

OTDK dolgozat

2015

**Egy fertőző gyermekbetegség alakulásának
modellezése és elemzése**

**Modelling and analysing the time series of
a contagious childhood disease**

Kézirat lezárva: 2014. június 19.

Tartalomjegyzék

Ábrák jegyzéke

Táblázatok jegyzéke

1. Bevezetés	1
2. Egyváltozós modellezés	4
2.1. Determinisztikus-sztokasztikus modell	5
2.1.1. Fehérzaj tesztek alkalmazása az összesített idősorra	5
2.1.2. Egységgyök tesztek alkalmazása	8
2.1.3. Periodicitás vizsgálata	11
2.1.4. Indikátor változós modell illesztése	12
2.1.5. Trigonometrikus trend beépítése	14
2.1.6. ARMA tagok identifikálása	21
2.2. Sztokasztikus modell	25
2.2.1. A SARMA tagok identifikálása	25
2.2.2. A SARIMA tagok identifikálása	27
2.3. A Poisson-regressziós megközelítés	29
2.3.1. Nemlineáris Poisson-regresszió bevezetése	29
2.3.2. Determinisztikus modell specifikáció	31
2.3.3. Nemlineáris Poisson-regresszió autoregresszív tagokkal	35
3. Egyváltozós előrejelzések és elemzések	39
3.1. A determinisztikus modell előrejelzései	39
3.2. A determinisztikus-sztokasztikus modell előrejelzései	41
3.3. A sztokasztikus modell előrejelzései	43
3.4. A determinisztikus nemlineáris Poisson-regresszió előrejelzései	44
3.5. A sztokasztikus Poisson-regresszió előrejelzései	46

4. Többváltozós modellezés és elemzések	48
4.1. Főkomponens elemzések	49
4.1.1. Főkomponensek a megyei idősorokra	49
4.1.2. Főkomponensek a szezonálisan differenciázott megyei idősorokra	51
4.2. A Moran-féle I-index	52
4.2.1. A Moran-féle I-index a megyei idősorokra	55
4.2.2. A Moran-féle I-index a szezonálisan differenciázott megyei idősorokra	56
4.3. A Geary-féle C-index	56
4.3.1. A Geary-féle C-index az idősorokra	57
4.4. A Bayes-i vektor autoregresszió	58
5. Összefoglalás	61
5.1. További kutatási lehetőségek	62
6. Irodalomjegyzék	64
Függelékek	I
A. Egyváltozós modellezés	I
A.1. Ljung-Box Q-statisztikák az összesített idősorra	I
A.2. Augmented Dickey-Fuller tesztek eredményei	II
A.3. Az országos idősorhoz tartozó periodogram	III
A.4. Periodikus dummy változós OLS modell jellemzői	IV
A.5. Az I. NLS modell <i>Gretl</i> optimalizációs kódja	V
A.6. A II. NLS modell <i>Gretl</i> optimalizációs kódja	V
A.7. A III. NLS modell <i>Gretl</i> optimalizációs kódja	VI
A.8. Ljung-Box Q-statisztikák a III. NLS modell maradékaira	VII
A.9. ARMA modellek a III. NLS modell maradékaira	VIII
A.10. Ljung-Box Q-statisztikák a kombinált modell maradékaira	IX
A.11. SARMA modellek a megbetegedés számra I.	X
A.12. SARMA modellek a megbetegedés számra II.	XI
A.13. SARIMA modellek a megbetegedés számra	XII
A.14. Ljung-Box Q-statisztikák a SARIMA I. modell maradékaira	XIII
A.15. Ljung-Box Q-statisztikák a SARIMA II. modell maradékaira	XIV
A.16. Determinisztikus Poisson-regresszió maximum likelihood kódrészlet	XV
A.17. Sztochasztikus Poisson-regresszió maximum likelihood kódrészlet	XV
A.18. Sztochasztikus Poisson-regresszió maradékához tartozó korrelogram	XVII

A.19.Determinisztikus modell optimalizációs kód	XVIII
A.20.Poisson-regresszió optimalizációs kód	XVIII
A.21.Determinisztikus modell rugalmasság kód	XIX
A.22.Poisson-regresszió rugalmasság kód	XIX
B. Többváltozós modellezés	XX
B.1. Deskriptív statisztikák a megyei idősorokra	XXI
B.2. Főkomponensek a megyei idősorokra	XXII
B.3. Forgatott komponens mátrix a megyei idősorokra	XXIII
B.4. Deskriptív statisztikák a szezonálisan differenciázott idősorokra	XXIV
B.5. Főkomponensek a szezonálisan differenciázott idősorokra	XXV
B.6. Forgatott komponens mátrix a szezonálisan differenciázott idősorokra	XXVI
B.7. Bináris területi súlymátrix Magyarországra	XXVII
B.8. A Moran-féle I-index értékét számító kód	XXVIII
B.9. A Moran-féle I-indexre felépített AR modell	XXIX
B.10.A Geary-féle C-index értékét számító kód	XXX
B.11.Bayes-i prior kód	XXXI

Ábrák jegyzéke

2.1. Az országos idősor	4
2.2. Az országos idősorhoz tartozó korrelogram	9
2.3. Az országos idősorhoz tartozó periodogram	11
2.4. Havi lebontású dummy változók alkalmazása	13
2.5. Az I. NLS modell becslései	16
2.6. A III. NLS modell becslései	21
2.7. A III. NLS és ARMA(2,2) modell becslései a tény adatokkal szemben	24
2.8. A SARIMA(3,1,3)(1,0,1) modell becslései a tény adatokkal szemben	28
2.9. A Poisson-regresszió becslései a tény adatokkal szemben	34
2.10. A sztochasztikus Poisson-regresszió becslései a tény adatokkal szemben	38
3.1. A III. NLS által becsült szélsőértékek	40
3.2. A III. NLS modell időindex szerinti rugalmassága	41
3.3. A III. NLS és ARMA(2,2) becslései	42
3.4. A SARIMA (3,1,3)(1,0,1) becslései	43
3.5. A determinisztikus Poisson-regresszió által becsült szélsőértékek	45
3.6. A determinisztikus Poisson-regresszió időindex szerinti rugalmassága	46
3.7. A sztochasztikus Poisson-regresszió becslései	47
4.1. Nem differenciázott főkomponensek alapján színezett térkép	51
4.2. Szezonálisan differenciázott főkomponensek alapján színezett térkép	53
4.3. Kapcsolati hálózat 6 megyére	54
4.4. A Moran-féle I-index a megyei idősorokra	55
4.5. A Moran-féle I-index a szezonálisan differenciázott megyei idősorokra	56
4.6. A Geary-féle C-index a megyei idősorokra	57
A.1. A III. NLS és ARMA(2,2) maradékhoz tartozó korrelogram	IX
A.2. A SARIMA (3,1,3)(1,0,1) maradékhoz tartozó korrelogram	XIII

A.3. A SARIMA (4,1,4)(1,0,1) maradékhoz tartozó korrelogram	XIV
A.4. A sztochasztikus Poisson-regresszió maradékához tartozó korrelogram	XVII
B.1. A Moran-féle I-indexre felépített modell maradékához tartozó korrelogram	XXIX

Táblázatok jegyzéke

2.1. Az ADF tesztek eredményei	10
2.2. A DHF tesztek eredményei	11
2.3. Az I. NLS modell jellemzői	15
2.4. A II. NLS modell jellemzői	18
2.5. A III. NLS modell jellemzői	20
2.6. A maradékra illesztett ARMA(2,2) modell	24
2.7. A SARIMA(3,1,3)(1,1,1) modell jellemzői	28
2.8. A determinisztikus Poisson–regressziós modell jellemzői	34
2.9. A sztochasztikus Poisson–regressziós modell jellemzői	37
A.1. Az összesített idősorhoz tartozó Q-statisztikák	I
A.2. Az országos adatokra vonatkozó periodogram értékei	III
A.3. Periodikus dummy változós OLS modell jellemzői	IV
A.4. A III. NLS modell maradék tagjához tartozó Q-statisztikák	VII
A.5. ARMA modellek a III. NLS modell maradékaira	VIII
A.6. A III. NLS és ARMA(2,2) modell maradékának Q-statisztikái	IX
7.7. SARMA modellek a megbetegedés számra I.	X
7.8. SARMA modellek a megbetegedés számra II.	XI
7.9. SARIMA modellek a megbetegedés számra III.	XII
A.10.A SARIMA (3, 1, 3)(1, 0, 1) modell maradékának Q-statisztikái	XIII
A.11.A SARIMA (4, 1, 4)(1, 0, 1) modell maradékának Q-statisztikái	XIV
A.12.A sztochasztikus Poisson–regresszió maradékához tartozó korrelogram	XVII
B.1. Deskriptív statisztikák a megyei idősorokra	XXI
B.2. Főkomponensek a megyei idősorokra	XXII
B.3. Forgatott komponens mátrix a megyei idősorokra	XXIII
B.4. Deskriptív statisztikák a szezonálisan differenciázott idősorokra	XXIV
B.5. Főkomponensek a szezonálisan differenciázott idősorokra	XXV
B.6. A Moran-féle I-indexre felépített AR(3) modell jellemzői	XXIX

- *Társadalomtudomány-e a közgazdaságtan?*
- *Nem.*
- *De hiszen emberek viselkedésével és emberi interakciók leírásával foglalkozik. Akkor mégis-csak igen?*
- *Nem. Ezek a módszerek és modellek a hagyományosan gazdaságinak, sőt szociológiaiinak tekintett jelenségek mellett alkalmasak biológiai és pszichológiai jelenségek leírására is.*

(Kőhegyi, 2013)

1. fejezet

Bevezetés

A dolgozatban a Magyarországon bejelentett bárányhimlős megbetegedések heti idősorait vizsgálom 1998 márciusa és 2013 augusztusa között. Az egyváltozós modellezés során az erre az időintervallumra kiterjedő adatok jelentik a tanulómintát. A többváltozós elemzések esetében egy szűkebb tanuló mintával dolgozok, ez a minta 2005 januárjától 2013 augusztusáig terjed. A 2013 szeptembere után felvett adatokat meghagyom a felépített modell becsléseinek jellemzésére. Az adatbázis teljesen egyedülálló, ugyanis az adatok egy része nem érhető el elektronikusan, és az elektronikus adatokból sem érhető el a kész adatbázis. Az adatbázis összeállítása során az Országos Epidemiológiai Központ – innentől kezdve OEK néven rövidítem — hetente megjelenő lapjának, digitálisan kiadott és papíralapú jelentéseit használtam fel – Epiinfo (OEK, 2013). A 2001. év előtti táblázatok digitálisan nem tette közzé az OEK, azonban digitalizált formában megszereztem a lap összes eddig megjelent számát, ezekhez online elérhetőséget készítettem, amely a források között elérhető (OEK, 2014). Ezekben a dokumentumokban hetekre lebontva megtalálható az összes adott héten regisztrált bejelentés kötelel megbetegedés, amely valamilyen fertőző betegséghez köthető. Ilyen betegség például a tuberkulózis, *HIV* illetve a vizsgálat témájának választott bárányhimlő.

A hazai bárányhimlős megbetegedés szám, mint induktív statisztikával módszerekkel elemezhető epidemiológiai jelenség az alábbi tényezők miatt elemezhető jól:

1. A betegséget nem övezi semmilyen intézményesült megkülönböztetés vagy megbélyegzés, azaz a betegség tényét nem fogják eltitkolni (Leary, Schreindorfer, 1998; Kelly, 1999).
2. Más hazai bejelentés kötelel fertőző betegségekkel ellentétben, az év minden időpontjában nagy számban fordul elő a populációban (OEK, 2013).

3. A hazai bejelentés köteles fertőző betegségek egy részével ellentétben, a bárányhimlő nem köthető csak az ország bizonyos városaihoz, megyéihez, azaz nem területi klaszterekben van jelen (Csepe, 2002).

4. A bejelentési kötelezettség miatt a teljes populáció megfigyeltnek tekinthető.

A hasonló nemzetközi kutatások egy része havi lebontású adatsorokat elemzett, amelyek az esetek többségében nem terjedtek ki a vizsgálat tárgyát képező ország egészére (Chan, Tian, 2011; Perez-Farinos, Ordobas, 2007). A kutatások egy másik része csak pár évre visszamenőleg elemezte a megbetegedések számát (Severine et al., 1998), a legalább egy évtizedet átfogó adatokat felhasználó elemzések esetében a felhasznált adatok gyakran aggregált éves értékek (Carol, Noah, 1988). A gyermekbetegségek viselkedésének sztochasztikus és determinisztikus módszerekkel történő elemzése az epidemiológiai irodalomban igen kedvelt (Bärbel, Grenfell, 2000; Louise et al., 1994). Az ilyen típusú elemzések alkalmazását indokolja az, hogy a jelenség egész egyszerűen öngerjesztőnek tekinthető a fertőzések miatt (sztochasztikus jelleg), emellett periodikusan visszatérő visszaesések, és adott hetekhez köthető csúcspontok figyelhetőek meg a tanítási szünetek közelében (determinisztikus jelleg). A betegség 2-3 hetes lappangási ideje mellett a heti lebontású idősor lehetővé teszi azt, hogy releváns, információ vesztést nem tartalmazó modellt alkossunk, erre a korábban említett havi lebontású idősorokkal foglalkozó tanulmányok szerzői nem tudtak kísérletet tenni. Maga a nagy mintaelemszám (hosszú idősor) önmagában egy előnyt jelent ezekkel a tanulmányokkal szemben.

A felhasznált adatbázis tartalmazza a megyékre és Budapestre lebontott idősorokat is 2005 januárjától kezdve, amely lehetővé teszi azt is, hogy ne csak az aggregált adatokat vizsgáljuk, és az esetleges területi autokorrelációt is elemezhessük.

A dolgozatban három kérdésre keressük a választ:

1. Az iskola és óvoda kezdés, illetve a nyári szünet hatása kimutatható-e az országos idősorban?
2. A lappangási időnek megfelelő sztochasztikus hatások vannak-e jelen az országos idősorban?
3. Kimutatható-e területi alapú együttmozgás a megyékre vonatkozó idősorokban?

A három feltett kérdésre adott hipotetikus válaszuk minden egyes esetben igen, a célunk ennek a három a hipotetikus válasznak a bizonyítása.

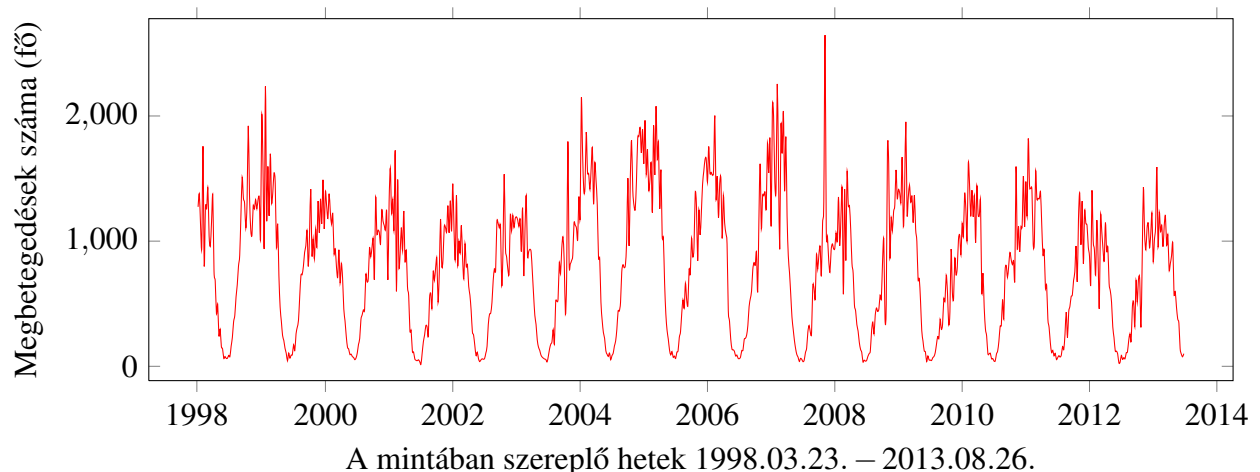
A dolgozat felépítése az alábbi: a 2. fejezetben három eltérő regressziós modellt építünk fel az aggregált országos idősorra. A három megközelítés egyaránt alkalmaz determinisztikus és sztochasztikus elemeket. Ezen módszerekkel az 1. és 2. hipotézis bizonyítása a célunk. Az egyváltozós előrejelzéseket a 3. fejezet tartalmazza. A 4. fejezetben a területi korrelációs kapcsolatokat vizsgáljuk, az alkalmazott módszertan ebben a fejezetben nem korlátozódik a regressziós modellezésre; ebben a fejezetben a 3. hipotézist akarjuk bizonyítani. A vektor autoregressziós modell mellett, területi autokorrelációs indexek és főkomponens elemzés is szerepel. Az 5. fejezet foglalja össze az eredményeinket és hozzájárulásainkat.

A dolgozatban szereplő ábrák a $\text{\LaTeX}$ *pgfplot* csomagjával készültek, mivel vektor grafikusak ezért tökéletesen nagyíthatóak a dolgozat elektronikus változatában, az ábrákon szereplő távolságok és arányok a nagyítástól nem torzulnak.

2. fejezet

Egyváltozós modellezés

Ebben a fejezetben az összesített – minden megbetegedést tartalmazó országos idősor elemzésével foglalkozunk. A modellezés megkezdése előtt érdemes megvizsgálni, hogy miképpen alakul a megbetegedések száma az idő függvényében ábrázolva.



2.1. ábra. Az országos idősor

A 2.1 ábra alapján érzékelhető, hogy az idősorban erős szezonális van jelen. Az is adódik, hogy a probléma kezelésére nem a legelőnyösebb heti lebontású indikátor változók alkalmazása, az idősort jellemző frekvencia miatt, hiszen hatalmas bevont prediktor számot jelentene az 51 indikátor. Az interakciókkal szerepeltetett polinomiális trendek esetében pedig nehéz lenne meghatározni a megfelelő illesztési pontokat a konvex és konkáv szakaszok között. A determinisztikus modellek mellett, a *SARIMA* modellezés sem feltétlenül elvetendő, azonban az egy évvel korábbi megbetegedéseknek nem biztos, hogy lehet köze, az adott évi megbetegedés számhoz. Mivel az idősor a megbetegedések darabszámát írj le egy harmadik – egy Poisson-regressziós megközelítés is alkalmazható lesz az idősorra. Mindezek miatt összesen

tehát három modellt fogunk vizsgálni: egy determinisztikus és sztochasztikus elemeket is tartalmazó modellt; egy teljesen sztochasztikus és egy Poisson-regressziót. Az első kettő így kapott modellt a *likelihood* alapú szelekciós kritériumokkal (*AIC*, *BIC*, *HIC*) lehet majd összehasonlítani az eltérő illesztési elvek miatt. A harmadikat pedig az illeszkedés jóságát jellemző mutatók (*MAPE*, *RMSE*) alapján lehet majd összehasonlítani az első két modellel.

2.1. Determinisztikus-sztochasztikus modell

Az első megközelítés a klasszikus Box-Jenkins (Box, Jenkins, 1970) metódus lesz, amely során a szezonális kezelést biztosító determinisztikus trendet nemlineáris legkisebb négyzetek módszerével fogjuk illeszteni (*NLS*).

A teljes eljárás sémája az alábbi lesz:

1. Fehérzaj vizsgálat az idősorra
2. Szezonális és nem szezonális egységgyök teszt alkalmazása
3. Szezonális kezelés
 - (a) Periodicitás vizsgálata
 - (b) Determinisztikus trend illesztése
 - (c) Fehérzaj vizsgálat a maradék tagra nézve
 - (d) Szükséges $ARMA(p, q)$ rendű tagok identifikálása
4. Újabb fehérzaj vizsgálat a maradékra nézve

2.1.1. Fehérzaj tesztek alkalmazása az összesített idősorra

Az összesített országos idősor esetében először tesztelni kell azt, hogy a folyamat pusztán fehérzajnak tekinthető-e, azaz alkalmas-e egyáltalán arra, hogy sztochasztikus modellt építsünk rá. Egy idősor fehérzajnak tekinthető, ha adott i késleltetés mellett állandó 0 az autokorrelációs együtttható nagysága, és konstans várhatóértéke van. Az első alkalmazott fehérzaj teszt a Ljung-Box (Ljung, Box, 1978) által bevezetett Q-statisztika lesz, amelynek a hipotézisrendszere az alábbi módon írható fel:

$$H_0 : \forall \rho_i = 0 \text{ és } i = 1, \dots, k$$

$$H_1 : \exists \rho_i \neq 0 \text{ és } i = 1, \dots, k$$

A nullhipotézis esetünkben úgy fogható fel, hogy az adott időszaki megbetegedés szám korrelálatlan az $i = 1, \dots, k$ időszakokkal korábban bejelentett bárányhimlős megbetegedések számával. A teszt segítségével tetszőleges k -ad rendig bezárólag az autokorrelációs kapcsolatot vizsgálhatjuk az idősorra nézve. A próbafüggvény értékét, az alábbi képlettel kaphatjuk meg:

$$Q_k = n \cdot (n + 2) \cdot \sum_{i=1}^k \frac{\rho_i^2}{n - i},$$

ahol n a minta elemszáma, i a késleltetés rendje, k a maximális késleltetés szám, és ρ_i az autokorrelációs együttható adott i késleltetés mellett. A vizsgálatot az első $k = 15$ késleltetésre fogjuk elvégezni, és a nullhipotézis esetleges elvetéséhez az $\alpha = 1\%$ szignifikancia szintet fogjuk választani. A kritikus értékek rendre χ^2 -eloszlást követnek k szabadsági fokkal, ahol k a késleltetések száma. Az előzetesen választott szignifikancia szint mellett, bármilyen tetszőleges $k = 1, \dots, 15$ késleltetés mellett elvetjük a nullhipotézist, hiszen az empirikus szignifikancia szint minden esetben $p = 0.000$. A megbetegedések száma időben nem korrelálatlan a Ljung-Box Q-teszt alapján. Az adott késleltetések melletti autokorrelációs együtthatókat, a próbafüggvények értékeit és az empirikus szignifikancia szinteket tartalmazó táblázat megtalálható a függelék A.1 alfejezetében.

Az idősor fehérzaj mivoltát egy Wald-F (Wald, 1943) teszt segítségével is tesztelhetjük, amelynek a hipotézisrendszere az alábbi:

$$H_0 : \forall \rho_i = 0 \text{ és } i = 1, \dots, k$$

$$H_1 : \exists \rho_i \neq 0 \text{ és } i = 1, \dots, k$$

A nullhipotézis értelmezése ebben az esetben az, hogy k késleltetés mellett a megelőző k időszakban történt megbetegedések számának nincsen szignifikáns autokorrelációja a tárgyidőszak megbetegedések számával. Az ellenhipotézis pedig az, hogy a k megelőző időszak közül, legalább az egyik időszak megbetegedés száma autokorrelált a tárgyidőszaki megbetegedés számmal.

A teszt megköveteli egy tengelymetszetet és k autoregresszív késleltetés tartalmazó regressziós modell *OLS* becslését, amelynek az általános alakja így írható fel:

$$Y_t = \beta_0 + \beta_1 \cdot Y_{t-1} + \beta_2 \cdot Y_{t-2} + \dots + \beta_k \cdot Y_{t-k} + u_t,$$

ahol Y_t a tárgyidőszaki megbetegedés szám, β_0 a tengelymetszet, $\beta_1, \dots, \beta_k$ az adott késleltetés melletti autokorrelációs együttható, $Y_{t-1}, \dots, Y_{t-k}$ az adott késleltetés melletti megbetegedés szám, és u_t a hibaterm. Esetünkben egy $k = 15$ késleltetést tartalmazó modellt kell megbecsülni,

és a próbafüggvény számításához az így kapott *OLS* modell többszörös determinációs együtt-hatójára (R^2) van szükségünk. A megbecsült modell illeszkedési mutatója ismeretében, a próbafüggvény általános alakja így adható meg:

$$\text{Wald-F} = \frac{R^2}{1 - R^2} \cdot \frac{n - k - 1}{k}$$

A képletben n a minta elemszáma, k a maximális késleltetés rendje, és R^2 a k késleltetést és tengelymetszetet tartalmazó *OLS* modell többszörös determinációs együtt-hatója. A teszt lényegében egy globális Wald-F teszt. A próbafüggvény értéke a $k = 15$ maximális késleltetéses modell esetében:

$$\begin{aligned}\text{Wald-F} &= \frac{0.8076}{1 - 0.8076} \cdot \frac{806 - 15 - 1}{15} \\ \text{Wald-F} &= 221.0686\end{aligned}$$

A kritikus értékek a próba esetében F -eloszlást követnek, $(k; n - k - 1)$ szabadsági fokkal, ahol n a mintaelemszám és k a késleltetések száma. A próbához tartozó empirikus szignifikancia szint $p = 0.000$ lesz, amely mellett a nullhipotézist minden szokásos szignifikancia szinten elvetjük. A következtetés ekkor az, hogy a k megelőző időszak közül, legalább az egyik időszaki megbetegedés szám szignifikáns összefüggésben van a tárgyidőszaki megbetegedés számmal. A következő fehérzaj teszt a Breusch-Godfrey Lagrange-multiplikátor teszt (Breusch, Pagan, 1980) lesz, amelynek a hipotézis rendszere az alábbi:

$$\begin{aligned}H_0 : \forall \rho_i &= 0 \text{ és } i = 1, \dots, k \\ H_1 : \exists \rho_i &\neq 0 \text{ és } i = 1, \dots, k\end{aligned}$$

Ez a teszt is megköveteli, hogy egy újabb csak tengelymetszetet tartalmazó regressziót írjunk fel a tárgyidőszaki változóra:

$$Y_t = \beta_0 + u_t$$

Ebben az esetben Y_t a tárgyidőszaki megbetegedés szám, β_0 a tengelymetszet és u_t a hibatag. Ezután lehetséges az, hogy a hibatagra fel lehet írni egy k rendbeli autoregresszív modellt, amely így néz ki:

$$u_t = \rho_1 \cdot u_{t-1} + \rho_2 \cdot u_{t-2} + \dots + \rho_k \cdot u_{t-k} + \varepsilon_t.$$

Az így kapott modell esetében u_t a tárgyidőszaki hibatag, $u_{t-1}, \dots, u_{t-k}$ a k rendben késleltetett hibatag, $\rho_{t-1}, \dots, \rho_{t-k}$ a megfelelő késleltetés melletti autokorrelációs együtt-ható és ε_t az ehhez a modellhez tartozó hibatag. Ekkor fel lehet írni egy *OLS* eljárással becsült regressziós modellt, a hibatagra ε_t , amely így fog kinézni:

$$\varepsilon_t = \beta_0 + \rho_1 \cdot \hat{u}_{t-1} + \rho_2 \cdot \hat{u}_{t-2} + \dots + \rho_p \cdot \hat{u}_{t-k} + \varepsilon_t$$

Az ehhez a modellhez tartozó többszörös determinációs együtthatóra (R^2) lesz szükségünk a teszthez, amelynek a próbafüggvénye a következő:

$$LM = (n - k - 1) \cdot R^2,$$

ahol R^2 a becsült modellhez tartozó determinációs együttható, n az elemszám - az idősor hossza, és k a bevont késleltetések száma. Az összesített idősorhoz tartozó próbafüggvény abban az esetben, ha $k = 15$ késleltetést választunk:

$$LM = (806 - 15 - 1) \cdot 0.807619$$

$$LM = 641.40811$$

Az empirikus szignifikancia szint $p = 0.000$, ami mellett a nullhipotézist minden szokásos szignifikancia szinten elvetjük. A következtetés ekkor is az, hogy a k megelőző időszak közül, legalább az egyik időszakra vonatkozó megbetegedés szám szignifikáns összefüggésben van a tárgyidőszaki megbetegedések számmal. Ez azt jelenti, hogy az idősor modellezésének van értelme, ugyanis nem valamilyen zaj a folyamat, amely a szórásával és várhatóértékével jellemezhető.

2.1.2. Egységgyök tesztek alkalmazása

Az előző alfejezet alapján adódik, hogy az idősor nem tekinthető fehér zajnak azonban az ARMA típusú modellezés megköveteli azt, hogy az idősor ne tartalmazzon semmilyen eredetű egységgyököt. A 2.1. ábra alapján adódik, hogy az idősorban nem lesz lineáris, exponenciális vagy valamiféle polinomiális trend, azonban meg van az esélye annak, hogy szezonális egységgyököt vagy szezonális eredetű trigonometrikus trendet tartalmaz az idősor. Az egységgyök teszt alkalmazása előtt indokolt a korrelogram vizsgálata az első $k = 15$ késleltetésre. A 15 késleltetés választását az alábbi indokolja: heti mintavételezésű éves szezonalitást tartalmazó idősor esetében $k = \frac{1}{2} \cdot \frac{52}{2} = 13$ késleltetés után cseng le először az autokorreláció függvény nullához (?). A 13. késleltetés után az autokorreláció függvénynek előjelet kell váltania. Egy tetszőleges késleltetés önmagában szignifikánsnak tekinthető, ha az alábbi kritikus értéknél nagyobb az autokorrelációs együttható nagysága:

$$|\rho_i| > \frac{Z_{1-\frac{\alpha}{2}}}{\sqrt{n}},$$

ahol ρ_i a korrelációs együttható, $Z_{1-\frac{\alpha}{2}}$ az α szignifikancia szint melletti normális eloszláshoz tartozó kritikus érték és n a teljes mintaelemszám, azaz az idősor hossza. Esetünkben $\alpha = 1\%$ mellett a kritikus érték az alábbi lesz:

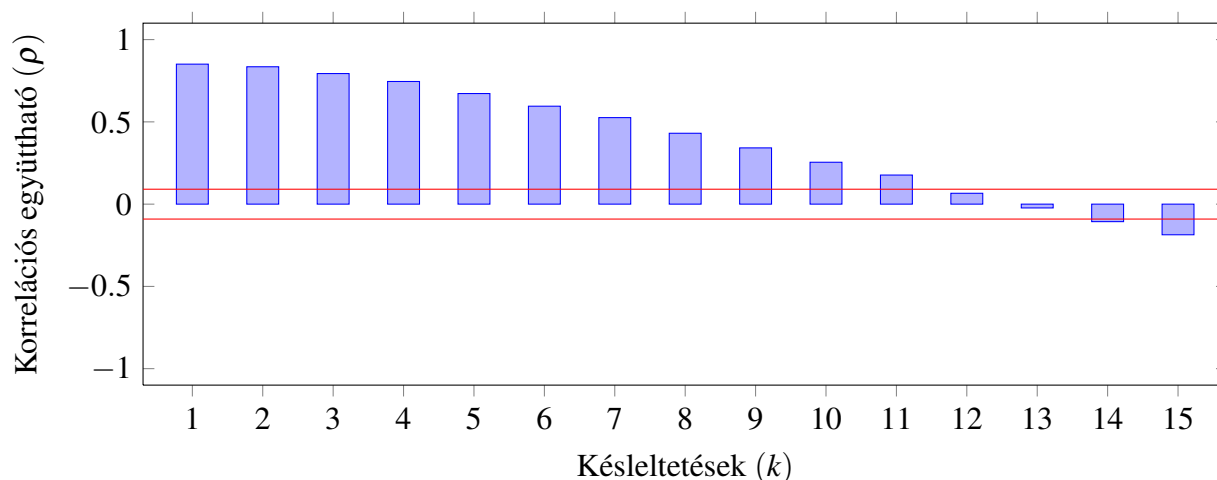
$$C_{1\%} = \frac{2.57583}{\sqrt{806}} = 0.09072$$

A 15 parciális teszt hipotézise így írható fel:

$$H_0 : \rho_i = 0$$

$$H_1 : \rho_i \neq 0$$

Ekkor az első 15 késleltetésre nézve az első 11 és az utolsó 2 parciális teszt esetében a nullhipotézist elvetjük. Azaz várható, hogy egy lassan lecsengő periodogram lesz az, amelyet kapunk. A különböző késleltetésekhez tartozó korrelációs együtthatók értékét a függelék A.1 alfejezete már tartalmazza, ugyanis a Ljung-Box Q-statisztikák becslése megkövetelte a korrelációs együtthatók becslését is már korábban.



2.2. ábra. Az országos idősorhoz tartozó korrelogram

A 2.2 ábrán szereplő korrelogram vizsgálata során tapasztaltak miatt indokolt lehet az egységgyök tesztek alkalmazása. Az erős szezonális jelenléte sejteti, hogy az egyszerű egységgyök tesztek nem lesznek alkalmasak arra, hogy a stacionaritást vizsgáljuk. Első lépésként egy Augmented Dickey-Fuller-féle teszt (Dickey, Fuller, 1976)¹ lesz az, amelyet végrehajtottunk. A teszt szezonális egységgyök kimutatására heti frekvenciájú idősorok esetében lényegében alkalmatlan, ennek ellenére a teljes körű elemzés megköveteli az alkalmazását.

Ebben az esetben az alábbi változatokat fogjuk használni:

1. Konstanssal bővített változat, 20 késleltetés
2. Konstanssal és trenddel bővített változat, 20 késleltetés
3. Konstanssal és kvadratikus trenddel bővített változat, 20 késleltetés

¹Innentől kezdve *ADF*.

Mind a három esetben így lehet formalizálni a nullhipotézist és a hozzá tartozó ellenhipotézist:

$$H_0 : \phi = 1 \quad H_1 : \phi \neq 1,$$

ahol ϕ az 1. késleltetéshez tartozó autoregressziós együttható. A nullhipotézisünk tehát az, hogy egységgyök van a folyamatban. A három teszt esetében az empirikus szignifikancia szintek a 2.1 táblázatban szerepelnek.

Teszt típusa	ADF tesztek
	<i>p</i> -értékek
Konstans	$6.181 \cdot 10^{-14}$
Konstans és trend	$9.865 \cdot 10^{-14}$
Konstans és kvadratikusan trend	$6.598 \cdot 10^{-26}$

2.1. táblázat. Az ADF tesztek eredményei

Azaz adódik, hogy bármely szokásos szignifikancia szint mellett elvethető az, hogy nem szezonális eredetű egységgyök van az összesített idősort jellemző folyamatban. A részletes output a függelék A.2 alfejezetben megtalálható. A következő lépés a Dickey-Hasza-Fuller-féle (Dickey et al., 1984) szezonális egységgyök teszt alkalmazása lesz az idősorra.² Ez a próbafüggvény nincsen beépítve a *Gretl* programcsomagba, azonban szabadon elérhető és integrálható az egységgyök tesztek közé. A programhoz tartozó bemeneti argumentumot vizsgáló kódrészletben engedélyeztem azt, hogy heti lebontású adatsorokon is lehessen egységgyök tesztet végezni, ugyanis az eredeti csomag csak havi és negyedéves adatok vizsgálatára képes. A standard $k = 5$ késleltetést tartalmazó *DHF* teszt mellett a szezonális *dummy* változókkal kibővített $k = 5$ késleltetést tartalmazó változatot is végrehajtjuk. Ebben az esetben a két tesztre az alábbi módon néz ki a nullhipotézis és a hozzá tartozó ellenhipotézis a késleltetésekre:

$$H_0 : \Phi = 1 \quad H_1 : \Phi \neq 1,$$

ahol Φ a szezonálisnak megfelelő autokorrelációs együttható. A nullhipotézisünk tehát az, hogy szezonális egységgyök van a folyamatban. A tesztek értékeit és a kritikus értékeket a 2.2 táblázat tartalmazza. A *DHF* próba szigorúan baloldali kritikus értékkel rendelkezik, ekkor a 2.2 táblázatban összefoglalt eredmények alapján mind a két esetben elutasítjuk a nullhipotézist, azaz szezonális egységgyök nincsen jelen a folyamatban. A következő lépés az idősorban jelenlévő szezonális identifikálása és eltávolítása lesz.

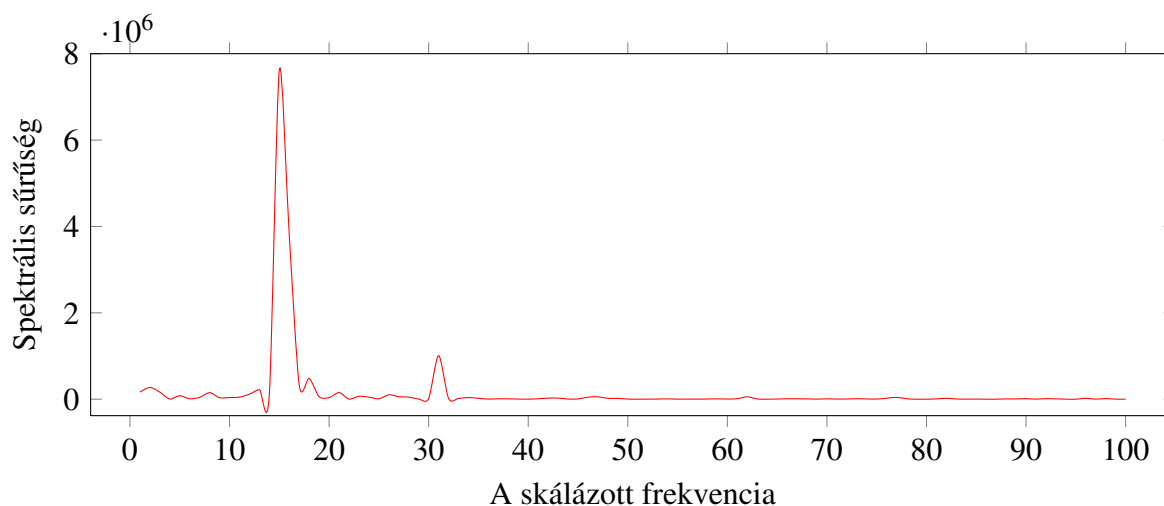
²Innentől kezdve *DHF*.

Teszt típusa	DHF tesztek	
	Teszt értéke	Kritikus érték
Konstans	-6.337	-3.400
Konstans és dummyk	-6.279	-4.150

2.2. táblázat. A DHF tesztek eredményei

2.1.3. Periodicitás vizsgálata

A szezonális identifikálására alkalmas a periodogram (Canova, 1996; Dickey, Akdi, 1999), amely segítségével meg tudjuk határozni, hogy az idősorban milyen frekvenciák mellett vannak erős ciklikus hatások, amelyektől az idősort meg kell tisztítani. A 2.1 ábra alapján már van egy erős sejtésünk, hogy éves szezonális van jelen az idősorban.



2.3. ábra. Az országos idősorhoz tartozó periodogram

A 2.3 ábrán szereplő periodogramon látszik, hogy a 15 egységnyi skálázott frekvenciánál globális maximum van, és később a 31 egységnyi skálázott frekvenciánál található még egy kevésbé kiemelkedő csúcs, amely lokális maximum, ez alapján és a többi csúcs hiánya alapján adódik, hogy 2 trigonometrikus görbével leírható a folyamat, és nem lesznek további harmonikusok. A periodogram által az első 31 skálázott frekvenciára kapott érték megtalálható a függelék A.3 alfejezetében. A későbbi modellezésnél kérdés lesz majd az, hogy a második görbe szerepeltetése indokolt-e a különböző determinisztikus modellek esetében. A csúcsokhoz tartozó skálázott frekvencia értékek (f_1) és (f_2), illetve a megfigyelések száma (n) alapján meghatározható a keresett szinuszgörbék periódusa (T_1 , T_2), az alábbi képlettel (Newton, 1994):

$$T_1 = \frac{n}{f_1} = \frac{806}{15} = 53.73 \text{ hét} \quad (\text{Az első görbe esetében})$$

$$T_2 = \frac{n}{f_2} = \frac{806}{31} = 26 \text{ hét} \quad (\text{A második görbe esetében})$$

Azaz a kapott eredmények konzisztensek azzal, amelyet a hasonló nemzetközi kutatások során kaptak (Chan, Tian, 2011), hiszen a periódusok lényegében éves, féléves hosszúságúak. Az eredmények alapján be lehetne vezetni két indexváltozót, amelyek január első hetével indulnak és periodikusan ismétlődnek, azonban az első periódus nem pontosan a második kétszerese. Emiatt pusztán a hetek számát lefedő t indexváltozót fogjuk bevezetni. Az így bevezetett index változó modellbe való beépítése két frekvenciaparaméterrel megoldja a periódusok közötti átszámíthatóság problémáját, hiszen a bevezetett frekvenciaparaméterek és trend elemek kezelik az első és második frekvenciához tartozó időszaki folyamatot. Korábban már említettem azt, hogy az $52 + 26$ *dummy* változó bevezetése a szabadságfok nagy mértékű csökkenésével járna. Emiatt lehet indokolt valamilyen eltolt trigonometrikus függvény alkalmazása, ez az *outlier* megbetegedés számokat is kezelné, amely a *dummy* változók értékét kitérítethetné. Az ilyen eltérő periódusokhoz eltérő frekvencia és amplitúdó paramétereket alkalmazó trigonometrikus függvényekre specifikált regressziós modellek alkalmazása a biostatistikában ráadásul elterjedt (Galbraith, 2005).

2.1.4. Indikátor változós modell illesztése

Az eltolt trigonometrikus függvény bevezetése előtt, érdemes megpróbálkozni azzal, hogy mi történik a havi lebontású *dummy* változók bevezetése esetében. A modell alapján kapott grafikus ábra is szemléltetheti már azt, hogy a trigonometrikus függvény bevezetése indokolt. Az így bevezetett havi lebontású logikai változók dekódolása az alábbi:

D_1 , Értéke 1, ha a hónap január.

$\vdots$

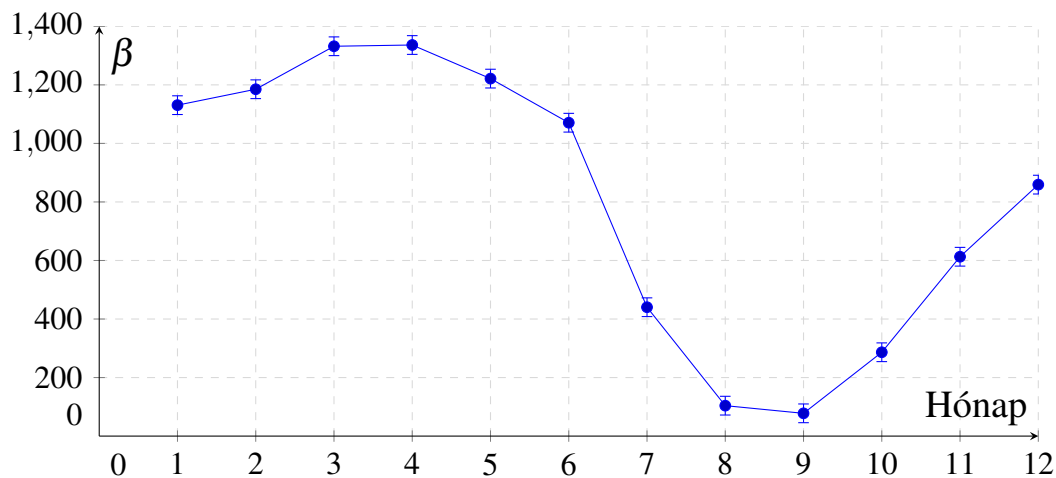
D_{12} , Értéke 1, ha a hónap december.

Ekkor az, alábbi *OLS* módszerrel is becsülhető regressziós modellt készíthetjük el:

$$Y_t = \sum_{i=1}^{12} \beta_i \cdot D_i + u_t,$$

ahol Y_t az adott heti összesített bárányhimlős megbetegedés szám, D_i az adott hónapra korábban bevezetett *dummy* változó, β_i a megfelelő *dummy* változóhoz tartozó paraméter és u_t a hibatag. A becsült modell esetében, az alábbi ábrán a kék pontok jelzik a becsült havi paraméterek

nagyságát, és a kék szintvonalak az $\alpha = 99\%$ megbízhatósági szinthez tartozó konfidencia intervallumokat. Az így kapott konfidencia intervallumok középpontjait – a paraméterek értékeit – a könnyedebb szemléltetés érdekében összekötöttem.



2.4. ábra. Havi lebontású dummy változók alkalmazása

Az így kapott modell pontos jellemzői, a paraméterek értékei, a szignifikancia szintek és egyéb jellemzők a függelék A.4 alfejezetében megtalálhatóak. Az összes paraméter szignifikáns lett, ebből adódik, hogy a havi átlagos megbetegedés számok szignifikánsan eltérnek egymástól. Azt is kimondhatjuk, hogy átlagosan a legtöbb bejelentett megbetegedés a tavaszi hónapokban történik, míg a legkevesebb augusztusban és szeptemberben. Kifejezetten érdekes a június és július között történő drasztikus visszaesés a megbetegedések átlagos számában. A kapott eredményeink alapján is indokolt lehet a trigonometrikus függvényforma választása, hiszen a 12 ponttal történő közelítés is szinuszos görbét ad.

A korábban kapott éves ciklust leíró időindex változót fogjuk beépíteni a modellbe, a modellezést elemi trigonometrikus függvénnyel fogjuk elvégezni. Azonban azt, hogy az idősor minimuma és maximuma pontosan melyik hetekben lesz nem tudjuk előre, ezért eltolás paramétereket vezetünk be. A periodogram alapján adódik, hogy tetszőlegesen választhatunk, hogy mind a két periódust beépítsük-e a modellbe. Első lépésként csak az egyiket szerepeltetjük a modellben, ezzel egy építkező modellszelekció lehetőségét adjuk meg.

2.1.5. Trigonometrikus trend beépítése

A periodogram, a korrelogram és a 2.4 ábra alapján az alábbi determinisztikus idősor modell felépítése lehet indokolt:

$$Y_t = \alpha + \underbrace{\beta_1 \cdot \sin\left(\frac{t}{\Psi_1} - \gamma_1\right)}_{\text{Az 1. periódus beépítése}} + u_t \quad (\text{NLS I. modell})$$

A felírt modellben Y_t az adott időszakai összesített bárányhimlős megbetegedés szám, α a tengelymetszet paraméter, t az indexváltozó, Ψ_1 az első periódus alapján létrehozott görbéhez tartozó frekvencia paraméter, β_1 az első periódus alapján létrehozott görbe amplitúdó paramétere, γ_1 a hozzá tartozó eltolási paraméter és u_t a hibatag.

A modell becslését csak nemlineáris legkisebb négyzetek módszerével (NLS) lehet végrehajtani (Lawson, Hanson, 1974), ugyanis paramétereiben nemlineáris, és nem linearizálható algebrai-lag. Az így becsült modellek esetében adott egy Θ paraméter halmaz, és egy adott nemlineáris Y függvényt próbálunk megbecsülni a paraméterek függvényében, egy adott időindex mellett – $Y = f(t; \Theta)$ – úgy, hogy a négyzetes eltérések összege a legkisebb legyen. A most létrehozott modell esetében a paraméter halmaz így írható le: $\Theta = \{\alpha, \Psi_1, \gamma_1, \beta_1\}$. A célunk tehát az, hogy a korábban választott függvényforma mellett, amely t és Θ értékeitől függ az alábbi optimalizációs problémát oldjuk meg (Cameron, Trivedi, 2005):

$$\hat{\Theta} = \arg \min_{\Theta} \left[\sum_{i=1}^n (Y_t - f(t; \Theta))^2 \right] \quad (2.1)$$

A becslési eljárás megköveteli azt, hogy manuálisan hozzunk létre a (2.1) problémából egy grádiens vektort a becslendő paraméterek szerinti parciális deriváltakból. Azaz esetünkben az összes paraméter szerinti parciális deriváltat elő kell állítani, és a grádiens alapján az NLS optimum feltétele rendezés után így adódik általánosan:

$$\frac{\partial \left[\sum_{i=1}^n (Y_t - f(t; \Theta))^2 \right]}{\partial \Theta} = 2 \cdot \left[\sum_{i=1}^n (Y_t - f(t; \Theta)) \right] \cdot \frac{\partial f(t; \Theta)}{\partial \Theta} = 0$$

Az általunk választott függvényforma mellett ezek az alábbiak lesznek:

$$\begin{aligned} 2 \cdot \left[\sum_{i=1}^n \left(Y_t - \left(\alpha + \beta_1 \cdot \sin\left(\frac{t}{\Psi_1} - \gamma_1\right) \right) \right) \right] &= 0 \\ 2 \cdot \left[\sum_{i=1}^n \left(Y_t - \left(\alpha + \beta_1 \cdot \sin\left(\frac{t}{\Psi_1} - \gamma_1\right) \right) \right) \right] \cdot \sin\left(\frac{t}{\Psi_1} - \gamma_1\right) &= 0 \end{aligned}$$

$$\begin{aligned}
& -2 \cdot \left[\sum_{i=1}^n \left(Y_t - \left(\alpha + \beta_1 \cdot \sin \left(\frac{t}{\Psi_1} - \gamma_1 \right) \right) \right) \right] \cdot \beta_1 \cdot \cos \left(\frac{t}{\Psi_1} - \gamma_1 \right) = 0 \\
& -2 \cdot \left[\sum_{i=1}^n \left(Y_t - \left(\alpha + \beta_1 \cdot \sin \left(\frac{t}{\Psi_1} - \gamma_1 \right) \right) \right) \right] \cdot \beta_1 \cdot \cos \left(\frac{t}{\Psi_1} - \gamma_1 \right) \cdot \frac{t}{\Psi_1^2} = 0
\end{aligned} \tag{2.2}$$

Az (2.2) egyenletrendszer megoldása analitikusan nem lehetséges, a megoldáshoz a *Gretl* NLS csomagja azonban felhasználható. Az iterációs folyamat megkezdéséhez egy tetszőleges x_0 kezdővektorra is szükségünk van, ez a paraméterek kezdőértékét tartalmazza. Ebben az esetben az $x_0 = [800, 800, -10, 10]$ indulóvektort választjuk, hiszen a nagyobb periódussal rendelkező görbét keressük. A *Gretl* programcsomag számára megadott induló vektor deklarációs és optimalizációs kódrészlet megtalálható a függelék A.5 alfejezetében. A numerikus Jacobi-mátrix ekkor előállítható, és a konvergencia folyamat sikeresen lefut 33 iterációs lépés alatt. Az adott *NLS* módszerrel kapott regresszió által magyarázott négyzetösszeg változását figyeljük. Egy *for* ciklusba beágyazva az *NLS* elven iteráló algoritmust, a modell stabilitása megvizsgálható azzal, hogy visszahelyettesítjük a kapott paramétereket induló paraméterként. A tolerancia szint, amelyenél az iteráció leáll az alábbi:

$$\text{Tolerancia} = 1.81899 \cdot 10^{-12}$$

A becsült modell jellemzői tehát az alábbiak:

	Paraméter	Std. hiba	t-próba	p-érték
α	793.554	10.143	78.229	0.000
β_1	646.937	14.962	43.237	0.000
γ_1	-6.114	0.042	-144.813	0.000
ψ_1	7.698	0.079	97.029	0.000
Függő változó átlaga	801.3449	Függő változó std. σ	528.1449	
SSR	65855883	Regresszió std. hiba	286.5564	
R^2	0.706713	Korrigált R^2	0.705616	
Log-likelihood	-5701.955	Akaike kritérium	11411.91	
Schwarz kritérium	11430.68	Hannan–Quinn	11419.12	
$\hat{\rho}$	0.506561	Durbin–Watson	0.986600	

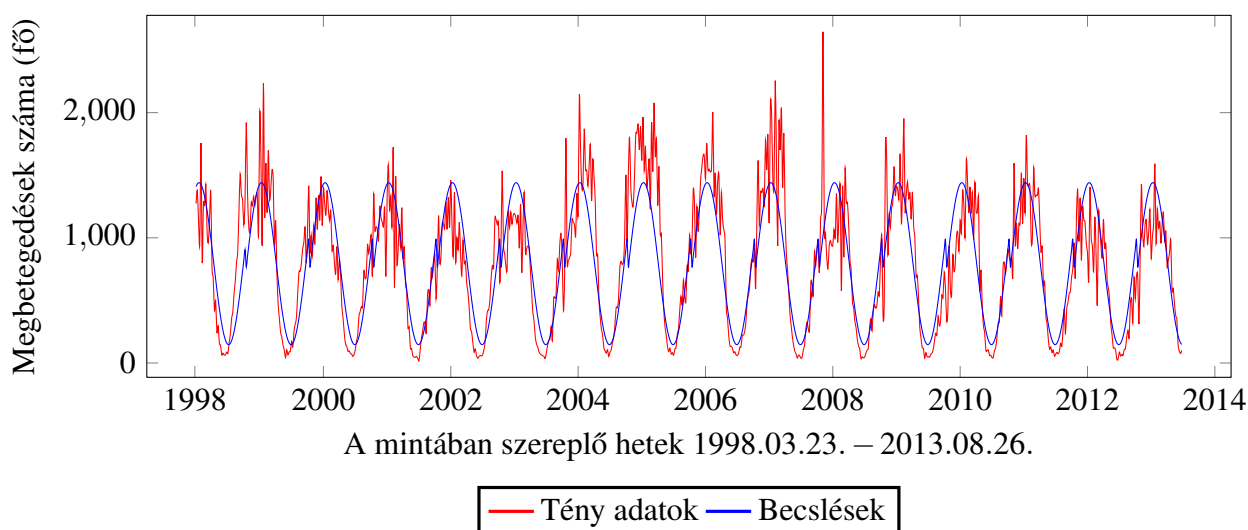
2.3. táblázat. Az I. NLS modell jellemzői

A modell jellemzőit összefoglaló 2.3 táblázat alapján az alábbiakat tudjuk: a parciális tesztek alapján elmondható, hogy a paraméterek bármilyen szokásos szignifikancia szinten szignifikánsak. Mivel a paraméterek nem változókhoz köthetőek, emiatt multikollinearitás sem lehet

annak az oka, hogy mindegyik ilyen mértékben szignifikáns. A bevezetett α tengelymetszet paraméter közelíti a megbetegedések több éves heti átlagát, de nem pontosan azt adja ki. A bevezetett egyéb paramétereknek ebben az esetben a beépített elemi szögfüggvény miatt nincsen önálló epidemiológiai tartalmú értelmezése. Az illeszkedést jellemző R^2 , és a modell szelekcióra alkalmas korrigált R^2 alkalmazása az idősoros esetekben nem előnyös, a később bevonásra kerülő autoregresszív tagok miatt csak a determinisztikus modellek esetében lesz releváns. A modellek összehasonlítását, szelekcióját a likelihood alapú információs kritériumok alapján fogjuk elvégezni. A magas $\hat{\rho}$ érték pozitív elsőrendű reziduális autokorrelációt sejtet, a reziduális autokorrelációt az alábbiak okozhatják (Hajdú, 2013):

1. Hibás függvényforma választása
2. Alacsony minta elemszám
3. Osztott késleltetésű folyamat
4. Kihagyott releváns magyarázóváltozók

A modell esetében a 2. és 3. nagy valószínűséggel nem okozhatja a reziduális autokorrelációt, hiszen a mintaelemszám ($n = 806$) a becsült paraméterek számához ($p = 4$) viszonyítva magas. Az osztott késleltetésű folyamat alkalmazását pedig semmilyen közvetlenül nem megfigyelhető változó jelenléte nem indokolja. A hibás függvényforma választását kizárhatjuk, hiszen a 2.5 ábra alapján is adódik, hogy a trend illeszkedése nagyon erős.



2.5. ábra. Az I. NLS modell becslései

Az elsőrendű reziduális autokorrelációt vizsgáló Durbin-Watson-féle teszthez tartozó hipotézisek:

$$H_0 : \rho_1 \neq 0 \quad H_1 : \rho_1 = 0,$$

ahol ρ_1 az elsőrendű reziduális autokorrelációs együttható. A megfelelő kritikus értékek $\alpha = 1\%$ -os szignifikancia szinten ebben az esetben, $p = 3$ becsült paraméter (tengelymetszet nélkül) és $n = 806$ mintaelemszám mellett:

$$D_L(806, 3)_{1\%} = 1.8763 \quad D_U(806, 3)_{1\%} = 1.8914$$

Az $\alpha = 1\%$ szignifikancia szinthez kritikus értékeket nem lehet a *Gretl* segítségével előállítani, ezért a *MATLAB* az, amelyet felhasználhatunk a kritikus értékek számításához. A döntésünk az, hogy $\alpha = 1\%$ mellett elfogadjuk a pozitív elsőrendű reziduális autokorreláció jelenlétét a modellben. A későbbiekben tehát mozgóátlagolású és autoregresszív késleltetések bevonása lehet indokolt. Mivel a két periódus felhasználásával történő modellezés összes kombinációját elő akarjuk állítani, ezért az egy időindexre felépített modellezés lehetősége a másik periódus esetében is adott lesz, ekkor az alábbi determinisztikus modellt lehet felépíteni:

$$Y_t = \alpha + \underbrace{\beta_2 \cdot \sin\left(\frac{t}{\Psi_2} - \gamma_2\right)}_{\text{A 2. periódus beépítése}} + u_t \quad (\text{A II. NLS modell})$$

A kapott modellben Y_t az adott időszak összesített bárányhimlős megbetegedés szám, α a tengelymetszet paraméter, t az indexváltozó, Ψ_2 a második periódus alapján létrehozott görbéhez tartozó frekvencia paraméter, β_2 a második periódus alapján létrehozott görbe amplitúdó paramétere, γ_2 a hozzá tartozó eltolási paraméter és u_t a hibatag. A becslési eljáráshoz tartozó optimalitási feltételek ebben az esetben:

$$\begin{aligned} 2 \cdot \left[\sum_{i=1}^n \left(Y_i - \left(\alpha + \beta_2 \cdot \sin\left(\frac{t}{\Psi_2} - \gamma_2\right) \right) \right) \right] &= 0 \\ 2 \cdot \left[\sum_{i=1}^n \left(Y_i - \left(\alpha + \beta_2 \cdot \sin\left(\frac{t}{\Psi_2} - \gamma_2\right) \right) \right) \right] \cdot \sin\left(\frac{t}{\Psi_2} - \gamma_2\right) &= 0 \\ -2 \cdot \left[\sum_{i=1}^n \left(Y_i - \left(\alpha + \beta_2 \cdot \sin\left(\frac{t}{\Psi_2} - \gamma_2\right) \right) \right) \right] \cdot \beta_2 \cdot \cos\left(\frac{t}{\Psi_2} - \gamma_2\right) &= 0 \quad (2.3) \\ -2 \cdot \left[\sum_{i=1}^n \left(Y_i - \left(\alpha + \beta_2 \cdot \sin\left(\frac{t}{\Psi_2} - \gamma_2\right) \right) \right) \right] \cdot \beta_2 \cdot \cos\left(\frac{t}{\Psi_2} - \gamma_2\right) \cdot \frac{t}{\Psi_2^2} &= 0 \end{aligned}$$

Ekkor az *NLS* becsléshez tartozó iterációs folyamat újra végrehajtható, ennek első lépése a paramétereket tartalmazó vektor deklarációja, kezdővektornak most az $x_0 = [800, -200, 15, -5]$

vektort választjuk, hiszen nem az előző optimalizáció során megkapott paramétereket keressük. A becslési eljárás ekkor megkezdődhet; felhasználva a paraméterekhez tartozó parciális deriváltakat és optimum feltételeket, amelyeket a (2.3) egyenletrendszer definiál. A változók deklarációját és a becslést végrehajtó *Gretl* kód megtalálható a függelék A.6 alfejezetében.

A numerikus Jacobi-mátrix ekkor előállítható, és a konvergencia folyamat sikeresen lefut 22 iterációs lépés alatt. A tolerancia szint, amelynél az iteráció leáll az alábbi:

$$\text{Tolerancia} = 1.81899 \cdot 10^{-12}$$

A becslült modell jellemzőit tartalmazza a 2.4 táblázat.

	Paraméter	Std. hiba	t -próba	p -érték	
	α	808.093	20.524	39.372	0.000
	β_2	−173.388	25.418	−6.821	0.000
	γ_2	15.347	0.308	49.879	0.000
	ψ_2	−4.468	0.471	−9.489	0.000
Függő változó átlaga		801.3449	Függő változó std. σ		528.1449
SSR		$2.11934 \cdot 10^8$	Regresszió std. hiba		514.0587
R^2		0.056161	Korrigált R^2		0.052631
Log-likelihood		-6172.984	Akaike kritérium		12353.97
Schwarz kritérium		12372.74	Hannan–Quinn		12361.17
$\hat{\rho}$		0.844587	Durbin–Watson		0.310049

2.4. táblázat. A II. NLS modell jellemzői

A modellben szereplő paraméterek, bármilyen standard szignifikancia szint mellett szignifikánsnak tekinthetők. Ennek ellenére a modell illeszkedése nagyon gyenge, messze elmarad az I. *NLS* modell által leírt görbe illeszkedésétől. Mindez a korrigált R^2 és az *SSR* érték alapján is látszik. Az azonos paraméter szám miatt itt az R^2 értékek összehasonlítása is megfelelne. A *likelihood* alapú információs kritériumok és a regresszió standard hibája is sokkal gyengébb illeszkedést mutat. Az első rendű reziduális korreláció mértéke is nagyobb, hiszen a becslült korrelációs együttható is nagyobb, a kritikus értékek nem változtak, hiszen azonos a becslült paraméterek száma, és az eredményváltozó sem változott, azaz ez a modell is legalább elsőrendű reziduális autokorrelációtól szenved.

A két periódus együttes alkalmazásával egy harmadik modell is előállítható, amely mind a két periódust tartalmazza, ennek az egyenlete így írható le:

$$Y_i = \underbrace{\alpha + \beta_1 \cdot \sin\left(\frac{t}{\Psi_1} - \gamma_1\right)}_{\text{Az 1. periódus beépítése}} + \underbrace{\beta_2 \cdot \sin\left(\frac{t}{\Psi_2} - \gamma_2\right)}_{\text{A 2. periódus beépítése}} + u_t \quad (\text{A III. NLS modell})$$

A kapott modellben Y_i az adott időszaki összesített bárányhimlős megbetegedés szám, α a tengelymetszet paraméter, t az indexváltozó, Ψ_1 az első görbéhez tartozó frekvencia paraméter, β_1 az első görbe amplitúdó paramétere, γ_1 a hozzá tartozó eltolási paraméter, Ψ_2 a második görbéhez tartozó frekvencia paraméter, β_2 a második görbe amplitúdó paramétere, γ_2 a hozzá tartozó eltolási paraméter és u_t a hibatag.

A modell becsléséhez az alábbi optimalitási feltételek tartoznak; a tengelymetszet paraméterre:

$$2 \cdot \left[\sum_{i=1}^n \left(Y_i - \left(\alpha + \beta_1 \cdot \sin\left(\frac{t}{\Psi_1} - \gamma_1\right) + \beta_2 \cdot \sin\left(\frac{t}{\Psi_2} - \gamma_2\right) \right) \right) \right] = 0$$

Az amplitúdó paraméterekre:

$$\begin{aligned} 2 \cdot \left[\sum_{i=1}^n \left(Y_i - \left(\alpha + \beta_1 \cdot \sin\left(\frac{t}{\Psi_1} - \gamma_1\right) + \beta_2 \cdot \sin\left(\frac{t}{\Psi_2} - \gamma_2\right) \right) \right) \right] \cdot \sin\left(\frac{t}{\Psi_1} - \gamma_1\right) &= 0 \\ 2 \cdot \left[\sum_{i=1}^n \left(Y_i - \left(\alpha + \beta_1 \cdot \sin\left(\frac{t}{\Psi_1} - \gamma_1\right) + \beta_2 \cdot \sin\left(\frac{t}{\Psi_2} - \gamma_2\right) \right) \right) \right] \cdot \sin\left(\frac{t}{\Psi_2} - \gamma_2\right) &= 0 \end{aligned}$$

Az eltolás paraméterekre:

$$\begin{aligned} -2 \cdot \left[\sum_{i=1}^n \left(Y_i - \left(\alpha + \beta_1 \cdot \sin\left(\frac{t}{\Psi_1} - \gamma_1\right) + \beta_2 \cdot \sin\left(\frac{t}{\Psi_2} - \gamma_2\right) \right) \right) \right] \cdot \beta_1 \cdot \cos\left(\frac{t}{\Psi_1} - \gamma_1\right) &= 0 \\ -2 \cdot \left[\sum_{i=1}^n \left(Y_i - \left(\alpha + \beta_1 \cdot \sin\left(\frac{t}{\Psi_1} - \gamma_1\right) + \beta_2 \cdot \sin\left(\frac{t}{\Psi_2} - \gamma_2\right) \right) \right) \right] \cdot \beta_2 \cdot \cos\left(\frac{t}{\Psi_2} - \gamma_2\right) &= 0 \end{aligned}$$

A frekvenciát kontrolláló paraméterekre:

$$\begin{aligned} -2 \cdot \left[\sum_{i=1}^n \left(Y_i - \left(\alpha + \beta_1 \cdot \sin\left(\frac{t}{\Psi_1} - \gamma_1\right) + \beta_2 \cdot \sin\left(\frac{t}{\Psi_2} - \gamma_2\right) \right) \right) \right] \cdot \beta_1 \cdot \cos\left(\frac{t}{\Psi_1} - \gamma_1\right) \cdot \frac{t}{\Psi_1^2} &= 0 \\ -2 \cdot \left[\sum_{i=1}^n \left(Y_i - \left(\alpha + \beta_1 \cdot \sin\left(\frac{t}{\Psi_1} - \gamma_1\right) + \beta_2 \cdot \sin\left(\frac{t}{\Psi_2} - \gamma_2\right) \right) \right) \right] \cdot \beta_2 \cdot \cos\left(\frac{t}{\Psi_2} - \gamma_2\right) \cdot \frac{t}{\Psi_2^2} &= 0 \end{aligned}$$

Az egyenlet NLS módszerrel történő becslése megkívánja a kezdő paramétereket tartalmazó vektor deklarálását. Ezt most az alábbiak választjuk meg $x_0 = [10, 10, 10, 10, 10, 10, 10]$; ekkor a korábban felírt parciális deriváltak alapján megadható maga a becslési eljárás is, amely segítségével az egyenlet megbecsülhető. A becslési eljáráshoz tartozó *Gretl* kód megtalálható a függelék A.7 alfejezetében. A becsült modell jellemzőit a 2.5 táblázat tartalmazza.

	Paraméter	Std. hiba	<i>t</i> -próba	<i>p</i> -érték	
	α	821.479	13.019	63.099	0.000
	β_1	−604.755	13.890	−43.540	0.000
	γ_1	34.279	0.080	429.091	0.000
	ψ_1	8.771	0.211	41.551	0.000
	β_2	−201.421	19.100	−10.546	0.000
	γ_2	12.666	0.186	68.283	0.000
	ψ_2	4.879	0.319	15.296	0.000
Függő változó átlaga	801.3449	Függő változó std. σ	528.1449		
SSR	57150099	Regresszió std. hiba	267.4454		
R^2	0.7454841	Korrigált R^2	0.743573		
Log-likelihood	-5644.815	Akaike kritérium	11303.63		
Schwarz kritérium	12372.74	Hannan–Quinn	11316.24		
$\hat{\rho}$	0.443867	Durbin–Watson	1.112259		

2.5. táblázat. A III. NLS modell jellemzői

A modelltől elmondható, hogy bármely a gyakorlatban használt szignifikancia szinten szignifikánsak a bevezetett paraméterek. A korrigált R^2 és az információs kritériumok alapján (Akaike, Hannan–Quinn és Schwarz–Bayes) a III. NLS modellt preferáljuk az I. NLS modellel szemben. Mindez azt jelenti, hogy a periodogramon megjelenő második lokális maximum nem az egy-éves trigonometrikus trend egy felharmonikusa volt, hanem egy önálló fél éves periódusokkal rendelkező trigonometrikus trend.

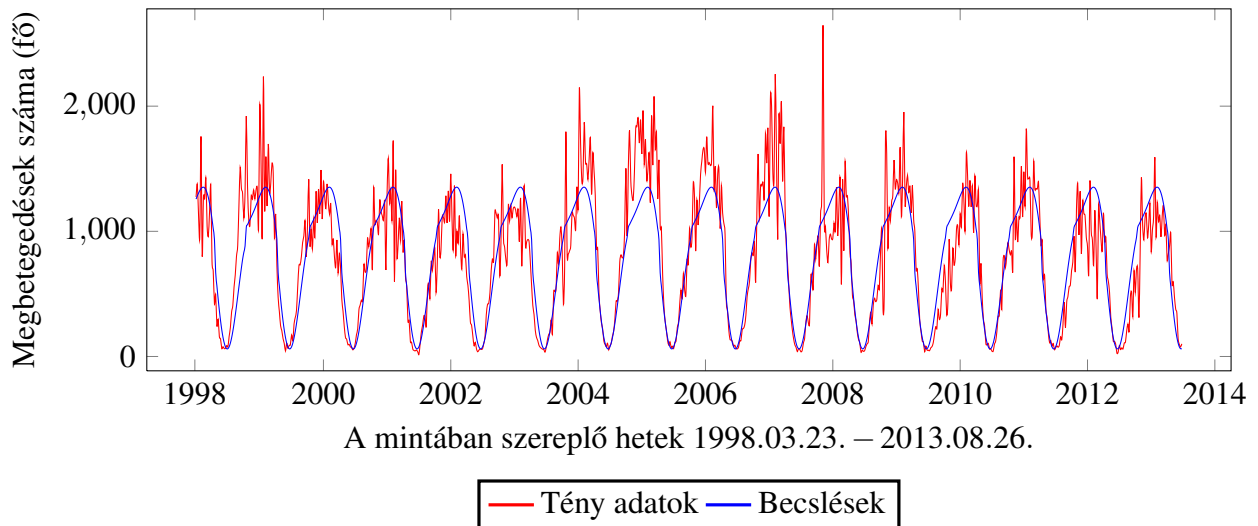
A becsült értékek egyike sem negatív, azaz epidemiológiai szempontból értelmes a modell, amit felépítettünk, hiszen az adott héten bejelentett megbetegedések száma nem lehet negatív semmiképpen. A III. NLS modell választását regressziós játékok (Pintér, 2006) alkalmazásával is meg lehet indokolni. Tegyük fel, hogy az első játékos az alábbi paraméterekből, és változókból álló halmaz $\{t, \beta_1, \gamma_1, \Psi_1\}$. A második játékos pedig az alábbi halmaz: $\{t, \beta_2, \gamma_2, \Psi_2\}$. Ekkor a kifizetés vektorok a többszörös determinációs együttható alapján:

$$v(\{\emptyset\}) = 0 \quad v(\{1\}) \approx 0.70671 \quad v(\{2\}) \approx 0.05616 \quad v(\{1, 2\}) \approx 0.74548$$

Ekkor a Shapley-értékek így adódnak a játékra:

$$Sh_1 = 0.69801 \quad Sh_2 = 0.04746$$

Azaz a játékosok értékelései az egy periódust felhasználó görbék illeszkedési mutatójától kevéssel térnek el, a függő változó varianciájának kis hányadát magyarázza mind a két modell. Mindez arra utal, hogy volt értelme a második görbe beépítésének ezen módszer szerint is, hiszen erős saját magyarázóerővel rendelkezik a két periódust tartalmazó III. *NLS* modellben is (Pintér, 2011). A becslési eljárással kapott determinisztikus trend és a megbetegedések száma egy közös grafikonon szerepel a 2.6 ábrán.



2.6. ábra. A III. *NLS* modell becslései

Az elsőrendű reziduális autokorrelációt tesztelő Durbin-Watson statisztikához tartozó kritikus értékek $\alpha = 1\%$ esetén a szignifikancia szinten ebben az esetben, $p = 6$ becsült paraméter (tengelymetszet nélkül) és $n = 806$ mintaelemszám mellett:

$$D_L(806, 6)_{1\%} = 1.8763 \quad D_U(806, 6)_{1\%} = 1.8914$$

Azaz szignifikáns pozitív elsőrendű reziduális autokorreláció van jelen a modellben.

2.1.6. ARMA tagok identifikálása

A Durbin-Watson teszt alapján már tudjuk, hogy az u_t hibatagban pozitív elsőrendű reziduális autokorreláció van jelen. A III. *NLS* modell maradékot modellezve egy olyan additív modellt hozhatunk létre, amely segítségével reziduális autokorreláció nélkül becsülhetjük a heti megbetegedés számot. A korábban alkalmazott fehérzaj tesztek, azaz a Wald-F teszt, a Lagrange-multiplikátor teszt és a Ljung-Box Q-teszt alkalmas a magasabb rendű autokorreláció jelenlétének tesztelésére is. Ekkor az u_t maradék tagra $k = 15$ késleltetésig fel lehet írni első lépésben

egy Ljung-Box Q-tesztet, a reziduális autokorreláció tesztelésére.

A próbafüggvény értékeit, az autokorrelációs függvény, a parciális autokorrelációs függvény értékeit és az empirikus szignifikancia értékeket a függelék A.8 alfejezetében csatoltam. A nullhipotézist az empirikus szignifikancia értékek alapján, a $k = 1, \dots, 15$ késleltetések mellett elvetjük, azaz legalább az egyik késleltetés szignifikáns lesz az adott maximális késleltetésig. A maradék taghoz tartozó parciális autokorreláció függvény 4 késleltetés után letörik, míg az autokorreláció függvény idővel lecseng.

A Wald-F próba esetében, a segédregressziós R^2 ismeretében adódik, hogy mi a próba függvény értéke:

$$\text{Wald-F} = \frac{R^2}{1 - R^2} \cdot \frac{n - k - 1}{k} = \frac{0.3605}{1 - 0.3605} \cdot \frac{806 - 15 - 1}{15} = 28.953932$$

A $p = 0.000$ empirikus szignifikancia szint alapján a nullhipotézist elvetjük. A következtetés ekkor az, hogy a k megelőző időszak közül, legalább az egyik késleltetés szignifikáns a maradék tagban. A következő reziduális autokorreláció teszt a Breusch–Godfrey Lagrange-multiplikátor teszt lesz, ehhez a teszthez is a segédregressziós többszörös determinációs együtthatóra lesz szükségünk, a próbafüggvény a következő:

$$\text{LM} = (n - k - 1) \cdot R^2 = (806 - 15 - 1) \cdot 0.3605 = 284.078267$$

A nullhipotézist a $p = 0.000$ empirikus szignifikancia szint alapján elvetjük. A következtetés ekkor az, hogy a k megelőző időszak közül, legalább az egyik időszakra vonatkozó késleltetett reziduális tag szignifikáns.

A következőkben tehát a maradék tagot fogjuk modellezni autoregresszív és mozgóátlagolású tagokkal, a parciális autokorreláció függvény letörése miatt a legmagasabb rendű késleltetés tartalmazó modell, amelynek illesztését vizsgálni fogjuk egy $ARMA(5,5)$ modell lesz. Ez biztosítja azt, hogy az esetleges magasabb késleltetések hatását is vizsgáljuk, annak ellenére, hogy a korábban említett letörés legfeljebb 4 késleltetés szerepeltetését indokolná. Az ilyen modellezést akkor tehetjük meg, ha a maradéktag nem tartalmaz egységgyököt, az ezzel kapcsolatos hipotézist már korábban elvetettük. A tisztán autoregresszív tagokat tartalmazó modelleket 5 késleltetésig fogjuk, vizsgálni és a tisztán mozgóátlagolású folyamatokat is legfeljebb 5 késleltetésig. Az autoregresszív és mozgóátlagolású tagokat tartalmazó modellek közül csak a szimmetrikusakkal foglalkozunk, amelyek esetében az autoregresszív és mozgóátlagolású tagok késleltetéseinek rendje megegyezik. Ennek a megfontolásnak az az alapja, hogy minden bevont késleltetés esetében tudjon a modell a hibákból tanulni. A becült modelleket jellemző

információs kritériumokat tartalmazó táblázat megtalálható a melléklet A.9 részfejezetében. Az összes modellt *Exact Maximum Likelihood* eljárással becsültük (*EML*).

Mivel az információs kritériumok közül a Schwarz-információs kritérium a legszigorúbb a szelekciós eljárások során, ezért ez alapján döntünk. Ennek a kritériumnak a minimuma ugyanazon modell esetében van, mint a Hannan-Quinn-információs kritérium minimuma – az $ARMA(2, 2)$ modell esetében. Mindez rendkívül érdekes ugyanis a kapott eredmény epidemiológiai szempontból is értelmes, ugyanis a legerősebben preferált modell esetében a tárgyidőszaki megbetegedés számra még hatást gyakorolnak azok a megbetegedések, amelyek két héttel korábban történnek. A bárányhimlő lappangási ideje is két és három hét – pontosabban 14-21 nap között van a legtöbb esetben (Mueller et al., 2008). Azaz a maradék modellezése során az alábbi modell felépítése lehet indokolt:

$$u_t = \alpha + \underbrace{\phi_1 \cdot u_{t-1} + \phi_2 \cdot u_{t-2}}_{\text{Autoregresszív tagok}} + \underbrace{\theta_1 \cdot \varepsilon_{t-1} + \theta_2 \cdot \varepsilon_{t-2}}_{\text{Mozgóátlagolású tagok}} + \varepsilon_t, \quad \varepsilon_t \sim WN(0, \sigma),$$

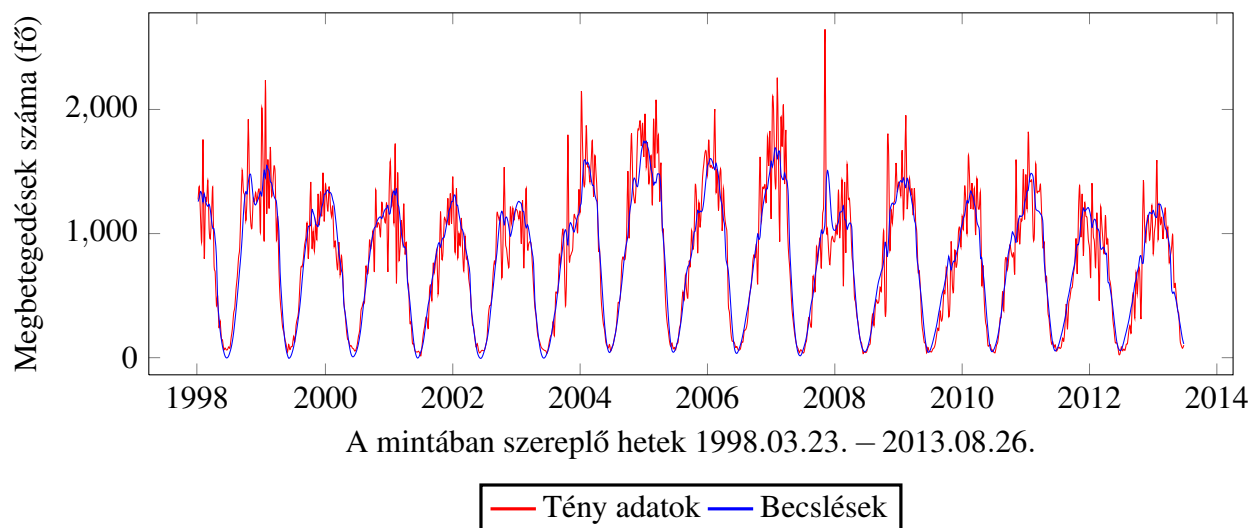
ahol u_t a III. *NLS* modell tévedése az adott időszakban, α a tengelymetszet, u_{t-1} a III. *NLS* modell egy időszakkal korábbi hibája, ϕ_1 a hozzá tartozó paraméter, u_{t-2} a III. *NLS* modell két időszakkal korábbi hibája és ϕ_2 a hozzá tartozó paraméter. A modellben ε_{t-1} az $ARMA(2, 2)$ modell 1 időszakkal korábbi hibáját, θ_1 a hozzá tartozó paramétert, ε_{t-2} az $ARMA(2, 2)$ modell 2 időszakkal korábbi hibáját, θ_2 a hozzá tartozó paramétert és ε_t az adott időszaki hiba, amit az $ARMA(2, 2)$ modell elkövet. A maradék tagról ebben az esetben azt feltételezzük, hogy nulla várhatóértékű fehérzaj és szórása pedig konstans σ .

Az így létrejövő modell esetében azonban már elvárás lesz az, hogy a maradéktagnak ne legyen reziduális autokorreláció semmilyen formában. Azaz a maradéktag fehérzaj mivoltát fogjuk tesztelni, a korábban felhasznált Ljung-Box Q-statisztikával első lépésben. Ekkor az u_t maradék tagra $k = 15$ késleltetésig fel lehet írni első lépésben egy Ljung-Box Q-tesztet, a reziduális autokorreláció tesztelésére. A próbafüggvény értékeit és az empirikus szignifikancia értékeit a mellékletekben csatoltam, a nullhipotézist az empirikus szignifikancia értékek alapján a $k = 1, \dots, 15$ késleltetések mellett elfogadjuk, azaz az egyik késleltetés sem lesz szignifikáns az adott maximális késleltetésig. A modell tehát mentes a reziduális autokorrelációtól $k = 15$ késleltetésig. Az autokorrelációfüggvény és a parciális autokorrelációfüggvény is ezt támasztja alá ugyanezen késleltetés szám mellett. A Q-statisztikákat tartalmazó *output* mellett a függelék A.10 alfejezetében megtalálható az autokorreláció függvény, a parciális autokorreláció függvény és az empirikus szignifikancia szintek is 15 késleltetésig. A becsült modellünk pontos jellemzőit a 2.6 táblázat foglalja össze.

	Paraméter	Std. hiba	z-próba	p-érték	
	α	-0.0080	31.1300	-0.0019	0.9985
	ϕ_1	1.3725	0.2134	6.4314	0.0000
	ϕ_2	-0.4307	0.1995	-2.1593	0.0308
	θ_1	-1.2010	0.2031	-5.9140	0.0000
	θ_2	0.4426	0.1367	3.2374	0.0012
Függő változó átlaga	$-2.84 \cdot 10^{-10}$		Függő változó std. σ	266.4468	
Innovációk átlaga	0.030740		Innovációk std. σ	215.7514	
Log-likelihood	-5475.575		Akaike kritérium	10963.15	
Schwarz kritérium	10991.30		Hannan-Quinn	10973.96	

2.6. táblázat. A maradékra illesztett ARMA(2,2) modell

Az így elkészített modell alapján meg tudjuk határozni a becsült megbetegedés értékeket a tanuló mintára vonatkoztatva abban az esetben, ha az ARMA(2,2) modell és a III. NLS modell becsléseit összeadjuk. A tanuló mintán túli adatokra vonatkozó becslések az előrejelzésekkel foglalkozó 3. fejezetben vannak. Az ex post előrejelzett és tény adatok a tanuló minta esetében a 2.7 ábrán szerepelnek.



2.7. ábra. A III. NLS és ARMA(2,2) modell becslései a tény adatokkal szemben

A tanulómintára történt becslések alapján látszódik, hogy a kombinált determinisztikus és sztochasztikus modell a megbetegedések szórásának változékonyságát nem tudja kezelni.

2.2. Sztochasztikus modell

A korábban keverten determinisztikus és sztochasztikus elemeket tartalmazó idősorral ellentétben most egy teljesen sztochasztikus idősori modellt vezetünk be. Az elemzési eljárás, amelyet felhasználunk a Box-Jenkins-féle módszer lesz, azaz lépéseink ebben az esetben az alábbiak:

1. Fehérzaj vizsgálat az idősorra
2. Szezonális és nem szezonális egységgyök teszt alkalmazása
3. Szükséges *SARMA* modell identifikálása
4. Fehérzaj vizsgálat a maradékra nézve

Az első 2 lépést, már korábban végrehajtottuk az előző modell felépítésekor. Mindezek miatt úgy vesszük, hogy az idősor nem fehérzaj és nem valamilyen egységgyök folyamat. A szezonális autoregresszív és mozgóátlagolású tagok bevezetésekor figyelni kell arra, hogy a tanulóminta megrövidül jelentős mértékben. Ez a probléma azért fontos mert a *likelihood* értékek nagyon erősen függenek a mintaelemszámtól, és emiatt mindez a *likelihood* alapú szelekciós kritériumok összehasonlíthatóságát is jelentősen befolyásolni fogja. Számunkra ez azt jelenti, hogy a modelleket csak regressziós standard hibák vagy a regressziós innovációk hibái alapján hasonlíthatjuk össze.

2.2.1. A SARMA tagok identifikálása

A következőkben tehát a megbetegedés számot fogjuk modellezni szezonális és nem szezonális autoregresszív mozgóátlagolású tagokkal. Az első vizsgált modell típus az alábbi általános alakban írható fel:

$$Y_t = \alpha + \underbrace{\sum_{i=1}^n \phi_i \cdot Y_{t-i} + \sum_{i=1}^n \theta_i \cdot u_{t-i}}_{\text{Nem szezonális tagok}} + \underbrace{\Phi_1 \cdot Y_{t-52} + \Theta_1 \cdot u_{t-52}}_{\text{Szezonális tagok}} + u_t$$

$$n = 1, \dots, 5 \quad u_t \sim WN(0, \sigma)$$

E modell esetében Y_t az adott heti megbetegedés szám, Y_{t-i} az i héttel korábbi megbetegedés szám, és ϕ_i a hozzá tartozó autoregresszív együttható. Az u_{t-i} a modell i héttel korábbi hibája, és θ_i a hozzá tartozó paraméter. Az Y_{t-52} az egy évvel korábbi megbetegedés számot jelöli, és Φ_{t-52} a hozzá tartozó paraméter, u_{t-52} a modell egy évvel korábbi hibáját jelöli, és Θ_{t-52} a hozzá tartozó paraméter. A maradék tagról, amelyet u_t jelöl feltételezzük, hogy 0 várhatóértékű,

σ szórású fehérzaj hiba. Az interakciókat, amelyek kialakulnak a nem szezonális és szezonális autoregresszív tagok között, nem szerepeltettem az egyenletben, és a következőkben szereplő sztochasztikus modellek esetében sem fogom. A másik vizsgált modell típus így jellemezhető:

$$Y_t = \alpha + \sum_{i=1}^n \phi_i \cdot Y_{t-i} + \sum_{i=1}^n \theta_i \cdot u_{t-i} + \sum_{j=1}^k \Phi_1 \cdot Y_{t-j.52} + \sum_{j=1}^k \Theta_1 \cdot u_{t-j.52} + u_t$$

$$n = 1, \dots, 5, \quad k = 1, 2, \quad u_t \sim WN(0, \sigma)$$

Ez a modell a két évvel korábbi megbetegedés számot, és a modell két évvel korábbi tévedéseit is figyelembe veszi. A legmagasabb rendű nem szezonális és szezonális késleltetést is tartalmazó modell, amelynek illeszkedését vizsgálni fogjuk egy $SARMA(5,5)(2,2)_{52}$ modell lesz – a maximálisan 5 nem szezonális késleltetés választását a parsimonius elv, és a 3 hetes maximális lappangási idő indokolja. A tisztán autoregresszív tagokat tartalmazó modelleket 5 nem szezonális késleltetésig fogjuk, vizsgálni és a tisztán mozgóátlagolású folyamatokat is legfeljebb 5 nem szezonális késleltetésig. Az autoregresszív és mozgóátlagolású tagokat tartalmazó modellek közül csak a szimmetrikusakkal foglalkozunk, amelyek esetében az autoregresszív és mozgóátlagolású tagok késleltetéseinek rendje megegyezik. A becsült modelleket jellemző információs kritériumok és standard hibák megtalálhatóak a függelék A.11 és A.12 alfejezetében. Azonban még egyszer érdemes hangsúlyozni, hogy a jelentősen megrövidülő minta miatt az információs kritériumok nem összehasonlíthatóak, emiatt minden modellenél jelöltem, hogy hány hétnyi rövidülés történt. Az összes vizsgált modell *EML* eljárással becslési eljárással állt elő. A kapott modellnek az alábbi feltételeket kell teljesítenie:

1. A maradék fehérzaj.
2. A parsimonius elvnek megfelel a modell.
3. A bevont változók szignifikánsak.
4. Erős illeszkedés, ha az 1., 2. és 3. pont teljesül.

A két szezonális késleltetést tartalmazó modellek esetében sérül a 2. és 3. szempont, emellett a magas paraméterszámot tartalmazó modellek esetében olyan szignifikáns magas rendű autoregresszív késleltetések jelennek meg a korrelogramon, amelyek jelenléte a folyamatban nem lehet indokolt. Az egyetlen szezonális késleltetést tartalmazó modellek esetében a maradék fehérzaj transzformációjának feltétele sérül, ez alacsony rendű szignifikáns reziduális autokorreláció jelenlétében nyilvánul meg. Emiatt a csak sztochasztikus késleltetéseket tartalmazó modellek esetében a differenciázás nélküli modelleket el kell vetni annak ellenére, hogy egységgyök nem volt a folyamatban.

2.2.2. A SARIMA tagok identifikálása

A *SARMA* modellezés során kiderült, hogy a differenciázatlan időseriesi modellezés nem megfelelő. A differenciázott modellek esetében az adott heti megbetegedés szám változást modellezzük, két típusú differenciázásra van lehetőségünk hetire és szezonálisra. A szezonális differenciázás esetében figyelni kell arra, hogy újabb nagy mértékű minta rövidüléssel jár. Abban az esetben, hogyha egyszer szezonálisan differenciázunk és egy szezonális késleltetés is szerepel a rövidülés már két évnek felel meg, ez 104 mintaelemet jelent. A differenciázás során csak szimmetrikus modellekkel fogunk foglalkozni, a két éves szezonális késleltetések alkalmazását már korábban elvetettük, és csak egy évvel korábbi szezonális késleltetéseket alkalmazunk. A nem szezonális autoregresszív és mozgóátlagos késleltetések esetében a legmagasabb késleltetés amelyet vizsgálunk legfeljebb ötöd rendű. Az ilyen modellek jellemzői megtalálhatóak a függelék A.13 alfejezetében. A megbecsült modellek közül a *SARIMA*(3, 1, 3)(1, 0, 1) és *SARIMA*(4, 1, 4)(1, 0, 1) alkalmazása merül fel. Ezek esetében az összes bevont késleltetés szignifikáns, és a maradék fehérzajnak tekinthető. A maradéktagok vizsgálatához tartozó Q-statisztikák, autokorrelációs, parciális autokorrelációs értékek és a korrelogramok a függelék A.14 és A.15 alfejezetében megtalálhatóak. A parsimonius szemlélet miatt, a szűkebb *SARIMA*(3, 1, 3)(1, 0, 1) modellt választjuk, az ehhez tartozó regressziós egyenlet a differenciákra felírva:

$$\Delta Y_t = \alpha + \sum_{i=1}^3 \phi_i \cdot \Delta Y_{t-i} + \sum_{i=1}^3 \theta_i \cdot u_{t-i} + \Phi_1 \cdot \Delta Y_{t-52} + \Theta_1 \cdot u_{t-52} + u_t$$

A szinteken megfigyelt megbetegedés számra vonatkozó egyenlet ekkor:

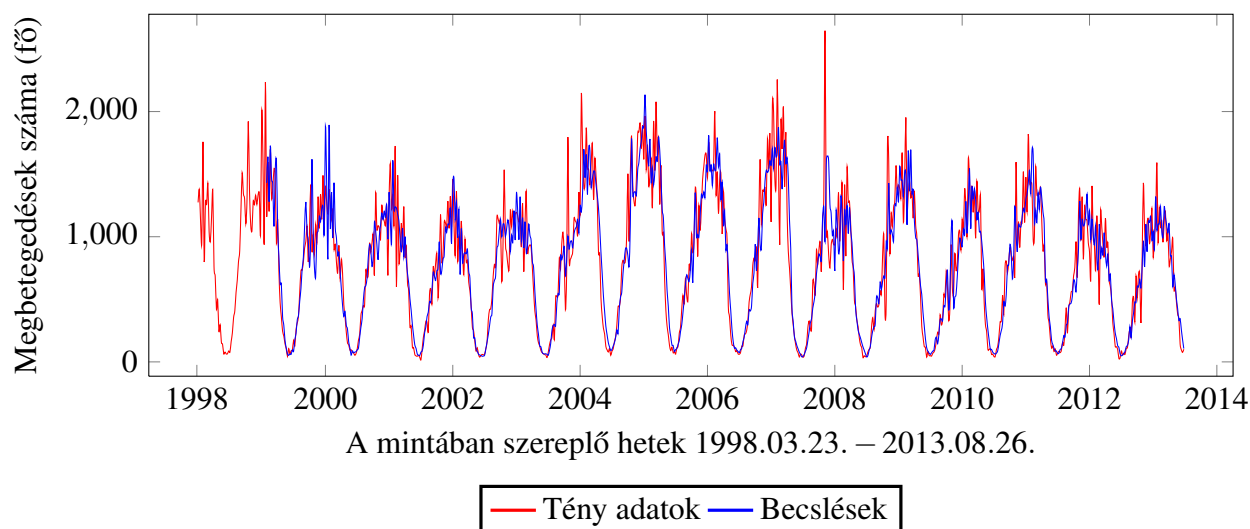
$$Y_t = Y_{t-1} + \alpha + \sum_{i=1}^3 \phi_i \cdot \Delta Y_{t-i} + \sum_{i=1}^3 \theta_i \cdot u_{t-i} + \Phi_1 \cdot \Delta Y_{t-52} + \Theta_1 \cdot u_{t-52} + u_t$$

Ezen modell pontos jellemzőit tartalmazza a 2.7 táblázat. A kapott eredményeink epidemiológiai szempontból is értelmesek, hiszen eszerint a nem szezonális késleltetések közül legfeljebb a 3 héttel korábbi megbetegedés számnak és hibáknak van hatása a jelenlegi megbetegedés szám változására, a bevont szezonális tagok pedig kezelik a megjelenő erős szezonális hatást. Ez szintén alátámasztja azt a megfogalmazott hipotézist, amely azt állítja, hogy a lappangási idő és a bevont autoregresszív késleltetések által lefedett időtartam között kapcsolat van, hiszen a bárányhimlő lappangási ideje 14 és 21 nap között van az esetek többségében. A regressziós paraméterek a standard szignifikancia szinteken mind szignifikánsak a tengelymetszetet leszámítva, a tengelymetszet benntartása azonban mindig indokolt, hiszen a nem megfigyelt hatásokat tartalmazza.

	Paraméter	Std. hiba	z -próba	p -érték	
	α	-1.9193	5.8146	-0.3301	0.7413
	ϕ_1	-1.2482	0.1722	-7.2479	0.0000
	ϕ_2	-0.5813	0.1996	-2.9117	0.0036
	ϕ_3	-0.1608	0.05692	-2.8247	0.0047
	Φ_1	0.8095	0.0259	31.1647	0.0000
	θ_1	0.5080	0.1753	2.8978	0.0038
	θ_2	-0.2261	0.0895	-2.5255	0.0116
	θ_3	-0.0661	0.03129	-2.1124	0.0173
	Θ_1	-0.4328	0.04298	-10.0665	0.0000
Függő változó átlaga	-2.846667	Függő változó std. σ	280.5876		
Innovációk átlaga	0.879488	Innovációk σ	213.1674		
Log-likelihood	-5085.762	Akaike kritérium	10191.52		
Schwarz kritérium	10237.73	Hannan-Quinn	10209.33		

2.7. táblázat. A SARIMA(3,1,3)(1,1,1) modell jellemzői

A modell becsléseinek és előrejelzéseink jellemzését, az ezzel foglalkozó 3. fejezetben helyeztem el. A 2.8 ábrán a differenciákból visszatranszformált becsült megbetegedés számok és a megfigyelt megbetegedés számok szerepelnek, szigorúan a tanuló mintára.



2.8. ábra. A SARIMA(3,1,3)(1,0,1) modell becslései a tény adatokkal szemben

2.3. A Poisson-regressziós megközelítés

Az adatsorunk csak nem negatív egészekből állhat természetéből adódóan, ebből az is adódik, hogy a függő változónkra olyan modellt építeni, amely negatív értékeket is felvehet nem előnyös. Adott, hogy a $[0, \infty]$ intervallumon vehet fel egész értékeket a megbetegedés szám; az ilyen modellezésre alkalmas lehet a Poisson-regressziós megközelítés, ahol a megbetegedések várhatóértékét fogjuk becsülni. Az ilyen modellezés alkalmazása a megbetegedések számának előrejelzésére ráadásul meglehetősen elterjedt az epidemiológiában (Parodi, Bottarelli, 2006; Bender, 2009). A korábban tapasztaltak alapján adódik, hogy a megbetegedés szám becsléséhez használt függvényforma nem feltétlenül lesz lineáris, emellett autoregresszív tagok bevezetése is indokolt lehet. A leggyakrabban használt idősorok vizsgálatára alkalmas statisztikai programcsomagok azonban beépítve nem tartalmaznak ilyen becselőfüggvényeket (*Eviews*, *Gretl*). Mindezek miatt be kell vezetni a nemlineáris függvényekre specifikált Poisson-regressziót.

2.3.1. Nemlineáris Poisson-regresszió bevezetése

A következőkben a Poisson-regresszió egy kiterjesztett, nem csak paramétereiben lineáris függvényekre alkalmazott (Greene, 2003) változatát fogjuk definiálni. Emellett egy megfelelő becslési eljárást is be fogunk mutatni, amely segítségével adott adathalmazból a becselőfüggvény tetszőleges függvényformára és paraméter számra előállítható.

Adott egy tetszőleges $h \in \mathbb{N}$ számú paraméterből álló Θ vektor és egy a magyarázó változóink értékeit tömörítő x vektor, amely tetszőleges k rendű késleltetéseket és egyéb, esetlegesen exogén változókat tartalmaz, összesen l darabot. A Θ_i paraméter jelöli a paramétereket tartalmazó vektor i -edik elemét. Azt feltételezzük, hogy a h darab paraméter és az l magyarázó változó vektor valamilyen nem feltétlenül lineárisan előállítható összetételét jelöli Ω . Az összetételben szerepelhet logaritmikus, exponenciális vagy akár trigonometrikus transzformáció is. Azaz tulajdonképpen az alábbi általános alakban megadott függvényt hozzuk létre:

$$\Omega = f(\Theta; x) \quad (2.4)$$

Az így előállított kompozícióval szemben az az elvárásunk lesz, hogy bármely paraméter szerint legalább egyszer differenciálható legyen. Ez a feltételezés a későbbi optimalizációs eljáráshoz szükséges. Az eredmény változó az adott t -edik heti megbetegedés szám (Y_t) várhatóértéke. Azt feltételezve, hogy a heti megbetegedéseknek Poisson-eloszlása van az alábbi módon definiálhatjuk a paraméterét:

$$\lambda = E(Y \mid x) \quad (2.5)$$

Ez természetesen megfogalmazható így is, mivel a természetes logaritmus a kapocsfüggvény (Madsen, Thyregod, 2011), amit alkalmazunk:

$$\lambda = e^\Omega = e^{f(\Theta; x)} \quad (2.6)$$

Ezzel az állítással megadtuk a Poisson-eloszlásunk λ paraméterét. Tudjuk, hogy ilyenkor a Poisson-eloszlás (Pintér, 2012) feltételes valószínűségi sűrűfüggvénye:

$$P(Y = y | \Theta; x) = \frac{\lambda^y \cdot e^{-\lambda}}{y!} \quad (2.7)$$

Ez tulajdonképpen felírható így is, a (2.5) és (2.7) egyenletek alapján:

$$P(Y = y | \Theta; x) = \frac{[E(Y | x)]^y \cdot e^{-E(Y | x)}}{y!} \quad (2.8)$$

Ez pedig tovább is felírható a (2.6) és (2.8) egyenletek alapján:

$$P(Y = y | \Theta; x) = \frac{e^{y \cdot \Omega} \cdot e^{-e^\Omega}}{y!} \quad (2.9)$$

Feltételezzük a továbbiakban azt, hogy n darab x vektorunk és y skalárunk van. (Ez gyakorlatban azt jelenti, hogy $(l + 1) \times n$ adatunk van). Ezeknek az együttes feltételes valószínűségi sűrűfüggvénye így adódik, rögzített Θ paraméter halmaz mellett, felhasználva a (2.9) egyenlet produktumát:

$$P(y_1, \dots, y_n | \Theta; x_1, \dots, x_n) = \prod_{i=1}^n \frac{e^{y_i \cdot \Omega_i} \cdot e^{-e^{\Omega_i}}}{y_i!} \quad (2.10)$$

Ekkor azt a Θ paraméter halmazt keressük, amely mellett a $\mathcal{L}$ jelölésű likelihood függvényünk értéke a legnagyobb, ez felírható így a (2.10) egyenletből, hiszen az ugyanez mint a (2.10) egyenlet csak másképpen írjuk fel a feltételes valószínűségeket:

$$\mathcal{L}(\Theta | y_1, \dots, y_n; x_1, \dots, x_n) = \prod_{i=1}^n \frac{e^{y_i \cdot \Omega_i} \cdot e^{-e^{\Omega_i}}}{y_i!} \quad (2.11)$$

Azaz tulajdonképpen az alábbi célfüggvényünk van:

$$\max \quad \mathcal{L}(\Theta | y_1, \dots, y_n; x_1, \dots, x_n) = \max \quad \prod_{i=1}^m \frac{e^{y_i \cdot \Omega_i} \cdot e^{-e^{\Omega_i}}}{y_i!}$$

A *likelihood* függvény logaritmusa, ugyanott veszi fel a maximumát, mint az eredeti függvény mert a logaritmus monoton transzformáció, ezért az úgynevezett *log – likelihood* függvényt alkalmazzuk:

$$\log \mathcal{L}(\Theta | y_1, \dots, y_n; x_1, \dots, x_n) = \ell(\Theta | y_1, \dots, y_n; x_1, \dots, x_n)$$

Mivel a logaritmus szeparábilisan additív, az alábbi függvényünk adódik a (2.11) egyenlet alapján:

$$\ell(\Theta|y_1, \dots, y_n; x_1, \dots, x_n) = \sum_{i=1}^n \left(y_i \cdot \Omega_i - e^{\Omega_i} - \log(y_i!) \right)$$

A Θ paramétereket tartalmazó vektor, az utolsó kifejezésben nem jelenik meg. Az ott szereplő kifejezés, az adatok rögzített mivolta miatt egy konstans. Emiatt mivel a kivonás önmagában csak egy monoton transzformáció, a kifejezés a függvényből elhagyható. Gyakorlatban az alábbi célfüggvényünk van:

$$\ell(\Theta|y_1, \dots, y_n; x_1, \dots, x_n) = \sum_{i=1}^n \left(y_i \cdot \Omega_i - e^{\Omega_i} \right)$$

Ez természetesen visszaírható, a korábban felírt kombinatív definíciónk (2.4) alapján:

$$\ell(\Theta|y_1, \dots, y_n; x_1, \dots, x_n) = \sum_{i=1}^n \left(y_i \cdot f(\Theta; x_i) - e^{f(\Theta; x_i)} \right)$$

A célfüggvényünk tehát az alábbi lesz:

$$\max \quad \ell(\Theta|y_1, \dots, y_n; x_1, \dots, x_n) = \max \quad \sum_{i=1}^n \left(y_i \cdot f(\Theta; x_i) - e^{f(\Theta; x_i)} \right) \quad (2.12)$$

A megoldást akkor tudjuk megkapni, ha teljesül az alábbi a paraméterek szerinti parciális deriváltakra (grádiens vektorra):

$$\begin{aligned} 0 &= \frac{\partial \ell(\Theta|y_1, \dots, y_n; x_1, \dots, x_n)}{\partial \Theta_1} \\ &\vdots \\ 0 &= \frac{\partial \ell(\Theta|y_1, \dots, y_n; x_1, \dots, x_n)}{\partial \Theta_h} \end{aligned} \quad (2.13)$$

A problémának, amelyet a (2.13) egyenletrendszer jelent nincsen egyértelmű zárt alakban megadható megoldása, még akkor sem, ha az $f(\Omega; x)$ függvény előre megadott (problémát okozhat az implicit függvények jelenléte és a magas paraméter szám). Egy nemlineáris programozási probléma felírása és egy iteratív optimalizálási megközelítés, azonban megoldást jelenthet a kérdésre.

2.3.2. Determinisztikus modell specifikáció

Azzal tisztában vagyunk, hogy az idősorunk nem valamilyen fehérzaj folyamat. Emellett nem tartalmaz sem szezonális, sem egyéb eredetű egységgyököt. Az idősor mindezek miatt determinisztikus és sztochasztikus tagokkal modellezhető. A korábban a felírt modell alapján tetszőleges függvényformát választhatunk, amelyet beágyazunk a Poisson-regresszióba. Első lépésben egy olyan determinisztikus trendet fogunk beépíteni a modellbe, amely felhasználja a

2.3 részfejezetben már bevezetett periodikusan ismétlődő index változót. A trigonometrikus függvények alkalmazását ugyanazzal lehet indokolni, mint a 2.1.4 részfejezetben – az éves és féléves periódusok meglétével. Emiatt az alábbi nemlineáris Poisson-regressziós modellt fogjuk felépíteni:

$$\lambda_t = E(Y_t | \alpha, \beta_1, \beta_2, \Psi_1, \Psi_2, \gamma_1, \gamma_2, t)$$

$$\lambda_t = e^{\alpha + \beta_1 \cdot \sin\left(\frac{t}{\Psi_1} - \gamma_1\right) + \beta_2 \cdot \sin\left(\frac{t}{\Psi_2} - \gamma_2\right)},$$

ahol Y_t az adott időszaki összesített bárányhimlős megbetegedés szám, α a tengelymetszet paraméter, t az indexváltozó, Ψ_1 az első görbéhez tartozó frekvencia paraméter, β_1 az első görbe amplitúdó paramétere, γ_1 a hozzá tartozó eltolási paraméter, Ψ_2 a második görbéhez tartozó frekvencia paraméter, β_2 a második görbe amplitúdó paramétere, γ_2 a hozzá tartozó eltolási paraméter. Ez tulajdonképpen az alábbi definíciót jelenti:

$$\Omega = f(\Theta; x) = \alpha + \beta_1 \cdot \sin\left(\frac{t}{\Psi_1} - \gamma_1\right) + \beta_2 \cdot \sin\left(\frac{t}{\Psi_2} - \gamma_2\right)$$

Az x vektor sorozat csak a t változókból áll, a paramétereket tartalmazó Θ vektor pedig az $\alpha, \beta_1, \gamma_1, \Psi_1, \beta_2, \gamma_2, \Psi_2$ paraméter halmazból. A maximalizálandó *log – likelihood* függvényünk ebben az esetben tehát:

$$\max \quad \ell(\alpha, \beta_1, \gamma_1, \Psi_1, \beta_2, \gamma_2, \Psi_2 | Y; t)$$

Ezen definíció és a 2.12 egyenlet alapján adódik:

$$\max \quad \sum_{i=1}^n Y_i \cdot \left(\alpha + \beta_1 \cdot \sin\left(\frac{t}{\Psi_1} - \gamma_1\right) + \beta_2 \cdot \sin\left(\frac{t}{\Psi_2} - \gamma_2\right) \right) -$$

$$- \sum_{i=1}^n e^{\alpha + \beta_1 \cdot \sin\left(\frac{t}{\Psi_1} - \gamma_1\right) + \beta_2 \cdot \sin\left(\frac{t}{\Psi_2} - \gamma_2\right)}$$

A megoldási algoritmus a BFGS (Broyden-Fletcher-Goldfarb-Shanno) grádiens eltéréseken alapuló optimalizáció lesz (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, Kettler, 1970), hiszen nincsen semmilyen lineáris vagy nemlineáris korlátunk. A korábban feltételezett első rendben történő differenciálhatósági feltétel ezért volt tehát fontos. Ehhez az optimalizáció előtt a 7 paraméter szerint a parciális deriváltakat manuálisan kell előállítani. A tengelymetszet és amplitúdó paraméterekre:

$$\frac{\partial \ell}{\partial \alpha} = \sum_{i=1}^n \left(Y_i - e^{\alpha + \beta_1 \cdot \sin\left(\frac{t}{\Psi_1} - \gamma_1\right) + \beta_2 \cdot \sin\left(\frac{t}{\Psi_2} - \gamma_2\right)} \right)$$

$$\frac{\partial \ell}{\partial \beta_1} = \sum_{i=1}^n \left(Y_i \cdot \sin\left(\frac{t}{\Psi_1} - \gamma_1\right) - \left(e^{\alpha + \beta_1 \cdot \sin\left(\frac{t}{\Psi_1} - \gamma_1\right) + \beta_2 \cdot \sin\left(\frac{t}{\Psi_2} - \gamma_2\right)} \right) \cdot \sin\left(\frac{t}{\Psi_1} - \gamma_1\right) \right)$$

$$\frac{\partial \ell}{\partial \beta_2} = \sum_{i=1}^n \left(Y_i \cdot \sin\left(\frac{t}{\Psi_2} - \gamma_2\right) - \left(e^{\alpha + \beta_1 \cdot \sin\left(\frac{t}{\Psi_1} - \gamma_1\right) + \beta_2 \cdot \sin\left(\frac{t}{\Psi_2} - \gamma_2\right)} \right) \cdot \sin\left(\frac{t}{\Psi_2} - \gamma_2\right) \right)$$

A periodicitást kontrolláló paraméterekre:

$$\begin{aligned}\frac{\partial \ell}{\partial \Psi_1} &= \sum_{i=1}^n \left(-Y_i \cdot \beta_1 \cdot \cos \left(\frac{t}{\Psi_1} - \gamma_1 \right) \cdot \frac{t}{\Psi_1^2} \right) + \\ &\quad + \sum_{i=1}^n \left(e^{\alpha + \beta_1 \cdot \sin \left(\frac{t}{\Psi_1} - \gamma_1 \right) + \beta_2 \cdot \sin \left(\frac{t}{\Psi_2} - \gamma_2 \right)} \cdot \beta_1 \cdot \cos \left(\frac{t}{\Psi_1} - \gamma_1 \right) \cdot \frac{t}{\Psi_1^2} \right) \\ \frac{\partial \ell}{\partial \Psi_2} &= \sum_{i=1}^n \left(-Y_i \cdot \beta_2 \cdot \cos \left(\frac{t}{\Psi_2} - \gamma_2 \right) \cdot \frac{t}{\Psi_2^2} \right) + \\ &\quad + \sum_{i=1}^n \left(e^{\alpha + \beta_1 \cdot \sin \left(\frac{t}{\Psi_1} - \gamma_1 \right) + \beta_2 \cdot \sin \left(\frac{t}{\Psi_2} - \gamma_2 \right)} \cdot \beta_2 \cdot \cos \left(\frac{t}{\Psi_2} - \gamma_2 \right) \cdot \frac{t}{\Psi_2^2} \right)\end{aligned}$$

Az eltolás paraméterekre:

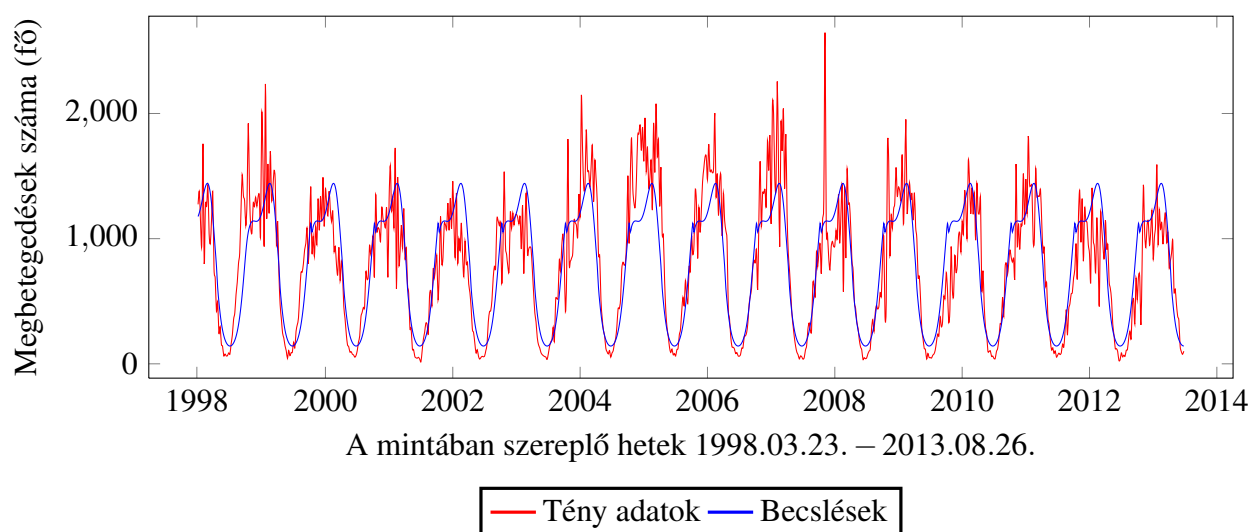
$$\begin{aligned}\frac{\partial \ell}{\partial \gamma_1} &= \sum_{i=1}^n \left(-Y_i \cdot \beta_1 \cdot \cos \left(\frac{t}{\Psi_1} - \gamma_1 \right) \right) + \\ &\quad + \sum_{i=1}^n \left(e^{\alpha + \beta_1 \cdot \sin \left(\frac{t}{\Psi_1} - \gamma_1 \right) + \beta_2 \cdot \sin \left(\frac{t}{\Psi_2} - \gamma_2 \right)} \cdot \beta_1 \cdot \cos \left(\frac{t}{\Psi_1} - \gamma_1 \right) \right) \\ \frac{\partial \ell}{\partial \gamma_2} &= \sum_{i=1}^n \left(-Y_i \cdot \beta_2 \cdot \cos \left(\frac{t}{\Psi_2} - \gamma_2 \right) \right) + \\ &\quad + \sum_{i=1}^n \left(e^{\alpha + \beta_1 \cdot \sin \left(\frac{t}{\Psi_1} - \gamma_1 \right) + \beta_2 \cdot \sin \left(\frac{t}{\Psi_2} - \gamma_2 \right)} \cdot \beta_2 \cdot \cos \left(\frac{t}{\Psi_2} - \gamma_2 \right) \right)\end{aligned}$$

Az így kapott egyenletrendszer alapján, ha a paraméterek indulóértékét egy kezdővektorba rendezzük, és a becslést végrehajthatjuk. A kezdő vektorunk komponensei önkényesen megválasztott 10 értékű skalárok lesznek, a becsléshez használt *Gretl* kód a függelék A.16 alfejezetében van elhelyezve. Az analitikus parciális deriváltakat használó *maximum likelihood* becslés során csak a paramétereket, a standard hibákat, a parciális próbákat és alapvető modell jellemzőket kapunk vissza. A becsült modell jellemzőit tartalmazza a 2.8 táblázat. A megbetegedések számának várhatóértékeire vonatkozó becsléseket mi magunk végezzük el emiatt. Az eltérő *likelihood* optimalizációs probléma miatt az információk kritériumok, és a *log – likelihood* érték nem hasonlítható össze a többi modell hasonló értékével. Az összes felhasznált paraméter rendkívüli mértékben szignifikáns a parciális tesztek empirikus szignifikancia értékei alapján. A magas pszeudo R^2 érték igen erős illeszkedésre utal, hiszen az aktuális modell az üres modell-től közelítőleg 76 százaléknnyira távolodott el. A korábban kiszámított kritikus értékek alapján a Durbin-Watson teszt statisztika első rendű reziduális korreláció jelenlétét jelezné, azonban az eltérő modellspecifikáció miatt nem feltétlenül megbízható ez az eredmény. Azonban a hibagram felírt, csak tengelymetszetet tartalmazó segédregresszió is reziduális autokorrelációt jelez. Mindez indokoltá teszi autoregresszív vagy mozgóátlagolású késleltetett tagok beépítését a modellbe valamilyen formában.

	Paraméter	Std. hiba	z-próba	p-érték	
	α	6.43825	0.00027	23530.0	0.0000
	β_1	−1.06141	0.00034	−3155.0	0.0000
	γ_1	9.30109	0.00057	16220.0	0.0000
	Ψ_1	8.19446	0.00128	6395.0	0.0000
	β_2	0.42433	0.00025	1725.0	0.0000
	γ_2	10.32350	0.00172	6015.0	0.0000
	Ψ_2	4.31450	0.00190	2269.0	0.0000
Függő változó átlaga	801.3449	Függő változó std. σ	528.1449		
McFadden R^2	0.759409	Korrigált R^2	0.759397		
Log-likelihood	-39388.44	Akaike kritérium	78780.88		
Schwarz kritérium	78790.27	Hannan–Quinn	78784.49		
$\hat{\rho}$	0.492406	Durbin–Watson	1.014972		

2.8. táblázat. A determinisztikus Poisson-regressziós modell jellemzői

A csak determinisztikus, nemlineáris függvényformával specifikált Poisson-regressziós modell becslései ábrázolhatóak a tanuló minta által lefedett intervallumon. A becslések alapján adódik, hogy a modell a minimumhelyeken felülbecsli a megbetegedés számot, azonban az ábra meglehetősen hasonlít a nemlineáris legkisebb négyzetek elvén becsült determinisztikus modell esetében kapott ábrára – ez a becslés szerepel a 2.9 ábrán. Ez az eredmény nem meglepő, hiszen nagy mintaelemszám esetén, adott Poisson-eloszlású valószínűségi változók sorozata közelíthető normális eloszlású valószínűségi változók egy sorozatával (Pintér, 2012).



2.9. ábra. A Poisson-regresszió becslései a tény adatokkal szemben

2.3.3. Nemlineáris Poisson-regresszió autoregresszív tagokkal

A továbbiakban az értelmes becslésekhez mindenképpen szükségünk van arra, hogy az erős pozitív autokorrelációt megszüntessük. A determinisztikus és sztochasztikus elemeket is tartalmazó modellben az $AR(4)$ modell is megfelelő lett volna, hiszen a maradék fehérzaj volt. Azonban az előnyösebb illeszkedést jelző információk kritériumok alapján az $ARMA(2,2)$ modellt választottuk. Mindezek miatt, feltételezhetjük, hogy elegendő egy autoregresszív késleltetéseket tartalmazó nemlineáris Poisson-regressziós modellt felépítése, amely nem tartalmaz mozgóátlagolású tagokat. A megbetegedések késleltetett száma helyett a késleltetett megbetegedés szám természetes logaritmusát alkalmazzuk. A nem transzformált késleltetéseket tartalmazó modellek esetében a becslés során a konvergencia még véletlenül generált vektorok esetében sem történik meg, egyes esetekben az optimalizációt a *Gretl* meg sem kezdi. A bevont változók mellett a várhatóérték így adódik:

$$\lambda_t = E(Y_t | \alpha, \beta_1, \beta_2, \Psi_1, \Psi_2, \gamma_1, \gamma_2, \phi_1, \dots, \phi_k, t, \log Y_{t-1}, \dots, \log Y_{t-k})$$

$$\lambda_t = e^{\alpha + \beta_1 \cdot \sin\left(\frac{t}{\Psi_1} - \gamma_1\right) + \beta_2 \cdot \sin\left(\frac{t}{\Psi_2} - \gamma_2\right) + \sum_{j=1}^k \phi_j \cdot \log Y_{t-j}},$$

ahol Y_t az adott időszaki összesített bárányhimlős megbetegedés szám, α a tengelymetszet paraméter, t az indexváltozó, Ψ_1 az első görbéhez tartozó frekvencia paraméter, β_1 az első görbe amplitúdó paramétere, γ_1 a hozzá tartozó eltolási paraméter, Ψ_2 a második görbéhez tartozó frekvencia paraméter, β_2 a második görbe amplitúdó paramétere, γ_2 a hozzá tartozó eltolási paraméter. A $\log Y_{t-k}$ jelöli a megbetegedés szám k héttel korábbi értékének természetes alapú logaritmusát, ϕ_j pedig a késleltetéshez tartozó autoregresszív paraméter. A természetes alapú logaritmus alkalmazása azt a feltételezést sejteti indirekt módon, hogy nem volt olyan hét, amikor nem jelentettek be megbetegedést, és nem is lesz ilyen a későbbiekben sem. A bevezetett λ paraméter, a korábban bevezetett jelölésrendszer szerint az alábbi definíciót jelenti:

$$\Omega = f(\Theta; x) = \alpha + \beta_1 \cdot \sin\left(\frac{t}{\Psi_1} - \gamma_1\right) + \beta_2 \cdot \sin\left(\frac{t}{\Psi_2} - \gamma_2\right) + \sum_{j=1}^k \phi_j \cdot \log Y_{t-j}$$

Az x vektor sorozat a $t, \log Y_{t-1}, \dots, \log Y_{t-k}$ változókból áll, a paramétereket tartalmazó Θ vektor pedig az $\alpha, \beta_1, \gamma_1, \Psi_1, \beta_2, \gamma_2, \Psi_2, \phi_1, \dots, \phi_k$ paraméter halmazból. A maximalizálandó log-likelihood függvényünk ebben az esetben tehát:

$$\max \ell(\alpha, \beta_1, \gamma_1, \Psi_1, \beta_2, \gamma_2, \Psi_2, \phi_1, \dots, \phi_k | Y; t, \log Y_{t-1}, \dots, \log Y_{t-k})$$

Ezen definíció és a 2.12 egyenlet alapján adódik:

$$\begin{aligned} \max \quad & \sum_{i=1}^n Y_i \cdot \left(\alpha + \beta_1 \cdot \sin \left(\frac{t}{\Psi_1} - \gamma_1 \right) + \beta_2 \cdot \sin \left(\frac{t}{\Psi_2} - \gamma_2 \right) + \sum_{j=1}^k \phi_j \cdot \log Y_{i-j} \right) - \\ & - \sum_{i=1}^n e^{\alpha + \beta_1 \cdot \sin \left(\frac{t}{\Psi_1} - \gamma_1 \right) + \beta_2 \cdot \sin \left(\frac{t}{\Psi_2} - \gamma_2 \right) + \sum_{j=1}^k \phi_j \cdot \log Y_{i-j}} \end{aligned}$$

A megoldási algoritmus újra a BFGS grádiens eltérésen alapuló optimalizáció lesz. Ehhez az optimalizáció előtt a paraméterek szerint a parciális deriváltakat manuálisan kell előállítani. Ebből adódik, hogy a természetes logaritmuson alapuló változó transzformáció nem nehezíti meg az analitikus deriváltak előállítását. A tengelymetszet és amplitúdó paraméterekre:

$$\begin{aligned} \frac{\partial \ell}{\partial \alpha} &= \sum_{i=1}^n \left(Y_i - e^{\alpha + \beta_1 \cdot \sin \left(\frac{t}{\Psi_1} - \gamma_1 \right) + \beta_2 \cdot \sin \left(\frac{t}{\Psi_2} - \gamma_2 \right) + \sum_{j=1}^k \phi_j \cdot \log Y_{i-j}} \right) \\ \frac{\partial \ell}{\partial \beta_1} &= \sum_{i=1}^n Y_i \cdot \sin \left(\frac{t}{\Psi_1} - \gamma_1 \right) - \\ & - \sum_{i=1}^n \left(e^{\alpha + \beta_1 \cdot \sin \left(\frac{t}{\Psi_1} - \gamma_1 \right) + \beta_2 \cdot \sin \left(\frac{t}{\Psi_2} - \gamma_2 \right) + \sum_{j=1}^k \phi_j \cdot \log Y_{i-j}} \right) \cdot \sin \left(\frac{t}{\Psi_1} - \gamma_1 \right) \\ \frac{\partial \ell}{\partial \beta_2} &= \sum_{i=1}^n Y_i \cdot \sin \left(\frac{t}{\Psi_2} - \gamma_2 \right) - \\ & - \sum_{i=1}^n \left(e^{\alpha + \beta_1 \cdot \sin \left(\frac{t}{\Psi_1} - \gamma_1 \right) + \beta_2 \cdot \sin \left(\frac{t}{\Psi_2} - \gamma_2 \right) + \sum_{j=1}^k \phi_j \cdot \log Y_{i-j}} \right) \cdot \sin \left(\frac{t}{\Psi_2} - \gamma_2 \right) \end{aligned}$$

A frekvenciát kontrolláló paraméterekre:

$$\begin{aligned} \frac{\partial \ell}{\partial \Psi_1} &= \sum_{i=1}^n \left(-Y_i \cdot \beta_1 \cdot \cos \left(\frac{t}{\Psi_1} - \gamma_1 \right) \cdot \frac{t}{\Psi_1^2} \right) + \\ & + \sum_{i=1}^n \left(e^{\alpha + \beta_1 \cdot \sin \left(\frac{t}{\Psi_1} - \gamma_1 \right) + \beta_2 \cdot \sin \left(\frac{t}{\Psi_2} - \gamma_2 \right) + \sum_{j=1}^k \phi_j \cdot \log Y_{i-j}} \right) \cdot \beta_1 \cdot \cos \left(\frac{t}{\Psi_1} - \gamma_1 \right) \cdot \frac{t}{\Psi_1^2} \\ \frac{\partial \ell}{\partial \Psi_2} &= \sum_{i=1}^n \left(-Y_i \cdot \beta_2 \cdot \cos \left(\frac{t}{\Psi_2} - \gamma_2 \right) \cdot \frac{t}{\Psi_2^2} \right) + \\ & + \sum_{i=1}^n \left(e^{\alpha + \beta_1 \cdot \sin \left(\frac{t}{\Psi_1} - \gamma_1 \right) + \beta_2 \cdot \sin \left(\frac{t}{\Psi_2} - \gamma_2 \right) + \sum_{j=1}^k \phi_j \cdot \log Y_{i-j}} \right) \cdot \beta_2 \cdot \cos \left(\frac{t}{\Psi_2} - \gamma_2 \right) \cdot \frac{t}{\Psi_2^2} \end{aligned}$$

Az eltolás paraméterekre:

$$\begin{aligned}\frac{\partial \ell}{\partial \gamma_1} &= \sum_{i=1}^n \left(-Y_i \cdot \beta_1 \cdot \cos \left(\frac{t}{\Psi_1} - \gamma_1 \right) \right) + \\ &+ \sum_{i=1}^n \left(e^{\alpha + \beta_1 \cdot \sin \left(\frac{t}{\Psi_1} - \gamma_1 \right) + \beta_2 \cdot \sin \left(\frac{t}{\Psi_2} - \gamma_2 \right) + \sum_{j=1}^k \phi_j \cdot \log Y_{i-j}} \right) \cdot \beta_1 \cdot \cos \left(\frac{t}{\Psi_1} - \gamma_1 \right) \\ \frac{\partial \ell}{\partial \gamma_2} &= \sum_{i=1}^n \left(-Y_i \cdot \beta_2 \cdot \cos \left(\frac{t}{\Psi_2} - \gamma_2 \right) \right) + \\ &+ \sum_{i=1}^n \left(e^{\alpha + \beta_1 \cdot \sin \left(\frac{t}{\Psi_1} - \gamma_1 \right) + \beta_2 \cdot \sin \left(\frac{t}{\Psi_2} - \gamma_2 \right) + \sum_{j=1}^k \phi_j \cdot \log Y_{i-j}} \right) \cdot \beta_2 \cdot \cos \left(\frac{t}{\Psi_2} - \gamma_2 \right)\end{aligned}$$

Egy tetszőleges autoregresszív paraméterre:

$$\frac{\partial \ell}{\partial \phi_j} = \sum_{i=1}^n \log Y_{i-j} \cdot \left(Y_i - e^{\alpha + \beta_1 \cdot \sin \left(\frac{t}{\Psi_1} - \gamma_1 \right) + \beta_2 \cdot \sin \left(\frac{t}{\Psi_2} - \gamma_2 \right) + \sum_{j=1}^k \phi_j \cdot \log Y_{i-j}} \right)$$

Az így kapott feltétekből álló egyenletrendszer és egy tetszőleges kezdővektor alapján a becslést végrehajthatjuk. A becsléshez használt *Gretl* kód a függelék A.17 alfejezetében van elhelyezve. A becsült modell jellemzőit a 2.9 táblázat tartalmazza.

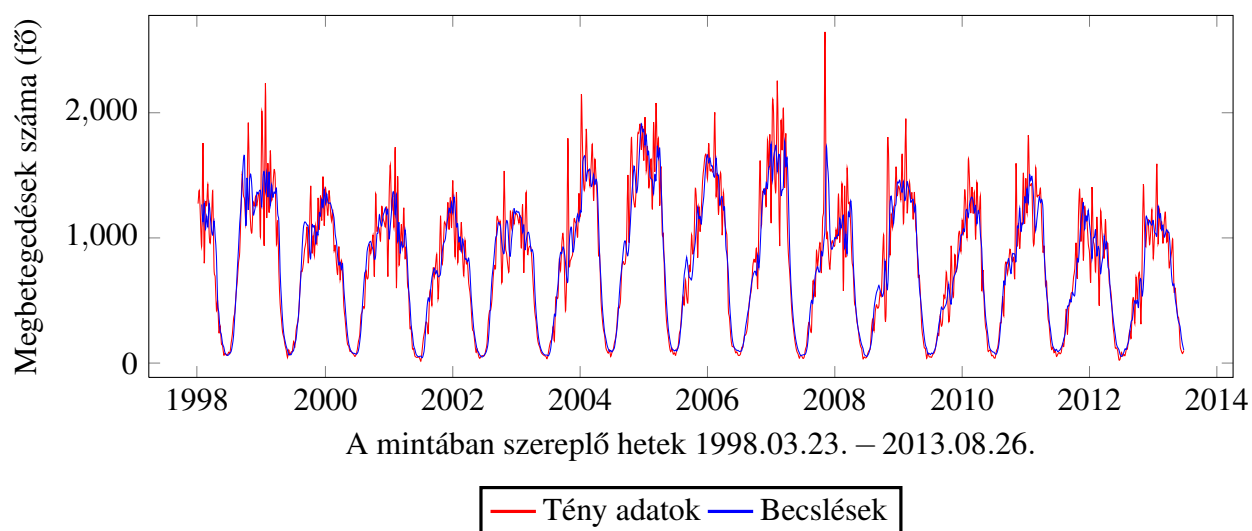
	Paraméter	Std. hiba	z-próba	p-érték
α	0.4274	0.0029	144.7758	0.0000
β_1	-8.8039	0.2094	-42.0508	0.0000
γ_1	1.7401	0.0017	1038.0070	0.0000
Ψ_1	-4.1012	0.0009	-4684.8847	0.0000
β_2	-8.8401	0.2095	-42.2058	0.0000
γ_2	-1.7458	0.0016	-1106.6524	0.0000
Ψ_2	4.1032	0.0009	4737.0732	0.0000
ϕ_1	0.3844	0.0005	784.9353	0.0000
ϕ_2	0.2970	0.0005	632.0621	0.0000
ϕ_3	0.2586	0.0005	566.3627	0.0000
McFadden R^2	0.855684	Korrigált R^2	0.855641	
Log-likelihood	-23509.36	Akaike kritérium	47032.72	
Schwarz kritérium	47065.52	Hannan-Quinn	47045.32	
$\hat{\rho}$	0.001542	Durbin-Watson	1.997916	

2.9. táblázat. A sztochasztikus Poisson-regressziós modell jellemzői

A 3 késleltetés alkalmazását az alábbiakkal lehet indokolni:

1. A 4 késleltetést, vagy több késleltetést tartalmazó modellek nem konvergálnak.
2. A maradéktag fehérzajnak tekinthető, nincsen szignifikáns reziduális autokorreláció a maradékban.
3. Az eredmény epidemiológiailag is értelmes.

A speciális modellspecifikáció miatt a paraméterek nem feltétlenül értelmezhetőek, azonban minden gyakorlatban alkalmazott szignifikancia szint mellett szignifikánsnak tekinthetőek. A McFadden-féle korrigált pszeudo R^2 közel 10 százalékpontos javulást mutat; a tárgy modell 85 százalékban közelítette meg a szaturált modellt. A *log – likelihood* érték és a 3 szelekciós kritérium is egyértelműen jelzi azt, hogy a modell magyarázó ereje nagy mértékben javult a pusztán determinisztikus tagokat tartalmazó Poisson-regresszióhoz viszonyítva. Az, hogy a 3 késleltetést tartalmazó modell esetében történik meg a maradék fehérzajjra transzformálása, újra alátámasztja a lappangási idővel kapcsolatos hipotézist. A késleltetések beépítése miatt a Durbin-Watson statisztika értéke nem értelmes, emiatt a szokásos Q-statisztikákat kérjük le az első 15 késleltetésre, a parciális autokorrelációs statisztikák, az empirikus szignifikancia szintek és a korrelogram a függelék A.18 alfejezetében találhatóak meg. A maradék ezen tesztek alapján fehérzajnak tekinthető, egyedül a 11. késleltetés esetében mutat enyhe szignifikáns pozitív reziduális autokorrelációt a parciális teszt. A tanuló mintára történő előrejelzéseket ábrázolhatjuk az idő függvényében úgy, hogy a tény értékeket is szerepeltetjük, ezt mutatja be a 2.10 ábra.



2.10. ábra. A sztochasztikus Poisson-regresszió becslései a tény adatokkal szemben

3. fejezet

Egyváltozós előrejelzések és elemzések

3.1. A determinisztikus modell előrejelzései

A nemlineáris legkisebb négyzetek módszerével becsült III. *NLS* modell esetében meghatározhatjuk azt, hogy a bevezetett index mely értékeire lesz várhatóan a legmagasabb és a legalacsonyabb az országos megbetegedés szám. Ez tulajdonképpen a szélsőérték helyek megkeresését jelenti, melyek meglétéhez optimalizálni kell a függvényt. Azért ezt a modellt választottam, mert itt a paraméterek becslését nem befolyásolják a bevont késleltetések, amelyek az index szerinti optimalizáció során kiesnek az egyenletből. A célfüggvényünk tehát:

$$Y = \alpha + \beta_1 \cdot \sin\left(\frac{t}{\Psi_1} - \gamma_1\right) + \beta_2 \cdot \sin\left(\frac{t}{\Psi_2} - \gamma_2\right) \rightarrow \max (\min)$$

A korlátozó feltétel pedig az alábbi:

$$t \in (1, 52)$$

Az optimalitási feltétel az, hogy az index szerinti parciális derivált nullával egyenlő:

$$\begin{aligned} 0 &= \frac{\partial Y}{\partial t} \\ 0 &= \beta_1 \cdot \cos\left(\frac{t}{\Psi_1} - \gamma_1\right) \cdot \frac{1}{\Psi_1} + \beta_2 \cdot \cos\left(\frac{t}{\Psi_2} - \gamma_2\right) \cdot \frac{1}{\Psi_2} \end{aligned}$$

Az optimalizációt végrehajtó *MATLAB* kódrészletek a függelék A.19 alfejezetében megtalálhatók. Az első optimalizációs probléma megoldása, és a célfüggvény érték ezen szélsőérték helyen:

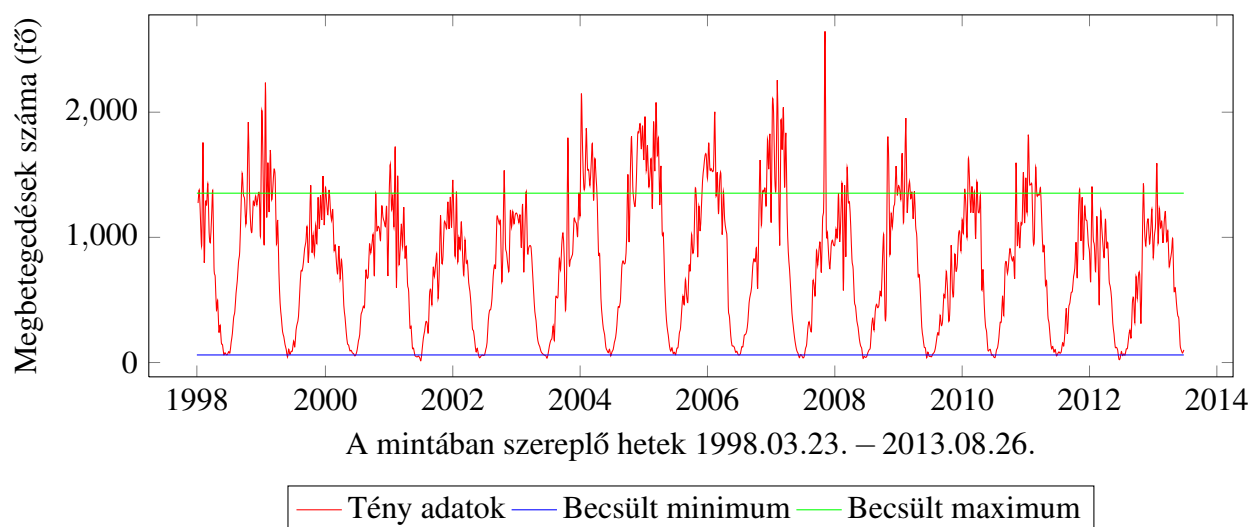
$$t \approx 16.8254 \qquad Y \approx 1353.0098$$

Azaz az átlagos megbetegedés szám március utolsó és április első heteiben lesz a legmagasabb. Ekkor a várható megbetegedés szám 1353 fő lesz. A második optimalizációs probléma

megoldása, és a célfüggvény érték ezen a szélsőérték helyen:

$$t \approx 36.4621 \quad Y \approx 60.5111$$

Azaz az átlagos megbetegedés szám szeptemberben lesz a legalacsonyabb. Mindez alátámasztja a kezdetben megfogalmazott hipotézist, hogy a megbetegedések száma erősen kötődik az iskolakezdéshez, hiszen a minimum pontosan szeptember elején van. A március végén megjelenő maximumot és letörést indokolhatja az, hogy a megbetegedések számát korlátozza az érintett populáció száma. A becsült megbetegedés szám szerepel az idő függvényében, a szélsőértékekkel együtt a 3.1 ábrán.



3.1. ábra. A III. NLS által becsült szélsőértékek

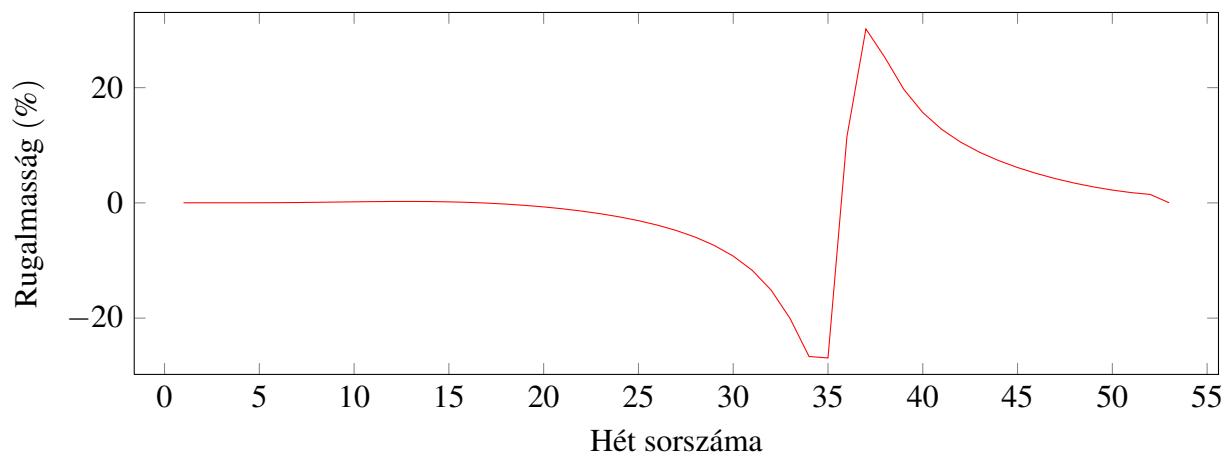
A minimum pontok esetében az eltérések nem jelentősek, azonban a maximumok esetében nagy tévedések vannak. A maximumot gyakran korábban veszi fel az idősor, az idő függő variancia problémáját ez jól mutatja.

A regressziós egyenlet alapján a rugalmasságot is ki lehet számítani minden hétre. Ez ebben az esetben így adódik:

$$\varepsilon_t = \frac{\partial Y_t}{\partial t} \cdot \frac{t}{Y_t} = \frac{t \cdot \left(\beta_1 \cdot \cos\left(\frac{t}{\Psi_1} - \gamma_1\right) \cdot \frac{1}{\Psi_1} + \beta_2 \cdot \cos\left(\frac{t}{\Psi_2} - \gamma_2\right) \cdot \frac{1}{\Psi_2} \right)}{\alpha + \beta_1 \cdot \sin\left(\frac{t}{\Psi_1} - \gamma_1\right) + \beta_2 \cdot \sin\left(\frac{t}{\Psi_2} - \gamma_2\right)} \quad (3.1)$$

Az időindex és a paraméterek korábbról adódtak, ezek alapján és az (3.1) egyenlet alapján a rugalmasság értékek egy *for* ciklus segítségével kiszámíthatóak. A számításokhoz tartozó *MATLAB* kód a függelék A.21 alfejezetében van elhelyezve. A rugalmasság értékek alapján adódik, hogy az iskolaév előrehaladtával átlagosan csökken a bejelentett új megbetegedések

száma, majd az iskolaév kezdetével robbanásszerűen növekedni kezd az átlagos bejelentett megbetegedés szám. Az abszolút értékben véve nagy, akár 20-30% körül mozgó rugalmasság értékek alapján adódik, hogy a folyamat rendkívül gyorsan változik iskolakezdéskor és a tanév végén is. A negatív és abszolút értékben magas rugalmasság értékek, amelyek a nyári heteket jellemzik arra utalnak, hogy a folyamat nagy sebességgel hal el. A hetek szerinti rugalmasság szerepel a hetek függvényében a 3.2 ábrán.



3.2. ábra. A III. NLS modell időindex szerinti rugalmassága

3.2. A determinisztikus-sztokasztikus modell előrejelzései

A pusztán determinisztikus III. NLS modell előrejelzéseit azért nem vizsgáljuk, mert reziduális autokorrelációtól szenved. Az előrejelzések jóságát jellemző illeszkedési mutatók részletes leírása megtalálható (Hajdú, 2013) művében. Az országos megbetegedés számot vizsgáló modellt az 1998 márciusa és 2013 szeptembere közötti adatok alapján parametrizáltuk. A tanuló mintán belüli előrejelzéseket jellemző hiba mutatók az alábbiak:

$$ME = 0.2326 \quad MPE = 5.6231\%$$

Az átlagolt hiba mértéke 0.2326 megbetegedés, az átlagolt százalékos hiba pedig 5.6231% az adott heti valós megbetegedés számhoz képest, ez arra utal, hogy a modell felül becsüli a megbetegedések számát a tanuló mintán. Az abszolút eltérések, és a négyzetes eltérések alapján számolt mutatók:

$$MAE = 149.9269 \quad RMSE = 216.0375 \quad MAPE = 26.1805\%$$

Az átlagos abszolút tévedés a becsült megbetegedés szám esetében közelítőleg 150 fő, a heti valós megbetegedés számra vetített százalékos érték pedig 26.1805%. Az átlagos négyzetes téved

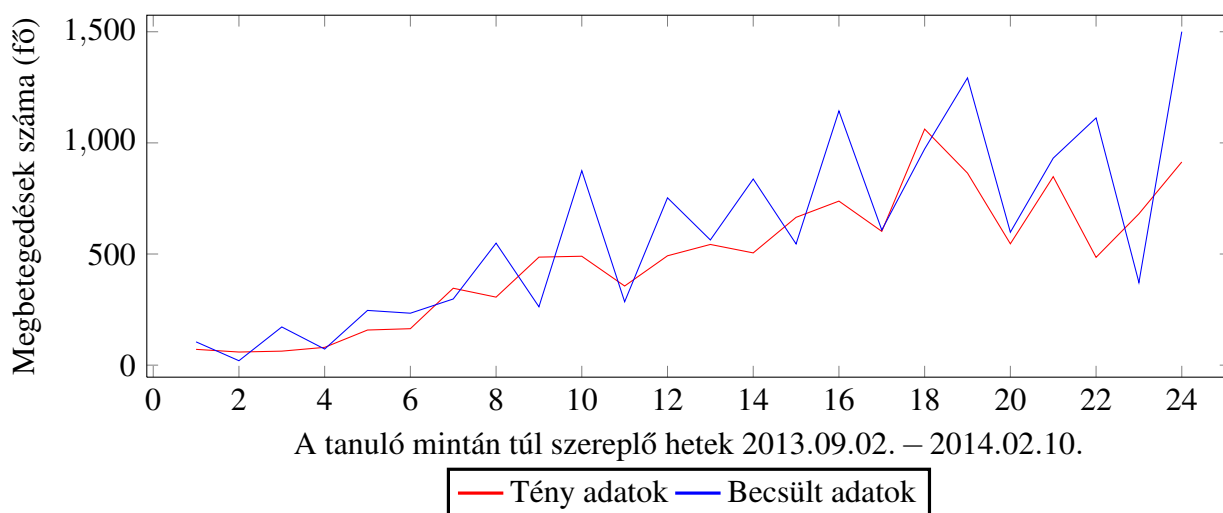
mértéke 216 fő, az is utal a változó szórásra, hogy ez nem egyezik az átlagos abszolút eltéréssel. A tanuló mintán túlra is vannak megfigyeléseink, és az utolsó adatfelvétel 2014. február 2. hetére esik. Ezen adatokkal és a korábban már parametrizált modellel statikus előrejelzést tudunk végezni a tanuló mintán túlra a III. NLS modell és az $ARMA(2,2)$ modell eredményeit felhasználva. Az előrejelzéseket, mindig a következő időszakra végezzük, a modell újra becslése nélkül. Az így kapott becslült megbetegedés számokat jellemző illeszkedési mutatók:

$$ME = 117.6262 \quad MPE = 28.7419\%$$

Azaz a modell nagyon erősen felülbecsli ezen hetekben a megbetegedés számot, és a felülbecslés mértéke átlagosan 29% körül mozog. Az abszolút és négyzeteltérések alapján is tudunk illeszkedési mutatókat számolni:

$$MAE = 193.2528 \quad RMSE = 264.8455 \quad MAPE = 47.7012\%$$

Az átlagos abszolút eltérés ebben az esetben kerekítve 193 fő, az átlagos négyzetes eltérés pedig kerekítve 265 fő, azaz a szórás nagy valószínűséggel nem állandó. Az átlagos abszolút százalékos eltérések az 50% körüli értéket közelítik. A kapott becslések szerepelnek a 3.3 ábrán.



3.3. ábra. A III. NLS és $ARMA(2,2)$ becslései

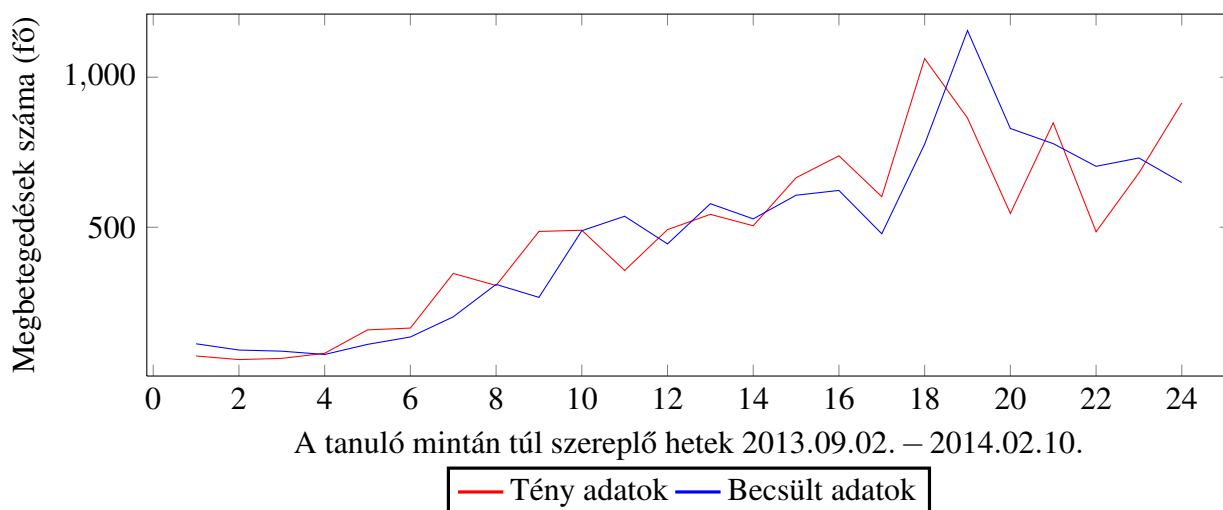
A 3.3 ábra alapján adódik, hogy a modell a közel fél éves tanuló mintán túli adatsoron felülbecsli a megbetegedés számot, ezt a tényt az abszolút eltéréseken alapuló illeszkedési mutatók is kimutatták. A sztochasztikus Poisson-regressziós modell és a $SARIMA$ modell, nem hibázik ilyen jól leírható struktúra szerint, azaz a hibák nem autokorreláltak ezen az intervallumon.

3.3. A sztochasztikus modell előrejelzései

A tisztán sztochasztikus modellt is az 1998 márciusától 2013 augusztusáig terjedő időszakon parametrizáltuk, ezen az időintervallumon a becslések hibáit jellemző mutatók:

$$ME = 0.8795 \quad MPE = 10.0424\%$$

Az átlagolt tévedések mérték az 1 megbetegedést közelíti, az átlagos százalékos tévedés pedig 10 százalék fölött van. Azaz a modellünk felül becsli a megbetegedés számot a tanuló mintán. A determinisztikus és sztochasztikus tagokat tartalmazó modellnél rosszabbul teljesít ez a modell, azonban a két modell esetében a rövidülések miatt nem ugyanakkora időszorra jeleztünk előre. A becsléseket, amelyeket a tanuló mintán parametrizált modellel végeztünk a 3.4 ábrán ábrázoltuk. A 3.4 ábrán megfigyelhető az, hogy nincsenek folytonos felül becslések az előrejelzett megbetegedés számban.



3.4. ábra. A SARIMA (3,1,3)(1,0,1) becslései

Az abszolút eltérések, és a négyzetes eltérések alapján az alábbi illeszkedési mutatók állíthatóak elő:

$$MAE = 149.9269 \quad RMSE = 216.0375 \quad MAPE = 26.1805\%$$

Ezen mutatók alapján azonban a pusztán sztochasztikus modell az, amely jobban teljesít a tanulómintára történő előrejelzéseket vizsgálva. Itt is rendkívüli módon eltér egymástól az átlagos abszolút eltérések és az átlagos négyzetes eltérések nagysága. A korábban bevezetett statisztikus módszerrel előrejelezhető az idősor alakulása 2014 februárjának közepéig, az így kapott előrejelzésekre ugyanezen mutatók kiszámolhatóak. Az előrejelzés hibákon alapuló mutatók így adódnak:

$$ME = -0.7194 \quad MPE = 5.5933\%$$

Azaz a tanuló mintán túlra jobban teljesít a teljesen sztochasztikus modell, és enyhén alulbecsli a megbetegedés számot, ellentétben a determinisztikus és sztochasztikus tagokat tartalmazó modellel, amely erősen felülbecsülte a megbetegedések számát. Az abszolút eltérések, és a négyzetes eltérések alapján az alábbi illeszkedési mutatók állíthatók elő a tanuló mintán túlra:

$$MAE = 106.6357 \quad RMSE = 145.5627 \quad MAPE = 25.5039\%$$

Ezek a mutatók is egy jobb illeszkedést jeleznek, azaz a tanuló mintán túli fél éves intervallumra jobban teljesít a sztochasztikus modell.

3.4. A determinisztikus nemlineáris Poisson-regresszió előrejelzései

A determinisztikus Poisson-regressziós modell esetében meghatározhatjuk, hogy az időindex mely értékeire lesz a legmagasabb és a legalacsonyabb az országos megbetegedés szám várhatóértéke. Ez tulajdonképpen a szélsőérték helyek megkeresését jelenti, melyek meglétéhez optimalizálni kell a függvényt, figyelve az intervallum korlátokra. Ez a probléma így formalizálható:

$$E(Y) = e^{\alpha + \beta_1 \cdot \sin\left(\frac{t}{\Psi_1} - \gamma_1\right) + \beta_2 \cdot \sin\left(\frac{t}{\Psi_2} - \gamma_2\right)} \rightarrow \max(\min)$$

A korlátozó feltétel pedig az alábbi:

$$t \in (1, 52)$$

Az exponenciális függvény azonban csak egy monoton transzformáció, a célfüggvényünk így is felírható:

$$\alpha + \beta_1 \cdot \sin\left(\frac{t}{\Psi_1} - \gamma_1\right) + \beta_2 \cdot \sin\left(\frac{t}{\Psi_2} - \gamma_2\right) \rightarrow \max(\min)$$

Az optimalitási feltétel tehát:

$$\begin{aligned} 0 &= \frac{\partial Y}{\partial t} \\ 0 &= \beta_1 \cdot \cos\left(\frac{t}{\Psi_1} - \gamma_1\right) \cdot \frac{1}{\Psi_1} + \beta_2 \cdot \cos\left(\frac{t}{\Psi_2} - \gamma_2\right) \cdot \frac{1}{\Psi_2} \end{aligned}$$

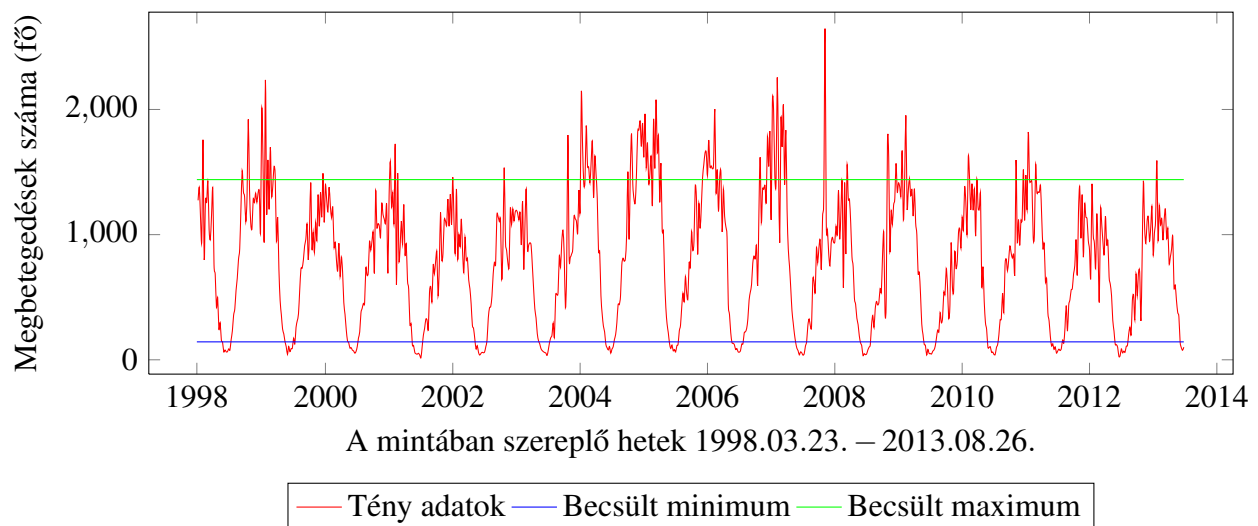
Az optimalizációt végrehajtó *MATLAB* kódrészletek a függelék A.20 részfejezetében megtalálhatóak. Az optimalizációs probléma egyik megoldása, és az eredeti célfüggvény érték ezen szélsőérték helyen:

$$t \approx 18.1357 \quad Y \approx 1439.517$$

Azaz a megbetegedés szám várhatóértéke március utolsó és április első heteiben lesz a legmagasabb. Ekkor a megbetegedés szám várhatóértéke 1440 fő lesz. Az optimalizációs probléma másik megoldása, és az eredeti célfüggvény érték ezen szélsőérték helyen:

$$t \approx 37.6974 \quad Y \approx 143.295$$

Azaz a megbetegedés szám várhatóértéke szeptemberben lesz a legalacsonyabb. Mindez ezen modell esetében is alátámasztja a kezdetben megfogalmazott hipotézist, hogy a megbetegedések száma növekszik az iskolakezdés után, hiszen a minimum szeptember elején van. A várhatóérték minimuma és maximuma a megfigyelésekkel együtt szerepel a 3.5 ábrán.



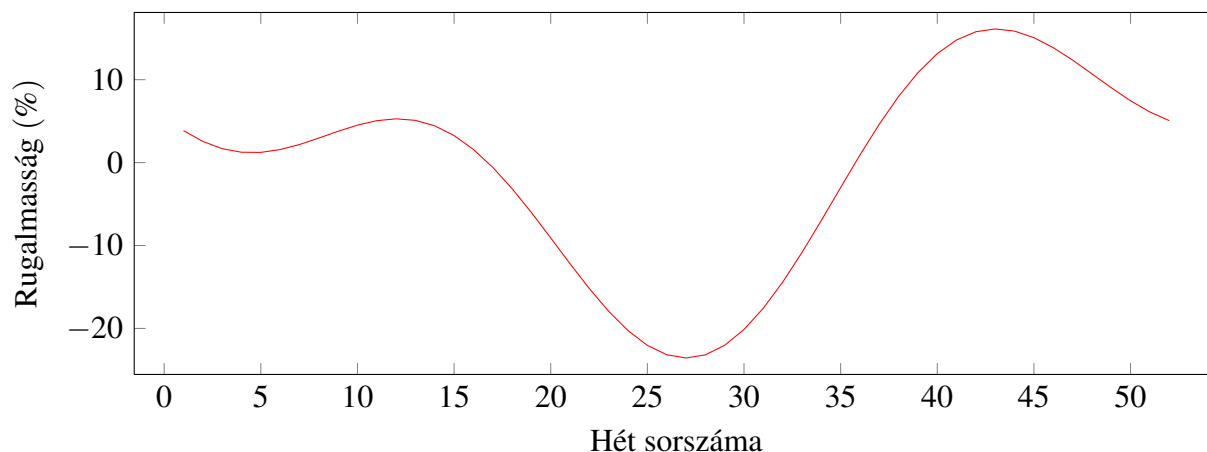
3.5. ábra. A determinisztikus Poisson–regresszió által becsült szélsőértékek

A regressziós egyenlet alapján a rugalmasságot is ki lehet számítani minden hétre, a rugalmasság ebben az esetben, a parciális deriválás és rendezés után így adódik:

$$\varepsilon_t = \frac{\partial Y_t}{\partial t} \cdot \frac{t}{Y_t} = t \cdot \left(\frac{\beta_1 \cdot \cos\left(\gamma_1 - \frac{t}{\Psi_1}\right)}{\Psi_1} + \frac{\beta_2 \cdot \cos\left(\gamma_2 - \frac{t}{\Psi_2}\right)}{\Psi_2} \right)$$

Az időindex és a paraméterek korábbról ismertek, ezek alapján és a rugalmasságot leíró egyenlet alapján a rugalmasság értékek egy *for* ciklus segítségével kiszámíthatóak. A számításokhoz tartozó *MATLAB* kód a függelék A.22 alfejezetében van elhelyezve. Az ehhez a modellhez tartozó rugalmasság értékek alapján is látszik, hogy az iskolaév előrehaladtával átlagosan csökken

a bejelentett új megbetegedések száma, majd az iskolaév kezdetével robbanásszerűen növekedni kezd. Az abszolút értékben véve nagy, akár 15-20% körül mozgó rugalmasság értékek alapján itt is adódik, hogy a folyamat rendkívül gyorsan változik iskolakezdekor és a tanév végén is. A negatív és abszolút értékben magas rugalmasság értékek, amelyek a nyári heteket jellemzik arra utalnak, hogy a folyamat nagyon nagy sebességgel hal el. Az időindex szerinti rugalmasság az idő függvényében ábrázolva szerepel a 3.6 ábrán.



3.6. ábra. A determinisztikus Poisson-regresszió időindex szerinti rugalmassága

3.5. A sztochasztikus Poisson–regresszió előrejelzései

Az autoregresszív tagokkal bővített Poisson-regressziós modellt is az 1998 márciusától 2013 augusztusáig terjedő idősoron parametrizáltuk, a becslések hibáit jellemző mutatók:

$$ME = -1.837 \cdot 10^{-12} \quad MPE = 8.7643\%$$

Az átlagolt tévedések mértéke a 0 megbetegedést közelíti, az átlagos százalékos tévedés pedig 10 százalék alatt van. Azaz a modellünk a tény adathoz viszonyítva felülbecsli a megbetegedés számot a tanuló mintán. A determinisztikus és sztochasztikus tagokat tartalmazó modellnél rosszabbul teljesít, azonban a tisztán sztochasztikus modellnél jobb. Az abszolút eltérések, és a négyzetes eltérések alapján az alábbi illeszkedési mutatók számolhatóak:

$$MAE = 148.15724 \quad RMSE = 222.4167 \quad MAPE = 23.9853\%$$

Ezen mutatók alapján azonban a pusztán sztochasztikus modell az, amely jobban teljesít a tanulómintán a Poisson-regressziónál. A Poisson-regresszióba ágyazott sztochasztikus modell, azonban jobban teljesít az első *NLS* és *ARMA* elven becsült modellnél. Ezen modellnél is rendkívüli módon eltér egymástól az átlagos abszolút eltérések és az átlagos négyzetes eltérések

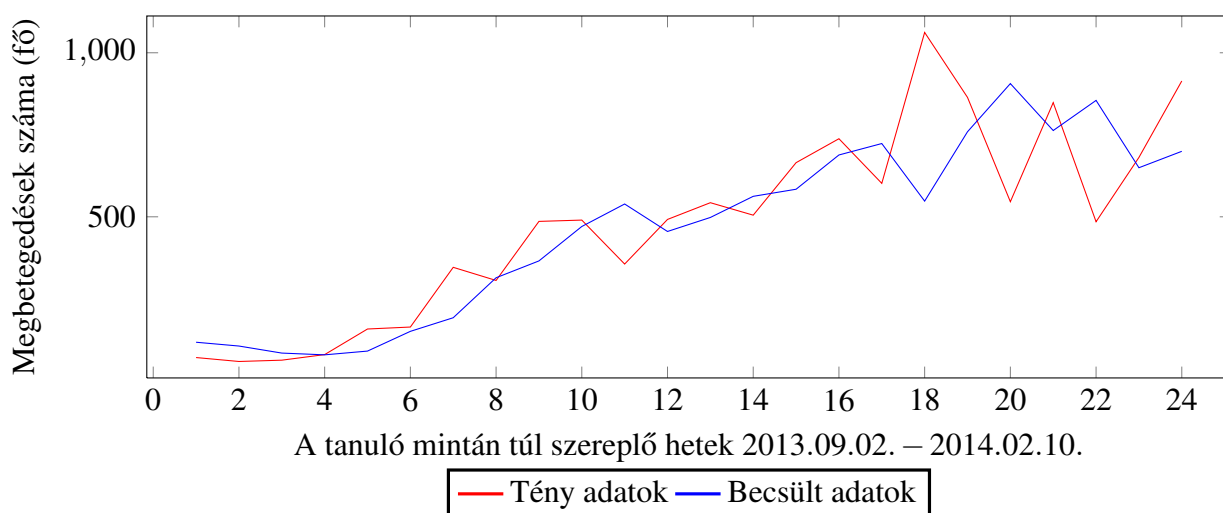
nagysága. A korábban bevezetett statikus módszerrel előrejelezhető az idősor alakulása 2014 februárjának közepéig, az így kapott előrejelzésekre ugyanezen mutatók kiszámolhatóak. Az előrejeles hibákon alapuló mutatók így adódnak:

$$ME = -4.5276 \quad MPE = 0.0897\%$$

Azaz a tanuló mintán túlra gyengébben teljesít min a teljesen sztochasztikus modell, és enyhén alulbecsli a megbetegedés számot mint a tisztán sztochasztikus modell. Az abszolút eltérések, és a négyzetes eltérések alapján az alábbi illeszkedési mutatók állíthatóak elő a tanuló mintán túlra:

$$MAE = 148.7322 \quad RMSE = 167.5533 \quad MAPE = 38.4423\%$$

Ezek a mutatók is egy jobb illeszkedést jeleznek mint a determinisztikus modell esetében, azonban a tanuló mintán túli féléves intervallumra gyengébben teljesít mint a sztochasztikus modell. Az így kapott becslések szerepelnek a 3.7 ábrán.



3.7. ábra. A sztochasztikus Poisson-regresszió becslései

4. fejezet

Többváltozós modellezés és elemzések

Ebben a fejezetben olyan elemzések és modellezések szerepelnek, amelyeknek a 19 megyére és Budapestre vonatkozó egyedi idősorok képezik az alapját. A felhasznált adatok, amely alapján az elemzések készültek, 2005.01.03–2013.08.26 között kerültek felvételre. Az adatbázisból így egyetlen adatfelvétel sem hiányzik, a megelőző időszakokra vonatkozóan gyakran csak az aggregált – országos – értékeket rögzítették. A fejezetben három eltérő szempontú megközelítést alkalmazok:

1. A megyéket leíró változókból kialakuló főkomponensek vizsgálatát
2. Globális területi autokorrelációs mutatók számítását
3. Bayes-i vektor autoregressziós modelleket

Az első módszer alapján megvizsgálható az, hogy mely megyék viselkednek hasonló módon, és mindez milyen összefüggésben van a térbeli kiterjedésekkel. A második módszer alkalmas a területi autokorreláció általános makro-szintű vizsgálatára. Visszaulva a korábbi eredményekre, ezzel a módszerrel vizsgálható az, hogy a nyári szünet alatt lazábbá válik-e a területi autokorreláció a megyék között. A harmadik módszer lehetőséget ad arra, hogy a területi és időbeli hatásokat a megyék között akár páronként is megvizsgálhassuk, a kapcsolatok irányát vagy kölcsönös mivoltát bizonyítsuk. A Bayes-i vektor autoregressziós modellezés keretét ad annak, hogy ne kelljen a nem releváns területi kapcsolatokhoz tartozó paramétereket megbecsülni. Az első és harmadik módszer viszonylag szabadon paraméterezhető és bővíthető, emiatt csak a legalapvetőbb megközelítések kerülnek bemutatásra a dolgozatban.

4.1. Főkomponens elemzések

A főkomponensek bevezetésének oka esetünkben nem az, hogy a magyarázóváltozók információ tartalmát szeretnénk tömöríteni és nem is ortogonális magyarázóváltozók bevezetése (Kovács et al., 2011). A célunk ezzel szemben az, hogy megmutassuk mely megyék megbetegedés számai korrelálnak egymással erősen, azaz mely megyékre vonatkozó idősorok viselkednek hasonlóan. A szezonális hatások által befolyásolt idősorok esetében a közös szezonális hatás miatt megjelenő korreláció erősen befolyásolhatja a kialakuló főkomponenseket (Joyeux, 1992), emiatt a szezonálisan differenciázott idősorokra is érdemes elvégezni a főkomponens elemzést. A szignifikánsan pozitív ferdeségű változók esetében természetes alapú logaritmus alkalmazása lehet majd szükséges, ha így a ferdeség abszolút értékben véve csökken (Kovács et al., 2011)

A főkomponensek alkalmazása előtt minden esetben meg kell vizsgálni azt, hogy a mintáink – ebben az esetben a megyei megbetegedés számok – milyen mértékben korreláltak, és a főkomponens elemzésnek alkalmazásának van-e értelme. Ezt az előfeltételt a Bartlett-féle próbával vizsgáljuk, a főkomponensek alkalmazását emellett a Kayser-Meyer-Olkin mértékkel lehet megindokolni, ha annak nagysága bizonyos küszöbnél nagyobb (Cureton, D’Agostino, 1983). Mind a három esetben a *varimax* rotációt alkalmazunk (Kaiser, 1958), ennek köszönhetően a változók különböző csoportjai jobban szét fognak válni. A főkomponensek szelektálásánál a Kaiser-szabályt alkalmazzuk, azaz azokat a változókat tartjuk meg, amelyekhez legalább $\lambda \geq 1$ sajátérték tartozik.

4.1.1. Főkomponensek a megyei idősorokra

A differenciázatlan megyei idősorok esetében a változók egyike sem mutat $\alpha = 1\%$ esetén szignifikáns pozitív ferdeséget, emiatt nem szükséges a megyéket leíró változók transzformációja. A leíró statisztikákat tartalmazó táblázat a függelék B.1 alfejezetében található meg. A Bartlett-féle teszt hipotézisei:

$$H_0 : C = I$$

$$H_1 : C \neq I$$

Azaz a nullhipotézis azt fogalmazza meg, hogy a korrelációs mátrix egy egységmátrix így a változók korrelálatlanok. A teszt értéke az alábbi:

$$\text{Bartlett-próba} = 6222.767$$

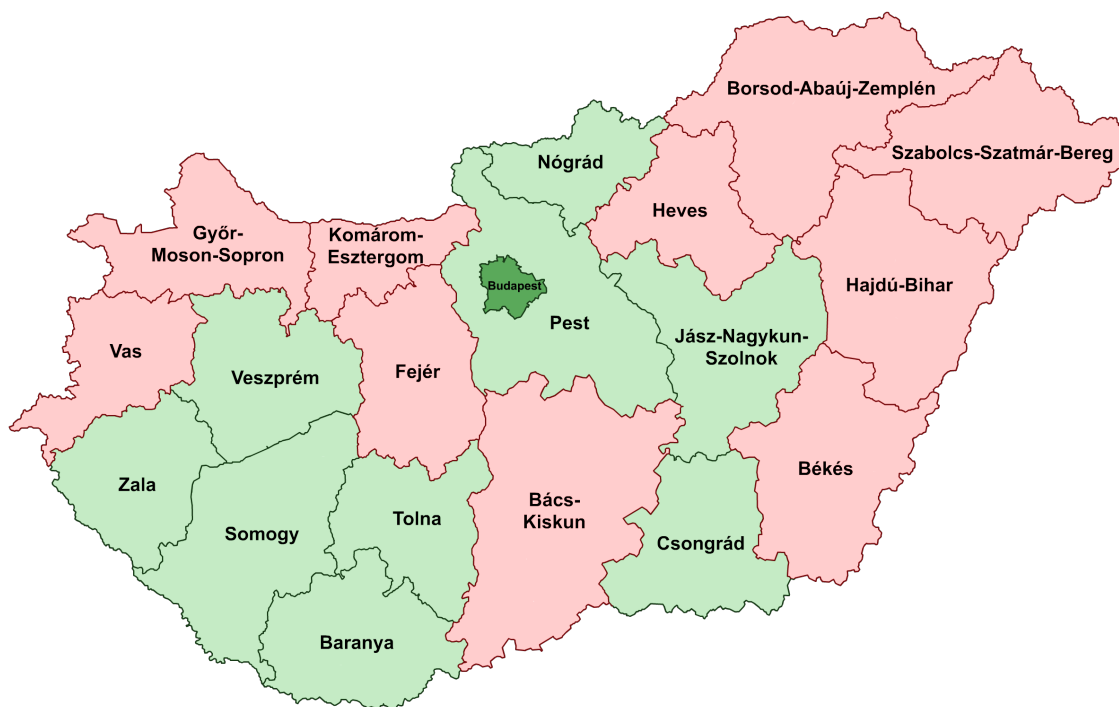
A teszt kritikus értéke χ^2 eloszlást követ és a kritikus érték szabadság foka a vizsgált változók számától (k) függ egyedül. A $p = 0.000$ empirikus szignifikancia mellett, a nullhipotézist elvetjük, a korrelációs mátrix nem egységmátrix. A főkomponensek alkalmazása a Kayser-Meyer-Olkin mutató alapján is indokolt, ugyanis ennek értéke 0.943, amely rendkívül erős korreláltságra utal.

A korábban választott $\lambda = 1$ küszöbérték mellett két főkomponens jön létre. A két főkomponens az eredeti variancia közel 60 százalékát magyarázza, azonban viszonylag egyenlően oszlik meg a két változó között a magyarázott variancia, hiszen a *varimax* rotáció ezt biztosítja. A forgatott főkomponensek alapján az első változó az eredeti variancia 33 százalékát magyarázza, a második csak a 27 százalékát. A főkomponenseket jellemző statisztikák megtalálhatóak a függelék B.2 alfejezetében. A *varimax* rotációval elforgatott komponens mátrix esetében meg lehet állapítani, hogy mely változókkal korrelálnak erősen a létrehozott főkomponensek. Erős korrelációnak vesszük azt, amikor a lineáris korrelációs együttható nagysága abszolút értékben meghaladja a $\rho = 0.5$ értéket. A komponensekkel a mi definíciónk szerint erősen korreláló változók:

1. főkomp.	Budapest	2. főkomp.	Budapest
	Bács–Kiskun		Baranya
	Békés		Csongrád
	Borsod–Abaúj–Zemplén		Jász-Nagykun-Szolnok
	Fejér		Nógrád
	Győr–Moson–Sopron		Pest
	Hajdú–Bihar		Somogy
	Heves		Tolna
	Komárom-Esztergom		Vas
	Pest		Veszprém
	Szabolcs–Szatmár–Bereg		Zala

A megyéket leíró változó halmazban két olyan változó van, amely jelentős hatást gyakorol mind a két forgatott főkomponensre – a budapesti és a pesti megbetegedés szám. A forgatott komponens mátrixhoz tartozó részletes korrelációs együtthatók megtalálhatóak a függelék B.3 alfejezetében. A forgatott mátrix alapján megállapítható, hogy az első főkomponensben szerepelnek a nyugat-magyarországi és kelet-magyarországi megyék, illetve Pest megye és Budapest. A második főkomponensben szerepelnek egyes közép és dél-nyugat magyarországi megyék, ezek alapján adódik, hogy a megbetegedésekre vonatkozó idősorok bizonyos jól körülhatárolható régiókban azonosan viselkednek.

A 4.1 ábrán is kivehető az, hogy főkomponensek szerint megszínezett térképen elkülönülnek azon eredeti változók, amelyek a két létrehozott főkomponenssel erősen korrelálnak.



4.1. ábra. Nem differenciázott főkomponensek alapján színezett térkép

4.1.2. Főkomponensek a szezonálisan differenciázott megyei idősorokra

A szezonális differenciázás hatására az eredetileg $n = 452$ hosszúságú idősoraink $n = 400$ hosszúságúra redukálódtak. A szezonálisan differenciázott megyei idősorok esetében a változók egyike sem mutat $\alpha = 1\%$ esetén szignifikáns pozitív ferdeséget, emiatt nem szükséges a megyéket leíró változók transzformációja. A leíró statisztikákat tartalmazó táblázat a függelék B.4 alfejezetében található meg. A Bartlett-féle teszt értéke az alábbi:

$$\text{Bartlett-próba} = 2102.251$$

Az empirikus szignifikancia szint értéke $p = 0.000$ azaz a nullhipotézist elvetjük, a korrelációs mátrix nem egységmátrix, a főkomponensek alkalmazása a Kayser-Meyer-Olkin mutató alapján is indokolt, ugyanis ennek értéke 0.684, ez azonban csak gyenge összefüggésre utal. A korábban választott $\lambda = 1$ küszöbérték mellett hét főkomponens jön létre. A hét főkomponens az eredeti variancia közel 65 százalékát magyarázza, azonban egyenlőtlenül oszlik meg a hét változó között a magyarázott variancia. A forgatott főkomponensek alapján az első változó az

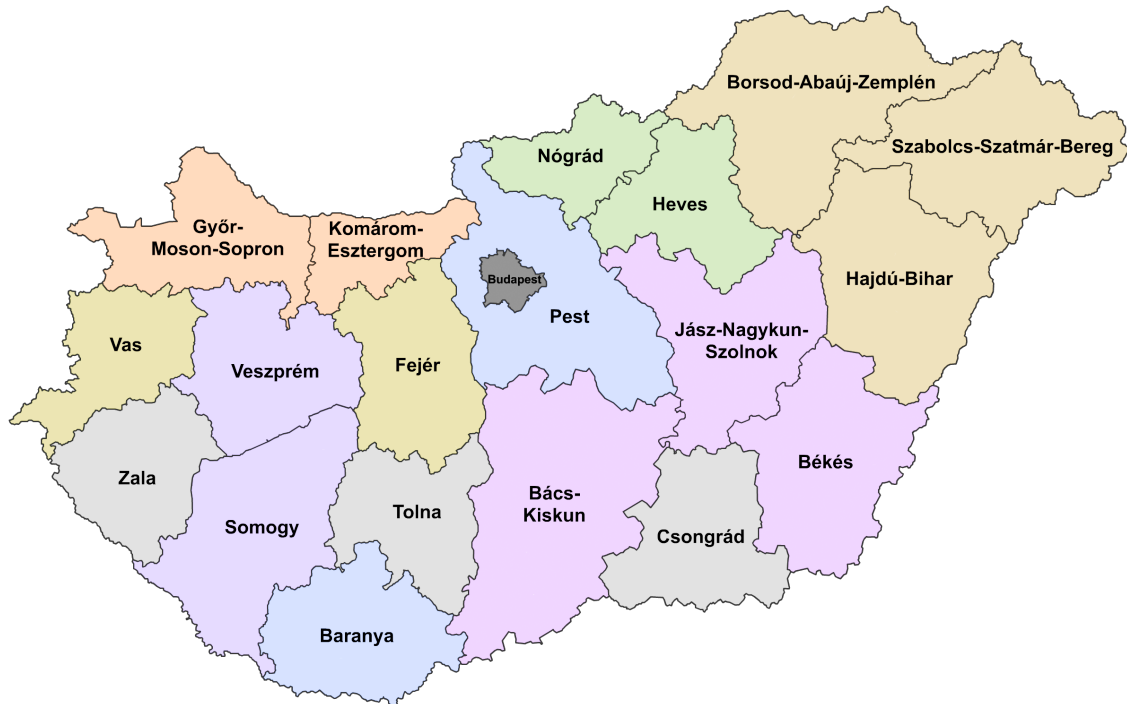
eredeti variancia 12 százalékát magyarázza, a leggyengébb információ tartalommal bíró hetedik főkomponens pedig közel a variancia 8 százalékát. A főkomponenseket jellemző statisztikák megtalálhatóak a függelék B.5 alfejezetében. A *varimax* rotációval elforgatott komponens mátrix alapján meg lehet állapítani, hogy mely változókkal korrelálnak erősen a létrehozott főkomponensek. Erős korrelációnak vesszük azt, amikor a lineáris korreláció mértéke abszolút értékben meghaladja a $\rho = 0.5$ értéket.

1. főkomp.	{ Somogy Veszprém	2. főkomp.	{ Fejér Vas
3. főkomp.	{ Borsod–Abaúj–Zemplén Hajdú–Bihar Szabolcs–Szatmár–Bereg	4. főkomp.	{ Győr–Moson–Sopron Komárom–Esztergom
5. főkomp.	{ Heves Nógrád	6. főkomp.	{ Baranya Pest
7. főkomp.	{ Bács–Kiskun Békés Jász–Nagykun–Szolnok		

A forgatott komponens mátrixhoz tartozó részletes értékek megtalálhatóak a függelék B.6 alfejezetében. A forgatott komponens mátrix alapján, megállapítható, hogy a hét létrejövő komponens hét különböző két vagy három megyéből álló régióval korrelál erősen, amelyek a hét esetből öt esetben szomszédosak. Mindez alátámasztja a dolgozat elején a harmadik kutatási kérdésre adott hipotetikus választ, azaz egyes esetekben kimutatható együttmozgás a megbetegedés számok között a szomszédos megyék esetében; ebben az esetben még a változások is együtt mozognak. Azaz a szezonálisan differenciázott idősor esetében az egymással érintkező megyék korrelálnak erősen, és határoznak meg főkomponenseket. Budapest, Csongrád, Tolna és Zala megye a hét létrejövő főkomponens egyikével sem korrelál erősen, emiatt a főkomponensek alapján megszínezett térképen szürke színnel szerepelnek. A 4.2 ábrán a hét főkomponens mindegyike egyedi színnel van jelölve a korábban már felhasznált térképen.

4.2. A Moran-féle I-index

A Moran-féle I-index (Moran, 1950) különböző területekről felvett keresztmetszeti adatok esetében mutatja meg azt, hogy a szomszédos területekről felvett adatok egy adott időpontban mennyire erősen korrelálnak egymással. A Geary-féle C-index mutatóval ellentétben a területi autokorreláció globális mutató száma. A Moran-féle I-index 1 értéket vesz fel abban az esetben,



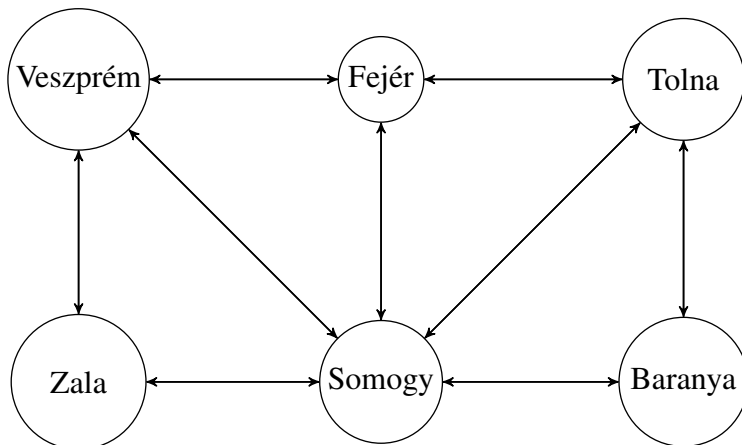
4.2. ábra. Szezonálisan differenciázott főkomponensek alapján színezett térkép

ha a szomszédos területi egységek által felvett érték tökéletesen pozitívan autokorrelált a vizsgált területi egység által felvett ismerv értékkel; ilyen függvényszerű a kapcsolat abban az esetben, ha minden érintkező megyében azonos lenne a megbetegedés szám. A Moran-féle I-index -1 értéket vesz fel abban az esetben, ha a szomszédos területi egységek által felvett ismerv értékek tökéletesen negatívan korreláltak egymással, negatívan területi autokorrelált például egy sakktábla bináris leképezése. A nulla értéket akkor veszi fel a mutató, amikor a megbetegedés számok egy adott időpontban függetlenek a szomszédos megyékben kialakuló megbetegedés számtól.

A területi kapcsolatokat írja le a súlymátrix, amelyet $\mathbf{W}$ jelöl, k területi egység esetében $k \times k$ méretű. A kapcsolat leírása sokféleképpen történhet: az érintkező felület nagyságával, az áthaladó forgalom nagyságával, vagy akár azzal, hogy a két terület érintkezik-e. A legelterjedtebb megoldás az, hogy a mátrixban 1 érték szerepel, ha két terület érintkezik, és 0 érték szerepel, ha két területi egységnek nincsen érintkező felülete. A mi esetünkben a keresztmetszeti adatfelvétel az adott heti megbetegedés szám egy megyében, a súly mátrix pedig a megyék közötti kapcsolatot írja le. A súlymátrix értelemszerűen egy szimmetrikus mátrix lesz, hiszen a kapcsolatok kölcsönösök, emellett a soraiban és oszlopaiban is csak 0 és 1 értékek szerepelnek. A keresztmetszeti elemzést ki lehet terjeszteni idő dimenzióba is, ha minden egyes hétre kiszá-

moljuk azt, hogy mekkora a területi autokorreláció mértéke.

A súlymátrix definiálását, az alábbi példával lehet szemléltetni; a Baranya, Tolna, Fejér, Veszprém, Zala és Somogy megyéket leíró kapcsolati hálózat:



4.3. ábra. Kapcsolati hálózat 6 megyére

Ekkor a $\mathbf{W}$ súlymátrix így adódik:

$$\mathbf{W} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 \end{pmatrix} \begin{matrix} \text{Somogy} \\ \text{Baranya} \\ \text{Tolna} \\ \text{Fejér} \\ \text{Veszprém} \\ \text{Zala} \end{matrix}$$

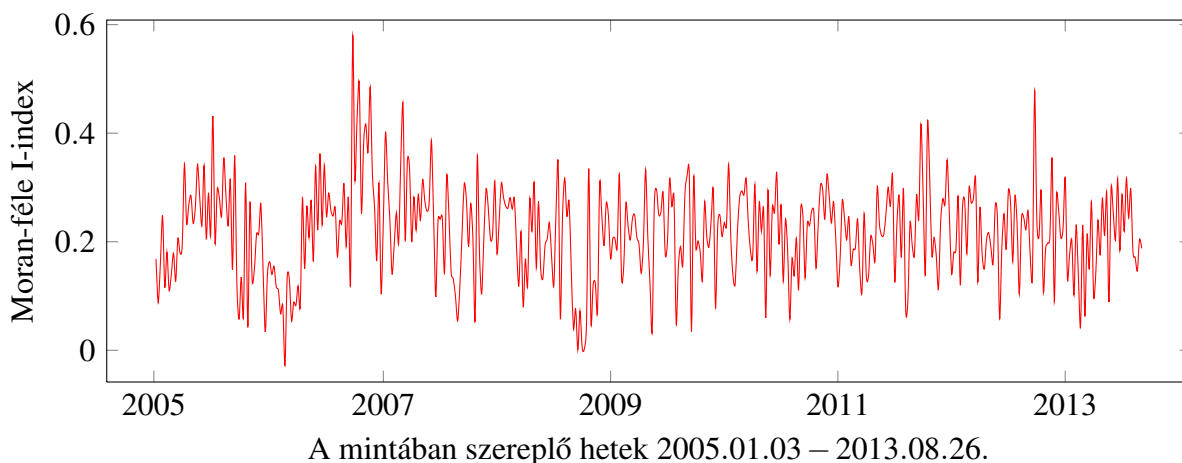
Ha az adatfelvételeket tartalmazó k hosszúságú vektort x jelöli, azaz k területi egységünk van, akkor a Moran-féle I-index így adódik, egy adott időpontra:

$$\mathcal{I} = \frac{k}{\sum_{i=1}^k \sum_{j=1}^k w_{ij}} \cdot \frac{\sum_{i=1}^k \sum_{j=1}^k w_{ij} \cdot (x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^k (x_i - \bar{x})^2} \quad (4.1)$$

A (4.1) képletben k a területi egységek száma, $\bar{x}$ az ismerv területi egységekre számolt számtani átlaga, w_{ij} a $\mathbf{W}$ súlymátrix i sorában és j oszlopában lévő eleme. Az x_i és x_j jelöli az ismerv megfelelő i -edik és j -edik index szerinti területegységhez tartozó értékét.

4.2.1. A Moran-féle I-index a megyei idősorokra

A számításokat a teljes hazai adatbázisra fogjuk elvégezni mind a 20 területi egységre, és mind a 452 időpontra. A 20 megye területi kapcsolatát leíró bináris mátrix megtalálható a függelék B.7 alfejezetében, a részletes sor címkékkel együtt. A számításokat végző *MATLAB* kód saját munka; a kódrészlet, amely minden egyes időpontra elvégzi a mutató számítását részletes magyarázattal együtt megtalálható a függelék B.8 alfejezetében. A mutató az idő függvényében szerepel a 4.4 ábrán.

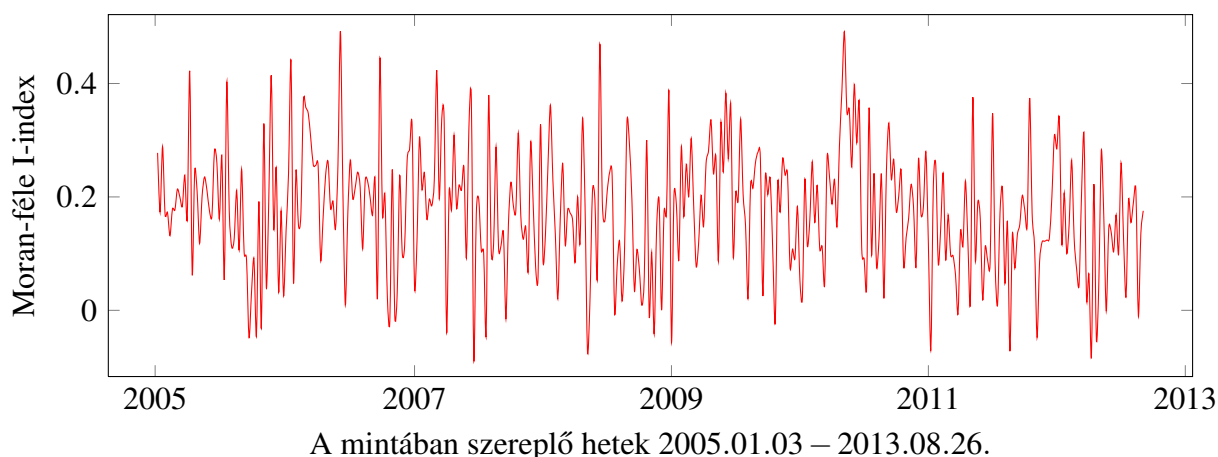


4.4. ábra. A Moran-féle I-index a megyei idősorokra

A Moran-féle mérték szerint a területi autokorreláció átlagos értéke a vizsgált időszakban $\overline{\rho_M} = 0.2176$ volt, azaz általában gyenge együttmozgás van az egymás mellett lévő megyék esetében. A legmagasabb területi autokorrelációs mérték $\rho_M = 0.5781$ volt, a legalacsonyabb pedig $\rho_M = -0.0281$ azaz jellemzően együtt mozgás van a szomszédos megyék között. A területi autokorrelációs index *ARMA* modellezhető, abban az esetben, ha a folyamat nem valamilyen egységgyök folyamat. Az Augmented-Dickey-Fuller tesztek alapján nem lesz egységgyök folyamat az indexsor. Ekkor az autokorrelációs függvény és a parciális autokorrelációs függvény alapján adódik, hogy az idősort egy *AR*(3) folyamat jól leírhatja. Az így kapott modell maradék tagja fehérzajnak tekinthetőek a korrelogram alapján. A kapott modell, az egységgyök tesztek eredményei és a maradék taghoz tartozó korrelogram a függelék B.9 részfejezetében megtalálhatóak. Azaz a területi korreláltság mértéke a három héttel korábbi területi korreláltság mértékétől függ, ez szintén alátámasztja a lappangási idővel kapcsolatos hipotézist, hiszen ha egy adott megye csoportban a megbetegedés szám három héttel korábban hasonlóan alakult, akkor a lappangási idő leteltével újra együtt fognak mozogni a megbetegedések.

4.2.2. A Moran-féle I-index a szezonálisan differenciázott megyei idősorokra

A számításokat ekkor a szezonálisan differenciázás miatt megrövidült adatbázisra fogjuk elvégezni mind a 20 területi egységre, és mind a 400 időpontra. A 20 megye területi kapcsolatát leíró bináris mátrix megtalálható a függelék B.7 alfejezetében, a részletes sor címkékkel együtt. A számításokat végző *MATLAB* kód esetében csak a bemeneti paramétereket kell megváltoztatni, a megváltozott kezdőértékeket megadó kódrészlet a függelék B.8 alfejezetében van. Az idő függvényében ábrázolt index szerepel a 4.5 ábrán.



4.5. ábra. A Moran-féle I-index a szezonálisan differenciázott megyei idősorokra

A Moran-féle mérték szerint a területi autokorreláció átlagos értéke a vizsgált időszakban $\overline{\rho_M} = 0.1748$ volt, azaz általában gyenge együttmozgás van a szomszédos megyék esetében megfigyelt szezonálisan differenciák esetében. A legmagasabb területi autokorrelációs mérték $\rho_M = 0.4925$ volt, a legalacsonyabb pedig $\rho_M = -0.0877$ azaz jellemzően a szezonális differenciák szintjén is együtt mozgás van a szomszédos megyék között. A szezonális differenciák alapján számolt indexsor fehérzaj folyamat $\alpha = 1\%$ szignifikancia szint mellett, ezért ez az idősor értelmes módon nem *ARMA* modellezhető.

4.3. A Geary-féle C-index

A Moran-féle I-index a területi autokorrelációt globálisan méri, és nem érzékeny a csak bizonyos egymással szomszédos területek esetében megjelenő területi autokorrelációra. A Geary-féle C-mutató is csak adott időpontra vonatkozó adatok esetében értelmezhető. A mutató kiszámításához szintén szükséges egy súlymátrix, ugyanúgy mint a másik területi autokorrelációs

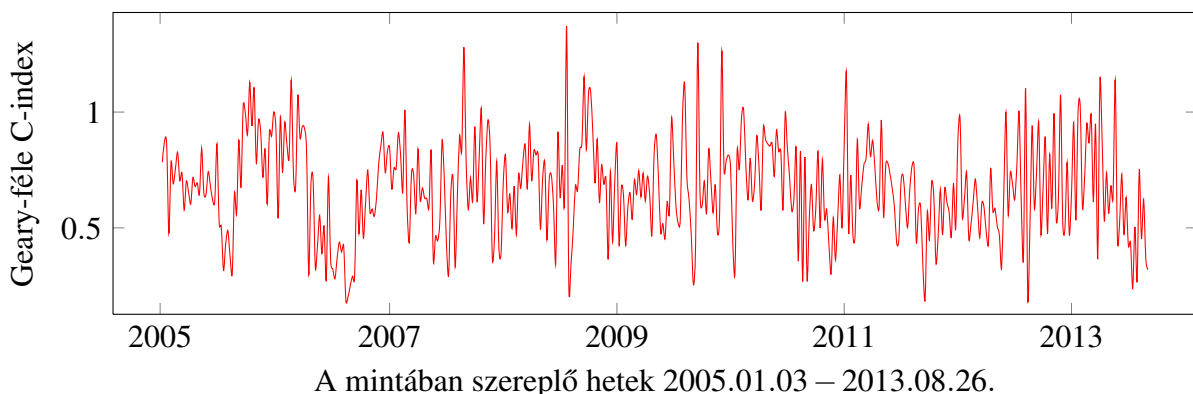
mutató esetében; a sor vagy oszlop szerinti normázás itt sem szükséges. A területi autokorreláció negatív, ha a $1 < \mathcal{C} \leq 2$ reláció áll, és pozitív abban az esetben, ha a mutatóra a $0 \leq \mathcal{C} < 1$ reláció áll, nincsen terület autokorreláció $\mathcal{C} = 1$ esetén.

$$\mathcal{C} = \frac{(k-1) \cdot \sum_{i=1}^k \sum_{j=1}^k w_{ij} \cdot (x_i - x_j)^2}{2 \cdot \left(\sum_{i=1}^k \sum_{j=1}^k w_{ij} \right) \cdot \sum_{i=1}^k (x_i - \bar{x})^2} \quad (4.2)$$

A (4.2) képletben k a területi egységek száma, $\bar{x}$ az ismerv területi egységekre számolt számtani átlaga, w_{ij} a súlymátrix i sorában és j oszlopában lévő eleme. Az x_i és x_j jelöli az ismerv megfelelő i -edik és j -edik index szerinti területegységhez tartozó értékét. A mutatószám számítását minden időszakra elvégző *MATLAB* kód saját munka; a függelék B.10 alfejezetében megtalálható azon kódrészletekkel együtt, amelyek az idősorokat betöltik. A felhasznált súlymátrix a megyék közötti bináris kapcsolatokat leíró mátrix marad, amely megtalálható a melléklet B.7 alfejezetében.

4.3.1. A Geary-féle C-index az idősorokra

A Geary-féle C-index a teljes megyékre lebontott idősorra kiszámolható. Ebben az esetben, ha az idő függvényében ábrázolva a mutatót, megkapjuk a 4.6 ábrát.



4.6. ábra. A Geary-féle C-index a megyei idősorokra

A mutató átlagos értéke a mintán $\overline{\rho_C} = 0.6705$. Ez közepesnél gyengébb pozitív területi autokorrelációt mutat, azaz a szomszédos megyékben a megbetegedés számok enyhén együtt mozognak, egy a kisebb csoportok együtt mozgására érzékeny mutató alapján is. A mutató minimuma, a $\rho_C = 0.1861$ érték, a maximuma pedig a $\rho_C = 1.3707$, ez alapján adódik, hogy az értékek egy nagyobb intervallumban szóródnak, a mutató nagyobb kilengését a kisebb csoportok erős területi korrelációja is okozhatja. Az idősor ebben az esetben is egy stacioner $AR(3)$

folyamatot követ, ez újfent arra utal, hogy a területi autokorreláció aktuális értéke függ a 3 héttel korábbi területi autokorreláció értékétől, ami a 3 héttel korábbi megyei megbetegedés számoktól, azaz indirekt módon megint a lappangási idő értékét kaptuk vissza.

4.4. A Bayes-i vektor autoregresszió

A több változós megközelítés miatt a megyék közötti kapcsolatot vektor autoregresszióval is meg lehetne vizsgálni. Az ökonometriában általánosan alkalmazott vektor autoregresszió azonban a tér és idő dimenzió vizsgálatára nem feltétlenül alkalmas, LeSage (2003) több érvet is felhoz a vektor autoregresszió alkalmazása ellen. Ezek a problémák a mi esetünkben is fennállnak. Abban az esetben, ha minden egyenletbe, minden megyére vonatkozóan 3 nem szezonális autoregresszív és 1 szezonális autoregresszív tagot építünk be, akkor összesen $p = 4 \times 20 \times 20$ paramétert kell megbecsülnünk, feltételezve, hogy nem alkalmazunk semmilyen tengelymetszetet.¹ A felhasznált szezonális paraméterek hatására ráadásul a meglévő $n = 452$ hosszúságú idősorok megrövidülnek egy teljes évre vonatkozó hét számnak megfelelően, és csak $n = 400$ hosszúságúak lesznek. Emellett az is feltételezhető, hogy nem feltétlenül mozog együtt két egymástól távol lévő megyében a szezonaritástól valamilyen módszerrel megszűrt megbetegedés szám.

A vizsgálat tárgyát is képező tér és idő dimenzióra kiterjedő adatsorok kapcsolatának vizsgálatára alkalmas lehet a priorokon alapuló Bayes-i vektor autoregresszió egy változata. A priorokon alapuló Bayes-i vektor autoregressziós modellek, amelyeket Doan et al. (1984) vezetett be, ugyanis az alábbi számunkra előnyös tulajdonságokkal rendelkeznek:

1. Már kezdetben redukáltak, nem minden térbeli kapcsolatot modelleznek.
2. Kevesebb paraméter becslése történik.
3. A térbeli kapcsolatok súlyként felhasználhatóak.

Az egyenletekbe bevont késleltetett magyarázóváltozókhoz rendelt autoregresszív paramétereket két csoportba osztjuk:

$$\beta_n \sim N(1, \sigma_{\beta_n}^2) \quad \beta_m \sim N(0, \sigma_{\beta_m}^2)$$

Minden egyenletben β_n típusú paraméter tartozik azokhoz a változók, amelyeket relevánsnak feltételezünk, ezek esetében a paraméter normális eloszlik, várhatóértéke 1 és $\sigma_{\beta_n}^2$ a varianciája.

¹Tengelymetszet alkalmazásakor $p = 5 \times 20 \times 20 = 2000$ paraméter becslünk meg.

Az általunk nem relevánsnak tartott változók pedig β_m típusú paraméterrel rendelkeznek; ezek paraméterek is normális eloszlással rendelkeznek a várhatóértékük 0 a varianciájuk pedig $\sigma_{\beta_m}^2$. A célunk a priorok szórásának becslése; amely egy adott egyenlet, megadott változó szerinti késleltetésére az alábbi módon történik (Doan et al., 1984):

$$\sigma_{ijk} = \Theta \cdot w(i, j) \cdot k^{-\phi} \cdot \left(\frac{\hat{\sigma}_{uj}}{\hat{\sigma}_{ui}} \right) \quad (4.3)$$

A (4.3) egyenletben σ_{ijk} jelöli az i változó j vektor autoregressziós egyenletben szereplő k késleltetéséhez tartozó szórását. A Θ skalár egy mesterségesen bevezetett úgynevezett *tightness* hiperparamétert jelöl, $w(i, j)$ egy adott $\mathbf{W}$ súlymátrix megfelelő eleme,² k a megadott késleltetésszám, a ϕ a mesterséges *decay* hiperparaméter, amely az autoregresszív folyamat lecsengését kontrollálja. A $\hat{\sigma}_{ui}$ becsült szórás az egyváltozós i változóra felírt $AR(1)$ modell paraméter becslésének hibája, $\hat{\sigma}_{uj}$ ugyanezen hiba a j változóra felírt $AR(1)$ modell esetében.

A hiperparaméterek, a súlymátrix és egyváltozós modellekhez tartozó szórások alapján tehát kiszámítható a priorok varianciája. A modell parametrizáció során legnagyobb mozgásterünk a súlymátrix megválasztásakor van; ugyanis a *decay* és *tightness* paramétereket értékét a legtöbb esetben az úgynevezett Minnesota-priorok alapján választják, amelyek a $\Theta = 0.1$ és $\phi = 1$ kezdő parametrizációt jelentik. A 4.2 alfejezetben szereplő példában a súlymátrix azonos súlyokat rendel a szomszédos megyékhez és a viszonyítási megyéhez, ez a Minnesota-prioroknak nem jellemzője (LeSage, Krivelyova, 1997). A becslésekhez módosított súlymátrixokat fogunk alkalmazni, a 0 súlyokat 0.1 súlyokra cseréljük:

$$\mathbf{W} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 \end{pmatrix} \sim \widehat{\mathbf{W}} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0.1 & 0.1 & 0.1 \\ 1 & 1 & 1 & 1 & 0.1 & 0.1 \\ 1 & 0.1 & 1 & 1 & 1 & 0.1 \\ 1 & 0.1 & 0.1 & 1 & 1 & 1 \\ 1 & 0.1 & 0.1 & 0.1 & 1 & 1 \end{pmatrix}$$

A Bayes-i vektor autoregressziós modellezés során a $k = 1, \dots, 6$ késleltetés tartalmazó modelleket vizsgáljuk, a felhasznált súlymátrix pedig az a bináris mátrix, amely a megyék közötti szomszédosságot jellemzi, és megtalálható a függelék B.7 alfejezetében. A magasabb késleltetés szám mellett a program lelassul nem becsli meg a paramétereket. A két skalár hiperparaméter értékét a Minnesota-priorok alapján választjuk meg, és a modellt a szezonálisan differenciázott 400 hét hosszúságú idősorokra írjuk fel. A becslések a LeSage (2003), által leírt *MATLAB* kóddal történnek. A parametrizációt elvégző kódrészlet megtalálható a függelék

²Maga a $\mathbf{W}$ súlymátrix is mesterséges hiperparaméter.

B.11 alfejezetében. A modell így is rendkívül magas paraméter számot becsül, emiatt a részletes outputokat a mellékletben nem csatoltam. A legtöbb szignifikáns kapcsolat a 3 késleltetést tartalmazó modellek esetében van. Az általunk választott $\alpha = 1\%$ szignifikancia szint mellett, a parciális próbák alapján az alábbi szignifikáns kapcsolatok vannak a szezonális differenciákra nézve, a változók saját késleltetéseit nem vizsgálva, baloldalon a függő, jobboldalon a magyarázó változók szerepelnek:

$$\begin{aligned}
 &\text{Budapest}_t - \text{Pest}_{t-1}, \text{Pest}_{t-2} \\
 &\text{Baranya}_t - \text{Somogy}_{t-1} \\
 &\text{Bács-Kiskun}_t - \text{Baranya}_{t-1} \\
 &\text{Borsod-Abaúj-Zemplén}_t - \text{Heves}_{t-1} \\
 &\text{Fejér}_t - \text{Tolna}_{t-1} \\
 &\text{Győr}_t - \text{Veszprém}_{t-1} \\
 &\text{Heves}_t - \text{Nógrád}_{t-1} \\
 &\text{Jász-Nagykun-Szolnok}_t - \text{Békés}_{t-1}, \text{Pest}_{t-1} \\
 &\text{Komárom}_t - \text{Fejér}_{t-1} \\
 &\text{Pest}_t - \text{Budapest}_{t-1}, \text{Budapest}_{t-2}, \text{Fejér}_{t-1} \\
 &\text{Tolna}_t - \text{Fejér}_{t-1}, \text{Somogy}_{t-1} \\
 &\text{Veszprém}_t - \text{Somogy}_{t-1}, \text{Vas}_{t-1}, \text{Zala}_{t-1}
 \end{aligned}$$

Azaz a prior szórásokon alapuló modell esetében a szignifikáns autoregresszív tagok az egyenletekben a szomszédos megyékből származnak. A kapcsolatok egy része azonban nem kölcsönös, aszimmetrikus kapcsolat van például Borsod-Abaúj-Zemplén megye és Heves megye között. A megyék saját autoregresszív késleltetéseitől az 1., 2. és egyes esetekben a 3. volt szignifikáns. Emiatt a modell a területi autokorreláció meglétéről szóló hipotézis mellett a lapangási idővel kapcsolatos hipotézist is alátámasztja.

5. fejezet

Összefoglalás

A dolgozatban három kutatási kérdésre kerestük a választ. Az előzetes feltételezésünk az volt, hogy a kérdésekre adott válasz minden esetben igen; ez az elemzések után sem változott meg. A kérdésekre adott hipotetikus válaszokat az alább eredmények támasztják alá:

1. kérdés – *Az iskola és óvoda kezdés, illetve a nyári szünet hatása kimutatható-e az országos idősorban?*

A nemlineáris legkisebb négyzetek módszerével becsült legerősebben illeszkedő trigonometrikus modell minimuma pont az iskolakezdés előtt van. Ugyanezen modell esetében a hetek száma szerinti rugalmasság alapján megállapítható, hogy a megbetegedések száma leggyorsabban a tanítási időszak végével csökken. A leggyorsabb, ugrásszerű növekedés a megbetegedés számban a nyári szünet utáni első hetekben történik. A trigonometrikus Poisson-regresszió módszerével becsült modell esetében is ugyanezeket tapasztaljuk, mind a szélsőértékek, mind a rugalmasságok esetében.

2. kérdés – *A lappangási időnek megfelelő sztochasztikus hatások vannak-e jelen az országos idősorban?*

A bárányhimlő lappangási ideje 14 és 21 nap között mozog, a heti frekvenciájú idősorokban az ennek megfelelő hatás 2 és 3 hetes sztochasztikus késleltetések formájában kell, hogy megjelenjen. A nemlineáris legkisebb négyzetek módszerével becsült legerősebben illeszkedő trigonometrikus modell maradékára illesztett sztochasztikus modell esetében ez teljesül. Az illesztett autoregresszív és mozgóátlagolású tagok esetében az ideális késleltetés szám 2 hét. A tisztán sztochasztikus modell esetében a nem szezonális tagok késleltetése 3 hét. A fehérzaj maradékkal rendelkező Poisson-regressziós modell esetében szintén 3 hétre visszamenő késleltetéseket szerepeltetünk. A területi autokorrelációs mutatók a 4. fejezetben szintén 3 késleltetésig voltak autokorreláltak. A Bayes-i vektor autoregressziós modellek esetében is az 1, 2 és 3 héttel korábbi megbetegedés számok

hatása volt meghatározó a tárgyidőszaki megbetegedés számra nézve, ha a szezonalitást kiszűrtük.

3. kérdés – *Kimutatható-e területi alapú együttmozgás a megyékre lebontott idősorokban?*

A szezonálisan differenciázatlan idősorokra számított főkomponensek alapján a megyék két csoportra oszthatóak. A szezonálisan differenciázott idősorokat vizsgálva a megyékből kisebb csoportok alakultak ki, ahol együtt mozgás volt, ezek a csoportok többségében olyan megyékből alakultak ki, amelyek rendelkeznek közös határfelülettel. A globális és lokális területi autokorrelációt jellemző indexek alapján enyhe pozitív területi autokorreláció jellemzi mind a megbetegedés számok alakulását, mind a szezonálisan differenciázott megbetegedés szám alakulását a 15 év alatt. Bayes-i vektor autoregressziós modellek alapján adódik, hogy egy adott megyében történő megbetegedés számra a megyén belüli megbetegedés számon kívül, a szomszédos megyék megbetegedés számai gyakorolhatnak hatást. Az ilyen területi autokorrelációs kapcsolatok azonban, nem minden esetben kölcsönösek.

A dolgozatban emellett több kisebb, önálló módszertani hozzájárulás is van. A 2.3 alfejezetben található általános alakban megadott nemlineáris Poisson-regresszió levezetés alapján tetszőleges nemlineáris Poisson-regressziós modell definiálható. A becsléshez pedig felhasználható a függelék A.16 és A.17 alfejezetében található *Gretl* kódrészlet. A Moran-féle I-index és a Geary-féle C-index számításához készített *MATLAB* kód, amely megtalálható a függelék B.8 és B.10 alfejezetében tömbfájlok meghívásával is futtatható, és idősorok területi autokorrelációjának elemzésére is alkalmas.

5.1. További kutatási lehetőségek

Az adatbázis és az Epiinfoban megtalálható egyéb adatbázisok alapján további elemzéseket is el lehet végezni, ezen elemzési lehetőségek közül néhány:

1. Egyéb nemlineáris modellspecifikációk az országos idősorra nézve.
2. A nem mozgó ünnepekhez kapcsolódó tanítási szünetek hatásának vizsgálata.
3. A mozgó ünnepekhez kapcsolódó tanítási szünetek hatásának vizsgálata.
4. Más rotációt alkalmazó főkomponens elemzések.
5. Eltérő súlymátrixok alkalmazás a területi autokorrelációs indexek számítása során.

6. Területi autokorreláció előrejelzése, a területi autokorrelációs modellekre felírt autoregresszív modellekkel.
7. A Bayes-i modellek más hiperparaméterekkel történő alkalmazása.
8. Előrejelzések a megyékre a Bayes-i vektor autoregressziós modellek segítségével.
9. Nemlineáris többváltozós modellek alkalmazása.
10. A budapesti *HIV* és tuberkulózis fertőzések vizsgálata.

6. fejezet

Irodalomjegyzék

- Ralf Bender. Introduction to the Use of Regression Models in Epidemiology. *Methods of Molecular Biology*, 471:179–95, 2009.
- George Box, Gwilym Jenkins. *Time Series Analysis: Forecasting and Control*. Wiley, 10 edition, 1970.
- Trevor Breusch, Adrian Pagan. The Lagrange Multiplier Test and its Applications to Model Specification in Econometrics. *The review of economic studies*, 47(1):239–253, 1980.
- Charles George Broyden. The Convergence of a Class of Double-Rank Minimization Algorithms. *Journal of the Institute of Mathematics and Its Applications*, 6:76–90, 1970.
- Finkenstädt Bärbel, Bryan Grenfell. Time Series Modelling of Childhood Diseases: a Dynamical Systems Approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 49:187–205, 2000.
- Colin Cameron, Pravin Trivedi. *Microeconometrics: Methods and Applications*. Cambridge University Press, 2005.
- Fabio Canova. Three Tests for the Existence of Cycles in Time Series. *Ricerche Economiche*, 50:135–162, 1996.
- Joseph Carol, Norman Noah. Epidemiology of Chickenpox in England and Wales, 1967-85. *British medical journal (Clinical research ed.)*, 27:673–689, 1988.
- Johnny Chan, Linwei Tian. Hospitalizations for Varicella in Children and Adolescents in a Referral Hospital in Hong Kong, 2004 to 2008: A Time Series Study. *BMC Public Health*, 11:366–375, 2011.
- Péter Csepe. HIV Risk Behaviour Among Gay and Bisexual Men in Budapest, Hungary. *International journal of STD & AIDS*, 13:192–200, 2002.
- Edward Cureton, Ralph D’Agostino. *Factor Analysis: an Applied Approach*. Hillside, 1983.

- David Alan Dickey, Yilmaz Akdi. Periodogram For Seasonal Time Series With a Unit Root. *Istatistik - Journal of the turkish statistical association*, 2:153–164, 1999.
- David Alan Dickey, Wayne Arthur Fuller. *Introduction to Statistical Time Series*. Wiley, 3 edition, 1976.
- David Alan Dickey, Wayne Arthur Fuller, David Hasza. Testing for Unit Roots in Seasonal Time Series. *Journal of the American Statistical Association*, 79:355–367, 1984.
- Thomas Doan, Robert Litterman, Christopher Sims. Forecasting and Conditional Projections Using Realistic Prior Distributions. *Econometric reviews*, 3:1–100, 1984.
- Roger Fletcher. A New Approach to Variable Metric Algorithms. *Computer Journal*, 13:317–322, 1970.
- Graham Galbraith. *Encyclopedia of Biostatistics: Trigonometric Regression*. Wiley and Sons, 2005.
- Daniel Goldfarb. A Family of Variable Metric Updates Derived by Variational Means. *Mathematics of Computation*, 24:23–26, 1970.
- William Greene. *Econometric Analysis*. Prentice-Hall, 2003.
- Ottó Hajdú. Statisztikai adatok ökonometriai elemzése, Egyetemi jegyzet 2. rész, Idősori modellek elmélete és alkalmazása, 2013.
- Roselyne Joyeux. Tests For Seasonal Cointegration Using Principal Components. *Journal of Time Series Analysis*, 13:109–118, 1992.
- Henry Kaiser. The Varimax Criterion for Analytic Rotation in Factor Analysis. *Psychometrika*, 23:187–200, 1958.
- Patricia Kelly. Isolation and Stigma: the Experience of Patients with Active Tuberculosis. *Journal of community health nursing*, 16:233–241, 1999.
- Erzsébet Kovács, Borbála Szüle, Vilmos Fliszár, Péter Vékás. *Pénzügyi adatok statisztikai elemzése: Egyetemi tankönyv*. Tanszék Kft., 2011.
- Gergely Kőhegyi. *Közgazdaságtan nem csak szemüvegeseknek*. ELTECON, 2013.
- Charles Lawson, Richard Hanson. *Solving Least Squares Problems*. Prentice–Hall, 1974.
- Mark Leary, Lisa Schreindorfer. The Stigmatization of HIV and AIDS: Rubbing Salt in the Wound. *HIV and social interaction*, 1:12–29, 1998.
- James LeSage. *Spatial econometrics*, 2003.
- James LeSage, Anna Krivelyova. A Spatial Prior for Bayesian Vector Autoregressive Models. *Journal of Regional Science*, 39:297–317, 1997.

- Greta Ljung, George Box. On a Measure of a Lack of Fit in Time Series Models. *Biometrika*, 65(2): 297–303, 1978.
- Kuhn Louise, Leslie Davidson, Maureen Durkin. Use of Poisson Regression and Time Series Analysis for Detecting Changes Over Time in Rates of Child Infection Following a Prevention Program. *American Journal of Epidemiology*, 10:943–955, 1994.
- Henrik Madsen, Poul Thyregod. *Introduction to General and Generalized Linear Models*. Chapman & Hall/CRC, 2011.
- Pat Moran. Notes on Continuous Stochastic Phenomena. *Biometrika*, 37:17–23, 1950.
- Niklaus Mueller, Donald Gilden, Randall Cohrs, Ravi Mahalingam, Maria Nagel. Varicella Zoster Virus Infection: Clinical Features, Molecular Pathogenesis of Disease, and Latency. *The Journal of Clinical Neurology*, 26:675–697, 2008.
- Joseph Newton. *TIMESLAB: A Time Series Analysis Laboratory*. Wadsworth & Brooks/Cole, 1994.
- OEK. Einfo 1998-2000. <http://www.oek.hu/>, November 2013.
- OEK. Einfo. <http://rozemberczki.wordpress.com/>, September 2014.
- Stefano Parodi, Ezio Bottarelli. Poisson Regression Model in Epidemiology - An Introduction. *Annals of the Faculty of Medicine Veterinary in Parma*, 26:25–44, 2006.
- Napoleon Perez-Farinos, Maria Ordobas. Varicella and Herpes Zoster in Madrid, Based on the Sentinel General Practitioner Network: 1997-2004. *BMC Public Health*, 7:59–65, 2007.
- Miklós Pintér. A Regressziós Játékok Alkalmazása Modellszelekcióra. 2006.
- Miklós Pintér. Regression Games. *Annals of Operations Research*, 186(1):263–274, 2011.
- Miklós Pintér. Valószínűségszámítás Jegyzet, 2012. URL <https://sites.google.com/site/miklospinter/home/okt>.
- Deguen Severine, Nguyen Phong Chau, Antoine Flahault. Epidemiology of Chickenpox in France. *Journal of epidemiology and community health*, 52:46–49, 1998.
- David Shanno, Paul Kettler. Optimal Conditioning of Quasi-Newton Methods. *Mathematics of Computation*, 24:657–664, 1970.
- Abraham Wald. Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large. *Transactions of the American Mathematical Society*, 54:426–482, 1943.

A. Függelék

Egyváltozós modellezés

A.1. Ljung-Box Q-statisztikák az összesített idősorra

Késleltetés	Ljung-Box Q-statisztikák			
	Autokorrelációs együttható ρ	Parciális autokorrelációs együttható $\rho(h)$	Q-statisztika	p -érték
1.	0.8509	0.8509	585.7203	0.000
2.	0.8350	0.4023	1150.5323	0.000
3.	0.7937	0.1154	1661.4230	0.000
4.	0.7455	-0.0282	2112.7072	0.000
5.	0.6722	-0.1685	2480.0648	0.000
6.	0.5953	-0.1879	2768.5239	0.000
7.	0.5259	-0.1017	2993.9627	0.000
8.	0.4308	-0.1549	3145.4522	0.000
9.	0.3419	-0.1200	3240.9687	0.000
10.	0.2546	-0.0691	3294.0229	0.000
11.	0.1768	-0.0007	3319.6207	0.000
12.	0.0668	-0.1338	3323.2822	0.000
13.	-0.0231	-0.1034	3323.7200	0.000
14.	-0.1064	-0.0547	3333.0335	0.000
15.	-0.1862	-0.0269	3361.5749	0.000

A.1. táblázat. Az összesített idősorhoz tartozó Q-statisztikák

A.2. Augmented Dickey-Fuller tesztek eredményei

Az összesített idősorra alkalmazott Augmented Dickey-Fuller teszt részletes eredményei:

```
1 Augmented Dickey-Fuller test for Osszes
2 including 19 lags of (1-L)Osszes (max was 20)
3 sample size 786
4 unit-root null hypothesis: a = 1
```

```
1 test with constant
2 model: (1-L)y = b0 + (a-1)*y(-1) + ... + e
3 1st-order autocorrelation coeff. for e: -0.002
4 lagged differences: F(19, 765) = 17.757 [0.0000]
5 estimated value of (a - 1): -0.264449
6 test statistic: tau_c(1) = -8.30666
7 asymptotic p-value 6.181e-14
```

```
1 with constant and trend
2 model: (1-L)y = b0 + b1*t + (a-1)*y(-1) + ... + e
3 1st-order autocorrelation coeff. for e: -0.003
4 lagged differences: F(19, 764) = 17.816 [0.0000]
5 estimated value of (a - 1): -0.267881
6 test statistic: tau_ct(1) = -8.38264
7 asymptotic p-value 9.865e-14
```

```
1 with constant and quadratic trend
2 model: (1-L)y = b0 + b1*t + b2*t^2 + (a-1)*y(-1) + ... + e
3 1st-order autocorrelation coeff. for e: -0.003
4 lagged differences: F(19, 763) = 17.879 [0.0000]
5 estimated value of (a - 1): -0.274177
6 test statistic: tau_ctt(1) = -8.47944
7 asymptotic p-value 6.598e-26
```

A.3. Az országos idősorhoz tartozó periodogram

Körfrekvencia (ω_i)	Skálázott frekvencia (f_i)	Periódus (T_i)	Spektrális sűrűség
0.0078	1	806	$1.65 \cdot 10^5$
0.01559	2	403	$2.73 \cdot 10^5$
0.02339	3	268.67	$1.63 \cdot 10^5$
0.03118	4	201.5	3746.4
0.03898	5	161.2	81383
0.04677	6	134.33	9044
0.05457	7	115.14	43666
0.06236	8	100.75	$1.52 \cdot 10^5$
0.07016	9	89.56	34534
0.07796	10	80.6	37932
0.08575	11	73.27	48269
0.09355	12	67.17	$1.17 \cdot 10^5$
0.10134	13	62	$2.21 \cdot 10^5$
0.10914	14	57.57	$1.25 \cdot 10^5$
0.11693	15	53.73	$7.62 \cdot 10^6$
0.12473	16	50.38	$3.79 \cdot 10^6$
0.13252	17	47.41	$3.31 \cdot 10^5$
0.14032	18	44.78	$4.80 \cdot 10^5$
0.14811	19	42.42	56899
0.15591	20	40.3	37403
0.16371	21	38.38	$1.53 \cdot 10^5$
0.1715	22	36.64	2712.2
0.1793	23	35.04	68743
0.18709	24	33.58	48088
0.19489	25	32.24	9104.3
0.20268	26	31	$1.02 \cdot 10^5$
0.21048	27	29.85	57185
0.21827	28	28.79	49126
0.22607	29	27.79	3912.1
0.23387	30	26.87	17978
0.24166	31	26	$1.01 \cdot 10^6$

A.2. táblázat. Az országos adatokra vonatkozó periodogram értékei

A.4. Periodikus dummy változós OLS modell jellemzői

	Paraméter	Std. hiba	<i>t</i> -próba	<i>p</i> -érték
D_1	1130.70	32.066	35.260	0.000
D_2	1185.09	32.312	36.676	0.000
D_3	1331.95	32.564	40.903	0.000
D_4	1336.22	31.362	42.607	0.000
D_5	1221.41	31.137	39.227	0.000
D_6	1070.72	31.362	34.141	0.000
D_7	440.32	31.362	14.040	0.000
D_8	103.98	30.490	3.411	0.000
D_9	77.77	32.066	2.425	0.015
D_{10}	286.45	32.312	8.864	0.000
D_{11}	612.80	32.312	18.965	0.000
D_{12}	859.18	32.312	26.590	0.000
Függő változó átlaga	801.3449	Függő változó std. σ	528.1449	
SSR	53885253	Regresszió std. hiba	260.5102	
R^2	0.9273909	Korrigált R^2	0.926384	
$F(12, 794)$	845.09601	<i>p</i> -érték (F)	0.000000	
Log-likelihood	-5621.109	Akaike kritérium	11266.22	
Schwarz kritérium	11322.52	Hannan–Quinn	11287.84	
$\hat{\rho}$	0.360512	Durbin–Watson	1.278914	

A.3. táblázat. Periodikus dummy változós OLS modell jellemzői

A.5. Az I. NLS modell *Gretl* optimalizációs kódja

```
1 genr alpha = 800 % Alfa deklaráció
2 genr beta1 = 800 % Béta 1 deklaráció
3 genr gamma1 = -10 % Gamma 1 deklaráció
4 genr psi1 = 10 % Pszi 1 deklaráció
5 nls Osszes = alpha+beta1*sin(t/psi1-gamma1) % Választott függvényforma
6 deriv alpha = 1 % Alfa der.
7 deriv beta1 = sin(t/psi1-gamma1) % Béta 1 der.
8 deriv gamma1 = -beta1*cos(t/psi1-gamma1) % Gamma 1 der.
9 deriv psi1 = -beta1*cos(t/psi1-gamma1)*(t)*psi1^(-2) % Pszi 1 der.
10 end nls % NLS ciklus lezárása
```

A.6. A II. NLS modell *Gretl* optimalizációs kódja

```
1 genr alpha = 800 % Alfa deklaráció
2 genr beta2 = -200 % Béta 2 deklaráció
3 genr gamma2 = 15 % Gamma 2 deklaráció
4 genr psi2 = -5 % Pszi 2 deklaráció
5 nls Osszes = alpha+beta2*sin(t/psi2-gamma2) % Választott függvényforma
6 deriv alpha = 1 % Alfa der.
7 deriv beta2 = sin(t/psi2-gamma2) % Béta 2 der.
8 deriv gamma2 = -beta2*cos(t/psi2-gamma2) % Gamma 1 der.
9 deriv psi1 = -beta2*cos(t/psi2-gamma2)*(t)*psi2^(-2) % Pszi 1 der.
10 end nls % NLS ciklus lezárása
```

A.7. A III. NLS modell *Gretl* optimalizációs kódja

```
1 genr alpha = 10 % Alfa deklaráció
2 genr beta1 = 10 % Béta 1 deklaráció
3 genr gamma1 = 10 % Gamma 1 deklaráció
4 genr psi1 = 10 % Pszi 1 deklaráció
5 genr beta2 = 10 % Béta 2 deklaráció
6 genr gamma2 = 10 % Gamma 2 deklaráció
7 genr psi2 = 10 % Pszi 2 deklaráció
8 nls Y = alpha+beta1*sin(t/psi1-gamma1)+beta2*sin(t/psi2-gamma2)
9 deriv alpha = 1 % Alfa der.
10 deriv beta1 = sin(t/psi1-gamma1) % Béta 1 der.
11 deriv gamma1 = -beta1*cos(t/psi1-gamma1) % Gamma 1 der.
12 deriv psi1 = -beta1*cos(t/psi1-gamma1)*(t)*psi1^(-2) % Pszi 1 der.
13 deriv beta2 = sin(t/psi2-gamma2) % Béta 2 der.
14 deriv gamma2 = -beta2*cos(t/psi2-gamma2) % Gamma 2 der.
15 deriv psi2 = -beta2*cos(t/psi2-gamma2)*(t)*psi2^(-2) % Pszi 2 der.
16 end nls % NLS ciklus lezárása
```

A.8. Ljung-Box Q-statisztikák a III. NLS modell maradékaira

Késleltetés	Ljung-Box Q-statisztikák			
	Autokorrelációs együttható ρ	Parciális autokorrelációs együttható $\rho(h)$	Q-statisztika	p -érték
1.	0.4439	0.4439	159.3799	0.000
2.	0.4687	0.3384	337.3586	0.000
3.	0.4443	0.2200	497.4259	0.000
4.	0.4364	0.1670	652.0696	0.000
5.	0.3775	0.0554	767.9130	0.000
6.	0.3223	-0.0186	852.4588	0.000
7.	0.3314	0.0385	941.9513	0.000
8.	0.2550	-0.0482	995.0071	0.000
9.	0.2269	-0.0390	1037.0876	0.000
10.	0.2149	0.0024	1074.8729	0.000
11.	0.2315	0.0630	1118.7757	0.000
12.	0.1426	-0.0572	1135.4633	0.000
13.	0.1236	-0.0447	1148.0011	0.000
14.	0.1137	-0.0198	1158.6337	0.000
15.	0.1045	0.0038	1167.6314	0.000

A.4. táblázat. A III. NLS modell maradék tagjához tartozó Q-statisztikák

A.9. ARMA modellek a III. NLS modell maradékaira

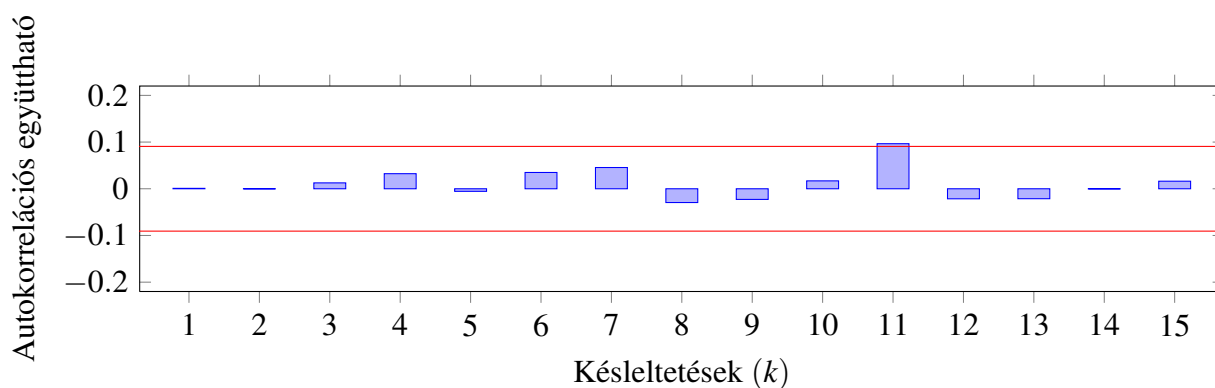
Modell	ARMA modell információk kritériumok			
	Log-likelihood	Akaike krit.	Schwarz krit.	Hannan–Q krit.
<i>AR</i> (1)	-5556.499	11119.00	11133.07	11124.40
<i>AR</i> (2)	-5507.591	11023.18	11041.95	11030.39
<i>AR</i> (3)	-5487.681	10985.36	11008.82	10994.37
<i>AR</i> (4)	-5476.281	10964.56	10992.72	10975.37
<i>AR</i> (5)	-5475.073	10964.15	10996.99	10976.76
<i>MA</i> (1)	-5592.666	11191.33	11205.41	11196.74
<i>MA</i> (2)	-5558.232	11124.46	11143.23	11131.67
<i>MA</i> (3)	-5537.393	11084.79	11108.25	11093.79
<i>MA</i> (4)	-5516.924	11045.85	11074.00	11056.66
<i>MA</i> (5)	-5502.205	-11018.41	11051.26	11031.02
<i>ARMA</i> (1,1)	-5482.777	10973.55	10992.32	10980.76
<i>ARMA</i> (2,2)	-5475.575	10963.15	10991.30	10973.96
<i>ARMA</i> (3,3)	-5475.552	10967.10	11004.64	10981.52
<i>ARMA</i> (4,4)	-5469.185	10958.37	11005.29	10976.39
<i>ARMA</i> (5,5)	-5465.524	10955.05	11011.35	10976.67

A.5. táblázat. ARMA modellek a III. NLS modell maradékaira

A.10. Ljung-Box Q-statisztikák a kombinált modell maradékaira

Késleltetés	Ljung-Box Q-statisztikák			
	Autokorrelációs együttható ρ	Parciális autokorrelációs együttható $\rho(h)$	Q-statisztika	p-érték
1.	0.0009	0.0009	0.0006	0.980
2.	-0.0002	-0.0002	0.0007	1.000
3.	-0.0126	-0.0126	0.1293	0.988
4.	0.0323	0.0323	0.9768	0.913
5.	-0.0055	-0.0056	1.0014	0.962
6.	-0.0349	-0.0351	1.9924	0.920
7.	0.0455	0.0466	3.6824	0.816
8.	-0.0295	-0.0310	4.3905	0.820
9.	-0.0228	-0.0234	4.8171	0.850
10.	0.0169	0.0207	5.0503	0.888
11.	0.0965	0.0927	12.6823	0.315
12.	-0.0215	-0.0219	13.0608	0.365
13.	-0.0213	-0.0170	13.4350	0.415
14.	-0.0003	-0.0029	13.4350	0.493
15.	0.0162	0.0112	13.6499	0.552

A.6. táblázat. A III. NLS és ARMA(2,2) modell maradékának Q-statisztikái



A.1. ábra. A III. NLS és ARMA(2,2) maradékhoz tartozó korrelogram

A.11. SARMA modellek a megbetegedés számra I.

SARMA modellek jellemzői						
Modell	Log-likelihood	Akaike	Schwarz	Hannan-Q.	Innovációs Std. hiba	Rövidülés
<i>SARMA</i> (1, 1)(1, 1)	-5130.539	10273.08	10300.82	10283.77	220.1979	53 hét
<i>SARMA</i> (2, 2)(1, 1)	-5105.217	10226.43	10263.42	10240.68	214.8443	54 hét
<i>SARMA</i> (3, 3)(1, 1)	-5084.985	10189.97	10236.18	10207.78	211.0329	55 hét
<i>SARMA</i> (4, 4)(1, 1)	-5069.879	10163.76	10219.20	10185.12	208.7005	56 hét
<i>SARMA</i> (5, 5)(1, 1)	-5059.135	10146.27	10210.93	10171.19	207.5932	57 hét
<i>SARMA</i> (1, 0)(1, 1)	-5209.949	10429.90	10453.02	10438.81	244.6883	53 hét
<i>SARMA</i> (2, 0)(1, 1)	-5133.304	10278.61	10306.34	10289.29	223.0204	54 hét
<i>SARMA</i> (3, 0)(1, 1)	-5102.242	10218.48	10250.83	10230.95	215.9382	55 hét
<i>SARMA</i> (4, 0)(1, 1)	-5081.049	10178.10	10215.06	10192.34	211.8321	56 hét
<i>SARMA</i> (5, 0)(1, 1)	-5067.552	10153.10	10194.67	10169.12	209.9394	57 hét
<i>SARMA</i> (0, 1)(1, 1)	-5249.496	10508.99	10532.12	10517.90	255.5092	53 hét
<i>SARMA</i> (0, 2)(1, 1)	-5211.449	10434.90	10462.65	10445.59	242.9359	54 hét
<i>SARMA</i> (0, 3)(1, 1)	-5181.408	10376.82	10409.19	10389.29	233.4472	55 hét
<i>SARMA</i> (0, 4)(1, 1)	-5157.948	10331.90	10368.90	10346.15	226.2955	56 hét
<i>SARMA</i> (0, 5)(1, 1)	-5150.870	10319.74	10361.37	10335.78	224.1809	57 hét

A.12. SARMA modellek a megbetegedés számra II.

SARMA modellek jellemzői						
Modell	Log-likelihood	Akaike	Schwarz	Hannan-Q.	Innovációs Std. hiba	Rövidülés
<i>SARMA</i> (1, 1)(2, 2)	-4734.735	9485.470	9521.890	9499.547	207.5390	105 hét
<i>SARMA</i> (2, 2)(2, 2)	-4721.198	9462.395	9507.906	9479.988	205.5377	106 hét
<i>SARMA</i> (3, 3)(2, 2)	-4714.056	9452.111	9506.707	9473.217	205.4209	107 hét
<i>SARMA</i> (4, 4)(2, 2)	-4702.202	9432.405	9496.080	9457.022	203.9227	108 hét
<i>SARMA</i> (5, 5)(2, 2)	-4690.561	9413.123	9485.871	9441.250	202.4929	109 hét
<i>SARMA</i> (1, 0)(2, 2)	-4803.151	9620.301	9652.169	9632.619	228.8157	105 hét
<i>SARMA</i> (2, 0)(2, 2)	-4749.280	9514.559	9550.968	9528.633	213.9509	106 hét
<i>SARMA</i> (3, 0)(2, 2)	-4718.205	9454.410	9495.356	9470.239	206.6438	107 hét
<i>SARMA</i> (4, 0)(2, 2)	-4706.344	9432.687	9478.169	9450.271	205.1362	108 hét
<i>SARMA</i> (5, 0)(2, 2)	-4699.677	9421.353	9471.368	9440.691	205.1584	109 hét
<i>SARMA</i> (0, 1)(2, 2)	-4847.203	9708.406	9740.283	9720.727	241.2681	105 hét
<i>SARMA</i> (0, 2)(2, 2)	-4820.311	9656.622	9693.054	9670.703	232.2006	106 hét
<i>SARMA</i> (0, 3)(2, 2)	-4788.201	9594.401	9635.387	9610.242	221.8186	107 hét
<i>SARMA</i> (0, 4)(2, 2)	-4765.937	9551.874	9597.414	9569.475	214.8942	108 hét
<i>SARMA</i> (0, 5)(2, 2)	-4759.012	9540.023	9590.117	9559.385	212.7846	109 hét

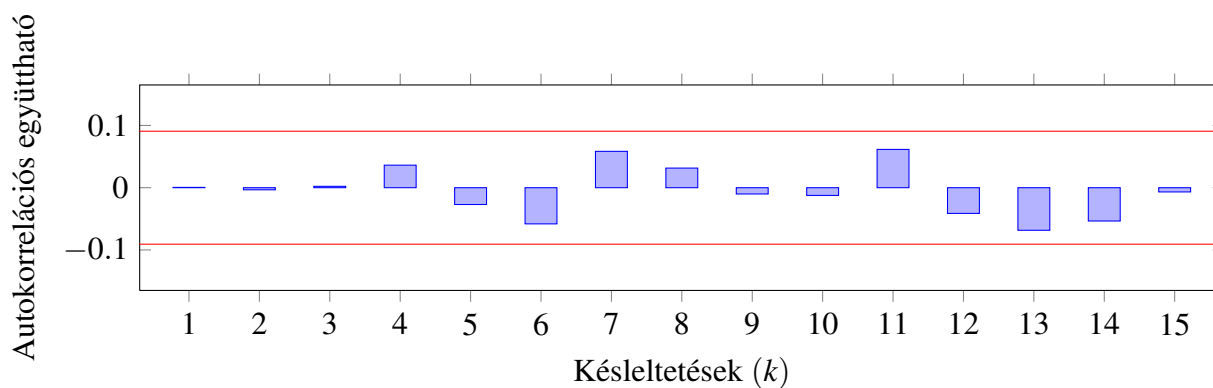
A.13. SARIMA modellek a megbetegedés számra

SARIMA modellek jellemzői						
Modell	Log-likelihood	Akaike	Schwarz	Hannan-Q.	Innovációs Std. hiba	Rövidülés
SARMA(1, 1, 1)(1, 0, 1)	-5121.335	10254.67	10282.41	10265.36	219.4989	54 hét
SARMA(2, 1, 2)(1, 0, 1)	-5099.929	10215.86	10252.83	10230.10	215.2742	55 hét
SARMA(3, 1, 3)(1, 0, 1)	-5085.762	10191.52	10237.73	10209.33	213.1674	56 hét
SARMA(4, 1, 4)(1, 0, 1)	-5073.624	10171.25	10226.67	10192.61	211.6481	57 hét
SARMA(5, 1, 5)(1, 0, 1)	-5065.623	10159.25	10223.89	10184.16	211.3012	58 hét
SARMA(1, 0, 1)(1, 1, 1)	-4774.535	9561.069	9588.384	9571.628	219.6632	105 hét
SARMA(2, 0, 2)(1, 1, 1)	-4761.573	9539.146	9575.554	9553.220	217.7415	106 hét
SARMA(3, 0, 3)(1, 1, 1)	-4749.248	9518.495	9563.992	9536.084	216.0278	107 hét
SARMA(4, 0, 4)(1, 1, 1)	-4736.248	9496.496	9551.074	9517.596	214.1158	108 hét
SARMA(5, 0, 5)(1, 1, 1)	-4731.367	9490.733	9554.388	9515.344	214.7015	109 hét
SARMA(1, 1, 1)(1, 1, 1)	-4762.011	9562.031	9589.338	9572.587	221.9633	105 hét
SARMA(2, 1, 2)(1, 1, 1)	-4758.735	9540.023	9576.420	9554.093	220.0087	106 hét
SARMA(3, 1, 3)(1, 1, 1)	-4752.217	9524.433	9569.916	9542.017	219.0708	107 hét
SARMA(4, 1, 4)(1, 1, 1)	-4741.261	9506.523	9561.084	9527.618	217.7712	108 hét
SARMA(5, 1, 5)(1, 1, 1)	-4725.735	9478.005	9541.640	9502.611	214.8325	109 hét

A.14. Ljung-Box Q-statisztikák a SARIMA I. modell maradáira

Késleltetés	Ljung-Box Q-statisztikák			
	Autokorrelációs együttható ρ	Parciális autokorrelációs együttható $\rho(h)$	Q-statisztika	p -érték
1.	-0.0005	-0.0005	0.0002	0.990
2.	-0.0034	-0.0034	0.0087	0.996
3.	0.0022	0.0022	0.0123	1.000
4.	0.0362	0.0361	1.0003	0.910
5.	-0.0270	-0.0270	1.5537	0.907
6.	-0.0581	-0.0579	4.1088	0.662
7.	0.0584	0.0584	6.7016	0.461
8.	0.0315	0.0302	7.4539	0.489
9.	-0.0101	-0.0081	7.5323	0.582
10.	0.0124	-0.0092	7.6489	0.663
11.	0.0615	0.0547	10.5334	0.483
12.	-0.0413	-0.0443	11.8346	0.459
13.	-0.0684	-0.0604	15.4133	0.282
14.	-0.0535	-0.0535	17.6062	0.225
15.	-0.0068	0.0157	17.6415	0.282

A.10. táblázat. A SARIMA (3,1,3)(1,0,1) modell maradékának Q-statisztikái

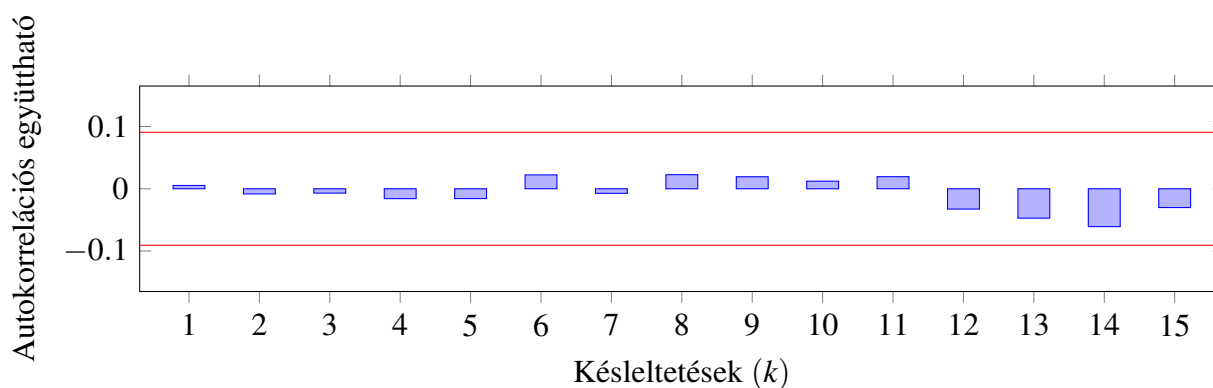


A.2. ábra. A SARIMA (3,1,3)(1,0,1) maradékhoz tartozó korrelogram

A.15. Ljung-Box Q-statisztikák a SARIMA II. modell maradáira

Késleltetés	Ljung-Box Q-statisztikák			
	Autokorrelációs együttható ρ	Parciális autokorrelációs együttható $\rho(h)$	Q-statisztika	p-érték
1.	0.0053	0.0053	0.0214	0.884
2.	-0.0083	-0.0084	0.0736	0.964
3.	-0.0070	-0.0069	0.1108	0.991
4.	-0.0159	-0.0159	0.3012	0.990
5.	-0.0158	-0.0158	0.4908	0.992
6.	0.0223	0.0222	0.8671	0.990
7.	-0.0073	-0.0080	0.9073	0.996
8.	0.0226	0.0227	1.2965	0.996
9.	0.0193	0.0187	1.5792	0.997
10.	0.0122	0.0127	1.6920	0.998
11.	0.0194	0.0205	1.9797	0.999
12.	-0.0326	-0.0324	2.7884	0.997
13.	-0.0472	-0.0448	4.4903	0.985
14.	-0.0608	-0.0611	7.3172	0.922
15.	-0.0302	-0.0309	8.0166	0.923

A.11. táblázat. A SARIMA (4,1,4)(1,0,1) modell maradékának Q-statisztikái



A.3. ábra. A SARIMA (4,1,4)(1,0,1) maradékhoz tartozó korrelogram

A.16. Determinisztikus Poisson-regresszió maximum likelihood kódrészlet

```
1 scalar a = 10 % Alpha skalár
2 scalar b1 = 10 % Béta 1 skalár
3 scalar g1 = 10 % Gamma 1 skalár
4 scalar p1 = 10 % Pszi 1 skalár
5 scalar b2 = 10 % Béta 2 skalár
6 scalar g2 = 10 % Gamma 2 skalár
7 scalar p2 = 10 % Pszi 2 skalár
8 mle loglik = Y*(a+b1*sin(t/p1-g1)+b2*sin(t/p2-g2))
9 -exp(a+b1*sin(t/p1-g1)+b2*sin(t/p2-g2)) % Loglikelihood célfüggvény
10 deriv a=Y-exp(a+b1*sin(t/p1-g1)+b2*sin(t/p2-g2)) % Alpha derivált
11 deriv b1=Y*sin(t/p1-g1) % Béta 1 szerinti derivált
12 -exp(a+b1*sin(t/p1-g1)+b2*sin(t/p2-g2))*sin(t/p1-g1)
13 deriv b2=Y*sin(t/p2-g2) % Béta 2 szerinti derivált
14 -exp(a+b1*sin(t/p1-g1)+b2*sin(t/p2-g2))*sin(t/p2-g2)
15 deriv p1=-Y*b1*cos(t/p1-g1)*(t/(p1*p1)) % Pszi 1 szerinti derivált
16 +exp(a+b1*sin(t/p1-g1)+b2*sin(t/p2-g2))*b1*cos(t/p1-g1)*(t/(p1*p1))
17 deriv p2=-Y*b2*cos(t/p2-g2)*(t/(p2*p2)) % Pszi 2 szerinti derivált
18 +exp(a+b1*sin(t/p1-g1)+b2*sin(t/p2-g2))*b2*cos(t/p2-g2)*(t/(p2*p2))
19 deriv g1=-Y*b1*cos(t/p1-g1) % Gamma 1 szerinti derivált
20 +exp(a+b1*sin(t/p1-g1)+b2*sin(t/p2-g2))*b1*cos(t/p1-g1)
21 deriv g2=-Y*b2*cos(t/p2-g2) % Gamma 2 szerinti derivált
22 +exp(a+b1*sin(t/p1-g1)+b2*sin(t/p2-g2))*b2*cos(t/p2-g2)
23 end mle
```

A.17. Sztochasztikus Poisson-regresszió maximum likelihood kódrészlet

```
1 scalar a = 10 % Alpha skalár
2 scalar b1 = 10 % Béta 1 skalár
3 scalar g1 = 10 % Gamma 1 skalár
4 scalar p1 = 10 % Pszi 1 skalár
5 scalar b2 = 10 % Béta 2 skalár
```

```

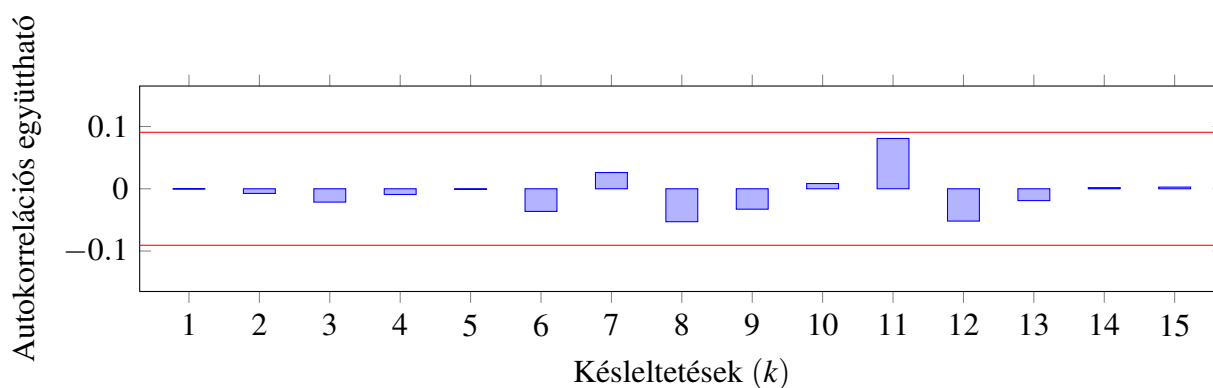
1 scalar g2 = 10 % Gamma 2 skalár
2 scalar p2 = 10 % Pszi 2 skalár
3 scalar h1 = 10 % Phi 1 skalár
4 scalar h2 = 10 % Phi 2 skalár
5 scalar h3 = 10 % Phi 3 skalár
6 mle loglik = Y*(a+b1*sin(t/p1-g1)+b2*sin(t/p2-g2)+h1*L1+h2*L2+h3*L3)
7 -exp(a+b1*sin(t/p1-g1)+b2*sin(t/p2-g2)+h1*L1+h2*L2+h3*L3)
8 % Loglikelihood célfüggvény
9 deriv a=Y-exp(a+b1*sin(t/p1-g1)+b2*sin(t/p2-g2)+h1*L1+h2*L2+h3*L3)
10 % Alpha derivált
11 deriv b1=Y*sin(t/p1-g1) % Béta 1 szerinti derivált
12 -exp(a+b1*sin(t/p1-g1)+b2*sin(t/p2-g2)+h1*L1+h2*L2+h3*L3)*sin(t/p1-g1)
13 deriv b2=Y*sin(t/p2-g2) % Béta 2 szerinti derivált
14 -exp(a+b1*sin(t/p1-g1)+b2*sin(t/p2-g2)+h1*L1+h2*L2+h3*L3)*sin(t/p2-g2)
15 deriv p1=-Y*b1*cos(t/p1-g1)*(t/(p1*p1)) % Pszi 1 szerinti derivált
16 +exp(a+b1*sin(t/p1-g1)+b2*sin(t/p2-g2)+h1*L1+h2*L2+h3*L3)
17 *b1*cos(t/p1-g1)*(t/(p1*p1))
18 deriv p2=-Y*b2*cos(t/p2-g2)*(t/(p2*p2)) % Pszi 2 szerinti derivált
19 +exp(a+b1*sin(t/p1-g1)+b2*sin(t/p2-g2)+h1*L1+h2*L2+h3*L3)
20 *b2*cos(t/p2-g2)*(t/(p2*p2))
21 deriv g1=-Y*b1*cos(t/p1-g1) % Gamma 1 szerinti derivált
22 +exp(a+b1*sin(t/p1-g1)+b2*sin(t/p2-g2)+h1*L1+h2*L2+h3*L3)
23 *b1*cos(t/p1-g1)
24 deriv g2=-Y*b2*cos(t/p2-g2) % Gamma 2 szerinti derivált
25 +exp(a+b1*sin(t/p1-g1)+b2*sin(t/p2-g2)+h1*L1+h2*L2+h3*L3)
26 *b2*cos(t/p2-g2)
27 deriv h1=Y-exp(a+b1*sin(t/p1-g1)+b2*sin(t/p2-g2)+h1*L1+h2*L2+h3*L3)
28 % Phi1 szerinti derivált
29 deriv h2=Y-exp(a+b1*sin(t/p1-g1)+b2*sin(t/p2-g2)+h1*L1+h2*L2+h3*L3)
30 % Phi2 szerinti derivált
31 deriv h3=Y-exp(a+b1*sin(t/p1-g1)+b2*sin(t/p2-g2)+h1*L1+h2*L2+h3*L3)
32 % Phi3 szerinti derivált
33 end mle

```

A.18. Sztochasztikus Poisson-regresszió maradékához tartozó korrelogram

Késleltetés	Ljung–Box Q–statisztikák			
	Autokorrelációs együttható ρ	Parciális autokorrelációs együttható $\rho(h)$	Q–statisztika	p–érték
1.	-0.0003	-0.0003	0.0001	0.993
2.	-0.0075	-0.0075	0.0449	0.978
3.	-0.0214	-0.0214	0.4136	0.937
4.	-0.0092	-0.0093	0.4817	0.975
5.	-0.0010	-0.0014	0.4826	0.993
6.	-0.0364	-0.0370	1.5549	0.956
7.	0.0261	0.0256	2.1049	0.954
8.	-0.0529	-0.0537	4.3737	0.822
9.	-0.0328	-0.0341	5.2468	0.812
10.	0.0084	0.0080	5.3048	0.870
11.	0.0808	0.0787	10.6187	0.476
12.	-0.0519	-0.0562	12.8165	0.383
13.	-0.0190	-0.0164	13.1103	0.439
14.	0.0018	0.0001	13.1128	0.518
15.	0.0027	0.0023	13.1189	0.593

A.12. táblázat. A sztochasztikus Poisson-regresszió maradékához tartozó korrelogram



A.4. ábra. A sztochasztikus Poisson-regresszió maradékához tartozó korrelogram

A.19. Determinisztikus modell optimalizációs kód

```
1 function fval=periodus1(x)
2 t1=x(1); % Időindex értékadás
3 beta1=-604.755; % Béta 1 deklarálás
4 beta2=-201.421; % Béta 2 deklarálás
5 gamma1=34.279; % Gamma 1 deklarálás
6 gamma2=12.666; % Gamma 2 deklarálás
7 psi1=8.771; % Pszi 1 deklarálás
8 psi2=4.879; % Pszi 2 deklarálás
9 fval(1)=beta1*cos(t/psi1-gamma1)/psi1
10 +beta2*cos(t/psi2-gamma2)/psi2; % Optimalizációs probléma
11 end
```

A gyökkeresést végrehajtó kódrészlet a függvényre:

```
1 [x,y,fval,exitflag]=fsolve('periodus1',35,15);
```

A.20. Poisson-regresszió optimalizációs kód

```
1 function fval=periodus2(x)
2 t1=x(1); % Időindex értékadás
3 beta1=-1.06141; % Béta 1 deklarálás
4 beta2=0.42433; % Béta 2 deklarálás
5 gamma1=9.30109; % Gamma 1 deklarálás
6 gamma2=10.3235; % Gamma 2 deklarálás
7 psi1=8.19446; % Pszi 1 deklarálás
8 psi2=4.3145; % Pszi 2 deklarálás
9 fval(1)=beta1*cos(t1/psi1-gamma1)/psi1
10 +beta2*cos(t1/psi2-gamma2)/psi2; % Optimalizációs probléma
11 end
```

A gyökkeresést végrehajtó kódrészlet a két függvényre:

```
1 [x,y,fval,exitflag]=fsolve('periodus2',35,15);
```

A.21. Determinisztikus modell rugalmasság kód

```
1 t=[1:52]; % Időindex értékadás
2 alpha= 821.479;% Alpha deklarálás
3 beta1=-604.755; % Béta 1 deklarálás
4 beta2=-201.421; % Béta 2 deklarálás
5 gamma1=34.279; % Gamma 1 deklarálás
6 gamma2=12.666; % Gamma 2 deklarálás
7 psi1=8.771; % Pszi 1 deklarálás
8 psi2=4.879; % Pszi 2 deklarálás
9 for i=1:52
10 epsilon(i)=t(i)*(beta1*cos((t(i)/psi1)-gamma1)/
11 psi1+beta2*cos((t(i)/psi2)-gamma2)/psi2)/
12 (alpha+beta1*sin((t(i)/psi1)-gamma1)+beta2*sin((t(i)/psi2)-gamma2));
13 end
```

A.22. Poisson-regresszió rugalmasság kód

```
1 t=[1:52]; % Időindex értékadás
2 beta1=-1.06141; % Béta 1 deklarálás
3 beta2=0.42433; % Béta 2 deklarálás
4 gamma1=9.30109; % Gamma 1 deklarálás
5 gamma2=10.3235; % Gamma 2 deklarálás
6 psi1=8.19446; % Pszi 1 deklarálás
7 psi2=4.3145; % Pszi 2 deklarálás
8 for i=1:52
9 epsilon(i)=t(i)*(beta1*cos(gamma1-(t(i)/psi1))/psi1+beta2
10 *cos(gamma2-(t(i)/psi2))/psi2);
11 end
```


B. Függelék

Többváltozós modellezés

B.1. Deskriptív statisztikák a megyei idősorokra

Változó	Minimum	Maximum	Átlag	Szórás	Ferdeség
Budapest	0	479	105.82	76.228	0.847
Baranya	0	194	36.41	33.820	1.263
Bács	0	238	38.50	35.217	1.403
Békés	0	271	31.30	39.273	1.363
Borsod	1	355	58.64	50.792	1.442
Csongrád	0	199	32.50	34.178	1.675
Fejér	0	164	35.06	32.329	1.037
Győr	0	181	43.82	36.908	0.955
Hajdú	0	262	49.79	45.227	1.341
Heves	0	210	30.43	32.529	1.001
Jász	0	224	42.42	37.773	1.043
Komárom	0	160	28.11	24.981	1.298
Nógrád	0	112	23.28	22.545	1.326
Pest	1	431	91.25	68.223	0.762
Somogy	0	155	29.51	27.430	1.321
Szabolcs	0	203	31.87	33.064	1.796
Tolna	0	131	22.10	24.045	1.600
Vas	0	141	23.65	25.932	1.579
Veszprém	0	230	43.02	42.130	1.488
Zala	0	216	20.88	22.044	1.391

B.1. táblázat. Deskriptív statisztikák a megyei idősorokra

B.2. Főkomponensek a megyei idősorokra

Komponens	Főkomponens jellemzők				
	Sajátérték	Variancia hányad %	Kumulatív variancia %	Forgatott variancia hányad %	Forgatott kumulatív variancia %
1	10.495	52.476	52.476	32.762	32.762
2	1.361	6.807	59.283	26.521	59.283
3	0.962	4.808	64.090		
4	0.884	4.421	68.512		
5	0.749	3.743	72.255		
6	0.666	3.330	75.585		
7	0.597	2.983	78.567		
8	0.545	2.723	81.290		
9	0.511	2.554	83.844		
10	0.448	2.241	86.086		
11	0.406	2.031	88.117		
12	0.360	1.800	89.916		
13	0.351	1.757	91.673		
14	0.318	1.588	93.261		
15	0.291	1.457	94.718		
16	0.264	1.322	96.040		
17	0.241	1.204	97.244		
18	0.228	1.140	98.384		
19	0.174	0.869	99.253		
20	0.149	0.747	100.000		

B.2. táblázat. Főkomponensek a megyei idősorokra

B.3. Forgatott komponens mátrix a megyei idősorokra

Megye	Forgatott komponens mátrix	
	1. komponens	2. komponens
Budapest	0.518	0.663
Baranya	0.463	0.586
Bács	0.672	0.338
Békés	0.713	0.183
Borsod	0.665	0.377
Csongrád	0.284	0.696
Fejér	0.802	0.183
Győr	0.805	0.348
Hajdú	0.747	0.258
Heves	0.581	0.373
Jász	0.485	0.608
Komárom	0.743	0.337
Nógrád	0.289	0.643
Pest	0.621	0.603
Somogy	0.414	0.651
Szabolcs	0.613	0.292
Tolna	0.135	0.766
Vas	0.214	0.692
Vesz prém	0.434	0.677
Zala	0.278	0.597

B.3. táblázat. Forgatott komponens mátrix a megyei idősorokra

B.4. Deskriptív statisztikák a szezonálisan differenciázott idő-sorokra

Változó	Minimum	Maximum	Átlag	Szórás	Ferdeség
Δ_{52} Budapest	-228	386	2.04	66.415	0.521
Δ_{52} Baranya	-100	165	0.33	33.916	0.578
Δ_{52} Bács	-210	130	-4.60	34.477	-0.336
Δ_{52} Békés	-236	146	-6.04	46.084	-0.992
Δ_{52} Borsod	-327	261	-3.59	52.392	-0.108
Δ_{52} Csongrád	-144	151	-0.92	40.415	0.103
Δ_{52} Fejér	-106	101	-3.62	33.581	-0.345
Δ_{52} Győr	-139	97	-4.89	31.981	-0.594
Δ_{52} Hajdú	-218	159	-4.07	46.476	-0.736
Δ_{52} Heves	-167	182	-2.43	43.108	0.092
Δ_{52} Jász	-148	125	-4.92	38.798	-0.490
Δ_{52} Komárom	-134	150	-4.75	23.968	-0.198
Δ_{52} Nógrád	-77	79	1.55	25.929	-0.280
Δ_{52} Pest	-288	313	0.72	58.082	-0.033
Δ_{52} Somogy	-118	114	-0.17	25.969	-0.126
Δ_{52} Szabolcs	-195	169	-1.37	38.645	-0.152
Δ_{52} Tolna	-98	118	2.01	29.097	0.589
Δ_{52} Vas	-124	140	-3.07	33.427	0.272
Δ_{52} Veszprém	-154	167	-3.67	41.513	0.320
Δ_{52} Zala	-143	195	-2.63	25.249	0.753

B.4. táblázat. Deskriptív statisztikák a szezonálisan differenciázott idő-sorokra

B.5. Főkomponensek a szezonálisan differenciázott idősorokra

Komponens	Főkomponens jellemzők				
	Sajátérték	Variancia hányad %	Kumulatív variancia %	Forgatott variancia hányad %	Forgatott kumulatív variancia %
1	3.458	17.291	17.291	11.519	11.519
2	2.812	14.058	31.349	9.878	21.397
3	1.793	8.966	40.315	9.412	30.808
4	1.661	8.304	48.620	9.309	40.117
5	1.183	5.916	54.536	8.789	48.906
6	1.103	5.516	60.052	8.537	57.443
7	1.060	5.299	65.351	7.908	65.351
8	0.894	4.470	69.821		
9	0.816	4.082	73.903		
10	0.756	3.780	77.683		
11	0.673	3.365	81.048		
12	0.610	3.050	84.098		
13	0.547	2.733	86.831		
14	0.473	2.365	89.196		
15	0.451	2.256	91.451		
16	0.413	2.066	93.517		
17	0.385	1.927	95.444		
18	0.341	1.707	97.151		
19	0.300	1.502	98.653		
20	0.269	1.347	100.000		

B.5. táblázat. Főkomponensek a szezonálisan differenciázott idősorokra

B.6. Forgatott komponens mátrix a szezonálisan differenci- ázott idősorokra

Megye	Forgatott komponens mátrix						
	1. komp.	2. komp.	3. komp.	4. komp.	5. komp.	6. komp.	7. komp.
Δ_{52} Budapest	0.414	0.363	0.444	0.072	0.109	0.198	0.043
Δ_{52} Baranya	-0.025	-0.108	-0.063	0.184	0.077	0.799	-0.020
Δ_{52} Bács	0.234	-0.180	-0.070	-0.297	0.074	0.166	0.667
Δ_{52} Békés	-0.025	0.374	0.108	0.214	0.122	0.636	0.213
Δ_{52} Borsod	-0.022	0.008	0.658	0.091	0.166	-0.084	0.004
Δ_{52} Csongrád	0.378	0.411	0.193	-0.229	0.252	-0.204	-0.027
Δ_{52} Fejér	-0.044	-0.766	0.089	0.289	0.081	-0.145	-0.021
Δ_{52} Győr	-0.196	-0.093	-0.012	0.701	0.193	0.024	0.138
Δ_{52} Hajdú	-0.449	0.197	0.530	0.249	0.068	0.080	0.038
Δ_{52} Heves	-0.076	-0.098	0.166	0.078	0.842	0.052	0.156
Δ_{52} Jász	0.208	0.067	-0.012	0.038	-0.077	0.080	0.843
Δ_{52} Komárom	0.092	0.090	0.338	0.649	-0.215	0.224	0.192
Δ_{52} Nógrád	0.191	0.269	-0.031	0.069	0.721	0.203	-0.138
Δ_{52} Pest	0.231	0.073	0.285	0.100	0.232	0.686	0.171
Δ_{52} Somogy	0.591	0.176	0.044	0.062	0.171	0.092	-0.059
Δ_{52} Szabolcs	0.167	-0.210	0.720	-0.172	-0.119	0.134	0.164
Δ_{52} Tolna	0.348	0.455	-0.015	-0.392	-0.182	0.450	0.084
Δ_{52} Vas	0.769	0.004	0.160	0.192	-0.252	0.096	0.087
Δ_{52} Veszprém	0.780	0.108	-0.050	0.022	-0.187	0.158	0.154
Δ_{52} Zala	0.483	0.103	0.177	-0.115	0.255	-0.148	0.419

B.7. Bináris területi súlymátrix Magyarországra

1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	Budapest
0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	Baranya
0	1	1	0	0	1	1	0	0	0	1	0	0	1	0	0	1	0	0	0	0	Bács
0	0	0	1	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	Békés
0	0	0	0	1	0	0	0	1	1	1	0	1	0	0	1	0	0	0	0	0	Borsod
0	0	1	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	Csongrád
0	0	1	0	0	0	1	0	0	0	0	1	0	1	1	0	1	0	1	0	0	Fejér
0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1	1	0	0	Győr
0	0	0	1	1	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0	Hajdú
0	0	0	0	1	0	0	0	0	1	1	0	1	1	0	0	0	0	0	0	0	Heves
0	0	1	1	1	1	0	0	1	1	1	0	0	1	0	0	0	0	0	0	0	Jász
0	0	0	0	0	0	1	1	0	0	0	1	0	1	0	0	0	0	0	1	0	Komárom
0	0	0	0	1	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	Nógrád
1	0	1	0	0	0	1	0	0	1	1	1	1	1	0	0	0	0	0	0	0	Pest
0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	1	1	1	Somogy
0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	Szabolcs
0	1	1	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	Tolna
0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	1	Vas
0	0	0	0	0	0	1	1	0	0	0	1	0	0	1	0	0	1	1	1	1	Veszprém
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	Zala

B.8. A Moran-féle I-index értékét számító kód

```
1 function [M] = moran(Adat,W)
2 c=length(transpose(Adat)); % Az idő dimenzió meghatározása
3 for t=1:c
4 X=Adat(t,:); % Az adatokat tartalmazó mátrix első sorának lekérése
5 k=length(X); % A kapott vektor hosszának meghatározása
6 osztó=sum(sum(W)); % A súlymátrix beösszegzése
7 atlag=mean(X); % Az adott időpontban vett térbeli átlag
8 for i=1:k
9 hanyados(i)=(X(i)-atlag)^2; % A második tört nevezője
10 end
11 for i=1:k
12 for j=1:k
13 szamlalo(i,j)=W(i,j)*(X(i)-atlag)*(X(j)-atlag); % A második tört szám.
14 end
15 end
16 sum(hanyados);
17 M(t)=(k/osztó)*(sum(sum(szamlalo))/sum(hanyados)); % Az index számítása
18 end
19 end
```

Az adatok meghívása a nem differenciázott idősorok esetében így történik:

```
1 W=csvread('proximity.csv',0,0);
2 Adat=csvread('megyek.csv',1,1);
```

Az adatok meghívása a szezonálisan differenciázott idősorok esetében így történik:

```
1 W=csvread('proximity.csv',0,0);
2 Adat=csvread('sezdiffmegyek.csv',1,1);
```

B.9. A Moran-féle I-indexre felépített AR modell

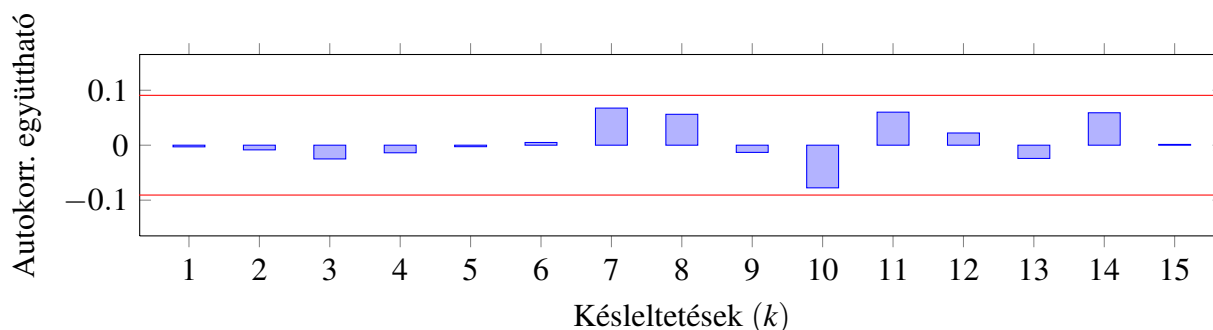
```

1 test with constant model: (1-L)y = b0 + (a-1)*y(-1) + ... + e
2 1st-order autocorrelation coeff. for e: 0.006
3 lagged differences: F(9, 431) = 5.985 [0.0000]
4 estimated value of (a - 1): -0.454962
5 test statistic: tau_c(1) = -5.41559
6 asymptotic p-value 2.67e-06
7 constant and trend model: (1-L)y = b0 + b1*t + (a-1)*y(-1) + ... + e
8 1st-order autocorrelation coeff. for e: 0.006
9 lagged differences: F(9, 430) = 5.929 [0.0000]
10 estimated value of (a - 1): -0.458116
11 test statistic: tau_ct(1) = -5.44036
12 asymptotic p-value 2.252e-05

```

	Paraméter	Std. hiba	<i>t</i> -próba	<i>p</i> -érték	
	α	0.2170	0.0079	27.4381	0.0000
	ϕ_1	0.1091	0.0461	2.3670	0.0179
	ϕ_2	0.2233	0.0452	4.9355	0.0000
	ϕ_3	0.1903	0.0462	4.1212	0.0000
Függő változó átlaga		0.217651	Függő változó std. σ		0.087023
Innovációk átlaga		0.109088	Innovációs std. hiba		0.080703
Log-likelihood		-496.1707	Akaike kritérium		982.3413
Schwarz kritérium		961.7729	Hannan–Quinn		974.2360

B.6. táblázat. A Moran-féle I-indexre felépített AR(3) modell jellemzői



B.1. ábra. A Moran-féle I-indexre felépített modell maradékához tartozó korrelogram

B.10. A Geary-féle C-index értékét számító kód

```
1 function [M] = geary(Adat,W)
2 c=length(transpose(Adat)); % Az idő dimenzió meghatározása
3 for t=1:c
4 X=Adat(t,:); % Az adatokat tartalmazó mátrix első sorának lekérése
5 k=length(X); % A kapott vektor hosszának meghatározása
6 for i=1:k
7 for j=1:k
8 szamlalo(i,j)=W(i,j)*((X(i)-X(j))^2); % A számláló értékei
9 end
10 end
11 szamlalo=sum(sum(szamlalo)); % A kapott értékek összegzése
12 atlag=mean(X);
13 for i=1:k
14 nevezo(i)=(X(i)-atlag)^2; % A nevező értékei
15 end
16 C(t)=(k-1)*szamlalo/(2*sum(sum(W))*sum(nevezo)); % A C mutató
17 end
18 end
```

Az adatok meghívása a nem differenciázott idősorok esetében így történik:

```
1 W=csvread('proximity.csv',0,0);
2 Adat=csvread('megyek.csv',1,1);
```

Az adatok meghívása a szezonálisan differenciázott idősorok esetében így történik:

```
1 W=csvread('proximity.csv',0,0);
2 Adat=csvread('szezdiffmegyek.csv',1,1);
```

B.11. Bayes-i prior kód

```
1 clear all
2 close all
3 clc
4 W=csvread('proximity2.csv',0,20);
5 Y=csvread('szездiff.csv',53,1);
6 nlag=3;
7 decay=1;
8 tightness=0.1;
9 result=bvar(Y,nlag,tightness,W,decay)
```