

Járványterjedési folyamatok paraméterbecslése többrétegű véletlen gráfokon

Önálló projekt II.

Krajnyák Bálint

Témavezető: Backhausz Ágnes

Bevezetés

Az Önálló Projekt második félévében folytattam az előző félévben elkezdett munkát, vagyis járványterjedési szimulációkat hajtottam végre, azonban a vizsgálat tárgya egy kicsit változott. Amíg előző félévben arra voltam kíváncsi, hogy az egyes paraméterek megváltoztatása milyen hatással van a járvány lefutására, addig ebben a félévben a fő kérdés az volt, hogy egy járványgörbe alapján hogyan tudjuk megbecsülni az egyes paramétereket.

Szakirodalom tekintetében elsősorban a [2]-es sorszámú forrásra támaszkodtam. Ezen cikkben négyrétegű hálózaton lettek szimulációk elvégezve MATLAB szoftver segítségével, amely programkódok velem is meg lettek osztva, és ezen kódok jelentették az általam elvégzett szimulációk kiindulópontját.

A beszámoló elején olvasható egy összefoglaló a modellről, a jelölésekről, a Gillespie-algoritmusról, illetve az előző félévi eredményekről. Ezen rész az előző félévi beszámolóban is olvasható volt, de úgy gondolom, hogy a teljesség kedvéért ezen beszámolóba is érdemes beletenni.

A modell

A járványterjedést kétrétegű véletlen gráfok segítségével modelleztük. Tekintünk egy populációt, amelyben minden egyes személyt megfeleltetünk a véletlen gráf egy csúcsának. A csúcsok között súlyozott élek mennek, mely élek megmutatják, hogy a populáció mely tagjai között van kapcsolat, és ezen kapcsolatnak mekkora az erőssége. Minél nagyobb ez a szám, annál erősebb a kapcsolat a 2 személy között.

A hálózat első rétegébe a háztartások tartoznak. Ezen réteg pontosan ugyanúgy lett megalkotva, mint ahogy a [2]-es cikkben a háztartások rétege elkészült. A szimulációk 1600 háztartásból álló hálózatra lettek lefuttatva, amely nagyjából 4000 fős populációt jelentett. Ez a létszám szimulációnként kicsit változott, mivel a hálózatban eltérő létszámú háztartások voltak, az egyes háztartások létszámai pedig valós magyarországi adatok alapján lettek kisorsolva. Mivel egy háztartás tagjai között elég erős kapcsolat van, ezért egy háztartáson belül bármely két személy között 1 súlyú él megy.

A második réteget pedig úgy alkottuk meg, hogy a populáció tagjait véletlenszerűen kisebb csoportokba soroltuk. Ezen csoportokat tekinthetjük például munkahelyeknek, iskoláknak. A sorsolást úgy végeztük el, hogy minden személy pontosan egy csoportba tartozzon, de az sem kizárható, hogy egy háztartásból több személy is ugyanazon csoportba került kisorsolásra. Ezen csoportokra is teljesül, mint ahogyan a háztartásokra is, hogy a csoport bármely 2 tagja között vezet él. Azonban ezen élek súlya szimulációnként eltérő.

Az SIR modell

A járványterjedés szimulációjára az SIR modellt használtuk. Ez azt jelenti, hogy a populáció egyedeit 3 csoportba soroljuk:

- **S** (Susceptible): fogékonyak
- **I** (Infected): fertőzők
- **R** (Recovered): gyógyultak

Minden olyan személyt, aki nem betegedett még meg, a fogékony kategóriába sorolunk. Miután megbetegszik, akkor átkerül a fertőző kategóriába, és miután meggyógyul, akkor pedig átkerül a gyógyultak kategóriába. Ezen modell szerint csak az előbb leírt sorrendben lehet egyik kategóriából a másikba átkerülni, tehát például, aki már meggyógyult nem tud újra megfertőződni. A személyek egyes állapotokban töltött idejét a fertőzési és gyógyulási ráták segítségével adhatjuk meg. Ha van egy él, amely egy fogékony és fertőző egyén között megy, akkor ezen élhez tartozik egy τ paraméterű, exponenciális eloszlású valószínűségi változó, amely megadja, hogy mikor adódik át a fertőzés ezen a csúcson. Ennek várható értéke $\frac{1}{\tau}$. Továbbá minden fertőző csúcshoz tartozik egy γ paraméterű, exponenciális eloszlású valószínűségi változó, amely a gyógyulás idejét adja meg, és melynek várható értéke $\frac{1}{\gamma}$.

A használt jelölések:

- τ : a csoporttagok közötti fertőzési ráta (élek súlya a gráfban), a háztartásokon belüli fertőzési ráta ettől lehet eltérő, amely a szimulációk során alapvetően egységesen 1 volt.
- γ : gyógyulási ráta
- R_{inf} : a fertőzésen átesettek száma
- I_{max} : a fertőzöttek maximális száma, vagyis egyazon időpontban a betegek maximális száma

A Gillespie-algoritmus

A járványterjedés szimulációjára a Gillespie-algoritmust futtattuk, amely az SIR modellt valósítja meg véletlen gráfon a következőképpen:

Minden egyes lépésben meghatározza, hogy a következő esemény mikor fog bekövetkezni. Ezen időt egy olyan exponenciális eloszlású valószínűségi változó adja meg, amelynek paramétere a fertőzési ráták, illetve gyógyulási ráták összege. Ezután pedig kisorsolja, hogy ezen esemény fertőzés vagy gyógyulás lesz, a fertőzési és gyógyulási ráták összege alapján. Gyógyulás esetén egyenletes eloszlás alapján választ egy csúcst a fertőzöttek közül, amelyek meggyógyul, fertőzés esetén, pedig a fogékony csúcsok közül választ egyet, annál nagyobb valószínűséggel, minél nagyobb az adott csúcsba a fertőző csúcsokból menő élsúlyok összege.

Az előző félévi eredmények összefoglalása

Az első félében elvégzett szimulációk eredményei alapján arra következtethetünk, hogy korlátozó intézkedések bevezetésével lehet lassítani a járványt. Azon szimulációk esetén, amikor a csoporttagok közötti τ értékeket változtattuk, akkor azt tapasztaltuk, hogy kisebb τ értékek esetén a járvány lefutása lassabb volt, a csúcson pedig kevesebben voltak betegek. Ez azt jelenti, hogy a maszkviselés egy hatásos védekezés lehet a járvánnyal szemben. Az eltérő τ értékek jelenthetik azt, hogy a csoport tagjai mennyire profi maszkot viseltek. Továbbá a csoportméretek csökkentésénél is hasonló tapasztalatokat fogalmazhatunk meg, vagyis hatásos védekezés lehet, ha a nagyobb csoportokat több kisebb csoportra osztják fel, például az iskolákban jobban odafigyelnek rá, hogy elkülönüljenek egymástól az osztályok.

Paraméterbecslés

Ahogy a bevezetésben is említettem, ebben a félévben a projekt fő célja annak vizsgálata, hogy egy adott járványgörbe alapján hogyan tudjuk meghatározni a járványterjedés egyes paramétereit. A [2]-es sorszámú cikk tartalmaz becsléseket a gyógyulási rátára, illetve a fertőzési rátára vonatkozóan.

A gyógyulási rátára vonatkozó becslés a következő:

$$\hat{\gamma} = \frac{Z_R}{\sum_{t_i \leq \bar{T}} I_{t_i} \cdot (t_i - t_{i-1})}$$

A fenti képletben Z_R jelöli az összes gyógyulási esemény számát a $[0, \bar{T}]$ időintervallumon, ahol $\bar{T} = \frac{5}{6} \cdot T$, az I_{t_i} pedig a fertőzöttek számát a t_i időpontban.

A fertőzési rátára pedig a következő becslést adja a [2]-es cikk:

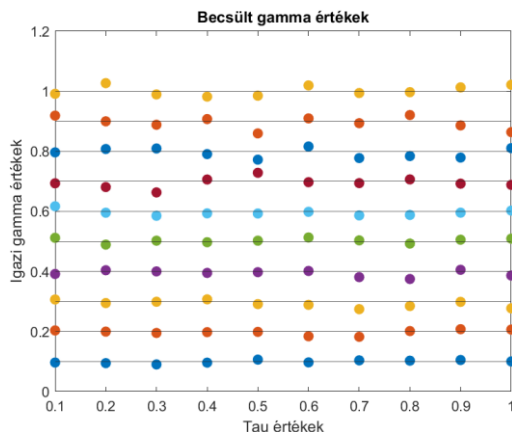
$$\hat{\tau} = \frac{Z_I}{\sum_{t_i \leq \bar{T}} E_{t_i}^{SI} \cdot (t_i - t_{i-1})}$$

Ebben a képletben Z_I jelöli az összes fertőződési esemény számát a $[0, \bar{T}]$ időintervallumon, $E_{t_i}^{SI}$ az SI élek számát a t_i időpontban.

Abban viszont nem lehetünk biztosak, hogy ezek a becslések a projekt keretében vizsgált gráf esetében is jó becslést fognak adni, mivel paraméterek ezen becsléseit egy teljes gráf alapján kapott ML-becslés adta eredményül, amely megfelelő becslésnek bizonyult a cikkben vizsgált nem teljes gráfra is. Azonban a most vizsgált gráf nem teljes, továbbá szerkezete eltér a [2]-es cikkben vizsgált gráf szerkezetétől is, emiatt az első lépés annak ellenőrzése, hogy ezen becslések megfelelőek-e ezen gráf esetén is.

Ezt úgy tettem meg, hogy az előző félévhez hasonlóan járványterjedési szimulációkat futtattam le MATLAB-ban, majd pedig kiszámítottam, hogy a fenti képletek alapján milyen becsléseket kapunk a fertőzési, illetve gyógyulási rátákra. A fertőzési, illetve gyógyulási ráta input paraméterek a szimulációk során, tehát számunkra ismertek, ezáltal könnyedén leellenőrizhetjük, hogy a kapott becslések mennyire adnak jó eredményt.

A gyógyulási ráta becslése: Az előző félévben a szimulációk során γ értéke egységesen 0,3-re lett beállítva, a paraméterérzékenység a τ , illetve a csoportméret változtatására lett vizsgálva. Viszont jelen esetben, mivel pont a gyógyulási rátát vizsgáljuk, ezért érdemes γ értékét is változtatni. A szimulációk során 0,1 és 1 közötti γ értékeket próbáltam ki.



1. ábra: γ értékek becslése a γ , illetve a τ értékek változásának függvényében

Az 1. ábra azt mutatja meg, hogy a cikkben található γ becsléssel milyen értékek jönnek ki a projekt során vizsgált szerkezetű gráf esetén. Összesen 100 becslést mutat meg, melyekben a τ és a γ értékek 0.1 és 1 között változnak. Az ábra x-tengelye felel meg az igazi τ értéknek, és az y-tengelye az igazi γ értéknek. Az ábrán lévő pontok ábrázolják a kapott becsléseket. Jelen szimulációk során csak a γ értékeket becsültük. A vízszintes vonalak jelzik az igazi γ értékeket. Az ábrán jól látható, hogy a kapott becslések viszonylag jól megközelítik az igazi γ értékeket.

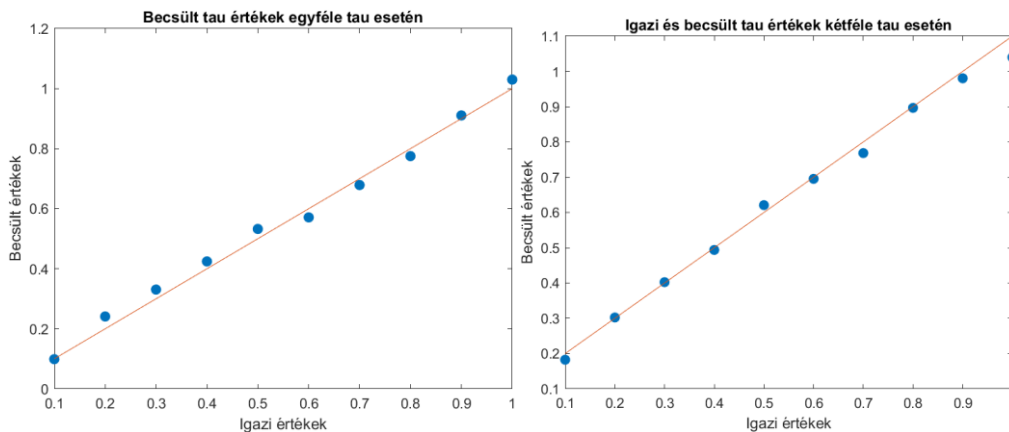
A fertőzési ráta becslése egységes τ esetén

A [2]-es cikk által adott τ becslés képletének alkalmazásához szükségünk van arra az információra, hogy az egyes időpontokban összesen hány SI él van a gráfon belül, azaz olyan él, amely egy fertőzött és egy fogékony állapotban lévő csúcst köt össze. Az SI élek számát korábbi szimulációk során nem kellett semmire felhasználni, úgyhogy a korábbi programkódokat ki kellett egészíteni az SI élek számának meghatározásával minden egyes időpontban.

A fertőzési ráta becslése egy kicsit bonyolultabb feladatnak bizonyult, mint a gyógyulási ráta becslése. A projekt során vizsgált gráfnak nem csak a szerkezete tér el a [2]-es cikkben vizsgált gráfétól, hanem még abban is eltér, hogy a fertőzési ráták értékét megkülönböztetjük a háztartások, illetve a csoportok esetén. Ezáltal kétféle fertőzési ráta is megjelenik, és az egy érdekes kérdés, hogy hogyan tudjuk ezt kezelni. A gyógyulási rátát a fertőzési ráták vizsgálata során egységesen 0,3-re állítottuk.

Első lépésben viszont érdemes azt megnézni, hogy működik-e a becslés abban az esetben, ha mindkét féle fertőzési rátát egységesre állítjuk. Ha ezen speciális esetben jó becslést kapunk akkor érdemes áttérni a kétféle τ érték kezelésének kérdésére.

Ezen szimulációk során mind a háztartásokon belül, mind a csoportokon belül 0,1 és 1 között választottam meg a fertőzési rátákat 0,1-es lépésközzel, vagyis mindkét típusú fertőzési ráta egységes volt. A 2. ábrán látható a szimulációk során kapott becslések és a tényleges értékek összehasonlítása, amelyeket a [2]-es cikkben található képlet segítségével határoztunk meg. Továbbá látható az ábrán az $y=x$ egyenes, vagyis ha jó a becslésünk, akkor a pontoknak ezen egyenes közelében kell elhelyezkedniük. Szerencsére valóban közel esetek a becsült értékek a valós értékekhez, tehát a becslés megfelelőnek bizonyult, tehát áttérhetünk a kétféle τ esetének vizsgálatára.



2-3. ábra: becsült és igazi τ értékek azon esetekben, amikor a háztartások közötti fertőzési ráta megegyezik a csoporttagok közötti fertőzési rátával, illetve amikor ezen fertőzési ráták eltérnek (a második esetben a csoporttagok közötti fertőzési rátát szeretnénk becsülni)

Kétféle τ kezelése

Ezt követően áttérhetünk a tényleges kérdésre, tehát, hogy hogyan tudjuk megbecsülni a fertőzési rátát, ha kétféle τ is van. Itt még érdemes azt is átgondolnunk, hogy mit is akarunk tulajdonképpen becsülni. A háztartásokon belül a fertőzési ráta egységesen 1, ezt ismerjük, tehát a csoportokon belüli τ értéket szeretnénk becsülni. Könnyedén átgondolható, hogy a [2]-es cikk becslése biztosan nem lesz jó a mi esetünkben. A csoporttagok közötti τ értékek 0,1 és 1 között mozognak, amíg a háztartások közötti fertőzési ráta 1. Vagyis a háztartásokon belül a fertőzési ráta értéke nagyobb, mint a csoportokon belül (attól az esettől eltekintve, amikor 1-nek választjuk a csoporttagok közötti fertőzési ráta értékét is). Ez azt jelenti, hogy a háztartásokon belül gyorsabban terjed a járvány, vagyis a képlet alapján kapott becslésnek valahol a csoporttagok közötti τ és a háztartáson belüli τ értékek között kell lennie.

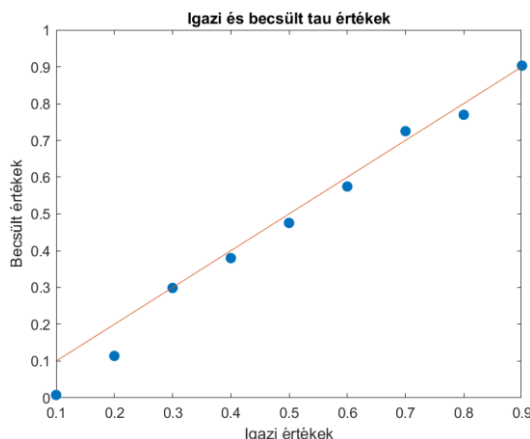
Ha elvégezzük a szimulációkat, és kiszámítjuk a [2]-es cikk becslését, akkor biztosan nem fogjuk megkapni a csoporttagok közötti τ becslését. Azonban nézzük meg, hogy milyen eredményt ad, hátha segítenek majd a kapott eredmények valamelyest, hogy hogyan érdemes a becslést módosítani ezen esetben.

A 3. ábrán láthatóak ezen szimulációk kapott becslések, illetve a valódi értékek a fertőzési rátára. Az ábrán látható egyenes egyenlete $y=x+0,1$. Az nagyon szembeütő, hogy az ábrázolt pontok nagyon jól illeszkednek ezen egyenesre, vagyis a cikkben olvasható becsléssel minden kipróbált τ esetén nagyjából 0,1-el kaptunk nagyobb becslést a tényleges értéknél. Ez alapján úgy gondoljuk, hogy a cikkben olvasható becslés egy jó kiindulópont lehet a mi esetünkben is, és reméljük, hogy a képlet módosítása által tudunk egy jó becslést adni a fertőzési rátára ezen esetében is.

Mivel az általunk vizsgált gráfban kétféle fertőzési ráta is adott, emiatt úgy is tekinthetünk a cikkben adott becslésre ebben az esetben, mint az átlagos fertőzési rátára, amely függ a csoportokon belüli, illetve a háztartáson belüli fertőzési rátáktól is. Mivel a háztartáson belüli fertőzési ráta ismert, ezért vissza lehet számolni a csoportokon belüli fertőzési rátát ezen becslés értékéből.

Tehát a következő lépéseket tesszük:

- kiszámoljuk a cikkben található becslést, most azonban ez az átlagos fertőzési rátára adja meg a becslést:
- $$\hat{\tau}_{\text{átlagos}} = \frac{Z_I}{\sum_{t_i \leq \bar{T}} E_{t_i}^{SI} \cdot (t_i - t_{i-1})}$$
- Az átlagos fertőzési rátát meg tudjuk adni a csoporton belüli élek, illetve a háztartáson belüli élek összes élhez viszonyított részarányának a segítségével, amelyet a következő egyenlet segítségével tudunk felírni:
- $$\hat{\tau}_{\text{átlagos}} = (1 - \text{csoporton belüli élek aránya}) \cdot 1 + (\text{csoporton belüli élek aránya}) \cdot \hat{\tau}_{\text{csoporton belüli}}$$
- Ezen egyenletben 2 ismeretlen van, azonban ha jó becslést tudunk adni az átlagos fertőzési rátára, akkor az egyenletet átalakítva ki tudjuk fejezni a csoporton belüli fertőzési rátát ennek segítségével:
- $$\hat{\tau}_{\text{csoporton belüli}} = \frac{\hat{\tau}_{\text{átlagos}} - (1 - \text{csoporton belüli élek aránya}) \cdot 1}{\text{csoporton belüli élek aránya}}$$



4. ábra: a csoporttagok közötti fertőzési rátára adott korrigált becslés összevetése a tényleges fertőzési rátákkal kétféle τ esetén

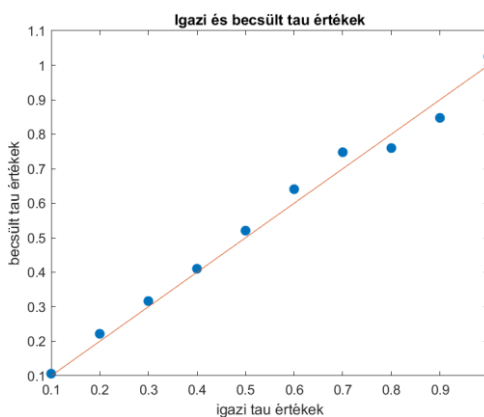
Az ábra alapján úgy néz ki, hogy az átlagos fertőzési rátából visszszámolt becslés viszonylag közel esett a tényleges csoporttagok közötti fertőzési rátához, viszont kis τ értékek esetén valamiért a becslés jelentősen eltért a ténylegesen kapott értéktől. Erre még nem sikerült rájönni, hogy kisebb τ értékek esetén mi okozza a rossz becslést. Egy lehetséges magyarázata ennek a rossz becslésnek az, hogy minél kisebb a csoporttagok közötti τ értéke, annál elszigeteltebbek a háztartások, és ezáltal esetleg ki is hal a járvány.

Háztartások egy egységként kezelve

Ezt követően módosítottam a korábbi modellen, és arra voltam kíváncsi, hogy a becslések így is megfelelőek maradna-e. A háztartásokra most egy egységként tekintettem, vagyis, hogyha egy személy a háztartás tagjai közül megbetegszik, akkor a háztartás összes többi tagja is megbetegszik, továbbá ha valamelyikük meggyógyul, akkor mindannyian meggyógyulnak. Vagyis ebben a módosított modellben nem egyének megbetegedéséről vagy gyógyulásáról beszélünk, hanem egész háztartásokéről. Ezáltal eltűnnek a modellből a háztartások közötti fertőzési ráták, és csak a csoporttagok közötti τ értékek maradnak meg. Így lényegében kiküszöböljük a kétféle τ értékből adódó problémát, de a háztartások közötti erősebb kapcsolat megmarad, mivel ezen modellben úgy tekintünk arra, hogyha a háztartás egyik tagja megbetegszik, hogy azonnal továbbadja a háztartás többi tagjának a betegséget.

A megvalósítás a következőképpen történt: Mivel a háztartásokat egy egységnek tekintjük, emiatt a háztartások tagjait összevontam egy csúcsba, ezáltal megszűntek a háztartások közötti 1 súlyú élek, de ezen csúcsból minden olyan csúcsba megy él, amelyből az eredeti gráfban a háztartás valamelyik tagjából ment él. Ezáltal keletkezett az eredetinel egy kisebb gráf, amelyben a csúcsok nem az egyéneknek felelnek meg, hanem a háztartásoknak.

A fertőzési ráta becslésénél, az SI élek megszámlálásánál többféleképpen is elindulhatunk. Első lépésben a háztartások gráfja alapján végeztem el a paraméterbecslést. Ezen esetben, ha megfertőződött egy háztartás, akkor 1 adódott hozzá a fertőzöttek számához, ha pedig meggyógyult, akkor pedig 1 adódott a gyógyultak számához. Ezzel a módszerrel az lehet a probléma, hogy nem veszi figyelembe, hogy a háztartások különböző létszámúak is lehetnek.



5. ábra: a tényleges, illetve a becsült fertőzési ráta azon esetben, amikor a háztartásokat egy egységnek tekintettük.

Az 5. ábrán láthatóak az igazi, illetve a becsült τ értékek, továbbá az $y=x$ egyenes. Mivel a kék pontok eléggé közel estek az egyeneshez, ezért úgy gondolom, hogy ez a becslés is jónak mondható. Azonban ezen modell nem veszi figyelembe, hogy a háztartások eltérő létszámúak lehetnek, ezáltal az egyes állapotokban lévő emberek száma nem a háztartások számával, hanem a benne élő emberek számával nő.

Ennek a módszernek egy lehetséges továbbfejlesztése a következő: Ha egy háztartás megbetegszik, akkor az SIR modellben nem a háztartások számát tartom nyilván, hanem az egyének számát. Vagyis a megbetegedés és gyógyulás háztartásonként történik, azonban nyilvántartjuk, hogy egy esemény során hány ember betegszik meg vagy pedig hány ember gyógyul meg. Mivel az előző esetben is elég jó becslést kaptunk a fertőzési rátára, feltételezhetően ez a módszer is jó becslést fog adni.

Jövőbeli tervek

A következő félévben szeretném folytatni a projekt eddigi 2 féléves munkáját. A paraméterbecslések során sikerült több esetben is jó becslést adni a járványterjedés paramétereire, azonban még maradtak nyitott kérdések. Érdemes lenne azzal foglalkozni, hogy kétféle fertőzési ráta esetén mi okozhatja, hogy kis τ értékekre nem működik a becslés, és hogyan lehetne ezen javítani. A fertőzési ráta becslésénél ismertnek tételeztük fel az SI élek számát, de az is egy érdekes kérdés, hogy vajon hogyan lehetne τ értékét megbecsülni abban az esetben, ha nem ismerjük az SI élek számát. Továbbá azzal is érdemes lenne foglalkozni, hogy mennyit javít a fertőzési ráta becslésén a háztartások létszámának ismerete azon esetben, amikor a háztartásokat egy egységként kezeltük.

Hivatkozások:

- [1] Backhausz Ágnes, Bognár Edit: Járványterjedés paramétereinek becslése többretegű véletlengráf-modellekben, Kézirat
- [2] Ágnes Backhausz, István Z. Kiss, Péter L. Simon: The impact of spatial and social structure on an SIR epidemic on a weighted multilayer network
- [3] Wasiur R. KhudaBukhsh, Boseung Choi, Eben Kenah, and Grzegorz A. Rempala: Survival dynamical systems: individual-level survival analysis from population-level epidemic models
- [4] Kiss., I.Z., Miller, J.C., Simon, P.L., Mathematics of Epidemics on Networks; From Exact to Approximate Models, Springer, Interdisciplinary Applied Mathematics 46, (2017).