**A székelyföldi megyék munkanélküliségi rátáinak összehasonlítása, statisztikai elemzése és előrejelzése. Az ARIMA és a mesterséges neuronháló-alapú modellek összehasonlítása.**

# 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.

Ebben az államvizsga dolgozatban a Hargita, Kovászna és Maros megye munkanélküliségi rátáit vizsgáljuk 2010 január és 2022 július között. Célunk, hogy statisztikai elemzést és előrejelzéseket készítsek ezekről az idősorokról, a jövőbeli munkanélküliségi rátákra vonatkozóan.

A dolgozatban két fő módszert alkalmazunk az előrejelzések elkészítéséhez: az ARMA (AutoRegressive Moving Average), valamint a mesterséges neurnhálón alapuló MLP (Multilayer Perceptron) és CNN (Convolutional Neural Network) modelleket.  
 Az ARIMA 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) területén széles körben használt módszerek, többek között az idősor előrejelzésére is.

A kutatás során egy Python programot készítettem az adatok beolvasásához, feldolgozásához és az eredmények megjelenítéséhez, ehhez egy Django keretrendszerben írt webalaklalmazás segítségével biztosítok felhasználói felületet.

Az kutatás során az INSS (National Institute of Statistics and Studies) adatait használom fel, amelyeket összehasonlítok majd a modellek által jósolt értékekkel, hogy megállapítsuk, melyik módszer nyújt jobb becslést.

## 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:

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

A **munkanélküliségi ráta** a munkanélküliek aránya a munkaerőben.

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.

## 1.2 Alapvető 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 ,201?)*

## Korreláció, autokorreláció, parciális autokorreláció.

**A korrelációs együttható** a kovarianciához hasonlóan megadja, hogy két változó között van-e lineáris összefüggés, és megadja annak mértékét:

*(Sándor – Tánczos ,201?)*

Tekintsük az alábbi többváltozós lineáris modellt:    
Azt a jelenséget, amikor a hibaváltozók korreláltak, vagyis létezik legalább két i ≠ j megfigyelés, amelyekre i és j kovarianciája nem nulla, vagyis , **autokorrelációnak nevezzük**.

Egy idősorban az egymás utáni értékek közötti korreláció hasznos információ lehet az idősor jövőbeli értékeinek előrejelzéséhez. Ha az értékek között erős korreláció van, az azt jelzi, hogy az egyik időpontban mért értékek segíthetnek megjósolni a következő időpontban mért értékeket. Az autokorreláció (ACF) és a parciális autokorreláció (PACF) segít meghatározni az idősor modelljének a típusát (AR, MA, ARMA). Egy **idősor autokorrelációs függvénye (ACF)** az autokorrelációk sorozata:  
 ahol  a k-ad rendű autokovariancia.   
Az ACF segítségével meglehet határozni egy mozgóátlag (MA) folyamatot. Ha egy stacionárius folyamat teljesíti azt, hogy pk = 0, k > q és pq ≠ 0, akkor az egy MA(q) folyamat. Ez utóbbi tulajdonság segít abban, hogy megkülönböztessük más (AR, ARMA) folyamatoktól és meghatározzuk annak rendjét.

A PACF

## 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.

A **mozgóátlag (MA)** azt jelenti, hogy az aktuális időpontbeli értéket a korábbi időpontbeli hibák lineáris kombinációjaként becsüljük meg. Az MA komponens arra utal, hogy az aktuális érték korrelál az előző időpontbeli hibákkal, és a *q* paraméter megadja a mozgóátlag rendszámát, vagyis, hogy hány korábbi hibaértéket használunk az aktuális érték becsléséhez.

Ezek a modellek mindig stacionáriusok, vagyis a folyamat autokorrelációi nem változnak az idő függvényében. Lényegében az AR(p) egy többváltozós lineáris modell, ahol a regresszorok (független változók) a függőváltozó késleltetett értékei. Egy AR(p) modellel meghatározott yt idősort AR(p) folyamatnak nevezünk. Egy ilyen folyamat paraméterei alapján eldönthető, hogy a folyamat stacionárius-e.

# A program működése

## Beolvasás

Az adatbeolvasásra egy függvényt írtam, amely három paramétert vár: a beolvasandó Excel file útvonalát (file), és a munkalap nevét, végül azt, hogy a az ábrázolás során milyen beosztás szerint legyen számozva az X tengely. A bemenő .xlsx fáj a következő módon kell legyen előkészítve, hogy a program jól tudja beolvasni:

|  |  |  |  |
| --- | --- | --- | --- |
| időszak | Kovászna | Hargita | Maros |
| 2010 január | 12.2 | 10.9 | 8.3 |
| 2010 február | 12.5 | 11.4 | 8.4 |
| 2010 március | 12.3 | 11.2 | 8.5 |
| 2010 április | 11.9 | 10.7 | 8.2 |
| 2010 május | 11.1 | 9.6 | 7.9 |
| 2010 június | 10.9 | 9.3 | 7.8 |

. táblázat: A bemeneti táblázat mintája (saját ábra)

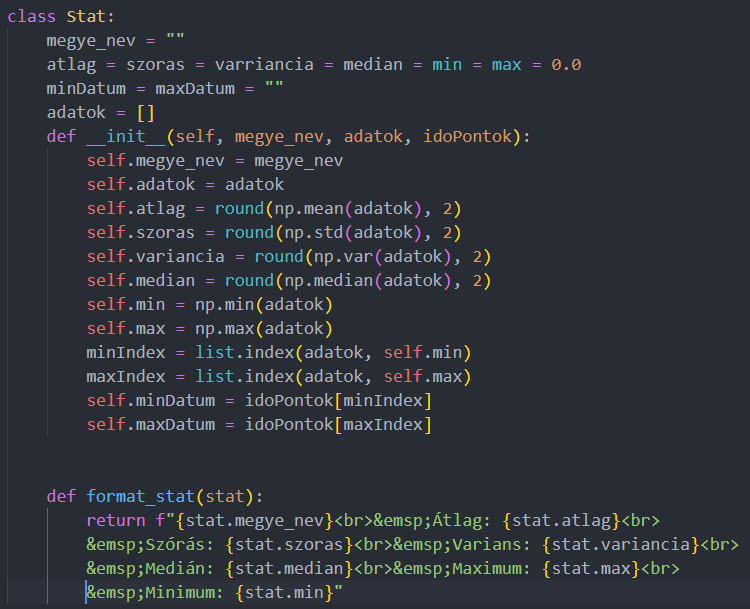
Tehát az első sor a fejléc, az első oszlop a megfigyelések időpontjai (ezek fogják a grafikonon az x tengely értékeit adni), míg a további három oszlop a három megye adatsorait tartalmazza. Fontos, hogy lehetőleg időpont szerint növekvő sorrendben legyenek a megfigyelések.



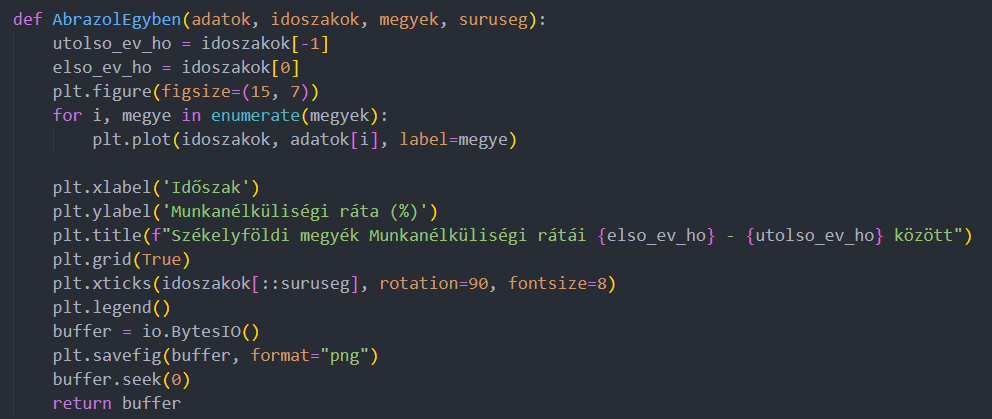
. ábra: A beolvasás kódja

A beolvasás egy html űrlapon keresztül történik, amiben a táblázatot kell feltölteni. A bevitt adatok POST kérésben továbbítódnak a kontroller felé, majd a feldolgozott adatokat egy másik nézetre átirányítva jeleníti meg. A fenti kódban látható, hogy a három megye adatsorait, megnevezéseiket és a megfigyelések dátumait külön listákban tárolom. Magát a beolvasást a panda csomag read\_excel függvénye végzi, amely visszaad egy DataFrame adatszerkezetet. Ezekkel a listákkal könnyen tudom elvégezni a további műveleteket a *numpy*, *pyplot*, *statistics* és *statmodels* nevű csomagok segítségével.

## A statisztikai mutatók számítása

Az 1.2-es pontban definiált statisztikai mutatókat, valamint a minimum- és maximumértékeket a numpy függvényei segítségével számítom ki mindegyik megyére. Az átláthatóság miatt csináltam egy osztályt a megyéknek, amely tárolja az idősort és a mutatókat kiszámolja:  


. ábra

* 1. Az idősorok ábrázolása

. ábra: Az idősort ábrázoló függvény

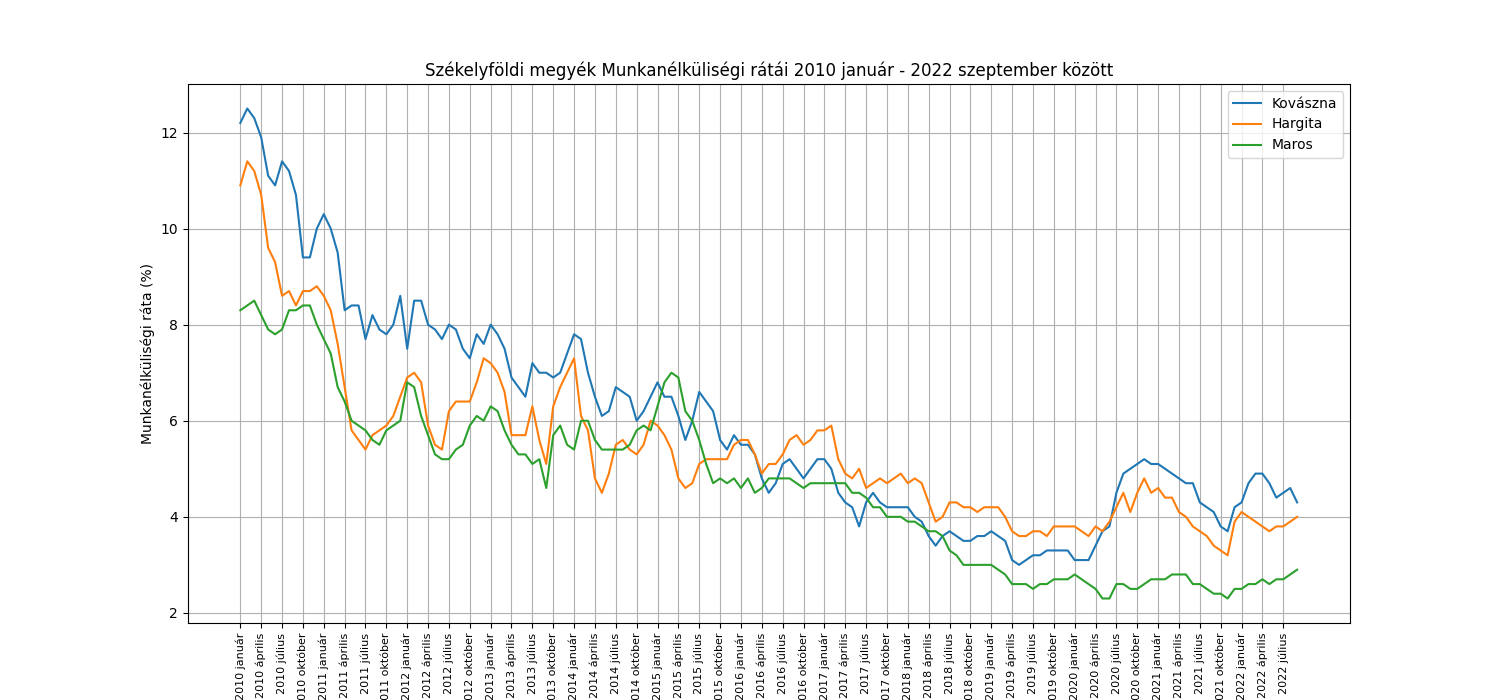
Mivel egyszerre több megye adatsoraival dolgoztam, ezért érdemesnek találtam egy olyan általános függvényt írni, amely paraméterei a következők:

* adatok: Egy listákat tartalmazó lista, amelyben a megyék idősorai találhatóak
* időszakok: Az a lista, amelyben a megfigyelések időpontjai vannak (legelső oszlop). Ez a grafikon X tengelyének feliratozásához kell.
* megyék: String lista, amely az adatok lista elemivel megegyező sorrendben adja meg a megyék neveit, hogy a jelmagyarázatban helyesen tudja feltüntetni az ábra.
* sűrűség: az X tengelyen hány hónaponként legyenek a segédvonalak – az átláthatóság miatt

A kapott adatoknak megfelelően feliratozom a digrammot, ezt követően a generált képet egy bájt bufferbe mentem el, amelyet base64 kódolású szövegként átadok a nézetnek, amit meg tud jeleníteni.

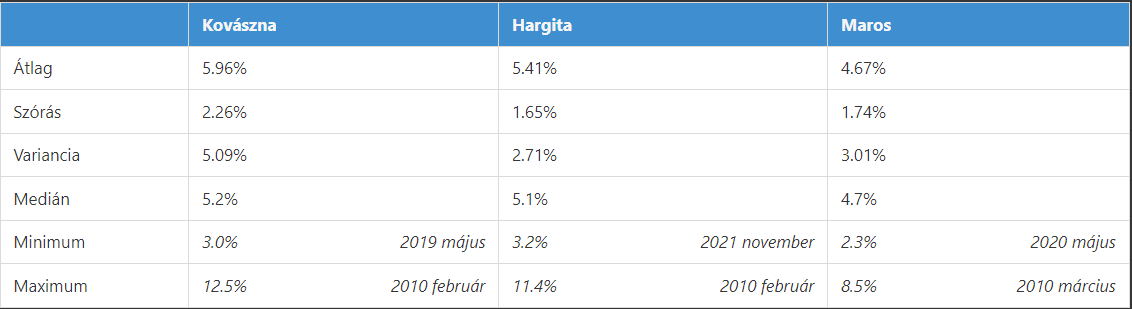
# Az adatok elemzése, a modellek és a paraméterek kiválasztása

A program következő diagrammon szemlélteti a beolvasott adatsorokat:

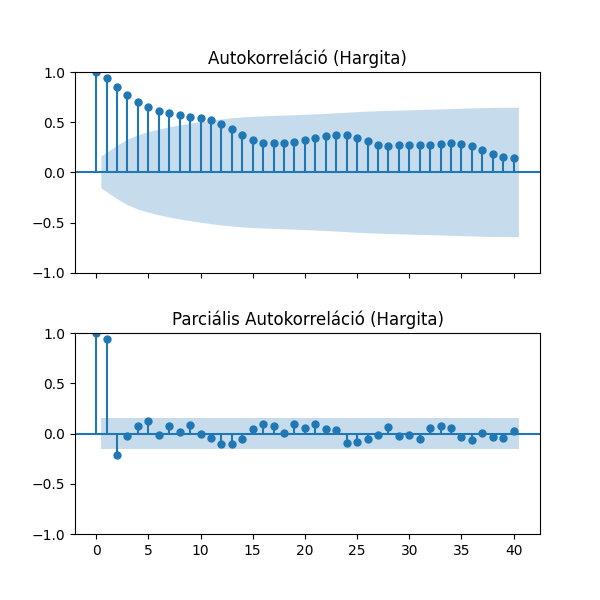
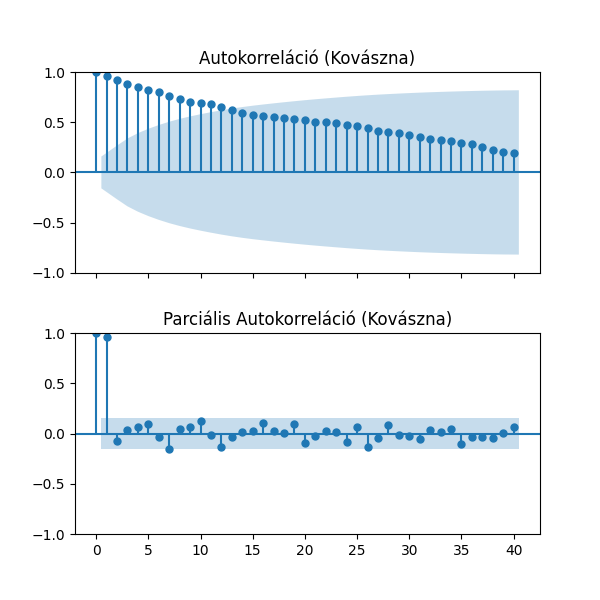
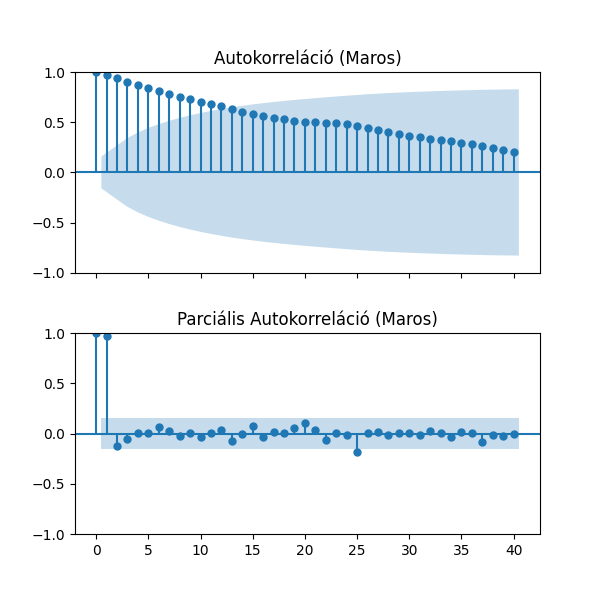


. ábra

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.



. ábra

A program a következő eredményeket adta ki a PACF és ACF függvényekre:  
A diagramok azt szemléltetik, hogy valószínűleg mindhárom megye esetében az AR(2) modellel érdemes becsülni, mivel a PACF első két értéke kimagaslóan magas (szignifikáns), a többi 0-hoz közeli érték.

# Előrejelzés

## 4.1 ARIMA

## 4.2 MLP, CNN

# Következtetések

# Bibliográfia

1. Biroul Internaţional al Muncii [↑](#footnote-ref-1)