Matematikai Közlemények
III. kötet, 2015

doi:10.20312/dim.2015.01

Adatelemzés mozgó statisztikákkal

Kalmár János
MTA CSFK GGI
kalmar@ggki.hu

ÖSSZEFOGLALÓ. A tanulmányban mozgó statisztikák segítségével határozom meg geofizikai és geodéziai idősorok 'eseményeit'. Ezek manuálisan, diagram-elemzéssel is megtalálhatók lennének, de az idősorok (többszázezres) hossza miatt indokolt az automatizált keresés megoldása.

ABSTRACT. In this study the „events" in geophysical and geodetic time series are identified by moving statistics. Although these events can be identified manually, the length of time series (several hundred thousands) justifies the automated search solution.

1. Bevezetés

Intézetünkben a közelmúltban merült fel két olyan probléma – nevezetesen geomágneses Q-kitörések illetve dőlésmérő méréshatár elmozdulások helyének meghatározása –, melyek kapcsán általánosítható eljárást fejlesztettem ki egydimenziós adatsorbeli „események" kimutatására (mintaillesztésre).

Idősorok valószínűségi eloszlását szokás jellemezni statisztikával, ami alatt az idősorból képlettel levezetett skalárt értünk. Ilyen nevezetes statisztika pl. a várható érték, szórás, regresszió, ferdeség vagy lapultság, két adatsor összevetésekor pedig a kovariancia és a konvolució.

Az adatsorokban található események nem feltétlen mutathatók ki a teljes adatsorra alkalmazott statisztikával [5], hanem általában csak egy részintervallumon alkalmazva mutatnak az átlagostól eltérő viselkedést (anomáliát), ezért mintaillesztéskor célszerűbb részintervallumokon számolt mozgó statisztikákat alkalmazni.

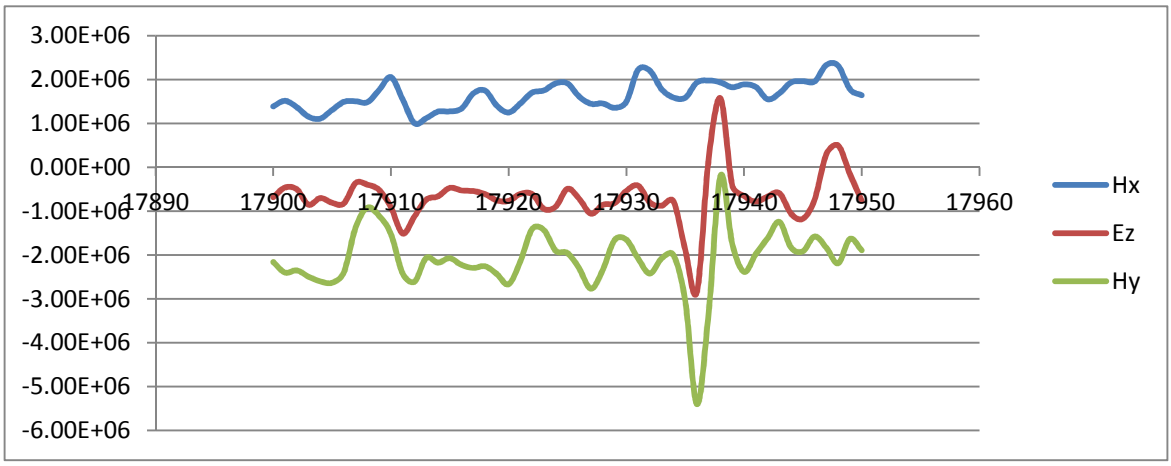
Tanulmányomban megmutatom, konkrét mintaillesztéskor hogyan célszerű kiválasztani a legjobb statisztikát és annak paramétereit, a mozgó intervallum hosszát és indikátor tartományát.

2. Q-kitörések geomágneses és geoelektromos adatsorokban

Schumann-rezonanciáknak [1] nevezzük a bolygófelszín és az ionoszféra által határolt gömbréteg elektromágneses sajátfrekvenciáit, amit a zivatartevékenység során keletkező villámok keltenek. A jelenség robusztus becslést ad a Föld troposzférájában lejátszódó globális időjárási folyamatokról a világ zivatartevékenységének idő- és térbeli változásán keresztül, valamint a Föld−ionoszféra üregrezonátor felső határoló régióját (ionoszferikus D-tartomány) érő extra-terresztrikus hatásokról.

A Schumann-rezonanciákban mutatkozó tranziens kitörések (1. ábra) kapcsolatban állnak a magas-légköri fényjelenségekkel (angol rövidítéssel: TLE). 1995-ben Boccippio és mások [2] megmutatták, hogy a leggyakoribb magas-légköri fényjelenség, a *vörös lidérc* pozitív töltésű, felhő-föld típusú villámlás során keletkezik. Ugyanekkor a Schumann-rezonanciák sávjában Q-kitörés jelentkezik. Más megfigyelések [3] rámutattak arra, hogy a vörös lidércek előfordulása és a Q-kitörések összefüggenek, így a Schumann-rezonanciákból nyerhető adatok felhasználhatók a lidércek globális előfordulásának becslésére [4].

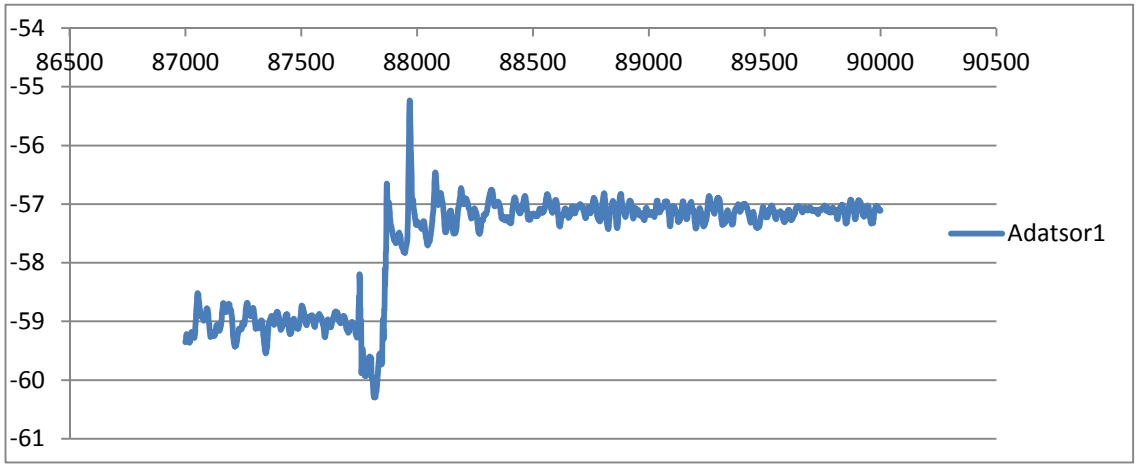
A Schumann-rezonancia mérése két geomágneses (H_x és H_y), és egy (E_z) geoelektromos csatornán történik. A Q-kitörés E_z-ben mindig jelentkezik és H_x, illetve H_y közül legalább egyben.



1. ábra. Q-kitörés esemény17936-nál

3. Méréshatár elmozdulás dőlésmérőnél

Geodéta kollégáim több módszerrel is igyekeznek kimutatni a Föld természetes felszíne vagy az épített műtárgyak lokális mozgását, deformációját. Egyik műszerük a dőlésmérő, mely a mérés síkjában megfigyelhető dőlést detektálja. A használt műszer sajátossága, hogy nagyobb kilengések a méréshatár elmozdulását okozhatják, amire a 2. ábrán látunk példát. Az adatsor feldolgozása előtt ezen lépcsőket korrigálni kell, hogy az elemzés ne vezessen téves következtetéshez.



2. ábra. Méréshatár elmozdulás esemény 87862-nél

4. A felhasznált statisztikai jelzőszámok és tulajdonságaik

Jelölje $\mathbf{E}[X]$ az X valószínűségi változó várható értékét, akkor X változó k-adrendű $\mathbf{M}_k$ centrális momentuma

$$\mathbf{M}_k = \mathbf{E}[(X - \mathbf{E}(X))^k] .$$

A *szórás* ($\mathbf{D}[X]$) négyzete (amit *varianciának* is neveznek) tulajdonképpen a másodrendű centrális momentum, azaz

$$\mathbf{D}[X]^2 = \mathbf{E}[(X - \mathbf{E}(X))^2] = \mathbf{M}_2.$$

A szórást az átlagtól való átlagos eltérésnek is tekinthetjük.

Az X valószínűségi változó *ferdesége* vagy *ferdeségi együtthatója* lényegében azt mutatja meg, mennyire szimmetrikus a valószínűségi változó eloszlása. Képlete:

$$\beta_1 = \mathbf{M}_3/\mathbf{M}_2^{3/2}.$$

Ha X sűrűségfüggvénye szimmetrikus (mint pl. a haranggörbe), akkor $\beta_1 = 0$, ha 'jobbra húz el', akkor $\beta_1 > 0$, ha 'balra húz el', akkor $\beta_1 < 0$, természetesen a deformációval arányosan.

Az X valószínűségi változó *lapultság*a, vagy *lapultsági mutatója* (másként *csúcsossága*, vagy *csúcsossági együtthatója*) lényegében a normális eloszlás sűrűségfüggvényéhez viszonyítja az X valószínűségi változó sűrűségfüggvényét. Képlete:

$$\beta_2 = \mathbf{M}_4/\mathbf{M}_2^2 - 3.$$

Normális eloszlás esetén $\beta_2 = 0$. A normális eloszlás haranggörbe sűrűségfüggvényénél 'csúcsosabb' (meredekebb) sűrűségfüggvényű eloszlások esetén $\beta_2 > 0$. A haranggörbénél 'laposabb' sűrűségfüggvényű eloszlások esetén $\beta_2 < 0$.

A konvolució két függvényhez egy harmadikat rendel hozzá. Generáló képlete:

$$(f * g)(t) = \int_{-\infty}^{\infty} f(\tau) g(t - \tau) d\tau .$$

Ha *f[i]* és *g[i]* ($i = 0, \ldots, k$) diszkrét idősorok (mérési értékek), akkor a konvolució képlete:

$$(f * g)[n] = \sum_{m=0}^{n} f[m] g[n - m] .$$

A generált idősor egy eleme felfogható úgy, hogy az az egyik idősor elemeinek súlyozott összege, ahol a súlyok fordított sorrendben lettek kiválasztva a második idősorból.

A lineáris regresszió modellje azt feltételezi, hogy a független $X = (x_1, x_2, \ldots, x_n)^T$ vektor és a függő y (skalár) változó között lineáris összefüggés van. Az

$$y(\mathbf{ß}, X) = ß_0 + \sum_{k=1}^{n} ß_k \cdot x_k$$

m elemű (X^i, y^i) i=1, …, $m \geq n$+1 mérési adatsor esetén a **ß** együtthatóvektor meghatározható pl. a legkisebb négyzetek elve alapján.

A $dy^i(\mathbf{ß}, X^i) = y^i - y(\mathbf{ß}, X^i)$, i=1, …, *m hibavektor* alapján ellenőrizhető a linearitás (és az illeszkedés) megléte.

A lineáris regresszió $y(\mathbf{ß}, X) = ß_0 + \sum_{k=1}^{n} ß_k \cdot x_k$ képlete y becslésére is felhasználható.

5. A mozgó statisztika lényege és alkalmazása

Egy statisztika egy képlet segítségével rendel hozzá egy (akár többdimenziós) adatsorhoz egy skalárt, ami az adatsor statisztikai jellemzője lesz.

‘Az ördög a részletekben rejlik’, azaz az adatsorokban rejtőző események nem feltétlen mutathatók ki a teljes adatsorra alkalmazott statisztikákkal, hanem általában csak egy részintervallumon mutatnak a szokásostól eltérő viselkedést (anomáliát).

A mozgó statisztika a teljes helyett csak egy rögzített hosszú részintervallumon számol, de ezt a részintervallumot végigcsúsztatja a teljes adatsoron, az így számolt statisztikák tehát pozíciókhoz rendelhetők, és maguk is egy adatsort alkotnak.

Ha jó statisztikát választottunk, akkor csak a keresett esemény környezetében lesz ‘találata’ a mozgó statisztikának, vagyis a (ritka) eseményt a mozgó statisztika viszonylag ritkán előforduló értékei jelzik.

Ha a számított mozgó statisztikának (mint adatsornak) előállítjuk a sűrűség függvényét (hisztogramját), akkor onnan leolvashatók azon ‘eseménygyanús’, kis gyakoriságú (indikátor) intervallumok, melyek általában a sűrűségfüggvény két szélén (a farkaknál) találhatók.

Ahhoz, hogy egy mozgó statisztikát és a hozzá tartozó indikátor intervallumot a keresett esemény kimutatására elfogadjunk, kézzel ellenőrizzük a kapott találatokat, minek alapján

- elfogadjuk a statisztikát és az indikátor intervallumot,
- módosítjuk az indikátor intervallumot,
- elvetjük a statisztikát.

Mivel egy esemény több (egymást részben átfedő) részintervallumon is látható, ezért pontos pozicionálása további (kézi vagy automatikus) elemzést igényel.

6. A Q-kitörések kimutatására használt statisztikák és indikátor intervallumaik

Konvolució: a három (Hx, Hy, Ez) adatsor közül legalább kettő (de Ez mindenképp) tartalmazza az eseményt, ezért együttes vizsgálatuk indokoltnak tűnt (3. ábra).

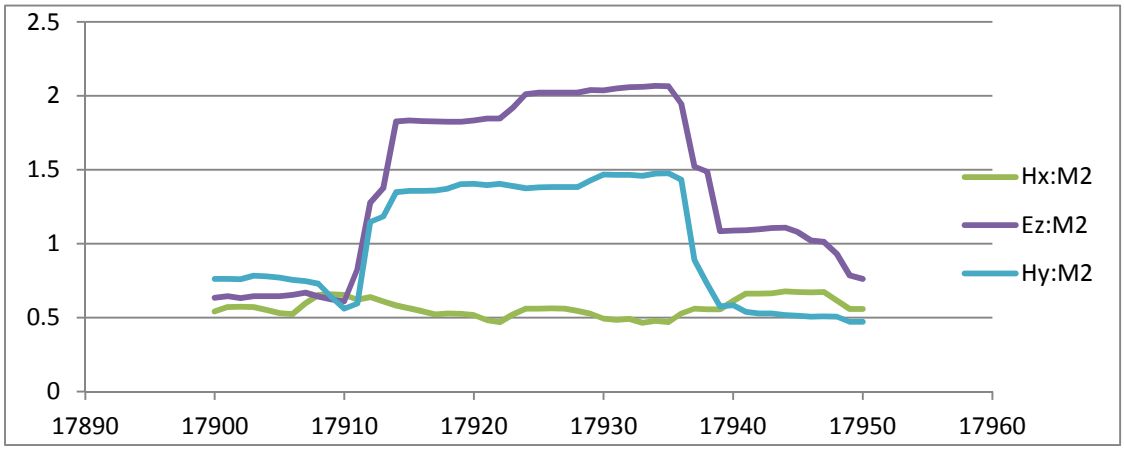
3. ábra. A normált *konvolució*k diagramja a 17936-os Q-kitörés környezetében

A *konvolució* az adatsorok skálázására érzékeny mutató, ami torzítaná az indikátor intervallumot, ezért normáljuk, vagyis elosztjuk az összetevő vektorok hosszával.

- I(Hx*Ez)I=(Hx*Ez)/(IHxI·IEzI) indikátor intervalluma: < 0,2.
- I(Hy*Ez)I=(Hy*Ez)/(IHyI · IEzI) indikátor intervalluma: < 0,2.

A Q-kitörés környezetében a *variancia* megnő (4. ábra).

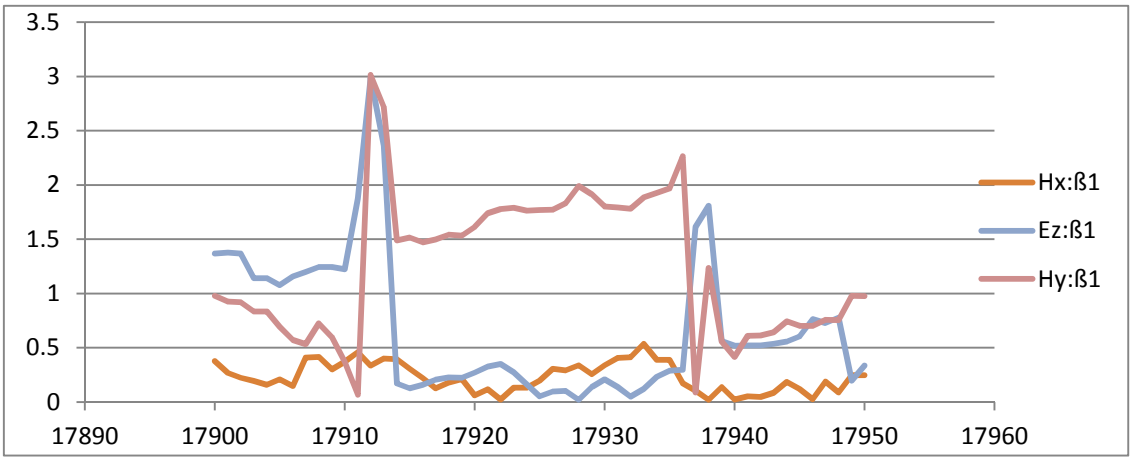
- Hx: M_2 indikátor intervalluma: > 1,4.
- Hy: M_2 indikátor intervalluma: > 1,4.
- Ez: M_2 indikátor intervalluma: > 2.



4. ábra. *Szórás* diagramok a 17936-os Q-kitörés környezetében

Ferdeség: csak az eloszlás pozitív farka utal találatra (5. ábra).

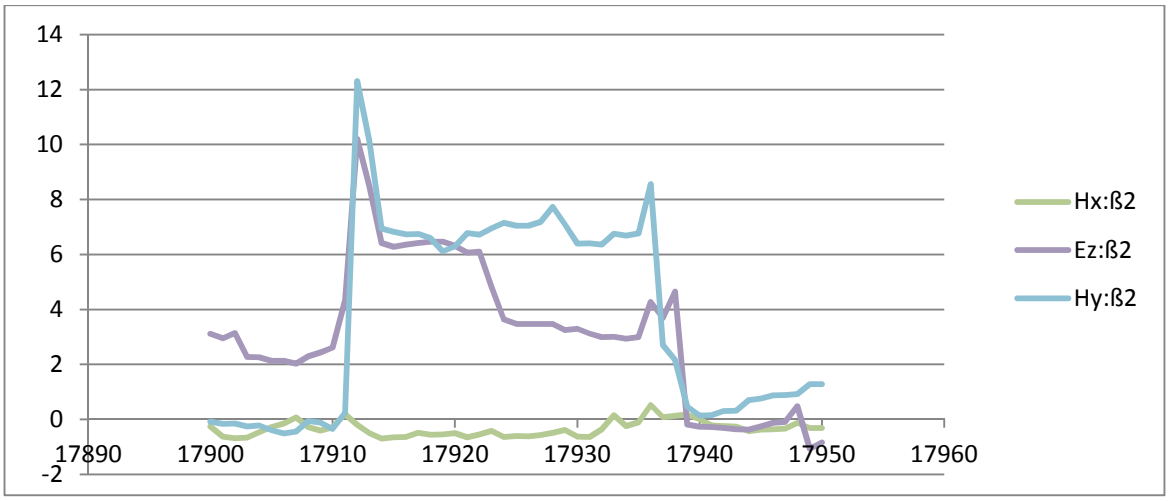
- Hx indikátor intervalluma: $ß_1 > 2$.
- Hy indikátor intervalluma: $ß_1 > 2$.
- Ez indikátor intervalluma: $ß_1 > 2$.



5. ábra. *Ferdeség* diagramok a 17936-os Q-kitörés környezetében

Csúcsosság: csak az eloszlás pozitív farka utal találatra.

- Hx indikátor intervalluma: $ß_2 > 8$.
- Hy indikátor intervalluma: $ß_2 > 8$.
- Ez indikátor intervalluma: $ß_2 > 8$.



6. ábra. *Csúcsosság* diagramok a 17936-os Q-kitörés környezetében

A *ferdeség és csúcsosság* mutatók többnyire megtalálták az ugrás környezetét, de előfordultak téves riasztások is. További sajátosságuk, hogy minden eseményre kétszer riasztanak:

- először amikor a részintervallum végén jelenik meg az esemény,
- másodszor, amikor a részintervallum elején jelenik meg az esemény.

7. A Q-kitörések kimutatásának logikai sémája

- Első lényeges döntés a mozgó statisztika részintervallum hosszának kiválasztása volt; ha ez túl rövid, akkor túlságosan érvényesülnek a lokális ‘eltérítő’ hatások, ha pedig túl hosszú, akkor egy részintervallum több ‘eseményt’ is tartalmazhat.
- Tranziens Q-kitörés keresésekor 25 mérés tett ki egy részintervallumot.
- A legfontosabb találatokat a I(Hx*Ez)I és I(Hy*Ez)I normált konvolúciók szolgáltatták – de nem feltétlen az esemény pontos helyén, hanem annak részintervallum sugarú környezetében jeleztek.
- A normált konvoluciók generálta hamis riasztásokat úgy szűrjük ki, hogy megnézzük a többi statisztikát (szórás, ferdeség, laposság) a vélt kitörés 25 mérés sugarú környezetében. Ha ott egyik sem jelzett, akkor a találatot elvetettük.
- A valószínűsíthető Q-kitörés pontos helyét úgy határozzuk meg, hogy az összetartozó találatokat tartalmazó 25 hosszú intervallumon meghatároztuk a mérések átlagát, és megkerestük azt a pontot, ahol ezen átlag és a mért érték eltérése maximális volt.
- Előfordulhat, hogy a Q-kitörés esemény közelében egyik konvolució sem riaszt; szerencsére ekkor a többi statisztika mindegyike jelzett a tranziens környezetében, vagyis a már bemutatott eltérés-számítással ekkor is meghatározható volt a Q-kitörés pontos helye.
- Összefoglalva, Q-kitörés esemény ott van,
  - ahol legalább 1 konvolúciós találat van, és a másik 3 statisztika közül legalább 1 találatot jelez, vagy
  - ahol nincs konvolúciós találat, de a többi statisztika mindegyike riaszt.
  - Szórás, ferdeség és csúcsosság statisztikák esetén akkor van találat, ha legalább két komponensben megjelenik, és közte van ez.

8. A méréshatár ugrások kimutatására használt statisztikák és indikátor intervallumaik

Nyilvánvaló, hogy a méréshatár ugrás pozíciójában a függvényérték bal- és jobboldali határértékének különbsége mutatja az ugrás nagyságát.

A vizsgált pontbeli függvényértéket ezért kétoldali *lineáris regresszióval* becsültük.

- Először a vizsgált ponttól balra található mérésekre illesztünk egyenest,
- majd a vizsgált ponttól jobbra található mérésekre illesztünk egyenest.
- A vizsgált pontban a két regressziós egyenes alapján becsült függvényértékek *dm* különbsége adja az ugrás nagyságát.
- Ha van tippünk a lépcsőmagasság alsó korlátjára, akkor ezt használhatjuk a regressziós becslések különbségének ellenőrzésére.
- Indikátor intervallumnak példánkban az *IdmI > 0,4* feltételt alkalmaztuk.

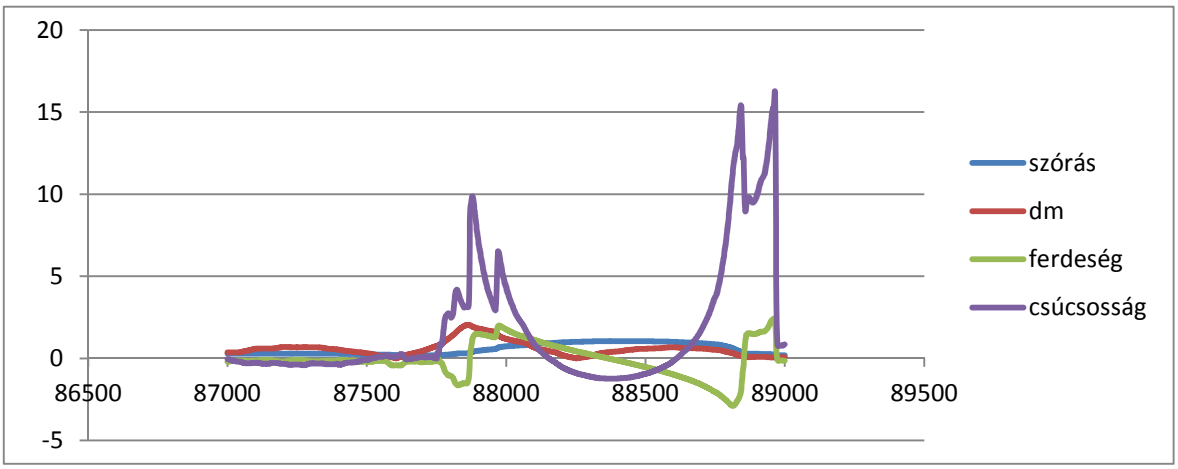
A *ferdeség*-mutató eloszlásának mindkét ‘farka’ generálhat találatot.

- Indikátor intervalluma: $ß_1 < -2$ vagy $ß_1 > 5$.
- A ferdeség–mutató általában jelzett az ugrás környezetében, de voltak téves riasztások is.

A *csúcsosság*-mutató eloszlásának mindkét ‘farka’ generálhat találatot.

- Indikátor intervalluma: $ß_2 < -1$ vagy $ß_2 > 8$.
- A csúcsosság–mutató általában jelzett az ugrás környezetében, de voltak téves riasztások is.

A *variancia* teljesen alkalmatlannak bizonyult a méréshatár ugrások kimutatására, ugyanis a mérések nem csak a keresett lépcsők környezetében adhatnak szokatlan szórást.



7. ábra. A használt statisztikák diagramja a 87862-es lépcső környezetében

9. A méréshatár ugrások kimutatásának logikai sémája

Első lényeges döntés a mozgó statisztika részintervallum hosszának kiválasztása volt; ha túl rövid, akkor túlságosan érvényesülnek a lokális ‘eltérítő’ hatások, ha pedig túl hosszú, akkor egy részintervallum több ‘eseményt’ is tartalmazhat.

A dőlésmérő adatsorában 1000 mérés képezett egy mozgó részintervallumot.

Ferdeség és csúcsosság statisztika esetén nem volt a priori tudásunk az indikátor intervallumról, ezért meghatároztuk a mutatók hisztogramját.

A hisztogramról úgy olvastuk le az *indikátor intervallum* korlátait, hogy mindkét farokba az összes mérés kb. 2%-a essen (sikertelenség, azaz túl sok, vagy túl kevés találat esetén, ezen paramétert módosítottuk).

Mindhárom statisztika (becsült lépcsőmagasság [*dm*], *ferdeség*, *csúcsosság*) esetén meghatároztuk azon pozíciókat, ahol a feltételek teljesültek.

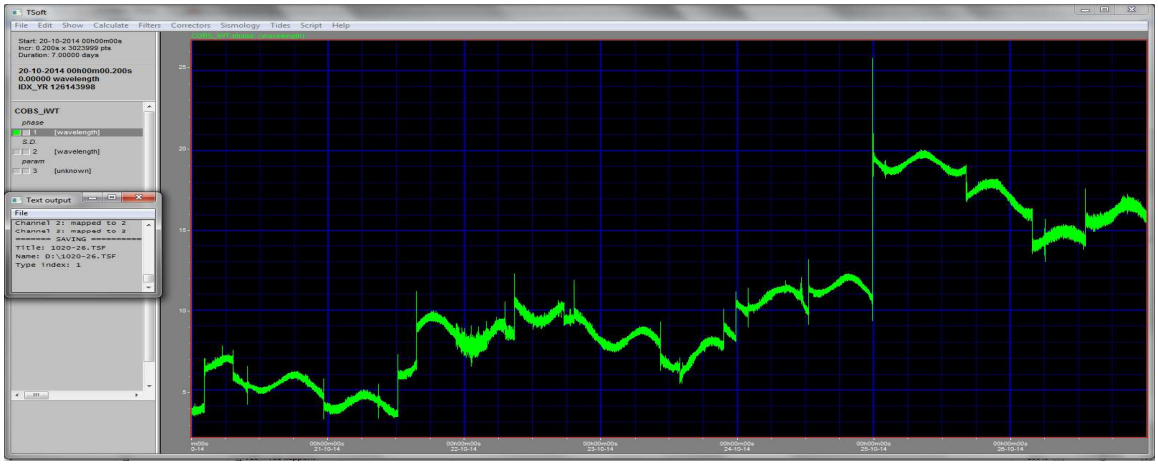
A pozíciók mindhárom mutató esetén zárt intervallumokba tömörültek, mert az esemény nem csak saját pozíciójában, hanem annak környezetében is anomáliát okozott a statisztikákban.

A statisztikák találati intervallumainak főpontját (az ugrás várt helyét) úgy határoztuk meg, hogy megkerestük a kétoldali lineáris becslések *dm* különbségének abszolút maximumát.

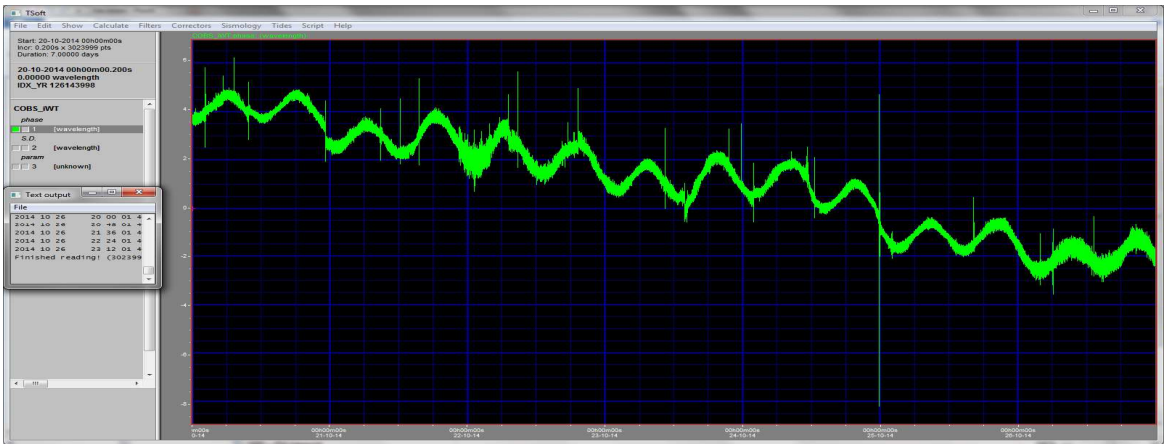
Azt tapasztaltuk, hogy a *dm* bázisú mutató minden alkalommal megtalálta az ugrás helyét, ha jól becsültük meg az ugrás minimális nagyságát (*dm* indikátor intervallumát).

A *ferdeség és csúcsosság* mutatók is többnyire megtalálták az ugrás környezetét, bár előfordultak téves riasztások is. További sajátosságuk, hogy minden eseményre kétszer riasztanak:

- először, amikor a részintervallum végén jelenik meg az esemény,
- másodszor, amikor a részintervallum elején jelenik meg az esemény.



8. ábra. az eredeti dőlésmérő adatsor



9. ábra. a dőlésmérő adatsor a lépcsők ignorálása után

10. Összefoglaló

Gyakori feladat, hogy hosszú, kézzel nehezen kiértékelhető mérési adatsorok eseményeinek pontos helyét automatikusan kell meghatározni, pl. online riasztás céljából.

A különböző mozgó statisztikák érzékenyek a lokális események előfordulására, ezért feladatonként megvizsgálandó, melyeket érdemes kimutatásukra felhasználni.

A mozgó statisztikák alkalmazásakor fontos meghatározandó paraméter a részintervallum hossza, illetve az indikátor intervallumok helyzete.

Egyes statisztikák a jól megválasztott paraméterek mellett is hibázhatnak: nem jeleznek egyes eseményeknél, vagy hamisan riasztanak.

Célszerű a vizsgálatoknál több statisztikát egyidejűleg figyelembe venni, mert kombinációjuk megbízhatóbb találatokhoz vezet.

Irodalomjegyzék

[1] **W. O. Schumann**, Über die strahlungslosen Eigenschwingungen einer leitenden Kugel, die von einer Luftschicht und einer Ionosphärenhülle umgeben ist, *Zeitschrift und Naturfirschung* 7a, (1952) 149–154. (doi: http://dx.doi.org/10.1515/zna-1952-0202)

[2] **D. J. Boccippio, E. R. Williams, S. J. Heckman, W. A. Lyons, I. T. Baker, R. Boldi**, Sprites, ELF transients, and positive ground strokes, *Science* 269 (5227) (1995), 1088–1091. (doi: http://dx.doi.org/10.1126/science.269.5227.1088)

[3] **C., E. Price, Greenberg, Y. Yair, G. Sátori, J. Bór, H. Fukunishi, M. Sato, P. Israelevich, M. Moalem, A. Devir, Z. Levin, J.H. Joseph, I. Mayo, B. Ziv, A. Sternlieb**, Ground-based detection of TLE-producing intense lightning during the MEIDEX mission on board the Space Shuttle Columbia, *G.R.L.* 31 (2004).

[4] **W. S. Hu, A. Cummer, W. A. Lyons, T. E. Nelson**, Lightning charge moment changes for the initiation of sprites, *G.R.L.* 29 (8) (2002), 1279. o.

[5] **Pödör Z.**, Töréspontok keresése meteorológiai idősorokban, és azok hatásainak vizsgálata, *Dimenziók* Matematikai Közlemények **II**. (2014), 35–43.