DIMENZIÓK
Matematikai Közlemények
XII. kötet, 2024 43–55

doi:10.20312/dim.2024.05

Bizonytalanság, avagy óvatosan a statisztikával[1]

Csanády Viktória
Soproni Egyetem, Faipari Mérnöki és Kreatívipari Kar, Alaptudományi Intézet
csanady.viktoria@uni-sopron.hu, 0009-0004-3461-4892

ÖSSZEFOGLALÓ. A statisztikai vizsgálatok alkalmazása során különböző instabil adatsorok esetén óvatosan kell eljárni, különösen akkor, ha ismeretlenek számunkra az egyes befolyásoló tényezők szerepei, és hatásainak nagysága. Időbeli folyamatok előrejelzése gyakran téves, íjesztő eredményeket produkálhat. Gyakori tünet azonban napjainkban, hogy egy adathalmazt közlő, nem kerülheti ki a statisztikai kiértékelést, ami szinte elvárás. Az alábbiak erre vonatkozóan mutatnak be egy példát, lehetőségeket elemezve egy instabil időbeli adatsor esetén, nagyméretű konklúziók levonása nélkül az előre jelzésre vonatkozóan, figyelmeztetve annak lehetséges hibájára.

ABSTRACT. In the case of various unstable data sets, we must proceed with caution when applying statistical tests, especially if the roles of individual influencing factors and the magnitude of their effects are unknown to us. Forecasting temporal processes can often produce wrong, frightening results. However, it is a common symptom these days that someone who communicates a set of data cannot avoid statistical evaluation, which is almost an expectation. The following presents an example of this, analyzing possibilities in the case of an unstable time series, without drawing large-scale conclusions about the forecast, warning of its possible error.

1. Bevezetés

A különböző meteorológiai adatok, illetve azok vizsgálata régóta foglalkoztatja a kutatókat. A történelmi Magyarország területén 1753-tól regisztráltak adatokat a Nagyszombati Érseki Egyetemen. A szomszédos Ausztria területén a Kremsmünsterben felépített Matematika Torony csillagvizsgálója büszkélkedhet Európa egyik leghosszabb idejű adatsorával, ami meghaladja a 255 évet. A hosszútávú idősorok fontossága napjainkban a klímaváltozás időszakában jelentősen megnövekedett. A klímaváltozás, és vegyük most csupán annak egyetlen elemét, a hőmérsékletet kihat az emberi életre és persze a teljes gazdaságra, így annak vizsgálata exponált feladat. Sajnos részletes adatsorokhoz hozzájutni általában nehézkes és költséges vállalkozás, azok számára, akiknek nem ez a tudományterülete. Így ha mégis igényünk lenne ilyen jellegű adathalmazra, elfogadjuk a kevésbé részletes letölthető állományt. Az alábbiakban nem kívánjuk, és nem is tudjuk a fent vázolt extrém problémát kielemezni, tegyék ezt az arra kitanított szakemberek. Így csupán egy hőmérsékleti adatsoron elmélkedünk, ennek során bemutatjuk az alkalmazható statisztikai vizsgálatok lehetőségeit és annak eredményeit. Ezen utóbbiak vizsgálata viszont tanulságos figyelmeztető példákkal támasztják alá azt a fontos tényt, hogy könnyen téves következtetések levonása történhet eredményekből,

[1] ENGLISH TITLE: Uncertainty, or be careful with statistics.
KULCSSZAVAK: staisztika, klímaváltozás, hőmérséklet.
KEYWORDS: statistics, climate change, temperature.

ha azokat csak, mint tisztán statisztikai eredményeket kezeljük, figyelmen kívül hagyva a tényleges folyamat lehetséges kimenetelét.

A vizsgálat tárgya egy 64 évet átölelő adatsor, éves átlaghőmérsékletre vonatkozóan. A mérések helye Kismarton (Eisenstadt), Ausztria, Soprontól légvonalban 17,5 km. Az adathalmaz a https://www.meteoblue.com internetes oldal szabadon hozzáférhető adatbankjából származik, ahonnét ingyenesen letölthető.

Az alábbiakban először bemutatásra kerül az adatsor grafikus prezentációja, annak teljes számszerű közlésétől eltekintünk. Ezt követően a 63 év során előforduló értékek normalitását vizsgáljuk függetlenül a feltételezett időbeli változástól. A normalitás vizsgálatot követi az időbeli folyamat tendenciájának leírása. Az adatsor jelentős mértékű szórása miatt indokolt az adatkiegyenlítés, hármasátlagolás majd centírozás. A centírozott adatsoron különböző modellek illesztése kerül bemutatásra. Ezt követően az adatsor 10 éves blokkokra bontással kerül vizsgálat alá melynek során a 10 éves átlagok felhasználásával illetve a hozzájuk tartozó konfidenciahatárral kimutatjuk a szélsőséges értékeket. Egy további kísérletben az idő intervallumonként bekövetkező hőmérséklet növekedés valamint az időintervallumban előforduló melegrekord értékek előfordulásának lehetséges kapcsolata kerül terítékre.

A vizsgálat céljai:

1. Mutassuk ki, hogy 63 év esetén az éves átlaghőmérsékleti adatok halmaza normális eloszlást követ!
2. Az időbeli folyamatot jellemezzük arra alkalmas matematikai modellel!

Az alkalmazott regressziós modellek:

- Lineáris függvény

- matematikai alakja:

$$y = b_1 \cdot x + b_0$$

- számítógépes alak:

$$Var2 = b1 \cdot Var1 + b0.$$

- Additív lineáris-trigonometrikus függvény

- matematikai alakja:

$$y = b_3 \cdot \sin(b_2 \cdot x) + b_1 \cdot x + b_0$$

- számítógépes alak:

$$Var2 = b3 \cdot \sin(b2 \cdot Var1) + b1 \cdot Var1 + b0.$$

- Transzformált exponenciális függvény

- matematikai alakja:

$$y = b_3 \cdot {b_2}^{x-b_1} + b_0$$

- számítógépes alak:

$$Var2 = b3 \cdot b2\text{^}(Var1 - b1) + b0.$$

- Telítési függvény

- matematikai alakja:

$$y = b_3 \cdot \left(1 - e^{-(b_2 \cdot x)^{b_1}}\right) + b_0$$

- számítógépes alak:

$$Var2 = b3 \cdot \left(1 - exp(-1 \cdot (b_2 \cdot Var1)\text{^}b_1)\right) + b0.$$

- Transzformált tangens hiperbolikusz függvény

- matematikai alakja:

$$y = b_3 \cdot th\left(b_2 \cdot (x - b_1)\right)b_0$$

- számítógépes alak:

$$Var2 = b3 \cdot TanH\left(b2 \cdot (Var1 - b_1)\right) + b0.$$

3. Adjuk meg a folyamatban előforduló szélsőséges éveket!
4. Igazoljuk az összefüggést az átlaghőmérséklet növekedése és szélsőséges melegrekordok gyakorisága között!

Az adathalmaz vizsgálata során alkalmazott software a STATISTICA, a modellek illesztésénél a szükséges kezdőértékek az adatsorból jól becsülhetők.

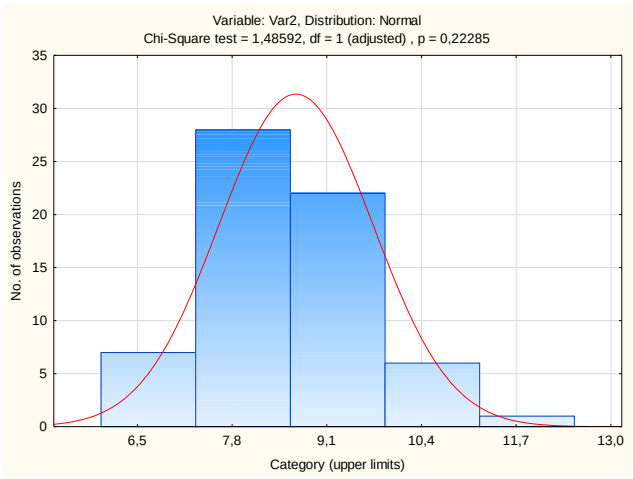
2. Számított eredmények, kiértékelés

2.1. 64 év hőmérsékleti átlagának normalitás vizsgálata

Az alábbi táblázatban a normalitás vizsgálat eredményei olvashatók, továbbá a hozzátartozó hisztogramm.

Upper Boundary	Variable: Var2, Distribution: Normal (átlaghő) Chi-Square = 1,48592, df = 1 (adjusted) , p = 0,22285				
	Observed Frequency	Cumulative Observed	Expected Frequency	Cumulative Expected	Observed-Expected
<= 7,30000	7	7	6,19277	6,19277	0,80723
8,60000	28	35	23,96242	30,15520	4,03758
9,90000	22	57	25,91365	56,06885	-3,91365
11,20000	6	63	7,38239	63,45123	-1,38239
< Infinity	1	64	0,54877	64,00000	0,45123

1. táblázat. Normalitás vizsgálati eredmények

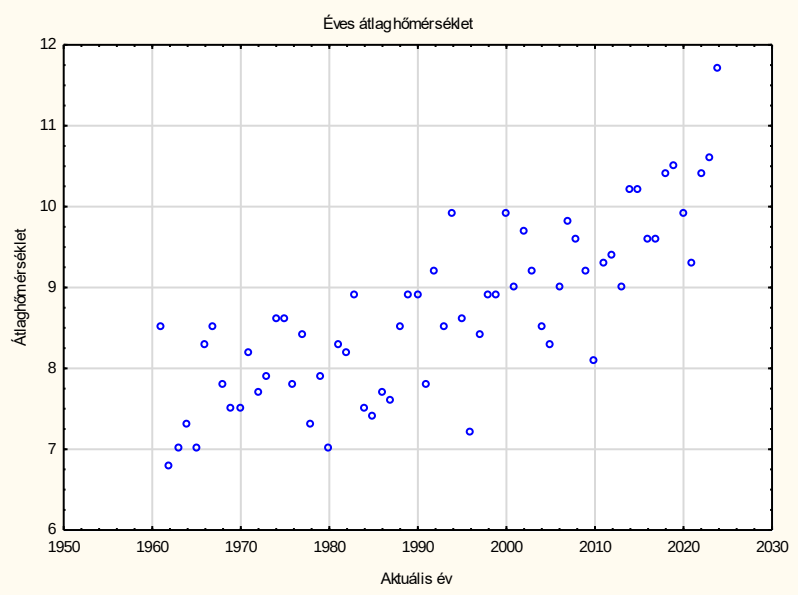


1. ábra. Normalitás vizsgálati hisztogram

Az előbbi számítások kimutatják, hogy az adathalmaz 5%-os tévedési szinten megfelel a feltételezett normális eloszlásnak. A hisztogram jól demonstrálja az egyes osztályok gyakoriságait, ami szerint a maximális gyakoriság a második osztályra jellemző és itt mutatkozik a legnagyobb eltérés az elméleti és tapasztalati gyakoriság között.

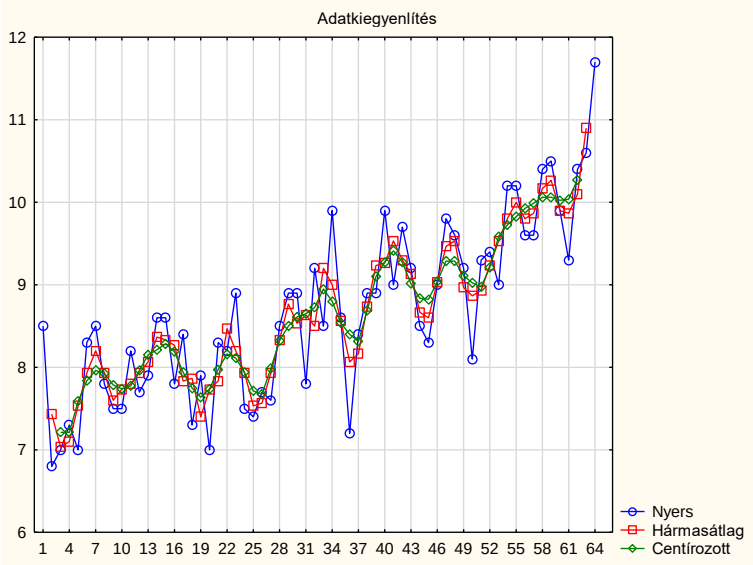
2.2. Az átlag adatsor regressziós vizsgálata

Az alábbi ábrán jól látható a nyers adatsor szóródása, ami arra a gondolatra vezeti a statisztikai kiértékelőt, hogy regresszió esetén a modell kiválasztása meglehetősen önkényes, a kiértékelő döntése lehet. Ez természetesen nem helytálló, hiszen a modellnek eleget kell tennie bizonyos folyamatjellemzőknek. Ebben az esetben azonban nagy a bizonytalanság, talán azt mondhatjuk ki, hogy a felső korlát indokolt, hiszen a hőmérsékleti növekedés nem mehet a végtelenbe, ez ugyancsak ijesztő lenne. A teljes folyamat leírása azonban nem várható el, sőt előrejelző trend sem számítandó a modell alapján a vizsgálati időszak rövidsége (~64 év) miatt.
Ami lehetséges, az a vizsgált időszak jellemzése egy arra alkalmas modell segítségével.
Mivel azonban az adathalmaz szóródása nagy, így érdemes elvégezni az adatkiegyenlítést melynek révén csillapítható az adatsor.



2. ábra. Nyers adatsor

A 3. ábra együttesen mutatja a nyers és a két csillapított adatsort, az áttekinthetőség érdekében az egymást követő évek átlaghőmérséklet adatait összekötve, így jobban elkülöníthetők az adatsorok. A csillapított adatsorokat vizsgálva a csillapítás mértéke már a hármasátlagolt adatsornál is jelentős, még markánsabb azonban a centírozás esetén. Az adatsorok kapcsán meg kell említeni, hogy a vizsgálati időintervallum 64 évként van feltüntetve, viszont az utolsó év csonka, ez azonban nem befolyásolja jelentős mértékben a számított eredményeket.

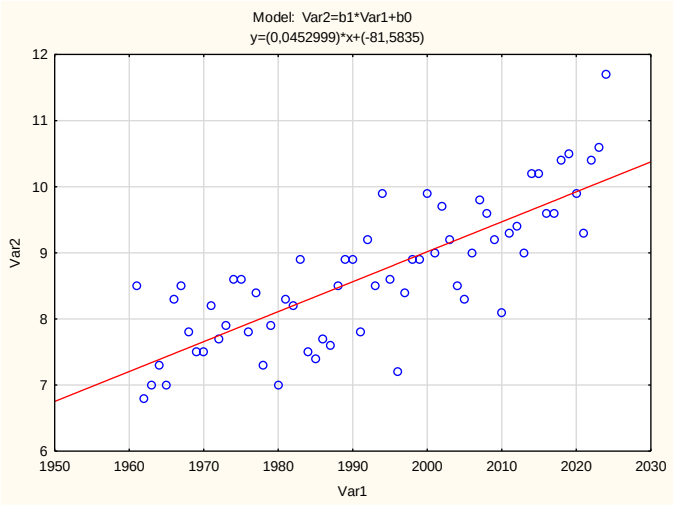


3. ábra. Nyers és csillapított adatsorok

Az adatelőkészítést követően először a legegyszerűbb lineáris modell illesztése kerül bemutatásra, mindhárom adathalmaz esetén.

- A nyers adatsor (Var2) eredménye:

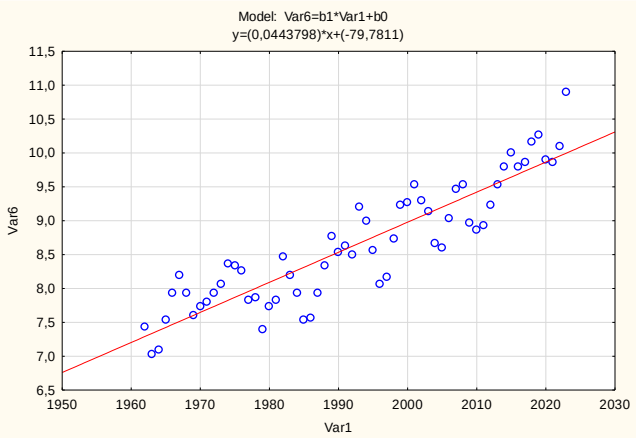
	Model: Var2=b1*Var1+b0 (átlaghő) Dep. var: Var2 Loss: (OBS-PRED)**2 Final loss: 25,797379350 R= ,79666 Variance explained: 63,467%	
N=64	b1	b0
Estimate	0,045300	-81,5835



4. ábra. Lineáris illesztés nyers adatsor esetén

- A hármasátlagolt adatsor (Var6) eredménye:

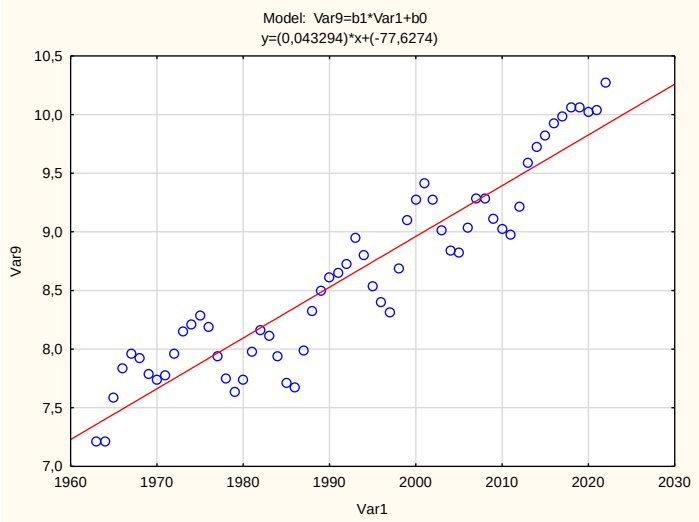
	Model: Var6=b1*Var1+b0 (átlaghő) Dep. var: Var6 Loss: (OBS-PRED)**2 Final loss: 9,080436767 R= ,90087 Variance explained: 81,156%	
N=62	b1	b0
Estimate	0,044380	-79,7811



5. ábra. Lineáris illesztés hármasátlagolt adatsorra

- A centírozott adatsor (Var9) eredménye:

	Model: Var9=b1*Var1+b0 (átlaghő) Dep. var: Var9 Loss: (OBS-PRED)**2 Final loss: 4,930583229 R= ,93406 Variance explained: 87,246%	
N=60	b1	b0
Estimate	0,043294	-77,6274

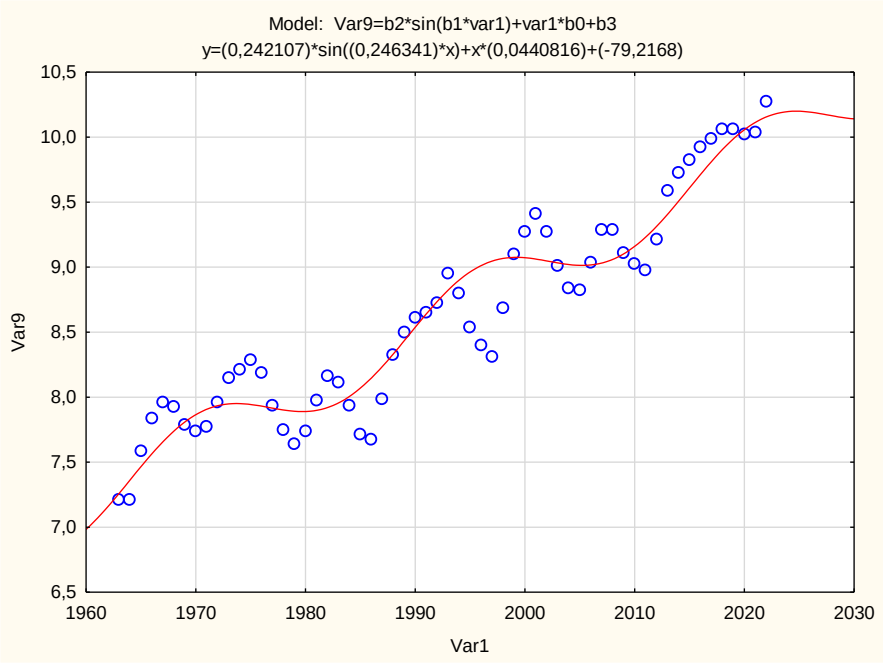


6. ábra. Lineáris illesztés centírozott adatsorra

Áttekintve a lineáris illesztések eredményeit arra a következtetésre jutunk, hogy a nyers adatsor vizsgálata az adatsor nagy szóródása miatt nem kedvező. A korrelációs együttható értéke ezt alátámasztja R=0,7967, R=0,9009, R=0,9341 nyers, hármasátlagolt, centírozott sorrendjében. Az R értékének növekedése azonban igazolja a csillapítás sikerességét, így a tendencia vizsgálatához elegendő a centírozott adatsor további vizsgálata, melynek során nem lineáris modellek illesztését hajtjuk végre.

- Additív lineáris-trigonometrikus függvény illesztésének eredménye:

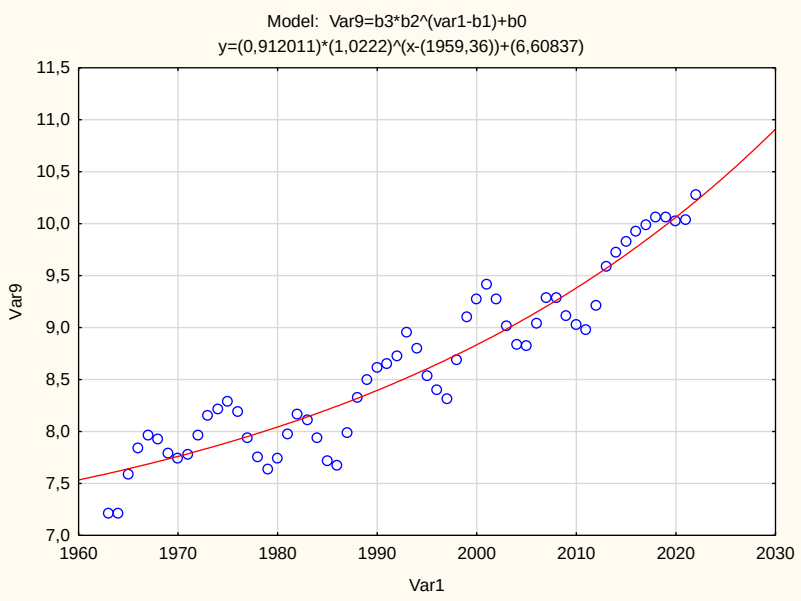
	Model: Var9=b2*sin(b1*var1)+var1*b0+b3 (átlaghő) Dep. var: Var9 Loss: (OBS-PRED)**2 Final loss: 3,214587391 R= ,95752 Variance explained: 91,685%			
N=60	b2	b1	b0	b3
Estimate	0,242107	0,246341	0,044082	-79,2168



7. ábra. Additív lineáris-trigonometrikus függvény illesztése

- Transzformált exponenciális függvény illesztésének eredménye:

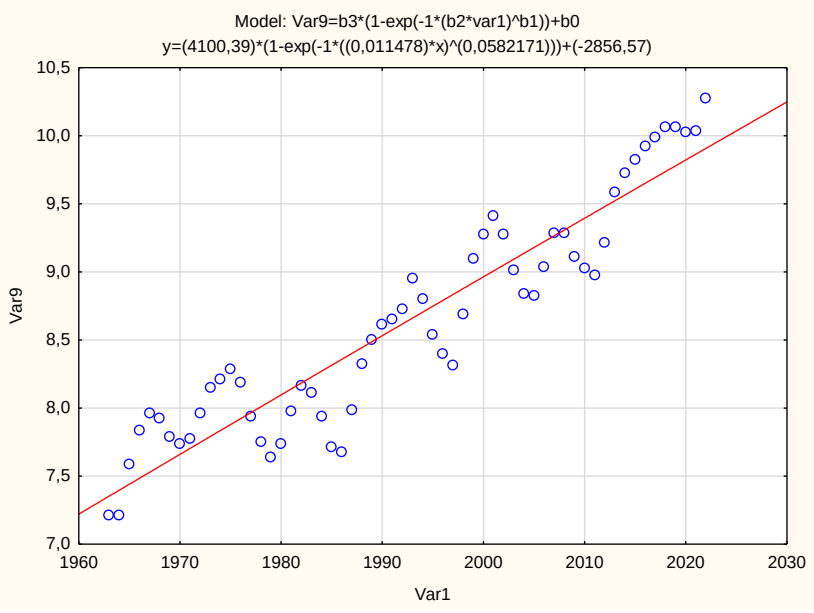
	Model: Var9=b3*b2^(var1-b1)+b0 (átlaghő) Dep. var: Var9 Loss: (OBS-PRED)**2 Final loss: 3,991091622 R= ,94698 Variance explained: 89,676%			
N=60	b3	b2	b1	b0
Estimate	0,912011	1,022198	1959,359	6,608373



8. ábra. Transzformált exponenciális függvény illesztése

- Telítési függvény illesztésének eredménye:

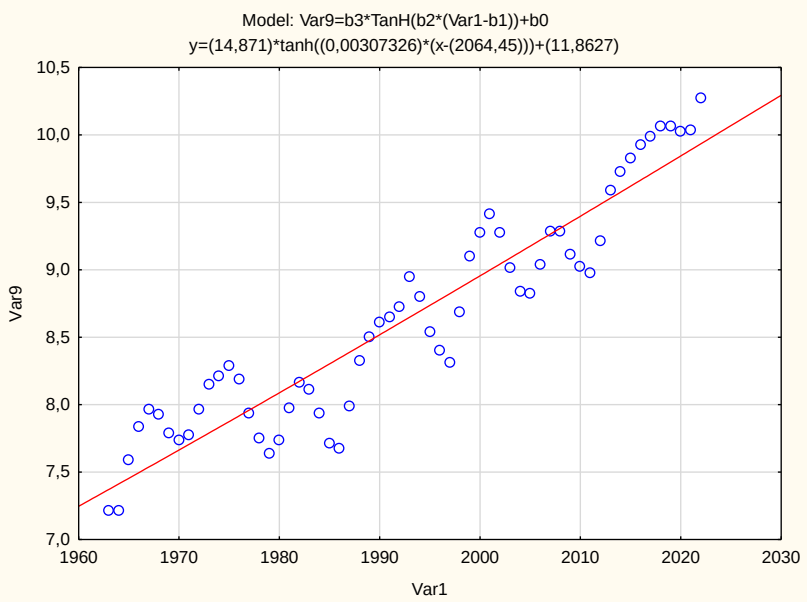
	Model: Var9=b3*(1-exp(-1*(b2*var1)^b1))+b0 (átlaghő) Dep. var: Var9 Loss: (OBS-PRED)**2 Final loss: 4,974480359 R= ,93345 Variance explained: 87,133%			
N=60	b3	b2	b1	b0
Estimate	4100,388	0,011478	0,058217	-2856,57



9. ábra. Telítési függvény illesztése

- Transzformált tangens hiperbolikusz függvény illesztésének eredménye:

	Model: Var9=b3*TanH(b2*(Var1-b1))+b0 (átlaghő) Dep. var: Var9 Loss: (OBS-PRED)**2 Final loss: 4,821994562 R= ,93556 Variance explained: 87,527%			
N=60	b3	b2	b1	b0
Estimate	14,87104	0,003073	2064,450	11,86269



10. ábra. Transzformált tangens hiperbolikusz függvény illesztése

Az illesztések illetve alkalmazott modellek esetén számított R értékek alkalmasak egy fajta rangsorolásra. Meg kell azonban jegyezni, hogy R mint korrelációs együttható nem tévesztendő össze a lineáris korrelációs együtthatóval, ami nyilván csak lineáris modellek esetén alkalmazható. Az alábbi táblázat összefoglalja a nem lineáris modellek R értékeit.

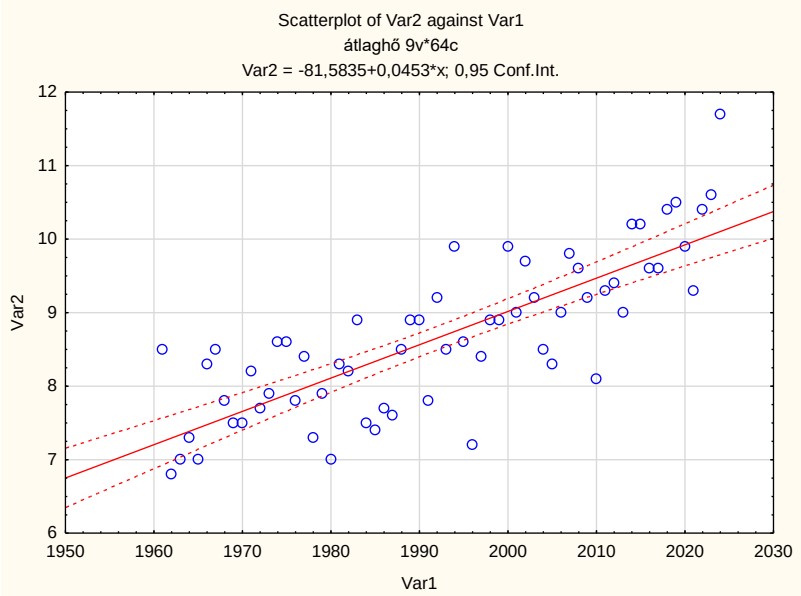
Illesztett modell típusok	*R*
Additív lineáris-trigonometrikus	0,9575
Exponenciális	0,9470
Telítési	0,9335
Transzformált tangens hiperbolikusz	0,9356

2. táblázat. Illesztési R értékek

A táblázati értékek egyértelműen arra utalnak, hogy a folyamatot legjobban leíró modell az additív lineáris-trigonometrikus modell. Ezt követi az exponenciális, majd az utóbbi kettő között az eltérés meglehetősen csekély. Meg kell azonban nyomatékosan jegyezni, hogy az első kettő nem korlátos modell, a vizsgált időintervallumra nézve azonban az illeszkedés statisztikai szempontból megfelelő.

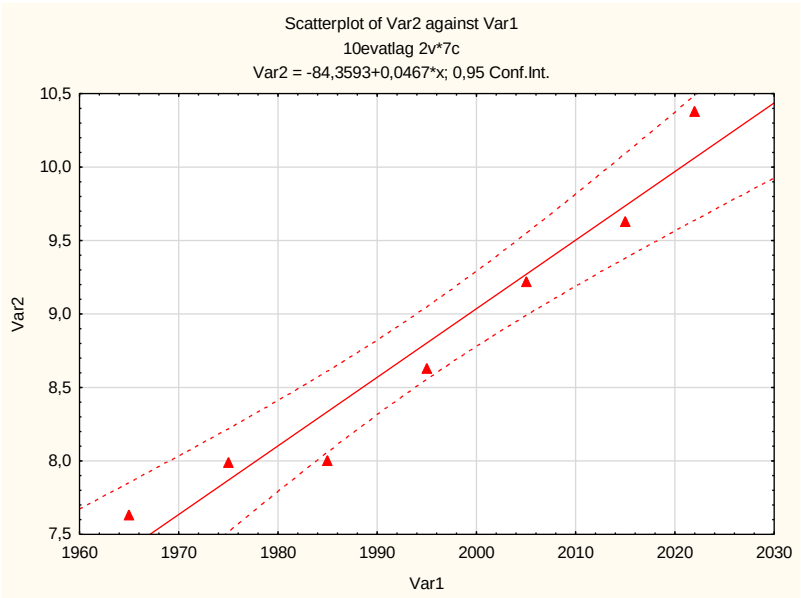
2.3. Szélsőséges éves átlaghőmérsékletek kimutatása

A szélsőséges adatok kimutatása egy bizonytalan, nagy szórással rendelkező adathalmaz esetén, ami időfüggő természeti folyamatból származik nehézkes feladat. Az alkalmazott módszer során az időintervallumra feltételezzük a linearitást és megadjuk a 95%-os konfidencia határokat.



11. ábra. Konfidencia határok a nyers adathalmaz esetén

Jól látható, hogy a módszer a nyers adathalmazra nem alkalmazható. A következőkben az adathalmazt tíz éves részintervallumokra bontjuk és az adatok tíz éves átlagaival kíséreljük meg a vizsgálatot.



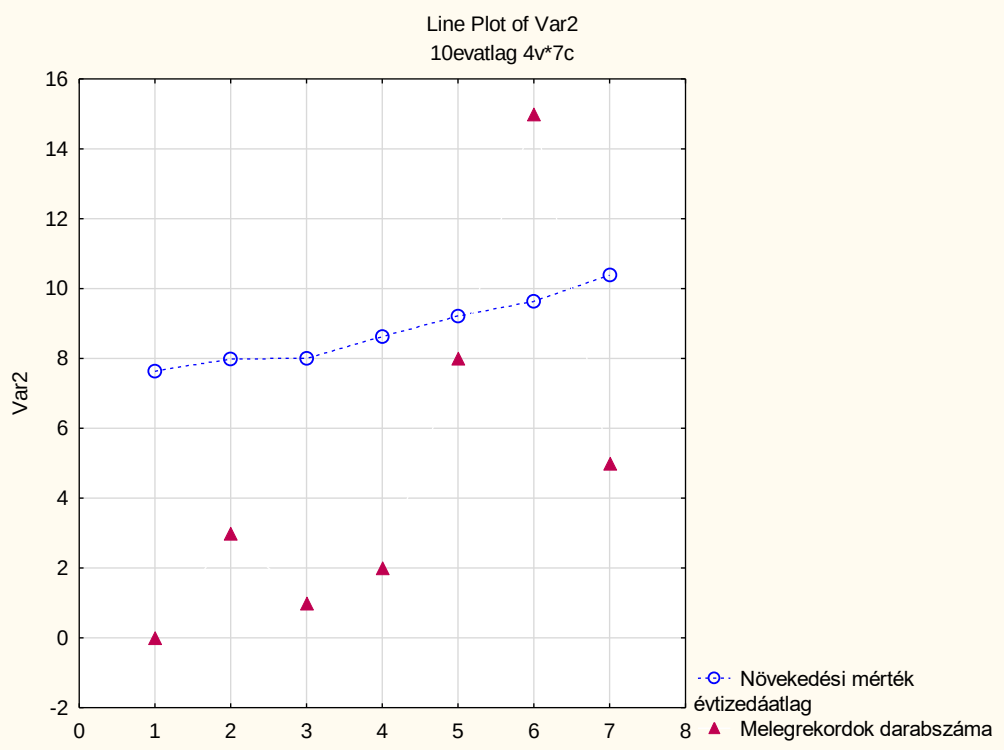
12. ábra. Konfidencia határok tízéves bontás esetén

A 12. ábra arra utal, hogy a vizsgált 10 éves részintervallumok (utolsó csonka 10 év) átlagértékei közül az 1980-1990 évtized hőmérséklet átlaga az alsó 95 %-s konfidenciahatár alá esik, a feltételezett tendenciának nem felel meg.

2.4. Évtizedek átlaghőmérsékletváltozásai és a melegrekordok előfordulásának számai

Az adatok részletes felsorolása nélkül - hiszen ezek szabadon hozzáférhetők az interneten – jegyezzük meg, hogy a vizsgált 64 évben az Európában mért melegrekordok száma 2024. július végéig 34 volt. Az abszolút maximum értéket, 49,1 C° –t 2021.07.20-án mérték Törökország európai területén. A Magyarországon mért eddigi legmagasabb hőmérséklet 41,9 C° Kiskunhalas 2007.07.20. mely értékkel a középmezőnyben vagyunk a 22. helyen. Természetesen a melegrekord nagyságrendje helyfüggő, de ne gondoljuk, hogy ez a tény feltétlenül szignifikáns. Ezt támasztja alá néhány példa, Ukrajna 42 C° 2010, Belgium 41,8 C° 2019 vagy éppen Svájc 41,5 C° 2003.

A földrajzi hely tehát nem feltétlenül mérvadó, a rekordok számának növekedése azonban azt a feltételezést kelti, hogy az emelkedő átlaghőmérséklettel szinifikáns. Feltételezve, hogy a birtokunkban lévő átlaghőmérsékleti adatsor növekedési tendenciája egy közepes tendenciának feltételezhető meghatároztuk a tízéves ciklusok átlaghőmérsékleteinek növekedési mértékeit, valamint utána kerestünk az egyes időszakokban bekövetkező európai hőmérsékleti rekordok számának, és feltételezzük a két változó szignifikáns kapcsolatát. Az alábbi ábrán az évtizedekhez kapcsolódó növekedési mértékek illetve az adott évtizedben előforduló melegrekordok db száma látható.



13. ábra. Átlaghőmérséklet növekedési mértéke és rekord számok

A 13. ábra nem igényel különösebb magyarázatot, arra vonatkozólag viszont, hogy a feltételezés a szignifikáns kapcsolatra elfogadható-e azt a választ adja, hogy nem.

Természetesen igaz az a tény is, hogy a vizsgált időintervallum rövid, emiatt az eredmények megbízhatósága megkérdőjelezhető.

2.5. Elemzés, értékelés

64 év átlaghőmérsékleteinek normalitás vizsgálati értékelése:
Az alkalmazott illeszkedés vizsgálattal kimutatható, hogy a vizsgált adatsor 5%-os tévedési szinten megfelel a nullhipotézisben feltételezett normális eloszlásnak. A tapasztalati gyakoriság maximuma a 2. osztályba esik, majd ezt követi kevés eltéréssel a középső 3. osztály. A tapasztalati és elméleti gyakoriságok különbsége mind két osztály esetén azonos abszolútértékben közelítőleg 4, a második osztályban (O-E) pozitív, a harmadikban negatív. A vizsgálat bár statisztikailag ugyan korrekt, belőle azonban messzemenő következtetéseket nem lehet levonni, annál is inkább mivel időfüggő adatsort vizsgáltunk!

Regressziós eredmények értékelése:
A vizsgálat első lépésében egy közönséges lineáris regressziót alkalmaztunk a nyers adatsorra, ami kimutatta, hogy ilyen formában csak a tendencia jóslása a rövid időintervallumra gyenge korrelációt mutat. Ez indokolta a bizonytalan nagy szórású adatsor csillapítását az adatkiegyenlítési módszerek alkalmazását. A már csillapított adatsor lineáris regressziós eredményei lényegesen szorosabb korrelációt mutattak. Mivel a linearitás nem feltételezhető, így további nem lineáris modellek alkalmazása következett. A négy felhasznált modell közül a legjobb eredményt a korrelációs együttható vonatkozásában a nem szokványos additív modell eredményezte, ami lineáris és trigonometrikus függvény kompozíciója. A modell természetesen csak erre a rövid vizsgált időtartamra vonatkoztatható, megjegyezve, hogy statisztikailag jó illeszkedést prezentál, de a tényleges folyamatra csak véletlenszerű a jó eredmény! Az ezt követő exponenciális modell alkalmasnak bizonyul a folyamat ezen időszakának tendenciális leírására, de nyilván hosszútávú elemzésre nem alkalmas. A telítési függvény és tangens hiperbolikusz függvény korlátos függvények, aszimptotikus jelleggel, így alkalmazásuk indokolt lehet a folyamat jellemzésére. Itt azonban ki kell jelenteni, hogy mindkét függvény esetében a vizsgált időintervallum adatai a függvények azon szakaszába esnek ahol a függvények grafikonjai jól láthatóan közel lineárisak.
A négy illesztés közül a fentiek figyelembe vételével kimondható, hogy a 64 éves vizsgálati időszak gyenge exponencialitást mutat, amit igazolnak az illesztett függvény paraméterei.

Szélsőséges átlagértékek kiszűrésének értékelése:
A nyers adathalmaz esetén a szélsőséges átlagok kiszűrése a felhasznált módszerrel nem járható. Az évtizedekre bontott adathalmaz esetén már kimutatható egy a konfidencia intervallumból kieső adat, azonban ne felejtsük el, hogy az adatok többszörös átlagolása révén a kapott eredmény elnagyolt. Ebben az esetben lehet, hogy alkalmasabb rövidebb időintervallumok alkalmazása, vagy adatkiegyenlítés használata a vizsgálat előtt.

Átlaghőmérséklet növekedése és rekordszámok összefüggésének értékelése:
A vizsgálat szignifikáns kapcsolatot keresett az évtizedes átlaghőmérséklet növekedése és az egyes évtizedekben előforduló melegrekordok darabszáma között. Két változó kapcsolatának vizsgálatára többféle módszer alkalmazható. Itt azonban csupán egy grafikus elemzés került bemutatásra. A grafikus eredményből feltételezhető a kezdeti állítás hamissága, miszerint a két vizsgált tényező között nincs szignifikáns kapcsolat.

3. Összefoglaló

A címből „Bizonytalanság, avagy óvatosan a statisztikával” kiolvasható, hogy a vizsgálat tárgya egy olyan adathalmaz, jelen esetben időfüggő adatsor, mely szélsőséges, jelentősen ingadozó, így vizsgálata, kielemzése nehézkes. Mivel hőmérsékleti átlagértékekről van szó, melyek ráadásul éves átlagok, így takarják a mögöttük rejlő további szélsőséges adatokat. A vizsgált időintervallum rövid, hiszen nincs egy század, de arra alkalmas, hogy jellemezhető legyen a vizsgált időszak trendje. A regressziós modell megválasztása ilyen esetben az adatsor külleme alapján történik, kijelentve azt, hogy csupán az adott intervallum jellegére utal, további előre jelzésekre nem alkalmas.

Az eredmények áttekintése után érdemes megjegyezni, hogy bár egy adatsorra szorosan illeszkedő matematikai modell is megalkotható, lásd additív modell - kevesebb adat esetén például akár elérhetjük, azt hogy egy alkalmazott polinom áthaladjon az egyes pontokon - statisztikailag tökéletes az illesztés, de a folyamat szempontjából értelmezhetetlen! Szem előtt kell tehát tartani a tényleges folyamat lehetséges alakulását, és eszerint választani a megfelelő modellt.

Szélsőséges adatok kiszűrése nehézkes feladat, adatkiegyenlítéssel és többszörös átlagolással viszont jelentősen csillapítjuk az adatsort, aminek következtében az torzul, ez azt eredményezheti, hogy a kapott szélsőséges érték valójában nem helytálló.

Mindezen tények arra utalnak, hogy a fentiekben vizsgált adatsorhoz hasonló adatsorok esetén óvatosan kell alkalmaznunk az egyes statisztikai módszereket, s bár lehet az eredmények statisztikai szempontból helyesek, szem előtt kell tartani a vizsgálat tárgyát figyelembe véve az arra vonatkozó szakmai ismereteket.

Irodalomjegyzék

[1] **Csanády, V., Horváth-Szováti, E., Szalay, L.**, Alkalmazott statisztika, Sopron, Nyugat-Magyarországi Egyetem Kiadó (2013), 175p.
[2] URL https://www.meteoblue.com
[3] **Csanády, V.,** Időjárás elemzés regressziós eljárás alkalmazásával, Dimenziók, Matematikai Közlemények (2064-2172): 3 pp 25-34 (2015)