fedlap vázlat

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

Tartalom

[1. Bevezetés 2](#_Toc151916108)

[1.1 A munkanélküliségi ráta meghatározása Romániában 3](#_Toc151916109)

[1.2 Felhasznált statisztikai mutatók és fogalmak 3](#_Toc151916110)

[2 Az idősorok elemzése 4](#_Toc151916111)

[3 Előrejelzés ARMA módszerrel 6](#_Toc151916112)

[3.1 Autoregresszív és mozgóátlag modellek (AR, MA, ARMA) 6](#_Toc151916113)

[3.2 Autokorrelációs és parciális autokorrelációs tesztek 7](#_Toc151916114)

[3.2.1 ACF 7](#_Toc151916115)

[3.2.2 PACF 8](#_Toc151916116)

[3.3 Előrejelzés pontosságának meghatározása a hibatesztek alapján 9](#_Toc151916117)

[4.1 AR/ARMA becslések 10](#_Toc151916118)

[4.2 MLP, CNN 11](#_Toc151916119)

[4 Következtetések 11](#_Toc151916120)

[5 Bibliográfia 11](#_Toc151916121)

# 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 regionális munkanélküliségi rátákkal foglalkozom. Hargita, Kovászna és Maros megye munkanélküliségi rátáit vizsgálom 2010 január és 2022 július között. Célom, hogy statisztikai elemzést és előrejelzéseket készítsek ezekről az idősorokról, a jövőbeli munkanélküliségi rátákra vonatkozóan, emellett egy olyan Django webalkalmazást igyekeztem készíteni, amely a feltöltött idősorokat tartalmazó erőforrás alapján az összes szükséges elemzést elvégzi és azokat megjeleníti, így sokkal könnyebbé és gyorsabbá téve az előrejelzések készítését.

A dolgozatban két fő módszert alkalmazok az előrejelzések elkészítéséhez: az ARMA (AutoRegressive Moving Average - Autoregresszív Mozgóátlag), valamint a mesterséges neurnhálón alapuló MLP (Multilayer Perceptron) és CNN (Convolutional Neural Network) modelleket Az ARMA modell statisztikai módszer, amely az idősorokat elemzi és előrejelzéseket készít azok alapján. Az MLP és a CNN pedig mesterséges neurális hálózatok, amelyek a gépi tanulás (ML – machine learning) területén széles körben használt módszerek, többek között idősorok előrejelzésére is. A dolgozatban kiderül az is, hogy az általam vizsgált esetben melyik módszerrel lehet megbízhatóbb előrejelzést készíteni.

Ide cikkek

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. A programban Felhasználtam a Python pandas, numpy, statsmodels, pyplot moduljait.

Az kutatás során a Román Statisztikai Hivatal (INSS - National Institute of Statistics and Studies) adatait használom fel, amelyeket összehasonlítok majd a modellek által jósolt értékekkel, hogy megállapítsam, melyik módszer nyújt pontosabb becsléseket.

## 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.

## 1.2 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:



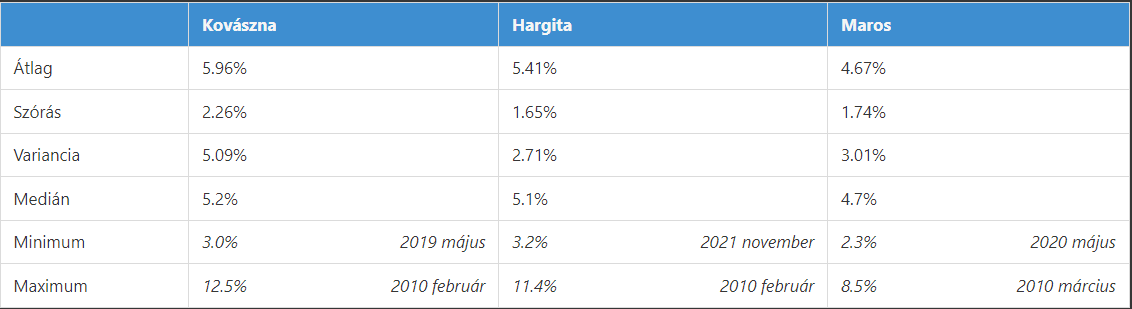
*A regionális munkanélküliségi ráták grafikonja*

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. (Steluța Enach 2014)

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 lépésben a program segítségével elvégeztem az autokorrelációs és parciális autokorrelációs teszteket.

# Előrejelzés ARMA módszerrel

Az előrejelzéshez meg kell találni azt a modellt a megfelelő paraméterrel, amelyik a legjobban illeszkedik az idősorra. A következő részben ismertetem azokat a modelleket, amelyeket számításba vettem, majd ezután azt a módszert, amely segítségével kiválasztottam ezeket a három megyére vonatkozóan.

## 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 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.



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

(Sándor, 2019)

## Autokorrelációs és parciális autokorrelációs tesztek

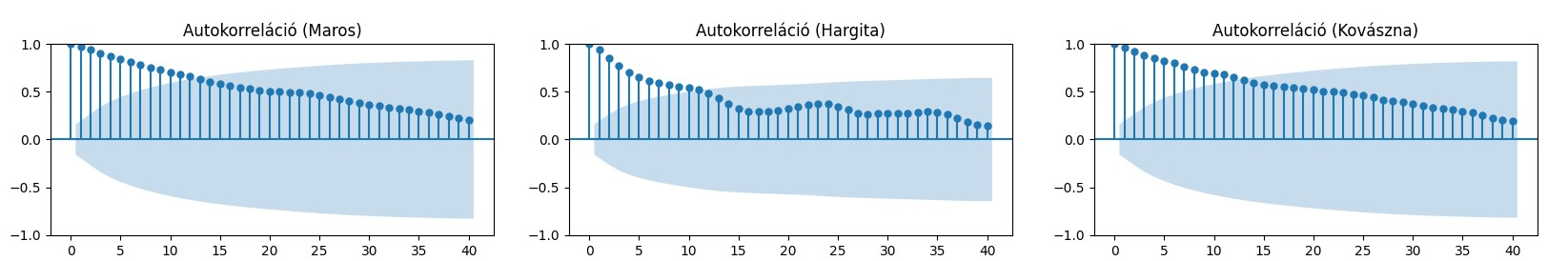
Ebben a részben megvizsgálom és elemzem az autokorrelációs (a továbbiakban ACF) és parciális autokorrelációs (a továbbiakban PACF) teszteket annak érdekében, hogy   
meg tudjam határozni a lehető legjobban illeszkedő idősor modelleket és   
azok paramétereit (AR(p), MA(q), ARMA(p,q)) az egyes idősorokra.

## 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:



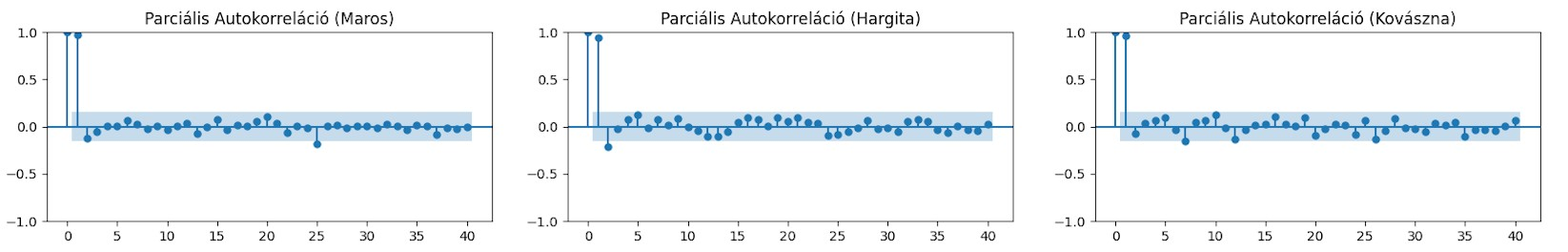
Elvégzett ACF tesztek a három megyére

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, fokozatok 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 stacionárius folyamat PACF értékei csak a p-edik időbeli lépésben nem nullák, tehát minden további lépésben megközelítőleg nullák, (tehát a) akkor AR(p) folyamatról van szó.

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 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), Hargitánál ARMA(1, 2), míg Kovászna megyénél ARMA(1, 1) modellt érdemes használni, tehát nem feltétlen adta a legjobb eredményt A következő pontban az előrejelzéseket vizsgálom ezeket a modelleket felhasználva.

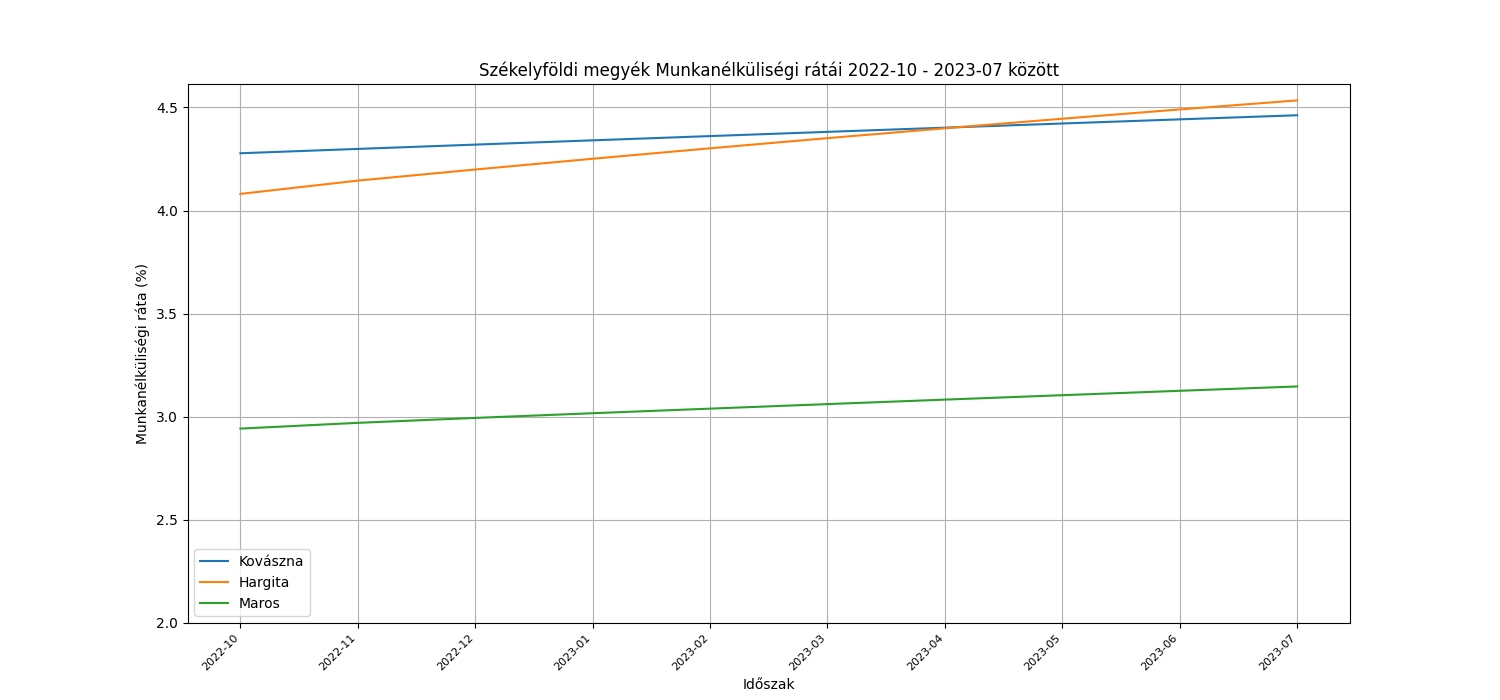
## 3.3 Előrejelzés pontosságának meghatározása a hibatesztek alapján

Az adatforrásunkból szándékosa meghagytam a legfrissebb 10 bejegyzést, hogy az egyes modellek becsléseinek hibáit meg tudjuk állapítani.

|  |  |  |  |
| --- | --- | --- | --- |
|  | 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 |

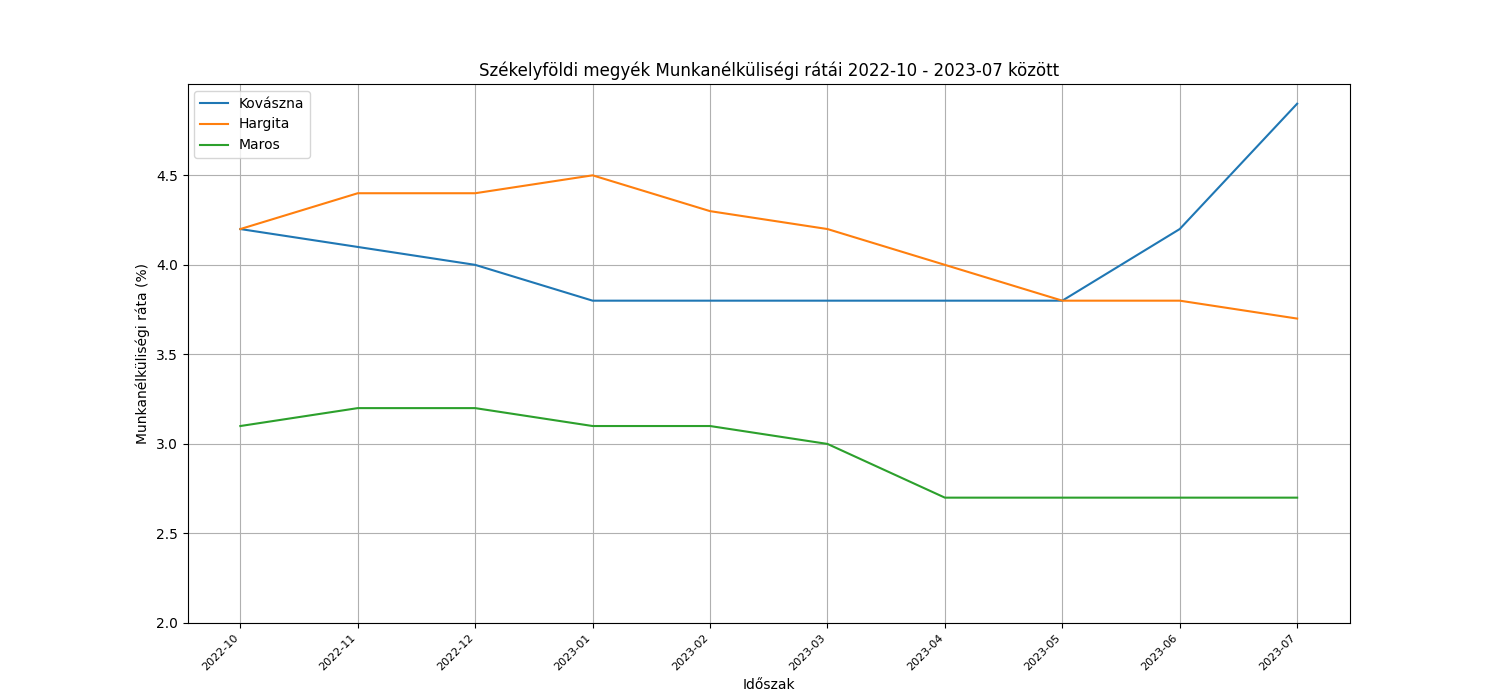
*12. ábra: a 2022 szeptember után mért értékek*

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



*13. ábra: ARMA előrejelzések grafikonon ábrázolva*

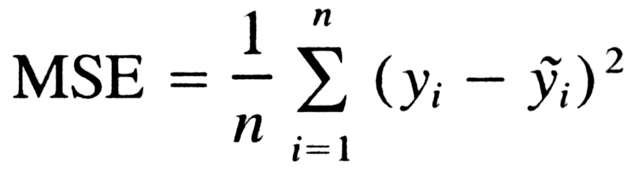
Viszonyításként az eredeti adatok így néznek ki grafikonon ábrázolva:



*14. ábra: A mért 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*

## 4.2 MLP, CNN

# Következtetések

# Bibliográfia

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)