# A bárányhimlő alakulása Magyarországon

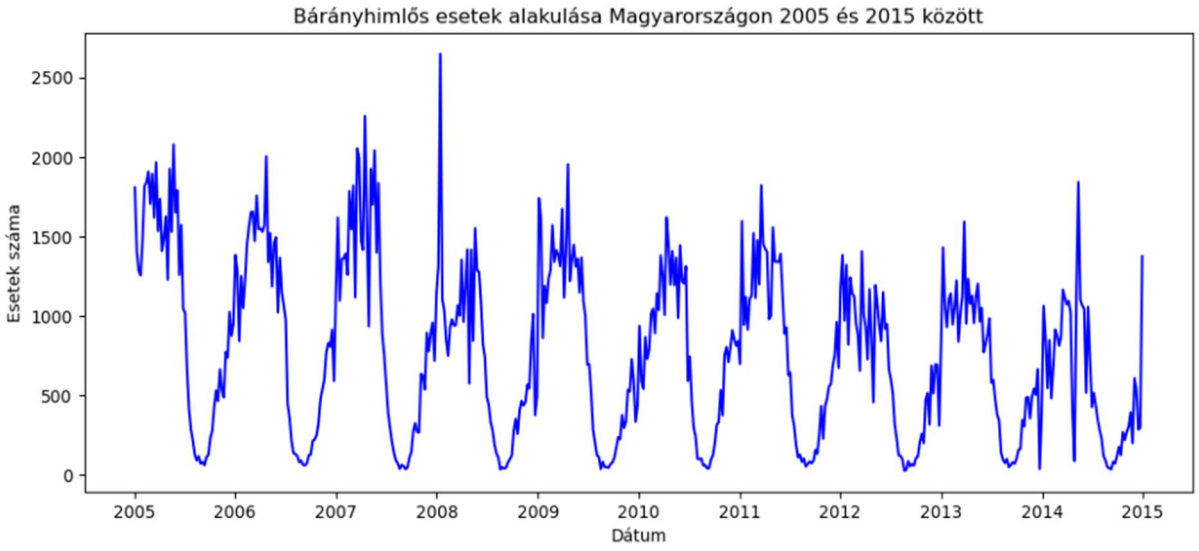
A bárányhimlő (latinul: varicella) az egyik leggyakoribb és legragályosabb, cseppfertőzéssel terjedő gyermekbetegség, amelyen a védőoltással nem védett emberek zöme még gyermekkorában átesik. Az elemzésem célja a betegség helyzetének, és alakulásának bemutatása hazánkban. További célom, hogy 20 hétnyi előrejelzést készítsek az esetek számára vonatkozóan.

Az adatok forrása egy, a magyarországi bárányhimlős eseteket magába foglaló térbeli-időbeli .csv állomány, ami 2005 és 2015 közötti, megyei szinten bejelentett esetek idősoraiból áll. Magára a fájlra a Kaliforniai Egyetem honlapján bukkantam.

Az adatok betöltése után láthatjuk, hogy a dataframe-ünk a dátumokat, az összes megyét és Budapestet kiemelten tartalmazó oszlopokból áll. 522 rekordunk van, ami ugyanennyi hétnek felel meg, így a 10 évnyi intervallumból eredő elvárásunknak megfelel az állomány.

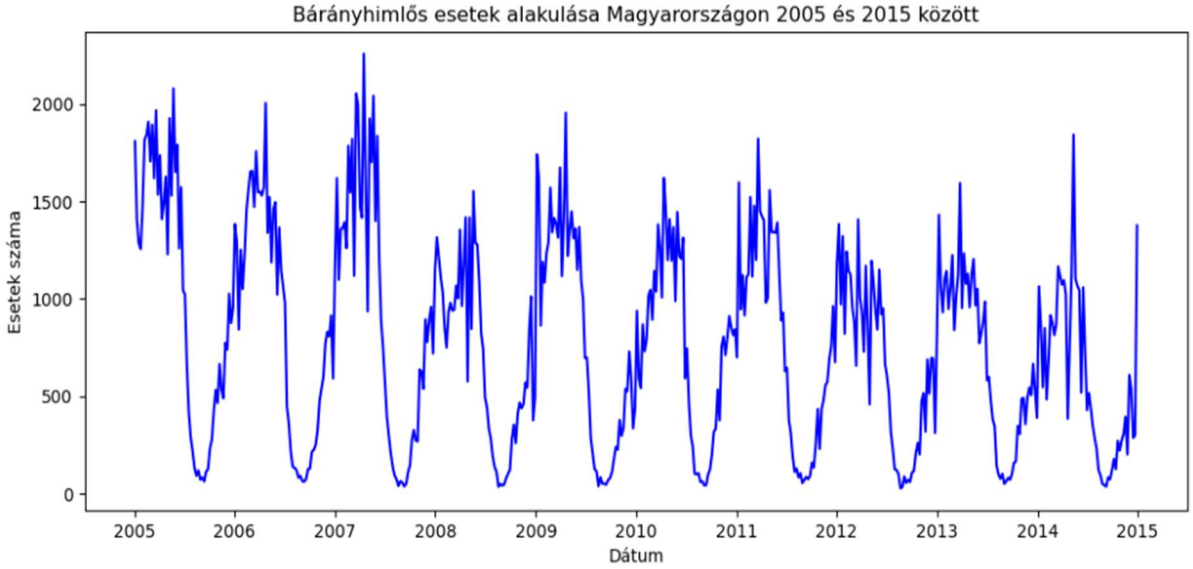
Mivel a dátum adatokat tartalmazó oszlop “object” típusú, így azt első lépésben a duplikációk, és a hiányzó értékek ellenőrzése után dátum típusúvá változtatjuk, illetve ebből indexet is képzünk.

Két konkrét ide kapcsolódó feladatunk is lehet. Az egyik a megyei szintű, a másik pedig az országos szintű esetszám előrejelzés. Mivel most az országos szintet szerettem volna elemezni, ezért létrehoztam egy új, összesített adatokat tartalmazó oszlopot, “ÖSSZES_ESET” néven.  Ezek után ábrázoltam az adatokat:

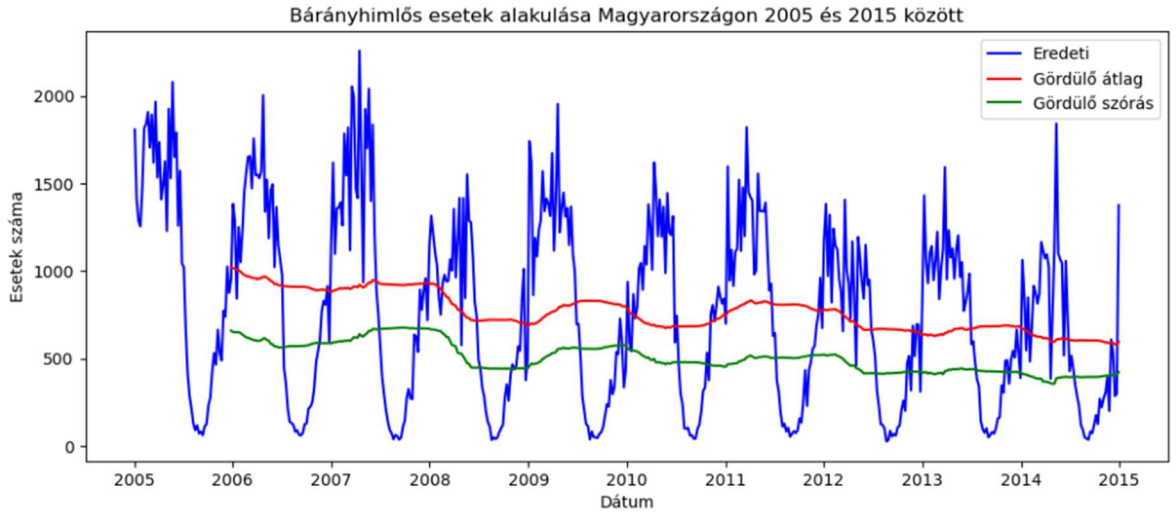



Felfedezhetjük, hogy az idősorban egy nagyon enyhe csökkenő tendencia, és szezonális ciklikusság rejlik. Egy ciklus pont egy év, azaz 52 hét. A legtöbb eset mindig az első negyedévben fordul elő. Ezen felül észrevehetjük, hogy van 3 feltűnően kiugró értékünk: egy rekord 2007. év végén, ami 2500 esetnél nagyobb számot rögzít, illetve két másik a 2013. év végén kezdődő periódusban, amelyek felemelkedő ágban 100nál is kevesebb adatot

rögzítettek. Mivel az esetszámok hetente kerültek rögzítésre, így előfordulhat, hogy ezek valós, és helytálló számok, de ettől függetlenül eltávolítottam őket. Ezt úgy tettem meg, hogy először a numpy könyvtár segítségével NaN értékűvé alakítottam, majd a pandas interpolate() funkciójával, lineáris interpolációval pótoltam őket. Ennek eredménye a viszonylag kiugrásmentes grafikon:

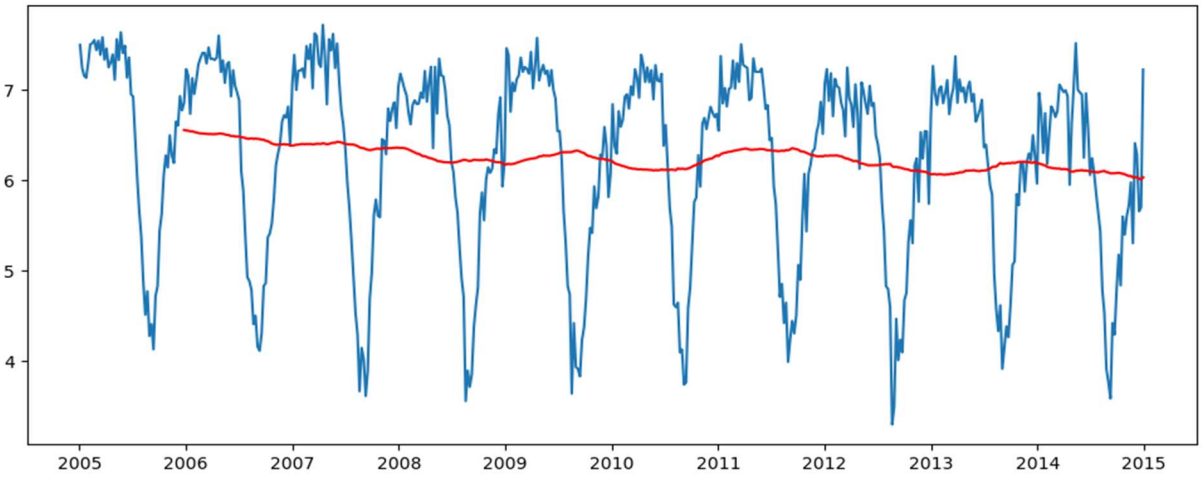



Néhány idősorelemzéshez használt modell - mint amilyet most én is használtam - stacionaritást feltételez. Abban az esetben, ha idősorunk állandó átlaggal, állandó szórással rendelkezik és az auto-kovariancia sem függ az időtől, akkor az stacionárius. Ennek ellenőrzésére kétféle módszert is bemutatok. Először grafikusan szemléltetem az átlagot, és a szórást egy évnyi ráhagyással.

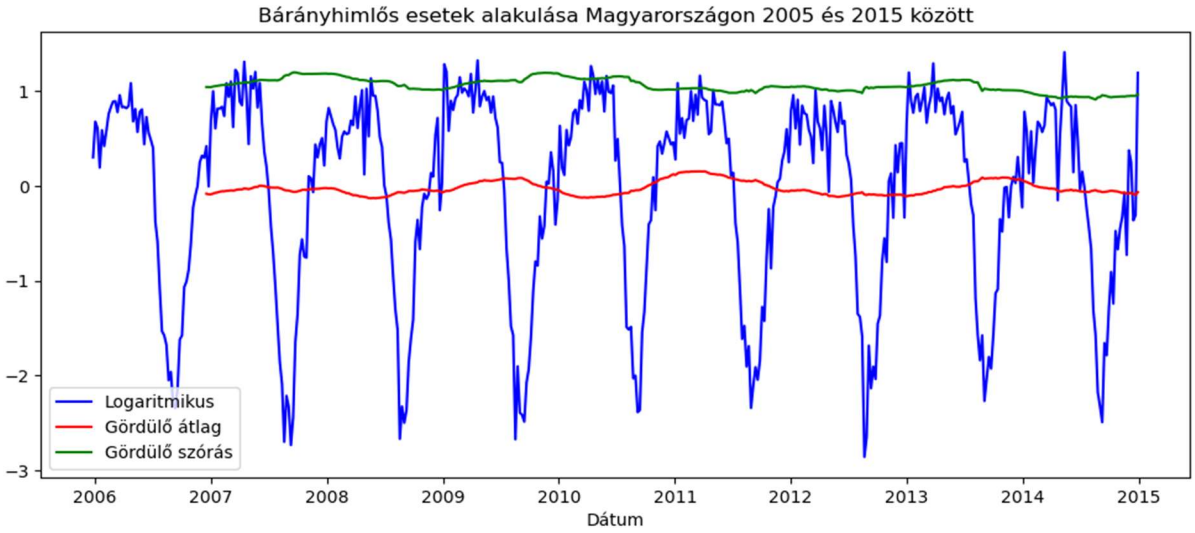



Látjuk, hogy sem a szórás, sem az átlag nem állandó, így ez az idősor még nem stacionárius. Több tényező is elronthatja a stacionaritást, például a trend vagy a szezonalitás.

A trend hatását többféle átalakítással is lehet csökkenteni, én az idősor elemeinek logaritmusát vettem.

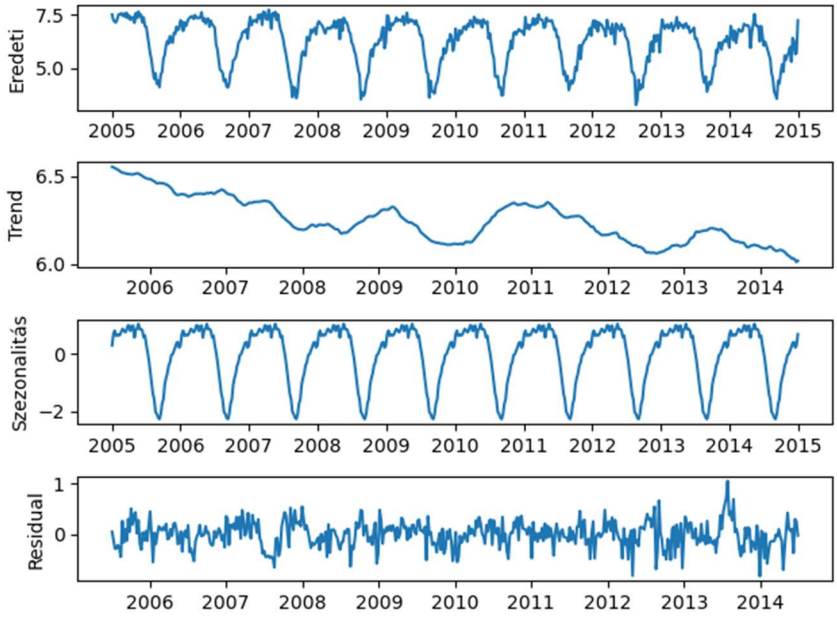



Némileg javult a gördülő átlag. Ha a logaritmus függvény és a mozgó átlagának különbségét képezzük, akkor az alábbi eredményt kapjuk:

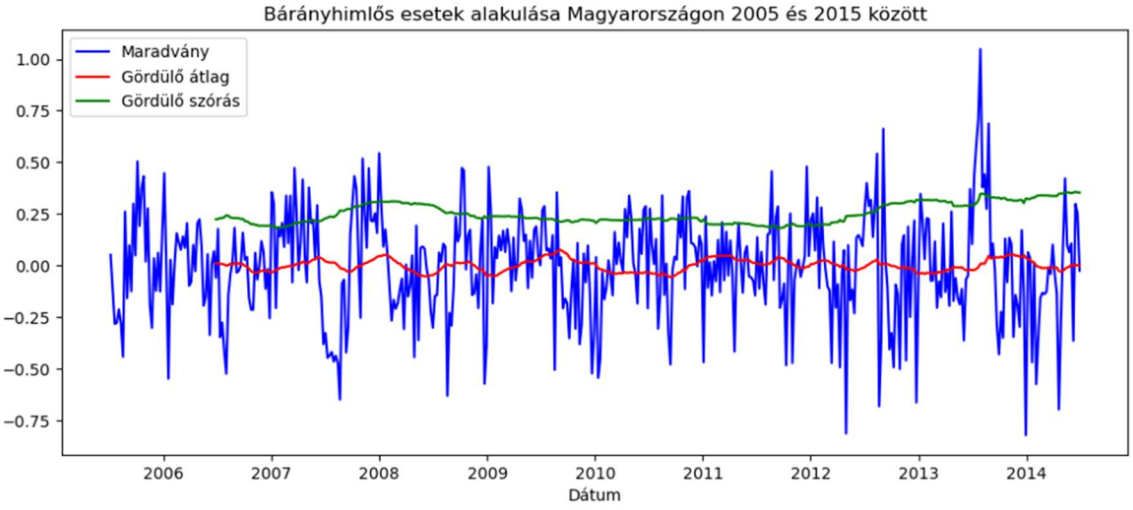



A következő módszer, amivel a szezonalitás hatását lehet csökkenteni az a dekompozíció. Ennek segítségével az idősor szétbonatható trend, szezonális és maradvány hatásokra.

Additív dekompozíció esetén **Y = T + Sz + R .**




Itt pedig a trendet, illetve szezonalitást elhagyva, és csak a maradványt használva az alábbi eredményt kapjuk.

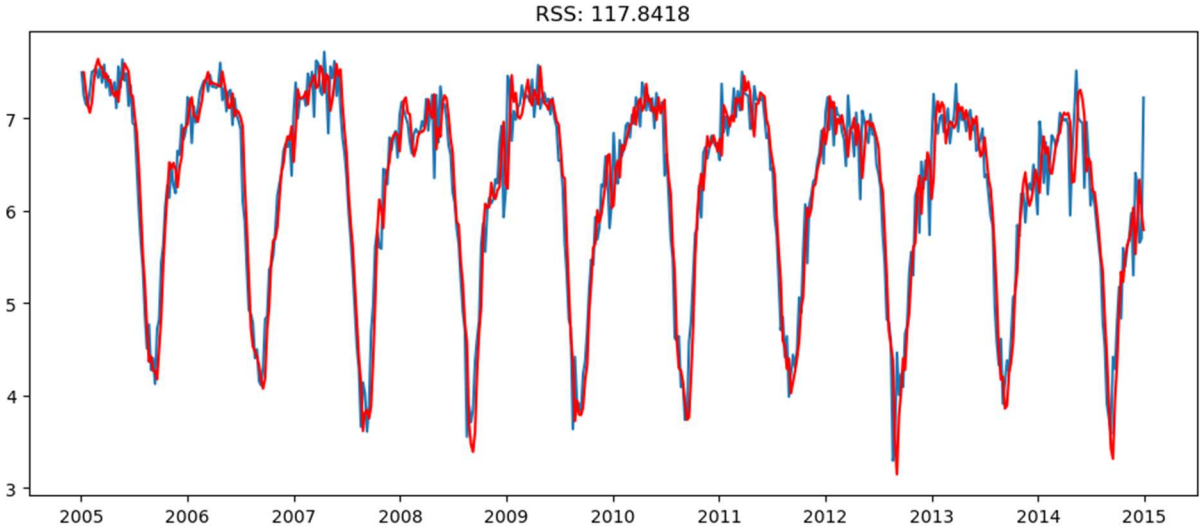



Ezt az eredményt most a Dickey-Fuller teszttel is leellenőriztem.

```
A Dickey-Fuller teszt eredménye:
Test Statistic                -7.634378e+00
p-value                        1.970140e-11
#Lags Used                     2.000000e+00
Number of Observations Used    4.670000e+02
Critical Value (1%)           -3.444431e+00
Critical Value (5%)           -2.867749e+00
Critical Value (10%)          -2.570077e+00
```

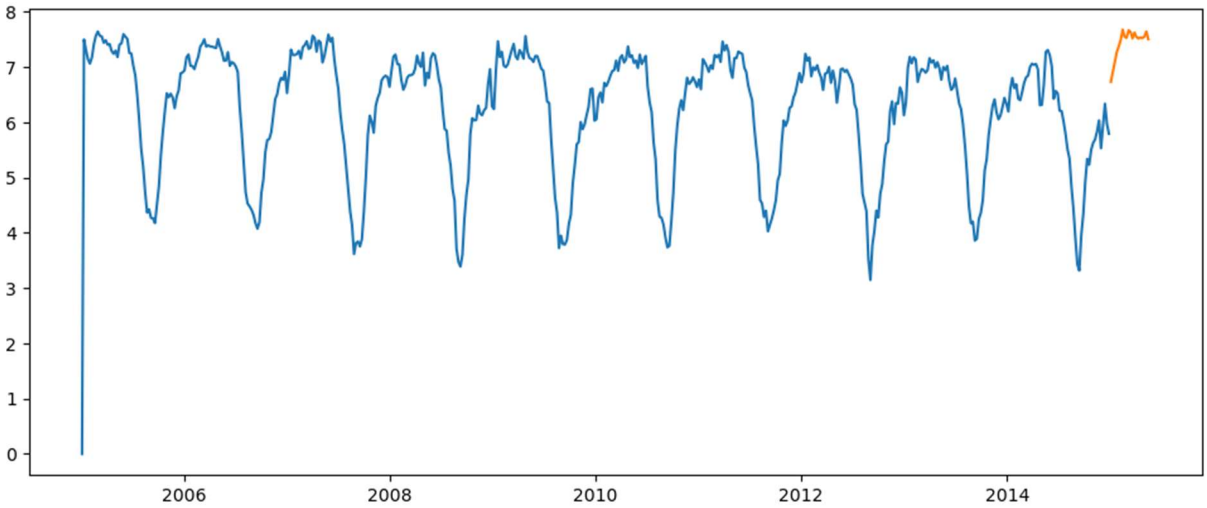
A teszt H0 hipotézise, hogy az idősor nem stacionárius. Abban az esetben ha a p-érték kisebb, mint 0,05, illetve a test érték kisebb, mint a kritikus értékek, akkor H0-t elutasítjuk, és az idősort stacionáriusként elfogadjuk. Így a mi esetünkben 99%-os szignifikancia szinten stacionárius az idősor.

Az előrejelzéshez az ARIMA (Autoregressive Integrated Moving Average) modell szezonalitás esetén alkalmasabb változatát, a SARIMA modellt használtam, ami két eljárás egymásba integrált verziója. Az AR modell egy regressziós modell, ami a múltbéli értékeket, az MA modell pedig a jelenlegi érték és a mozgóátlag hibája közötti függőséget használja fel. A legideálisabb paraméterek megkereséséhez az auto_arima() funkciót hívtam segítségül, amik alkalmazásával elért eredményem az alábbi:




Ebben az esetben az RSS (residual sum of squares) 117,84 lett. Ez az érték annál kisebb minél pontosabb a illeszkedés.

A modell elkészítése, és a predikció után 20 hétnyi intervallumot jeleztem előre, amit az alábbi grafikonon ábrázoltam.



Mint látható az esetszám a jövőben sem csökken radikálisan. A bárányhimlő ellen létezik védőoltás, ami Magyarországon 1998-óta elérhető. A vélemények az oltás szükségességéről megoszlanak. A fő ellenérv, hogy a bárányhimlő gyors lefolyású, enyhe betegség, azonban átlagban minden tizedik esetben fellép valamilyen szövődmény, amelyek eltérő súlyosságúak lehetnek. A 2006-os évben 444 beteget kellett bárányhimlő miatt kórházban kezelni Magyarországon. 2019 szeptemberétől a bárányhimlő elleni vakcina felkerült a kötelező oltások listájára, így ezt követően javulás várható.