

OTDK - dolgozat

2024

Budapesti Corvinus Egyetem



Anomália-detekciós eljárások összehasonlítása mortalitási idősorokon

Lukács Emma

Alkalmazott közgazdaságtan BSc III.

Konzulens: Dr. Ferenci Tamás

Kézirat lezárának dátuma: 2023. április 17.

Absztrakt

A járványkezelés és közegészségügyi hatásvizsgálatok kiemelt fontosságú mutatószáma a többlethalálozás. Ennek oka, hogy ez egy olyan kimeneti mérőszám, amin lemérhetjük az intézkedés vagy járvány potenciális hatását. A kutatás során négy eltérő anomália-detekciós eljárást hasonlítottam össze különböző szempontok szerint. Az anomáliák vizsgálatának relevanciája a többlethalálozási elemzések tárgyterületén abból fakad, hogy a halálozási számokban jelentős ingadozások figyelhetők meg a nyári hőhullámok alatt, valamint a téli influenza-szezonban, amiknek bekövetkezése és mértéke szabálytalan mintázatot követ. A mortalitás elemzésére rendkívül nagy mennyiségű adat áll rendelkezésre. Ezek az algoritmusok lehetővé teszik, hogy az adattömegből viszonylag rövid idő alatt releváns információt nyerjünk ki az anomáliákra vonatkozóan. A tanulmány fő témája a k-NN, LOF, izolációs erdő, valamint Seasonal-Hybrid ESD algoritmusok összehasonlítása volt. A kutatás során megállapítottam, hogy ezeknek az algoritmusoknak a különböző típusú anomáliákra való érzékenysége és számításigénye is rendkívül heterogén. Az eredmények részletesen bemutatják, melyik eljárás milyen jellegű anomáliák észlelésében volt hatékony.

Kulcsszavak: anomália-detekció, k-NN, LOF, izolációs erdő, S-H-ESD, mortalitás

JEL klasszifikáció: C49

Tartalomjegyzék

Absztrakt	I
Bevezetés	1
1. Szakirodalmi áttekintés	3
1.1. A többlethalálozás fontossága	3
1.2. Különböző modellezési megközelítések	4
2. Módszertan	7
2.1. Az anomáliákról általában	7
2.2. Legközelebbi szomszéd analízisen alapuló detekció	7
2.3. Lokális kiugrási faktor (LOF – Local Outlier Factor) eljárás	8
2.4. Izolációs erdő	9
2.5. Seasonal Hybrid ESD (S-H-ESD)	10
3. Elemzés	12
3.1. Felhasznált adatok	12
3.2. Trend és szezonaritás	12
3.3. Detekció és következtetések	13
3.4. A besorolás menete	14
4. Eredmények	19
4.1. Eljárások számításigénye nagy mennyiségű adat mellett	19
4.2. Anomália profilok hasonlósága	19
4.3. Futásidő, mint hatékonysági szempont	21
4.4. Impulzussal terhelt adatok becslése	23
4.5. Kitekintés, alkalmazási lehetőségek	24
5. Összegzés	26

Ábrák jegyzéke

1. Maradéktagok idősora	14
2. k-NN alapú becslés eredménye egy adott idősoron	15

3.	LOF alapú becslés eredménye egy adott időszoron	16
4.	Izolációs erdő alapú becslés eredménye egy adott időszoron	17
5.	S-H-ESD alapú becslés eredménye egy adott időszoron	18
6.	A modellek által becsült heti profilok összehasonlítása	20
7.	Futásidő a bemeneti adatok méretének függvényében	22
8.	Futásidő a nagyságrendileg megegyező módszerekre szűkítve	22
9.	Impulzussal terhelt becslés	24

Táblázatok jegyzéke

1.	1250 időszorra vonatkozó számítási idő	19
2.	Modellek futásideje különböző méretű inputok mellett	23

Bevezetés

A járványkezelés és közegészségügyi hatásvizsgálatok alapvető kérdései közé tartozik, hogy mi alapján ítéltető meg egy beavatkozás hatékonysága. Szükség van egy olyan kimeneti mutatóra, amin lemérhetjük az intézkedés potenciális hatását. Ennek legfőbb nehézségeivel az 1. fejezetben részletesen foglalkozom.

Az egyik erre alkalmas mutató a többlethalálozás, aminek meghatározásához minden esetben elengedhetetlen az adott területen, korábbi adatok alapján várható halálozási számok ismerete. Ennek pontos meghatározása elősegíti és hitelesen alátámasztja az intézkedések kifejtett hatásáról alkotott véleményeket, segít felmérni különböző egészségügyi jelenségek veszélyeit és bizonyos attribútumait. A várható halálozás megbecslésére statisztikai módszerek széles tárháza áll rendelkezésre. Ezeken a modelleken belül azonban sokszor kis hangsúlyt kapnak a télen és nyáron előforduló, szabálytalan rendszerességgel bekövetkező ingadozások. Kutatásom során különböző anomália-detekciós eljárásokat hasonlítottam össze. Minden detekció alapja egy hosszútávú kvadratikusan trend, valamint egy harmonikus tag becslése volt, ami a szezonális hatásokat hivatott megragadni. Ezeken felül marad egy olyan komponens, ami magában foglalja a nyári hőhullámok és téli influenza-szezon okozta eltérő valószínűséggel bekövetkező kilengéseket. Ezeknek a mintázatoknak a felismerése a legtöbb esetben szabad szemmel is viszonylag egyszerű, de fontos, hogy nem feltétlenül csak extrém értékek léphetnek fel, hanem kisebb mértékben kilógó, de az alapmintázattól jelentősen eltérő megfigyelések is. A másik fontos szempont, hogy nagy mennyiségű idősor esetén nem hatékony mindegyiknél külön elemezni, hogy hol jelenhetnek meg ezek az outlierok. Ezt a folyamatot jelentősen leegyszerűsítheti az anomália detekciós algoritmusok használata, amiknek célja kifejezetten az ilyen, normálistól eltérő értékek felkutatása az adathalmazban. A kutatás során négy különböző anomália detekciós eljárást hasonlítottam össze annak tekintetében, hogy a halálozási számok esetében mennyire hatékonyan képesek meghatározni a releváns anomáliákat.

A tanulmány fő témája a k-NN, LOF, izolációs erdő, valamint Seasonal-Hybrid ESD algoritmusok összehasonlítása különböző szempontok szerint. Megvizsgáltam, melyik modell tudja leghatékonyabban megállapítani a teljes megfigyelt időszakon egy adott pontról, hogy anomália-e, milyen jellegzetes csoportosulásokra érzékeny leginkább az adott eljárás – például nagy súlyt fektet-e a szezonális ingadozásokra, vagy csak a drasztikus kilengéseket észleli –, valamint melyik tudja legjobban jellemezni adott területet a kiugró

értékek alapján. Ezen kívül figyelembe vettem az algoritmusok számításigényét is a teljes lefutáshoz szükséges idő tekintetében, hiszen nagy adattömegek esetében ez a szempont komoly relevanciával bír. Azt is megvizsgáltam, mennyiben befolyásolja a futásidőt a bementi adattábla mérete. Konklúzióként összefoglalom, melyik módszer milyen körülmények között volt hatékony.

A kutatás fő motivációja a legideálisabb anomália detekciós eljárás megadása kifejezetten mortalitási idősorokra vonatkozóan, aminek segítségével minden bemeneti adatbázisra felírható egy olyan anomália profil, ami akár komolyabb becslések alapjául is szolgálhat, ezzel elősegítve, hogy a hatásvizsgálat a lehető legobjektívebb eredményt szolgáltatassa.

A tanulmány további része az alábbi módon épül fel. Az 1. fejezet ismerteti az alapvető fogalmakat, alkalmazhatósági kérdéseket, valamint kifejezetten többlethalálozás becslésével foglalkozó elemzések eredményeit. Ezek után a 2. fejezetben részletesen foglalkozom a módszertannal, a különböző eljárások statisztikai háttérével és ezek alkalmazhatóságával ebben a specifikus környezetben. A 3. fejezet ismerteti a felhasznált adatbázist, a különböző eljárások besorolási mechanizmusát, végül egy példán keresztül szemlélteti működésüket. A 4. fejezet tartalmazza az algoritmusok összehasonlítását számos különböző szempont szerint, végül az 5. fejezet összefoglalja a tanulmányt.

1. Szakirodalmi áttekintés

A fejezetben bemutatok pár olyan tanulmányt, amik kifejezetten többlethalálozási becslésekkel és a várható halálozás idősorának modellezésével foglalkozik. Ezek a munkák nem csak azért kiemelten fontosak, mert támpontot adnak az elemzés további menetéhez, hanem a kutatás kiterjesztéséhez is alapot szolgáltatnak.

1.1. A többlethalálozás fontossága

A többlethalálozás gyakran alkalmazott mutatószám, de kifejezetten nagy jelentősége volt a SARS-CoV-2 pandémia idejében. Komoly feladatnak bizonyult ugyanis egy működőképes, országok közötti releváns összehasonlításra alkalmas statisztikát kialakítani. A COVID-19 járvány adatbázisaival több probléma is akadt: a pandémia okozta halálok felvétele nem történt meg a tesztelés hiányából fakadóan vagy az alacsony teszthatékonyság miatt, sok esetben történthetett félrediasztizálás, valamint nehéz identifkálni a kifejezetten a járvány vagy felülfertőzés és egyéb szövődmények okozta halálozás közötti vékony határvonalat. Főleg a járvány kitörésének kezdetén, a különböző országokban a rutin regisztrációs rendszerekben a halálokok megállapításának módja jelentősen eltért, akár országokon belül is váltakozhatott (Páldy és Bobvos, 2013).

A különböző megközelítések közé tartozik például, hogy azt vették figyelembe, ha a beteg bármikor a halál beállta előtt COVID-19 pozitív teszteredményt mutatott: vagy a halálozás bekövetkezését megelőzően – egy előre meghatározott időszakban – állapították meg a fertőzöttséget, vagy nem volt pozitív teszteredménye, de a tünetei COVID-19 megbetegedésre utaltak (Beaney et al., 2020).

Az Egészségügyi Világszervezet kiadott egy közleményt COVID-19 halálesetek besorolására vonatkozóan. Ebben úgy határozza meg, hogy a pandémia hatásaként jegyezhető egy haláleset, amennyiben a halál okaként nagy biztossággal megállapítható a járvány, beleértve a megerősített és a feltételezett eseteket is. Amennyiben a COVID-19 hozzájáruló tényező, de nem a halálhoz közvetlenül vezető ok, nem tekinthető relevánsnak. Ez az ajánlás azonban relatíve későn – 2020 áprilisának végén – érkezett, így sok ország nem adaptálta megfelelően a már kialakult rutinba (World Health Organization, 2020).

Ez egyesek szakemberek szerint arra enged következtetni, hogy a világjárványnak a halálozási számokra gyakorolt hatásáról az egyetlen megbízható adat a teljes többlethalálozási

mutató, amely definíció szerint figyelembe veszi a nem azonosított COVID-19 haláleseteket, valamint a lezárással és/vagy az egészségügyi szolgáltatásokban bekövetkezett változások okozta további haláleseteket is (Pearce et al., 2020).

1.2. Különböző modellezési megközelítések

A többlethalálozás kiszámításának kritikus kérdése, hogy adott területen mennyi az átlagosan várható halálok száma, hiszen ehhez tudjuk viszonyítani azt a többletet, amit esetleges külső beavatkozás vagy járvány okoz.

Kirpich et al. (2022) tanulmányukban kifejtik, hogy az országok (vagy régiók) közötti formális összehasonlítások alapvetően standardizált pontszámok segítségével végezhetők el. A P-pontszámok lehetővé teszik, hogy a különböző időszakokra és különböző népesedésű régiókra vagy országokra vonatkozó halálozási számok összehasonlíthatóak legyenek. Ezek a következőképpen adhatók meg:

$$\mathcal{P}(t_i) = \frac{x(t_i) - \hat{\mu}(t_i)}{\hat{\mu}(t_i)}$$

ahol $x(t_i)$ a valós, $\hat{\mu}(t_i)$ pedig a modell által előrejelzett népességhalandóság t_i vizsgált időszak esetében az azt megelőző n év adatainak felhasználásával végzett modellillesztés alapján.

Az a modell, amit végül a becsült $\hat{\mu}(t_i)$ értékek előállítására választunk a megfigyelt időszakon belül, jelentőséget kell, hogy tulajdonítson a vizsgált populációban egyidejűleg zajló (egymással versengő) folyamatoknak, és be kell építenie ezek mortalitásra gyakorolt hatásainak szétválaszthatóságát. Úgy gondolják, a modellnek ahhoz, hogy valóban hatékony lehessen, tartalmaznia kell legalább a mortalitás éves szezonálisát, a lehetséges késleltetetteket, valamint a vizsgált populáció szerkezetében és méretében felmért változásokat. A modellezésre a tanulmányban két idősoros technikát használtak: egy dekompozíciós előrejelző modellt és egy autoregresszív integrált mozgóátlag (ARIMA) modellt.

Az első modell Taylor és Letham (2018) tanulmánya alapján a következőképpen definiált:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t$$

Itt $g(t)$ a trendfüggvény, amely az idősorok értékének nem periodikus változásait modellezi, $s(t)$ pedig a periodikus változásokat (pl. heti és éves szezonalitást), végül $h(t)$ a extrém események (vagy ünnepnapok) hatásait jelenti, amelyek potenciálisan szabálytalan ütemezésből

származhatnak. Az ε_t hibatag minden olyan egyedi változást jelöl, amelyet a modell nem tartalmaz.

A modell specifikációja egy általánosított additív modellre (GAM) emlékeztet. A GAM-formulának megvan az az előnye, hogy egyszerűen végezhető rajta dekompozíciós eljárás és szükség esetén rugalmasan fogad be új paramétereket, például ha a szezonális valamilyen új forrását azonosítják. A GAM-on alapuló illesztések emellett rendkívül gyorsnak is számítanak (Hastie és Tibshirani, 1987).

Egy teljesen eltérő modellezési megközelítés az ARIMA-modell. Ez három paraméter (p , d , q) segítségével adható meg. A p paraméter határozza meg a modell késleltetését, azaz azoknak a korábbi megfigyeléseknek a számát, amelyeket az aktuális megfigyelés becslésére használnak fel. A d paraméter a modell "fokát" jelöli, vagyis azt, hogy hányszor kell differenciálni a változót a szükséges stacionaritási modellfeltevés biztosításához. Például $Y_t, Y_{t-1}, \dots, Y_1$ idősor definiálása mellett $d = 1$ -es specifikációban a modellezni kívánt változó már nem $y(t) = Y(t)$, hanem $Y_t - Y_{t-1}$. A szükséges differenciák számának függvényében az idősort n -ed rendben integrálnak nevezzük. A harmadik paraméter (q) a modellen belüli mozgóátlag-paraméterek számát jelöli. A teljes egyenlet általánosított formában:

$$y_i = c + \sum_{i=1}^p \varphi_i y_{t-1} + \sum_{i=1}^1 \theta_i \varepsilon_{t-1} + \varepsilon_t$$

Az ARIMA kiterjesztéseként előfordul, hogy az eredeti idősoron felül egy externális változót is bevonnak a becslésbe. Ez könnyen implementálható a korábbiakban kifejtett egyenlet kiegészítésével. Ezt a modellt ARIMAX-nak nevezzük, melynek általános felírása:

$$y_i = c + \sum_{i=1}^p \varphi_i y_{t-1} + \sum_{i=1}^1 \theta_i \varepsilon_{t-1} + \sum_{i=1}^M \beta_m X_{m,t} \varepsilon_t$$

Acosta és Irizarry (2022) módszerükben Poisson regressziós becsléseket vettek alapul. Az eseményhatást az idő – matematikai definíció szerint – sima függvényeként modellezték. Mivel a természetes változékonyság több varianciát foglal magában, mint amit a Poisson-modell előrejelez, valamint a napi adatok időbeli korrelációt mutatnak, olyan vegyes hatású modellt alakítottak ki, amely egy autoregresszív folyamatot is tartalmaz. A Poisson regresszió az általánosított lineáris modellek családjába tartozik. Azt feltételezi, hogy az Y eredményváltozó Poisson-eloszlású, valamint, hogy várható értékének logaritmusát ismeretlen paraméterek lineáris kombinációjával modellezhető.

A napi halálozások számára felírt modell ezek alapján így néz ki a megfigyelési időszakra, ahol $t = 1, \dots, T$:

$$Y_t | \varepsilon_t \sim \text{Poisson}(\mu_t[1 + f(t)]\varepsilon_t)$$

A μ_t várható érték itt tovább bontható, azaz magában foglalja a modell dekompozíciós jellegét:

$$\mu_t = N_t \exp\{\alpha(t) + s(t) + w(t)\}$$

N_t a populációt jelöli t időpontban, amely egyfajta eltolásként kezelt paraméter, $\alpha(t)$ egy lassan mozgó trendet határoz meg, ami a szekuláris változásokat magyarázza – például javuló egészségügyi tendenciákat –, $s(t)$ egy szezonális trendet reprezentáló éves periodikus függvény, végül $w(t)$ a hét napjának hatása.

A felsorolt modellek módszertanilag jelentősen eltérnek egymástól, de az alapgondolat mindegyikben hasonló. A cél az idősor felbontása valamilyen logika szerint olyan főbb komponensekre, amik azt a lehető legjobban képesek leírni. A különböző megközelítések áttekintése alapján arra jutottam, a legtöbb modellben többletinformációként bevonható lenne egy olyan változó, ami a hőmérsékleti vagy influenzaszezon hatásából fakadó kiugrásokat ragadja meg. Ezek az anomáliák ugyanis a halálozási idősorok sajátjai, amit adott esetben akár indokolatlanul tulajdoníthatnak az elemzések a vizsgált külső tényezőnek vagy beavatkozásnak annak ellenére, hogy nem abból fakadnak. Elképzelhető olyan hierarchikus modellspecifikáció is, ami adott terület és időszak kiugrásokra való hajlamosságát valamilyen jól illeszkedő valószínűségi eloszlás meghatározásával építi be az elemzésbe.

2. Módszertan

Ebben a részben összefoglalom, mit tekintünk általában véve anomáliának. Ezután pedig részletesen kifejtem az egyes algoritmusok működési elveit és felépítését.

2.1. Az anomáliákról általában

Az anomália detekció a statisztikai módszereknek egy olyan ága, amelynek célja az outlierok felismerése az adatokban. Az idősorok esetében anomáliákról akkor beszélhetünk, ha a megfigyelt időszakban előfordul olyan pont vagy szekvencia, ami szignifikánsan eltér az adatok normál viselkedésétől.

[Shaukat et al. \(2021\)](#) tanulmányukban három anomáliatípust különböztetnek meg.

A *pontanomália* egy olyan megfigyelést takar, ami valamilyen előre meghatározott szempont alapján eltér az adathalmaz többi elemétől. Ebben az esetben az algoritmus egyszerű lefutású, bizonyos limitérték meghatározását követően egyértelműen, a többi adattól függetlenül kezelve meg tudja állapítani a besorolt elemről, hogy anomália-e.

Az ún. *kontextusfüggő anomáliák* csak az őket körülvevő adatok függvényében detektálhatók. Nagyobb kihívást jelentenek azok a kontextuális anomáliák, amelyek a normál tartományon belül jelentkeznek, de nem felelnek meg az elvárt időbeli mintázatnak.

Az utolsó kiemelt kategória a *kollektív anomáliák* csoportja, ahol az előbbiektől eltérően nem adatpontok, hanem csak azok halmazai detektálhatók. Külön-külön vizsgálva a halmaz pontjai megfelelnének az elvárt mintázatnak, de együttesen kiemelendők valamilyen csoportjellegük alapján. Az utóbbi két esetben, azaz a szekvenciális anomáliák észlelésekor, gyakori hogy az algoritmusok az adatok részsorozatait olyan pontokká alakítják át, amik az adott részsorozatra vonatkozó információkat tartalmazzák. Ezután már a pontanomáliák detektálására szolgáló módszereket a szekvenciális anomáliák esetében is lehet alkalmazni.

2.2. Legközelebbi szomszéd analízisen alapuló detekció

Az anomália detektálás egyik legintuitívabb módszere a pontok közötti távolságok vizsgálata. Ennek a megközelítésnek az egyik legnépszerűbb típusa a k legközelebbi szomszéd távolságán alapuló eljárás. Legyen adott egy k paraméter, a megfigyelések összességét magába foglaló M halmaz, valamint egy távolságot mérő függvény $d(\cdot, \cdot)$. Egy adott t megfigyelés $dis_k(t)$ -vel jelölt k -távolsága az elem M -beli k -adik legközelebbi szomszédjától vett távolságával egyenlő.

A távolság mértéke az elemzés során az euklideszi távolság volt, ami két pont között az alábbi módon adható meg: $d(x_1, x_2) = \sqrt{(x_1 - x_2)^2}$.

Amennyiben nem csak egy szomszédától, hanem egy egész meghatározott szomszédságtól vett távolságot szeretnénk vizsgálni, az egy N távolsági mátrix segítségével adható meg, ami matematikai jelölésszerrel a következőképpen írható fel. A mátrix minden pontban az adott k -adik szomszédától vett távolságot tartalmazza, ahol i a megfigyelések száma.

$$N = \begin{bmatrix} dis_1(1) & dis_2(1) & dis_3(1) & \dots & dis_k(1) \\ dis_1(2) & dis_2(2) & dis_3(2) & \dots & dis_k(2) \\ dis_1(3) & dis_2(3) & dis_3(3) & \dots & dis_k(3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ dis_1(i) & dis_2(i) & dis_3(i) & \dots & dis_k(i) \end{bmatrix} \quad (1)$$

Ezek után az egyes elemekhez tartozó $knn(i)$ értéket a mátrix sorátlagai adják meg. Azaz a pont detektálásakor az eljárás a szomszédoktól való távolságokat együttesen veszi figyelembe. Minél kisebb ez a távolság, annál biztosabban állítható, hogy a t -t normál esetként definiált adatpontok veszik körül, ami növeli az esélyét annak, hogy t maga is normális elem. A korábbi feltételek fennállása mellett vegyünk egy δ távolsági határt. Ebben az esetben egy t elem k -NN anomáliának tekinthető, ha a hozzárendelt érték meghaladja δ -t, egyébként pedig k -NN normalitásnak. Ez az anomáliafogalom az alacsony sűrűségű területek detekcióját jelenti a gyakorlatban. A δ egy hiperparaméter, amit az elemző határoz meg, elvi szinten a pontosság és szenzitivitás trade-off döntési változója.

2.3. Lokális kiugrási faktor (LOF – Local Outlier Factor) eljárás

A k -NN módszertan egyfajta kiegészítése az ún. LOF analízis. Ez egy olyan nem felügyelt anomália-felismerő módszer, amely kiszámítja egy adott adatpont sűrűségbeli eltérését a szomszédaihoz képest. Azokat az eseteket tekinti kiugrónak, amelyek sűrűsége lényegesen alacsonyabb, mint a szomszédaié. Az N -ként definiált mátrixban M halmaz több eleméhez is ugyanazt a $dis_k(t)$ távolságot rendelhetjük hozzá, azaz nem egy kölcsönösen egyértelmű hozzárendelést definiálunk. Ezzel szemben a LOF értékek egyediek. Első lépésben itt is kinyerhetjük k legközelebbi szomszédától való távolság értékét a távolságmátrixból. Adott o elem k szomszédságába tartozó megfigyelésekhez viszonyított sűrűségét szeretnénk megállapítani. A mérték az elérhetőségi távolság mutatóján alapszik, amit a $\max\{dis_k(o), d(t, o)\}$ függvény

ad meg. Egy t megfigyelés elérhetőségi távolsága o -tól a két objektum valódi távolsága, de legalább az o elem k -távolságával adható meg. A lokális elérhetőségi sűrűség a LOF alapja, ami az adott objektum szomszédjaitól vett átlagos elérhetőségi távolságának inverzével számítható.

$$lrd(i) = \frac{1}{\frac{1}{k} \sum_{o \in N} \frac{\max\{dis_k(o), d(i,o)\}}{|N|}} \quad (2)$$

$$lof(i) = \frac{1}{k} \sum_{o \in N} \frac{lrd_k(o)}{lrd_k(i)} \quad (3)$$

A képlet azt mutatja, hogy az egyes pontok lokális elérhetősége a szomszédos pontok lokális elérhetőségének átlagával összehasonlítva megadja a lokális kiugrási faktor értékét. Ezek alapján δ határ mellett definiálhatjuk t elemet, mint LOF anomália, abban az esetben, ha $lof_k(t) > \delta$ (Teng, 2010).

2.4. Izolációs erdő

A korábban bemutatott legközelebbi szomszédoktól számított távolságon alapuló módszerekkel szemben több ellenérv is felhozható. Mindegyik azon a mechanizmuson alapszik, hogy felépíti a normálisként definiált elemek profilját, majd az anomáliákat olyan elemekként azonosítja, amelyek nem felelnek meg ezeknek a normális attribútumoknak. Az anomáliafelismerő képességük általában egy olyan algoritmus mellékterméke, amit eredetileg más célra terveztek. Éppen ebből kifolyólag ez a megközelítés közel sem az anomáliák detektálására optimalizált. Ezzel szemben az izoláción alapuló algoritmusok egy teljesen más irányt képviselnek, aminek nincsen köze semmilyen korábban kifejtett távolságmértékhez (Liu et al., 2012).

Az izolációs erdő egy ensemble technika, azaz egy hiperparaméter segítségével magadott számú izolációs fa bizottsági szavazata alapján kapnak az egyes pontok egy-egy anomália értéket. Az izolációs fák felépítése megegyezik egy BST-vel (binary search tree). A bináris fa olyan fa adatszerkezet, ahol minden csúcsnak legfeljebb 2 gyereke van, amiket leggyakrabban bal és jobb oldali gyerekeknek nevezünk. A bináris keresőfa a bináris fák egy speciális típusa, amelyben egy csomópont bal oldali gyerekének értéke kisebb, mint a csomópont értéke, a jobb oldali gyerekének értéke pedig nagyobb, mint a csomópont értéke. Ezt a tulajdonságot BST-tulajdonságnak nevezzük és lehetővé teszi a fa elemeinek hatékony keresését, beszúrását és törlését (Cormen et al., 2001).

Legyen $M = \{m_1, m_2, \dots, m_n\}$ d -dimenziós pontok halmaza és $M' \subset M$ valódi részhalmaz.

Ekkor az izolációs fa az alábbi adatstruktúra alapján áll elő:

1. A fában található minden T csúcsra igaz, hogy vagy egy külső csúcs, ami levél vagy egy belső csúcs, ami rendelkezik egy döntési teszttel és pontosan 2 gyerekkel.
2. Adott csúcsnál lévő teszt egy q ismérvből és egy p vágási értékből épül fel. A q ebben az esetben maga az eredményváltozó, tehát a halálozások száma adott pontban, p pedig az ismerv tartományában felvett véletlen érték. Ez alapján a $q < p$ teszt meghatározza a pont besorolását T_l -be vagy T_r -be.

Az algoritmus a fát úgy növeszti, hogy iteratívan felosztja M' -t q ismerv mentén egészen addig, míg (i) a csúcsnak egy megfigyelése marad, (ii) minden csúcsnál lévő pontnak ugyanaz az értéke. Egy teljesen kifejlett izolációs fában M halmaz minden pontja valamely külső csúcsban izolált. Az intuíciónak megfelelően anomáliáknak tekinthetők azok a pontok, amik egyszerűbben izolálhatók, azaz kisebb a hozzájuk rendelt úthossz ($h(m_i)$) a fában. Ezzel a mutatóval azonban az a probléma, hogy erősen függ a tesztminta elemszámától, így valamilyen formában standardizálása van szükség ahhoz, hogy egymással összehasonlítható anomália értékeket tudjunk megadni. Ennek kulcsa, hogy az izolációs fa esetében egy külső csúcsra érkezés a BST sikertelen keresésének felel meg, aminek pedig megadható az átlagos úthossza:

$$c(\psi) \begin{cases} 2H(\psi - 1) - 2(\psi - 1)/\psi & \text{ha } \psi > 2 \\ 1 & \text{ha } \psi = 2 \\ 0 & \text{egyébként} \end{cases} \quad (4)$$

H a harmonikus szám, ami $H(i) = \ln(i) + \gamma$ formában becsülhető, ahol γ az Euler–Mascheroni állandó, ψ pedig a teszthalmaz mérete. Ebből már megadható egy általánosított anomália érték a teljes izolációs erdőre vonatkozóan:

$$s(m, \psi) = 2^{-\frac{E(h(m))}{c(\psi)}} \quad (5)$$

ahol $E(h(m))$ az épített fák $h(m)$ értékeinek átlaga (Liu et al., 2008).

2.5. Seasonal Hybrid ESD (S-H-ESD)

Az összes vizsgált eljárás közül talán az S-H-ESD bizonyul a legkomplexebbnek. A módszer kifejezetten olyan idősorok kezelésére optimalizált, ahol nagy arányban fordulnak elő anomáliák.

Első lépésben az algoritmus beépített dekompozíciója fut le az adatokon, ami a LOESS simítóeljáráson alapuló STL módszer. Ennek köszönhetően a dekompozíció nem csak egyszerű multiplikatív vagy additív megközelítéseket tud alkalmazni a szezonális leválasztásában, hanem ezeknél komplexebb függvényeket is. A reziduális értékek meghatározása után az Extreme Studentized Deviate (ESD) teszt segítségével adhatók meg az anomáliák. Annak kiküszöbölése érdekében, hogy csak egyetlen statisztikailag szignifikáns anomália detektálása menjen végbe, a teszt az előre meghatározott k számú érték esetében adja meg a következő kritikus értéket és tesztstatisztikát:

$$C_k = \frac{\max_k |m_k - \bar{m}|}{s} \quad (6)$$

$$\lambda_k = \frac{(n-k)t_{p,n-k-1}}{\sqrt{(n-k-1+t_{p,n-k-1}^2)(n-k+1)}} \quad (7)$$

ahol m_k a megfigyelési halmaz egy kritikus eleme, n a megfigyelések száma, $t_{p,n-k-1}$ pedig a választott α szignifikancia szintnek megfelelően $1 - \frac{\alpha}{2(n-k-1)}$ -ban adja meg a t -eloszlás megfelelő értékét. A k értéke az algoritmuson belül határozódik meg, nem hiperparaméter. Ennek módja, hogy az algoritmus minden tesztlépésben meghatározza a Q teszthalmazt, valamint az Ω halmazt, ami a már kiválasztott anomália elemeket tartalmazza. Ekkor két eltérésvektort definiál a két csoportra: $\Gamma = |Q_i - \mu_Q|$ és $Y = |\Omega_i - \mu_\Omega|$. Az iteráció addig tart, míg az alábbi feltétel teljesül a vektorok szórásaira:

$$\sqrt{2}\sigma_Y \geq \sigma_\Omega \quad (8)$$

Az alternatív hipotézis ebben az esetben, hogy legfeljebb k számú anomália van az adatokban (Hochenbaum et al., 2017) (Chhikara és Feiveson, 1980).

3. Elemzés

Ebben a fejezetben bemutatom a felhasznált adatok forrását, ismertetem a trend és szezonális meghatározásának menetét, valamint minden algoritmus esetében szemléltetem a detekciós eredményeket.

3.1. Felhasznált adatok

Az elemzés során felhasznált adatok mind az [Eurostat \(2023\)](#) oldaláról származnak. A „Deaths by week, sex, 5-year age group and NUTS 3 region” nevű adatbázisban finom felbontásban szerepelnek, heti frekvenciában 35 ország NUTS 3-as régiókra lebontott adatai, 18 korcsoportot megkülönböztetve.

Ez az adatbázis rendkívül nagy mennyiségű adatot tartalmaz, így a számítási kapacitás korlátozottsága miatt a besorolhatatlan életkor kategóriába tartozó egyéneket kivettem az elemzésből. Minden modellt a teljes megfigyelt időszakon futtattam, elsősorban regionális bontásban, de a függvények a legtöbb esetben lehetőséget adnak további, életkor szerinti finomításra is.

3.2. Trend és szezonális

A következőkben összefoglalom a tanulmány során alkalmazott modellek pontos specifikációit és ezek matematikai hátterét. Minden modell alapja a hosszú távú trend és szezonális hatások meghatározása volt. Jelölje t_i és t_i^2 az adott területen megfigyelt időszak kérdéses megfigyeléséhez tartozó időindexét és annak négyzetét. Ekkor a becsült trend a következő regressziós egyenlet alapján áll elő:

$$y_i^{trend} = \alpha_i + \beta_1 t_i + \beta_2 t_i^2 + \varepsilon_i \quad (9)$$

Jelölje w_i a megfigyelt időszakban az adott megfigyeléshez tartozó hét indexét. Ekkor a szezonáliséra vonatkozó becslési egyenlet felírható egy harmonikus regressziós modell segítségével. Fourier tizennyolcadik századi kutatásai óta ismert, hogy egy diszkrét sorozatot reprodukálni lehet véges számú szinuszoid hullámforma szuperpozíciójával, egy átlagra vonatkozó korrekciós taggal kiegészítve. Kezdetben a harmonikus analízis célja az ismeretlen periodicitás megbecslése volt, aminek legfőbb eszköze a periodogram. Az utóbbi időben

azonban a figyelem ismert periódus melletti ciklusbecslésekre irányult a módszertan kapcsán (Doran és Quilkey, 1972).

A a szezonális időszorelemzésben használt modellekkel – pl. SARIMA – ebben az esetben az a probléma, hogy míg az éves szezonális havi adatokra megfelelően illeszthetőek autoregresszorok és mozgóátlagok a 12 többszöröseiben, addig a nem egészértékű mintakeresés kifejezetten hatékonytalan és nehezen kezelhető ebben a modellkörnyezetben. A korábban kifejtett Fourier-tétel alapján minden folytonos periodikus függvény felbontható az alábbi formában, ahol P a frekvencia alapján meghatározott konstans (Champeney, 1987):

$$f(x) = A_0 + \sum_{n=1}^{\infty} \left(A_n \cos \left(\frac{2\pi nx}{P} \right) + B_n \sin \left(\frac{2\pi nx}{P} \right) \right) \quad (10)$$

A halálozási adatok szezonálisát tekintve jó közelítésnek bizonyult már az $n = 1$ mellett felírható egyenlet, azaz a szezonális hatásokat a következőképpen adtam meg:

$$y_i^{sazon} = \alpha + \beta_1 \cos \left(\frac{2\pi w_i}{52,25} \right) + \beta_2 \sin \left(\frac{2\pi w_i}{52,25} \right) + \epsilon_i \quad (11)$$

Grenander és Rosenblatt (2008) kimutatta, hogy ezeknek a paramétereknek a becslése a legkisebb négyzetek módszerével (OLS) aszimptotikusan BLUE becslést eredményez, függetlenül a hibatag kovariancia struktúrájától. Ezek után meghatározható a szezonális- és trendhatásoktól megszabadított reziduum r_i tag, amire alkalmazhatók az anomália detekciós eljárások.

3.3. Detekció és következtetések

Az elemzés során felhasznált adatokat még a járvány kezdete előtti években vizsgáltam, azaz 2019 előtt, hogy az eredmények egy általánosnak tekinthető időszak alapján álljanak elő. Az első lépés a korábban meghatározott trend- és szezonhatások leválasztása volt. A detekciós eljárásokat az így kapott reziduum tagon futtattam. Ezt követően minden megfigyeléshez rendelkezésre állt egy anomália érték, ami önmagában egyfajta anomália-esélyt ad meg, azaz meghatározza, mennyire lehetünk biztosan abban, hogy adott pont eltér az adatok normál viselkedésétől. Ez alapján egyértelműen meghatározhatóvá vált, hogy a pontot az eljárás az anomáliák közé sorolta-e vagy sem. A cél ezek után annak megállapítása volt, hogy milyen jellegű anomáliák felismerésében erős adott módszertan, valamint milyen következtetések vonhatók le azzal kapcsolatban, hogy melyik hetek esetében jellemzőbbek kiugrások és hol fordulnak elő elsősorban normál értékek. Mindez lehetőséget ad arra,

hogy adott területen kialakítható legyen egy éves, heti bontásban feltüntetett profil, ami a kiugrásokra való hajlamosságot foglalja össze. A hetekre meghatározott értékek az anomália értékekből az alábbi képlet segítségével adhatók meg, ahol z a tanuló időszakban adott hét előfordulásainak száma, $anom_i$ pedig adott eljárással becsült anomália érték:

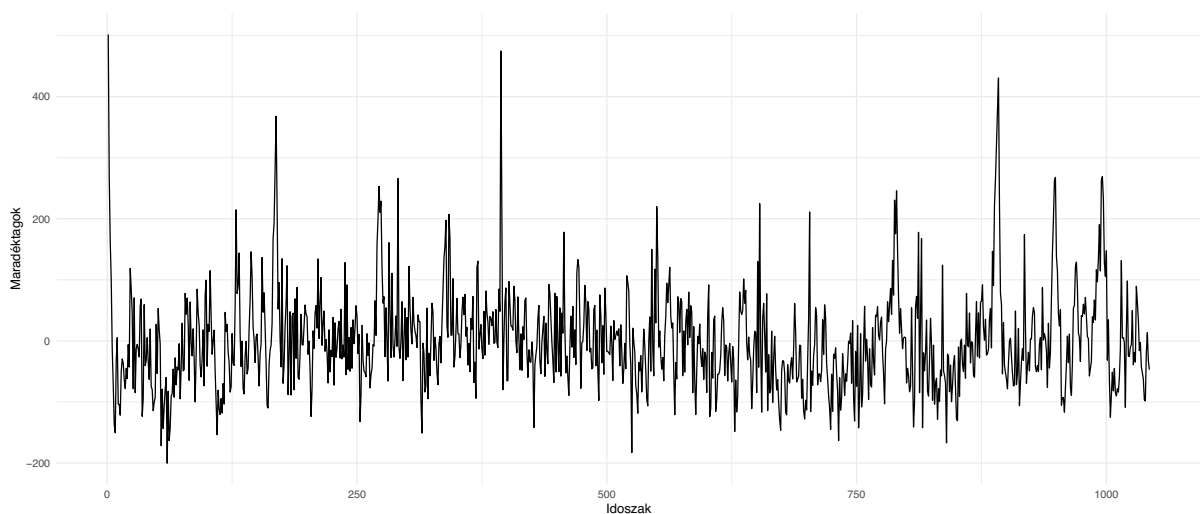
$$W_w^{anom} = \frac{1}{z} \sum_{i=1}^z anom_{w,i} \quad (12)$$

Ez a heti bontásban vett anomália-hajlamosság minden területi és életkori specifikációban jellemezni tudja, hogyan alakulnak az éven belüli ingadozások. Ezek a profilok arra is alkalmasak lehetnek, hogy a különböző detekciós eljárások hasonlóságát objektíven felmértjük.

3.4. A besorolás menete

Ebben a részben összefoglalom az anomáliaként azonosított megfigyelések kiválasztásának pontos módszerét, majd egy példán keresztül szemléltetem ennek eredményét minden eljárás esetében. A szemléltetés céljából kiválasztott terület Magyarország, ahol a férfiak halálozási számait vettem figyelembe 2000-től egészen 2019-ig. Azért tartottam ideálisnak ezt az idősort, mert megfigyelhetőek benne felső és alsó kiugró értékek is, valamint jól példázza az alapvető különbségeket a módszerekben. A detekciós eljárásokat rendre a trend- és szezonhatástól megtisztított maradéktagokon futtattam. A választott idősor esetében ezeket a reziduumokat az 1. ábra szemlélteti.

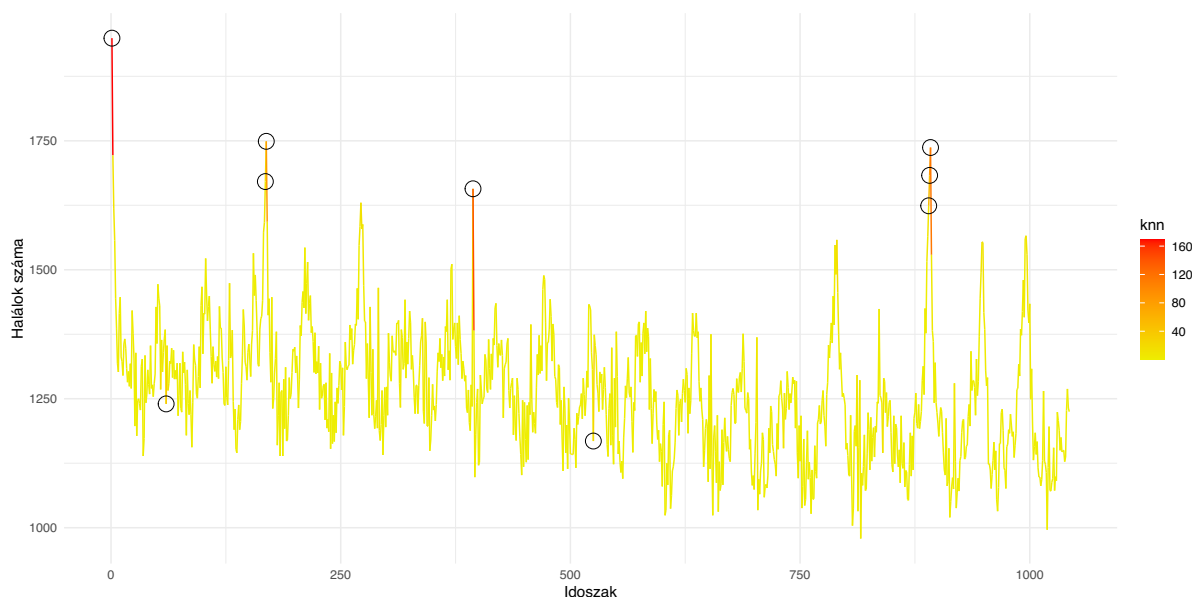
1. ábra. Maradéktagok idősora



Forrás: saját szerkesztés

Az első felírt modell a k-NN távolságokon alapuló detekciós módszer. Az elemzés során 10 szomszédot vettem figyelembe, ami intuíció alapján az időszakot követő és azt megelőző 3 hónap adatainak felelne meg. A becslésben ezeket az értékeket használom fel. A detekció során az elemző maga határozza meg azt a vágási értéket, ami fölött anomáliaként definiál egy pontot. Az elemzési helyzet és a hibás besorolás költségeinek függvényében ez a vágási pont egyszerűen módosítható. Az igazán nagy pontosságot igénylő elemzések esetében ezeket a paramétereket ajánlott lehet hangolni. A hangolás elsősorban a gépi tanulási modellekben alkalmazott módszer, melynek során bizonyos intervallumon ciklusosan megadhatjuk a becslés pontosságát tetszőleges mutató alapján, majd a legkedvezőbb eredmény melletti paramétert használjuk a továbbiakban. Ebben az esetben én a 20-as értéket választottam vágási pontnak, amit a becslt anomália értékek eloszlásának segítségével határoztam meg. A 20 olyan határt definiál, ami feletti anomália értéket csak az adatok nagyjából 1%-a kapott.

2. ábra. k-NN alapú becslés eredménye egy adott időszoron



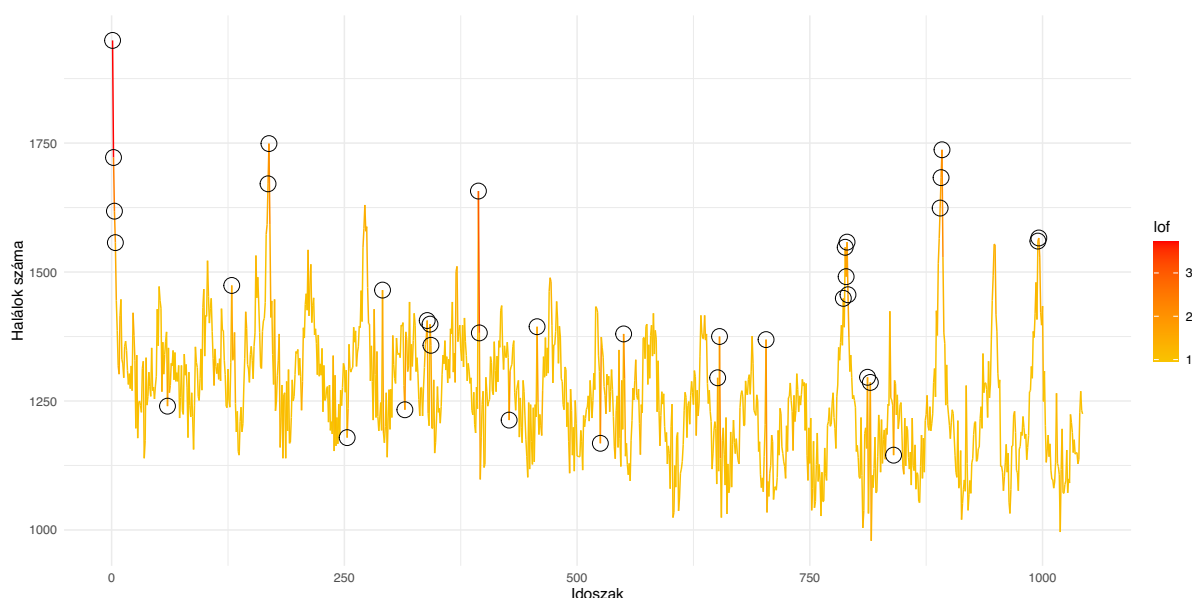
Forrás: saját szerkesztés

A 2. ábrán a k-NN alapján becslt anomália értékek és azok a pontok láthatók, amiket az algoritmus az anomália kategóriába sorolt. A k-NN értékek rendkívül széles skálán mozognak. Eloszlásuk erőteljesen jobbra elnyúló. Ennek az a következménye, hogy az extrém felső értékek rendkívül nagy jelentőséget kapnak és valamivel kisebb mellettük a relevanciája a közepesen kiugró értékeknek. Az eljárás, a választott vágási érték mellett, a pontok viszonylag kis arányát sorolta az anomáliák közé, de a legjelentősebb ingadozásokat már így is sikeresen detektálta.

Az előzőekből származtatható a LOF alapú modell, ami rendkívül hasonlóan épül fel, azzal a különbséggel, hogy itt az anomália értékek között a – szintén N távolságmátrixból kinyert – $lof(i)$ mutatók szerepelnek. A LOF alapján való besorolás egyszerűbbnek bizonyul. A legtöbb esetben az alábbi három kategóriát határozzák meg.

1. $lof(i) \sim 1$ a szomszédokhoz hasonló sűrűséget jelent.
2. $lof(i) < 1$ a szomszédokhoz képest magasabb sűrűséget jelent (belső pont).
3. $lof(i) > 1$ a szomszédokhoz képest alacsonyabb sűrűséget jelent (külső pont).

3. ábra. LOF alapú becslés eredménye egy adott idősoron



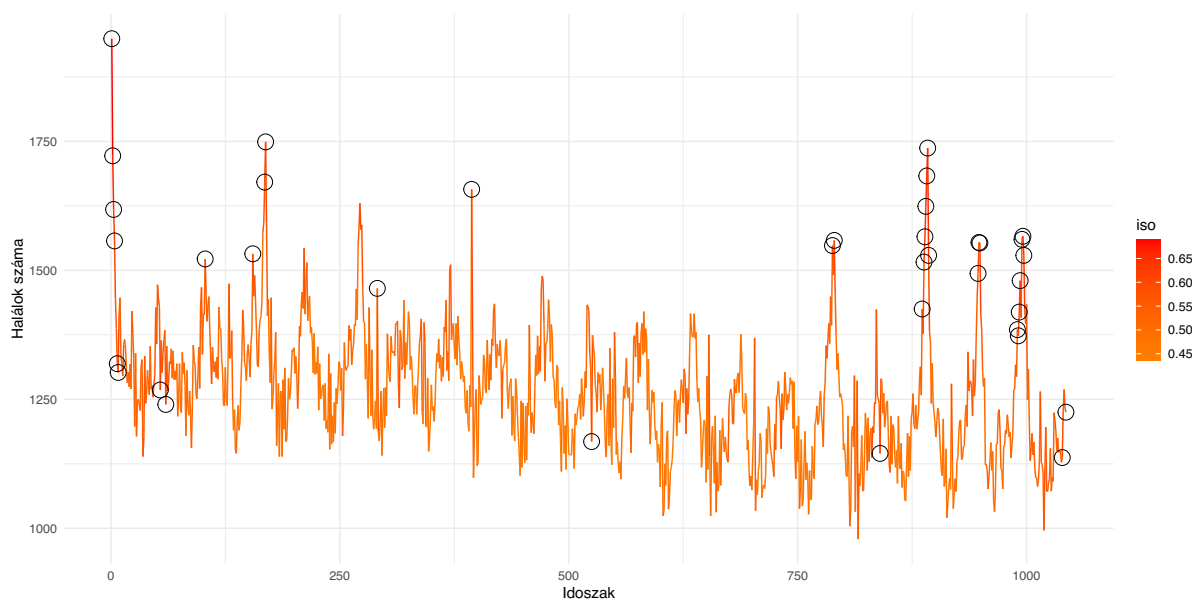
Forrás: saját szerkesztés

A 3. ábra alapján megállapítható, hogy a LOF segítségével anomáliaként meghatározott értékek jóval nagyobb arányt tesznek ki a teljes idősorban, mint a k -NN anomáliák. Az eljárás nem csak az egyértelműen kiugró értékeket azonosította, hanem a normálisnak vélt sávon belül is felismert olyan eseteket, amik valamilyen módon kilógnak a mintázatból. Ezek általában a szezonális periódusán kívül eső kisebb pozitív és negatív irányú kilengések voltak.

Az izolációs erdő anomália értékeinek számítása a (5)-ös egyenlet alapján meghatározott távolságmértékkel történik. A besorolás általános menete, hogy az anomáliákhoz minimálisan -1-es pontszámot rendelhetünk, a normál pontokhoz pedig legfeljebb 1-est. A két határ között az anomália érték annak megfelelően alakul, hogy mennyire valószínűsíthető róla, hogy outlier.

Így azt mondhatjuk, 0,5 feletti értékek esetében már nagy az esélye annak, hogy anomáliáról beszélünk.

4. ábra. Izolációs erdő alapú becslés eredménye egy adott idősoron

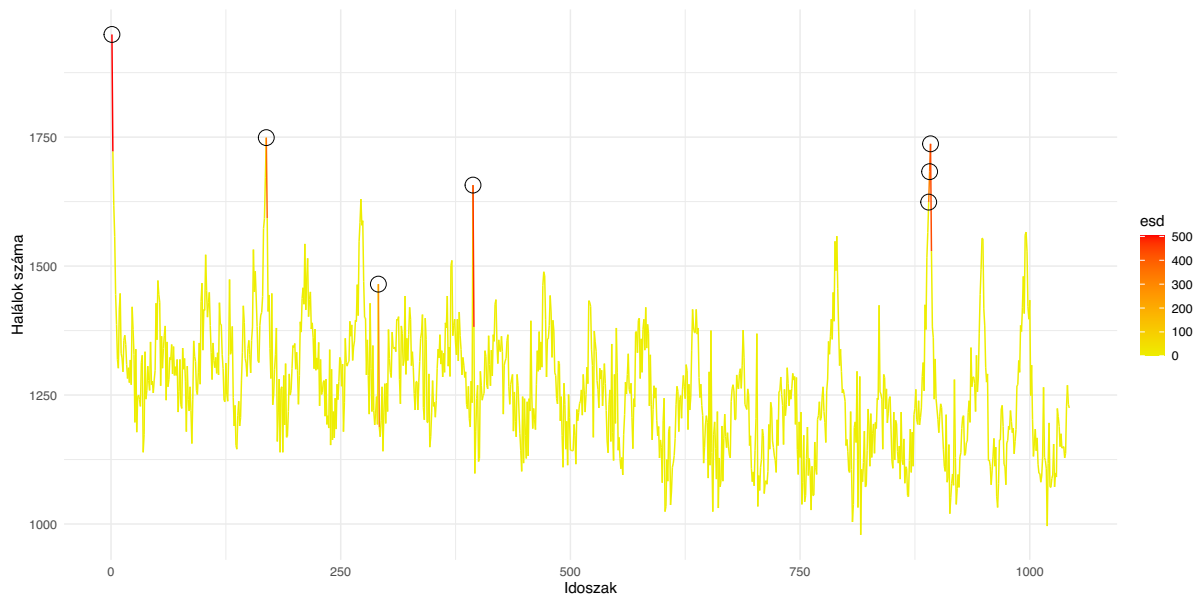


Forrás: saját szerkesztés

A 4. ábra azt mutatja, hogy az izolációs erdő sikeresen detektálta a felső anomália értékeket nem csak abban az esetben, mikor ezek drasztikusan kiugróak voltak, hanem a nagyobb anomáliákat kísérő kisebb jelentőségű kilengéseket is. Az eljárás az alacsonyabb tartományokban is relevánsnak tekinthető eseteket talált, azonban jóval kevesebbet, mint a LOF. Az elemzés során 150 izolációs fa együttes eredményét vettem figyelembe, így mivel az eredmény nem determinisztikus, a bizottság méretének növelésével ponosítható lehet a detekció. Ez azonban jeletős számítási kapacitásveszteséggel járna.

Végül az S-H-ESD alapú modell teljesítményét vizsgáltam. Ez az algoritmus a saját beépített simítóeljárásának elvégzése után detektált anomáliákat építi be a becslőfüggvénybe. Itt az elemzőnek nincs szabad keze a határvonal meghatározásában, hiszen ez a korábban kifejtett statisztikai teszt alapján történik. Az anomáliák túlzottan magas arányban történő detektálását pedig a (8)-as feltétellel gátolja a rendszer. Az itt hozzárendelt anomália értékek már a várható halálozási számtól vett eltérést mutatják minden esetben.

5. ábra. S-H-ESD alapú becslés eredménye egy adott időszoron



Forrás: saját szerkesztés

Az 5. ábrán látható eredmény igazolja, hogy ez az eljárás valóban a drasztikusan kiugró értékek detektálására optimalizált. A beépített függvény azt is külön jelzi, hogy mekkora arányban sorolta a modell a pontokat az anomália kategóriába. Az alsó és normál tartományban az algoritmus nem detektált egy pontot sem. Ez azt jelenti, hogy az eljárás azokban az esetekben optimális, mikor valóban csak a drasztikus, normál profiltól jelentősen eltérő esetekre vagyunk kíváncsiak.

4. Eredmények

A továbbiakban összefoglalom az algoritmusok összehasonlításának eredményeit. A különböző megfigyelési szempontok közé tartozott a futásidő vizsgálata 1250 idősor esetében, a felépített profilok különbözőségének szemléltetése, valamint a számítási idő változása a bemenet méretének függvényében.

4.1. Eljárások számításigénye nagy mennyiségű adat mellett

Az eljárások futásidejét első lépésben regionális és nemi bontásban vett idősorokon vizsgáltam. Ezekből összesen 1250 állt rendelkezésre. Megfigyeltem, az egyes módszereknek mennyi időre van szüksége ahhoz, hogy minden idősorra elvégezzék a detekciót és felépítsék a heti anomália-hajlamosságra vonatkozó profilt. Ennek eredményeit az 1. táblázat foglalja össze.

1. táblázat. 1250 idősorra vonatkozó számítási idő

	k-NN	LOF	ESD	IsoFor
Futásidő (perc)	1,267	1,656	4,135	52,08

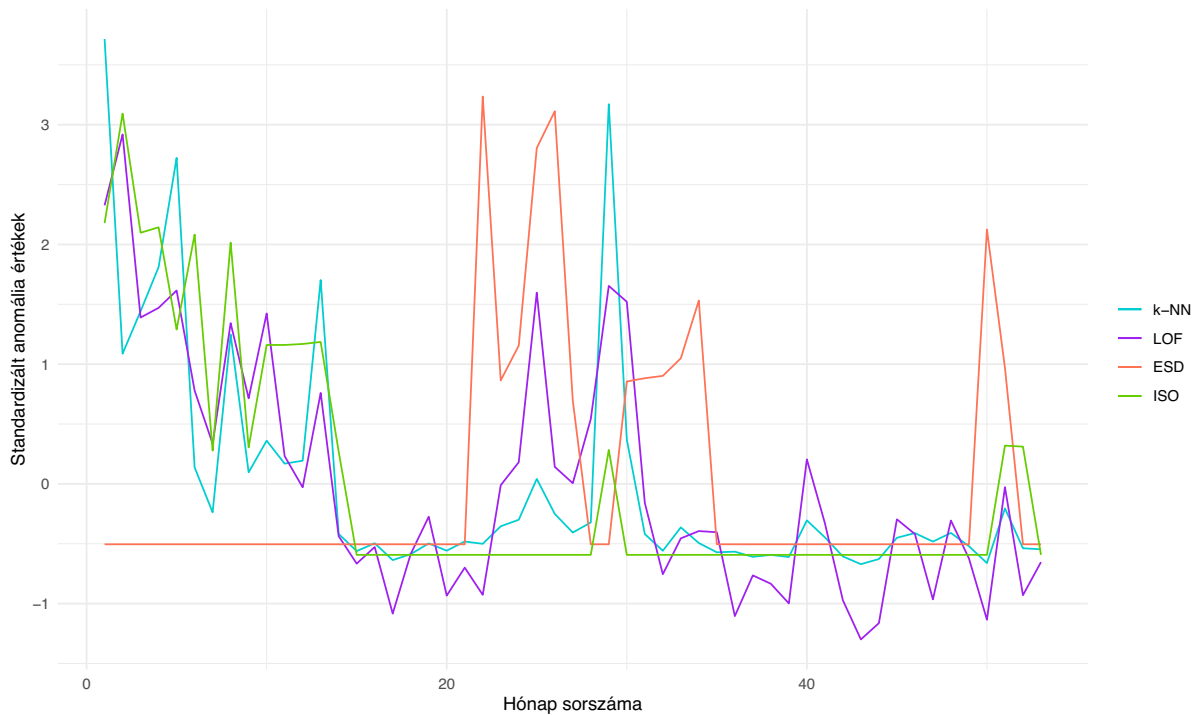
Forrás: saját szerkesztés

Látható, hogy az értékek rendkívül széles skálán mozognak. A k-NN és LOF kifejezetten gyorsnak tekinthető, hiszen nagyjából 1,5 perc szükséges ahhoz, hogy 1250 adatsorról az összes fent említett információt összefoglalják. Az ESD ezeknél valamivel több ideig, nagyjából 4 percre számol, de ez továbbra is elfogadható eredmény. Az izolációs erdő időigényessége azonban nagyságrendileg meghaladja az összes többi algoritmusét. A közel egy órás futásidő természetesen a sok fából álló, bizottsági döntési mechanizmusból fakad, de megkérdőjelezhető, hogy a magas számítási időt képes-e kompenzálni a detekciós hatékonyság.

4.2. Anomália profilok hasonlósága

Az algoritmusok összehasonlítását azzal folytattam, hogy megvizsgáltam, ugyanazoknak a jelenségeknek tulajdonítanak-e nagy súlyt a különböző eljárások. Annak érdekében, hogy ezt megállapítsam, kiszámítottam az algoritmusok által becsült anomália profilokat, majd ezeket összehasonlítottam egymással. Az eredményeket vizuálisan a 6. ábra foglalja össze.

6. ábra. A modellek által becsült heti profilok összehasonlítása



Forrás: saját szerkesztés

Az összehasonlítást úgy végeztem, hogy minden algoritmus esetében az egyes hetekhez hozzárendelt anomália értékeket összegyűjtöttem egy külön táblázatba. Ezek a változók azonban jelentősen eltérő skálákon mozogtak, így a releváns összehasonlítás érdekében standardizáltam őket. Míg az első negyedévben a k-NN, LOF és izolációs erdő értékei hasonlóan mozognak, addig az ESD itt nem vár nagy kiugrásokat. A téli anomáliákat az ESD az év eleje helyett inkább év végére csoportosítja. A másik intenzív időszak a nyári hetekhez köthető. A kora nyári kiugrásokat az ESD jelentősen felülértékeli a többi algoritmusnál. Ugyanez a helyzet a nyár végi időszakban a k-NN-nel. Az ESD még egy őszi hullámot is beépít a profilba, ami a többi eljárás esetében egyáltalán nem jelenik meg. Itt is bebizonyosodik, hogy az egyetlen algoritmus, ami az alsóbb tartományoknak is komolyabb súlyt ad, az a LOF.

Mindez a viszonylagos függetlenség hozzásegíthet egy olyan algoritmus kialakításához, ami ötvözi bizonyos eljárások erősségeit és segít hatékonyabbá tenni a csak egy megközelítés alapján detektáló módszereket. Ezzel a problémával érdemes lehet behatóbban foglalkozni, amit a kitekintés során részletesebben ki is fejtek.

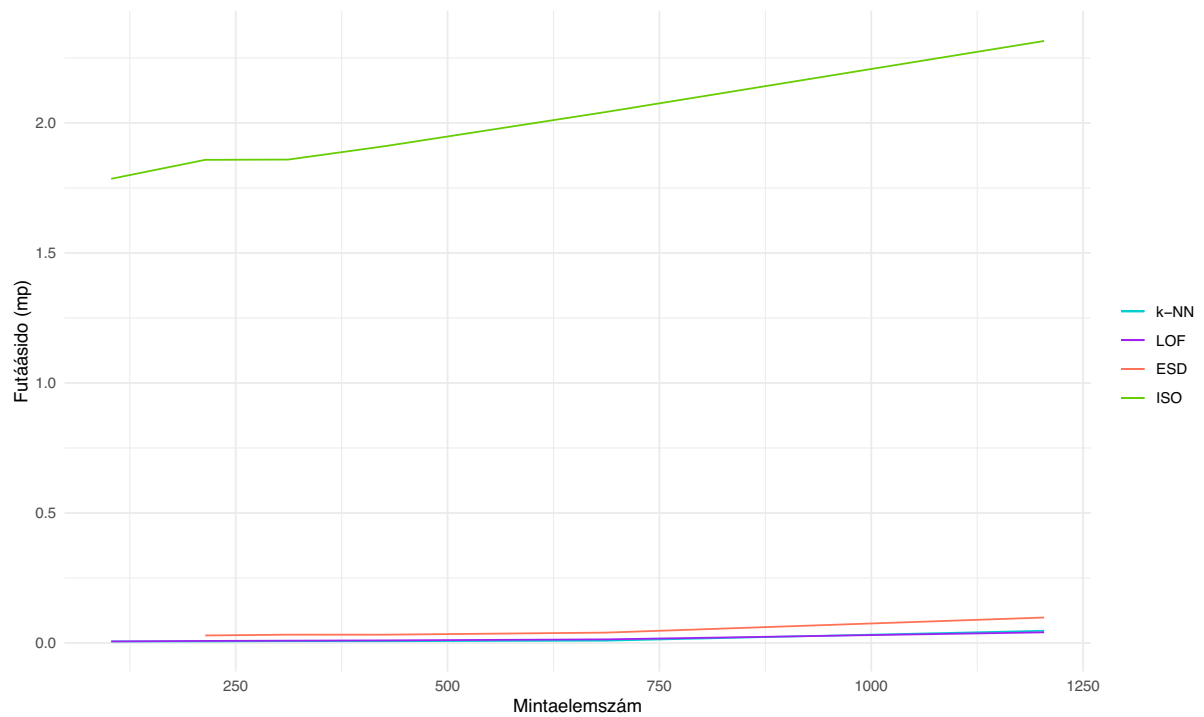
4.3. Futásidő, mint hatékonysági szempont

Az algoritmuselméletben az algoritmus naiv matematikai megfogalmazása: egy leírás, amely egyértelműen magadja, hogy adott input esetén milyen elemi lépések sorozatával jutunk el az eredményig. Az algoritmusok egyik legalapvetőbb meghatározandó tulajdonsága a nagy ordó. Az O jelölést az algoritmusok osztályozására használják aszerint, hogy milyen teljesítményt nyújtanak a bemeneti adathalmaz méretétől függően, amelyen dolgoznak. Ez jellemzően azt jelenti, hogy mind a térbeli, mind az időbeli viselkedésüket egyszerre elemezzük (Chivers és Sleightholme, 2015).

Meg kell jegyezni azonban, hogy a nagy ordó csak olyan esetekben releváns, ha az input mérete a végtelenhez közelít, azaz nagyságrendileg jóval meghaladja ennek a tanulmánynak a kereteit. Gayathri Devi et al. (2011) tanulmányukban úgy fogalmazzák, hogy ez a mérték adott függvény korlátozott viselkedését írja le, amikor az argumentum egy adott értékhez vagy a végtelen felé tendál, azaz egy aszimptotikus felső korlátként fogható fel. Azért tartottam mégis érdemének kitérni a hatékonyságnak erre a mértékére, hiszen az anomália detektálás fontossága jelentősen felértékelődik Big Data módszerek esetén. A többlethalálózás becslésének tárgyterületén is előfordulhatnak olyan esetek, mikor kifejezetten fontos, hogyan viselkedik egy adott algoritmus extrém körülmények között. A korábban leírtakból az is látszik, hogy az itt vizsgált néhány eljárás esetén is jelentős különbségek mutatkoznak a futásidőben, éppen ezért érdeklődésre adhat okot, hogy ezek pontosan miből származhatnak.

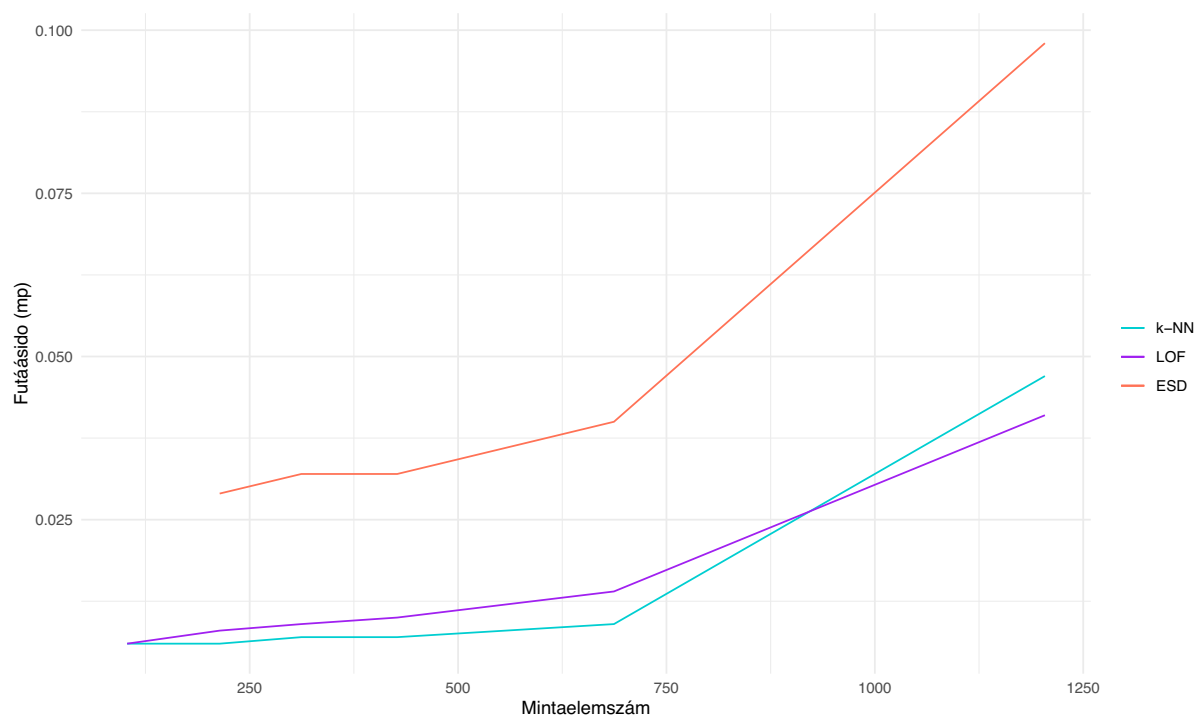
A következőkben az elérhető adatok segítségével vizsgáltam a futásidő és bemenet méretének kapcsolatát az általam vizsgált algoritmusok esetében. Ennek eredménye a 7. ábrán látható. A 8. ábrán az átláthatóság érdekében kiemeltem a három, nagyságrendileg megegyező időigényű algoritmus diagramját. Itt is bebizonyosodik a korábbi feltételezés az izolációs erdő relatív lassúságával kapcsolatban, hiszen az eredmények megbecsléséhez általánosan nagyjából negyvenszer annyi időre van szükség, mint a többi módszernél.

7. ábra. Futásidő a bemeneti adatok méretének függvényében



Forrás: saját szerkesztés

8. ábra. Futásidő a nagyságrendileg megegyező módszerekre szűkítve



Forrás: saját szerkesztés

A felvázolt kapcsolatok alapján az látszik, hogy az elvárásoknak megfelelően mindegyik

függvény monoton növekvő. Érdekes azonban, hogy az izolációs erdő esetében az a növekedés laposabbnak bizonyul, mint bármelyik másikon. A legmeredekebbnek tűnő függvény a vizsgált tartományon az ESD-hez tartozik. A k-NN és a LOF egymáshoz hasonló, kisebb mértékű növekedési tendenciákat mutat a vonaldiagramok alapján. Az ESD a legkisebb mintaelemszám mellett a kevés információból adódóan nem tudott detekciót végezni.

2. táblázat. Modellek futásideje különböző méretű inputok mellett

Mintaelemszám	k-NN	LOF	ESD	IsoFor
103	0.006	0.006		1.785
214	0.006	0.008	0.029	1.858
312	0.007	0.009	0.032	1.859
427	0.007	0.01	0.032	1.911
687	0.009	0.014	0.04	2.042
1204	0.047	0.041	0.098	2.315

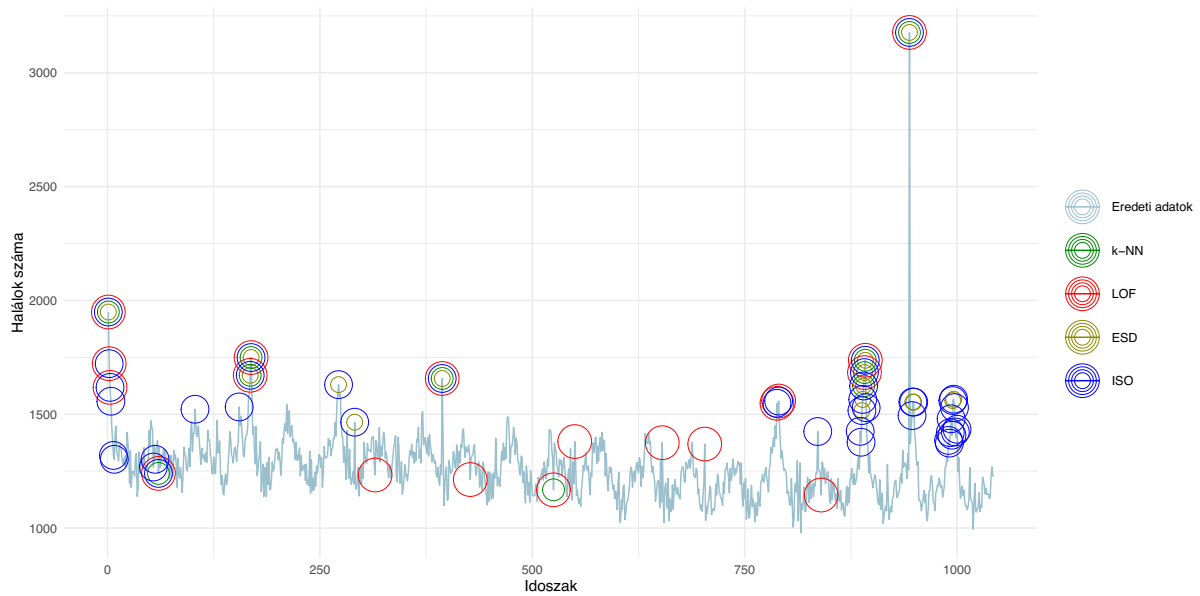
Forrás: saját szerkesztés

A 2. táblázatból azonban kiolvasható, hogy míg az izolációs erdő esetében a tartomány két végpontja közötti növekedés nagyjából 1,3-szoros volt, addig a k-NN-nél ez az érték jóval magasabb, 7,8-szoros. A növekedés az ESD esetében sem elhanyagolható, közel 4-szeres. Bár az izolációs erdő itt kedvező tulajdonságokkal rendelkezik, a mintaelemszámtól való viszonylagos függetlenség nem feltétlenül tud kompenzálni a drasztikus különbségeket, maximum extrém mennyiségű adat esetében. Az algoritmusok szélsőségek közötti viselkedésére azonban a korábban leírtak szerint nagyon nehéz ilyen kevés információ alapján bármilyen becslést adni.

4.4. Impulzussal terhelt adatok becslése

Az utolsó szempont, ami alapján összehasonlítottam a különböző módszertanokat az volt, milyen hatékonysággal képesek egy impulzussal terhelt adatsorral dolgozni. Az idősorba itt mesterségesen belevittem egy impulzust egy adott hétre vonatkozóan. A bevitt hatás drasztikus, az átlagos halálozási számot 1,5 interkvartilis terjedelemmel haladja meg. Az első vizsgálati kérdés természetesen az volt, hogy adott algoritmus felismeri-e az extrém adatot, de ezen felül arra is kíváncsi voltam, mennyiben befolyásolja ez az eredeti becslést. Az előzőekhez hasonlóan itt is a 2000-től 2019-ig vett magyarországi férfi halálozási adatokkal dolgoztam. Az eredmények a 9. ábrán láthatók.

9. ábra. Impulzussal terhelt becslés



Forrás: saját szerkesztés

Az egyes körök a jelmagyarázatnak megfelelően azt jelzik, hogy melyik pontot melyik algoritmus azonosította anomáliaként. A legtöbb jelentős kilengést minden eljárás sikeresen detektálta. Nagy mértékű különbségek lépnek fel azonban a normál tartományon belül, valamint az alsó kilengések esetében. A LOF ezekben jóval magasabb arányban detektált, mint bármelyik másik algoritmus. A besorolt pontok általában kisebb mértékű felső irányban kilógó értékek, amik a szezonálitáson kívül esnek. Az ezeknél valamivel jelentősebb pozitív irányú potenciális anomáliákat az izolációs erdő határozta meg. Ezeknél a pontoknál kifejezetten nehéz megállapítani, hogy a döntés jogosnak bizonyult-e. Az ESD és k-NN elsősorban a felső, jelentős mértékben kilógó értékeket ismerte fel, mint ahogy az eredeti esetben is. Összességében majdnem minden algoritmus fenntartotta az eredeti idősorban detektált értékeket, azaz az impulzus nem befolyásolta jelentősen a végeredményüket. Más a helyzet azonban az izolációs erdőnél, hiszen itt az impulzus bevonása azt eredményezte, hogy az algoritmus jóval kevesebb alsó anomália pontot határozott meg.

4.5. Kitekintés, alkalmazási lehetőségek

A részletes összehasonlítás során kiderült, hogy az elemzésbe bevont eljárások egészen más jellegzetességekre érzékenyek. Az S-H-ESD kifejezetten erős a felső anomáliák detektálásában, míg a LOF jóval nagyobb hangsúlyt fektet a szezonális mintázatokról vett eltérésekre, valamint

az alsó tartományokra. Ebből következik, hogy érdemes lehet kialakítani egy olyan komplex eljárást, ami ötvözi a két detekciós algoritmust, vagyis az alsó tartományokban LOF, míg a felsőbb tartományban S-H-ESD anomáliákat határoz meg. Ez azért is lehet működőképes, mert a becslési eredmények különbözőségét a nulla közeli korrelációs együttható is alátámasztja. A tanulmány kibővítéseként érdemes lehet tehát egy ilyen eljárás kialakításával próbálkozni, majd az itt meghatározott szempontok szerint összevetni a már létező detekciós módszerekkel, hogy valóban sikerülhet-e javulást elérni.

Úgy gondolom, az itt felállított, hetekre bontott anomália profilok hasznosak lehetnek komplexebb mortalitást becslő modellekben. Kifejezetten fontos lehet, hogy az elemzés során ezektől az „endogén” anomália hatásoktól tisztítottan vizsgáljuk az idősorokat, mert nagy mértékben befolyásolhatják a vizsgált hatásról alkotott képet. Mivel ezek az anomáliák általában felső kilengések, a valódi hatás mértékét feltehetően felfelé torzítják azok a modellek, amik ezekkel nem számolnak külön. A létező mortalitás predikciós eljárások általában lehetőséget nyújtanak további extra információk beépítésére, így ennek megfelelő alapot szolgáltathatnak az anomália-detekciós eljárások alapján felállított profilok.

5. Összegzés

A kutatás során négy adnomália-detekciós eljárást hasonlítottam össze különböző szempontok szerint. Az anomáliák vizsgálatának relevanciája a többlethalálozási elemzések tárgyterületén abból fakad, hogy a halálozási számokban jelentős ingadozások figyelhetők meg a nyári hőhullámok alatt, valamint a téli influenza-szezonban, amiknek bekövetkezése és mértéke szabálytalan mintázatot követ. A mortalitás elemzésére rendkívül nagy mennyiségű adat áll rendelkezésre. Ezek az algoritmusok lehetővé teszik, hogy az adattömegből viszonylag rövid idő alatt releváns információt nyerjünk ki.

A tanulmányban az összehasonlításba bevont algoritmusok a következők voltak. A legközelebbi szomszédoktól számított távolságon alapú eljárások közül a k-NN és LOF detekció, a bináris keresőfák módszerén alapuló izolációs erdő, valamint a saját simítóeljárást és ESD teszteket alkalmazó S-H-ESD. A kutatás során megállapítottam, hogy ezeknek az algoritmusoknak a különböző típusú anomáliákra való érzékenysége és számításigénye is rendkívül heterogén. A k-NN és S-H-ESD kifejezetten a magasabb tartományokban bekövetkező, drasztikus kilengéseket észlelik leghatékonyabban. A kisebb kiugrásokat és szezonális mintázattól eltérő jelenségeket az izolációs erdő és LOF észleli magas arányban. Az alsóbb tartományokban szintén ezek az eljárások bizonyultak érzékenyebbnek, a LOF azonban jóval nagyobb számban detektál itt értékeket. A becsült anomáliák alapján felállított hetekre lebontott éves profilok ennek megfelelően szintén jelentősen eltértek egymástól, azonban összességében mindegyiknél megjelent a nyári és téli időszakok anomáira való hajlamossága. A futásidő vizsgálata során megállapítottam, hogy míg a k-NN, LOF és S-H-ESD nagyságrendileg egy kategóriába sorolható, az izolációs erdő lefutása jóval több időt vesz igénybe. A leggyorsabb algoritmusnak a k-NN bizonyult.

Hivatkozások

- Acosta, R. J. és Irizarry, R. A. (2022). A flexible statistical framework for estimating excess mortality. *Epidemiology*, 33(3):346–353.
- Beaney, T., Clarke, J. M., Jain, V., Golestaneh, A. K., Lyons, G., Salman, D., és Majeed, A. (2020). Excess mortality: the gold standard in measuring the impact of covid-19 worldwide? *Journal of the Royal Society of Medicine*, 113(9):329–334. PMID: 32910871.
- Champeney, D. C. (1987). *Fourier series*, pages 155–169. Cambridge University Press.
- Chhikara, R. S. és Feiveson, A. L. (1980). Extended critical values of extreme studentized deviate test statistics for detecting multiple outliers. *Communications in Statistics - Simulation and Computation*, 9(2):155–166.
- Chivers, I. és Sleightholme, J. (2015). *An Introduction to Algorithms and the Big O Notation*, pages 359–364. Springer International Publishing, Cham.
- Cormen, T., Leiserson, C., Press, M., Rivest, R., Stein, C., és Company, M.-H. P. (2001). *Introduction To Algorithms*. Introduction to Algorithms. MIT Press.
- Doran, H. E. és Quilkey, J. J. (1972). Harmonic analysis of seasonal data: Some important properties. *American Journal of Agricultural Economics*, 54(4):646–651.
- Eurostat (2023). Deaths by week, sex, 5-year age group and nuts 3 region. [Adatbázis. Letöltve: 2023-04-01]
https://ec.europa.eu/eurostat/databrowser/view/demo_r_mweek3/default/table?lang=en.
- Gayathri Devi, S., Selvam, K., és Rajagopalan, S. P. (2011). An abstract to calculate big o factors of time and space complexity of machine code. In *International Conference on Sustainable Energy and Intelligent Systems (SEISCON 2011)*, pages 844–847.
- Grenander, U. és Rosenblatt, M. (2008). *Statistical Analysis of Stationary Time Series*. AMS Chelsea Publishing Series. American Mathematical Society.
- Hastie, T. és Tibshirani, R. (1987). Generalized additive models: Some applications. *Journal of the American Statistical Association*, 82(398):371–386.

- Hochenbaum, J., Vallis, O. S., és Kejariwal, A. (2017). Automatic anomaly detection in the cloud via statistical learning.
- Kirpich, A., Shishkin, A., Weppelmann, T. A., Tchernov, A. P., Skums, P., és Gankin, Y. (2022). Excess mortality in belarus during the covid-19 pandemic as the case study of a country with limited non-pharmaceutical interventions and limited reporting. *Scientific Reports*, 12(1):5475.
- Liu, F. T., Ting, K. M., és Zhou, Z.-H. (2008). Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422.
- Liu, F. T., Ting, K. M., és Zhou, Z.-H. (2012). Isolation-based anomaly detection. *ACM Trans. Knowl. Discov. Data*, 6(1).
- Páldy, A. és Bobvos, J. (2013). Többlethalalózás Európában 2020.10. és 2021.18. hét között az euromomo hálózat alapján = excess mortality anomalies in europe in the period of week 10 of 2020 and week 18 of 2021 based on the euromomo network. *EGÉSZSÉGTUDOMÁNY : A MAGYAR HIGIÉNIKUSOK TÁRSASÁGA TUDOMÁNYOS ÉS TOVÁBBKÉPZŐ FOLYÓIRATA = HEALTH SCIENCE : SCIENTIFIC AND POSTGRADUATE JOURNAL OF HUNGARIAN SOCIETY OF HYGIENE*, 65(2):4–18.
- Pearce, N., Lawlor, D. A., és Brickley, E. B. (2020). Comparisons between countries are essential for the control of COVID-19. *International Journal of Epidemiology*, 49(4):1059–1062.
- Shaukat, K., Alam, T. M., Luo, S., Shabbir, S., Hameed, I. A., Li, J., Abbas, S. K., és Javed, U. (2021). A review of time-series anomaly detection techniques: A step to future perspectives. In Arai, K., editor, *Advances in Information and Communication*, pages 865–877, Cham. Springer International Publishing.
- Taylor, S. J. és Letham, B. (2018). Forecasting at scale. *The American Statistician*, 72(1):37–45.
- Teng, M. (2010). Anomaly detection on time series. In *2010 IEEE International Conference on Progress in Informatics and Computing*, volume 1, pages 603–608.
- World Health Organization (2020). *International Guidelines for Certification and Classification (Coding) of COVID-19 as Cause of Death Based on ICD International Statistical Classification of Diseases*.