fedlap vázlat

**A SZÉKELYFÖLDI MEGYÉK MUNKANÉLKÜLISÉGI RÁTÁINAK ELŐREJELZÉSE, ARMA ÉS A MESTERSÉGES NEURONHÁLÓ-ALAPÚ MODELLEK ALAPJÁN.**

Tartalom

[1. Bevezetés 2](#_Toc152352500)

[1.1 Szakirodalmi áttekintés 2](#_Toc152352501)

[1.2 A munkanélküliségi ráta meghatározása Romániában 3](#_Toc152352502)

[1.3 Felhasznált statisztikai mutatók és fogalmak 4](#_Toc152352503)

[2 Az idősorok elemzése 5](#_Toc152352504)

[3 Előrejelzés ARMA módszerrel 6](#_Toc152352505)

[3.1 A stacionaritás vizsgálata 6](#_Toc152352506)

[3.2 Autoregresszív és mozgóátlag modellek (AR, MA, ARMA) 7](#_Toc152352507)

[3.3 Autokorrelációs és parciális autokorrelációs tesztek 9](#_Toc152352508)

[3.3.1 ACF 9](#_Toc152352509)

[3.3.2 PACF 9](#_Toc152352510)

[3.3 Előrejelzés pontosságának meghatározása a hibatesztek alapján 11](#_Toc152352511)

[4 Előrejelzés ANN módszerekkel 13](#_Toc152352512)

[4.1 MLP ……………………………………………………………………………………13](#_Toc152352513)

[4.2 CNN 13](#_Toc152352514)

[5 Következtetések 13](#_Toc152352515)

[5. Irodalomjegyzék 14](#_Toc152352516)

# Bevezetés

A munkanélküliség hosszú ideje jelentős gazdasági mutató és központi téma a közgazdaságtani kutatásokban. A munkanélküliség alakulása és változása az adott régió gazdasági egészségét tükrözi, és fontos információkat szolgáltathat a gazdasági kilátásokról. Az idősorok elemzése és az előrejelzés az egyik kulcsfontosságú eszköz a munkanélküliség változásainak megértésében és a gazdasági intézkedések megalapozásában.

IDE MAJD kellene pár cikk a romániai munkanélküliségről, hogy miért aktuális téma…

Ebben az államvizsga dolgozatban Székelyföld regionális munkanélküliségi rátáival foglalkozom. Hargita, Kovászna és Maros megye havi munkanélküliségi rátáit vizsgálom 2010 január és 2022 szeptember között. Az adatokat Románia országos statisztikai hivatalának (Institutul Național de Statistică) hivatalos oldaláról töltöttem le. Célom, hogy statisztikai elemzést készítsek ezekről az idősorokról, valamint megvizsgáljam, hogy ezen idősorok esetében, tíz hónap távlatában az ARMA (autoregresszív mozgóátlag), vagy a mesterséges neurális hálózatokon (ANN) alapuló modellek: CNN (Convolutional Neural Network), MLP (Multilayer Perceptron) nyújtanak pontosabb előrejelzéseket. A „teszt periódusnak” a 2022 október és 2023 július közötti adatokat használtam, ezeket jósoltattam meg a különböző modellekkel és hasonlítottam össze.

A kutatás során egy Python programot készítettem az adatok beolvasásához, feldolgozásához és az eredmények megjelenítéséhez, ehhez egy Django keretrendszerben írt webalkalmazás segítségével biztosítok felhasználói felületet, ezzel szemléletesebbé és egyszerűbbé téve a különböző statisztikai számításokat. A webalkalmazás lényegében bármennyi és bármilyen idősort képes elemezni a megfelelően előkészített adatforrásokból, tehát a jövőben még fel lehet használni más tematikájú elemzésekhez is.

# Szakirodalmi áttekintés

Madaras (2018) Hargita és Brassó megye esetében végzett hasonló regionális kutatást, és megállapította, hogy rövidtávon a mesterséges neuronháló-alapú NAR (nemlineáris autoregresszív) modell, középtávon az ARMA modell nyújtott pontosabb becséleseket.

Davidescu et al. (2021) Románia országos munkanélküliségi rátáival végeztek kutatást, SARIMA, SETAR, Holt-Winters, ETS, és NNAR modelleket hasonlítottak össze, amelyek közül több szempont alapján is az NNAR (neuronhálós autoregresszív) modell mutatkozott a legjobbnak.   
Madaras (2014) ARIMA (1, 1, 4) típusú autoregressziós modellel előrejelzést készített és megjósolta a munkanélküliek számának növekedését 2013 július és 2014 február között Romániában.

Mulaudzi & Ajoodha (2020) Dél-Afrika országos munkanélküliségi rátáinak előrejelzéséhez használtak többféle modellt (ARIMA, ENet, LASSO, LSTM, RNN, Holt-Winters, stb.), amelyekből az MLP (multi-layer perceptron) modell három rejtett réteggel pontosabb előrejelzést adott, mint a többi.

Tufaner & Sözen (2021) Törökország esetében hasonlítottak össze egy MLP két rejtett rétegű, és ARIMA (3, 1, 2) modellt, itt szintén az MLP volt a jobb.

## A munkanélküliségi ráta meghatározása Romániában

A dolgozat során felhasznált adatok a Romániai Statisztikai Hivataltól (INSTITUTUL NATIONAL DE STATISTICA) származnak. Az ő módszertanuk a következőképpen definiálja a munkanélküliséget és a munkanélküli rátát:

A **munkanélküliek** a nemzetközi meghatározás szerint (BIM[[1]](#footnote-1)) szerint azok a 15-74 év közötti személyek, akik egyidejűleg teljesítik a következő három feltételt:

* munkanélküliek (a mérés pillanatában nincs bejelentett munkahelyük)
* a következő két héten belül munkába tudnának állni
* az elmúlt négy hétben aktívan munkát kerestek.

A **munkanélküliségi ráta** tulajdonképpen egy százalékos arányszám: a munkanélküliek arányát számolja ki a munkaerőhöz viszonyítva egy adott térségre. Tehát beszélhetünk országos, regionális vagy megyei munkanélküliségi rátáról.

A gazdaságilag aktív népesség a bázisidőszakban az áruk és szolgáltatások előállítására rendelkezésre álló munkaerőt biztosító valamennyi személyt magában foglalja, beleértve a foglalkoztatottakat és a munkanélkülieket is. (INSSE, 2016)

Tehát a munkanélküliségi ráta megmutatja, hogy a munkaképes lakosság hány százaléka nem rendelkezik a mérés pillanatában munkahellyel, viszont tudna és akarna dolgozni.

A következő fejezetekben ismertetem a statisztikai mutatókat, modelleket, amelyeket felhasználtam a dolgozat során.

## Felhasznált statisztikai mutatók és fogalmak

**Átlag**: Az egyik legismertebb statisztikai mutató. Jelöljünk n megfigyelést x1, x2, ..., xn -nel. Ekkor az átlaguk:

**Szórás**: A szórás egy szóródási mutató, ami azt méri, hogy a megfigyelések mennyire esnek távol az átlagtól. Két minta lehet lényegesen különböző, még ha az átlaguk egyforma is, ugyanis az egyik minta megfigyelései eshetnek jóval közelebb az átlaghoz, mint a másik minta megfigyelései. A szóródás méréséhez az eltérések négyzetét veszi figyelembe, mert ezáltal fejezhető ki az összes megfigyelés távolsága az átlagtól. A szórás kiszámításához négyzetgyököt vonunk az eltérések négyzetének átlagából:

**A variancia (szórásnégyzet):**

Egy másik fontos szóródási mutató a variancia, ami egyszerűen a szórás négyzete, vagyis az átlagtól valóeltérések négyzeteinek az átlaga:

**Medián:** Jelöljük x1, x2, ..., xn –nel a megfigyeléseket, és jelöljük x(1), x(2), ... , x(n) –nel ugyanezeket a megfigyeléseket növekvő sorrendben. Tehát, x(1) a legkisebb, x(2) a következő, ... és x(n) legnagyobb:

x(1) ≤ x(2)  ≤ ⋯ ≤ x(n)

A medián a sorrendbe állított x1, x2, ..., xn megfigyelések középső megfigyelése. Ha n páratlan, akkor egészen egyszerű; a medián a (n + 1) / 2 sorrendű megfigyelés.

Ha n páros, akkor 2 középső megfigyelés van, n/2 és (n / 2) + 1, tehát a medián:

(Sándor & Tánczos, 2019)

# Az idősorok elemzése

A webalkalmazásomba feltöltöttem a 2010 január – 2022 szeptember közötti munkanélküliségi ráta adatsorokat a három megyére. A feldolgozás után a program a következő eredményeket jelenítette meg:



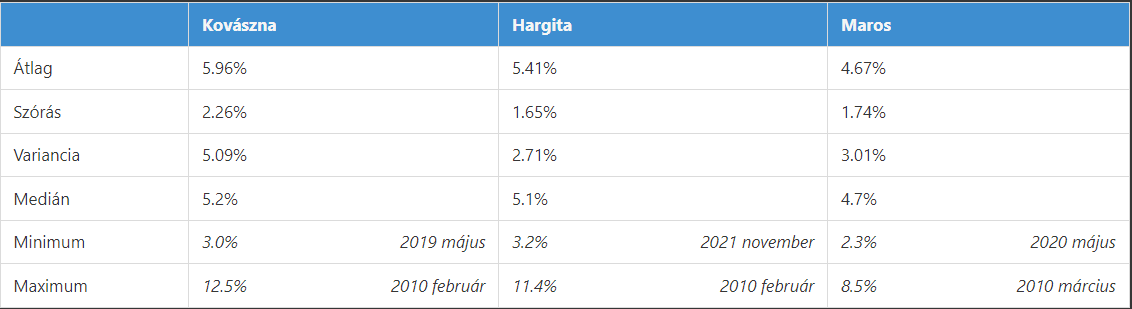
*A regionális munkanélküliségi ráták grafikonja*

Látszik, hogy Maros megyében szinte végig a legalacsonyabb a munkanélküliségi ráta, míg Kovásza megyében a legmagasabb. Sok periódusban megfigyelhető, hogy télen magasabb a mutató, mint nyáron, ilyenkor többen kérnek munkanélküliségi segélyt. Szerencsére a 2008-as válság óta folyamatosan csökken az arány, viszont a Covid járvány idején sajnos egy nehezebb periódus jelei láthatóak.

Kovászna és Hargita megyében a vizsgált időszakban 2010 februárjában volt a legmagasabb a munkanélküliségi ráta (12.5% valamint 11.4%), míg Maros megyében 2010 márciusában, 8.5%-os értékkel. Ez bizonyára a 2008-ban kirobbant gazdasági világválság hatása, amely elérte Romániát is, viszont az akkori vezetőség ezt kezdetben nem látta be. Románia erősen érintett volt a gazdasági válság által. Az ország nagymértékben ki volt téve az ingatlanspekulációnak, és külföldi banki tőkének. Az ország gazdasága jellemzően az alacsony és közepes képzettségű munkaerőt használó, viszonylag kevés technológiát felhasználó és kevés hozzáadott értékű iparágakon alapult. A gazdasági recesszió miatt rengeteg munkahely szűnt meg, vagy jelentősen csökkentette dolgozóinak létszámát. (Georgeta, 2015)

Azonban 2020 tavaszáig összeségében nézve folyamatosan csökkent a munkanélküliek száma, ez jól leolvasható a grafikonról, a gazdaság folyamatosan fejlődött, ehhez a technológiai fejlődés is hozzájárult. Maros megyében 2020 májusában volt a legalacsonyabb a mutató, 2.3%, Hargita megyében 2021 novemberében 3.2%, míg Kovászna megyében 2019 májusában 3% volt. Sajnos a koronavírus járvány miatt

A következő táblázat szemlélteti a három idősor átlagát, szórását, szórásnégyzetét (variancia), mediánját, minimum és maximum érétkeit.



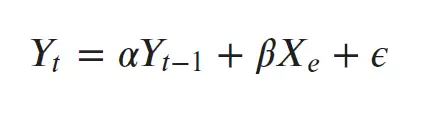
*Az idősorok statisztikai mutatói*

A következő lépésekben megvizsgálom ADF és KPSS teszt segítségével az idősorok stacionaritását, ismertetem és a program segítségével elvégeztem az autokorrelációs és parciális autokorrelációs teszteket, ezek segítségével megállapítom azokat az AR/MA/ARMA modelleket, amelyek a legjobban illeszkednek a megyék idősoraira, ezekkel pedig előrejelzéseket készítek.

# Előrejelzés ARMA módszerrel

## A stacionaritás vizsgálata

A stacionaritás az idősorok statisztikai tulajdonságainak időbeni állandóságát vagy közelítő állandóságát jelenti. Egy stacionárius idősor esetén az várható értéke, varianciája és autokorrelációs függvénye állandó, vagy csak időben állandó konstans eltolódásokkal változik. Stacionárius idősorok könnyebben modellezhetők és előrejelzhetők. A stacionaritás meglétét többféleképpen meg lehet állapítani, én a következő teszteket használtam:

* Augmented Dickey-Fuller (ADF) teszt: Ez a teszt azt vizsgálja, hogy az autoregresszív modellben lévő egységgyökér (unit root) mellett az idősor rendelkezik-e egyéb autoregresszív szerkezettel. A nullhipotézis az, hogy az idősor rendelkezik egy egységgyökérrel, tehát nem stacionárius.
* Kwiatkowski-Phillips-Schmidt-Shin (KPSS) teszt: Ezzel a teszttel azt lehet vizsgálni, hogy az idősor szigorúan stacionárius-e. A nullhipotézis az, hogy az idősor szigorúan stacionárius, tehát nincs egységgyökér.

A programom segítségével elvégeztem a beolvasott adatsorokra ezeket a teszteket, és a következő eredményeket kaptam:

. táblázat

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | ADF Statisztika | ADF p-érték | KPSS Statisztika | KPSS p-érték |
| Kovászna | -2.79 | 0.06 | 1.53 | 0.01 |
| Hargita | -2.64 | 0.08 | 1.48 | 0.01 |
| Maros | -2.37 | 0.15 | 1.65 | 0.01 |

ADF Statisztika: Ez a teszt az autoregresszív komponens jelenlétét vizsgálja az idősorban. Ha az ADF statisztika abszolútértéke elég nagy és a p-érték kisebb a kritikus értéknél (általában 0.05), akkor elutasíthatjuk a nullhipotézist, ami azt jelenti, hogy az idősor stacionárius.

ADF p-érték: Ez az érték a nullhipotézis elutasításához szükséges szignifikanciaszintet méri. Általában, ha a p-érték kisebb, mint 0.05, akkor elutasítjuk a nullhipotézist.

KPSS Statisztika: A KPSS teszt arra irányul, hogy vizsgálja, vannak-e olyan periódusos elemek az idősorban, amelyek kiszélesítik a várt értéket. Ha a statisztika nagyobb a kritikus értéknél, akkor az idősor nem stacionárius.

KPSS p-érték: Ez is a nullhipotézis elutasításának szignifikanciaszintjét méri. Általában, ha a p-érték kisebb, mint 0.05, akkor elutasítjuk a nullhipotézist.

A táblázatban látható eredmények alapján minden megye esetén az ADF és KPSS p-értékeket összehasonlítva, mindegyik esetben elutasíthatjuk a nullhipotézist, vagyis **az idősorok stacionáriusnak tekinthetők**.

## Autoregresszív és mozgóátlag modellek (AR, MA, ARMA)

**Az autoregresszió (AR)** azt jelenti, hogy az aktuális időpontbeli értéket a korábbi időpontbeli értékek határozzák meg. Az AR komponens arra utal, hogy az aktuális érték korrelál az előző időpontbeli értékekkel. A "p" paraméter megadja az autoregressziós rendszámot, vagyis hány darab előző időpontbeli értéket használunk az aktuális érték becsléséhez.

Egy p-rendű autoregresszív modellt, amelynek jelzése AR(p), a következőképpen értelmezünk:

Az ismeretlen paraméterek (autoregresszív együtthatók) és az a hibaváltozó, amit **fehérzajnak** feltételezünk, vagyis olyan folyamat, amelynek várható értéke 0, varianciája konstans és autokorrelációja 0, valamint a hibaváltozó kovarianciája az minden késleltetett értékével 0.   
Az AR(p) tulajdonképpen egy többváltozós **lineáris** modell, ahol a regresszorok (független változók) a függőváltozó késleltetett értékei.

Egy AR(p) folyamat akkor stacionárius, ha:

* Az autoregresszív paraméterek abszolútértékei kisebbek, mint 1
* Ezen paraméterek összegének abszolútértékei kisebb, mint 1

A **mozgóátlag** **(MA)** azt jelenti, hogy az aktuális időpontbeli értéket a korábbi időpontbeli hibák lineáris kombinációjaként becsüljük meg. Az MA arra utal, hogy az aktuális érték korrelál az előző időpontbeli hibákkal, és az "q" paraméter megadja a mozgóátlag rendszámát, azaz hány korábbi hibaértéket használunk az aktuális érték becsléséhez. Egy q-rendű mozgóátlag folyamat (MA(q)) jelölése:

Ezek a modellek mindig **stacionáriusok**, vagyis a folyamat autokorrelációi nem változnak az idő függvényében, tehát az idősorban nincsenek trendek is szezonális mintázatok

A két folyamat kombinációja az **ARMA (p, q)** (autoregresszív mozgóátlag) folyamat, amely komplexebb idősorokat is képes leírni csupán két paraméterrel.

ahol p az autoregressziós folyamat rendje, q a mozgóátlag folyamat rendje, az idősorozat aktuális értéke, a konstans érték, az autoregresszív együtthatók, az adott időpontbeli fehérzaj, a mozgóátlag együtthatók.

Egy ARMA folyamat akkor stacionárius, ha az AR komponense stacionárius.

(Sándor, 2019)

## Autokorrelációs és parciális autokorrelációs tesztek

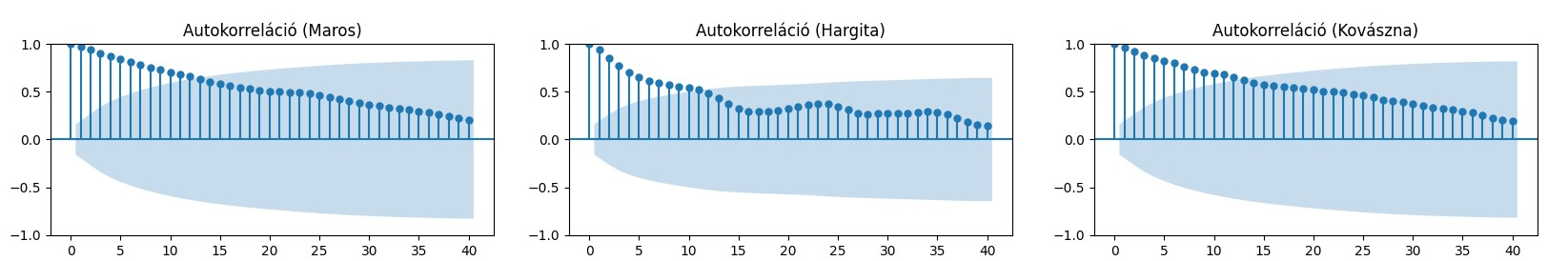
Ebben a részben megvizsgálom és elemzem az autokorrelációs (a továbbiakban ACF) és parciális autokorrelációs (a továbbiakban PACF) teszteket annak érdekében, hogy   
meg tudjam határozni a lehető legjobban illeszkedő idősor modelleket és   
azok paramétereit (AR(p), MA(q), ARMA(p, q)) az egyes idősorokra.

## ACF

Egy idősor autokorrelációs függvénye (ACF) az autokorrelációk sorozata:  
 ahol a k-ad rendű autokovariancia.  
Az ACF segít azonosítani a mozóátlag (MA) folyamatot azáltal, hogy megmutatja, hány időegységnyi korreláció van az aktuális és az előző időpontok között, miközben figyelmen kívül hagyja a köztes időpontokat. Ha egy stacionárius folyamat ACF-je teljesíti azt a feltételt, hogy

(Sándor, 2019)

Az általam vizsgált idősorok autokorrelációs tesztjei a következőképpen néznek ki:



Elvégzett ACF tesztek a három megyére

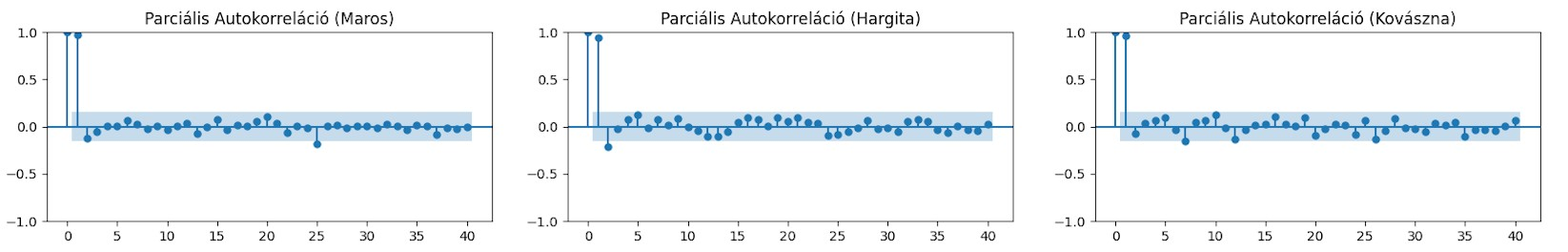
Ha AR folyamat lenne, akkor az első néhány lépés után az autokorrelációk értékei hirtelen zuhannának, viszont a fent látható grafikonok nem ezt mutatják, hanem lineáris, fokozatos csökkenést, emiatt kizárható az, hogy MA(q) folyamatokról lenne szó bármelyik megye esetében is.

## PACF

Egy AR(1) folyamat esetében a pk autokorrelációkfolyamatosan (exponenciálisan) csökkennek. Általában minden AR(p) folyamatnak megvan ez a tulajdonsága, viszont p > 1 rendű folyamatok esetében a csökkenés nem feltétlenül monoton. Ha egy stacionárius folyamat PACF értékei csak a p-edik időbeli lépésben nem nullák (tehát minden további lépésben megközelítőleg nullák), akkor AR(p) folyamatról van szó.

(Sándor, 2019)

Az általam vizsgált idősorok parciális autokorrelációs tesztjei a következőképpen néznek ki:



Elvégzett PACF tesztek a három megyére

Itt mindhárom megye esetében az látszik a PACF tesztek eredményein, hogy az első kettő lépésben az autokorrelációs érték 1, míg az összes többiben elhanyagolhatóak az autokorrelációs kapcsolatok, tehát valószínűleg AR(2) folyamatról beszélünk.

A grafikonok azt szemléltetik, hogy valószínűleg mindhárom megye esetében az AR(2) modellel érdemes próbálkozni az előrejelzéshez, viszont megnéztem még az ARMA(1, 1) és ARMA(1, 2) eseteket is, majd ezekből kiválasztottam a legkisebb AIC[[2]](#footnote-2) (Akaike Information Criterion) értékű modellt a Sándor (2019) alapján, mert valószínűleg ez a modell fog a legjobban illeszkedni az adott idősorra. A következő táblázatban összefoglalom, hogy a különböző modellekre milyen AIC értékeket kaptunk az egyes megyék esetében (a minimum AIC-k celláit beszíneztem):

|  |  |  |  |
| --- | --- | --- | --- |
| Modell | AIC | | |
| Maros | Hargita | Kovászna |
| AR (1) | 9.65 | 128.96 | 123.91 |
| AR (2) | 1.59 | 114.34 | 124.09 |
| ARMA (1,1) | 2.15 | 116.51 | 123.70 |
| ARMA (1,2) | 3.81 | 113.73 | 124.86 |

*A lefuttatott modell tesztek eredményei*

Azt látjuk, hogy az előrejelzéseknél Maros megyénél AR(2), viszont Hargitánál ARMA(1, 2), míg Kovászna megyénél ARMA(1, 1) modellt érdemes használni, nem az AR(2)-t.

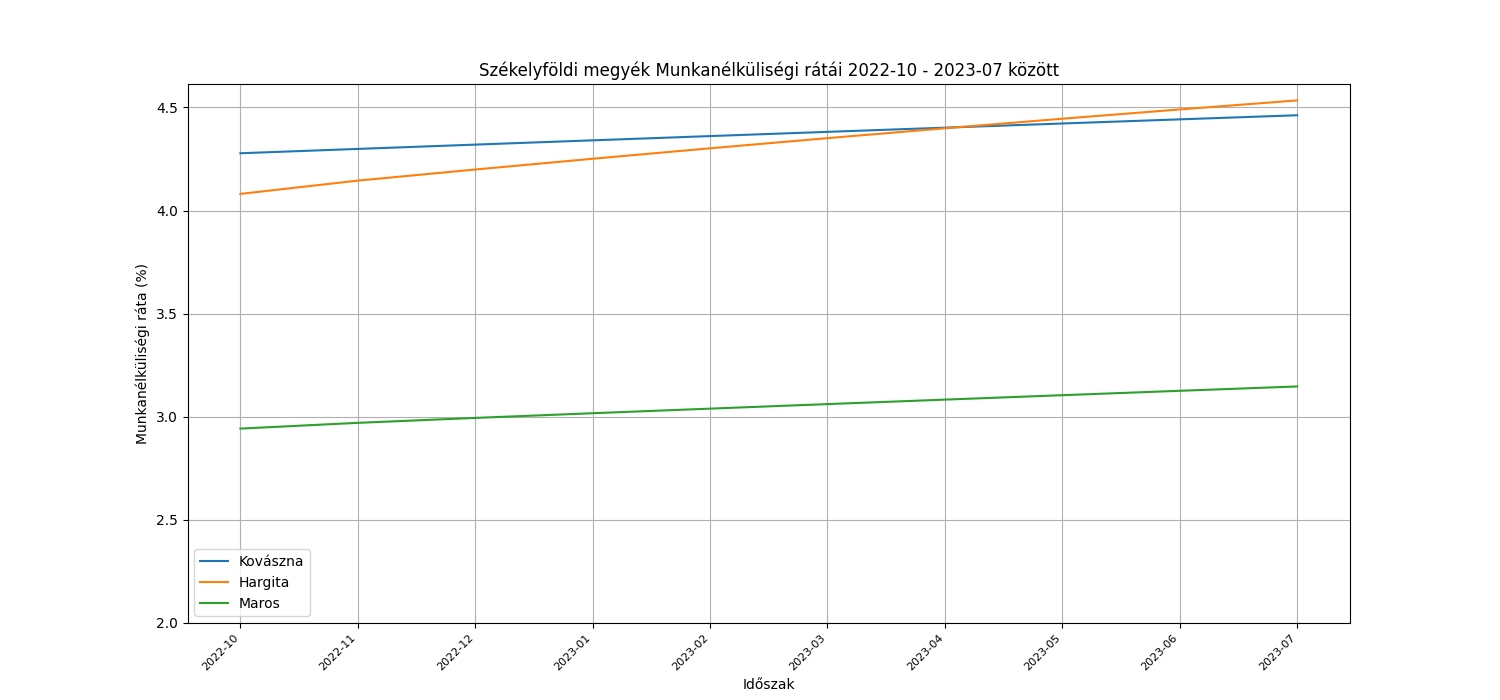
## 3.3 Előrejelzés pontosságának meghatározása a hibatesztek alapján

Az adatforrásból szándékosa meghagytam a legfrissebb 10 bejegyzést, hogy az egyes modellek becsléseinek hibáit meg tudjam állapítani a valódi adatokhoz képest.

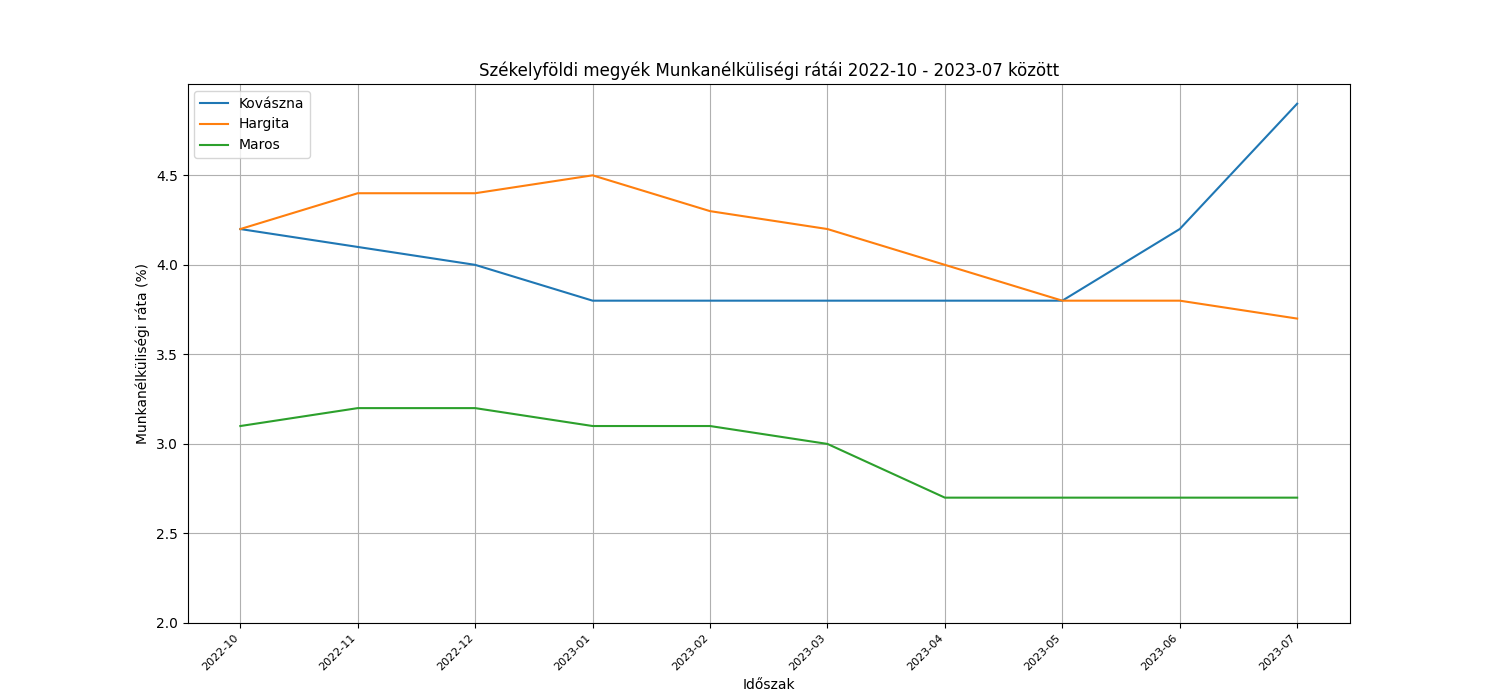
|  |  |  |  |
| --- | --- | --- | --- |
|  | Kovászna | Hargita | Maros |
| 2022 október | 4,2 | 4,2 | 3,1 |
| 2022 november | 4,1 | 4,4 | 3,2 |
| 2022 december | 4 | 4,4 | 3,2 |
| 2023 január | 3,8 | 4,5 | 3,1 |
| 2023 február | 3,8 | 4,3 | 3,1 |
| 2023 március | 3,8 | 4,2 | 3 |
| 2023 április | 3,8 | 4 | 2,7 |
| 2023 május | 3,8 | 3,8 | 2,7 |
| 2023 június | 4,2 | 3,8 | 2,7 |
| 2023 július | 4,9 | 3,7 | 2,7 |

*12. ábra: a 2022 szeptember után mért értékek*

Miután a programon keresztül kértem a kiválasztott modellekre a 10 hónapos előrejelzéseket, a következő adatokat kaptam:



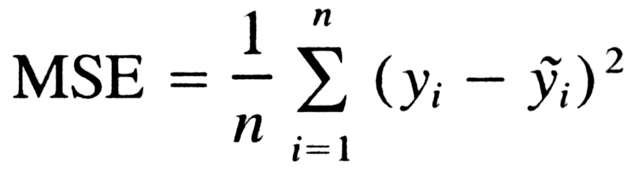
*13. ábra: ARMA előrejelzések grafikonon ábrázolva*



*14. ábra: A valós adatok grafikonon ábrázolva*

A becslések pontosságának megállapításához két mutatót használtam:

* az **átlagos négyzetes eltérést** (Mean Squared Error, MSE), amely a tényleges és becsült adatok közötti különbségek négyzeteinek az átlaga:



* **a relatív átlagos négyzetes** eltérések gyökét (Relative Root Mean Square Error, RRMSE), amely relatívan, normalizálva adja meg a hibákat, százalékos értékben. Minél kisebb ez két mutató, a modell annál jobb becslést ad.

A képletekben n az adatok száma, yi a valóságos, a becsült érték az i-edik mintában. a valóságos értékek átlaga. A következő táblázatban összefoglaltam az MSE és RRMSE értékeket az egyes modellekre:

|  |  |  |
| --- | --- | --- |
| Modell | MSE | RRMSE |
| Kovászna ARMA(1, 1) | 0.209 | 0.113 |
| Hargita ARMA(1, 2) | 0.1953 | 0.107 |
| Maros AR(2) | 0.0826 | 0.0974 |

*MSE és RRMSE értékek*

# Előrejelzés ANN módszerekkel

## MLP

## 4.2 CNN

# Következtetések

# 5. Irodalomjegyzék

Batuhan TUFANER, M., & SÖZEN, İ. (2021). Forecasting Unemployment Rate in the Aftermath of the Covid-19 Pandemic: The Turkish Case. *İzmir Journal of Economics*, Volume: 36 Issue: 3, 685 - 693, 02.09.2021.

Davidescu, A. A., Apostu, S.-A., & Paul, A. (2021. március 9). Comparative Analysis of Different Univariate Forecasting Methods in Modelling and Predicting the Romanian Unemployment Rate for the Period 2021–2022. *Entropy, 23 (3)*, 325.

Georgeta, E. S. (2015). The economic and social situation in Romania. Európai Unió: European Economic and Social Committee.

Madaras. (2014). A gazdasági válság hatása a munkanélküliség alakulására országos és megyei szinten Romániában. *Közgazdász Fórum, 17 (1-2)*, 136-149.

Madaras, S. (2018). Forecasting the regional unemployment rate based on the Box-Jenkins methodology vs. the Artifi cial Neural Network approach. Case study of Brașov and Harghita counties. *Közgazdász Fórum (Forum on Economics and Business)*, 66.

Mulaudzi, R., & Ajoodha, R. (2020). An Exploration of Machine Learning Models to Forecast the Unemployment Rate of South Africa: A Univariate Approach. IEEE.

Sándor, Z. (2019). *Bevezetés az ökonometriába.* Sepsiszentgyörgy: T3 Kiadó.

Sándor, Z., & Tánczos, L. J. (2019). *Gazdasági statisztika jegyzet.* Sepsiszentgyörgy: T3 Kiadó.

1. Biroul Internaţional al Muncii [↑](#footnote-ref-1)
2. *Az AIC (Akaike Information Criterion) célja az, hogy a modellek közötti összehasonlításra szolgáljon, figyelembe véve a modell illeszkedését és a paraméterszámot. Az AIC egy olyan kompromisszumot keres, amely az illeszkedés minőségét és a modell bonyolultságát egyensúlyozza.* [↑](#footnote-ref-2)