

OTDK-dolgozat

2025

BUDAPESTI CORVINUS EGYETEM

ADATELEMZÉS ÉS INFORMATIKA INTÉZET



**KORONAVÍRUS TERJEDÉSÉNEK ELŐREJELZÉSE INTERNETES
KERESÉSI ADATOK ALAPJÁN A MAGYAR RÉGIÓKBAN**

**PREDICTING THE SPREAD OF CORONAVIRUS IN HUNGARIAN
REGIONS BASED ON INTERNET SEARCH DATA**

Büki Fanni

Konzulens: Kovács László

Kézirat lezárásának dátuma: 2023. 04. 17.

Tartalom

1. Bevezetés	1
2. A szakirodalmi áttekintés.....	2
2.1. A Google Trends bemutatása	2
2.2. A Google Trends és más internetes keresési indexek korábbi alkalmazásainak bemutatása a járványterjedés előrejelzésének területén.....	2
2.3. A keresési index kompozíció relevanciája	5
2.4. A keresési index kompozíció alapötlete.....	6
2.5. A keresési index kompozíció felépítése	7
2.6. Az idősorelemzés releváns alapfogalmainak bemutatása	8
2.7. Vektor Autoregresszió (VAR) modell bemutatása	10
2.7.1. Granger-okság és az Impulse-Response Function	11
2.8. Általánosított additív modell (GAM) bemutatása	12
3. Adatok bemutatása	15
3.1. Google Trends adatok	15
3.2. Koronamonitor adatok	16
4. Az elemzés folyamata.....	17
5. A módszertan alkalmazása, eredmények ismertetése.....	18
5.1. Közép-Magyarország: Budapest	20
5.2. Közép-Magyarország: Pest vármegye.....	26
5.3. Nyugat-Dunántúl.....	26
5.4. Dél-Dunántúl.....	27
5.5. Dél-Alföld	28
5.6. Közép-Dunántúl, Észak-Magyarország, Észak-Alföld:	29
6. Egy főre jutó GDP és a Granger-okság kapcsolata	29

7. Összegzés.....	32
8. Irodalomjegyzék.....	34
9. Mellékletek.....	37
1. számú melléklet: Keresési kifejezések volumenindexei régiókként.....	37
2. számú melléklet: Stacionaritás tesztek p-értékei régiókként	45
3. számú melléklet: Keresési index kompozíciók és koronavírus esetszámok IVF ábrái régiókként csoportosítva	49

Táblázatok jegyzéke

1. TÁBLÁZAT STACIONARITÁS TESZTEK P-ÉRTÉKEI ELSŐ ÉS MÁSODIK DIFFERENCIÁLÁS UTÁN BUDAPEST ESETÉBEN	17
2. TÁBLÁZAT VAR ÉS GAM MODELLEKBEN ELVÉGZETT TESZTEK EREDMÉNYEI.....	20
3. TÁBLÁZAT STACIONARITÁS TESZTEK P-ÉRTÉKEI ELSŐ ÉS MÁSODIK DIFFERENCIÁLÁS UTÁN PEST VÁRMEGYE ESETÉBEN.....	45
4. TÁBLÁZAT STACIONARITÁS TESZTEK P-ÉRTÉKEI ELSŐ ÉS MÁSODIK DIFFERENCIÁLÁS UTÁN DÉL-ALFÖLD RÉGIÓ ESETÉBEN.....	45
5. TÁBLÁZAT STACIONARITÁS TESZTEK P-ÉRTÉKEI ELSŐ ÉS MÁSODIK DIFFERENCIÁLÁS UTÁN DÉL-DUNÁNTÚL RÉGIÓ ESETÉBEN	46
6. TÁBLÁZAT STACIONARITÁS TESZTEK P-ÉRTÉKEI ELSŐ ÉS MÁSODIK DIFFERENCIÁLÁS UTÁN NYUGAT-DUNÁNTÚL RÉGIÓ ESETÉBEN	46
7. TÁBLÁZAT STACIONARITÁS TESZTEK P-ÉRTÉKEI ELSŐ ÉS MÁSODIK DIFFERENCIÁLÁS UTÁN ÉSZAK-ALFÖLD RÉGIÓ ESETÉBEN.....	47
8. TÁBLÁZAT STACIONARITÁS TESZTEK P-ÉRTÉKEI ELSŐ ÉS MÁSODIK DIFFERENCIÁLÁS UTÁN ÉSZAK-MAGYARORSZÁG RÉGIÓ ESETÉBEN.....	47
9. TÁBLÁZAT STACIONARITÁS TESZTEK P-ÉRTÉKEI ELSŐ ÉS MÁSODIK DIFFERENCIÁLÁS UTÁN KÖZÉP-DUNÁNTÚL RÉGIÓ ESETÉBEN	48

Ábrajegyzék

1. ÁBRA RÉGIÓNKÉNTI HATÁSOK A VAR ÉS GAM MODELLEK EREDMÉNYEI ALAPJÁN	19
2. ÁBRA BUDAPESTI KORONAVÍRUS ESETSZÁMOK IVF ÁBRÁJA A KERESÉSI INDEX KOMPOZÍCIÓBÓL ÉRKEZŐ SOKK HATÁSÁRA (95%-OS KONFIDENCIA-INTERVALLUM)	22
3. ÁBRA BUDAPESTI KERESÉSI INDEX KOMPOZÍCIÓ IVF ÁBRÁJA A KORONAVÍRUS ESETSZÁMOKBÓL ÉRKEZŐ SOKK HATÁSÁRA (99%-OS KONFIDENCIA-INTERVALLUM)	23
4. ÁBRA NEMLINEÁRIS (GAM) HATÁSOK BUDAPESTEN	25
5. ÁBRA NEMLINEÁRIS (GAM) HATÁSOK NYUGAT-DUNÁNTÚL RÉGIÓBAN	27
6. ÁBRA NEMLINEÁRIS (GAM) HATÁSOK DÉL-ALFÖLD RÉGIÓBAN	28
7. ÁBRA EGY FŐRE JUTÓ GDP ÉS GRANGER-OKSÁGI WALD-TEST P-ÉRTÉKEK RÉGIÓK SZERINT	30
8. ÁBRA EGY FŐRE JUTÓ GDP ÉS GRANGER-OKSÁGI WALD TEST P-ÉRTÉKEK RÉGIÓK SZERINT (LOGARITMÁLT VÁLTOZAT)	31
9. ÁBRA KERESÉSI KIFEJEZÉSEK VOLUMENINDEXEI BUDAPESTEN I.	37
10. ÁBRA KERESÉSI KIFEJEZÉSEK VOLUMENINDEXEI BUDAPESTEN II.	37
11. ÁBRA KERESÉSI KIFEJEZÉSEK VOLUMENINDEXEI PEST VÁRMEGYÉBEN I.	38
12. ÁBRA KERESÉSI KIFEJEZÉSEK VOLUMENINDEXEI PEST VÁRMEGYÉBEN II.	38
13. ÁBRA KERESÉSI KIFEJEZÉSEK VOLUMENINDEXEI NYUGAT-DUNÁNTÚL RÉGIÓBAN I.	39
14. ÁBRA KERESÉSI KIFEJEZÉSEK VOLUMENINDEXEI NYUGAT-DUNÁNTÚL RÉGIÓBAN II.	39
15. ÁBRA KERESÉSI KIFEJEZÉSEK VOLUMENINDEXEI DÉL-DUNÁNTÚL RÉGIÓBAN I.	40
16. ÁBRA KERESÉSI KIFEJEZÉSEK VOLUMENINDEXEI DÉL-DUNÁNTÚL RÉGIÓBAN II.	40
17. ÁBRA KERESÉSI KIFEJEZÉSEK VOLUMENINDEXEI KÖZÉP-DUNÁNTÚL RÉGIÓBAN I.	41
18. ÁBRA KERESÉSI KIFEJEZÉSEK VOLUMENINDEXEI KÖZÉP-DUNÁNTÚL RÉGIÓBAN II.	41
19. ÁBRA KERESÉSI KIFEJEZÉSEK VOLUMENINDEXEI ÉSZAK-MAGYARORSZÁG RÉGIÓBAN I. ..	42
20. ÁBRA KERESÉSI KIFEJEZÉSEK VOLUMENINDEXEI ÉSZAK-MAGYARORSZÁG RÉGIÓBAN II. ..	42
21. ÁBRA KERESÉSI KIFEJEZÉSEK VOLUMENINDEXEI ÉSZAK-ALFÖLD RÉGIÓBAN I.	43
22. ÁBRA KERESÉSI KIFEJEZÉSEK VOLUMENINDEXEI ÉSZAK-ALFÖLD RÉGIÓBAN II.	43
23. ÁBRA KERESÉSI KIFEJEZÉSEK VOLUMENINDEXEI DÉL-ALFÖLD RÉGIÓBAN I.	44
24. ÁBRA KERESÉSI KIFEJEZÉSEK VOLUMENINDEXEI DÉL-ALFÖLD RÉGIÓBAN II.	44
25. ÁBRA PEST VÁRMEGYE KORONAVÍRUS ESETSZÁMAINAK IVF ÁBRÁJA A KERESÉSI INDEX KOMPOZÍCIÓBÓL ÉRKEZŐ SOKK HATÁSÁRA (95%-OS KONFIDENCIA-INTERVALLUM)	49
26. ÁBRA NYUGAT-DUNÁNTÚL KERESÉSI INDEX KOMPOZÍCIÓ IVF ÁBRÁJA A KORONAVÍRUS ESETSZÁMOKBÓL ÉRKEZŐ SOKK HATÁSÁRA (90%-OS KONFIDENCIA-INTERVALLUM)	49
27. ÁBRA DÉL-DUNÁNTÚL KORONAVÍRUS ESETSZÁMAINAK IVF ÁBRÁJA A KERESÉSI INDEX KOMPOZÍCIÓBÓL ÉRKEZŐ SOKK HATÁSÁRA (95%-OS KONFIDENCIA-INTERVALLUM)	50

1. Bevezetés

Dolgozatom célja a koronavírus terjedésének előrejelzése, internetes keresési adatok alapján, a járvány különböző hullámaiban, magyarországi régiók szerint külön-külön csoportosítva. Az előrejelzést a Li, Liang, Li, Liu, Feng, és Shao (2020) tanulmánya által bemutatott internetes keresési kifejezésekből készített keresési index kompozíció magyar adatokra készített változatának segítségével végzem, melynek felépítését a 2.5. fejezetben fejtem ki részletesen. Elemzésem további célkitűzése a régiók közti predikciós minőségben tapasztalható eltérések szignifikanciájának vizsgálata.

Napjainkban már szigorú korlátozások nélkül élhetjük életünket, viszont bármikor előfordulhat, hogy egy újabb COVID-19 variáns vagy egy hasonlóan veszélyes új vírustörzs miatt ismét szükség lehet a korábbiakhoz hasonló intézkedésekre. Ezek szigorúságának és ütemezésének meghatározásához hasznos lehet egy olyan eszköz, mely segítségével megbecsülhetjük a vírus terjedésének ütemét.

Az utóbbi években, különösen a COVID-19 vírus járvány kitörése óta, egyre több figyelmet kaptak a Google Trends, illetve más keresőmotorok volumenadatai, melyeket egyre több kutatásban alkalmaznak különböző fertőző vírusok terjedésének prognosztizálására (Ben, Xin, Chen, Jiang, Yuan, Su, Christiani, Zhang, Du & Wang, 2022; Fantazzini, 2020). Dolgozatomban az előrejelzéshez magyar nyelvű internetes keresési adatokat használok a Google Trends által szolgáltatott keresési volumenadatok alapján. Szemben a hivatalos statisztikákkal, az internetes keresési adatok közel valós időben elérhetőek, ami előnyt jelent az előrejelzés szempontjából. A magyar és angol nyelvű szakirodalom általam jelenleg áttekintett része alapján, korábban ilyen elemzés magyarországi adatokra még nem készült, így a kutatás egyedülálló lehet hazai viszonylatban.

A kiinduló adatok és az elvégzett elemzés kódja a következő linken tekinthető meg:

https://unicorvinus-my.sharepoint.com/:f:/g/personal/fanni_buki_stud_uni-corvinus_hu/Ei1L5xERSmhNuD0ULtTKYwsBFY6WmF-4ZwmxwV_WXcjsQ?e=SGWNbG

2. A szakirodalmi áttekintés

2.1. A Google Trends bemutatása

A Google Trends valós Google keresési adatokból készített, nem szűrt, anonim, témák szerint kategorizált és aggregált minta. Ez az adattömeg lehetővé teszi, hogy a Google kereséseket globális, régió, megye, vagy akár város szinten megvizsgáljuk. A volumenadatok egyes kifejezésekre vagy ritkábban azok kombinációira érhetők el (Rogers, 2016).

Ezen felül a megfigyelések kiválasztott időszakra és tetszőleges fogalomra megtekinthetők, illetve letölthetők a Google Trends oldaláról. Már letöltés nélkül is reális képet kaphatunk az adatok alakulásáról, mivel az oldal beépített diagramon időszorként ábrázolja a feltételeknek megfelelően szűrt adatokat. A keresett fogalmakra vonatkozó megfigyelések 2004-től kezdve érhetők el (Google, 2022b).

Az idősoroknak kétféle verziója létezik: valós idejű és nem valós idejű. A valós idejű adatok 7 napra visszamenőleg tekinthetők meg. A nem valós idejű adatok egy ettől eltérő random mintában szerepelnek, ezek az adatok 2004-től egészen $t - 72$ óráig elérhetők. Elemzésem során ez utóbbit fogom használni (Google, 2022a).

Adatvédelmi okokból az egyes időpontokban nem a konkrét gyakoriság, hanem egy 1-100-ig terjedő skálára átszámított érték látható, ez az úgynevezett keresési volumenindex (röviden volumenindex). Ezen skála maximum értéke mindig az adott időszak keresési gyakoriságainak maximuma az adott területen, ehhez viszonyítják az intervallumba eső többi időpont keresési volumeneit (Rogers, 2016).

2.2. A Google Trends és más internetes keresési indexek korábbi alkalmazásainak bemutatása a járványterjedés előrejelzésének területén

McGough, Brownstein, Hawkins és Santillana (2017) kísérletet tettek a szűnyogokon keresztül terjedő, Latin-Amerikában járványt okozó Zika vírus terjedésének internetes adatok alapján történő előrejelzésére. A hivatalos kormányzati rendszerek időbeli korlátjai és a helyzet súlyossága miatt szükségessé vált egy olyan eszköz létrehozása, mellyel megkönnyíthetik a vírus terjedésének valós időbeli követését, ezzel lehetővé téve a terjedés sebességének csillapítását. A tanulmány megállapította, hogy az internetes adatok a jelenlegi rendszer kiegészítőjeként használhatók. Alkalmazásuk kulcsfontosságú a járvány terjedési ütemének megfigyelésében és annak mérséklésében.

Hasonló megközelítéssel készült a Lymelight is, amely egy új, internetes keresésre épülő, adat alapú módszer a Lyme-kór terjedésének valós idejű követéséhez. A modell hasznos kiegészítés lehet, a sokszor egy évnél is hosszabb késleltetéssel megjelenő állami riportokhoz képest (Sadilek, Hswen, Bavadekar, Shekel, Brownstein & Gabrilovich, 2020).

Tajvanban demenciával kapcsolatos kutatást végeztek, melynek keretei között a demencia fogalom keresési volumenindexének és a demenciás esetek, illetve a járóbeteg-ellátásban ápoltak száma közti keresztkorrelációt vizsgálták, melynek fogalmát a 2.6. alfejezetben fejtem ki részletesen. A kutatás eredményei szerint a demencia kifejezés keresési volumenindexe alapján előre jelezhetőek a demenciás esetek és a járóbeteg-látogatások száma is. Ez segítheti az egészségügyi szervek felkészülését a szűrővizsgálatok lebonyolítására, mivel előzetes képet kapnak a várható terhelésről (Wang, Chen, Yu & Chen, 2015).

Syamsuddin, Fakhrudin, Sahetapy-Engel és Soewono (2020) tanulmányukban a Dengue-láz új esetszámainak és a betegséggel kapcsolatos legnépszerűbb kifejezések (*demam berdarah dengue* és *dbd*) keresési volumenindexének kointegrációját (hosszútávú közösnek tekinthető pályával rendelkezését) vizsgálták az indonéziai Bandungban. A kointegrációt Engle-Granger teszttel vizsgálták. Az eredmények alapján az idősorok kointegráltak, tehát van közösnek tekinthető hosszútávú pályájuk, így hibakorrekciós (ECM) modellt építettek. A modellben a hibakorrekciós tényező nem lett szignifikáns, de rövid távon Granger-okság kimutatható volt. Granger-okság akkor áll fenn két idősor között, ha az egyik idősor legalább olyan jól előre jelezhető egy másik idősorral, mint a saját múltjával. A fogalmat a 2.7.1. alfejezetben részletesebben is kifejtem. Az eredmények alapján tehát a keresési adatok hasznosnak bizonyultak a Dengue-láz járvány kezdeti előrejelzéséhez.

Xu, Wang, Yang, Hou, Sun, Zhang, Cao, Hou, Wang, Cai & Wang 2019-es kutatásában megfigyelhető, hogy a megközelítés és a módszer is hatással lehet a végeredmény értelmezhetőségére, használhatóságára. A kínai polgárok rákos elváltozásokkal/betegségekkel kapcsolatos internetes keresései és a rákos megbetegedés miatt életüket veszített emberek számának kapcsolatát vizsgáló tanulmány például nem használt időbeli késleltetéseket a számítások elvégzésekor. Így az eredmények sem megbízhatóak, hiszen a betegek leggyakrabban nem abban a hónapban haláloznak el, amelyikben a keresést végrehajtották.

A következő, Litvániában készült tanulmány az influenzajárvány előrejelzését tűzte ki célul. Austys és Burneikaitė 2019-es tanulmányukban a litván Google Trends keresési volumenindexek segítségével dolgoztak. A lehetséges keresési kifejezéseket más nyelvű

tanulmányokban leggyakrabban használt keresési kifejezések litván nyelvre fordításával határozták meg. Ezt követően letöltötték a Google Trendsről a litván keresési kifejezésekhez tartozó volumenindexeket 5 évre visszamenőleg. Ezután kiszámították a keresztkorrelációs értékeket minden keresési kifejezés volumenindexe és a valós fertőzöttek száma között. A legmagasabb korrelációs együtthatóval rendelkező kifejezés keresési volumenindexe és a két héttel későbbi valós fertőzöttek száma alapján készítettek egy autoregresszív integrált mozgóátlag (ARIMA) modellt. Ezután az előrejelzett értékek és a megfigyelt betegek száma közti keresztkorreláció alapján megállapították, hogy a Google Trends adatok potenciális segítséget nyújthatnak a litván influenzajárvány alakulásának előrejelzésében. Litvánia lakossága mindössze 2,8 millió fő, melyhez hasonló nagyságrendű Magyarország 9,69 millió fős lakossága (Központi Statisztikai Hivatal, 2022b). Mivel Magyarország populációja nagyobb, mint Litvániáé, így ez a kutatás arra enged következtetni, hogy a feltehetőleg nagyobb volumenű magyar internetes keresési adatokra is érdemes lehet hasonló elemzést végrehajtani.

Ben et al. 2022-ben megjelent kutatásában a bélrendszeri megbetegedési tünetek keresési volumenindexeinek előrejelző hatását vizsgálták mind globális, mind regionális COVID-19 járványkitörésekre. A kutatást földrajzilag az Amerikai Egyesült Államokra, az Egyesült Királyságra és Ausztrálira, valamint Kínára szűkítették. Mivel Kínában a Baidu keresőmotor használata a legelterjedtebb, globális szinten pedig a Google-t használják a legtöbben, ezért ebben a tanulmányban mindkettőt alkalmazták. Egyes tünetekre vonatkozó GAM modellekkel vizsgálták a tünetek előrejelző képességét. Az eredmények arra utalnak, hogy a bélrendszeri megbetegedési tünetek közül különösen a *hasmenés* és az *ízlelés elvesztése* kifejezések keresési volumenindexei segíthetnek a koronavírus terjedésének monitorozásában és a COVID-19 járvány kitörésének akár három hetes időtávlatú előrejelzésében.

Fantazzini (2020) elemzésében 158 ország adatait felhasználva készített és versenyeztetett 18 modellt, kéthetes előrejelzési késleltetéssel. A kutatás során a tüdőgyulladás és a koronavírus keresési kifejezések keresési volumenindexeinek predikciós erejét vizsgálta a koronavírus fertőzöttek számát illetően. Annak érdekében, hogy pontos képet kapjon, az egyes keresési kifejezések keresési volumenindexeit, kizárólag a Google által egészség kategóriába sorolt keresésekre szűkítette. Erre azért volt szükség, mert a hírek befolyásolhatják az emberek keresési szokásait, így ugrásszerűen növekedhetnek az egyes kifejezések keresési volumenei. A fenti lépéssel elkerülték, hogy a hírekkel kapcsolatos keresési volumenek belekerüljenek az elemzésbe. Az eredmények alapján elmondható, hogy az egészség kategórián belül a

koronavírus és a *tüdőgyulladás* Google keresési volumenindexe sok országra nézve jól használható az új napi esetszámok előrejelzésére.

2.3. A keresési index kompozíció relevanciája

A következő fejezetekben központi szerepet játszik az úgy nevezett keresési index kompozíció, mely az adott fertőző betegség szempontjából releváns keresési kifejezések keresési volumenének információtartalmának tömörítését célozza.

Huang, Luo, Duan, Zhang, Zhang, Tang, Smith, Li és Zou 2020-ban megjelent tanulmánya a Baidu Keresési Index és a HIV/AIDS, szifilisz, illetve gonorrhoea fertőzöttek számának kapcsolatát vizsgálta Kínában. A gondosan kiválasztott kulcsszavakból keresési index kompozíciót készítettek, melyet a később komponált előrejelző vektor autoregresszió (VAR) modell egyik fő magyarázóváltozójaként használtak. A VAR modell fogalmát és elméleti hátterét a 2.7. alfejezetben mutatom be részletesen. Ezen kívül az eredményváltozó (új diagnosztizált esetek száma) időben eltolt változatai kerültek magyarázóváltozóként a modellbe. A kialakított VAR modellel készített előrejelzésekből az a következtetés vonható le, hogy a Baidu keresési index használható az új HIV/AIDS, szifilisz és gonorrhoea fertőzöttek számának monitorozására és előrejelzésére Kínában.

Li, Liu, Feng, Ning, Ou, Sun, Wei, Liang & Shao (2019) a Nature folyóiratban megjelent tanulmánya, a HIV/AIDS Baidu Index keresési volumenek és a ténylegesen megfigyelt betegek száma közti kapcsolatot vizsgálta. A fent említett Huang et al. 2020-as tanulmányában bemutatott elvhez hasonlóan kompozíciós keresési indexet alkottak a szerzők, mely később az épített modell változójaként szolgált. VAR modell helyett ebben a tanulmányban külső változókkal kiegészített autoregresszív integrált mozgóátlag (ARIMAX) modellt készítettek a kutatók. A VAR és az ARIMAX modell annyiban különböznek egymástól, hogy a VAR modellben az eredményváltozó és a külső magyarázóváltozók késleltetettjei egyaránt szerepelnek, az ARIMAX modellben pedig csak az előbbiek, illetve a külső magyarázóváltozó változó t -beli értéke. Az eredmények alapján a Baidu kompozíciós index-szel kibővített ARIMAX modell az egyszerű autoregresszív integrált mozgóátlag (ARIMA) modellhez képest kisebb Akaike információs kritérium értékkel, melynek fogalmát a 2.7. alfejezetben részletesen kifejtem, és pontosabb előrejelzett értékekkel rendelkezik.

2.4. A keresési index kompozíció alapötlete

A keresési index kompozíció alapötletét és felépítését szeretném Li et al. (2020) és Huang et al. (2020) tanulmánya alapján részletesebben bemutatni. Az elemzésemben én is ezen tanulmányokban leírt módszer alapján készítettem el a keresési index kompozíciót.

Li et al. 2020-as tanulmányukban az internetes keresési adatok COVID-19 járvánnyal kapcsolatos predikciós erejét vizsgálták. Mindezt az előző fejezetben említett keresési index kompozíció megalkotásával tették. Az itt használt kompozíciós index elve megegyezik a fent említett, többségében HIV/AIDS, illetve szifilisz és gonorrhoea fertőzöttek számának előrejelzésére irányuló tanulmányokban használtakal (Huang et al., 2020; Li et al., 2019). Ezt a kompozíciót később magyarázóváltozóként használták a készített modellekben.

Li et al. 2020-ban megjelent tanulmányában a Google Trends helyett Baidu keresési indexet használtak. Míg a Google Trends érték a keresési érdeklődést reprezentálja egy 0 és 100 között terjedő skálán, ahol 100 az adott időszak keresési volumeneinek maximuma, addig a Baidu keresési index egy-egy keresési kifejezés keresési gyakoriságának súlyát reprezentálja (Ben et al., 2022).

A kompozíció összeállításakor a cél az, hogy az a téma szempontjából, jelen helyzetben a COVID-19 szempontjából legrelevánsabb keresési kifejezéseket tartalmazza. Li et al. (2020) tanulmányában többféle módszer együttesét használták a megfelelő kifejezések meghatározásához. Elsősorban egy kínai weboldalról gyűjtötték össze a potenciális kulcsszavakat, majd ezek listájához hozzáadták egy szemantikus korreláció analízis eredményeként kapott kifejezéseket. Az elemzéshez felhasznált oldalakra néhány példa a Baidu, Microblog, Wikipedia és a WeChat. Így sikerült összegyűjteni tizennégy COVID-19-cel kapcsolatos keresési viselkedést tükröző kifejezést. Mivel nem minden kifejezés kapcsolódik közvetlenül a COVID-19 vírushoz, így tovább szűkítették a mintát releváns kulcsszavakra.

A szűkítéshez a keresési kifejezések Baidu keresési indexeinek különböző mértékű késleltetettjei és az esetszámok közti keresztkorrelációt vizsgálták. Az eredmények alapján eltávolították a gyűjteményből azokat a szavakat, melyekre a keresztkorrelációs együttható minden késleltetés esetén abszolút értékben kisebb volt, mint 0,4. Ezt követően, minden kifejezés azon késleltetettjét használták fel, amely legszorosabban korrelált az esetszámokkal (Huang et al., 2020).

A kutatás által alkalmazott, fentebb ismertetett, szűrési eljárás alapján COVID-19-cel legtöbbször asszociált kifejezések a következők (Li et al., 2020):

- az új típusú koronavírus jellemzői
- új típusú koronavírus tünetek
- az új típusú koronavírus lefolyása (tükörfordítás: Melyik napon a legsúlyosabbak az új típusú koronavírus tünetei)
- az új típusú koronavírus terjedése valós időben
- az új típusú koronavírus első tünetei.

Az eredményül kapott öt keresési kifejezést használták fel a keresési index kompozíció kialakításához.

Dolgozatomban a tizennégy elemű kiinduló halmazba tartozó kifejezések magyar fordításának magyarországi régió szintű Google Trends keresési volumenindexeit fogom megvizsgálni a fent leírt szűrési módszerrel, a 2020 október 1. és 2022 február 28. közötti időszakra vonatkozóan.

2.5. A keresési index kompozíció felépítése

Huang et al. (2020) tanulmányában az egyes kifejezések súlyát a kifejezésenkénti legmagasabb abszolút értékű korrelációs együttható (ρ) alapján határozták meg, ezzel az értékkel mérték egy adott kifejezés a hatásnagyságát.

A súlyok számítását az alábbi képlet alapján végezték:

$$Súly_i = \frac{\rho_i}{\sum_{i=1}^n \rho_i} \quad (1)$$

ahol n a kulcsszavak száma és ρ_i az i -edik kulcsszó legmagasabb abszolút értékű korrelációs együtthatója (Huang et al., 2020).

A keresési index kompozíció képlete:

$$COVID - 19 \text{ Keresési Index} = \sum_{i=0}^n súly_i kulcsszó_i \quad (2)$$

ahol n a kulcsszavak száma, $súly_i$ és $kulcsszó_i$ az i -edik kulcsszó súlyát és napi Baidu indexét (BDI) reprezentálja (Huang et al., 2020).

2.6. Az idősorelemzés releváns alapfogalmainak bemutatása

Elemzésem során többféle modellt építettem és tesztet alkalmaztam, melyek egységes fogalmi keretben vett tárgyalásához elengedhetetlen az idősorelemzés néhány alapfogalmának bemutatása.

Idősorok esetén alapfogalomnak tekintjük a lagot, amely magyarra fordítva késleltetést jelent. Egy változó k -adik lagja/késleltetettje, az adott változó k időegységgel korábbi értékét jelöli (Kirchgässner, Wolters & Hassler, 2013).

A stacionárius, avagy azonos eloszlású idősoroknak két fajtáját különböztetjük meg. Erős stacionaritás esetén az idősor eloszlása időben állandó, tehát annak valószínűsége, hogy az idősor felvesz egy tetszőleges x értéket, minden t időpontra nézve azonos módon számítható, a számítás módja független t -től. Mivel ez a feltétel a gyakorlatban legtöbb esetben nem, vagy csak nagyon nehezen ellenőrizhető, ezért bevezették a gyenge stacionaritás fogalmát. Ha egy idősor gyengén stacionárius, akkor annak várható értéke és szórása időben állandó konstans, t -től független érték. Gyenge stacionaritásra léteznek tesztek, melyeket a későbbiekben használni is fogok. Ilyen teszt a Kwiatkowski–Phillips–Schmidt–Shin (kpss) és az Augmented Dickey–Fuller (adf) is. Kpss teszt nullhipotézise (H_0) szerint az idősor stacionárius viselkedésű, alternatív hipotézise (H_1) szerint nem stacionárius. Az adf teszt hipotézisei ellentétes jelentéssel bírnak, nullhipotézis esetén (H_0) nem stacionárius, míg alternatív hipotézis esetén (H_1) stacionárius viselkedésűnek tekinthető az idősor. Érdekes minkét tesztet alkalmazni, így minkét logika mentén megvizsgálhatjuk az idősor stacionaritás szempontjából. Amennyiben az idősor nem stacionárius, differenciálással sok esetben azzá alakítható. A differenciált idősor elemeit az $\Delta Y_t = Y_t - Y_{t-1}$ képlet alapján számoljuk (Woodward, Gray & Elliott, 2017).

A hamis korrelációk felfedezésében fontos szerepet játszanak a stacionárius idősorok. Az ilyen esetekben, úgy tűnik, mintha az idősorok korrelálnának egymással, azonban valójában nincs köztük valós kapcsolat csupán a trendjük vagy szezonális mintázatuk azonos. Erre nagyon jó példa az Egyesült Államokbeli Maine régió területén nyilvántartott válási arány és az egy főre jutó margarinfogasztás kapcsolata 2000 és 2009 között. Ha a két idősor értékeit egy diagramon ábrázoljuk láthatjuk, hogy meglepően hasonló tendenciát mutatnak. A korrelációs együttható értéke a két változó között 0,9925, tehát ezek alapján majdnem függvényszerű kapcsolat kell legyen a két változó között. Azonban, ha belegondolunk a valóságban a két jelenség aligha áll kapcsolatban egymással. Ezt könnyen bizonyítani is tudjuk stacionaritás-vizsgálattal. Amennyiben idősoraink nem stacionáriusak, szinte teljesen biztosak lehetünk benne, hogy

hamis korrelációval állunk szemben. Az ilyen és ehhez hasonló félrevezető jelenségeket kiküszöbölhetjük, amennyiben a nem stacionárius idősorokat differenciálással stacionerré alakítjuk elemzés előtt (*Spurious correlations*, n.d.).

Az autokorrelálatlan idősorok esetében, az adott Y_t elem értéke független attól, hogy mi volt a Y_{t-k} -adik elem. Amennyiben az idősorban trendszerű változások figyelhetőek meg, szinte biztosan kijelenthető, hogy az adott idősor nem *autokorrelálatlan*, tehát adott Y_t elem korrelál valamelyik késleltetettjével. Autokorrelálatlanságot az autokorreláció függvény (acf) és a parciális autokorreláció függvény (pacf) segítségével vizsgáljuk. A pacf parciális része arra vonatkozik, hogy az Y_t és Y_{t-k} időpontok közti korrelációt úgy adja meg, hogy kiszűri az összes többi időpont hatását 1-től $t - (k - 1)$ és $(k + 1)$ -től T késleltetett időpontokig (Kirchgässner et al., 2013).

Amennyiben az adott idősor hibatagjai stacionáriusak és autokorrelálatlanok, akkor a hibatagok fehérzajnak nevezhetők. A fehérzaj tesztek azt vizsgálják, hogy egy adott késleltetésig nézve nullának vehetők-e az autokorrelációk együttesen. Ilyen teszt például a Ljung-Box teszt és a Breusch–Godfrey teszt is. Mindkét teszt esetén a nullhipotézis (H_0) szerint fehérzaj viselkedésűek az idősor hibatagjai (Shumway & Stoffer, 2000).

Breusch-Godfrey teszt egy stacioner idősorra OLS regressziót készít, melynek képlete:

$$\hat{u}_t = \alpha_0 + \alpha_1 X_{1t} + \dots + \alpha_k X_{kt} + \dots + \rho_1 \hat{u}_{t-1} + \rho_p \hat{u}_{t-p} + e_t \quad e_t \sim WN(0, \sigma_e^2) \quad (8)$$

ahol u_t az eredményváltozó, és u_t lagjai a magyarázóváltozók, k az aktuális késleltetés száma, amelyek maximum p lagig nyúlnak vissza a múltba, X_{kt} pedig az esetleges szezonális vagy trend hatást jelölő dummy vagy időindex változókat jelzi. Nullhipotézise (H_0), hogy az u_t -re felírt regresszió $R^2 = 0$, tehát az u_t idősor a fehérzajnak tekinthető a nem megfigyelt sokaságban (Woodward et al., 2017).

Keresztkorreláció esetében az idősorok közti kapcsolatot vizsgáljuk. Azt mutatja meg, hogy van-e X_{t-k} elem és Y_t között kapcsolat, ha igen, akkor milyen irányú és erősségű. A keresztkorrelációkból leszűrhető, hogy egy Y_t idősor mekkora k késleltetés esetén korrelál legerősebben egy másik, X_t idősorral. Az idősorok közötti időben eltoltsági kapcsolatok előzetes vizsgálataként is tekinthetünk rá, melyek a VAR modell alapjait képezik (Kirchgässner et al., 2013).

2.7. Vektor Autoregresszió (VAR) modell bemutatása

A VAR modell több stacionárius idősor együttes vizsgálatakor alkalmazható. Alapötlete arra épül, hogy az idősorokat egy olyan rendszerben magyarázzuk meg, ahol minden idősor saját egyenlettel rendelkezik. A különbség a klasszikus autoregresszív modellekhez képest az, hogy az egyes egyenletekben az eredményváltozó saját múltja mellett a rendszerben lévő többi idősor késleltetettjei is szerepelnek. Az eredményváltozók vektort alkotnak az említett egyenletrendszerben. Minden esetben egy meghatározott k lagig (késleltetésig) nyúlunk vissza a múltba, melynek értéke erősen függ a megfigyelések számától (Woodward et al., 2017).

A VAR modell képlete két idősor esetén:

$$Y_t = \beta_0 + \beta_{11}Y_{t-1} + \beta_{12}Y_{t-2} + \dots + \beta_{1k}Y_{t-k} + \beta_{21}X_{t-1} + \beta_{22}X_{t-2} + \dots + \beta_{2k}X_{t-k} + u_t \quad (6)$$

$$X_t = \alpha_0 + \alpha_{11}Y_{t-1} + \alpha_{12}Y_{t-2} + \dots + \alpha_{1k}Y_{t-k} + \alpha_{21}X_{t-1} + \alpha_{22}X_{t-2} + \dots + \alpha_{2k}X_{t-k} + v_t \quad (7)$$

Ahol Y és X az eredményváltozók, u_t és v_t a hibatagok, t a jelen időpontot jelöli, k pedig a késleltetések száma (Woodward et al., 2017).

A modell fontos alapfeltevései közé tartozik, hogy a hibatagok (u_t, v_t) együttesen fehérzaj viselkedésűek. Ennek tesztelésére alkalmas a Portmanteau Test, melynek nullhipotézise (H_0) szerint a hibatagok együttesen fehérzajnak tekinthetők, tehát a hibatag nemcsak a saját múltjával, de a többi hibataggal sem korrelál. Az együttes hibatag fehérzaj teszt helyett, az egyenletek hibatagjainak viselkedését külön-külön is vizsgálhatjuk a 2.6. alfejezetben említett Breusch-Godfrey teszt segítségével. Ezt a tesztet egy VAR modell esetében az egyes egyenletek hibatagjaira külön-külön kell alkalmazni. Egy másik fontos feltevése a VAR modellnek, hogy az idősorok melyek kapcsolatát leírja stacionáriusak legyenek. Ez a feltétel fontos szerepet játszik a hamis korrelációk kiküszöbölésében, amely a végeredmény szempontjából fontos, hiszen csak így állapíthatunk meg tényleges összefüggéseket. A modell együtthatóit OLS módszerrel becsüljük meg (Kirchgässner et al., 2013).

A k értékének meghatározására az *információs kritériumokat* alkalmazzuk. Az információs kritériumok, csak úgy, mint az általános regresszió esetében, a magyarázóváltozók számának mérlegelésére alkalmasak. Legtöbb esetben az Akaike, a Hannan-Quinn és a Bayes-Schwarz információs kritériumokat együttesen alkalmazzák. Mindegyik mutató az OLS becslés célfüggvényéből, az összesített négyzetes hibából (SSE, Sum of Squared Errors) indul ki, amelyet az egyenletben szereplő paraméterek számának egy függvényével büntetnek. A különbség köztük a büntetőtagok szigorúsága. Amíg az Akaike viszonylag megengedő, a

Hannan-Quinn már szigorúbb, a Bayes-Schwarz pedig a legkevésbé toleráns egy új magyarázóváltozó modellbe építése esetén (Hastie, Tibshirani, & Friedman, 2011).

Az egyes információs kritériumok képletei:

$$AIC = \frac{SSE}{n} e^{2p/n} \quad (3)$$

$$HQC = \frac{SSE}{n} (\ln n)^{2p/n} \quad (4)$$

$$BIC = \frac{SSE}{n} n^{p/2} \quad (5)$$

ahol p a becsült paraméterek (β -k) száma tengelymetszettel együtt, SSE a modell teljes négyzetes hibája, n pedig az elemszám (Wood, Goude & Shaw, 2015).

Ahogy a képletekben látható, minden információs kritérium számításának alapja az SSE azaz a modell teljes hibája. Ezt kiegészíti minden esetben egy p -től függő büntetőtag. Összességében elmondható, hogy amennyiben két modellt összehasonlítunk, azt tekinthetjük jobbnak, melynek esetében az IC-k értéke kisebb, tehát összességében kevesebbet hibázik. Ezen elv alapján tudjuk kiválasztani a késleltetések optimális számát VAR modellek esetén is, hiszen ebben az esetben egy-egy késleltetés hozzáadása eredményváltozók számától függően több új magyarázóváltozó bevonásával is járhat (Shumway & Stoffer, 2000).

2.7.1. Granger-okság és az Impulse-Response Function

Két idősor közti Granger-okság vizsgálatkor azt tudjuk meghatározni, hogy Y idősor aktuális értékére szignifikáns hatással bír-e X idősor valamelyik késleltetettje. Amennyiben igen azt mondhatjuk, hogy X idősor Granger okozza Y idősort. A Granger-okságot egy hipotézisvizsgálattal tesztelhetjük. A Wald-teszt nullhipotézise szerint (H_0) egy VAR modell egyenletben, ahol Y_t az aktuális eredményváltozó, a magyarázóváltozó (X_t) összes késleltetettjének együtthatója nulla a teljes sokaságban. Ekkor X nem Granger okozza Y -t. Az alternatív hipotézis (H_1) szerint X Granger okozza Y -t, tehát az egyenletben szereplő magyarázóváltozónak van legalább egy olyan késleltetése melynek együtthatója nem nulla a mintán kívüli világban (Woodward et al., 2017).

Az impulzus-válaszfüggvény (IVF, angolul Impulse-Response Function, IRF) azt mutatja meg, hogy ha a modellben lévő változók valamelyikét t időszakban egy szórásnyi sokk éri, az milyen hatással lesz $t + k$ időpontokban az összes többi változóra, illetve önmagára. Sokkolással foglalkozni abban az esetben érdemes, ha változók között kimutatható Granger-okság. Az

úgynevezett „sokkolási sorrend” (melyik változóval kezdjük a sokkolást, és milyen sorrendben folytatjuk több változó esetén a kezdeti t időpontban) fontos lehet az IVF-ek esetén, ugyanis a kezdeti sokkokat a változók közti egyidejű korrelációs mátrix *Cholesky felbontásával* határozzuk meg, ami érzékeny változóink sorrendjére a korrelációs mátrixban. Ezt a VAR modellnek átadott változók sorrendjével lehet befolyásolni. R nyelvben az irf függvény *ortho* paraméterének igazra állítása esetén az eredmény függ a sokkolási sorrendtől, míg hamis érték esetén a kapott eredmény sorrendfüggetlen (Kirchgässner et al., 2013).

A VAR modell alkalmas Granger-oksági kapcsolatok vizsgálatára. Azonban ahhoz, hogy valós kapcsolatot tudjunk felfedezni, a modellnek meg kell felelnie a fent említett alapfeltevéseknek. Ezt könnyen lehet igazolni az R nyelvben található `serial.test` függvénnyel, amely a Portmanteau és Breusch-Godfrey tesztek együttesen használja. Amennyiben a `serial.test` függvény eredményeképp elfogadjuk a nullhipotézist, az azt jelenti, hogy a modellünk hibatagjai együttesen fehérzajnak tekinthetők, tehát a hibatag a saját múltján kívül más egyenletek hibatagjával sem korrelál. Amennyiben el kell utasítanunk a nullhipotézist, még mindig lehet csupán a Breusch-Godfrey teszt segítségével vizsgálni az egyes egyenletek hibatagjait külön-külön (Shumway & Stoffer, 2000).

A modell stabilitásának tesztelésekor elmaradhatatlan az egységgyökök vizsgálata. A VAR modell csak abban az esetben tekinthető stabilnak, ha a tisztán autoregresszív részei az egyenleteknek (amelyekben az eredményváltozót saját késleltetettjeivel magyarázzuk) stacionáriusnak tekinthetők. Ezt az adott egyenlethez tartozó karakterisztikus polinom segítségével könnyen megvizsgálhatjuk. A polinom általános alakja (9) módon megadható:

$$z^k - \beta_{11} \times z^{k-1} - \beta_{12} \times z^{k-2} - \dots - \beta_{1k} \times z^{k-k} = 0 \quad (9)$$

Amennyiben a (9) kifejezésben minden z az egységkörön belül található, tehát abszolút értékben kisebb, mint egy, a modell stabilnak tekinthető (Woodward et al., 2017).

2.8. Általánosított additív modell (GAM) bemutatása

A dolgozatomban nem lineáris késleltetett hatásokat is vizsgáltam, így a VAR modell mellett általánosított additív modellt (angolul Generalized Additive Model, továbbiakban GAM) alkalmaztam.

A GAM használható nem-lineáris hatások modellezésére, ahol a modellben az egyes magyarázóváltozók marginális hatása együtthatók helyett tetszőleges $f(x)$ függvénnyel kifejezhető. A függvényekre csupán annyi megkötés vonatkozik, hogy a reprodukáló magfüggvényű Hilbert-térben legyenek (Kovács, 2022).

A modell általános alakja (3) kifejezéssel adott:

$$n(Y) = \varepsilon + \beta_0 + \sum_{j=1}^p \beta_1 X_j \quad (3)$$

A modellnek nem feltétele az $f(x)$ függvény pontos képletének ismerete. Ez gyakran egy spline függvény. A spline függvény több polinom (bázisfüggvények) lineáris kombinációja, azaz súlyozott összege. Ezekben a szakaszos függvényekben léteznek szakaszhatárok (knots), itt váltunk az egyes függvénytípusok között, illetve ide tartoznak az értelmezési tartomány szélei is (Hastie et al., 2011).

A spline függvény meghatározása három lépésből áll, melyek közül azonban egyik sem egyszerű. Elsőként meg kell határoznunk a bázisfüggvények számát, majd azok formáját, végül meg kell választanunk függvények közti határokat (knots) és meg kell becsülni súlyukat a lineáris kombinációban. A harmadik lépésre, azaz a súlyok meghatározására szerencsére alkalmazható az Ordinary Least Squares (OLS) módszer, melyet leggyakrabban a lineáris regresszió együtthatóinak becslésére használnak (Wood, 2017).

A teljes spline függvény meghatározására megoldást jelent az úgynevezett Thin Plate Splines (TPS) simító eljárás. A Thin Plate Splines függvénye:

$$\min ||Y - f||^2 + \lambda \int f''(x)^2 dx \quad (4)$$

Melyben az első rész funkciója a pontonkénti optimalizálás, a második rész pedig büntető tagként funkcionál az inflexiós pontok számának függvényében. A TPS módszerrel a büntetőtag jelenléte miatt a spline függvények száma és a csomópontok (knots) is automatikusan meghatározásra kerülnek a becslés során (Wood et al., 2015).

A becslés elméletben pontonként történik, azonban nem lehet annyi paraméterünk a gyakorlatban ahány megfigyeléssel rendelkezünk, hiszen, ha minden pontra megadunk egy értéket, akkor a modell tökéletesen fog illeszkedni a mintánkra, melynek következtében

túlillesztetté és a mintán kívüli világban előrejelzés szempontjából pontatlanná válik. Ennek elkerülése érdekében inkább a következő kifejezést minimalizáljuk:

$$\min ||Y - E_k \Psi_k||^2 + \lambda \Psi_k^T E_k \Psi_k \quad (5)$$

A formula megértésének szempontjából fontos E egy $n \times n$ dimenziós mátrix melynek sorai a megfigyelések és oszlopai az adott megfigyeléshez tartozó függvényértékek, tehát a mátrix minden megfigyelésre leírja az illesztett függvény helyettesítési értékét. Az (5) egyenletben E_k egy $n \times k$ dimenziós mátrix, ami legjobban közelíti a meghatározott k dimenziós térben az eredeti E mátrixot euklideszi normában. Ez alapján, meg kell keresnünk az első k legnagyobb sajátértékhez tartozó sajátvektort, mellyel az eredetihez képest a legtöbb információt meg tudjuk őrizni k dimenzióban. Ezen k sajátvektorok alkotják az E_k közelítő mátrixot. A lenti (6) egyenletben a végső $f(x)$ függvényértékek a csökkentett dimenziójú E_k mátrix oszlopértékeinek lineáris kombinációjaként állnak össze, melynek súlyai a Ψ_k -k (Kovács, 2022):

$$f(x) = E_k \times \Psi_k \quad (6)$$

Azért fontos, hogy E_k az E mátrix k dimenziós térben megadott legjobb közelítése legyen, mivel így tudjuk megőrizni a legtöbb információt az eredeti mátrixhoz képest. E_k esetében k egy irányadó szám, amit úgy választunk meg, hogy az a legkisebb olyan értéket vegye fel, ahol a két mátrix különbsége még nem tér el szignifikánsan a nullától. Így elérjük azt az állapotot, ahol a függvényreprezentáció méretének $n \times n$ dimenzióról $n \times k$ dimenzióra csökkentésével, kevesebb paraméter segítségével közelítőleg ugyanakkora mértékű információt tudunk megőrizni, ugyanakkor a túlillesztés problémáját is kiküszöböltük (Wood, 2017).

Ezután egy χ^2 próbával célszerű illeszkedésvizsgálatot végrehajtani, mellyel validálhatjuk a korábban kiszámított k darab töréspont helyességét. A teszt nullhipotézise, hogy az eredeti E és E_k mátrix közti eltérés az euklideszi normában nem szignifikáns. A λ paraméterrel pedig szabályozhatjuk, hogy mennyire vesszük szigorúan a büntetőtag értékét az (4) egyenletben. Amennyiben λ nagyobb egynél, szigorúan büntetjük az inflexiós pontok számának növekedését, azonban, ha kisebb, mint egy, akkor nem tartjuk olyan fontosnak a büntetőtag értékét, az OLS célfüggvény minimalizálásához képest. Az λ paraméter értékét leggyakrabban keresztvalidációval választjuk ki (Wood et al., 2015).

A változószelekció során a sajátérték- sajátvektor felbontásban megengedjük, hogy a sajátérték nulla értéket vegyen fel. Amennyiben E mátrixnak létezik olyan sajátértéke melynek értéke 0, abban az esetben a sajátértéket egy kis értékű ε taggal megnöveljük, majd így vesszük be az E mátrix sajátérték- sajátvektor felbontásába. A számítást követően a kapott E_k mátrix segítségével írjuk fel a (4) problémát, amivel (5) feladathoz jutunk. Ezzel megengedjük azt is, hogy az eredmény az azonosan nulla függvény legyen bizonyos magyarázóváltozók esetén. Ezt nevezzük konstans nulla hatásnak. Amennyiben ilyen eredményt kapunk, a magyarázóváltozó nem bír szignifikáns hatással az outputra nézve, így a modellparaméterek becslése során el is hagyhatjuk azt a modelltől (Hastie et al., 2011).

3. Adatok bemutatása

Az elemzés során felhasznált adatok két fő forrásból származnak. A keresési kifejezések megye szintű heti volumenadatait (keresési volumenindexeit) a Google Trends oldaláról, míg a megye szintű napi új koronavírus esetszámok adatait az atlatszo.hu Koronamonitorjáról töltöttem le. Az adatokat a 2020. október 1. és 2022.február 28. közti időszakra vonatkozóan vizsgáltam, így a koronavírus esetszámok heti szintű aggregálását követően 73 megfigyelésből állnak az adattáblák (Google, 2022b; Átlátszó, 2020).

Földrajzi kontrollváltozóként a Központi Statisztikai Hivatal (KSH) oldalán található 2021-es üzleti évre vonatkozó egy főre jutó régiós GDP adatokat használtam, szintén heti szintű aggregálást alkalmazva (Központi Statisztikai Hivatal, 2022a).

3.1. Google Trends adatok

A keresési volumenadatok esetén a korábban említett Li et al. (2020) tanulmányban eredményül kapott keresési kifejezéseket vettem alapul. Rövid kutatás eredményeképp láthatóvá vált, hogy a tanulmányban használt kifejezések csupán egy töredéke keltette fel a magyar internethasználók érdeklődését. Ezek közé tartoznak a koronavírus tünetek és a covid tünetek. A többi kifejezés esetében csak Budapesten keletkezett az elemzéshez szükséges adatmennyiség, így más, Magyarország teljes területén népszerű COVID-19-hez kapcsolódó kifejezések után kutattam (Google, 2022b).

A magyarországi legnépszerűbb kifejezések a következők a vizsgált időszakra vonatkozóan:

- koronavírus tünetek/ covid tünetek

- koronavírus teszt/ covid teszt
- koronavírus lappangási idő/ covid lappangási idő
- koronavírus hírek/ covid hírek
- koronavírus gyógyszer/ covid gyógyszer
- delta variáns
- poszt covid
- pcr teszt

A magyar keresések esetében a koronavírus és a covid kifejezéseket legtöbb esetben szinonimaként használják, így ezeket azonos kifejezésnek tekintettem, és átlagoltam a keresési volumenindexeiket. Az egyes keresési kifejezések volumenindexei régiókként csoportosítva az 1. számú mellékletben található (Google, 2022b).

3.2 Koronamonitor adatok

A Koronamonitor oldalán található megfigyelések esetén 2021 júliustól kezdődően a hétvégi adatok hiányoznak, azokat feltehetőleg csak hétfőnként közzétették. Így, a konzisztencia érdekében a hiányzó adatokat nullaként kezeltem. A heti szintű aggregálás miatt ez nem okoz torzítást (Átlátszó, 2020).

A Google Trends, a Koronamonitor és a KSH oldalán is megye szintű megfigyelések érhetők el. A Google Trends adatok esetében egyes kifejezések esetén léteznek olyan megyék melyek esetében vagy nem volt elegendő keresési adat ahhoz, hogy a Google Trends megjelenítse őket, vagy nagyon hiányosak voltak több időpontban is. Az előbbi esetben az adott megye keresési adatait nullának tekintettem. Ebből fakadóan arra a döntésre jutottam, hogy a megfelelő nagyságú volumenek és ezáltal stabil eredmények elérése érdekében szükséges az aggregáció. Átlagolással régiós szintre aggregáltam a keresési kifejezések volumenindexeit, ezáltal megtartva a volumenindex jellegét az idősoroknak, a koronavírus esetszámokat pedig régióként összegeztem. Az aggregálás során Budapestet külön egységként kezeltem (Átlátszó, 2020; Google, 2022b; Központi Statisztikai Hivatal, 2022a).

Az idősorok előkészítése során stacionaritás vizsgálatot végeztem, melyből kiderült, hogy összesen két differenciálásra is szükség van ahhoz, hogy minden idősor stacionárius legyen. Ebből adódóan a további elemzéseket az idősorok kétszer differenciált változatán végeztem el.

Budapest esetében az idősorok stacionaritás tesztjeinek p-értékei differenciálás előtt, első differenciálás után, illetve a második differenciálás után az alábbi 1. táblázatban láthatók:

Budapest	Első verzió		Első differenciálás után		Második differenciálás után	
változó	adf p-értéke	kpss p-értéke	adf p-értéke	kpss p-értéke	adf p-értéke	kpss p-értéke
Budapest_c	0,626	0,010	0,367	0,1	0,01	0,1
gyogy	0,117	0,100	0,010	0,1	0,01	0,1
hir	0,230	0,010	0,010	0,1	0,01	0,1
lapp	0,024	0,100	0,010	0,1	0,01	0,1
teszt	0,073	0,067	0,010	0,1	0,01	0,1
tunet	0,063	0,064	0,010	0,1	0,01	0,1
delta	0,523	0,026	0,010	0,1	0,01	0,1
pcr	0,022	0,022	0,010	0,1	0,01	0,1
poszt	0,418	0,084	0,010	0,1	0,01	0,1

1. táblázat Stacionaritás tesztek p-értékei első és második differenciálás után Budapest esetében

Forrás: saját szerkesztés R outputok alapján

Ahogy látható, a második differenciálás után minden idősor esetén az adf teszt értéke 0,01, így nullhipotézist elutasítván az idősorokat stacionáriusnak tekinthetjük. A kpss teszt esetében a nullhipotézist minden idősor esetében elfogadjuk, mely szintén azt jelenti a tesztek 2.6. fejezetbeli bemutatása alapján, hogy az idősorok kivétel nélkül stacionáriusnak tekinthetők.

A további régiókra vonatkozó összesítő táblázatok a 2. számú mellékletben találhatók.

4. Az elemzés folyamata

Az adatok előkészítése, és az alapfeltételek, mint például a stacionaritás vizsgálata után, elsőként a keresztkorrelációkat vizsgáltam az egyes keresési kifejezések volumenindexei és a koronavírus esetszámok között minden régió esetén. A megadott maximum késleltetések számaként 7 késleltetést engedélyeztem, ugyanis a később építendő VAR modell és a megfigyelések számának tekintetében maximum $73/5=14,6 \sim 14$ változója lehet a modellnek. Efelett már könnyen túlillesztetté (túl sok paraméter a megfigyelések számához képest), és így előrejelzés szempontjából használhatatlanná válna a modell. A 14 magyarázóváltozót egy VAR modellben például két magyarázóváltozó és 7 késleltetés esetén el is érjük. Az adott régió legerősebb keresztkorrelációs (abszolút értékben 0,3-nél nagyobb) értékeihez tartozó keresési kifejezések, bekerültek a régió keresési index kompozíciójába. A keresési index kompozíció és az egyes kifejezésekhez tartozó súlyok minden régió esetében a 2.5-ös pontban ismertetett képletek alapján kerültek számításra.

A kompozíció elkészítése után, következő lépésként régiók szerint készítettem modelleket: elsőként minden esetben VAR modellt építettem az információs kritériumok által javasolt késleltetésszámmal. Amennyiben az egyes VAR egyenletekben található szignifikáns késleltetett magyarázóváltozók, és a Granger-oksági teszt eredményének értelmében azok valóban befolyásolják az eredményváltozó aktuális értékét, akkor tovább vizsgáltam a hatást az impulzus-válasz függvény (IVF) segítségével. Az IVF-ben egy X idősor értékeinek egy szórásnyi sokkolásából fakadó hatását láthatjuk a választott válaszváltozóra nézve. Jelen esetben csak két idősorom van, így a modellben található szignifikáns késleltetett magyarázóváltozóktól függ, melyik idősor változóit sokkolom. Amennyiben a VAR modell mindkét egyenletében található szignifikáns magyarázóváltozó, két IVF készül. Az IVF függvényben azt ortho paraméter értékét minden esetben igazra állítottam, ugyanis így a konfidencia intervallumok stabilabban becsülhetők, mint a sorrendfüggetlen hamis paraméter használatakor kapott esetben. Ennek következményeképp azonban a kezdeti t időpontbeli sokkolási sorrendet is figyelembe kell venni az eredmények szempontjából.

A VAR modell összefoglalójának ismerete szükséges a GAM modellek számának megállapításához is. Abban az esetben, ha a VAR modell egyenletrendszerében voltak szignifikáns változók, ezek esetleges további nem lineáris tulajdonságainak vizsgálatára GAM modelleket készítettem. Egyéb esetben csupán a keresztkorrelációk vizsgálata során kigyűjtött fogalmak, keresőkifejezések azon késleltetettje került a GAM modellbe, melyhez a legnagyobb keresztkorrelációs együttható tartozott. A következőkben bemutatom a fent említett általános elemzési folyamat régióspecifikus elemeit, földrajzi területegységek szerint csoportosítva.

5. A módszertan alkalmazása, eredmények ismertetése

Az eredmények értelmezése során erősen szignifikánsnak tartottam azokat az eredményeket, melyek 1%-os szignifikancia-szint mellett is szignifikánsnak tekinthetők. Közepes hatásnak minősítem az 1-5% közötti szignifikancia-szinteken szignifikánsnak bizonyuló eredményeket. Enyhe mértékűnek tekintem azokat a hatásokat, amelyek 5-10% közötti szignifikancia-szinteken szignifikánsak. Amennyiben egy hatás a szokásos szignifikancia-szinteken (1- 10%) nem bizonyul szignifikánsnak, úgy tekintem, hogy a hatás nem létezik.

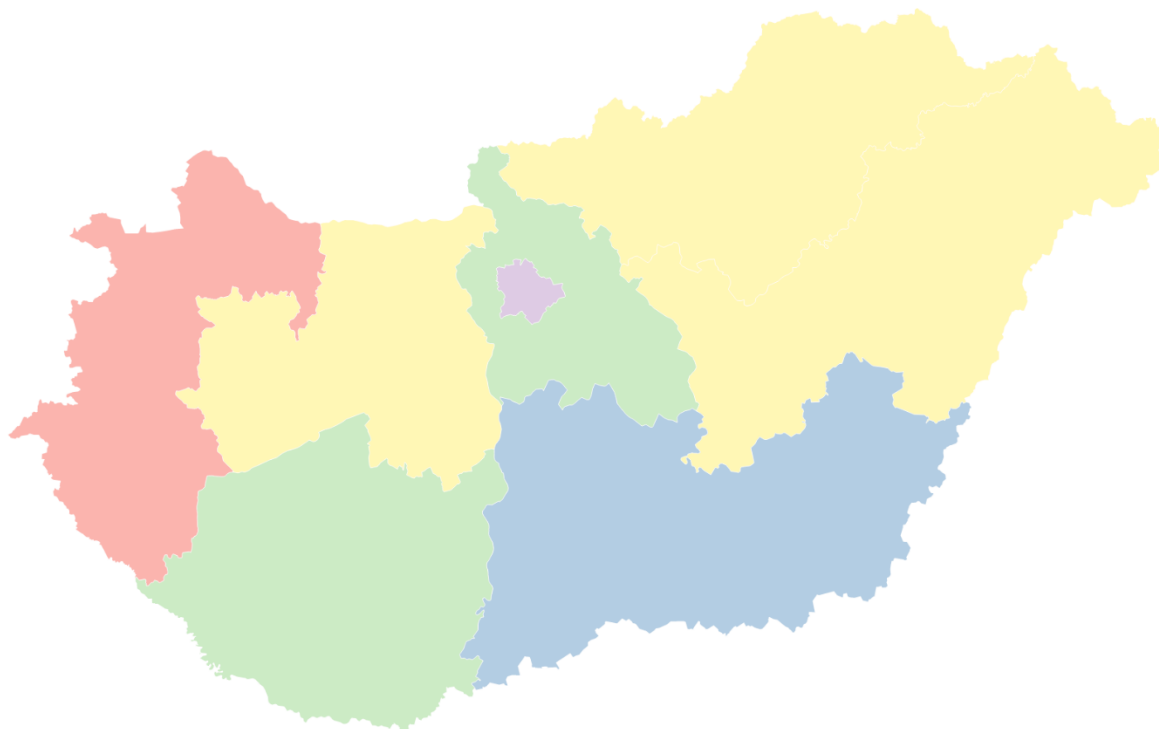
A hatások ábrázolása régiókként a VAR és GAM modellek eredményei alapján az 1. ábrán látható.

Eredmények

VAR és GAM modellek

HATÁSOK

Fordított irány* + GAM GAM Keresési index kompozíció -> esetszám Mindkét irány + GAM Nincs hatás



* esetszámok keresési index kompozícióra gyakorolt hatása

A készítéshez használt program: Datawrapper

1. ábra Régiókénti hatások a VAR és GAM modellek eredményei alapján

Forrás: Saját szerkesztés R outputok alapján Datawrapperrel

Ahogy az 1. ábrán látható, Dél-Dunántúl régió és Pest vármegye esetében fedezhető fel az előzetesen várt hatás, azaz az adott régió keresési index kompozíció koronavírus esetszámokra gyakorolt késleltetett egyirányú hatása. Tehát amennyiben nő a keresési index kompozícióba tartozó keresési kifejezések iránti érdeklődés, várhatóan növekedni fog néhány héttel később a koronavírus fertőzöttek száma. Fordított hatásnak tekinthető a régiós COVID-19 esetszámok késleltetett hatása a keresési index kompozícióba tartozó szavak keresési intenzitására. Többek között ilyen hatás látható Nyugat-Dunántúl régió esetében. Az ilyenfajta hatások általában a társadalom által érzékelt csúcspontot jelezhetik, mivel legtöbbször az esetszámok növekedésének hatására a hetekkel későbbi keresési index kompozíció népszerűsége csökken (negatív előjelű, fordított a hatás), tehát a társadalom által érzékelt járvány csillapodik.

GAM hatásról akkor beszélhetünk, ha az előbbi hatások valamelyike nem lineáris tulajdonságokkal rendelkezik, vagy a régiós keresési indexbe tartozó kifejezések közül egy vagy több önállóan nemlineáris hatással bír az esetszámokra. GAM hatások megfigyelhetők Nyugat-Dunántúlon, Dél-Alföldön és Budapesten is. Budapest esete különleges ugyanis itt nem csak a várt egyirányú hatás, de fordított és nemlineáris hatás is megfigyelhető. A nincs hatás jelző azt jelenti, hogy a fent felsoroltak közül egyik sem jellemző az adott régióra, másnéven egyik adott régióra készített modell egyenleteiben sem található a szokásos szignifikancia-szinteken szignifikánsnak tekinthető magyarázóváltozó. Az egyes régiók modelljeinek átfogó eredményei a 2. táblázatban találhatók.

Régió	Wald teszt p- értéke	Portmanteau többváltozós fehérzaj teszt p-értéke	Karakterisztikus polinom gyökeinek tartománya	5%-on szignifikáns GAM splineok száma	Breusch- Godfrey teszt p- értéke (GAM)
Budapest	0,035	0,278	0,6899558-0,7793697	1	0,180
Pest	0,015	0,355	0,5662304-0,8278476	0	0,038
Nyugat-Dunántúl	0,072	0,307	0,5847606-0,9174678	1	0,114
Dél-Dunántúl	0,044	0,700	0,6122421-0,8092364	0	0,162
Közép-Dunántúl	0,385	0,808	0,5797103-0,8352681	0	0,323
Észak- Magyarország	0,165	0,887	0,4586053-0,7566639	0	0,010
Észak-Alföld	0,222	0,981	0,6382955-0,8349954	0	0,196
Dél-Alföld	0,115	0,975	0,6043404-0,6882772	2	0,059

2. táblázat VAR és GAM modellekben elvégzett tesztek eredményei

Forrás: saját szerkesztés R outputok alapján

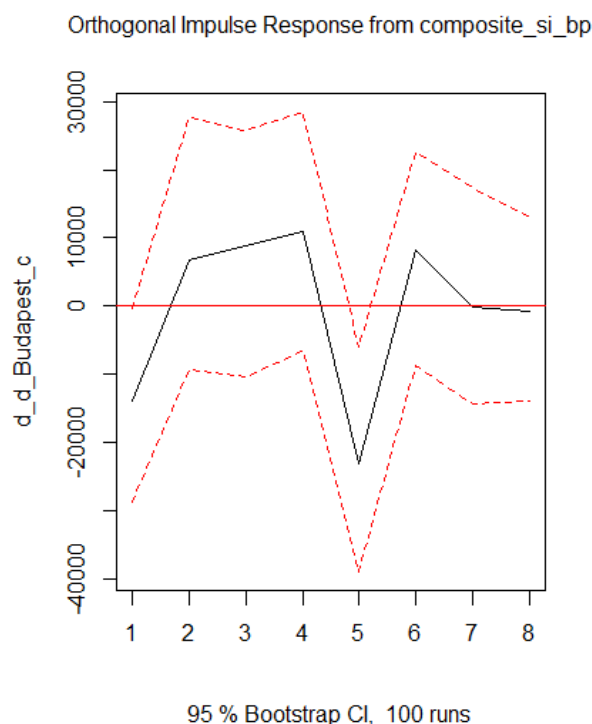
5.1. Közép-Magyarország: Budapest

Mivel Budapest rendelkezik a régiók közül a legtöbb különféle hatással (egyirányú, fordított, GAM), ezért szeretném az eredmények részletes szemléltetését a fővárossal kezdeni. Budapest esetében a keresztkorrelációk vizsgálata során a koronavírus/covid lappangási idő, a koronavírus teszt, a delta variáns, a PCR teszt és a poszt covid keresési kifejezések korreláltak legjobban a fertőzöttek számával. A város esetében a keresztkorrelációs együttthatók értéke rendre -0,331, -0,315, 0,301, -0,347 és -0,465. A hozzájuk tartozó késleltetések száma rendre 5, 4, 3, 2 és 4.

A felsorolt keresési kifejezésekből a 2.5. pontban ismertetett képletek alapján elkészítettem Budapest keresési index kompozícióját. Ezután megépítettem a VAR modellt, melyben a késleltetések maximális száma négy az információs kritériumok alapján. A modell stabilnak

mondható, mivel a karakterisztikus polinomjának gyökei az egységkörön belül találhatók. A Portmanteau-próba esetében a nullhipotézist el tudjuk fogadni, tehát a modell hibatagjai együttesen fehérzajnak tekinthetők. A pontos értékek a fejezet elején közölt 2. táblázatban találhatók.

A Granger-okság vizsgálat során, mivel a VAR modell mindkét egyenletében volt szignifikáns magyarázóváltozó, ezért kétféle oksági kapcsolatot teszteltem. Elsőként a keresési index kompozíció esetszámokra gyakorolt hatását vizsgáltam. Az eredmények alapján 5%-os szignifikancia-szint mellett a Wald-teszt nullhipotézise elvethető, tehát a budapesti keresési index kompozíció Granger-okozza a budapesti koronavírus esetszámokat. Mivel 5%-os szignifikancia-szinten fennáll a Granger-okság ezért az IVF hatásokat is megvizsgáltam. A 2. ábrán látható esetben az IVF kezdeti t időpontjában elsőként a keresési index kompozíciót sokkolom majd ezt követően a koronavírus esetszámokat. Azonos idejű együtt-mozgások vizsgálatakor különösen fontos a jó sokkolási sorrend megválasztása. A jelenlegi sokkolási sorrend eredményeképp koronavírus esetszám már a nulladik időpontban reagál a keresési index kompozíció sokkjára, így, a koronavírus esetszámok IVF-je a keresési index kompozíció sokkjainak hatására nem a nullából fog indulni, tehát a keresési index kompozíció hatása már a nulladik időpontban is látható koronavírus esetszámokon. Ezt a sorrendet azért tartom logikusnak, mivel a keresési index kompozíció szignifikáns magyarázóváltozónak bizonyult a koronavírus fertőzöttek számának tekintetében, és logikus feltételezés, hogy ez a hatás azonos időben is jelentkezhessen. Tehát, a hatás pontos kimutatása érdekében elsőként feltehetőleg az esetszámokra ható keresési index kompozíciót kell sokkolni. Az IVF-ek esetén a konfidencia-intervallumot 95%-osra állítottam, hogy összhangban legyen a Wald-teszt p -értékével, amely az 1- és 5% közti tartományba esik (0,035).



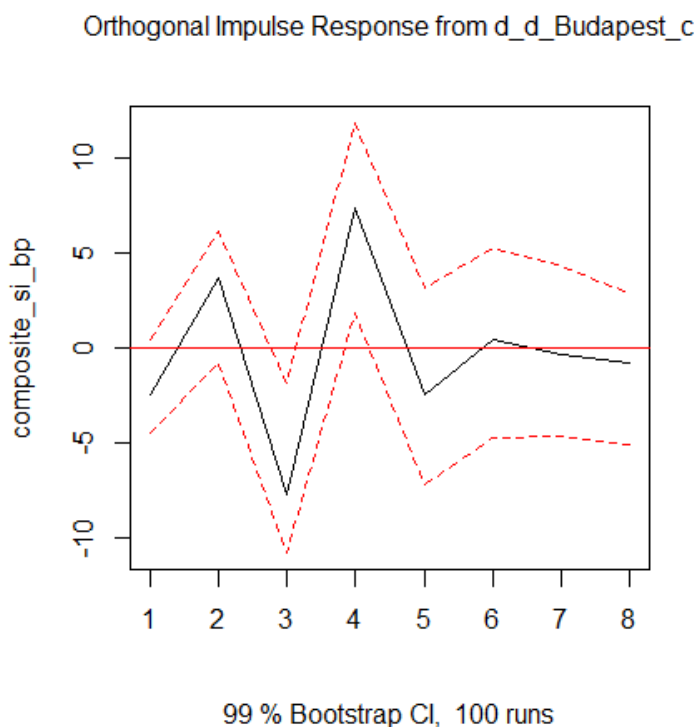
2. ábra Budapesti koronavírus esetszárok IVF ábrája a keresési index kompozícióból érkező sokk hatására (95%-os konfidencia-intervallum)

Forrás: R output saját szkript alapján

A 2. ábrán látható, hogy a keresési index kompozícióban bekövetkező egy szórársnyi sokk átlagosan öt héttel később a koronavírus fertőzöttek számában is szignifikáns csökkenést okoz. Tehát elmondható, hogy a kompozícióba tartozó kifejezések volumenindexeinek növekedése átlagosan a járvány lecsengését jelzi előre. Az, hogy a keresési index kompozíció ellentétesen (tehát negatívan) hat az esetszárokra, azt jelentheti, hogy a koronavírus kifejezések keresési gyakoriságai nagy valószínűséggel a járvány tetőpontja környékén kezdtek növekedni. Ez történhetett a média által közvetített hírek, vagy az internetezők környezetében kiugróan magas fertőzöttek számának hatására is. Az is előfordulhat, hogy az adott személy észlel tüneteket magán és szeretne tesztel megbizonyosodni az egészségügyi állapotáról. Ezután várhatóan öt héttel következik a lecsengés, tehát a társadalom több, mint egy hónappal az esetszárok mérséklődése előtt érdeklődik kifejezetten a koronavírus és az azzal kapcsolatos kifejezések iránt.

Ezután megvizsgáltam, hogy a koronavírus esetszárok Granger-okozzák-e a keresési index kompozíció keresési intenzitását. A Wald-teszt 0,00021-es nagyságú p-értéke alapján minden szokásos szignifikanciaszinten (1-10%) elfogadható az az állítás, hogy a budapesti koronavírus

esetszámok Granger-okozzák a budapesti keresési index kompozíciót. Így, megvizsgáltam az IVF hatásokat:



3. ábra Budapesti keresési index kompozíció IVF ábrája a koronavírus esetszámokból érkező sokk hatására (99%-os konfidencia-intervallum)

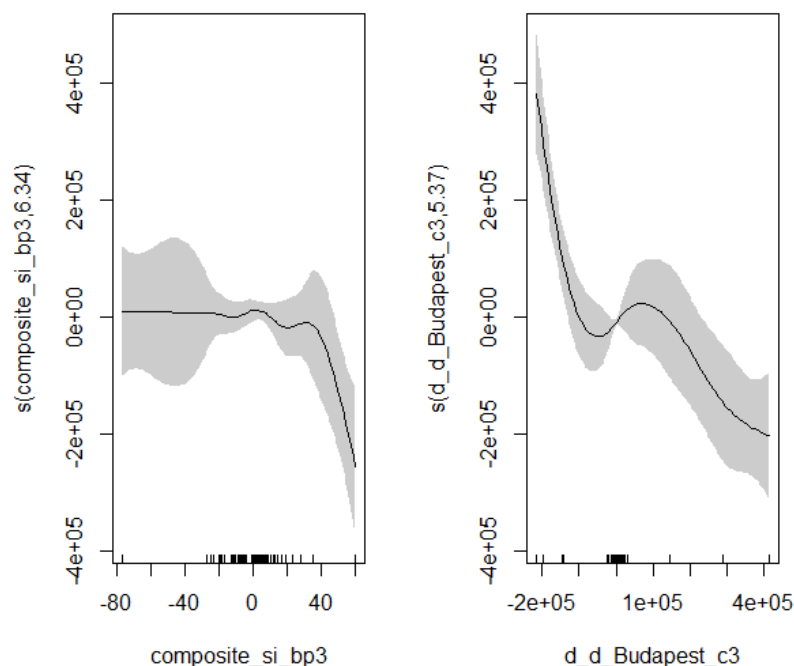
Forrás: R output saját szkript alapján

Ebben az esetben először az esetszámot sokkoljuk t időpontban, majd a keresési index kompozíciót, így már a keresési index kompozíció IVF-jének nulladik időpontjában látható az esetszámokból érkező sokk hatása a keresési index kompozícióra. Ez a sorrend azért logikus, mivel a keresési index kompozíció modellegyenletében az esetszámok szignifikáns magyarázóváltozóként vannak jelen, így előbb az esetszámoknak kell változni, tehát ezt a változót kell elsőként sokkhatás érje ahhoz, hogy a keresési index kompozícióra gyakorolt hatását teljes mértékben megfigyelhessük.

A 3. ábrán látható a koronavírus esetszámokat majd a keresési index kompozíciót t időpontban érő egységnyi sokk hatása. A sokkolás időpontjához képest átlagosan három héttel később csökkenés, egy hónappal később pedig növekedés figyelhető meg a keresések számában. Itt valószínűleg a járvány felfutó szakaszában vagyunk, ahol a fertőzöttek számának növekedése még nem okoz a keresési kifejezések esetében növekedést, sőt a lappangó időszakban

kifejezetten kevésbé érdeklődnek az internetezők a vírus iránt. Azonban az esetszámok növekedése után egy hónap elteltével már akkora mértékű a vírus terjedése, hogy a társadalom jelentős része észleli. Elképzelhető, hogy az egyének családjában vagy környezetében megfertőződött valaki, vagy a hírekben értesülnek az új fertőzöttek növekvő számáról, ezért érdeklődni kezdenek a vírus különböző aspektusai iránt. Ez azt jelentheti, hogy Budapest régió esetében a fertőzöttek számának növekedése átlagosan a járvány társadalom által észlelt hullámának felfutó szakaszát jelzi előre.

Mivel a VAR modell mindkét egyenletében található szignifikáns magyarázóváltozó, ezért két GAM modellt is építettem ezen változók esetleges nem lineáris tulajdonságainak vizsgálata szempontjából. Annak elkerülése érdekében, hogy az esetleges autoregresszív hatások a külső magyarázóváltozó eredményváltozóra gyakorolt hatásában közvetett hatásként megjelenjenek, a fertőzöttek számának modellje esetében az eredményváltozó harmadik, illetve keresési index kompozícióra épített GAM esetében az eredményváltozó második késleltetettjét is beépítettem a modellbe. Az eredményváltozónak azért ezen késleltetettjei kerültek az elsőként említett GAM modellbe, mivel a budapesti VAR modell fertőzöttek számára épített egyenletében a keresési index kompozíció harmadik késleltetettje bizonyult szignifikánsnak, így a keresési index kompozíció harmadik késleltetettje mellett az eredményváltozó ezen késleltetését kell beépíteni a GAM modellbe, így ezen magyarázóváltozó hatásai nem kerülnek át közvetetten a keresési index kompozíció hatásába. A keresési index kompozícióra épített GAM modell egyenletébe hasonló okból került bele magyarázóváltozóként a kompozíció második késleltetettje is az esetszámok második késleltetése mellett. Itt a kompozíció ezen késleltetésének közvetett hatását szerettük volna kiküszöbölni, így az nem kerül bele más magyarázóváltozók tapasztalt hatásába, jelen esetben az esetszámok második késleltetettjének keresési index kompozícióra gyakorolt hatásába. A keresési index kompozíciót eredményváltozóként tartalmazó modell esetében nem tapasztalható nemlineáris tulajdonság. A budapesti fertőzöttek számára épített GAM modell esetében 3%-os szignifikanciaszinten szignifikánsnak tekinthető a budapesti keresési index kompozíció harmadik késleltetése. A GAM hibatag viselkedése nem sérti meg a modellfeltételeket, mivel a Breusch–Godfrey teszt eredményeképp fehérzajnak tekinthetők minden szokásos szignifikancia-szinten. A nemlineáris hatások az 4. ábrán kerültek ábrázolásra.



4. ábra nemlineáris (GAM) hatások Budapesten

Forrás: R output saját szkript alapján

Ahogy a 4. ábrán látható a keresési index kompozíció harmadik késleltetése nem lineáris hatást gyakorol a budapesti esetszámokra. Különösen negatív a hatása a három héttel későbbi esetszámokra abban az esetben, ha a keresési index kompozícióba tartozó kifejezések keresési intenzitása extrém mértékben növekszik. A fenti IVF hatásoknál említett ellentétes hatást egészíti ki ez a nem lineáris hatás. A keresési index kompozíció növekedése után öt héttel bekövetkező esetszámokban megfigyelhető lineáris tendenciájú csökkenést kiegészíti a hirtelen extrém mértékű keresési intenzitás növekedést három héttel követő nagy mértékű (nemlineáris) csökkenés az esetszámokban. Tehát, a keresések kiugróan magas növekedése a korábbiakat megerősítve feltehetőleg a hullám csúcspontját jelzi, majd három héttel később hirtelen csökkennek az esetszámok, amely a lecsengés kezdetét jelzi. A lineáris változat valószínűleg azért is jelzi későbbre a hullám lecsengését, mivel ezt a nem-lineáris hatást lineárisaként próbálja megjeleníteni. Az VAR és a GAM hatások alapján tehát összességében elmondható, hogy a keresési index kompozíció normál és extrém mértékű növekedése is az esetszámok várható csökkenését, tehát a hullám lecsengését jelzik előre Budapest esetében.

5.2. Közép-Magyarország: Pest vármegye

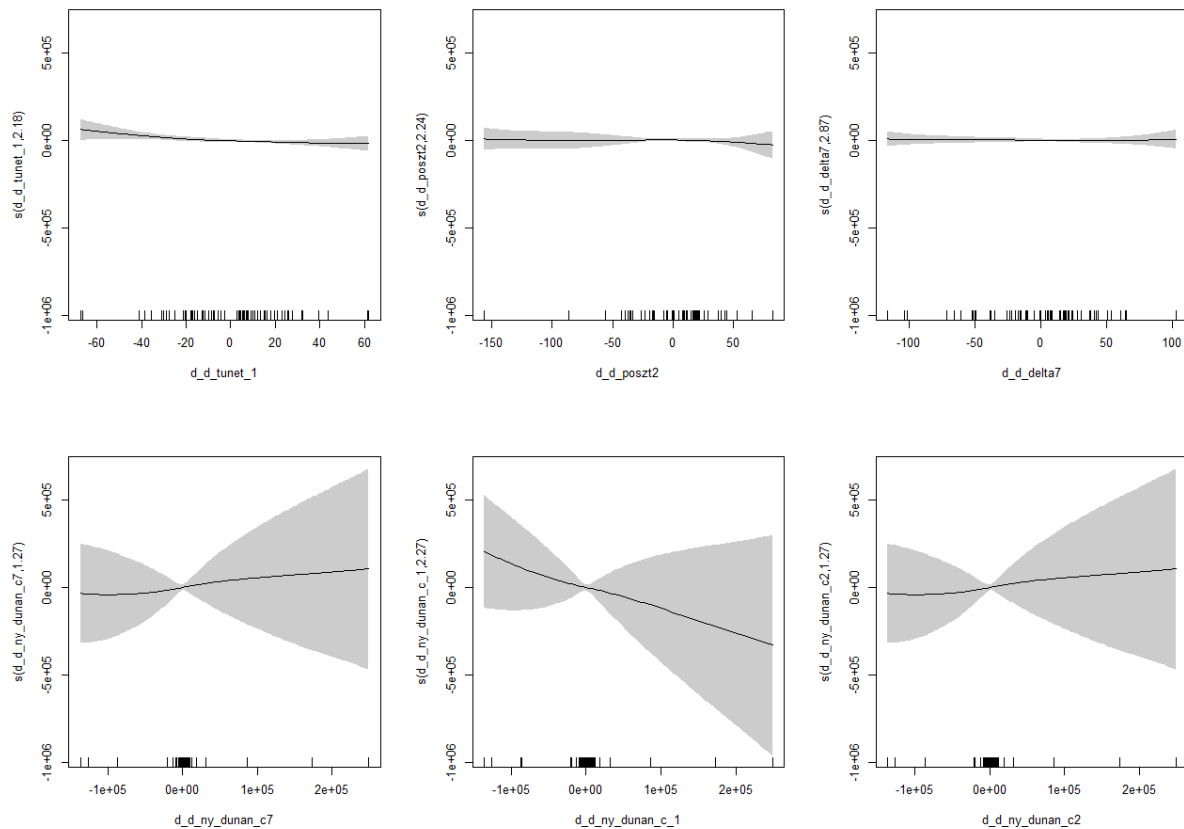
Közép-Magyarország régió Budapesttől különválasztott változata esetében, a régió keresési index kompozíciója Granger-okozza az esetszámokat, tehát valós egyirányú késleltetett hatást gyakorol a keresési index kompozíció az esetszámokra. Az IVF esetén a sokkolási sorrendben elsőként a keresési index kompozíciót sokkoltam, majd az esetszámokat t időpontban, így az esetszámok IVF-jén már látható a nulladik időpontbeli index kompozíció sokkjának hatása is. A gondolatmenet tehát egyezik a Budapest esetén alkalmazottal. A 3. számú mellékletben látható Közép-Magyarország régióra vonatkozó, IVF ábra alapján, a keresési index kompozíció feltehetőleg a járvány csúcspontját jelzi előre négy hetes késleltetéssel, ugyanis a sokkolástól számított negyedik héten az esetszámok átlagosan megnövekednek, azonban az ötödik heti késleltetés esetében az esetszámok már csökkenő tendenciát mutatnak.

5.3. Nyugat-Dunántúl

Nyugat-Dunántúl régió esetében a Granger-okság vizsgálat eredményeként kapott 7,2%-os p -érték azt jelzi, hogy 10%-os szignifikanciaszint mellett a fertőzöttek száma Granger-okozza a keresési index kompozícióba tartozó kifejezések keresési intenzitásának növekedését. Így elmondható, hogy Nyugat-Dunántúl régió esetében valóban fordított hatás figyelhető meg. Az impulzus-válasz függvény esetében elsőként az esetszámot majd a keresési index kompozíciót érte egységnyi sokk t időpontban, így a keresési index kompozíció IVF-jében már látható a kezdeti időpontban is az esetszám sokkjának keresési index kompozícióra gyakorolt hatása. Az esetszámok növekedését követően átlagosan hat hetes késleltetéssel a keresési index kompozíció keresési intenzitása nő, így feltehetőleg az esetszámok növekedése Nyugat-Dunántúl régióban másfél hónap késéssel érezhető a társadalom számára. A régióhoz tartozó IVF ábra a 3. számú mellékletben tekinthető meg.

Azonban nem csak lineáris hatások fedezhetők fel Nyugat-Dunántúl régióban. A kifejezéseket egyesével magyarázóváltozóként tartalmazó, esetszámokra épített GAM modellbe a magyarázóváltozók maximális korlátja miatt a legkisebb keresztkorrelációs együtthatóval rendelkező kifejezés nem került be. A végső modellben a koronavírus/covid tünetek kifejezés első késleltetése bizonyult szignifikáns magyarázóváltozónak 5%-os szignifikancia-szint mellett. A modell hibatagjai fehérzajnak tekinthetők, így a modellfeltételek nem sérülnek. Ahogy az 5. ábrán látható, hogy van egy alacsony határérték a koronavírus tünetek keresésének

számában, amelyet átlépve a keresésekhez képest egy hét csúszással a koronavírus esetszám a szokásosnál némileg jobban csökken.



5. ábra Nemlineáris (GAM) hatások Nyugat-Dunántúl régióban

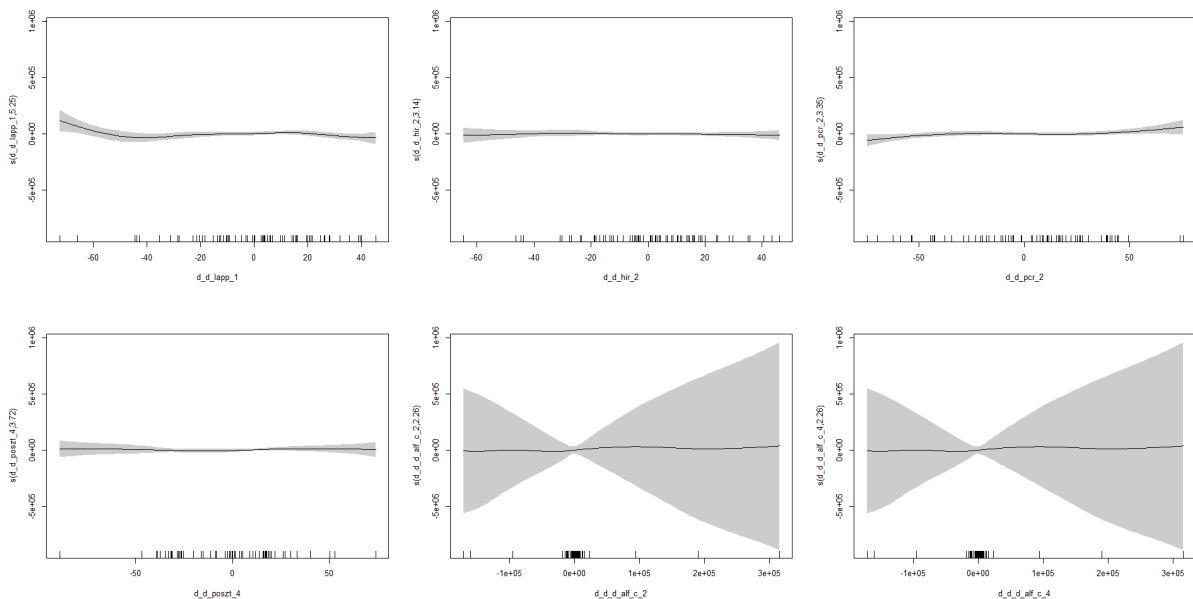
Forrás: R output saját kód alapján

5.4. Dél-Dunántúl

Dél-Dunántúl régió VAR modelljében 5%-os szignifikancia-szint mellett Granger-okság áll fenn, tehát a Dél-Dunántúl-i keresési index kompozíció és az esetszámok között felfedezhető egyirányú késleltetett ok-okozati kapcsolat. A VAR modellben található sok szignifikáns magyarázóváltozó ellenére az IVF esetében kizárólag a sokkolást követő második héten látható növekedés az esetszámokban. Tehát, a keresési index kompozíció egységnyi növekedése esetén a fertőzöttek számának emelkedése várható, kéthetes késleltetéssel, így a kompozíció az esetszámok növekedésének előfutárjaként tekinthető a Dél-Dunántúlon.

5.5. Dél-Alföld

Dél-Alföld esete egészen különleges: kizárólag enyhe nem lineáris hatások figyelhetők meg ezen régió esetében. A régió individuális keresési kifejezéseinek volumenindexeinek felhasználásával, a koronavírus esetszámot eredményváltozóként tartalmazó GAM modellt készítettem. A magyarázóváltozók maximális számának megkötése miatt a kifejezések közül kettő (az esetszámokkal abszolút értékben leggyengébb keresztkorrelációs együtthatóval rendelkező kettő) nem került be a modellbe. A végső modellben két keresési kifejezés is szignifikáns magyarázóváltozónak bizonyult 5%-os szignifikancia-szinten. A kifejezések közül szignifikánsnak minősül *covid/koronavírus lappangási idő* első késleltetettje, és a *PCR teszt* második késleltetettje. A Breusch-Godfrey teszt 5,9%-os p-értéke ebben az esetben arra utal, hogy az eredményeket kellő óvatossággal kell kezelni, nem minden szokásos szignifikancia-szinten (1-10%) fogadható el az állítás, miszerint a hibatagok fehérzajnak tekinthetők. 5%-os és ennél alacsonyabb szignifikancia-szinteken azonban még elfogadható a fenti állítást.



6. ábra Nemlineáris (GAM) hatások Dél-Alföld régióban

Forrás: R output saját szkript alapján

Ahogy az ábrán látható, a koronavírus/covid lappangási idő keresési kifejezés volumenindexének alacsony értékről történő extrém növekedése átlagosan negatív hatást fejt ki az esetszámokra egy hetes késleltetéssel. A PCR teszt kifejezés esetében a volumenindex viszonylag magas értékről extrém magas értékre növekedése átlagosan a fertőzöttek számának növekedéséhez vezet két héttel a kifejezés keresési intenzitásának növekedését követően.

5.6. Közép-Dunántúl, Észak-Magyarország, Észak-Alföld:

Közép-Dunántúl régió esetében nem található szignifikáns magyarázóváltozó egyik modellben sem, így a keresési index kompozíció és az esetszámok között nem fedezhető fel kapcsolat.

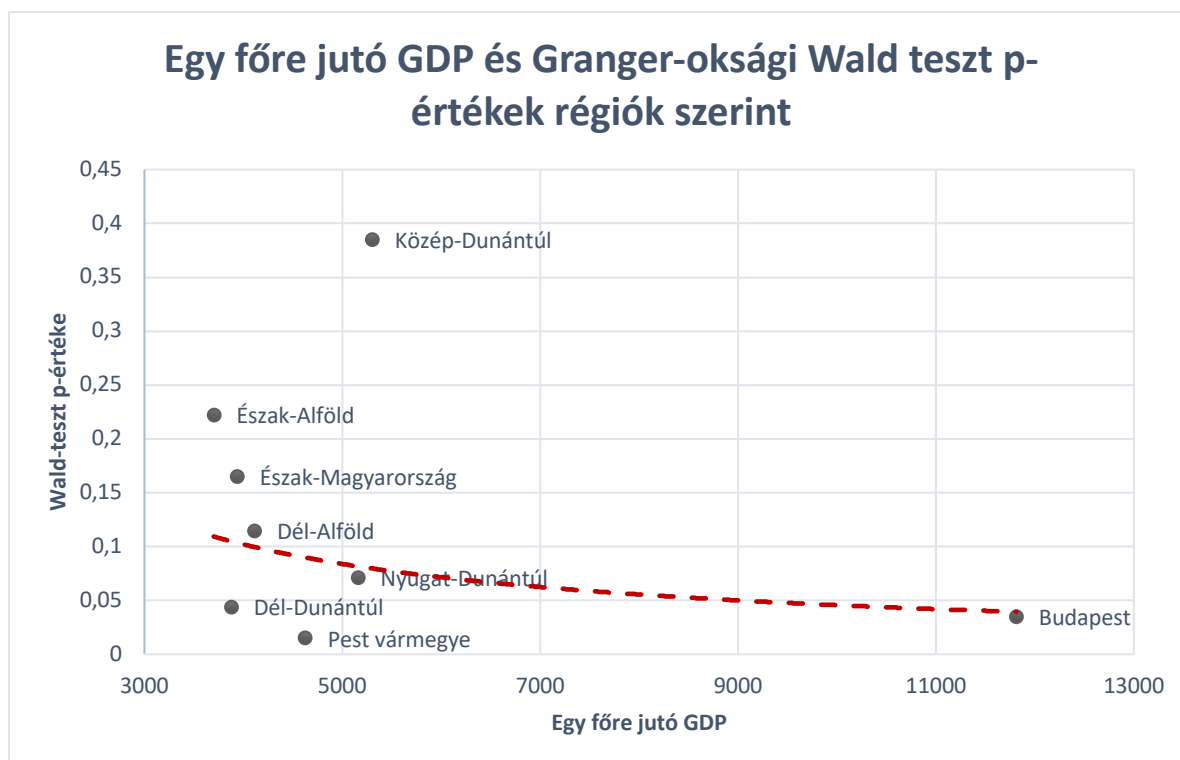
Habár Észak-Magyarország régióban a VAR modell esetszám egyenletében voltak szignifikáns magyarázóváltozók (keresési index kompozíció második és harmadik késleltetettje), Granger-okság nem figyelhető meg a keresési index kompozíció és a fertőzöttek száma között. Így, a két változó között Észak-Magyarországon sem fedezhető fel ok-okozati kapcsolat.

Észak-Alföld VAR modelljének esetében felfedezhető szignifikáns magyarázóváltozóként a keresési index kompozíció első késleltetése, azonban Granger-okság a korábbi két régióhoz hasonlóan nem áll fenn. A keresési index kompozíció és a fertőzöttek száma tehát itt sem áll ok-okozati kapcsolatban.

Az egyes keresési kifejezések esetszámokra gyakorolt nemlineáris hatásainak vizsgálata szintén nem vezetett eredményre ezen régiók esetében, így elmondható, hogy az egyes régiók keresési index kompozíciójába került keresési kifejezések külön-külön sem hatnak a fent említett régiók esetszámaira.

6. Egy főre jutó GDP és a Granger-okság kapcsolata

A korábban említett földrajzi kontrollváltozó (egy főre jutó GDP) és VAR modellek esetében vizsgált Granger-okságok közti kapcsolatot vizsgálom régiókként ebben a fejezetben. A kapcsolatvizsgálatot első körben egy pontdiagram segítségével valósítom meg, amelyen a Granger-oksági Wald-teszt p-értékeit vizsgálom egy főre jutó GDP függvényében régiókként. Emellett a két változó közti korrelációs kapcsolatot is vizsgálom.



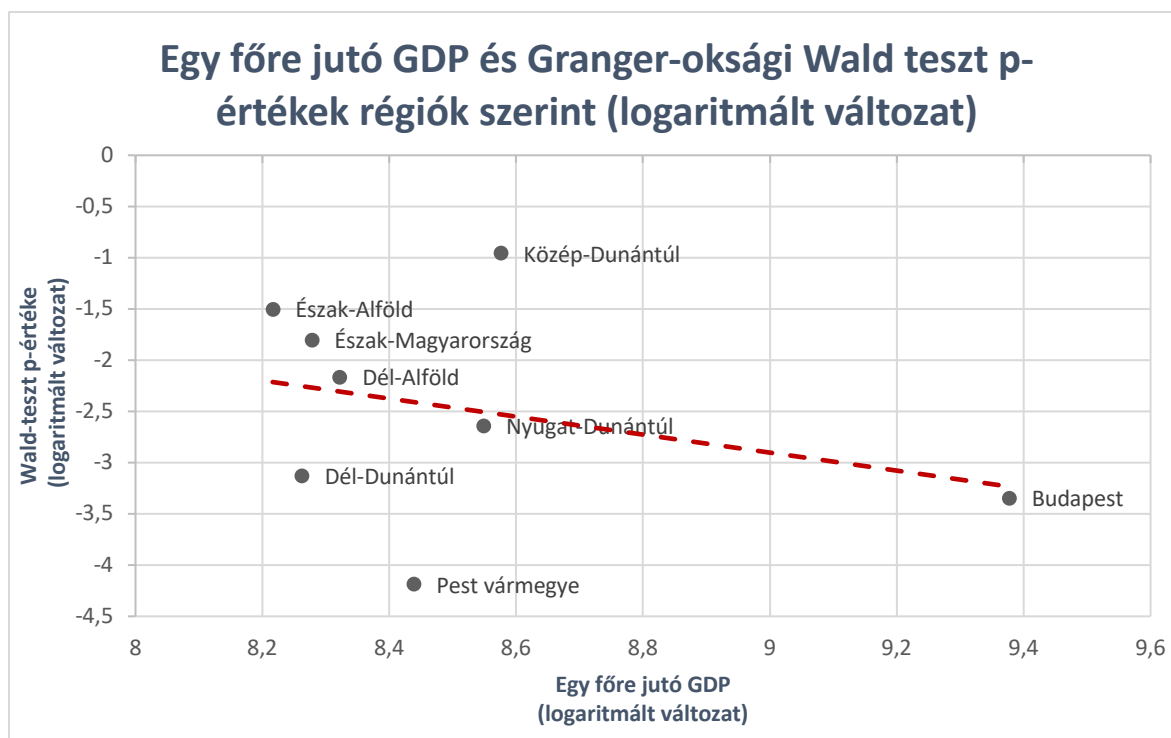
7. ábra Egy főre jutó GDP és Granger-oksági Wald-teszt p-értékek régiók szerint

Forrás: saját szerkesztés R outputok alapján

A 7. ábráról leolvasható, hogy az egy főre jutó GDP növekedésével egy-egy kivételtől eltekintve csökken a Wald-teszt p-értékének nagysága. Tehát Közép-Dunántúl régió kivételével általánosan elmondható a régiókról, hogy minél nagyobb az egy főre jutó GDP az adott régióban, annál inkább szignifikáns a régió modelljében megfigyelhető Granger-oksági kapcsolat. Ez alól kivétel Közép-Dunántúl, hiszen magas GDP-vel rendelkezik a többi régióhoz képest, ugyanakkor a Wald-teszt p-értéke nagyon magas, tehát nem fedezhető fel Granger-okság ebben a régióban.

A kiugró értékeket egy logaritmussal lehet kezelni mindkét változó esetében. Így egy log-log elven illesztett regressziós függvényt kapunk, ebből következik, illetve az ábrán is látható, hogy az egy főre jutó GDP és a Wald-teszt p-értéke közötti kapcsolat közel sem lineáris, a log-log kapcsolatból kiindulva sokkal inkább hatványkitevős trendet vesz fel.

Nézzük meg mindkét változónk logaritmusa esetén a kapcsolatot:



8. ábra Egy főre jutó GDP és Granger-oksági Wald teszt p-értékek régiók szerint (logaritmált változat)

Forrás: saját szerkesztés R outputok alapján

Ahogy a 8. ábrán látható a két változó között valóban hatványkitevős kapcsolat áll fenn, hiszen a logaritmált változók ábrázolása esetén a trend kisimul, sokkal inkább közelít a lineárishoz. Tehát a gazdasági fejlettség növekedésével hatványosan csökken a Wald-teszt p-értéke.

A korrelációk vizsgálata során az eredeti adatsorok közti korrelációs együttható értéke -0.2500976. Ez a negatív együttható is alátámasztja a korábbi megállapítást, miszerint az egy főre jutó GDP növekedésével átlagosan csökken a Wald-teszt p-értéke. A két változó közti kapcsolat enyhének minősül, azonban már közelíti a 0,3-as határt, amelytől közepesen szoros kapcsolatnak számít két változó kapcsolata. A logaritmált változók esetében a korrelációs együttható értéke -0.3116609, amely már közepesen szoros kapcsolatnak számít. Ezzel is alátámaszthatjuk, hogy az adatsorok logaritmálása szükséges volt.

A fent megfigyelt tendenciák feltehetőleg azért alakulnak így, mivel azok a régiók, amelyekben magas az egy főre jutó GDP jellemzően fejlettebbek is mint azok a régiók, ahol az egy főre jutó GDP alacsonyabb. A fejlettebb régiókban jobbak az ott élő emberek életkörülményei, amelyhez a mai társadalmi helyzetben hozzátartozik az internetelérés (Minges, 2016). Az általam vizsgált korábbi tanulmányok esetében ilyen jelenség feltehetőleg nem volt tapasztalható, hiszen minden kutatás rendkívül fejlett régiókra korlátozódott.

Mivel a fejlettebb régiókban több ember rendelkezik internethozzáféréssel, így a keresések gyakorisága is jellemzően nagyobb. Több megfigyelés esetén pontosabb elemzéseket lehet készíteni, így könnyebben megállapíthatunk erősebb ugyanakkor stabil összefüggéseket, mint a kisebb mintával rendelkező régiók esetében. Így megállapítható, hogy az előrejelző modellek, feltehetőleg azokban a régiókban tudnak hatékonyan működni, ahol magas az internetpenetráció. Azokban a régiókban, ahol kisebb az internetpenetráció mértéke, érdemesebb inkább az egyváltozós időszorosmodelleket (például Autoregresszív Integrált Mozgóátlag (ARIMA) modelleket) használni a járványkezelésre.

7. Összegzés

Dolgozatom célja a koronavírus esetszámok előrejelzése internetes keresési volumenadatok alapján magyarországi régiók szerint. Az előrejelzés során a Google Trends keresési volumenindex adatokat, illetve a Koronamonitor koronavírus esetszámait felhasználva vizsgáltam a kérdést, miszerint van-e szignifikáns hatása a keresési adatoknak a koronavírus fertőzöttek számára az egyes régiókban.

Az elemzés során egyes régiók adatai alapján épített VAR és GAM modellek eredményei alapján Budapesten érzékelhető a legtöbb fajta, és egyben legstabilabb hatás az összes régió közül. Budapest esetében a keresési index kompozíció lineáris hatások révén feltehetőleg a járványhullám lecsengését várhatóan öt hetes késleltetéssel jelzi előre. Ezt erősíti meg a GAM modell esetében látott nemlineáris hatás, melynek értelmében a keresési intenzitás extrém mértékű növekedését követően várhatóan három hetes késleltetéssel kezd hirtelen lecsengeni a járvány. Ezen felül fordított hatás is megfigyelhető, ahol az esetszámok növekedésének hatására három héttel később csökkennek majd négy héttel később emelkednek a keresési index kompozíció volumenadatai, így feltehetőleg jelezve a járvány társadalom által észlelt felfutó szakaszát. Közép-Magyarország Budapest nélkül vett területén a keresési index kompozíció feltehetőleg a járvány tetőpontját jelzi előre, ugyanis négy héttel a keresések intenzitásának növekedését követően az esetszámok növekednek, öt héttel utána pedig csökkenni kezdenek. Nyugat-Dunántúl esetében feltehetőleg az esetszámok növekedését követő hatodik héten válik érzékelhetővé a társadalom számára a járvány felfutása. Ugyanakkor nemlineáris hatás is felfedezhető, a koronavírus tünetek kifejezés keresési intenzitásának extrém növekedését követő héten már észrevehető a szokásosnál nagyobb visszaesés az esetszámokban. Dél-Dunántúl esetében a keresési index kompozícióba tartozó kifejezések népszerűségének növekedése várhatóan kéthetes késleltetéssel jelzi előre a régió koronavírus esetszámainak

növekedését, ezzel feltehetően a járvány kezdeti, felfutó szakaszát jelzi előre. Dél-Alföld régióban kizárólag nemlineáris hatások figyelhetők meg. A *koronavírus/covid lappangási idő* keresési kifejezés gyakoriságának extrém mértékű növekedése várhatóan egy hetes késleltetéssel negatív hatást gyakorol az esetszámokra. A *PCR teszt* kifejezés esetében ellentétes a hatás, ugyanis itt a volumenindex magas értékről hirtelen még magasabb értékre ugrása átlagosan a fertőzöttek számának növekedését jelzi előre kéthetes késleltetéssel. Észak-Alföld, Észak-Magyarország és Közép-Dunántúl régiók esetében nem figyelhető meg semmilyen szignifikáns hatás.

A keresési index kompozíciók, illetve egyes önálló keresési kifejezések és az esetszámok egymásra gyakorolt késleltetett hatásainak elemzését követően a régió szintű adatok segítségével megvizsgáltam a VAR modellek Granger-oksági Wald-teszt p-értékei és az egy főre jutó GDP közti kapcsolat jellegét és szorosságát. Az eredmények alapján elmondható, hogy az egy főre jutó GDP növekedésével arányosan hatványosan csökkennek a Wald-teszt p-értékek, tehát a Granger-oksági kapcsolatok egyre stabilabbak. Az egy főre jutó GDP és a Wald-teszt p-értékek között közepesen szoros kapcsolat áll fenn. Ez nagyrészt annak köszönhető, hogy fejlettebb régiók nagyobb internetpenetrációval rendelkeznek, így több keresési adat áll rendelkezésre. Ennek ismeretében érdemes lehet a modelleket csupán a magas internetpenetrációval rendelkező régiók esetében alkalmazni, a kisebb internetpenetrációval rendelkező régiók esetében pedig egyváltozós idősormodelleket (például ARIMA modellt) alkalmazni. A régiós egy főre jutó GDP és a Wald-tesztek p-értéke közti kapcsolat jellege azért is számít érdekes felfedezésnek, mivel a korábban általam vizsgált kutatások esetében a vizsgált régiók nagymértékű fejlettségének és internetpenetrációjának következtében ilyen egyenlőtlenség a kapcsolatok stabilitásában feltehetőleg nem volt tapasztalható.

Összességében eredményesnek tekintem a kutatást, hiszen nem csupán a kezdeti felvetést sikerült több régióban is igazolni, miszerint a keresési kifejezésekből épített keresési index kompozíció előrejelzi a régiós esetszámok alakulását, de több régió esetében nem-lineáris, illetve fordított hatások is felfedezhetők voltak, melyek még inkább betekintést adnak a járvány terjedésébe és a társadalom sajátos reakcióiba. Emellett a régiók közti szignifikanciák eltéréseinek vizsgálata is eredményesnek bizonyult. Úgy gondolom, hogy több szempontból is előnyös lehet a kutatás során épített vagy hasonló modellek használatával felkészülni egy esetleges újabb koronavírus hullámra. Érdemes lehet továbbá vizsgálni, hogy a jelenlegi modellek, szükség esetén kisebb módosításokkal más járványokra is robusztusok-e.

8. Irodalomjegyzék

Austys, D., & Burneikaitė, M. (2019). Google Trends and forecasting of influenza epidemics in Lithuania. *European Journal of Public Health*, 29(Supplement_4), ckz185.403. <https://doi.org/10.1093/eurpub/ckz185.403>

Átlátszó. (2020, április 26). *Koronatérkép | atlo*. <https://atlo.team/koronaterkep/> Letöltés dátuma: 2023. március 20.

Ben, S., Xin, J., Chen, S., Jiang, Y., Yuan, Q., Su, L., Christiani, D. C., Zhang, Z., Du, M., & Wang, M. (2022). Global internet search trends related to gastrointestinal symptoms predict regional COVID-19 outbreaks. *Journal of Infection*, 84(1), (pp. 56–63). <https://doi.org/10.1016/j.jinf.2021.11.003>

Fantazzini, D. (2020). Short-term Forecasting of the COVID-19 Pandemic using Google Trends Data: Evidence from 158 Countries. *Applied Economics*, 59, (pp. 1–26). <https://doi.org/10.2139/ssrn.3671005>

Google. (2022a). *FAQ about Google Trends data—Trends Help*. <https://support.google.com/trends/answer/4365533?hl=en> Letöltés dátuma: 2022. november 6.

Google. (2022b). *Google Trends*. <https://trends.google.com/trends/explore?q=covid&geo=HU> Letöltés dátuma: 2022. november 13.

Hanck, C., Arnold, M., Gerber, A., & Schmelzer, M. (2021). *Introduction to Econometrics with R*. Chair of Econometrics, Department of Business Administration and Economics, University of Duisburg-Essen. <https://www.econometrics-with-r.org/>

Hastie, T. J., Tibshirani, R., & Friedman, J. (2011). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Huang, R., Luo, G., Duan, Q., Zhang, L., Zhang, Q., Tang, W., Smith, M. K., Li, J., & Zou, H. (2020). Using Baidu search index to monitor and predict newly diagnosed cases of HIV/AIDS, syphilis and gonorrhea in China: Estimates from a vector autoregressive (VAR) model. *BMJ Open*, 10(3), e036098. <https://doi.org/10.1136/bmjopen-2019-036098>

Kirchgässner, G., Wolters, J., & Hassler, U. (2013). *Introduction to modern time series analysis*. Springer Science & Business Media. <https://doi.org/10.1007/978-3-642-33436-8>

Kovács, L. (2022). Feature selection algorithms in generalized additive models under concavity. *Computational Statistics*. Paper: online first, 33 p. <https://doi.org/10.1007/s00180-022-01292-7>

Központi Statisztikai Hivatal. (2022, január 1.a). *21.1.2.2. Egy főre jutó bruttó hazai termék vármegye és régió szerint*. https://www.ksh.hu/stadat_files/gdp/hu/gdp0078.html

Központi Statisztikai Hivatal. (2022, január 1.b). *Magyarország népességének száma nemek és életkor szerint*. <https://www.ksh.hu/interaktiv/korfak/orszag.html>

Li, K., Liang, Y., Li, J., Liu, M., Feng, Y., & Shao, Y. (2020). Internet search data could Be used as novel indicator for assessing COVID-19 epidemic. *Infectious Disease Modelling*, 5, (pp.848–854). <https://doi.org/10.1016/j.idm.2020.10.001>

Li, K., Liu, M., Feng, Y., Ning, C., Ou, W., Sun, J., Wei, W., Liang, H., & Shao, Y. (2019). Using Baidu Search Engine to Monitor AIDS Epidemics Inform for Targeted intervention of HIV/AIDS in China. *Scientific Reports*, 9(1), (pp. 1-12). <https://doi.org/10.1038/s41598-018-35685-w>

McGough, S. F., Brownstein, J. S., Hawkins, J. B., & Santillana, M. (2017). Forecasting Zika Incidence in the 2016 Latin America Outbreak Combining Traditional Disease Surveillance with Search, Social Media, and News Report Data. *PLOS Neglected Tropical Diseases*, 11(1), e0005295. <https://doi.org/10.1371/journal.pntd.0005295>

Minges, M. (2016, január 31). *Exploring the relationship between broadband and economic growth*. World Bank. <https://documents.worldbank.org/pt/publication/documents-reports/documentdetail/178701467988875888/Exploring-the-relationship-between-broadband-and-economic-growth> Letöltés dátuma: 2023. április 8.

Rogers, S. (2016, július 1). *What is Google Trends data—And what does it mean?* Google News Lab. <https://medium.com/google-news-lab/what-is-google-trends-data-and-what-does-it-mean-b48f07342ee8> Letöltés dátuma: 2022. november 13.

Sadilek, A., Hswen, Y., Bavadekar, S., Shekel, T., Brownstein, J. S., & Gabrilovich, E. (2020). Lymelight: Forecasting Lyme disease risk using web search data. *Npj Digital Medicine*, 3(1), (pp. 1-12). <https://doi.org/10.1038/s41746-020-0222-x>

Shumway, R. H., & Stoffer, D. S. (2000). *Time series analysis and its applications* (3rd ed.). Springer.

Spurious correlations. (n.d.). <http://tylervigen.com/spurious-correlations> Letöltés dátuma: 2023 április 5.

Stock, J. H., & Watson, M. W. (2015). *Introduction to Econometrics* (3rd updated ed.). Pearson Education Limited.

Syamsuddin, M., Fakhruddin, M., Sahetapy-Engel, J. T. M., & Soewono, E. (2020). Causality Analysis of Google Trends and Dengue Incidence in Bandung, Indonesia With Linkage of Digital Data Modeling: Longitudinal Observational Study. *Journal of Medical Internet Research*, 22(7), e17633. <https://doi.org/10.2196/17633>

Tse, G., & McLean, L. (2022). Seasonal Trends in Pediatric Respiratory Illnesses: Using Google Trends to Inform Precision Outreach. *Pediatric Emergency Care*, 38(2), e752–e755. <https://doi.org/10.1097/PEC.0000000000002442>

Venables, W. N., & Smith, D. M. (2010). *An Introduction to R*. <https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf>.

Wang, H.-W., Chen, D.-R., Yu, H.-W., & Chen, Y.-M. (2015). Forecasting the Incidence of Dementia and Dementia-Related Outpatient Visits With Google Trends: Evidence From Taiwan. *Journal of Medical Internet Research*, 17(11), e4516. <https://doi.org/10.2196/jmir.4516>

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R* (2nd ed.). Chapman and Hall/CRC. <https://doi.org/10.1201/9781315370279>

Wood, S. N., Goude, Y., & Shaw, S. (2015). Generalized additive models for large data sets. *Journal of the Royal Statistical Society. Series C*, 64(1), (pp. 139–155).

Woodward, W. A., Gray, H. L., & Elliott, A. C. (2017). *Applied time series analysis with R*. CRC press.

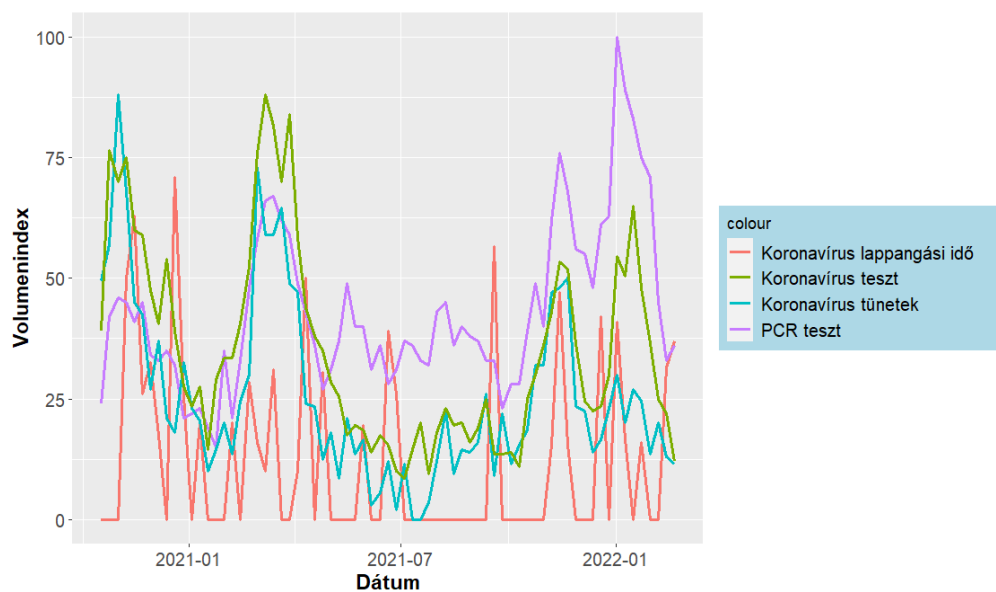
Xu, C., Wang, Y., Yang, H., Hou, J., Sun, L., Zhang, X., Cao, X., Hou, Y., Wang, L., Cai, Q., & Wang, Y. (2019). Association Between Cancer Incidence and Mortality in Web-Based Data in China: Infodemiology Study. *Journal of Medical Internet Research*, 21(1), e10677. <https://doi.org/10.2196/10677>

9. Mellékletek

1. számú melléklet: Keresési kifejezések volumenindexei régiókként

Közép-Magyarország: Budapest

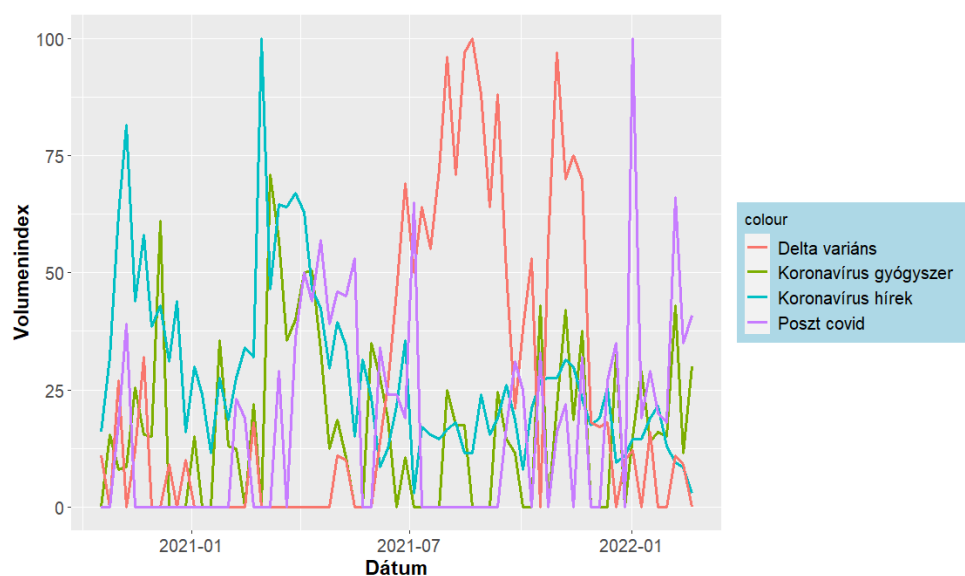
Keresési kifejezések volumenindexei Budapesten I.



9. ábra Keresési kifejezések volumenindexei Budapesten I.

Forrás: R output saját szkript alapján

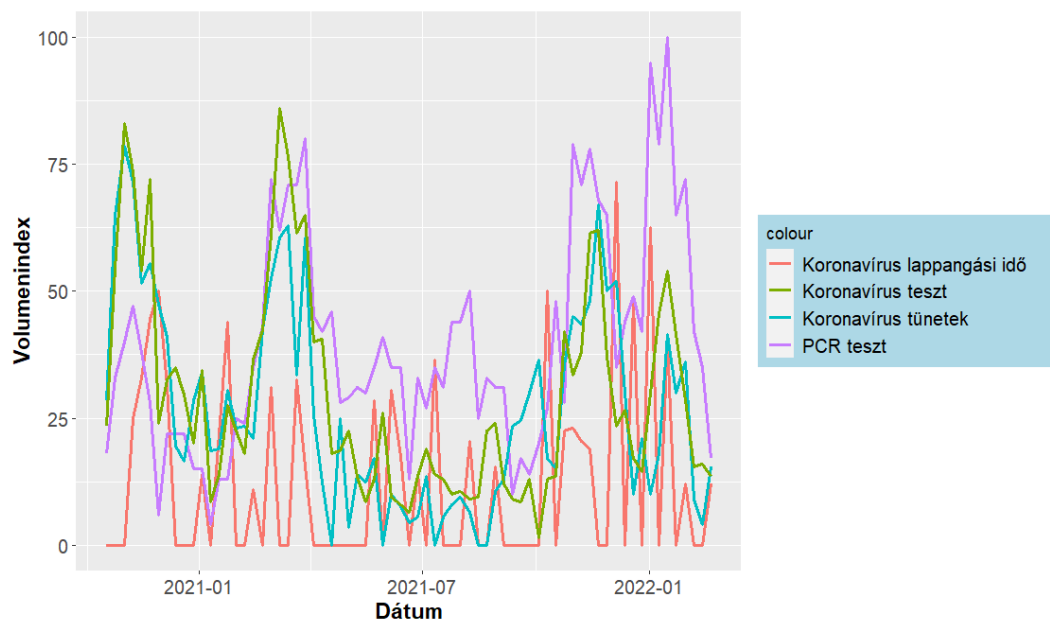
Keresési kifejezések volumenindexei Budapesten II.



10. ábra Keresési kifejezések volumenindexei Budapesten II.

Forrás: R output saját szkript alapján

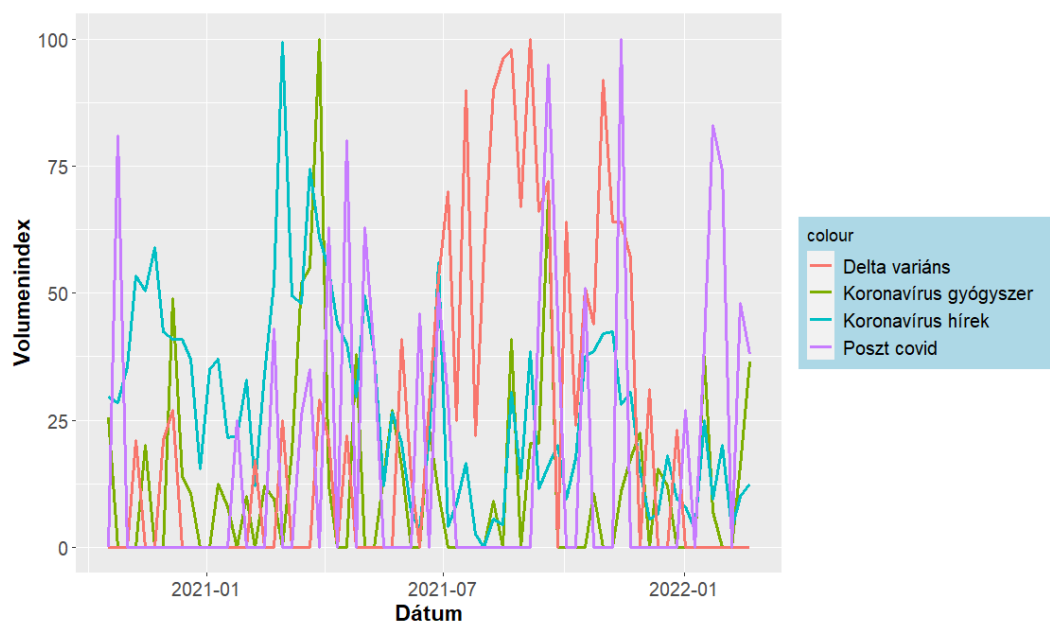
Keresési kifejezések volumenindexei Pest vármegyében I.



11. ábra Keresési kifejezések volumenindexei Pest vármegyében I.

Forrás: R output saját szkript alapján

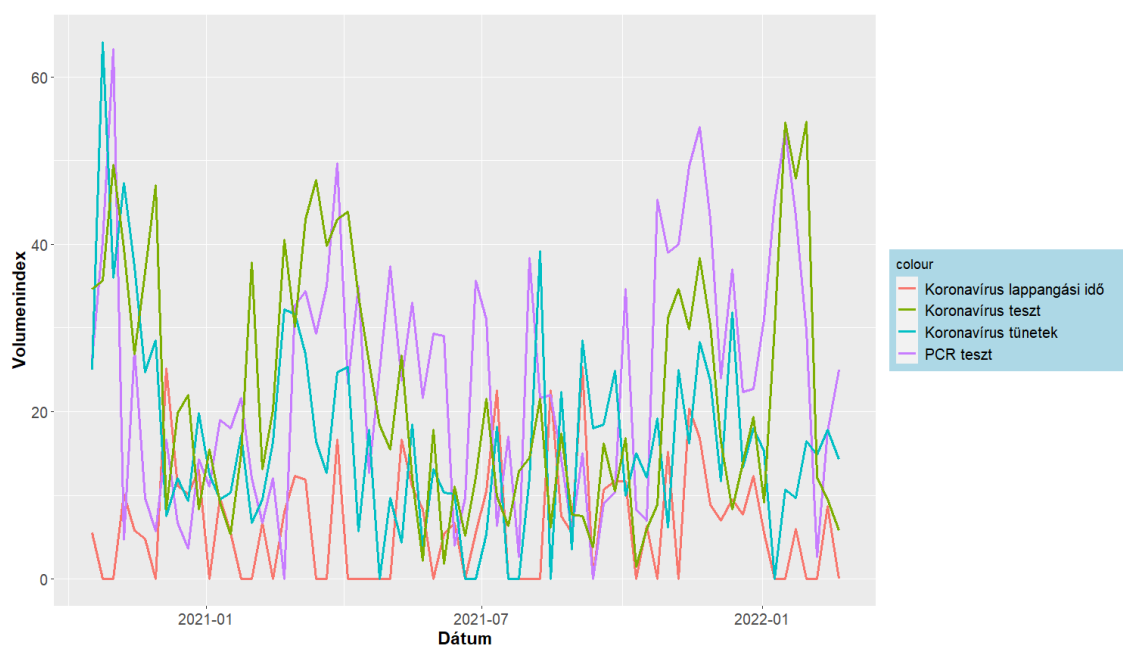
Keresési kifejezések volumenindexei Pest vármegyében II.



12. ábra Keresési kifejezések volumenindexei Pest vármegyében II.

Forrás: R output saját szkript alapján

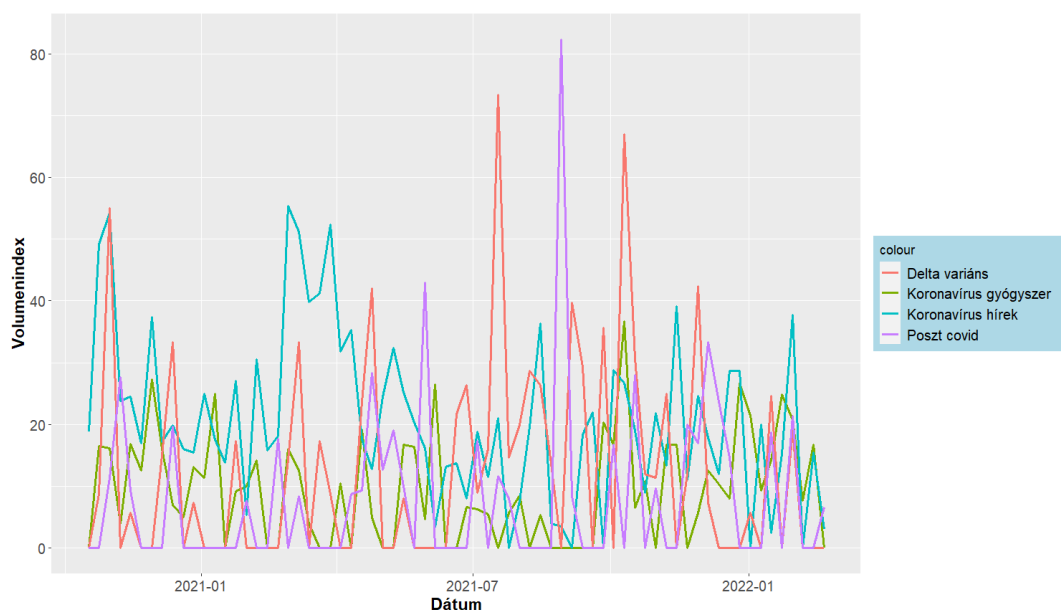
Keresési kifejezések volumenindexei Nyugat-Dunántúlon I.



13. ábra Keresési kifejezések volumenindexei Nyugat-Dunántúl régióban I.

Forrás: R output saját szkript alapján

Keresési kifejezések volumenindexei Nyugat-Dunántúlon II.

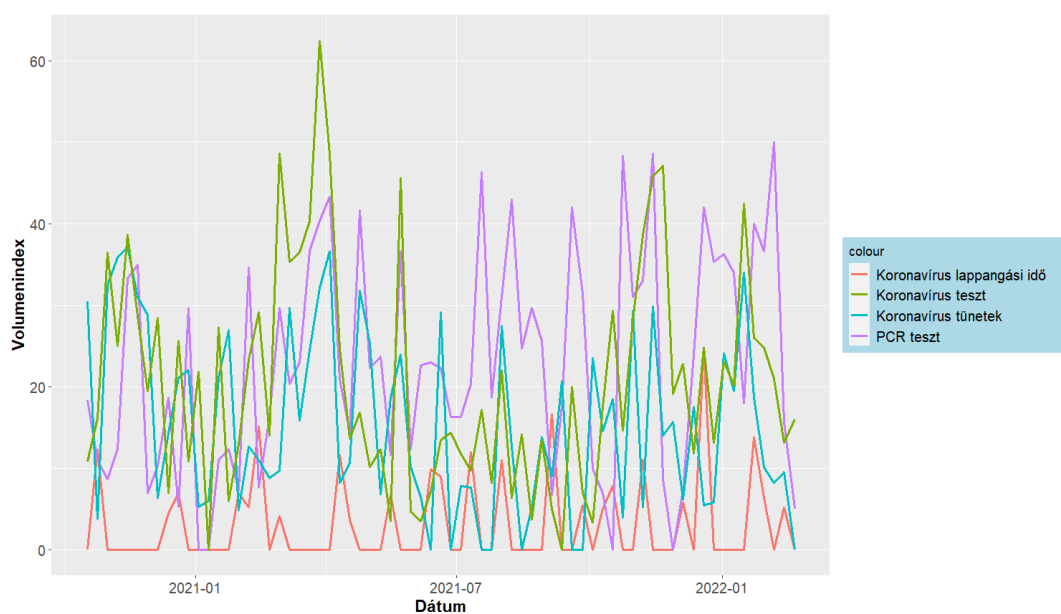


14. ábra Keresési kifejezések volumenindexei Nyugat-Dunántúl régióban II.

Forrás: R output saját szkript alapján

Dél-Dunántúl

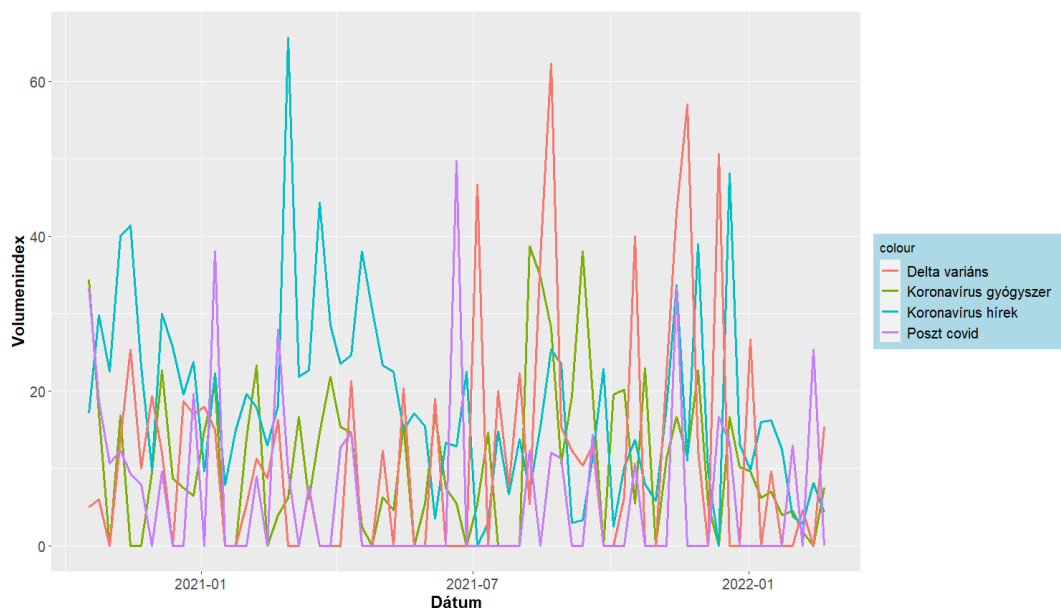
Keresési kifejezések volumenindexei Dél-Dunántúlon I.



15. ábra Keresési kifejezések volumenindexei Dél-Dunántúl régióban I.

Forrás: R output saját szkript alapján

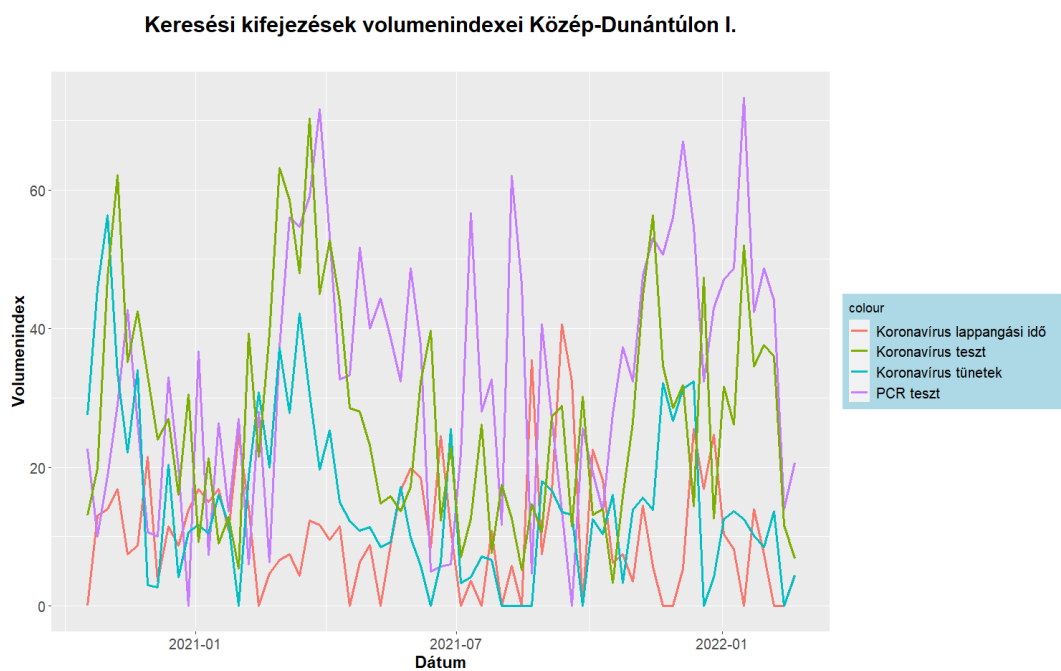
Keresési kifejezések volumenindexei Dél-Dunántúlon II.



16. ábra Keresési kifejezések volumenindexei Dél-Dunántúl régióban II.

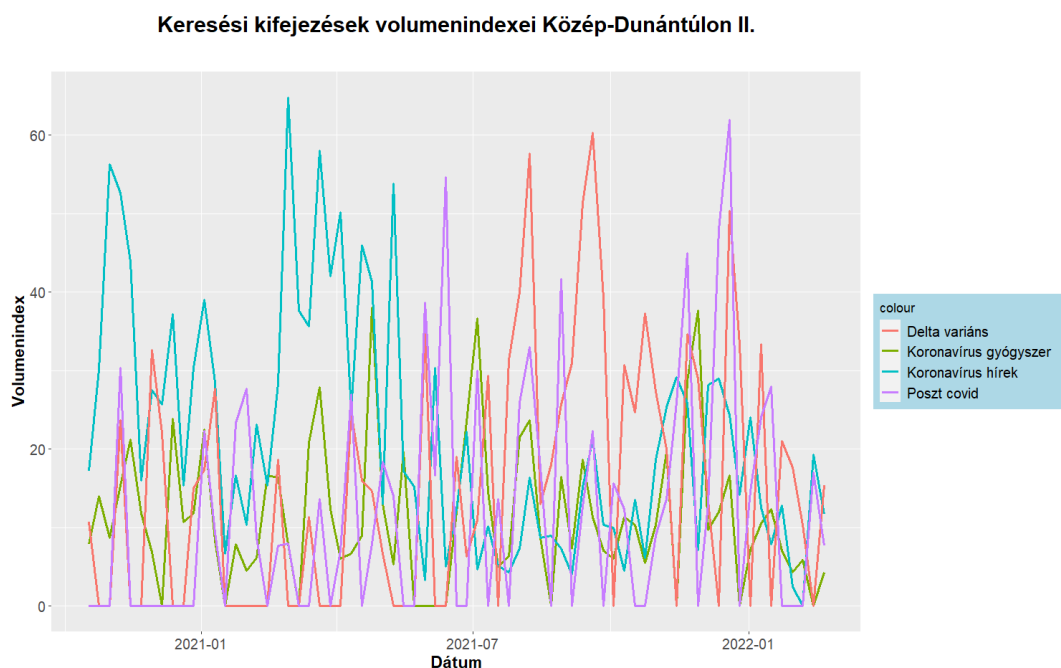
Forrás: R output saját szkript alapján

Közép-Dunántúl



17. ábra Keresési kifejezések volumenindexei Közép-Dunántúl régióban I.

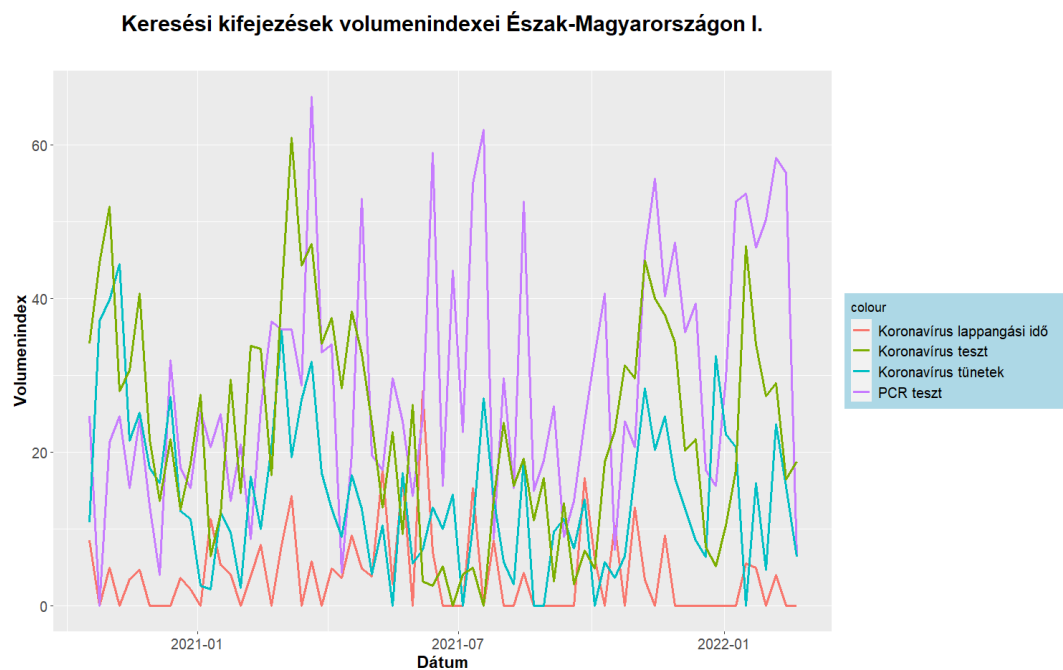
Forrás: R output saját szkript alapján



18. ábra Keresési kifejezések volumenindexei Közép-Dunántúl régióban II.

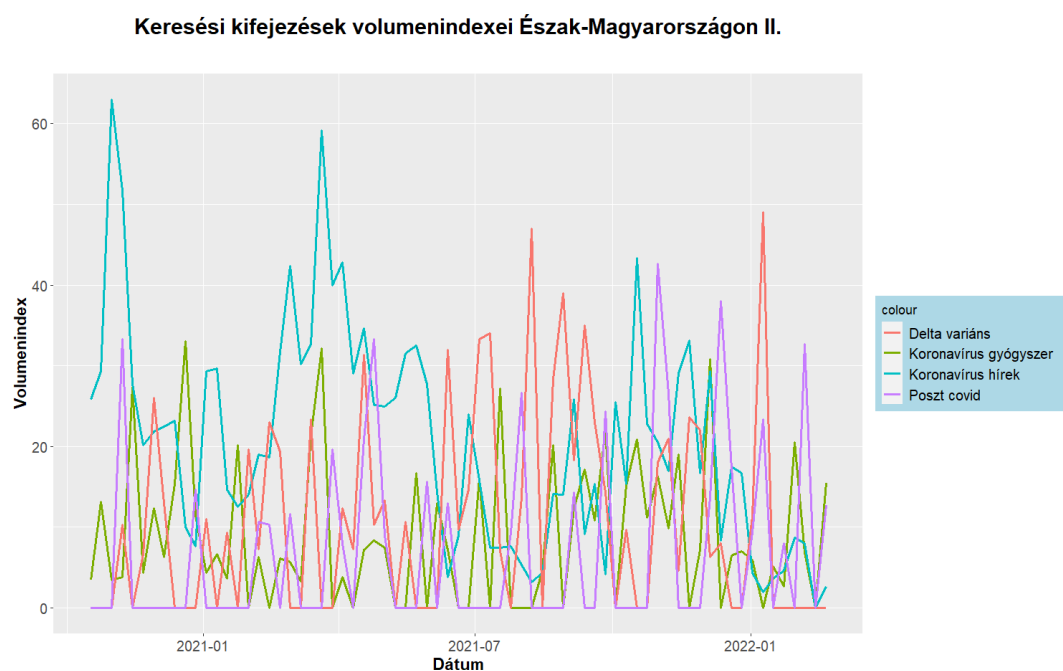
Forrás: R output saját szkript alapján

Észak-Magyarország



19. ábra Keresési kifejezések volumenindexei Észak-Magyarország régióban I.

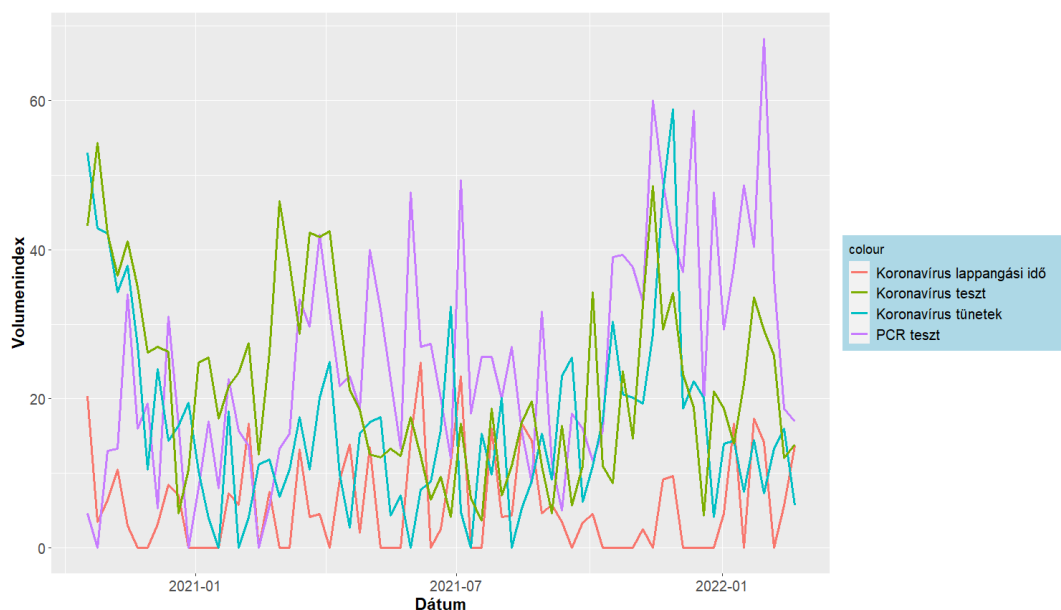
Forrás: R output saját szkript alapján



20. ábra Keresési kifejezések volumenindexei Észak-Magyarország régióban II.

Forrás: R output saját szkript alapján

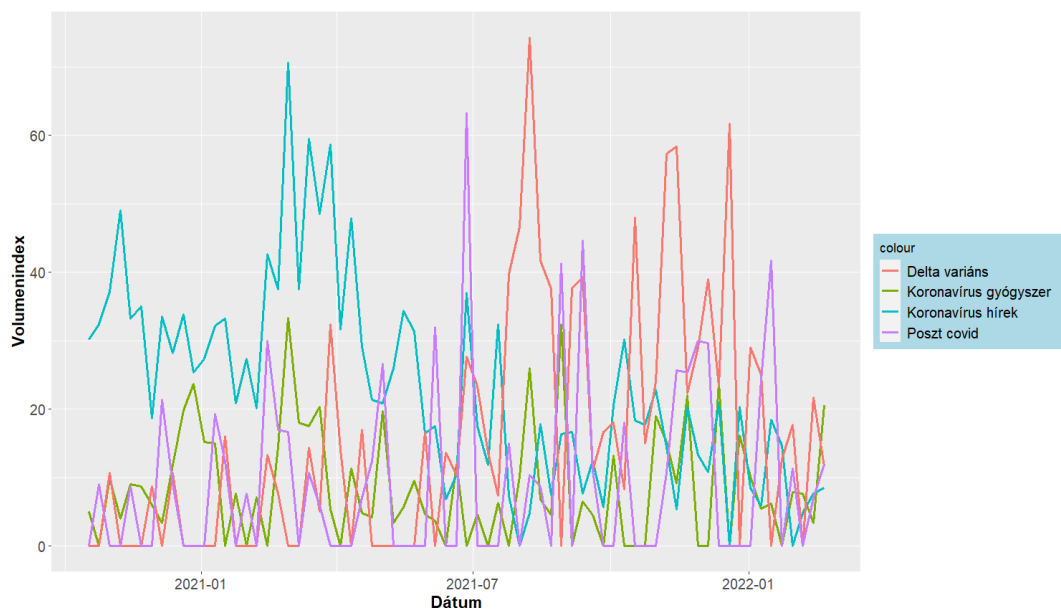
Keresési kifejezések volumenindexei Észak-Alföldön I.



21. ábra Keresési kifejezések volumenindexei Észak-Alföld régióban I.

Forrás: R output saját szkript alapján

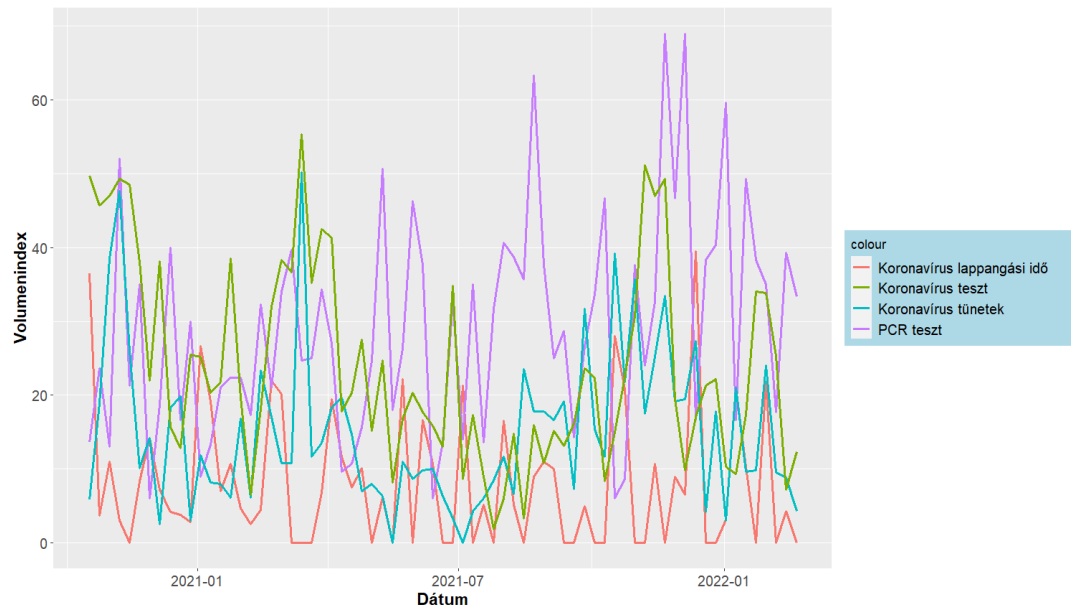
Keresési kifejezések volumenindexei Észak-Alföldön II.



22. ábra Keresési kifejezések volumenindexei Észak-Alföld régióban II.

Forrás: R output saját szkript alapján

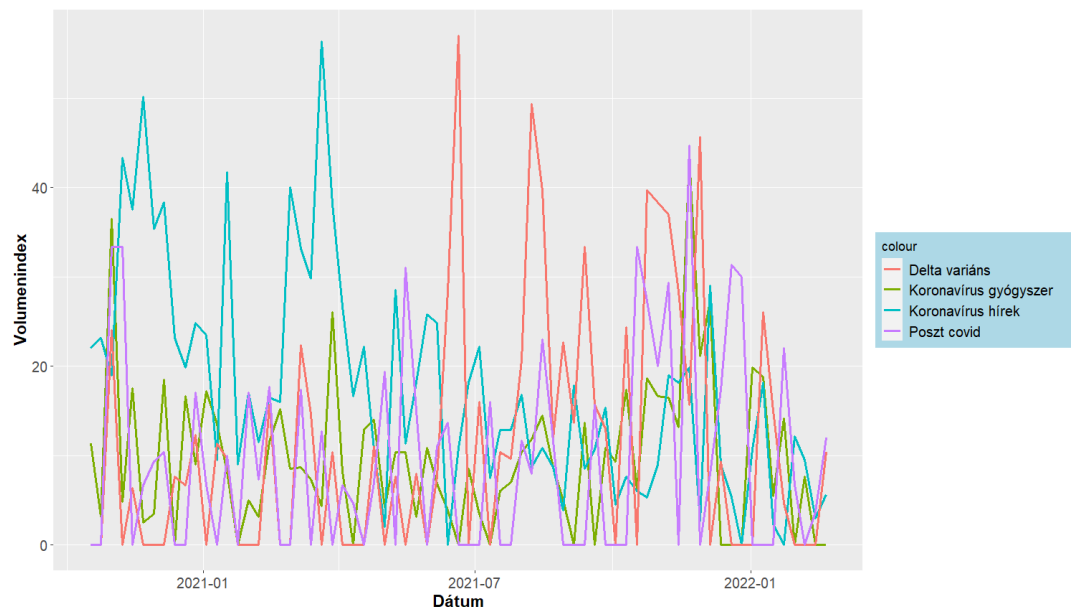
Keresési kifejezések volumenindexei Dél-Alföldön I.



23. ábra Keresési kifejezések volumenindexei Dél-Alföld régióban I.

Forrás: R output saját szkript alapján

Keresési kifejezések volumenindexei Dél-Alföldön II.



24. ábra Keresési kifejezések volumenindexei Dél-Alföld régióban II.

Forrás: R output saját szkript alapján

2. számú melléklet: Stacionaritás tesztek p-értékei régiókként

Pest	Első verzió		Első differenciálás után		Második differenciálás után	
változó	adf p-értéke	kpss p-értéke	adf p-értéke	kpss p-értéke	adf p-értéke	kpss p-értéke
Pest c	0,590	0,010	0,388	0,1	0,01	0,1
gyogy	0,052	0,100	0,010	0,1	0,01	0,1
hir	0,093	0,010	0,010	0,1	0,01	0,1
lapp	0,021	0,100	0,010	0,1	0,01	0,1
teszt	0,022	0,100	0,010	0,1	0,01	0,1
tunet	0,093	0,088	0,010	0,1	0,01	0,1
delta	0,756	0,040	0,015	0,1	0,01	0,1
pcr	0,020	0,046	0,010	0,1	0,01	0,1
poszt	0,025	0,100	0,010	0,1	0,01	0,1

3. táblázat Stacionaritás tesztek p-értékei első és második differenciálás után Pest vármegye esetében

Forrás: saját szerkesztés R outputok alapján

Dél-Alföld	Első verzió		Első differenciálás után		Második differenciálás után	
változó	adf p-értéke	kpss p-értéke	adf p-értéke	kpss p-értéke	adf p-értéke	kpss p-értéke
d alf c	0,706	0,010	0,399	0,1	0,01	0,1
gyogy	0,097	0,100	0,010	0,1	0,01	0,1
hir	0,010	0,010	0,010	0,1	0,01	0,1
lapp	0,010	0,100	0,010	0,1	0,01	0,1
teszt	0,093	0,054	0,010	0,1	0,01	0,1
tunet	0,344	0,100	0,010	0,1	0,01	0,1
delta	0,083	0,055	0,010	0,1	0,01	0,1
pcr	0,010	0,010	0,010	0,1	0,01	0,1
poszt	0,204	0,100	0,010	0,1	0,01	0,1

4. táblázat Stacionaritás tesztek p-értékei első és második differenciálás után Dél-Alföld régió esetében

Forrás: saját szerkesztés R outputok alapján

Dél-Dunántúl	Első verzió		Első differenciálás után		Második differenciálás után	
változó	adf p-értéke	kpss p-értéke	adf p-értéke	kpss p-értéke	adf p-értéke	kpss p-értéke
d_dunan_c	0,664	0,010	0,529	0,1	0,01	0,1
gyogy	0,334	0,100	0,010	0,1	0,01	0,1
hir	0,414	0,012	0,010	0,1	0,01	0,1
lapp	0,010	0,100	0,010	0,1	0,01	0,1
teszt	0,096	0,100	0,010	0,1	0,01	0,1
tunet	0,056	0,045	0,010	0,1	0,01	0,1
delta	0,261	0,100	0,010	0,1	0,01	0,1
pcr	0,025	0,027	0,010	0,1	0,01	0,1
poszt	0,010	0,100	0,010	0,1	0,01	0,1

5. táblázat Stacionaritás tesztek p-értékei első és második differenciálás után Dél-Dunántúl régió esetében

Forrás: saját szerkesztés R outputok alapján

Nyugat-Dunántúl	Első verzió		Első differenciálás után		Második differenciálás után	
változó	adf p-értéke	kpss p-értéke	adf p-értéke	kpss p-értéke	adf p-értéke	kpss p-értéke
ny_dunan_c	0,670	0,010	0,493	0,1	0,01	0,1
gyogy	0,233	0,100	0,010	0,1	0,01	0,1
hir	0,367	0,033	0,010	0,1	0,01	0,1
lapp	0,017	0,100	0,010	0,1	0,01	0,1
teszt	0,050	0,100	0,010	0,1	0,01	0,1
tunet	0,090	0,074	0,010	0,1	0,01	0,1
delta	0,231	0,100	0,010	0,1	0,01	0,1
pcr	0,028	0,100	0,010	0,1	0,01	0,1
poszt	0,016	0,100	0,010	0,1	0,01	0,1

6. táblázat Stacionaritás tesztek p-értékei első és második differenciálás után Nyugat-Dunántúl régió esetében

Forrás: saját szerkesztés R outputok alapján

Észak-Alföld	Első verzió		Első differenciálás után		Második differenciálás után	
változó	adf p-értéke	kpss p-értéke	adf p-értéke	kpss p-értéke	adf p-értéke	kpss p-értéke
e alf c	0,750	0,010	0,465	0,1	0,01	0,1
gyogy	0,010	0,100	0,010	0,1	0,01	0,1
hir	0,190	0,010	0,010	0,1	0,01	0,1
lapp	0,084	0,100	0,010	0,1	0,01	0,1
teszt	0,160	0,047	0,010	0,1	0,01	0,1
tunet	0,130	0,100	0,010	0,1	0,01	0,1
delta	0,517	0,010	0,010	0,1	0,01	0,1
pcr	0,324	0,010	0,010	0,1	0,01	0,1
poszt	0,010	0,100	0,010	0,1	0,01	0,1

7. táblázat Stacionaritás tesztek p-értékei első és második differenciálás után Észak-Alföld régió esetében

Forrás: saját szerkesztés R outputok alapján

Észak-Magyarország	Első verzió		Első differenciálás után		Második differenciálás után	
változó	adf p-értéke	kpss p-értéke	adf p-értéke	kpss p-értéke	adf p-értéke	kpss p-értéke
e mo c	0,672	0,010	0,426	0,1	0,01	0,1
gyogy	0,091	0,100	0,010	0,1	0,01	0,1
hir	0,338	0,013	0,010	0,1	0,01	0,1
lapp	0,392	0,100	0,010	0,1	0,01	0,1
teszt	0,132	0,100	0,010	0,1	0,01	0,1
tunet	0,165	0,093	0,010	0,1	0,01	0,1
delta	0,052	0,100	0,010	0,1	0,01	0,1
pcr	0,104	0,010	0,010	0,1	0,01	0,1
poszt	0,010	0,040	0,010	0,1	0,01	0,1

8. táblázat Stacionaritás tesztek p-értékei első és második differenciálás után Észak-Magyarország régió esetében

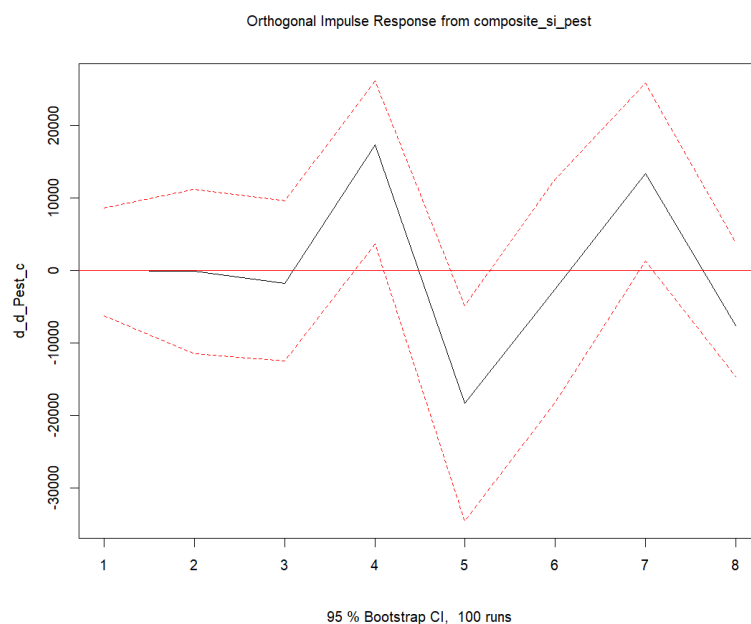
Forrás: saját szerkesztés R outputok alapján

Közép-Dunántúl	Első verzió		Első differenciálás után		Második differenciálás után	
változó	adf p-értéke	kpss p-értéke	adf p-értéke	kpss p-értéke	adf p-értéke	kpss p-értéke
k_dunan_c	0,588	0,010	0,489	0,1	0,01	0,1
gyogy	0,018	0,100	0,010	0,1	0,01	0,1
hir	0,358	0,010	0,010	0,1	0,01	0,1
lapp	0,024	0,100	0,010	0,1	0,01	0,1
teszt	0,069	0,100	0,010	0,1	0,01	0,1
tunet	0,031	0,079	0,010	0,1	0,01	0,1
delta	0,486	0,010	0,010	0,1	0,01	0,1
pcr	0,184	0,062	0,010	0,1	0,01	0,1
poszt	0,172	0,026	0,010	0,1	0,01	0,1

9. táblázat Stacionaritás tesztek p-értékei első és második differenciálás után Közép-Dunántúl régió esetében

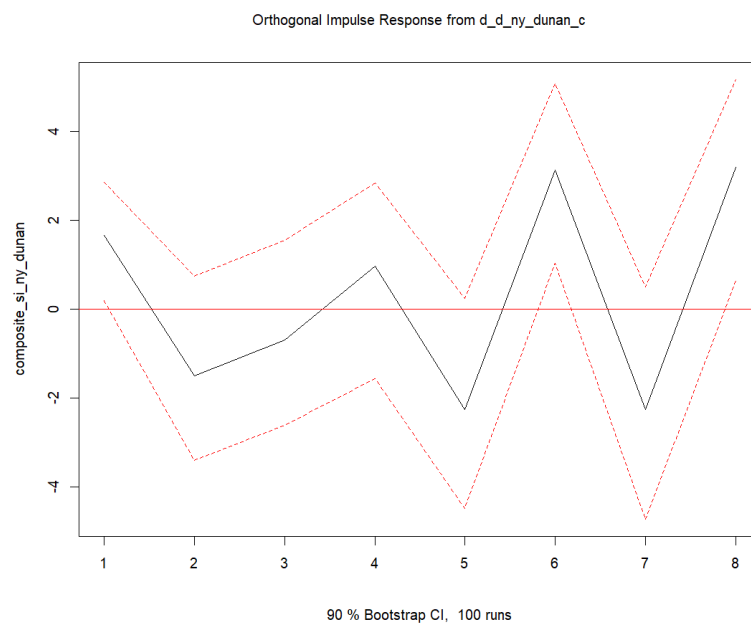
Forrás: saját szerkesztés R outputok alapján

3. számú melléklet: Keresési index kompozíciók és koronavírus esetszámok IVF ábrái régiókként csoportosítva



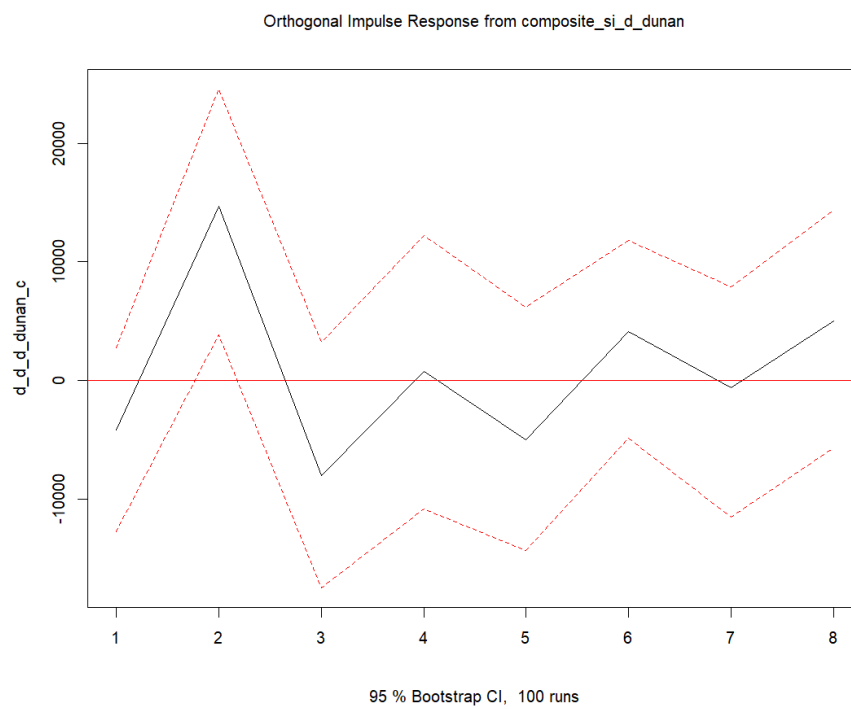
25. ábra Pest vármegye koronavírus esetszámainak IVF ábrája a keresési index kompozícióból érkező sokk hatására (95%-os konfidencia-intervallum)

Forrás: R output saját szkript alapján



26. ábra Nyugat-Dunántúl keresési index kompozíció IVF ábrája a koronavírus esetszámokból érkező sokk hatására (90%-os konfidencia-intervallum)

Forrás: R output saját szkript alapján



27. ábra Dél-Dunántúl koronavírus esetszámainak IVF ábrája a keresési index kompozícióból érkező sokk hatására (95%-os konfidencia-intervallum)

Forrás: R output saját szkript alapján