Sporteredmények modellezése extrém érték modellekkel

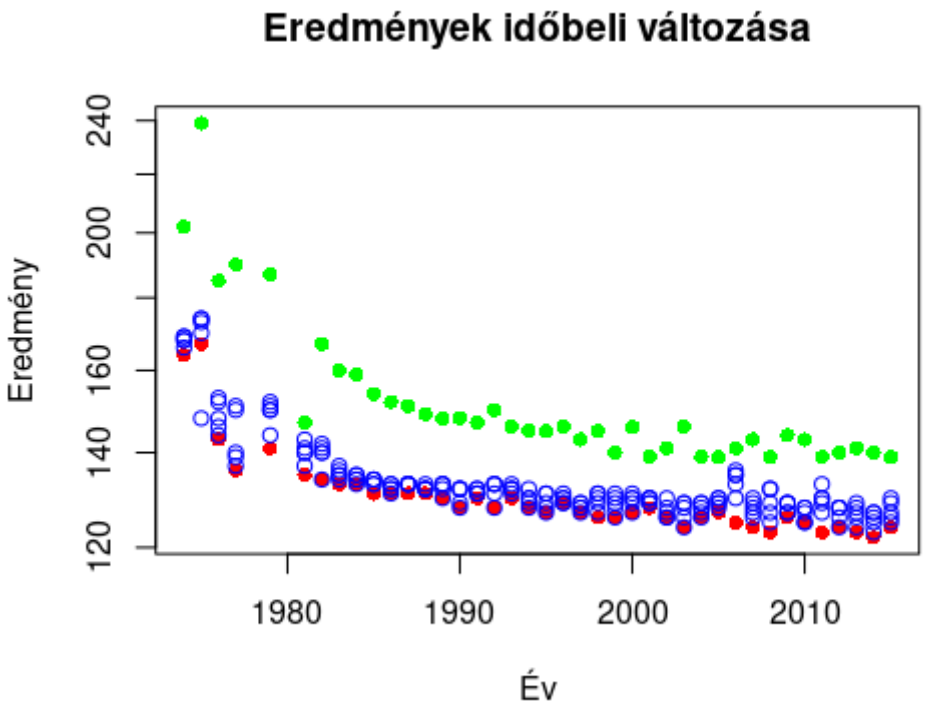
Csáfordi József András

Korábbi félév összefoglalása

A projektem első félévében megvizsgáltuk a Berlini maraton győtes eredményeit és ezt megpróbáltuk modellezni extrém-érték modellekkel. Említést tettünk a blokk-maximum módszerről és ennek a hasznosságáról, illetve a GEV modellelről, majd ennek segítségével megpróbáltunk előrejelzést adni a jövőbeli győztesek eredményeire. Különböző paraméterekkel próbáltuk felépíteni a modellünket, amiknek a becsléseit egy iteratív becslési segítségével határoztuk meg. A félév végén pedig megvizsgáltunk egy időfüggő hisztogrammot, az időszak végére számolt sűrűségfüggvénnyel és az úgynevezett "ablak-módszerrel" egy 5 éves visszatérési szintet 95%-os konfidencia intervallum mellett.
Ebben a félévben a cél az volt, hogy többdimenziós esetben is használható modelleket kapjunk. Ennek több módját is megvizsgáltuk: különböző maratonokon elért eredmények, illetve egy maratonon belül több eredményt vettünk figyelembe, nem csak a győztes időt.

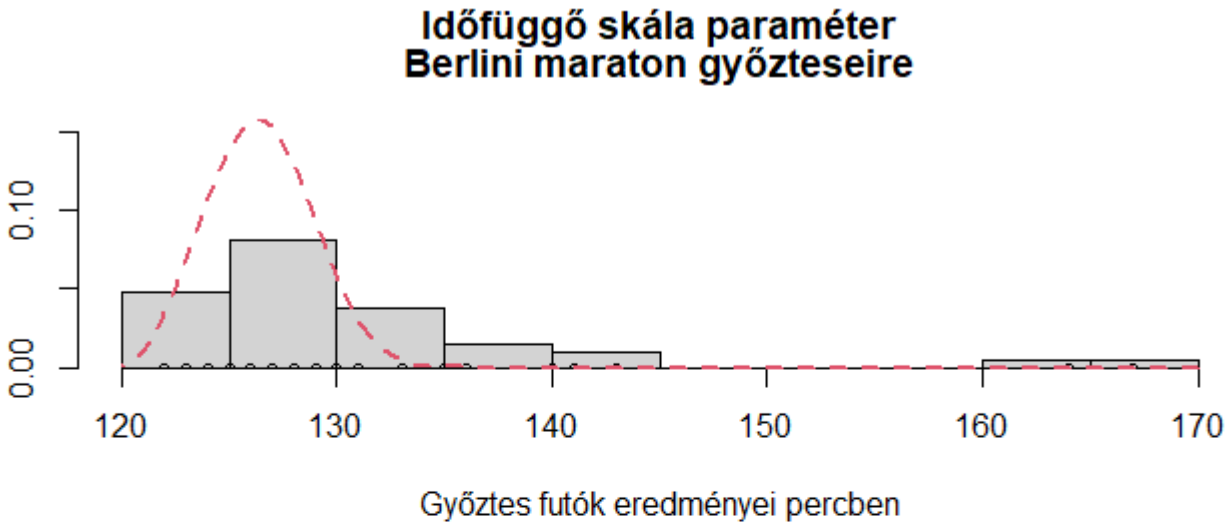
Adatok és feldolgozásuk

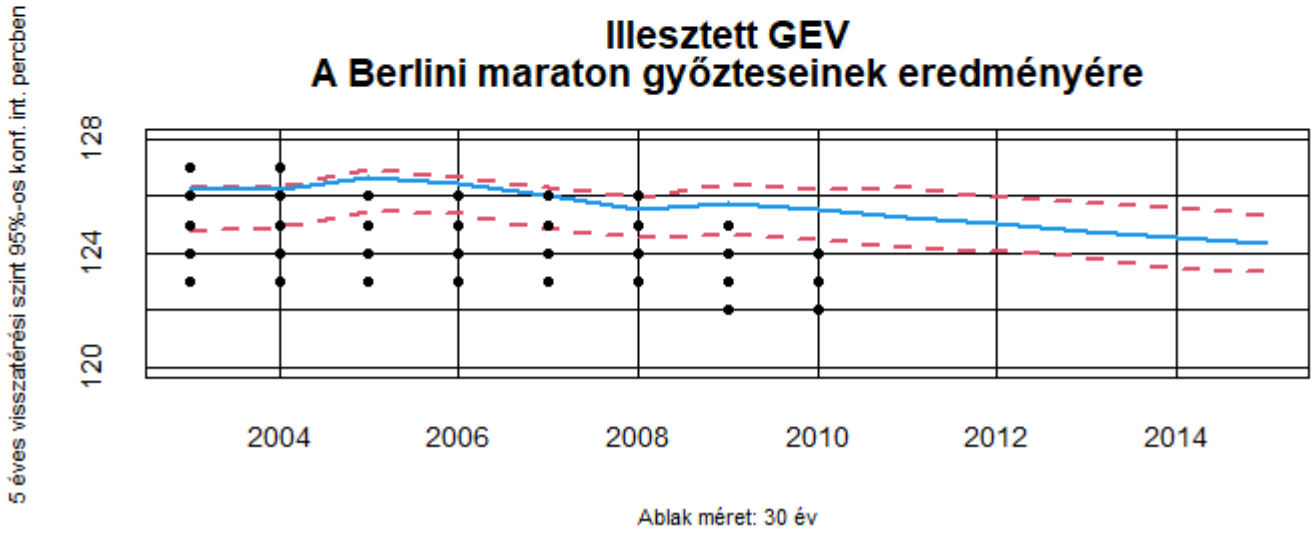
Az első félév során a férfi győztes eredményeket vettük alapul. Ebben a félévben több adatot is figyelembe vettünk, ezért kibővítettük az adathalmazt a női győztesek és a férfi első 6 helyezettjének az idejével. (1. kép) Az ábrán zöld ponttal jelöltük a női győztesek idejét, illetve kékkel a férfi időket. Az ábráról jól látható, hogy a nők esetében is, mint a férfi versenyzők esetében a 80-as évek előtt a győztes idők sokkal magasabbak voltak, illetve a szórásuk is nagyobb, mint a 2000-es évek után. Az adatok feldolgozásához először használjuk a már korábban említett blokk-maximum módszert. Minden évet válasszunk egy különálló blokknak és minden blokkban külön vegyük a legkisebb adatot. Ezzel a módszerrel megkapjuk az első félévben használt adatokat, amit az ábrán piros színnel jelöltünk.



GEV-eloszlás a férfi győztesekre

Az előző pontban bemutatott blokk-maximum módszer segítségével megkaptuk az előző félévben használt adatokat, így fel tudjuk írni a már ott bemutatott Időfüggő hisztogrammot és a visszatérési szintet 95%-os konfidencia intervallummal:





Hill becslés és GPD modell

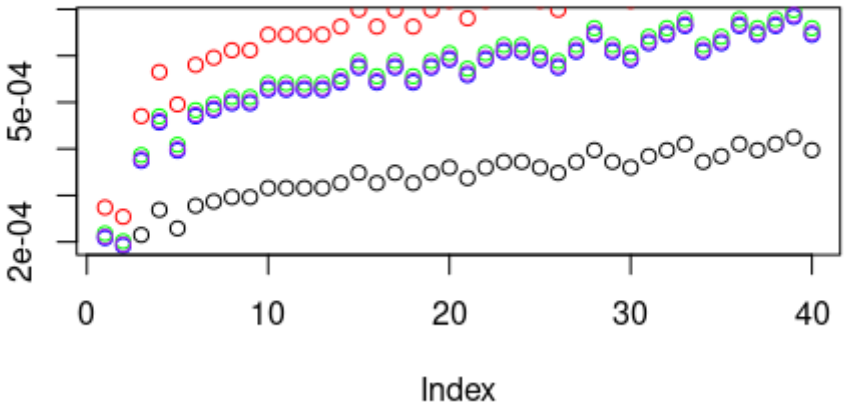
Egy meghatározott küszöbszint feletti adatok elemzésére alkalmas módszer a GPD modell. Az általános Pareto-eloszlás akalparaméterének meghatározásához használjuk a Hill-becslést

Hill becslés statisztikái:

- k = 3 : 0.0584321700696554
- k = 4 : 0.0697264387131906
- k = 5 : 0.0840435273494733
- k = 6 : 0.10002464470257
- k = 7 : 0.117255422261156
- k = 8 : 0.135595967463689

- k = 9 : 0.133035208982828
- k = 10 : 0.140232425814622
- k = 11 : 0.0749156572637565
- k = 12 : 0.0681051429670514
- k = 13 : 0.0624297143864638
- k = 14 : 0.0652903014099972
- k = 15 : 0.0606267084521403
- k = 16 : 0.0565849278886643
- k = 17 : 0.0607704159895331
- k = 18 : 0.0649778260792625
- k = 19 : 0.0613679468526368
- k = 20 : 0.0581380549130244
- k = 21 : 0.0552311521673732

Minden k értékre megkaptuk a becsült alakparamétert, amit a GPD modellben használhatunk:



Az alábbi ábrán az általános Pareto eloszlás pontjai láthatóak, különböző alak- és mérték- paraméterekkel.

Következő félévi tervek

A következő félévben érdemes lenne megvizsgálni, hogy többdimenziós modelleket alkalmazva, hogy viselkednek nem csak a győztesek eredményei, hanem a legjobbak eredményei is illetve, hogy a többi futó eredménye milyen kapcsolatban állhat a győztes idővel.

A modellünk továbbfejlesztését is célul tűztük ki, hiszen a GEV modell pontatlan eredményt adhat kevés adat esetén. Illetve még érdemes lenne megvizsgálni a GPD modellt úgy, hogy csak az alakparamétert használjuk.

Irodalomjegyzék

[1] Embrechts, Paul, Claudia Klüppelberg, and Thomas Mikosch. Modelling extremal events: for insurance and finance. Vol. 33. Springer Science Business Media, 2013.

[2] https://hu.wikipedia.org/wiki/Berlini_maraton

[3] https://www.bmw-berlin-marathon.com/en/impressions/statistics-and-history/results-archive/

[4] Coles, Stuart, et al. An introduction to statistical modeling of extreme values. Vol. 208. London: Springer, 2001.

[5] Brilhante, M. Fátima, M. Ivette Gomes, and Dinis Pestana. "A simple generalisation of the Hill estimator." Computational Statistics Data Analysis 57.1 (2013): 518-535.

[6] Arnold, Barry C. "Pareto distribution." Wiley StatsRef: Statistics Reference Online (2014): 1-10.