a vonatkozó adatvédelmi elvárásoknak megfelelően leszűrtem, az alábbiak szerint.

Arra voltam kíváncsi, hogy hányan voltak mikorra beosztva, és ezek közül hányan nem jöttek be, előzetes értesítés nélkül (a pontos személy nem, csak sokaság vizsgálható – ld. Nagylétszámú termelővállalatok állományi kapacitás becslése című cikk 5. fejezete [3]). Időtáv szerint nem releváns, hogy a beosztás mennyivel a műszakkezdés előtt történt, csak az, hogy végül hányan vették fel a műszakot, illetve a nem megjelent állományból hányan nem szóltak előtte, hogy nem fognak dolgozni.

A pontosság kedvéért érdemes megvizsgálni, hogy miért mindegy a beosztás időpontja, amennyiben az minimum 96 órával műszakkezdés előtt történt, és honnan tudható, hogy nem szóltak időben előre a távolmaradásról, továbbá, hogy mit jelent időben szólni.

Beosztás időpontja: vállalatonként eltér a gyakorlat. A legelterjedtebb módszer – ahol van szoftveres segítség, mint például a vizsgált cég esetében használt EASE++ –, hogy egy évre előre beosztják a teljes állományt és minden munkarendet. A munkavégzés napjának közeledtével, de maximum egy héttel azt megelőzően [10], a kivételek (szabadság, betegség, kilépés stb.) és aktuális üzleti kihívások alapján pontosítják a beosztást.

Helyenként létezik a csak heti-kétheti beosztás, míg máshol havi-negyedévi, majd heti pontosítás. A lényeg, hogy teljesüljön a legalább 96 órás kritérium.

Időben szólni: a Munka törvénykönyve vonatkozó paragrafusai szerint a napi munkaidő kezdetét megelőzően legalább 96 órával korábban jelezni kell a munkavállaló felé, ha a beosztásában változás következik be. Ellenkező esetben - a módosításba történő beleegyezése esetén is – pótlék jár neki az új beosztás szerinti ledolgozott óráira, határidőn túli átosztás miatt [11]. Tekintettel arra, hogy ez a munkáltató számára kerülendő többletköltséget jelent, az „időben szólni, hogy nem fogok tudni munkába állni" minimum 4 nappal korábbi bejelentést feltételez.

Honnan tudható, hogy nem szólt időben: fontos tisztázni egy adattömeg elemzésekor, hogy milyen információk állnak rendelkezésünkre, illetve azok milyen releváns jelentést hordoznak magukban. Ez esetben, anonimizált adatokról lévén szó, a beosztás/átosztás időpontja utal a felvetett kérdésre. Amennyiben ugyanis az adott műszakkezdést megelőző 96 órán belül történt a módosítás – akár a műszakot követő időszakban (azaz az óraadat negatív szám) –, úgy a fentieket kimerítettük, a feltevés igazolt.

Alább a személyes adatokat nem tartalmazó, részlegesen szűrt, kiindulási Excel fájl egy részletének képe látható.

date	emp_id	name	wsbu_created_on	ea_status	ea_created_on	Egységenkénti beosztás Hány nap telt el a beosztott nap é.a beosztás között	Egységentkénti beosztást módosították Egyéni beosztásra ennyi nappal az adott munkanap előtt	Távollét kiírása az adott munkanaphoz képest napban
2019-01-02 00:00	511038 12 órás de (06:00-18:00)		2017-09-26 08:19	2	2019-01-29 08:14	463	EMPTHY	-27
2019-01-03 00:00	511038 12 órás de (06:00-18:00)		2017-09-26 08:19	2	2019-01-29 08:14	464	EMPTHY	-26
2019-01-04 00:00	511038 12 órás de (06:00-18:00)		2017-09-26 08:19	2	2019-01-29 08:14	465	EMPTHY	-25
2019-01-05 00:00	511038 Pihenőnap		2017-09-26 08:19	2	2019-01-29 08:14	466	EMPTHY	-24
2019-01-06 00:00	511038 Pihenőnap		2017-09-26 08:19	2	2019-01-29 08:14	467	EMPTHY	-23
2019-01-07 00:00	511038 Pihenőnap		2017-09-26 08:19	2	2019-01-29 08:14	468	EMPTHY	-22
2019-01-08 00:00	511038 12 órás du (18:00-06:00)		2017-09-26 08:19	2	2019-01-29 08:14	469	EMPTHY	-21
2019-01-09 00:00	511038 12 órás du (18:00-06:00)		2017-09-26 08:19	2	2019-01-29 08:14	470	EMPTHY	-20
2019-01-10 00:00	511038 12 órás du (18:00-06:00)		2017-09-26 08:19	2	2019-01-29 08:14	471	EMPTHY	-19
2019-01-11 00:00	511038 Pihenőnap		2017-09-26 08:19	2	2019-01-29 08:14	472	EMPTHY	-18
2019-01-12 00:00	511038 Pihenőnap		2017-09-26 08:19	2	2019-01-29 08:14	473	EMPTHY	-17
2019-01-13 00:00	511038 Pihenőnap		2017-09-26 08:19	2	2019-01-29 08:14	474	EMPTHY	-16
2019-01-14 00:00	511038 12 órás de (06:00-18:00)		2017-09-26 08:19	2	2019-01-29 08:14	475	EMPTHY	-15
2019-01-15 00:00	511038 12 órás de (06:00-18:00)		2017-09-26 08:19	2	2019-01-29 08:14	476	EMPTHY	-14
2019-01-16 00:00	511038 12 órás de (06:00-18:00)		2017-09-26 08:19	2	2019-01-29 08:14	477	EMPTHY	-13
2019-01-17 00:00	511038 Pihenőnap		2017-09-26 08:19	2	2019-01-29 08:14	478	EMPTHY	-12
2019-01-18 00:00	511038 Pihenőnap		2017-09-26 08:19	2	2019-01-29 08:14	479	EMPTHY	-11
2019-01-19 00:00	511038 Pihenőnap		2017-09-26 08:19	2	2019-01-29 08:14	480	EMPTHY	-10
2019-01-20 00:00	511038 12 órás du (18:00-06:00)		2017-09-26 08:19	2	2019-01-29 08:14	481	EMPTHY	-9
2019-01-21 00:00	511038 12 órás du (18:00-06:00)		2017-09-26 08:19	2	2019-01-22 10:25	482	EMPTHY	-1
2019-01-22 00:00	511038 12 órás du (18:00-06:00)		2017-09-26 08:19	2	2019-01-22 10:27	483	EMPTHY	0
2019-01-23 00:00	511038 Pihenőnap		2017-09-26 08:19			484	EMPTHY	EMPTHY
2019-01-24 00:00	511038 Pihenőnap		2017-09-26 08:19			485	EMPTHY	EMPTHY
2019-01-25 00:00	511038 Pihenőnap		2017-09-26 08:19			486	EMPTHY	EMPTHY
2019-01-26 00:00	511038 12 órás de (06:00-18:00)		2017-09-26 08:19	2	2019-01-22 10:27	487	EMPTHY	4
2019-01-27 00:00	511038 12 órás de (06:00-18:00)		2017-09-26 08:19	2	2019-01-22 10:27	488	EMPTHY	5
2019-01-28 00:00	511038 12 órás de (06:00-18:00)		2017-09-26 08:19	2	2019-01-22 10:27	489	EMPTHY	6
2019-01-29 00:00	511038 Pihenőnap		2017-09-26 08:19			490	EMPTHY	EMPTHY
2019-01-30 00:00	511038 Pihenőnap		2017-09-26 08:19			491	EMPTHY	EMPTHY
2019-01-31 00:00	511038 Pihenőnap		2017-09-26 08:19			492	EMPTHY	EMPTHY
2019-02-01 00:00	511038 12 órás du (18:00-06:00)		2017-09-26 08:19	2	2019-01-22 10:27	493	EMPTHY	10
2019-02-02 00:00	511038 12 órás du (18:00-06:00)		2017-09-26 08:19	2	2019-01-22 10:27	494	EMPTHY	11
2019-02-03 00:00	511038 12 órás du (18:00-06:00)		2017-09-26 08:19	2	2019-01-22 10:27	495	EMPTHY	12
2019-02-04 00:00	511038 Pihenőnap		2017-09-26 08:19			496	EMPTHY	EMPTHY
2019-02-05 00:00	511038 Pihenőnap		2017-09-26 08:19			497	EMPTHY	EMPTHY
2019-02-06 00:00	511038 Pihenőnap		2017-09-26 08:19			498	EMPTHY	EMPTHY
2019-02-07 00:00	511038 12 órás de (06:00-18:00)		2017-09-26 08:19	2	2019-01-22 10:27	499	EMPTHY	16
2019-02-08 00:00	511038 12 órás de (06:00-18:00)		2017-09-26 08:19	2	2019-01-22 10:27	500	EMPTHY	17
2019-02-09 00:00	511038 12 órás de (06:00-18:00)		2017-09-26 08:19	2	2019-01-22 10:27	501	EMPTHY	18
2019-02-10 00:00	511038 Pihenőnap		2017-09-26 08:19			502	EMPTHY	EMPTHY

1. ábra: Részlegesen szűrt adatokat tartalmazó Excel-részlet, forrás: saját szerkesztés

Az adatszűrést követően az Excelt konvertáltam CSV (Comma-Separated Values) text fájlba, amit a későbbi felhasználás miatt célszerű megtenni. Maradhatott volna Excelben is, de a matematikai programnyelvben történő további használat miatt döntöttem az egyszerűbben kezelhető CSV mellett. [12]

nyers_szűrt_adatok_anonim

Egységenkénti beosztás Hány nap telt el a beoszto	Egységentkénti beosztást módosították Egyéni beosztásra ennyi nappal az adott	Távollét kiírása az adott munkanaphoz képest nap
523	0	2
91	2	0
5	0	0
6	1	1
7	2	2
464	-1	-1
70	0	1
71	1	2
444	0	-1
408	2	2
61	2	2
238	0	0
238	0	0
44	-1	-1
84	-1	-1
528	0	0
48	0	0
49	1	1
147	-1	-1
148	0	0
149	1	1

2. ábra: Szűrt adatokat tartalmazó CSV-részlet, forrás: saját szerkesztés

Az értelmezett és szűrt CSV formátumú adathalmazt betöltöttem R programnyelvbe, feldolgozásra.

4. MATEMATIKAI LEVEZETÉS

Besenyei szerint: „Az AR folyamatokkal általában azokat az idősorokat modellezhetjük, amelyekről feltehetjük, hogy jelen idejű értékeik alakulásában a közvetlen múlton kívül a véletlen hiba is beleszól” [13]

Az elmélet gyökere egészen az 1920-as évekre nyúlik vissza, ám a Box és Jenkins által kidolgozott ARIMA modellekkel vált lehetségessé idősorokra vonatkozó összetettebb elemzés elvégzése. [14] Széleskörű elterjedését az informatika fejlődése hozta el az elmúlt évtizedekben, és azon tulajdonsága révén vált népszerűvé, hogy matematikai szempontból jól kezelhetők, és a folyamatok egy elég általános osztályát képviselik, mindamellett jól is automatizálható maga az elemzési eljárás. Utóbbi tulajdonságát is látni fogjuk, először azonban tisztázom az alapfogalmakat.

Autoregresszív folyamat: az diszkrét paraméterű Y_t sztochasztikus folyamatot k-ad rendű autoregresszív folyamatnak nevezzük, ha [15]

$$Y_t = \alpha_1 \times Y_{t-1} + \cdots + \alpha_k \times Y_{t-k} + \varepsilon_t$$

Ahol:

- konstansok α_i
- fehér zaj (várható értéke 0, szórása σy) Y_t

Mozgóátlag folyamat: az diszkrét paraméterű Y_t sztochasztikus folyamatot k-ad rendű mozgóátlag folyamatnak nevezzük, ha [16]

$$Y_t = \beta_0 \times U_t + \beta_1 \times U_{t-1} + \cdots + \beta_k \times U_{t-k}$$

Ahol:

- konstansok β_k
- diszkrét fehér zaj (várható érték 0, szórás σu) U_t

AR és MA folyamatokra jellemző, egymásból kifejezhetők, és mindkét esetben különböző rendeket különböztethetünk meg:

- AR(p)
- MA(q), ahol p és q a folyamat rendjét jelenti

ARMA modell: autoregresszív és mozgóátlag modellek (Autoregressive and Moving Average) a sztochasztikus idősorelemzés leginkább elterjedt módszere, amely az AR és MA folyamatokat egyesíti. [17]

Az autoregresszív (AR) modelltag az idősor jelenlegi értékét saját előző értékeinek függvényében fejezi ki;

A mozgóátlag (MA) modelltag az idősor jelenlegi értékét a jelenlegi és a múltbeli véletlen változók függvényében fejezi ki.

A paraméterek megállapítása általában empirikus idősor alapján történik, azaz ARMA (p,q): [18]

$$Y_t = \alpha_1 \times Y_{t-1} + \alpha_2 \times Y_{t-2} + \cdots + \alpha_p \times Y_{t-p} + \varepsilon_t + \beta_1 \times \varepsilon_{t-1} + \cdots + \beta_q \times \varepsilon_{t-q}$$

Ahol:

- fehér zaj ε_t

ARIMA (p,d,q): autoregresszív integrált mozgóátlag modell (Autoregressive Integrated Moving Average), mely megengedi a stacionárius transzformációkat (differenciálás, logaritmizálás) is. [19]

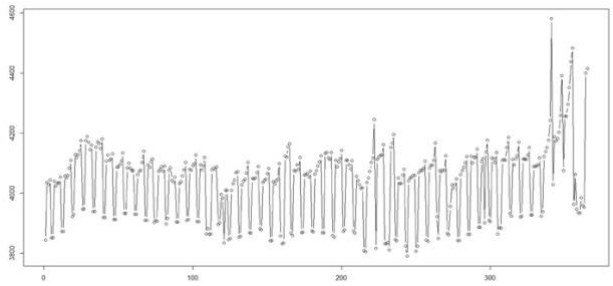
- p = autoregresszió rendje
- d = differenciák száma
- q = mozgóátlag rendje

Léteznek még fentieken kívül FARIMA, SARIMA, VARIMA és egyéb modellek is, melyekre azonban nem térek ki azok kutatásomra vonatkozó alacsony relevanciája miatt. [20]

4.1 Stacionaritás vizsgálat

Láttuk, hogy a különféle modellek szabályos alkalmazhatóságának vannak kritériumai. Az ARMA modell esetén a függvény stacionaritása a feltétel. Ebben az összefüggésben ez azt jelenti, hogy az idősor jellemzői időben állandók, azaz függetlenek a t időváltozótól. [21]

Jellege miatt, ránézésre a mi adatsorunk is stacionáriusnak tűnik (bár a vége felé kissé kiugrik), ám ez nem elég a feltétel teljesítésének igazolására.



3. ábra: Teljes tapasztalati idősor az összbeosztottakra, forrás: saját szerkesztés

A megfelelő igazolásra vannak különböző statisztikai próbák, melyeket helyesen alkalmazva, megkapjuk a választ arra a kérdésre, hogy az idősorunk valójában stacionárius-e vagy sem.

Ezek jellemzően és jellegű, egymást kizáró, ám egyben kiegészítő feltevések igazolásán alapuló algoritmusok, melyek matematikai levezetése túlmutat jelen cikk keretein. Létezésüket, és használatuk, továbbá értelmezésük módját azonban ismerni kell, hogy az adott szoftverben megfelelően alkalmazni tudjuk őket. [22]

A próbák statisztikai alapon működnek, és nem tudjuk a becslésük eloszlását amely alapján tudnánk a valószínűségüket. A megfelelően konzervatív megközelítés miatt három tesztet használok, és csak akkor fogadom el az eredményt, ha mindhárom megegyezik. [23] Az elemzést végzőn múlik az elfogadási döntés, én azonban a konzervatív utat választottam. A vizsgálathoz az alábbi teszteket használtam:

- Augmented Dickey-Fuller (ADF) teszt [24]
- Phillips-Perron Unit Root (PP) teszt [25]
- Kwiatkowski-Phillips-Schmidt-Shin (KPSS) teszt [26]

A teszt típusára vonatkozóan figyelembe kell venni, hogy az ADF és PP egységgyök (Unit Root) típusú tesztek, azaz a próba nullhipotézise szerint az idősor nem tekinthető stacionáriusnak, transzformációra van szükség. A KPSS teszt ehhez képest ellenkező eredmény esetén adja ugyanazt a konklúziót. R-ben futtatás után az alábbiakat kaptam:

```
> adf.test(x1)

        Augmented Dickey-Fuller Test

data:  x1
Dickey-Fuller = -3.9952, Lag order =
7, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(x1) : p-value smaller
than printed p-value
> pp.test(x1)

        Phillips-Perron Unit Root Test

data:  x1
Dickey-Fuller Z(alpha) = -180.65,
Truncation lag parameter = 5,
p-value = 0.01
alternative hypothesis: stationary

Warning message:
In pp.test(x1) : p-value smaller than
printed p-value
> kpss.test(x1)

        KPSS Test for Level
Stationarity

data:  x1
KPSS Level = 0.94032, Truncation lag
 parameter = 5, p-value = 0.01

Warning message:
In kpss.test(x1) : p-value smaller
than
 printed p-value
```

4. ábra: Statisztikai tesztek az eredeti idősorra, forrás: saját szerkesztés

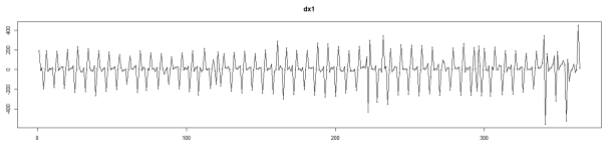
- ADF-teszt: alternatív hipotézis az, hogy stacionárius p=0,01, azaz , tehát az igaz, p értékre szignifikáns az eltérés, stacionaritás igazolt [27]
- PP-teszt: alternatív hipotézis az, hogy stacionárius p=0,01, azaz , tehát az igaz, p értékre szignifikáns az eltérés, stacionaritás igazolt
- KPSS-teszt: nem egységgyök típusú teszt, tehát fordítva működik p=0,01, azaz , tehát az igaz. Ez alapján azonban az idősor egységgyök, azaz nem stacionárius

Fenti eredmény alapján két megválaszolandó kérdés van:

- Elfogadjam-e a 2:1 arányú, stacionaritásra utaló végeredményt vagy sem, illetve
- nem elfogadás esetén elvessem-e az autoregresszív és mozgóátlag modellekkel való további vizsgálatokat.

Az első kérdésre a szakirodalom jellemző válasza, hogy érdemes „hinni” a negatív eredménynek, és nem elfogadni a függvény stacionárius mivoltát. [28] Ezek alapján én is így tettem, ismét hangsúlyozandó, hogy akár bármelyik fenti teszt eredménye önmagában is tekinthető volna eredménynek, a már leírt bizonytalanságok észben tartásával. A rossz döntés kockázata az egyértelműen rossz, vagy fals pozitív elemzési végeredmény kockázata. Ebből az eredményből az következik, hogy az idősorra ARMA modellt nem tudunk illeszteni, mert az eredeti függvény nem felelt meg a stacionaritás kritériumának. A fejezet elején leírtak szerint csak szabálytalanul lehetne alkalmazni a modellt.

Az ARIMA modell azonban épp ilyen esetekre áll rendelkezésünkre, így a következő lépésben, mivel lineáris trendtagunk van (ld. 3.1.5), differenciálom az idősort, majd azt is megvizsgálom. Az alábbi ábra az egyszeres differenciálás utáni idősort ábrázolja.



5. ábra: Az egyszeresen differenciált idősor, forrás: saját szerkesztés

Érdemes megjegyezni, hogy ez a lépés típusú e^x időfüggvényekre nem lenne eredményesen alkalmazható a differenciálás után megmaradó eredeti függvényjelleg miatt, azokban az esetekben más módszer áll rendelkezésünkre. [29]

Nem ránézésre kell eldöntenünk a függvény jellegét, ám ezen az ábrán már igen szembetűnő a stacionárius jelleg. Lefuttatva a tesztet a már differenciált idősorra, a következőket kapjuk:

```
> adf.test(dx1)

        Augmented Dickey-Fuller Test

data:  dx1
Dickey-Fuller = -8.3927, Lag order =
7, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(dx1) : p-value smaller
than printed p-value
> pp.test(dx1)

        Phillips-Perron Unit Root Test

data:  dx1
Dickey-Fuller Z(alpha) = -285.67,
Truncation lag parameter = 5,
p-value = 0.01
alternative hypothesis: stationary

Warning message:
In pp.test(dx1) : p-value smaller than
printed p-value
> kpss.test(dx1)

        KPSS Test for Level
Stationarity

data:  dx1
KPSS Level = 0.030187, Truncation lag
parameter = 5, p-value = 0.1

Warning message:
In kpss.test(dx1) : p-value greater
than printed p-value
```

6. ábra: Statisztikai tesztek a differenciált idősorra, forrás: saját szerkesztés

- ADF-teszt: stacionaritás igazolt
- PP-teszt: stacionaritás igazolt
- KPSS-teszt: p=0,1, azaz $p > 0{,}05$, tehát az igaz. Ez H_0 alapján a stacionaritás igazolt

Mindhárom statisztikai teszt a függvény stacionaritását igazolta, ami alapján már egyértelműen kimondható a stacionárius jelleg.

Miután egyszeres differenciálás útján értem el a stacionárius jelleget, megvan a d paraméterünk, d=1. Ebből az következik, hogy tudok ARIMA modellt illeszteni, és a következő lépésekben megkeresem a p és q paramétereket.

4.2 Paraméter meghatározás

Mint azt az idősorelemzés kapcsán már többször tapasztaltuk, több úton lehet elindulni ez esetben is. Erre a feladatra is található több, már meglévő és alkalmazható algoritmus (pl. Schwarz, Akaike, Hannan – Quinn stb.). [30]

Az én választásom a Hyndman-Khanadakar algoritmusra esett, amelynek alkalmazási feltétele, hogy d_{max} =2 legyen, tehát maximum másodrendű differenciálással elért stacionaritás esetén használható szabályosan. [31]

Az elemzés lépései a következők: [32]

1. „d” paraméter vizsgálat: ezt a feltételt a mi adatsorunk d=1 értékkel teljesíti (ld. fent).
2. Az ARIMA-modell felírása, azaz az idősor paramétereinek és a leírására alkalmas modellnek a meghatározása. Ennek során 4 modell illesztése: ARIMA(0,d,0)(2,d,2)(1,d,0)(0,d,1). Ha d=0 vagy 1, akkor (0,d,0)konstans nélkül is illesztünk, ez esetben (ami a mi esetünk is), összesen 5 modellt.
3. A kapott öt érték közül megkeressük a legkisebb értéket, és vele elkezdjük a modell illeszkedésének tesztelését, javítását. Erre is több eszköz létezik, én az Akaike-féle információ kritériumot, az AIC-t alkalmazom.

   Az AIC (Akaike Information Criterion) egy mérőszám (2002 óta egy továbbfejlesztett, azaz korrigált AIC (AICc)), ami adott idősorra megmutatja, hogy egy modell mennyire illeszkedik jól. [33]
4. Kiválasztjuk a legkisebb értéket, variáljuk p és q értékét +/– 1-gyel, megnézzük arra az összeget, és tesszük ezt mindaddig, amig nem találunk olyan modell-t, amire nincs lokálisan kisebb AICc összeg.
5. Előrejelzés készítése az eredmény alapján.

Fenti lépéssor szerint a Hyndman-Khanadakar algoritmussal végigszámolva, az alábbiakat kaptam R-ben:

```
ARIMA(2,1,3)

Coefficients:
          ar1      ar2      ma1
ma2      ma3
      -0.4421  -0.7685  -0.1510
0.0744  -0.7584
s.e.   0.0486   0.0670   0.0551
0.0788   0.0458

sigma^2 estimated as 10819:  log
likelihood=-2206.02
AIC=4424.04   AICc=4424.28
BIC=4447.43
```

7. ábra: p és q tagok eredménye Hyndman-Khanadakar alapján számítva, forrás: saját szerkesztés

Két AR és három MA tag lett, ebből p=2, q=3. Korábbról d=1, így 2,1,3 típusú ARIMA modellt kaptam, mely felírva ARIMA (2,1,3,).

Vegyük észre, hogy AIC=4424,04 és AICc=4424,28 között ez esetben csupán 0,24 a különbség. Más esetben nagyobb eltérést is adhat a két mérőszám, de nem nagyságrendit (nekem a számítások során 20 körül volt a legnagyobb eltérés). Mindebből az következik, hogy amennyiben erre az idősorra csak az AIC-t használnánk, nem kapnánk modellalkotás során szignifikánsan különböző eredményt (az AIC megalkotója, Hirotugu Akaike a korrigált mérőszámot kisebb adatsorokra értelmezte, ahol az AIC és AICc közötti különbség nőhet). [33]

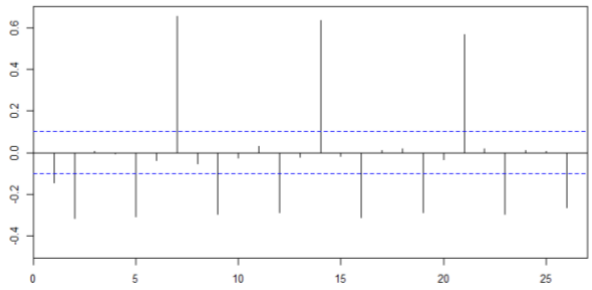
4.3 Autokorreláció vizsgálat (ACF és PACF)

ARMA és ARIMA modellekről beszélve fontos ismerni egy másik gyakran használt módszert, mellyel szintén el lehet dönteni, hogy szükséges-e a differenciálás, azaz stacionárius-e egy idősor, illetve következtethetünk az AR és MA tagokra is: az ACF függvényről van szó.

A vizsgálat lényege, hogy az eredeti idősorra felvesszük az ACF függvényt, és az autokorrelációs együtthatók értékeinek jellege alapján (majdnem egyformák, vagy csak lassan, esetleg gyorsan csökkennek) eldönthető, hogy indokolt-e a differenciaképzés. Ezt mindaddig folytatjuk (általában maximum 3-szor), ameddig nem kapunk stacionárius jellegre utaló korrelogramot. [34]

Az úgynevezett részleges autokorreláció függvény (PACF) az ACF függvényből számítható ki, és jellemzően az AR együtthatókat határozza meg, így a szignifikáns értékei alapján becsülhető az illesztendő modell AR tagjainak száma. [35]

Az ábrák alapján történő elemzés azonban gyakorlatot igényel, ezáltal kevésbé automatizálható, mint az előző pontban ismertetett eljárás. ARMA típusú elemzéseknél azonban sokszor találkozhatunk vele, és némi elemzési gyakorlatot követően érdekes támpontokat tud nyújtani az idősor jellegét illetően, így a teljeskörűség jegyében magam is felrajzoltam őket, a már differenciált függvényre.

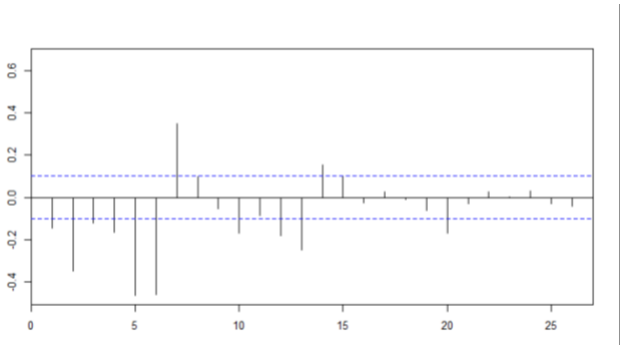


8. ábra: ACF diagram az egyszeresen differenciált idősorra, forrás: saját szerkesztés

Az ábrán látjuk, hogy nincs szignifikáns lecsengés, tehát stacionárius az idősor, továbbá 7-es lag-nél van egy jelentős korreláció. Tekintettel arra, hogy most nem dekompozíciós modellvizsgálatot csinálok, nem kötelező vele foglalkoznom.

Egy másik típusú függvény, az úgynevezett parciális autokorreláció függvény (PACF) függvény, melynek célja, hogy a magasabb rendű autokorrelációk hatását megtisztítsa az alacsonyabb rendű autokorrelációk hatásaitól, ezáltal segíti az összefüggések megértését. Úgy is fogalmazhatunk, hogy felszínre hozza a mélyebben rejlő korrelációkat. A PACF az autokorreláció függvényből számítható ki, és az autoregresszív (AR) tag p kezdeti értékének eldöntésében segít a szignifikáns értékei alapján történő becsléssel. [36]

A mi differenciált idősorunk PACF függvényképe az alábbi:



9. ábra: PACF diagram az egyszeresen differenciált idősorra, forrás: saját szerkesztés

Ebből két jelenséget vehetünk észre. Sinusos lecsengésű, ami szintén a (differenciált) idősor stacionaritására utal, illetve a fentebb már konstatált szezonalitás hatására az autokorrelációs együtthatók értékei a szezonális komponens hatásának megfelelően hullámoznak. [36]

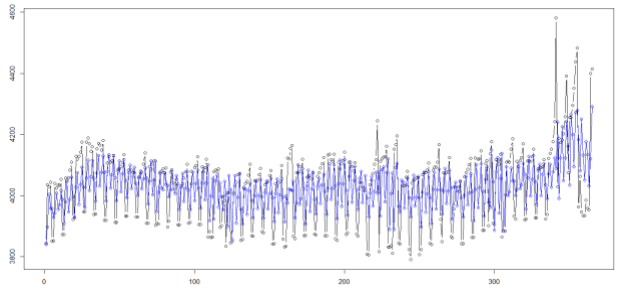
Tekintve, hogy a modell tagszámait korábban már megfelelő bizonyossággal kiszámoltuk, illetve az ACF és PACF is igazolták a modell helyességét, következő lépés a modellalkotás.

4.4 Modellalkotás

Az ARIMA (2,1,3) modell az eredeti képlet szerint

$$Y_t = \alpha_1 \times Y_{t-1} + \alpha_2 \times Y_{t-2} + \cdots + \alpha_p \times Y_{t-p} + \varepsilon_t + \beta_1 \times \varepsilon_{t-1} + \cdots + \beta_q \times \varepsilon_{t-q}$$

p=2 és q=3 értéket ad. Ezt R-ben lefuttatva és ábrázolva, az összbeosztottak idősorára illesztem a kapott ARIMA modellt (kékkel).



10. ábra: Eredeti idősorra illesztett ARIMA modell, forrás: saját szerkesztés

Szembetűnő jelenség, hogy az eredeti idősor képéhez hasonlóan, a modell is mutat egyfajta sztochasztikus jelleget.

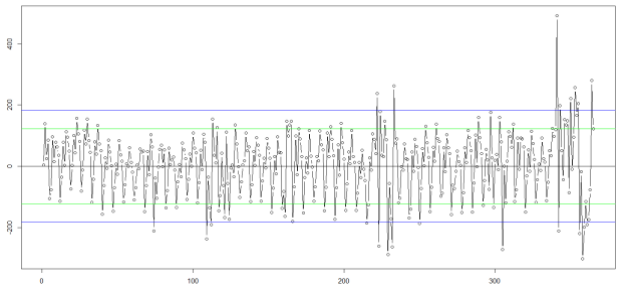
Ez az ARMA típusú módszerek azon elvéből adódik, mely szerint a véletlennek is jelentős hatást tulajdonítanak, a modellalkotás jelentős eleme a hibatag karakterisztikájának reprodukálása – ha az egyszerűség

kedvéért a determinisztikus szemlélet elnevezései szerint fogalmazunk.

Az eredmény tehát egy kevésbé „művi" kinézetű modell, annak vizsgálata azonban még hátravan, hogy matematikailag mennyire jó.

4.5 Hibatag vizsgálat

A modell hibatag vizsgálata során a fenti modellel korrigált eredeti idősorra bejelöltem zölddel az 1-szeres, és kékkel a 1,5-szeres szórást.



11. ábra: Eredeti idősor ARIMA modellel korrigálva, forrás: saját szerkesztés

A kiugró elemeket a következők miatt nem jelöltem külön: az egyszeres szóráson kívül 79 darab pont esik – az összesen 26 darab outliernél lényegesen több – , ami 78,4%-os egyezőség. Másfélszeres szórásnál 22 darab kívül eső elemmel 93,9% a becslés pontossága.

Ezek alapján kijelenthető, hogy az ARIMA modell 90% feletti szignifikanciája kiváló eredmény.

5. ÖSSZEGZÉS, KÖVETKEZTETÉS

Az idősorelemzés témakörét kutatva, szembetűnő a megközelítések és módszerek nagy (és dinamikusan növekvő) száma, továbbá az elemzések elvégzésének nagy szabadságfoka. Nem létezik olyan módszertan, melyet bármilyen idősor esetén alkalmazva, garantált jó eredményt kapnánk, azaz az idősorelemzés nehezen automatizálható. Természetesen, valamilyen megfelelően szűkített feltételrendszer szerinti idősorok esetén némileg árnyaltabb a kép, de a kutatói tapasztalat és lelkiismeretesség elkerülhetetlen a mindenkori céloknak megfelelő eredmény elérése céljából.

A jelen cikkben elemzett munkavállalói beosztásokat tartalmazó idősor, bár konkrét vállalat konkrét számadatait tartalmazza, az általános ipari beosztási gyakorlatot is jól szemlélteti, annak jogszabályi és szakmai korlátai miatt. Másképpen fogalmazva, a különböző cégek beosztási idősorainak abszolútértékei és bizonyos komponensei változhatnak ugyan, de jellege nagymértékben nem.

Annak érdekében, hogy kutatási feladatomat a fentebb említett elvárható lelkiismeretességgel és alapossággal végezzem, megvizsgáltam az adott idősor elemzésére szabályosan alkalmazható, sztenderd eljárásokat. Ezek alapján az ARMA típusú modellalkotás tűnt megfelelőnek, és sikerült is bebizonyítanom a feltevés helyességét. A végül eredményt hozó ARIMA modell 90% feletti szignifikanciája önmagában igen erős. A feltételezésem, miszerint létezik egy adott munkavállalói állomány jövőbeni rendelkezésre állásához optimálisan alkalmazandó, meglévő sztenderd idősorelemző modell, igazoltam. A témakör kutatójaként fontos is ismerni a módszert, ugyanis bizonyos karakterisztikájú idősorok esetén lehetséges, hogy ezt az utat kell járni.

FELHASZNÁLT IRODALOM

[1] PORTFOLIO: Már nem csak a munkaerőhiány sújtja a magyar gazdaságot, https://www.portfolio.hu/gazdasag/mar-nem-csak-a-munkaerohiany-sujtja-a-magyar-gazdasagot.267235.html, Letöltés ideje: 2020. január 31., 22:49, 2017. november 8.

[2] KSH: Összefoglaló táblák (STADAT) - Idősoros éves adatok – Munkaerőpiac, 2020., https://www.ksh.hu/stadat_eves_2_1, Letöltés ideje: 2020. szeptember 28., 22:43

[3] FEHÉR, A., „Nagylétszámú termelővállalatok állományi kapacitásbecslése", Bánki Reports: 2020. – befogadás alatt

[4] BISHOP, C. M., „Neural networks for pattern recognition", [ISBN 978-0198538493], Clarendon Press, 1995.

[5] NILSSON, N. J., „Introduction to Machine Learning", Stanford University: California, 2005.

[6] SCHULZ, H., BEHNKE, S., "Deep Learning", KI - Künstliche Intelligenz. 26 (4) pp. 357–363., [doi:10.1007/s13218-012-0198-z. ISSN 1610-1987], 2012.

[7] HUNYADI, L., „Bayesi gondolkodás a statisztikában", Statisztikai Szemle, 89. évfolyam 10–11. szám

[8] ZHANG, G. P., „Time series forecasting using a hybrid ARIMA and neural network model", [PII: S0925-2312(01)00702-0], Department of Management, J. Mack Robinson College of Business, Georgia State University, University Plaza: Atlanta, 2001

[9] BESENYEI, L., DOMÁN, Cs., „Üzleti prognózisok idősoros modelljei", Digitális Tankönyvtár: 2011., 3. fejezet, Letöltés ideje: 2020. október 6., 09:50.

[10] LOGIN: „Ki, mikor, mit és hol? Lehet tudatosan tervezni.", https://login.hu/hu/tartalom/shift.html, Letöltés ideje: 2020. október 20., 11:30.

[11] 2012. évi I. törvény a munka törvénykönyvéről, "50/96.§

[12] DATA HUB: „CSV – Comma Separated Values", https://datahub.io/docs/data-packages/csv, Letöltés ideje: 2020. szeptember 13., 21:00.

[13] GÉCZY-PAPP, R., „Autoregresszív és mozgóátlag folyamatok", Miskolci Egyetem Gazdaságtudományi Kar Gazdaságelméleti és Módszertani Intézet: Miskolc, https://docplayer.hu/79561095-Autoregresszivesmozgoatlag-folyamatok.html, Letöltés ideje: 2020. október 7., 11:32.

[14] STELLWAGEN, E., TASHMAN, L., „ARIMA: The Models of Box and Jenkins", 2013.01., https://www.researchgate.net/publication/285902264_ARIMA_The_Models_of_Box_and_Jenkins#fullTextFileContent, Letöltés ideje: 2020. október 7., 14:04.

[15] NORTH, G. R., „Statistical Methods, Data Analysis", ScienceDirect: 2015, https://www.sciencedirect.com/topics/earth-and-planetary-sciences/autoregressive-process, Letöltés ideje: 2020. október 8., 08:10.

[16] STAT510: „Applied Time Series Analysis", Penn State Eberly College of Science, https://online.stat.psu.edu/stat510/lesson/2/2.1, Letöltés ideje: 2020. október 8., 08:36.

[17] BROCKWELL, P. J.; DAVIS, R. A., "Time Series: Theory and Methods", (2nd ed.). New York: Springer. pp. 273. [ISBN 9781441903198], 2009.

[18] HANNAN, E. J., „Multiple time series. Wiley series in probability and mathematical statistics", New York: John Wiley and Sons, 1970.

[19] SAS INSTITUTE: „Notation for ARIMA Models", Time Series Forecasting System, https://support.sas.com/documentation/cdl/en/etsug/63939/HTML/default/viewer.htm#etsug_tffordet_sect016.htm, Letöltés ideje: 2020. október 10., 13:01.

[20] TUTORIALS POINT: „Time Series – Variations of ARIMA", https://www.tutorialspoint.com/time_series/time_series_variations_of_arima.htm, Letöltés ideje: 2020. október 10., 15:51.

[21] BME: „Idősorelemzés – előadás", 2016.10.15., https://math.bme.hu/~ftamas/szrmea/szrmea789.pdf, Letöltés ideje: 2020. október 13., 08:43.

[22] Sné KRISZT, É., VARGA, E., Vné KENYERES, E., KORPÁS, A., CSERNYÁK, L., „Általános statisztika II.", [ISBN 978-963-

19-2781-8], Nemzeti Tankönyvkiadó Rt.: Budapest, 1997., 8. fejezet

[23] KWIATKOWSKI, D., PHILLIPS, P. C. B., SCHMIDT, P., SHIN, Y., „Testing the null hypothesis of stationarity against the alternative of a unit root", Journal of Econometrics 54., pp. 159-178., North-Holland 1992.

[24] R: „Augmented Dickey-Fuller Test", http://finzi.psych.upenn.edu/R/library/tseries/html/adf.test.html, Letöltés ideje: 2020. október 16., 09:00.

[25] R: „Phillips-Perron Unit Root Test", http://finzi.psych.upenn.edu/R/library/tseries/html/pp.test.html, Letöltés ideje: 2020. október 16., 09:20.

[26] R: „Kwiatkowski-Phillips-Schmidt-Shin Test", http://finzi.psych.upenn.edu/library/aTSA/html/kpss.test.html, Letöltés ideje: 2020. október 16., 10:02.

[27] STEINBACH, M., KUMAR, V., TAN, P-N., „Bevezetés az adatbányászatba", Panem Könyvkiadó Kft., 2011., C. függelék - Valószínűségszámítás és statisztika

[28] EVIEWS: „Unit Root Testing", http://www.eviews.com/help/helpintro.html#page/content/advtimeser-Unit_Root_Testing.html, Letöltés ideje: 2020. október 17., 18:09.

[29] MATHREFERENCE: „Deriválttáblázat", https://www.mathreference.org/index/page/id/54/lg/hu, Letöltés ideje: 2020. október 18., 10:19.

[30] SHITTU, O. I., „Comparison of Criteria for Estimating the Order of Autoregressive Process: A Monte Carlo Approach", European Journal of Scientific Research [ISSN 1450-216X], Vol.30 No.3, pp.409-416, 2009.

[31] HYNDMAN, R. J., KHANDAKAR, Y., „Automatic Time Series Forecasting: The forecast Package for R", July 2008, Journal of statistical software 26, [DOI: 10.18637/jss.v027.i03]

[32] TALAGALA, T. S., HYNDMAN, R. J., ATHANASOPOULOS, G., „Meta-learning how to forecast time series", Monash University, Department of Econometrics and Business Statistics, [ISSN 1440-771X]

[33] DATE, S., „The Akaike Infromation Criterion", Towards, Data Science: 2019.11.09., https://towardsdatascience.com/the-akaike-information-criterion-c20c8fd832f2, Letöltés ideje: 2020. október 20., 21:12.

[34] R: „Auto- and Cross- Covariance and -Correlation Function Estimation", https://stat.ethz.ch/R-manual/R-devel/library/stats/html/acf.html, Letöltés ideje: 2020. október 21., 00:36.

[35] HOLMES, E. E., SCHEUERELL, M. D., WARD, E. J., „Applied Time series analysis", 2020.02.03., 4.4 fejezet, https://nwfsc-timeseries.github.io/atsa-labs/sec-tslab-correlation-within-and-among-time-series.html, Letöltés ideje: 2020. október 22., 18:40.

[36] HYNDMAN, R. J., „Better ACF and PACF plots, but no optimal linear prediction", Electronic Journal of Statistics, Vol. 0 (0000), [ISSN: 1935-7524, DOI: 10.1214/154957804100000000]