**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 az elmúlt 10 év során. Célunk, hogy elemző és előrejelző módszerek segítségével kielégítő elemzést nyújtsunk ezekről az idősorokról, és előrejelzéseket készítsünk a jövőbeli munkanélküliségi rátákra vonatkozóan.

A dolgozatban két fő módszert alkalmazunk az előrejelzések elkészítéséhez: az ARIMA (AutoRegressive Integrated Moving Average) és az MLP (Multilayer Perceptron) modelleket. Az ARIMA modell statisztikai módszer, amely az idősorokat elemzi és előrejelzéseket készít azok alapján. Az MLP pedig egy mesterséges neurális hálózat, amely a gépi tanulásban széles körben használt módszer többek között az idősor előrejelzésére is.

A kutatás során egy Python programot készítettünk az adatok beolvasásához, feldolgozásához és az eredmények megjelenítéséhez. A program lehetővé teszi, hogy többféle adatsort dolgozzunk fel és különböző időszakokra számoljunk előre.

Az eredmények értékeléséhez az INSS (National Institute of Statistics and Studies) adatait használjuk, és összehasonlítjuk az előrejelzéseinket a regisztrált munkanélküliségi rátákkal, hogy megállapítsuk, melyik módszer nyújt jobb becslést.

A következő fejezetekben részletesen bemutatjuk a módszereket, ismertetjük a program működését, továbbá az adatok elemzését, az előrejelzéseket és az eredmények értékelését, valamint kiértékeljük a kutatás eredményeit és következtetéseit.

## 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

**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?)*

## AR, MA, ARMA, ARIMA

**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: 

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 az "q" paraméter megadja a mozgóátlag rendszámát, azaz hány korábbi hibaértéket használunk az aktuális érték becsléséhez.



**Az integráció (I)** azt jelenti, hogy az idősorozatot különbségként vagy differenciálként kezeljük. Az I komponens segít megszüntetni az idősorozatban található trendeket vagy szezonális mintázatokat. Az "d" paraméter megadja az integrációs rendszámot, ami azt jelenti, hány alkalommal végezzük el az idősorozat differenciálását. (Sándor, Tánczos 201?)

## ACF, PACF

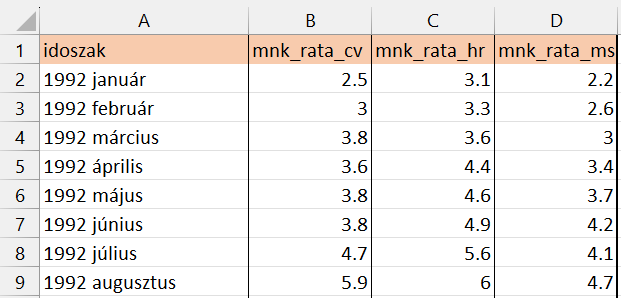
…

# A program működése

## Beolvasás

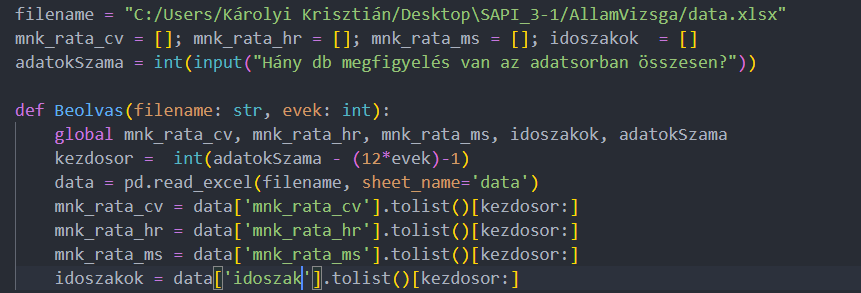
Az adatbeolvasásra egy függvényt írtam, amely két paramétert vár: a beolvasandó Excel file útvonalát (file), és azt, hogy a legutolsó megfigyeléstől számítva hány évre visszamenőleg dolgozza fel ad adatsort. Ez azért hasznos szerintem, mert az én esetemben a letöltött adatforrás több, mint 30 év havi munkanélküliségi rátáit tartalmazta, viszont jó, ha nem kell módosítsuk a bemeneti fájlt, ehelyett inkább kezelje azt rugalmasan a program. Ha ugyanazon idősorból más és más sokaságú megfigyelést vizsgálunk, az teljesen más eredményekhez vezethet a teszteket és előrejelzéseket tekintve.

A bemeneti .xlsx fáj a következő módon kell legyen előkészítve, hogy a program jól tudja beolvasni:



1. ábra: A bemeneti táblázat mintája (saját ábra)

Tehát az első sor mindenképpen fejléc, az első oszlopban a megfigyelések időpontjai (ezek fogják a grafikonon az x tengely értékeit adni), míg a további három oszlop a három megye adatsorait tartalmazza. Fontos, hogy időpont szerint növekvő sorrendben legyenek a megfigyelések, illetve a munkalap neve „data” kell legyen. A program bekéri beolvasás előtt azt, hogy hány db megfigyelés lesz ahhoz, hogy megfelelően tudja kiválasztani a kívánt szakaszt az adatsorból.

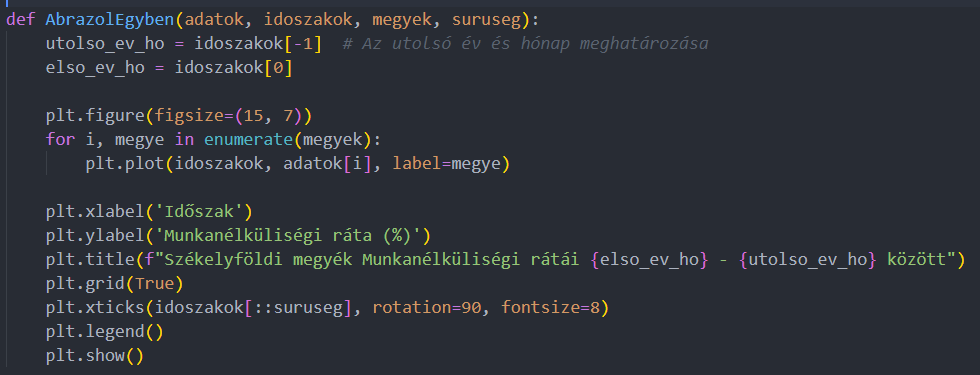


2. ábra: A beolvasás kódja

A fenti kódban látható, hogy a három megye adatait és a megfigyelések dátumait külön listákban tárolom. Magát a beolvasást a panda csomag read\_excel függvénye végzi, amely visszaad egy összetett adatszerkezetet (DataFrame), amely megőrzi az excel táblázat szerkezetét, így az eredeti táblázat fejlécei szerint el lehet érni a különböző adatsorokat. Ezekkel a listákkal már 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. Ezek a függvények mind az akutális adatsort kérik be. Például, Hargita megye adatsorának a szórása: np.std(mnk\_rata\_hr).

* 1. Az idősorok ábrázolása

3. 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

# Az adatok elemzése

# Előrejelzés

## 4.1 ARIMA

## 4.2 MLP

# Következtetések

# Bibliográfia

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