

TDK dolgozat

Nemoda Dóra

DEBRECENI EGYETEM

Informatikai Kar

Alkalmazott Matematika és Valószínűesszámitás Tanszék



TUDOMÁNYOS DIÁKKÖRI DOLGOZAT

VALÓSZÍNŰSÉGI SZÉLSEBESSÉG ELŐREJELZÉS

Konzulens:

Dr. Baran Sándor

egyetemi docens, DE IK

Dr. Horányi András

meteorológus, OMSZ

Szerző:

Nemoda Dóra

II. éves alk. mat. MSc
hallgató

**Debrecen
2011**

Tartalomjegyzék

Előszó	4
1. Bevezetés	6
2. BMA modell	8
2.1. Bayesian Model Averaging (BMA)	8
2.2. Paraméterek becslése: Maximum likelihood módszer EM algoritmussal . . .	9
2.3. Felcserélhető változók	12
2.4. Gamma modell	12
2.5. Előrejelzések és illeszkedésvizsgálat	15
3. Szélesebbség BMA előrejelzése	17
3.1. Adataink	17
3.2. A nyers ensemble vizsgálata	20
3.3. A tanuló periódus meghatározása	20
3.4. A modellillesztés eredményei	22
4. Összefoglaló	25
5. Saját munka leírása	27
A. Függelék: További előrejelzések	28
Ábrák jegyzéke	35

Előszó

Az időjárási előrejelzések a fizikai folyamatok leírása révén olyan matematikai módszereken alapulnak, melyek segítségével le lehet írni a légkör jövőbeli állapotát. Meteorológiai előrejelzéseket a légköri folyamatokat leíró parciális differenciálegyenletek megoldásával kaphatunk. Az egyenletek azonban nem lineárisak, s ezért megoldásuk analitikusan nem származtathatóak (hanem közelítő, numerikus megoldásra van szükség). Az egyik legfőbb nehézség azonban az, hogy a megfigyelések száma többnyire nem elégséges, így a légkör pillanatnyi állapota nem adható meg tökéletesen, azaz a modell kiindulási állapotát sohasem tudjuk teljesen pontosan meghatározni. Az előrejelzések eredményeit az említett adathiány, vagy bármilyen más alkalmazott közelítés hibája vagy éppen a kerekítési hibák is jelentősen befolyásolják. Ezen korlátozó tényezők felismerése vezetett ahhoz az eljáráshoz, hogy nem egy, hanem több előrejelzést készítenek figyelembe véve az összes számbavehető bizonytalansági tényezőt. Ennek a módszernek egyik megvalósítása az, hogy meghatározzák a legvalószínűbb kiindulási állapotot és ennek értékeit hibahatáron belül megváltoztatják, perturbálják, majd az így kapott kezdeti feltételek mindegyikéből készítenek egy-egy előrejelzést. Így tehát az előrejelzések együttesét kapjuk meg, s innen ered a módszer elnevezése is: ensemble (együttes) előrejelzés. Az így kapott előrejelzésegyüttest felhasználva képet kaphatunk az előrejelzett időjárási változók valószínűségi eloszlásáról. Éppen ezért az ensemble technika különösen hasznos, hiszen kiegészítő információt szolgáltat a várható jelenségek bekövetkezési valószínűségéről is. Ez az információ különösen értékes lehet szélsőséges időjárási helyzetek esetében.

A dolgozat célja meteorológiai (időjárási) változók jövőbeli értékeire vonatkozó valószínűségi sűrűségfüggvények megadása, melyeket a előrejelzésül szolgáló ensemble tagokból származtatunk. Ennek egyik lehetséges módja a Bayesian Model Averaging (BMA), melyet Raftery és társai (2005) mutatták be, mint egy az ensemble előrejelzések utófeldolgozására szolgáló módszert.

Magyarországon az Országos Meteorológiai Szolgálat (OMSZ) az ún ALADIN HUNEPS LAMEPS (Limited Area Modelling Ensemble Prediction System) előrejelzési rendszert használja (Hágel, 2010; Horányi és társai, 2011). Az ALADIN HUNEPS rendszer egy rövidtávú ensemble előrejelző rendszer, melynek alapja az ALADIN (Aire Limitée Adaptation dynamique Développement InterNational) rövidtávú, korlátos tartományú előrejelző modell (Hágel, 2010; Horányi és társai, 2011). Az OMSZ-nél a LAMEPS kísérleteket a PEACE (Prevision d'Ensemble a Courte Echéance), az ARPEGE francia globális modellre épülő, 10+1 tagú (10 felcserélhető tagot perturbált analízisből kiindulva készítenek, a 11-ik, az ún. kontrol előrejelzés számára pedig a perturbálatlan analízis szolgál kezdeti feltételeket), rövidtávú

globális ensemble rendszer előrejelzéseinek dinamikai leskálázásával (egy kisebb tartományra "koncentráló" előrejelzési modell alkalmazásával) kezdték. A globális rendszer minden egyes tagja kezdeti-, és peremfeltételt szolgáltat a korlátos tartományú, rövidtávú ensemble rendszer egy-egy tagja számára, így az ALADIN modellre épülő LAMEPS rendszer is 10+1 tagból áll és ez fogja képezni a vizsgálandó valószínűségi sűrűségfüggvény meghatározásának alapját. Végül pár mondatban kifejttem az általam végzett munkát a dolgozatban.

A dolgozat négy fő részből áll, az 1. fejezetben röviden ismertetem a szélsőérték előrejelzési problémát és magát az ensemble előrejelzési módszert. A következő fejezetben bemutatom a BMA modellt. Ez utóbbi több alfejezetből áll, először az általános modellt tárgyalom, majd ismertetem a paraméterek becslésére használható eszközöket és végül ejtek pár szót a felcserélhető tagok esetéről is. Azt követően egy speciális esetre is megnézem a modellt, amikor a minta Gamma eloszlást követ. Végül bemutatom a modell illeszkedésvizsgálatához szükséges legfontosabb mutatókat. A 3. fejezet tartalmazza az általunk elvégzett statisztikai elemzést melyhez az R programnyelv ensembleBMA csomagját használtuk (Fraley és társai, 2009, 2011). Ebben először bemutatom az általunk felhasznált adatsort, melyekből a BMA módszer segítségével készítünk szélsőérték előrejelzéseket. Megvizsgálom a modell illeszkedését és itt közölöm a módszer alkalmazása után kapott eredményeinket, észrevételeinket és következtetéseinket is. A dolgozatot egy összefoglaló zárja, amit a saját munka beutatása követ.

Ezúton szeretném megköszönni konzulensemnek, Dr. Baran Sándor egyetemi docensnek, a segítségét, támogatását és a belém vetett bizalmát. Köszönetet szeretnék továbbá mondani külső konzulensemnek Dr. Horányi Andrásnak az OMSZ meteorológusának, aki rengeteg segítséget nyújtott a munkám meteorológiai részének pontosabb megismerésében. Külön köszönet Mile Máténak és Szűcs Mihálynak, akik a feldolgozandó adatokat bocsátották a rendelkezésemre.

1. fejezet

Bevezetés

Az időjárási elemek közül kutatásunkban a szélre koncentrálunk, vagyis a továbbiakban szélesebesség előrejelzéseket fogunk vizsgálni. A szél az egyik kiemelt fontosságú előrejelzési paraméter, amely a gazdaság több ágazatában is fontos szerepet játszik. Itt talán elegendő csak a repülésmeteorológiai alkalmazásokra gondolni (viharos szél előfordulása le- és felszállás során, az optimális repülési útvonal tervezése vagy akár a megfelelő leszállópálya kiválasztása). A szélinformációk utáni keresletet jelentősen megnövelték a széleenergia hasznosítása körüli kérdések, különös tekintettel a szélparkokból várható széleenergia becslésére. Mindezen előrejelzésekhez és alkalmazásokhoz a valószínűségi információk elengedhetetlen adalékot szolgáltatnak, ezért ezek előállítása és optimalizálása kiemelten fontos feladat.

Dolgozatunkban kizárólag rövidtávú (0-48 óra) előrejelzésekkel foglalkozunk. A szélesebesség rövidtávú előrejelzési módszereiről Giebel (2003) ad kiváló összefoglalót.

Az időjárást előrejelzése során számszerű előrejelző modellek eredményeit vesszük alapvetően figyelembe. Az ensemble előrejelzések a 'hagyományos' előrejelzések olyan kiterjesztései, amelyek több előrejelzés futtatása révén megadják a várható időjárás bekövetkezési valószínűségét is. Az ensemble előrejelzési adatokból származtatott empirikus valószínűségi sűrűségfüggvényt a BMA módszer segítségével (és megfigyelési adatok bevonásával) pontosítjuk tovább. A gyakorlatban a szélesebesség előrejelzés statisztikai interpretációja során két jelentős probléma merül fel: először is a szélesebesség sűrűségfüggvénye jelentős mértékben ferde, másodsor a szélesebesség megfigyelési értékek más időjárási változókhoz képest eléggé széles skálán mozognak.

A BMA módszert, mint már említettük, Raftery és társai (2005) mutatták be, mint egy rövidtávra (0-48 órás időtartamú) vonatkozó statisztikai utófeldolgozó módszer, mely numerikus előrejelzésekből származó ensemble értékek feldolgozásából szolgáltatja az időjárási előrejelzéseket. Bármely időjárási változó BMA által készített előrejelzett sűrűségfüggvénye az ensemble tagokhoz tartozó sűrűségfüggvények súlyozott átlagaként áll elő, melynek alapja egy egyéni torzítással korrigált előrejelzés, ahol a súlyok tekinthetők a modellünk posteriori valószínűségeiknek, melyek megmutatják, hogy az előrejelzések egy adott tanuló periódus alatt mennyivel járulnak hozzá az együttes eloszláshoz. A Raftery és társai (2005) által készített BMA modellt eredetileg olyan időjárási változókra használták, melyek eloszlása közel normális, ilyen például a hőmérséklet illetve a tengerszinti légnyomás. A későbbiekben ezt

az eljárást Slougher és társai (2007) terjesztették ki egyéb változókra is, például szél erősség előrejelzésre. Slougher és társai (2010) csapadék előrejelzésekkel foglalkoztak, ahol úgynevezett kevert eloszlást alkalmaztak, amelyre jellemző, hogy az értékei pozitív valószínűséggel lehetnek nullák, amikor pedig az érték nem nulla, akkor erősen aszimmetrikus.

2. fejezet

BMA modell

A következő fejezet öt alrészből tevődik össze: a BMA modell ismertetése normális eloszlású minta esetén; a modellben fellépő paraméterek becslési eljárásainak bemutatása (Maximum likelihood módszer és az az EM algoritmus); a felcserélhető ensemble tagok esetének kezelése; Gamma eloszlású időjárási mennyiségek esete; a modellilleszkedés vizsgálata.

2.1. Bayesian Model Averaging (BMA)

Az általános statisztikai analízis, mint például a regresszió analízis, tipikusan egy feltételezett statisztikai modellből származik. A lehetséges modellek közül általában nem a legjobbat sikerül kiválasztani, csak egy megközelítőleg jó eredményt adót. Ezzel a pontatlansággal számolnunk kell a következtetések levonásakor. A BMA módszer úgy küzd le ezt a problémát, hogy nem "egy legjobb" modellt választ, hanem modellek együttesét veszi figyelembe.

Legyen y egy időjárási változó előrejelzése, melyet K darab statisztikai modell $M_1, \dots, M_K$ és egy tanuló adatsor alapján kívánunk meghatározni. Tanuló adatokon az előző n napra vonatkozó ensemble tagokat és tényleges megfigyeléseket értjük. Ez általában 10-45 nap között van. A teljes valószínűség tétele alapján az előrejelzendő y mennyiség sűrűségfüggvénye $p(y)$ a következő formulával adott:

$$p(y) = \sum_{k=1}^K p(y|M_k)p(M_k|y^T),$$

ahol $p(y|M_k)$ az előrejelzendő mennyiség sűrűségfüggvénye, ami csak az M_k modelltől függ, a $p(M_k|y^T)$ súly pedig annak a posteriori valószínűsége, hogy a tanuló adatok alapján az M_k modell a megfelelő választás. Ez a súly megmutatja, hogy az M_k modell mennyire illeszkedik jól a tanuló adatokra. Természetesen $\sum_{k=1}^K p(M_k|y^T) = 1$. A BMA módszer által adott együttes sűrűségfüggvény az egyes modellekből származó sűrűségfüggvények súlyozott átlagaként áll elő, mindegyiket akkora súllyal bevéve az együttes sűrűségfüggvénybe, amekkora a posterior valószínűsége volt. A modell elméletileg optimális tulajdonságokkal rendelkezik, továbbá mind valós mind szimulált adatok esetében jó eredményeket produkál.

Visszatérve a modell részletezésére, tekintsük y -t mint előrejelzési változót és legyenek $f_1, \dots, f_K$ az ensemble előrejelzések. Előfordulhat, hogy az ensemble tagokra torzítás korrekciót kell alkalmaznunk, ekkor a korrigált ensemble tagokkal dolgozunk tovább. Tegyük fel, hogy f_k a legjobb előrejelzés az ensemble tagok közül, és legyen $g_k(y|f_k)$ az y f_k -ra nézett feltételes sűrűségfüggvénye. A BMA előrejelző modell általános alakban a következőképpen írható fel:

$$p(y|f_1, \dots, f_K) = \sum_{k=1}^K w_k g_k(y|f_k),$$

ahol w_k annak a valószínűsége, hogy a tanuló adatok alapján az f_k a legjobb előrejelzés. Mivel az w_k -k valószínűségek, ezért nemnegatívak és összegük 1, azaz $\sum_{k=1}^K w_k = 1$. Az w_k -k meghatározását később részletezzük.

Az előrejelző eloszlás segítségével könnyen megkaphatjuk a vizsgált időjárási mennyiség pont előrejelzéseit (várható érték, medián), valamint a vele kapcsolatos események valószínűségeit (például hőmérséklet alacsonyabb, mint 273 K). Ugyanezen mennyiségek természetesen meghatározhatóak az $f_1, \dots, f_K$ nyers ensemble alapján is, azokat egy statisztikai mintának tekintve.

Amikor a hőmérsékletet vagy a tengerszinti légnyomást szeretnénk előrejelezni, gyakran ésszerűnek tűnik a feltételes sűrűségfüggvényt normális eloszlással közelíteni. Feltesszük tehát, hogy $g_k(y|f_k)$ normális eloszlású, $a_k + b_k f_k$ várható értékkel és σ szórással (Raftery és társai, 2005). Ennek alapján

$$y|f_k \sim \mathcal{N}(a_k + b_k f_k, \sigma^2).$$

Ebben az esetben a BMA által adott várható érték az y feltételes várható értéke, az adott ensemble előrejelzésekre nézve:

$$E[y|f_1, \dots, f_K] = \sum_{k=1}^K w_k (a_k + b_k f_k).$$

2.2. Paraméterek becslése: Maximum likelihood módszer EM algoritmussal

Tekintsük azt az esetet, amikor a feltételes sűrűségfüggvények normális eloszlásúak, azaz modellünk

$$p(y|f_1, \dots, f_K) = \sum_{k=1}^K w_k g_k(y|f_k, a_k, b_k, \sigma^2), \quad (2.2.1)$$

ahol g a $\mathcal{N}(a_k + b_k f_k, \sigma^2)$ sűrűségfüggvénye. Ez hőmérséklet és tengerszinti nyomás esetén optimális döntésnek tűnik, de szélsébség és csapadék esetén már más eloszlásra lesz szükség. Az alapgondolat minden eloszlás esetén hasonló lesz a következőhöz. Először azt nézzük meg, hogy becsüljük meg a modell által felhasznált $a_k, b_k, w_k, k = 1, \dots, K$, és σ^2 paramétereket, az ensemble előrejelzéseket és a tényleges megfigyeléseket tartalmazó tanuló adatokat alapul véve. Jelöljük s -el és t -vel a megfigyelés helyét és idejét, azaz f_{kst} jelölje az s helyre és t

időpontra készített k -adik ensemble tagot. Jelölje továbbá y_{st} az ugyanezen helyen és időben mért tényleges megfigyeléseket.

Először a fő paraméterekkel foglalkozunk. Az a_k -t és b_k -t lineáris regresszióval becsüljük meg, melynek képlete

$$y_{st} = a_k + b_k f_{kst} + \varepsilon_{st},$$

ahol y_{st} a tényleges megfigyelés minden s és t értékre, és ε_{st} az ún. hibatag. Ezek a paraméterek függnak az ensemble tagoktól, éppen ezért külön kiszámítjuk őket minden egyes ensemble tagra, felhasználva a megfigyelt tényleges adatokat, mint függő változót, és a szélesség előrejelzéseket (f_{kst}) mint független változót. Jelölje $\hat{a}_k$ és $\hat{b}_k$ a megfelelő becsléseket, valamint a továbbiakban legyen

$$g_k(y_{st}|f_{kst}, \sigma^2) := g_k(y_{st}|f_{kst}, \hat{a}_k, \hat{b}_k, \sigma^2).$$

Az $w_1, \dots, w_K$ és σ becslését maximum likelihood módszerrel végezzük a tanuló adatokat alapul véve, feltételezve azok tértől és időtől való függetlenségét. Az együttes sűrűségfüggvényre a statisztikában általában használatos a likelihood függvény elnevezés is. A $w_1, \dots, w_K$ és σ^2 paraméterekhez tartozó L likelihood függvény következő:

$$L(w_1, \dots, w_K, \sigma^2) := \prod_{s,t} p(y_{st}|f_{1st}, \dots, f_{Kst}) = \prod_{s,t} \sum_{k=1}^K w_k g_k(y_{st}|f_{kst}, \sigma^2),$$

ahol $p(y|f_1, \dots, f_K)$ az y $f_1, \dots, f_K$ -ra nézett feltételes sűrűségfüggvénye, ezek szorzata pedig az y együttes sűrűségfüggvénye. Az L likelihood függvény maximumhelye adja meg az L maximum likelihood becslést. Mivel a logaritmus függvény szigorúan monoton növekvő, így L maximum helyének meghatározása egyenértékű a

$$\mathcal{L}(w_1, \dots, w_K, \sigma^2) := \log L(w_1, \dots, w_K, \sigma^2) = \sum_{s,t} \log \left(\sum_{k=1}^K w_k g_k(y_{st}|f_{kst}, \sigma^2) \right) \quad (2.2.2)$$

loglikelihood függvény maximum helyének megkeresésével (ahol a logaritmus a e alapú logaritmust jelöli) és ahol az összegzést a tanuló adat összes s és t értékre megcsináljuk. Bizonyos feltételek mellett paramétereink becslését megkaphatjuk a következő likelihood egyenletrendszer megoldásával:

$$\frac{\partial \mathcal{L}(w_1, \dots, w_K, \sigma^2)}{\partial w_i} = 0, \quad i = 1, \dots, K, \quad \text{és} \quad \frac{\partial \mathcal{L}(w_1, \dots, w_K, \sigma^2)}{\partial \sigma^2} = 0.$$

Mivel a (2.2.2) loglikelihood függvény maximuma általában nem határozható meg analitikusan, ezért az ún. EM-algoritmus fogjuk használni. Az Expectation-Maximization (EM) algoritmus egy széleskörben alkalmazott módszer a maximum likelihood becslés megtalálására, ami akkor is alkalmazható, amikor nem teljes az adatállományunk, vagy amikor már a Newton-Ralphson módszer túl bonyolult lenne (McLachlan and Krishnan, 1997). Az EM algoritmus minden iterációja felváltva használ két lépést; az E-lépést (Expectation - számítási) és az M-lépést (Maximization - maximalizálás). Az E-lépés a paraméter korábbi becslése alapján rekonstruálja a hiányzó adatokat feltételes várható érték képzéssel, míg az M-lépés az ilyen módon kiegészített teljes adatrendszerből meghatározza a likelihood-függvény

maximumhelyét. Nem csak hiányzó adatok esetében alkalmazható az EM algoritmus, hanem cenzorált vagy keverékfelbontás esetén is (McLachlan and Krishnan, 1997, Section 1.4.3).

Tekintsük a (2.2.2) loglikelihood függvényt, és vezessük be z_{kst} nem megfigyelhető változókat, ahol $z_{kst} = 1$ ha a k -adik ensemble tag az s helyen és t időben mért legjobb megfigyelés, egyébként $z_{kst} = 0$. Minden egyes (s, t) párra a $(z_{1st}, \dots, z_{Kst})$ koordináták közül pontosan egy egyenlő 1-el, az összes többi 0. Így jutunk el a teljes loglikelihood függvényhez:

$$\mathcal{L}_C(w_1, \dots, w_K, \sigma^2) = \sum_{s,t} \sum_{k=1}^K z_{kst} \log g_k(y_{st} | f_{kst}, \sigma^2),$$

aminek már meg tudjuk határozni a maximumát.

A $w_1, \dots, w_K, \sigma^2$ -ra adott kezdeti értékekből indulunk ki. Az E-lépésben z_{kst} -t a paraméterek aktuális értékét felhasználva becsüljük meg, megjegyezve, hogy a z_{kst} -re adott becslés általában nem egész érték, függetlenül attól, hogy ténylegesen 0-t vagy 1-et vehet fel értéként. Az M-lépésben a z_{kst} pillanatnyi értékével becsüljük meg az ismeretlen paramétereket.

A (2.2.1) BMA modellre, az E-lépés a következő képlettel adható meg

$$\hat{z}_{kst}^{(n+1)} = \frac{w_k^{(n)} g_k(y_{st} | f_{kst}, \sigma^{(n)})}{\sum_{i=1}^K w_i^{(n)} g_i(y_{st} | f_{kst}, \sigma^{(n)})},$$

ahol az n felső index mutatja az EM algoritmus n -edik iterációs lépését, $g_k(y_{st} | f_{kst}, \sigma^{(n)})$ normális eloszlású $\hat{a}_k + \hat{b}_k f_{kst}$ várható értékkel és $\sigma^{(n)}$ szórással, és a becslésekhez y_{st} használjuk fel. Az M-lépés w_k és σ becsléséből áll, felhasználva z_{kst} aktuális becsléseit, nevezetesen $\hat{z}_{kst}^{(n+1)}$ -t. Így

$$\begin{aligned} \hat{w}_k^{(n+1)} &= \frac{1}{n} \sum_{s,t} \hat{z}_{kst}^{(n+1)}, \\ \hat{\sigma}^{2(n+1)} &= \frac{1}{n} \sum_{s,t} \sum_{k=1}^K \hat{z}_{kst}^{(n+1)} (y_{st} - f_{kst})^2, \end{aligned}$$

ahol n a tanuló adat megfigyeléseinek a száma.

Az E- és M-lépéseket addig iteráljuk, amíg a log-likelihood függvény értékei, a paraméter becslések értékei és a $\hat{z}_{kst}^{(n)}$ értékek már csak egészen kicsit változnak. A loglikelihood értékek garantáltan növekedni fognak minden egyes iterációs lépésben, amiből arra következtetünk, hogy általában ez az érték a likelihood függvény lokális maximumához konvergál. A globális maximumhoz való konvergencia nem következik, így a megoldás erősen függ a kezdeti értékektől. Ezen probléma kiküszöbölésére a kezdeti értékeket múltbeli tapasztalatok alapján választunk, ezek közel jó megoldást fognak szolgáltatni.

2.3. Felcserélhető változók

Eddig olyan esetekben alkalmaztuk a módszert, amikor az ensemble tagok egyértelműen különböztek egymástól. Ebben a részben azt mutatjuk meg, hogyan módosul a módszer,

amikor az ensemble tagok között levő különbségek csak a véletlenszerű perturbációból származhatnak: az ensemble előrejelzések többsége ez utóbbi típusba sorolható (és ez így van az általunk vizsgált esetben is). Ezekben az esetekben tekintsük őket felcserélhetőnek, és megköveteljük, hogy a BMA várható érték és a torzítási paraméterek, valamint a BMA súlyok minden ensemble tag esetén azonosak legyenek egymással. A torzítás paramétereit az előzőekhez hasonlóan lineáris regresszióval becsüljük. A csapadék és szélerősség előrejelzéshez használt Gamma modellben további feltétel még (Sloughter és társai (2007), Sloughter és társai (2010)), hogy a felcserélhető tagok szórás paraméterének egyenlőnek kell lennie.

Vegyünk egy m tagú ensemble-t, ahol az előrejelzések K darab felcserélhető csoportba sorolhatók. Legyen az k -adik felcserélhető csoportnak $m_k \geq 1$ darab felcserélhető tagja, vagyis $\sum_{k=1}^K m_k = m$. Legyen továbbá $f_{k,\ell}$ a k -adig csoport ℓ -edik tagja. Ekkor a BMA keverék eloszlás sűrűségfüggvénye a következő alakban írható fel:

$$p(y|(f_{k,\ell})_{k=1,\dots,K,\ell=1,\dots,m_k}; (\theta_k)_{k=1,\dots,K}) = \sum_{k=1}^K \sum_{\ell=1}^{m_k} w_k g_k(y|f_{k,\ell}, \theta_k),$$

ahol θ_k a paraméterek halmaza, $w_i \geq 0$ és $\sum_{k=1}^K w_k = 1$. Megjegyezzük ismét, hogy a csoportokon belül a BMA súlyok, a sűrűségfüggvények és a paraméterek mind megegyeznek. A teljes loglikelihood függvény a korábbiakhoz hasonlóan

$$\mathcal{L}_C(w_1, \dots, w_K; (\theta_k)_{k=1,\dots,K}) = \sum_{s,t} \left(\sum_{k=1}^K \sum_{\ell=1}^{m_k} z_{k\ell st} \log g_k(y_{st}|f_{k\ell st}, \theta_k) \right).$$

Az EM algoritmus számításainál az E-lépést módosítsuk a következőre:

$$\hat{z}_{k\ell st}^{(n+1)} = \frac{w_k^{(n)} g_k(y_{st}|f_{k\ell st}, \theta_k^{(n)})}{\sum_{p=1}^K \sum_{q=1}^{m_k} w_p^{(n)} g_p(y_{st}|f_{pqst}, \theta_p^{(n)})}$$

Az M-lépésben a súlyokat megkaphatjuk a megfelelő felcserélhető csoportba tartozó ensemble tagok átlagaként:

$$\hat{w}_k^{(n+1)} = \frac{1}{n} \frac{1}{m_k} \sum_{s,t} \left(\sum_{\ell=1}^{m_k} \hat{z}_{k\ell st}^{(n+1)} \right).$$

A szükséges paraméterek aktuális becsléseiket, az $\mathcal{L}_C(\hat{w}_1, \dots, \hat{w}_K; (\theta_k)_{k=1,\dots,K})$ teljes loglikelihood függvény maximumhelyének meghatározásával kapjuk.

2.4. Gamma modell

Haslett és Raftery (1989) a szélsőbességi értékek négyzetgyökét normális eloszlásúnak modellezte. Korábban a szélsőbességi értékeket általában Weibull eloszlásúnak tekintették, de Tuller és Brett (1984) feljegyezte, hogy nem mindig teljesülnek annak a feltételei, hogy a modell illeszkedjen a Weibull eloszlásra. Ezért a Weibull eloszlást úgy általánosítjuk, hogy ha a megfigyelt szélsőbesség értékeket hatványozzuk, akkor azok Gamma eloszlásúnak tekinthetők, ezzel kapunk egy eloszlást, ami jó közelítést fog adni. A Gamma eloszlás a nyers

szélsébségi mintára általában jobb illeszkedést ad mint bármely hatványraemeléses módszer. Ezen információk tudatában a szélsébség sűrűségfüggvényét Gamma eloszlással fogjuk modellezni. Legyen y valószínűségi változó α -ad rendű β paraméterű Gamma eloszlású, sűrűségfüggvénye a következő

$$g(y) = \frac{1}{\beta^\alpha \Gamma(\alpha)} y^{\alpha-1} \exp(-y/\beta),$$

ha $y > 0$, és $g(y) = 0$ egyébként. Ezen Gamma eloszlás várható értéke $\mu = \alpha\beta$ és szórásnégyzete $\sigma^2 = \alpha\beta^2$.

Sloughter és társai (2010) előzetes elemzései azt mutatták, hogy a megfigyelt szélsébségi adatok az előrejelzések megközelítőleg lineáris függvényei, és ugyanez mondható el a szórásukról is. Ezek alapján a szélsébségnek a k -adik ensemble tagon alapuló sűrűségfüggvényére a következő modellt kapjuk:

$$g_k(y|f_{k\ell}; b_{0k}, b_{1k}, c_{0k}, c_{1k}) = \frac{1}{\beta_{k\ell}^{\alpha_{k\ell}} \Gamma(\alpha_{k\ell})} y^{\alpha_{k\ell}-1} \exp(-y/\beta_{k\ell}), \quad (2.4.1)$$

ahol a várható érték ($\mu_{k\ell} = \alpha_{k\ell}\beta_{k\ell}$) és a szórás ($\sigma_{k\ell}^2 = \alpha_{k\ell}\beta_{k\ell}^2$) az ensemble tagok előrejelzéseitől ($f_{k\ell st}$) függ, nevezetesen:

$$\mu_{k\ell} = b_{0k} + b_{1k}f_{k\ell st},$$

és

$$\sigma_{k\ell}^2 = c_{0k} + c_{1k}f_{k\ell st}.$$

Itt feltesszük, hogy minden ensemble tag szórás paramétere konstans. Így egyszerűsödik a modell, mivel csökkentjük a becsülendő paraméterek számát, ezzel téve könnyebbé a paraméterbecslési eljárásokat, és csökkentve a becslési hibát. A tapasztalatok alapján ez a kikötés nem befolyásolja lényegesen az eredményeket (Sloughter és társai, 2010). Ezek után c_{0k} és c_{1k} -t helyettesítsük c_0 illetve c_1 -al. További egyszerűsítés lehet ha b_{0k} és b_{1k} paraméterek nem függnek a csoportoktól, azaz helyettük b_0 és b_1 értékeket használunk.

A paraméterek becslését b_{0k} és b_{1k} -val kezdjük, ezeket lineáris regresszióval becsüljük meg:

$$y_{st} = b_{0k} + b_{1k}\bar{f}_{kst} + \varepsilon_{st},$$

ahol

$$\bar{f}_{kst} = \frac{1}{m_k} \sum_{\ell=1}^{m_k} f_{k\ell st},$$

és ε_{st} az ún. hibatag. Ezek a paraméterek függnek az ensemble tagoktól, éppen ezért külön kiszámítjuk őket minden egyes ensemble tagra, felhasználva a megfigyelt tényleges adatokat, mint függő változót, és a szélsébség előrejelezéseket ($f_{k\ell st}$) mint független változót. Jelölje $\hat{b}_{0k}$ és $\hat{b}_{1k}$ a megfelelő becsléseket, valamint a továbbiakban legyen

$$g_k(y_{st}|f_{k\ell st}, c_0, c_1) := g_k(y_{st}|f_{k\ell st}, \hat{b}_{0k}, \hat{b}_{1k}, c_0, c_1).$$

A még fennmaradó $w_1, \dots, w_K, c_0$ és c_1 paraméterek becslését a normális eloszlású esethez hasonlóan maximum likelihood módszerrel végezzük a tanuló adatokat alapul véve. A

$w_1, \dots, w_K, c_0$ és c_1 paraméterekhez tartozó $\mathcal{L}$ (nem teljes) log-likelihood függvénye a most következő:

$$\mathcal{L}(w_1, \dots, w_K, c_0, c_1) = \sum_{s,t} \log \left(\sum_{k=1}^K \sum_{\ell=1}^{m_k} w_k g_k(y_{st} | f_{k\ell st}; c_0, c_1) \right).$$

Megjegyezzük, hogy számításaink elvégzésekor a 0.1 m/s alatti értékeket 0-nak tekintettük. A log-likelihood módszernél az egyes megfigyelések logaritmusát kell meghatározni, melyet nem lehet elvégezni, amikor az adott megfigyelés 0 értéket vesz fel.

Olyan Gamma eloszlású mintánál, amikor az adatállomány tartalmaz nulla értéket, Wilks (2006) javasolta, hogy a paraméterek becslésére használt maximum likelihood módszernél minden nulla értéket helyettesítsünk olyan értékekkel, amelyet nullára kerekítenénk, jelen esetben 0 és 0.1 között értékkel. Tekintsük a 0 értéket felvevő y_{st} megfigyeléseket, és helyettesítsük sűrűségfüggvényüket a következő valószínűséggel

$$p(y_{st} | (f_{k\ell st})_{k=1, \dots, K; \ell=1, \dots, m_k}) = P(0.1 | (f_{k\ell st})_{k=1, \dots, K; \ell=1, \dots, m_k}),$$

ahol

$$P(0.1 | (f_{k\ell st})_{k=1, \dots, K; \ell=1, \dots, m_k}) := \int_0^{0.1} p(y_{st} | (f_{k\ell st})_{k=1, \dots, K; \ell=1, \dots, m_k}) dy.$$

Vezessük be ismét $z_{k\ell st}$ nem megfigyelhető változókat, ahol $z_{k\ell st} = 1$ ha az $f_{k\ell st}$ ensemble tag az y_{st} legjobb becslése, egyébként $z_{k\ell st} = 0$. Minden egyes (s, t) párra pontosan egy $\{z_{k\ell st}\}_{k=1, \dots, K; \ell=1, \dots, m_k}$ érték egyenlő 1-el, az összes többi 0. Ekkor a teljes loglikelihood függvény a következő alakú

$$\mathcal{L}_C(w_1, \dots, w_K, c_0, c_1) = \sum_{s,t} \left(\sum_{k=1}^K \sum_{\ell=1}^{m_k} z_{k\ell st} \log(g_k(y_{st} | f_{k\ell st}; c_0, c_1)) \right).$$

Tekintsük a $w_1, \dots, w_K, c_0$ és c_1 paraméterek egy kiinduló becslését. Az E-lépésben $z_{k\ell st}$ a paraméterek pillanatnyi becslésével kiszámítható, azaz

$$\hat{z}_{k\ell st}^{(n+1)} = \frac{w_k^{(n)} g_k(y_{st} | f_{k\ell st}; c_0^{(n)}, c_1^{(n)})}{\sum_{p=1}^K \sum_{q=1}^{m_k} w_p^{(n)} g_p(y_{st} | f_{k\ell st}; c_0^{(n)}, c_1^{(n)})},$$

ahol az n -edik index az algoritmus n -edik lépésére utal, éppen ezért $w_k^{(n)}$ a n -edik iterációja a w_k -nak. $g_k(y_{st} | f_{k\ell st})$ -t definiáljuk a (2.4.1) képlettel, ha y_{st} nem nulla, vagy $G_k(0.1 | f_{k\ell st})$ ha y_{st} nulla értéket vesz fel, ahol

$$G_k(0.1 | f_{k\ell st}) = \int_0^{0.1} g_k(y | f_{k\ell st}) dy.$$

Az M-lépésben a $w_1, \dots, w_K, c_0$ és c_1 paraméterek függvényeként kapott loglikelihood függvényt maximalizáljuk, a $z_{k\ell st}$ becslésének ismeretében figyelembe véve az adatállományt és az előző becsléseket. Akár $z_{k\ell st}$ akár $w_1, \dots, w_K, c_0$ és c_1 paramétereket használjuk a loglikelihood

függvényénél, maximalizálásnál ugyanazt az értéket fogjuk kapni, legyen tehát $z_{k\ell st} = \hat{z}_{k\ell st}^{(n+1)}$. Ezért

$$\hat{w}_k^{(n+1)} = \frac{1}{n} \frac{1}{m_k} \sum_{s,t} \sum_{\ell=1}^{m_k} \hat{z}_{k\ell st}^{(n+1)},$$

ahol n a tanuló adatsor megfigyeléseinek a száma.

Végül $\hat{c}_0^{(n+1)}$ és $\hat{c}_1^{(n+1)}$ becslését $\mathcal{L}_C(\hat{w}_1^{(n+1)}, \dots, \hat{w}_K^{(n+1)}, c_0, c_1)$ maximumának meghatározásával kapjuk.

2.5. Előrejelzések és illeszkedésvizsgálat

Ha ismert az előrejelzendő időjárási mennyiség eloszlása, abból nagyon egyszerűen juthatunk pontelőrejelzésekhez, kiszámítva például a mediánt vagy a várható értéket. Megtudjuk határozni az eloszlás kvantiliseit vagy például annak a valószínűségét, hogy az előrejelzés egy bizonyos tartományba esik (pl. 10 m magasságban a szélsősebesség nagyobb, mint 12 m/s).

Ugyanezeket a mennyiségeket természetesen a nyers ensemble alapján is meghatározhatjuk, úgy tekintve arra, mint egy statisztikai mintára.

Az illeszkedésvizsgálat célja az, hogy megmutassuk, az általunk meghatározott előrejelző eloszlás, illetve a belőle származó pontelőrejelzések, mennyire közelítik meg a tényleges adatokat. Jelölje y_{st} az s helyen és t időben mért tényleges megfigyeléseket. Jelölje továbbá $p_{st}(y)$ illetve $P_{st}(y)$ az s helyen és t időre való előrejelzés becsült sűrűség- illetve eloszlásfüggvényét. Legyen $\hat{y}_{st}$ a becsült $p_{st}(y)$ sűrűségfüggvényből vagy a nyers ensemble alapján számított pontbecslés és jelölje továbbra is n a mintaelemszámot.

Pontbecslések összehasonlítására illetve illeszkedésének vizsgálatára általában az átlagos abszolút eltérést (Mean Absolute Error - MAE),

$$\frac{1}{n} \sum_{s,t} |y_{st} - \hat{y}_{st}|$$

és az átlagos négyzetes eltérés négyzetgyökét (Root Mean Square Error - RMSE) használjuk

$$\sqrt{\frac{1}{n} \sum_{s,t} (y_{st} - \hat{y}_{st})^2}.$$

A medián előrejelzések esetén a MAE, míg a várható érték illetve átlag esetén az RMSE az optimális illeszkedési mutató (Gneiting, 2011).

Valószínűségi eloszlások illeszkedésének vizsgálatára egy lehetséges mutató az átlagos CRPS (Continuuous Rank Probability Score) (Gneiting and Raftery, 2007), azaz

$$\frac{1}{n} \sum_{s,t} crps(P_{st}, y_{st}),$$

ahol

$$crps(P, y) = \int_{-\infty}^{\infty} (P(x) - 1_{\{x \geq y\}})^2 dx.$$

Nyers ensemble esetén az átlagos CRPS mutatót az ensemble empirikus eloszlásfüggvényéből számoljuk. Nyilvánvaló módon mindhárom fenti mutató esetén az alacsonyabb értékek jelzik a jobb illeszkedést.

Egy másik, a valószínűségi előrejelzések illetve a nyers ensemble tagok illeszkedésének vizsgálatára szolgáló mutató a lefedettség, melyen azt értjük, hogy a tényleges megfigyelések mekkora arányban esnek például a 67% -os vagy a 90% -os előrejelzési intervallumba. Optimális esetben ezek az arányok legalább 67% illetve 90%, itt a minél magasabb értékek számítanak jónak. Vizsgálhatjuk továbbá, hogy átlagosan milyen széles például a 67% -os vagy a 90% -os előrejelzési intervallum. Minél keskenyebb, annál pontosabb az előrejelzés (Gneiting and Raftery, 2007).

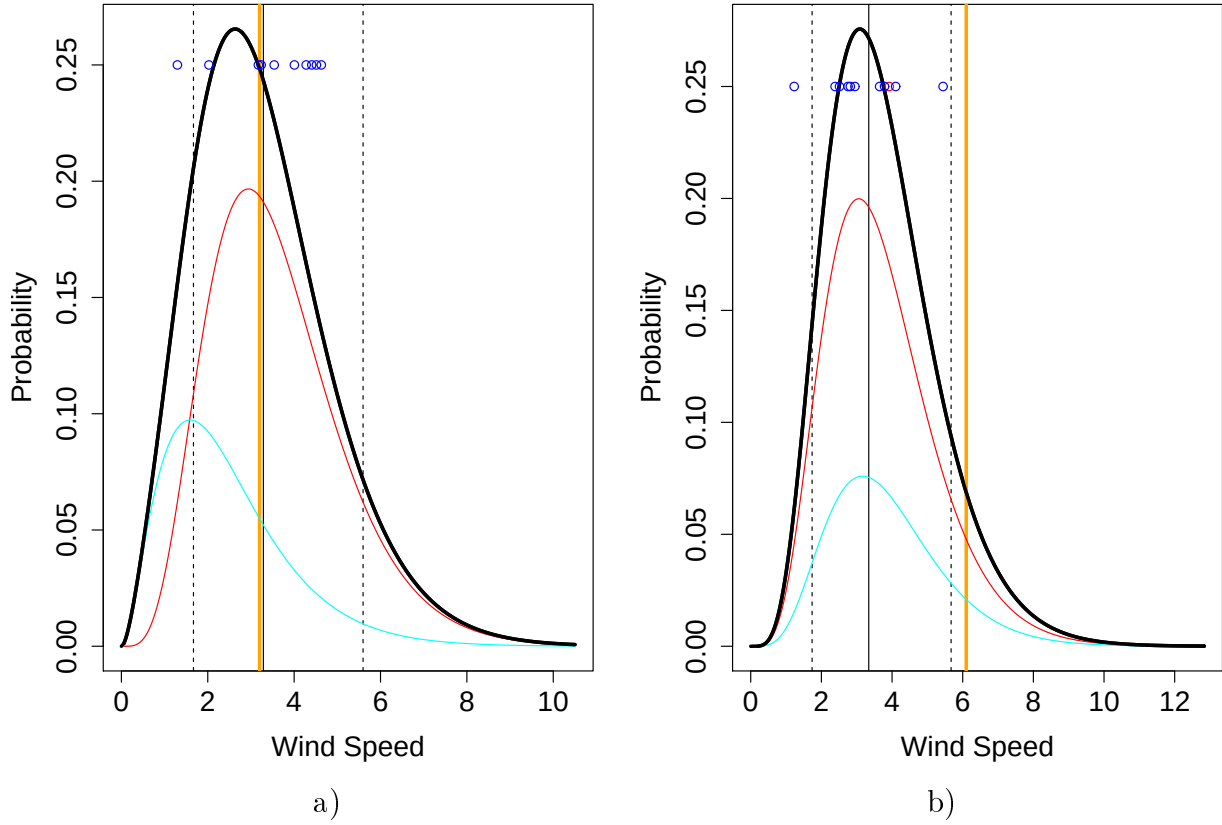
3. fejezet

Szélsébség BMA előrejelzése

3.1. Adataink

A BMA módszer alkalmazásához ténylegesen mért szélsébségeket és az OMSZ LAMEPS rendszere által szolgáltatott szélsébség előrejelzéseket használtunk (Hágel, 2010; Horányi és társai, 2011). Az előrejelzési adatállományunk 2010. október 1. és 2011. március 25. közötti időszakra vonatkozó 42 órás ensemble előrejelzéseket tartalmazza 10 magyarországi nagyváros esetén (Miskolc, Szombathely, Győr, Budapest, Debrecen, Nyíregyháza, Nagykanizsa, Pécs, Kecskemét, Szeged). Az ensemble 1+10 tagból áll, a kontrol (control) tagból, amit a perturbálatlan kezdeti értékekből számítottak ki; valamint 10 olyan előrejelzésből (lam1,..., lam10), ami perturbált kezdeti értékek alapján lett meghatározva. Ez utóbbiak egymástól nem megkülönböztethetők (bármelyik lehet tehát a ténylegesen bekövetkező szél érték), ezeket felcserélhetőnek tekinthetjük (Fraley és társai, 2010). Az adatállomány ezenkívül tehát tartalmazta az egyes napokon és az egyes állomásokon mért tényleges szélsébségeket is (observation). Nem elhanyagolható továbbá ezen adatok keletkezésének dátuma (date), valamint a megfigyeléseknek helyet adó város kódja (station) és földrajzi koordinátái (latitude és longitude). Hiányzó adatokat 2010.10.18-án és 2011.02.15-én találhatunk, 2010.11.20-án pedig egyetlen előrejelzés sem állt rendelkezésünkre. A szél sebességét a meteorológiában m/s-ban mérik, számításaink elvégzésekor a 0.1 m/s alatti értékeket 0-nak tekintettük. Az alábbi táblázat az adatállomány első 11 sorát mutatja.

"control"	"lam01"	"lam02"	...	"lam10"	"observation"	"date"	"station"	"latitude"	"longitude"
1.45958	1.8572	1.59668	...	1.88054	0.50	2010100112	12772	48.10	20.8
1.11019	1.1671	1.39280	...	1.29920	2.70	2010100112	12812	47.20	16.6
1.12803	1.5023	1.19865	...	1.34658	1.00	2010100112	12822	47.70	17.6
0.32109	0.3940	0.90456	...	0.72474	2.00	2010100112	12843	47.50	19.1
1.37952	1.2760	2.06969	...	1.77216	1.50	2010100112	12882	47.50	21.6
1.40861	1.3269	1.85133	...	1.47112	1.00	2010100112	12892	47.90	21.6
1.79307	1.8275	1.62762	...	1.51466	2.00	2010100112	12925	46.50	17.0
2.07634	2.1199	1.24950	...	1.94657	2.10	2010100112	12942	46.10	18.2
0.90836	0.9370	1.16169	...	1.00753	1.00	2010100112	12970	46.90	19.7
1.19373	0.8519	1.75918	...	1.02291	2.80	2010100112	12982	46.30	20.2
2.59177	1.2057	1.50121	...	1.53215	1.10	2010100212	12772	48.10	20.8



3.1. ábra. A szélsébség előrejelzésének BMA sűrűségfüggvényei, az egyes csoportok sűrűségfüggvényei, az ensemble tagok, a BMA medián előrejelzés és a tényleges megfigyelések Debrecen városára a) 2010.12.30; b) 2011.03.17

Az adatainkra a következő modellt írtuk fel. Legyen az y szélsébségi mennyiség sűrűségfüggvénye

$$p(y|f_c, f_{\ell 1}, \dots, f_{\ell 10}; b_{0c}, b_{1c}, b_{0\ell}, b_{1\ell}, c_0, c_1) =$$

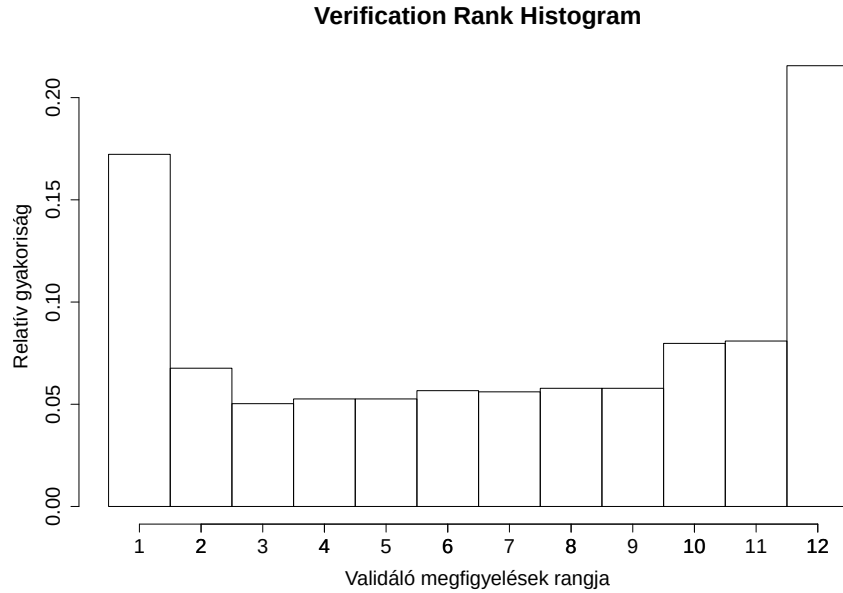
$$= wg_c(y; f_c, b_{0c}, b_{1c}, c_0, c_1) + \sum_{j=1}^{10} \frac{1-w}{10} g_{\ell}(y; f_{\ell j}, b_{0\ell}, b_{1\ell}, c_0, c_1),$$

ahol f_c jelöli az ensemble control tagját, $f_{\ell 1}, \dots, f_{\ell 10}$ pedig a 10 felcserélhető ensemble tagot. Legyen továbbá $g_c(y; f_c, b_{0c}, b_{1c}, c_0, c_1)$ egy $b_{0c} + b_{1c}f_c$ várható értékű és $c_0 + c_1f_c$ szórású Gamma eloszlás sűrűségfüggvénye; $g_{\ell}(y; f_{\ell j}, b_{0\ell}, b_{1\ell}, c_0, c_1)$ pedig hasonlóan az előzőhöz $b_{0\ell} + b_{1\ell}f_{\ell j}$ várható értékű és $c_0 + c_1f_{\ell j}$ szórású Gamma eloszlás sűrűségfüggvénye. A modellezés során feltettük, hogy nincs különbség a két csoport várható érték paramétere között, azaz $b_{0c} = b_{0\ell} = b_0$ és $b_{1c} = b_{1\ell} = b_1$.

A paramétereink becsléseit az elméleti részben leírtakhoz hasonlóan végezzük el. A becslésekhez itt is egy n napos tanuló adatsort veszünk a mintából, ez az adatsor tartalmazza

	2010.12.30			2011.03.17		
	medián	átlag	megfigyelés	medián	átlag	megfigyelés
Ensemble	4.0059	3.5697	3.2	2.9492	3.2323	6.1
BMA	3.2876	3.4965	3.2	3.3409	3.5623	6.1

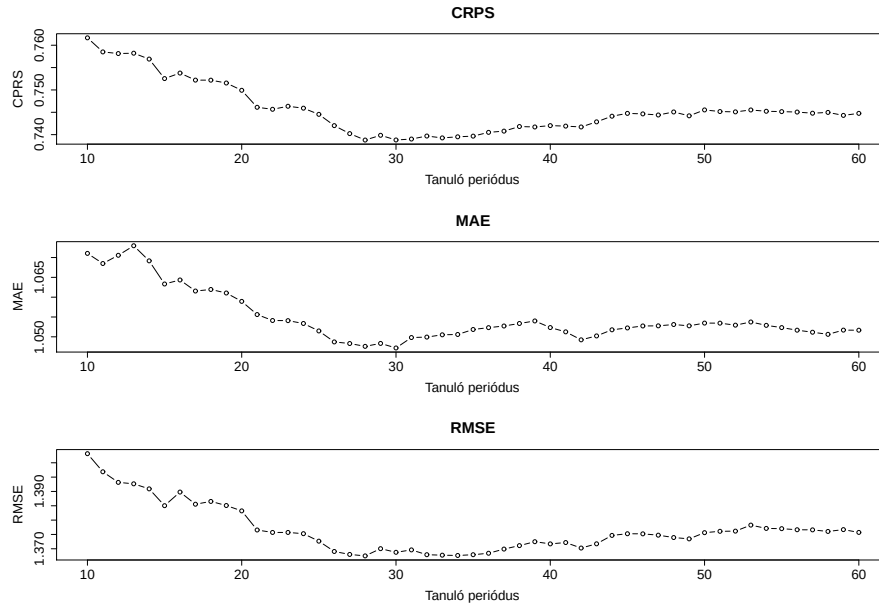
3.1. táblázat. Debreceni előrejelzések 2010.12.30-re és 2011.03.17-re



3.2. ábra. Verification Rank Histogram

a tényleges megfigyeléseket és az ensemble előrejelzéseket az előző n napra. A b_0 illetve b_1 becsléseit tehát lineáris regresszióval végezzük, míg w és c_0, c_1 becsléseit maximum likelihood becsléssel, illetve a likelihood függvény maximumát EM algoritmussal határozzuk meg.

Illusztrációként tekintsük a 3.1 ábrát, ahol két adott dátumra (2010.12.30 és 2011.03.17) vonatkozó debreceni előrejelzések sűrűségfüggvényeit ábrázoltuk. A fekete vonal az együttes sűrűségfüggvény; a piros az ensemble kontrol tagjának a sűrűségfüggvénye, a kék pedig az ensemble 10 felcserélhető tagjainak az együttes sűrűségfüggvénye. Az ábrán látható körök az ensemble értékei, a piros a kontrol tagé, míg a 10 kék kör a felcserélhető tagoké. Az ábrán megfigyelhető függőleges vonalak közül a narancssárga a tényleges megfigyelést, a fekete a medián előrejelzést, míg a szaggatott vonalak az alsó illetve felső decilist jelölik. A 3.1 a) ábrán egy olyan esetet mutattunk be, amikor a tényleges megfigyelés az ensemble által lefedett tartományba esik. Láthatjuk, hogy habár az ensemble tagok szórása nem jelentős, a BMA medián előrejelzés majdnem pontosan megegyezik a tényleges megfigyeléssel. A 3.1 táblázat azt mutatja, hogy a BMA előrejelzések jobbak, mint az ensemble tagokból számított megfelelő mennyiségek. A 3.1 b) ábra egy ettől egy lényegesen különböző esetet mutat, amikor bár az ensemble előrejelzéseknek nagyobb a szórása, azok mindegyike alulbecsli a tényleges megfigyelést. Itt mind az ensemble, mind pedig a BMA előrejelzések nagyon rosszak (lásd 3.1 táblázat), de az ensemble átlag és medián távolabb van a valóságtól, mint BMA



3.3. ábra. A szélesség BMA előrejelzésének CRPS, valamint MAE és RMSE értékei (medián előrejelzés) a tanuló periódus hosszának függvényében

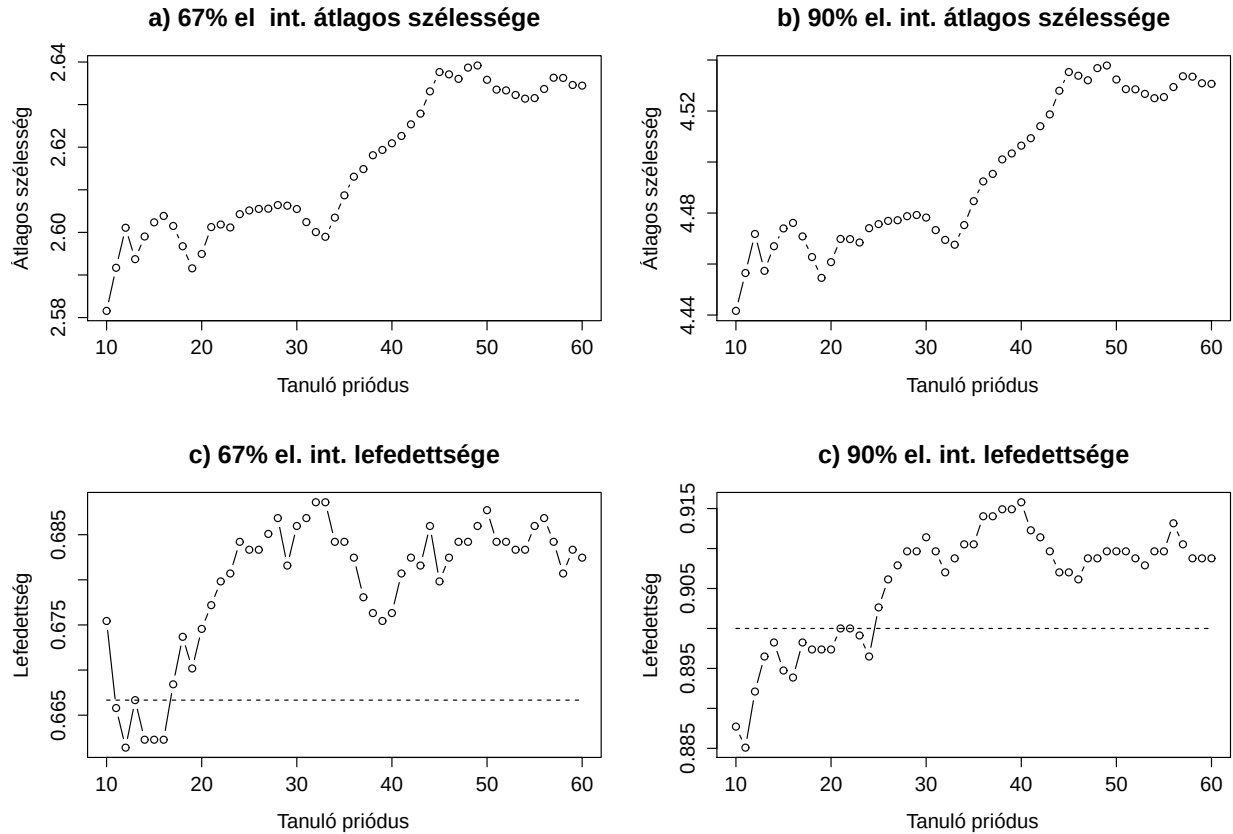
sűrűségfüggvényből származó értékek.

3.2. A nyers ensemble vizsgálata

A nyers ensemble előrejelzésekkel kapcsolatban általános probléma, hogy nem elég nagy a szórásuk és nem kalibráltak. Ez utóbbi azt jelenti, hogy a tényleges megfigyelések nem elég nagy arányban esnek a legkisebb és legnagyobb ensemble tagok közé. Ennek vizsgálatára szolgál a Verification Rank hisztogram, amely a tényleges megfigyelések és az ensemble tagok egymáshoz képesti elhelyezkedését mutatja. Meghatározásához először számítsuk ki minden egyes vizsgált helyen és időpontban a tényleges megfigyelés rangját, azaz az ensemble tagokkal közösen alkotott rangsorban elfoglalt helyét. Ezeket a rangokat azután egy hisztogrammal ábrázolhatjuk, ami megfelelően szóródó ensemble tagok esetén közel egyenletes eloszlást mutat. A 3.2 ábra alapján látható, hogy ez esetünkben messze nem teljesül, az ensemble tagok többnyire vagy mind alul vagy mind felülbecslik a tényleges megfigyelést. Mindezek mellett az ensemble kalibráltsága sem megfelelő, hiszen a tényleges megfigyelés az esetek 61.21% -ában esik az ensemble által lefedett tartományba.

3.3. A tanuló periódus meghatározása

Első lépésként meg kell határoznunk a modellillesztéshez szükséges optimális tanuló periódus hosszát, majd ennek tudatában ismertetjük a konkrét eredményeinket. A tanuló periódus hosszának meghatározása azért fontos, mert ezen adatok alapján fogjuk elvégezni a paramé-

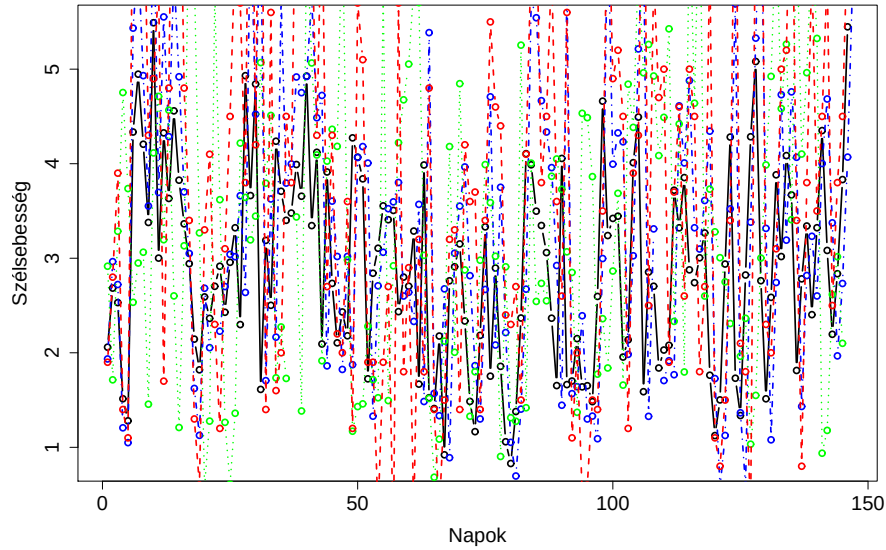


3.4. ábra. A BMA előrejelzések 67% és 90%-os lefedettsége és az előrejelző intervallumok átlagos szélessége a tanuló periódus hosszának függvényében

terek becslését, azaz meghatározni a BMA súlyait, a szórást és a korrigált együtthatókat. A tanuló periódus meghatározásához ugyanarra az adatsorra többször (a tanuló periódus 10-60 napnak feltételezve) lefutattuk a modellt, és megvizsgáltuk a kapott eredményeket.

A 3.3 ábrán a CRPS, a BMA medián előrejelzés MAE, valamint a BMA várható érték RMSE illeszkedési mutatójának értékeit ábrázoltuk a tanuló periódus hosszának függvényében. Ahogy a 2.5 fejezetben említettük, ezek az értékek a „minél kisebb annál jobb” elvet követik, azaz annál jobb egy modell illeszkedése minél alacsonyabbak. Leolvasható, hogy a CRPS és az RMSE 28 napnál veszi fel a minimumát (0.7388 és 1.3675 értékekkel). A MAE a minimumát 30 napnál veszi fel (1.0472), míg a második legkisebb értékét 28 napnál veszi fel (1.0476). Ezekből arra következtetünk, hogy a 28 napos tanuló idő tűnik megfelelőnek.

A 3.4(a) és (b) ábrán a 67% illetve 90% -os előrejelzési intervallumok átlagos szélességét ábrázoltuk a tanuló időszak hosszának függvényében, a 3.4(c) és (d) ábrán pedig a lefedettségeket, azaz a tényleges megfigyelések arányát a 67% illetve 90% -os előrejelzési intervallumokban. Az 3.4(c) ábra alapján látható, hogy a 67% -os lefedettség már 17 napos tanuló idővel eléri a 67% -ot. 90% -os előrejelzési intervallum esetében ez a határ 24 napnál van (3.4(d) ábra), és a tanuló periódus hosszának növelésével mindkét lefedettség növekszik. Így az előző eredményekkel összevetve, a 28 napos tanuló periódus megfelelő választásnak



3.5. ábra. Előrejelzések Debrecenre a 2010.10.29 és 2011.03.25 közötti időszakra (BMA medián előrejelzések: fekete; a tényleges megfigyelések: szaggatott piros; az ensemble tagok mediánjai: pontozott zöld; és az ensemble tagok átlaga: szaggatott kék)

tűnik.

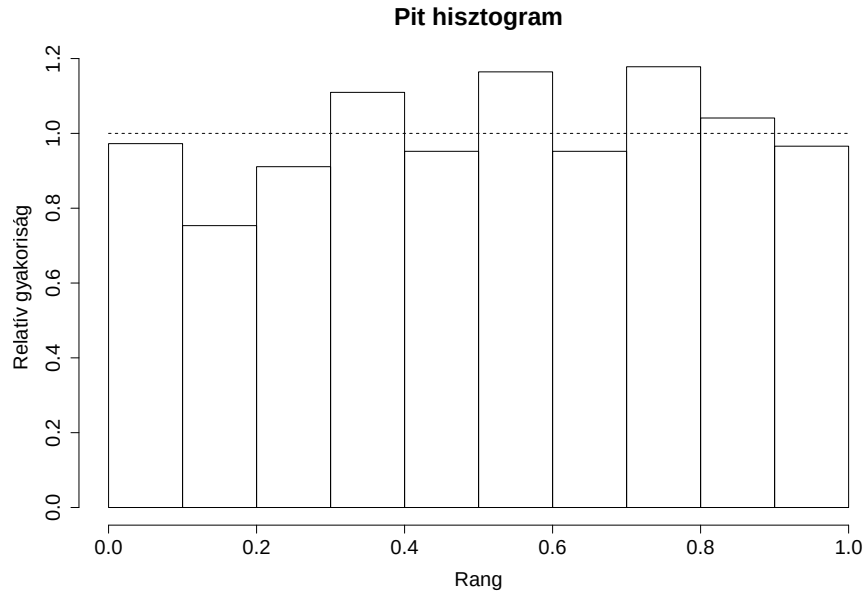
3.4. A modellillesztés eredményei

A 3.5 ábrán a Debrecenre adott különböző előrejelzéseket ábrázoltuk 2010.10.29 és 2011.03.25 időszakra valamint a tényleges megfigyeléseket. Látható, hogy a BMA előrejelzések kevésbé változékonyak, mint a nyers ensemble tagokból számolt mennyiségek, ezért a tényleges megfigyelések szélsőséges értékeit nem követik jól. Azt is megfigyelhetjük viszont, hogy az átlagos értékeket jobban becsülik, mint az ensemble átlagok és mediánok. A további 9 város előrejelzései a Függelékben találhatóak.

Az illeszkedési mutatók értékeit is kiszámítottuk, és igen jó eredményeket kaptunk, melyeket a 3.2 táblázatban foglaltunk össze. Mint látható, a BMA előrejelzések átlagos CRPS értéke a nyers ensemble tagokból számított érték 88.11%-a. Megfigyelve a pontelőrejelzések mutatóit leolvasható, hogy a legjobb illeszkedést a BMA medián előrejelzése mutatja, mely-

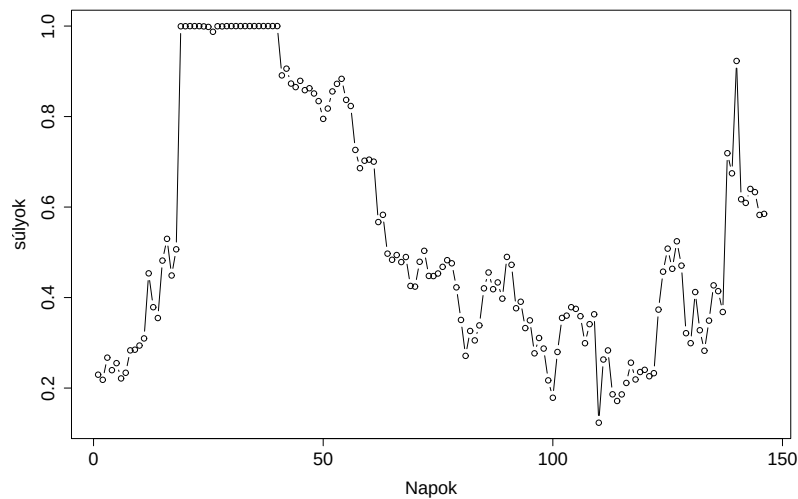
	CRPS	MAE		RMSE		lefedettség		átl. szélesség	
		medián	átlag	medián	átlag	67%	90%	67%	90%
Ensemble	0.8599	1.1215	1.1090	1.4634	1.4440	0.3870	0.5514	1.4388	2.2001
BMA	0.7577	1.0678	1.0763	1.4213	1.4067	0.6808	0.9034	2.6359	4.5297

3.2. táblázat. Illeszkedési mutatók

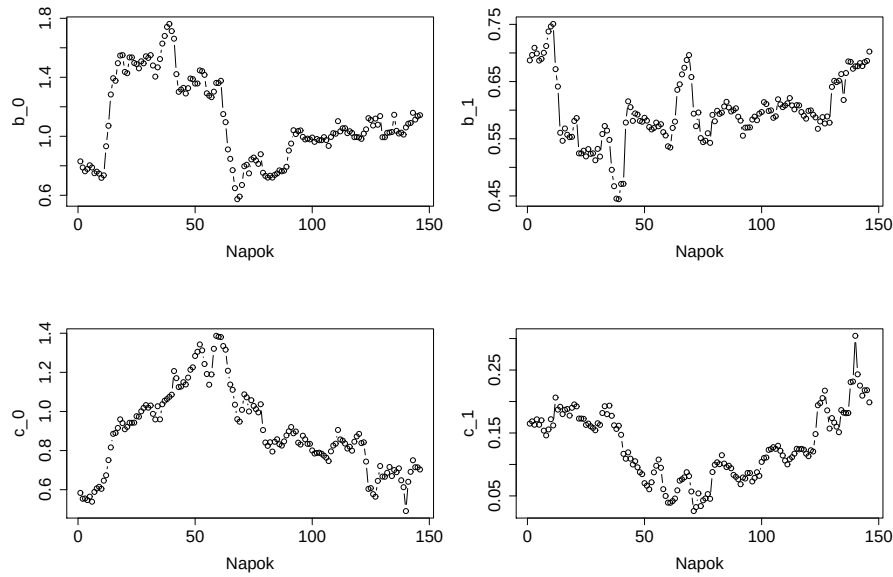


3.6. ábra. Pit hisztogram

nek mint az RMSE mind pedig a MAE értéke alacsonyabb, mint az ensemble átlag és medián megfelelő értékei. A leglényegesebb különbség azonban a lefedettségekben mutatkozik, ahol a BMA sűrűségfüggvény hozza a kívánt értékeket, míg a nyers ensemble igen távol áll tőlük. Az előrejelző intervallumok átlagos szélességét megvizsgálva azonban azt tapasztaljuk, hogy a BMA sűrűségfüggvényhez tartozó értékek majdnem kétszeresei a nyers ensemble értékeinek. Ennek oka lehet a nyers ensemble tagoknak a korábban már említett kis szóródása. Megállapíthatjuk tehát, hogy a BMA előrejelzés kalibráltsága megfelelő, azonban nem elég éles.



3.7. ábra. Ensemble kontrol tagjának súlyai



3.8. ábra. A Gamma modell paraméterei

Megvizsgáltuk a PIT hisztogramot is, amely úgy áll elő, hogy az adott előrejelzett eloszlásfüggvénybe behelyettesítjük a neki megfelelő tényleges megfigyelést, és ábrázoljuk az így kapott minta hisztogramját. Jól illeszkedő eloszlásfüggvény esetén ennek egyeneses eloszlásúnak kell lennie. Az 3.6 ábráról már látszik, hogy ez sajnos nem teljesül. Lefuttatva a Kolmogorov-Szmirnov próbát $p = 0.0222$ miatt, ugyan el kell vetnünk az egyenletességet, de a 3.2 ábrához képest igen jelentős előrelépést értünk el.

A 3.7 ábrán a BMA módszer során a kontrol tagnak kiosztott súlyokat mutatjuk be. Láthatjuk, hogy az esetek 84.56%-ban tényleges keveredés van, mindössze 23 esetben (16%) $w \geq 0.995$, kivéve 2010.11.20-át, amikor $w = 0.9873$, ez az a nap, amikor az összes ensemble előrejelzés hiányzott az adatállományból. Valóban igaz tehát az, hogy a BMA módszer által készített sűrűségfüggvény a kontrol tag sűrűségfüggvénye és a 10 felcserélhető ensemble taghoz tartozó sűrűségfüggvények egy súlyozott átlagaként áll elő. Végezetül a 3.8 ábrán az előrejelző eloszlásfüggvény várható értékének (felső két ábra, b_0 , b_1) és szórásának paramétereit (alsó két ábra, c_0 , c_1) láthatjuk. Összességében ezek a paraméterek nem változnak lényegesen a különböző napok esetén. A modellre nézve ez azt jelenti, hogy maguk a paraméterek meglehetősen stabilak és az egyes napokhoz tartozó becsült várható értékek és szórások viselkedését az ensemble tagok határozzák meg.

Összefoglalásképpen elmondhatjuk, hogy a nyers ensemble tagok BMA módszerrel történő utófeldolgozása javított az előrejelzések pontosságán. Megállapíthatjuk, hogy a 28 napos tanuló periódus optimális választásnak tűnik. A PIT hisztogram azt mutatja, hogy a BMA előrejelző eloszlás illeszkedése még nem tökéletes, a modell esetünkben apróbb módosításokra szorulhat.

4. fejezet

Összefoglaló

Az időjárási változók rövidtávú ensemble előrejelzése manapság a meteorológia egyik fő kutatási területe. Operatív ensemble előrejelző rendszer Magyarországon csak néhány éve működik (Hágel, 2010; Horányi és társai, 2011). A valószínűségi időjárás előrejelzés és azon belül az ensemble előrejelzések statisztikai utófeldolgozása szintén egy új, igen intenzíven kutatott terület (Gneiting and Raftery, 2005). Az utófeldolgozás egyik módja a Raftery és társai (2005) által bevezetett BMA módszer, ami az előrejelzendő időjárási mennyiség sűrűségfüggvényének egy becslését állítja elő ez előző néhány nap (tanuló periódus) előrejelzései és tényleges megfigyelései alapján.

Hazánkban valószínűségi időjárás előrejelzéssel először konzulenseim, Dr. Horányi András és Dr. Baran Sándor kezdtek el foglalkozni bő egy évvel ezelőtt és ebbe a munkába vontak be engem is. Jelen TDK dolgozat a szélsőbesség BMA előrejelzésével kapcsolatos tapasztaltokat ismerteti, amihez az adatokat az OMSZ bocsátotta rendelkezésünkre. A vizsgált adatállományunk 10 magyar nagyváros 42 órás ensemble szélsőbesség előrejelzéseit tartalmazza a ténylegesen mért szélsőbességekkel egyetemben a 2010. október 1. és 2011. március 25. közötti időszakra.

A dolgozat első részében egy rövid összefoglalót adok az ensemble előrejelzésekről és a különböző időjárási mennyiségek BMA modelljeiről. Ezután részletesen ismertetem a BMA módszert, megadva a modell paramétereinek becslési technikáit, valamint azt, miképpen lehet mérni az előrejelző eloszlás, illetve a belőle számolt pontelőrejelzések (átleg, medián) illeszkedését.

A dolgozat második része foglalkozik a tényleges adatelemzéssel, amit az R programnyelv ensembleBMA csomagjával (Fraley és társai, 2009, 2011) végeztünk. Itt először megvizsgáltuk, mennyire illeszkednek az ensemble előrejelzések a tényleges megfigyelésekhez és megállapítottuk, hogy az ensemble tagok szóródása nem megfelelő. Ezután 10–60 napos tanulóperiódussal ugyanarra az adatsorra elvégeztük a modellillesztést és a korábban ismertett illeszkedési mutatók felhasználásával úgy találtuk, hogy a 28 napos tanuló periódus az optimális. Végezetül részletesen elemeztük a 28 napos tanulóperiódussal a teljes adatsorra elvégzett modellillesztést és megállapítottuk, hogy a BMA sűrűségfüggvény mediánja jobb előrejelzéseket ad, mint a hagyományosan használt ensemble átlag és ensemble medián. El kell azonban mondanunk azt is, hogy habár a BMA eloszlás jobban illeszkedik a tényleges

megfigyelésekre, mint az ensemble tagok empirikus eloszlása, az illeszkedés azonban mégsem tökéletes. Ez azt jelzi, hogy a modell apróbb módosításokra szorulhat, amik későbbi vizsgálatok tárgyát képezik.

További kutatási iránya ezen kívül, hogy a módszert kipróbáljuk más időjárási mennyiségekre is, például a hőmérsékletre és a csapadékmennyiségre. A cél az lenne, hogy az ensemble előrejelzések BMA módszerrel történő statisztikai utófeldolgozását a jövőben operatív módon is használni lehessen a meteorológiai gyakorlatban.

5. fejezet

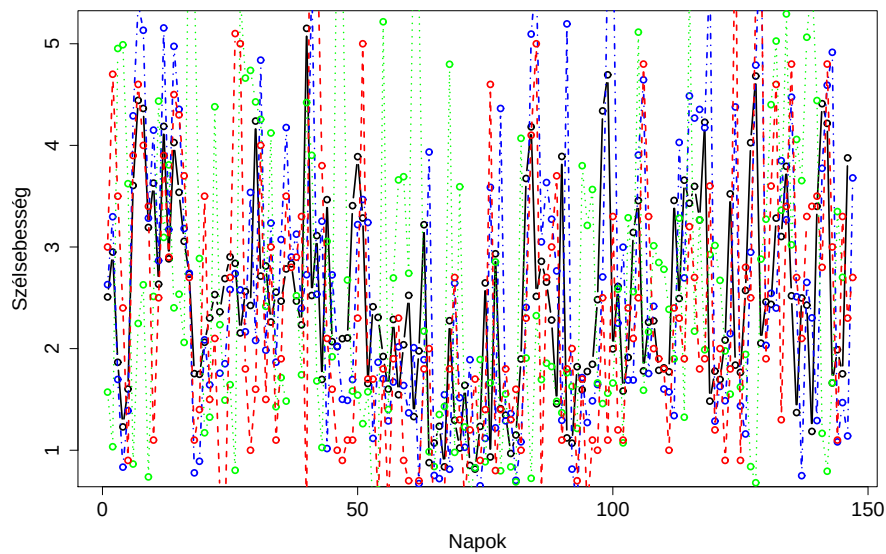
Saját munka leírása

A BMA módszer elméletének megismerése után szükség volt a módszer gyakorlati alkalmazására, azaz az OMSZ által küldött konkrét adatállományra szélsőérték előrejelzéseket adni. Az ehhez szükséges statisztikai vizsgálatokhoz mint már említettük, az R programnyelv `ensembleBMA` programcsomagját használtuk. Ez a csomag azonban nem tartalmazta az összes számunkra szükséges eljárást, így jónéhányat magunknak kellett leprogramoznunk. Ezen a feladat jelentős részét, az adatelemzést, valamint a kapott eredmények értékelésének túlnyomó részét a szerző végezte.

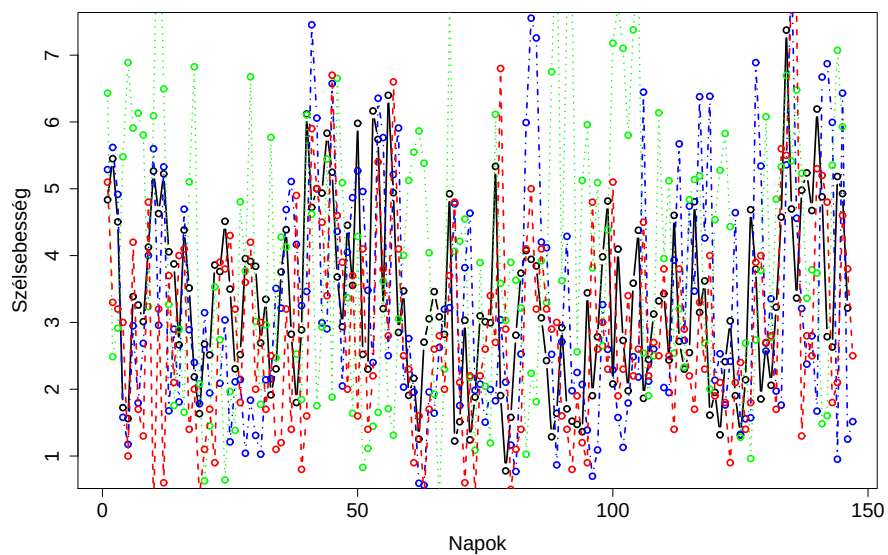
A. Függelék

További előrejelzések

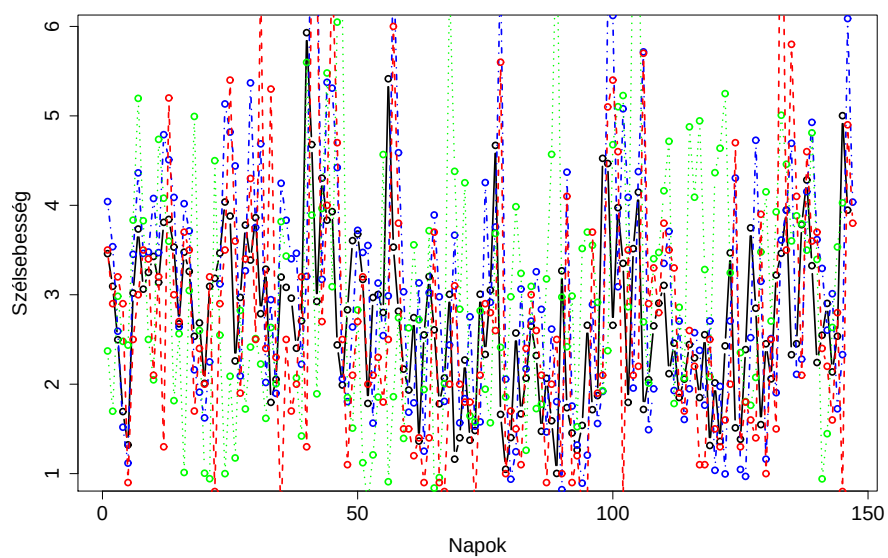
Az alábbiakban megadjuk 9 nagyváros szélsébség előrejelzéseit a 2010. október 29. és 2011. március 25. közötti időszakra. A jelölések ugyanazok, mint a 3.5 ábra esetén (BMA medián előrejelzések: fekete; a tényleges megfigyelések: szaggatott piros; az ensemble tagok mediánjai: pontozott zöld; és az ensemble tagok átlaga: szaggatott kék).



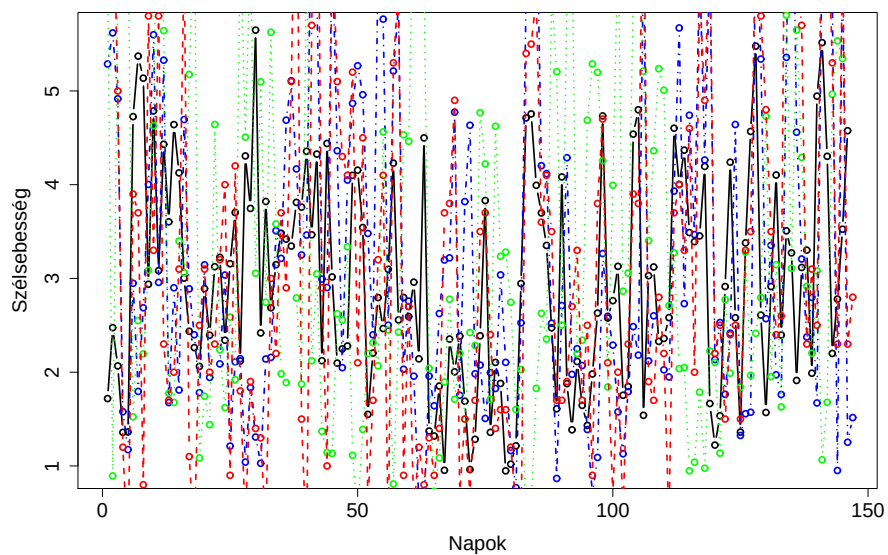
A.1. ábra. Előrejelzések Miskolcra



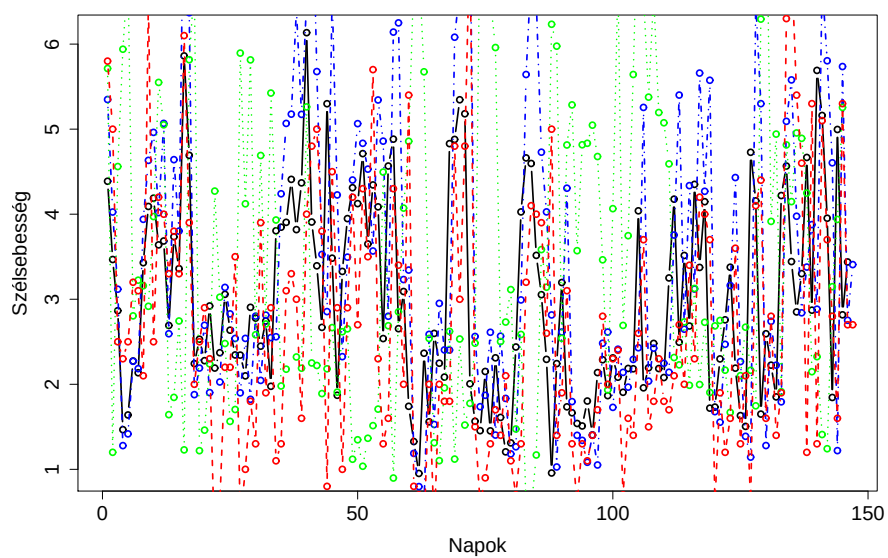
A.2. ábra. Előrejelzések Győrre



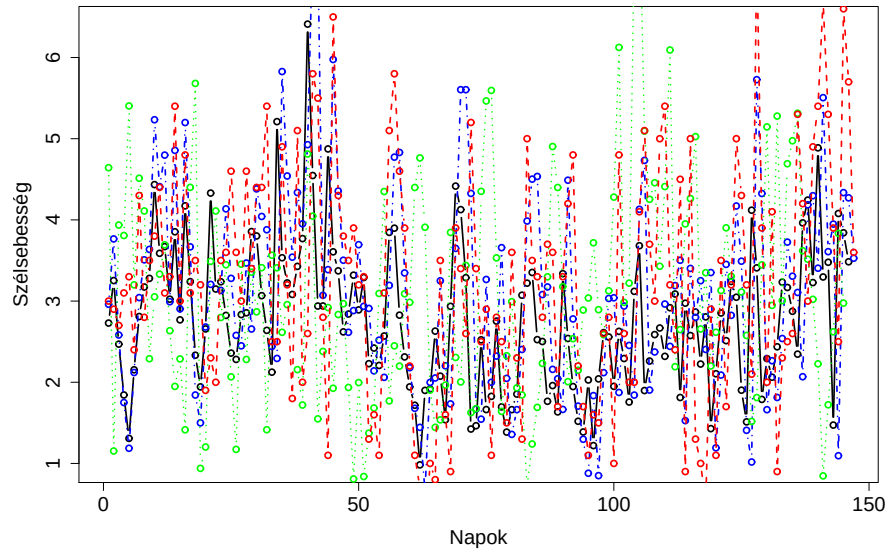
A.3. ábra. Előrejelzések Budapestre



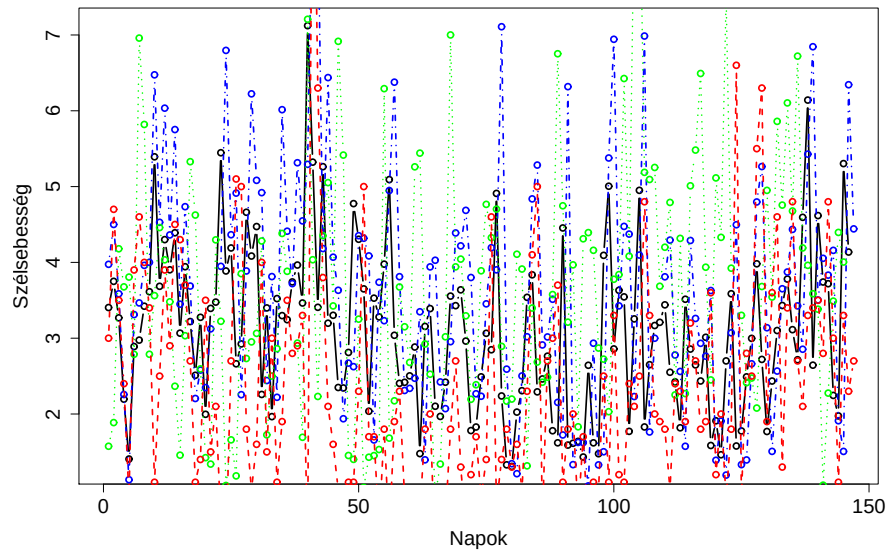
A.4. ábra. Előrejelzések Nyíregyházára



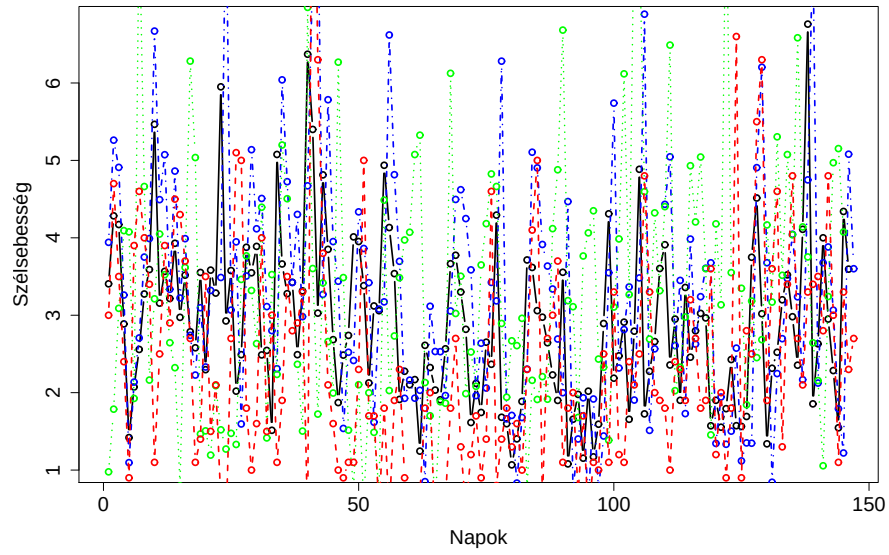
A.5. ábra. Előrejelzések Nagykanizsára



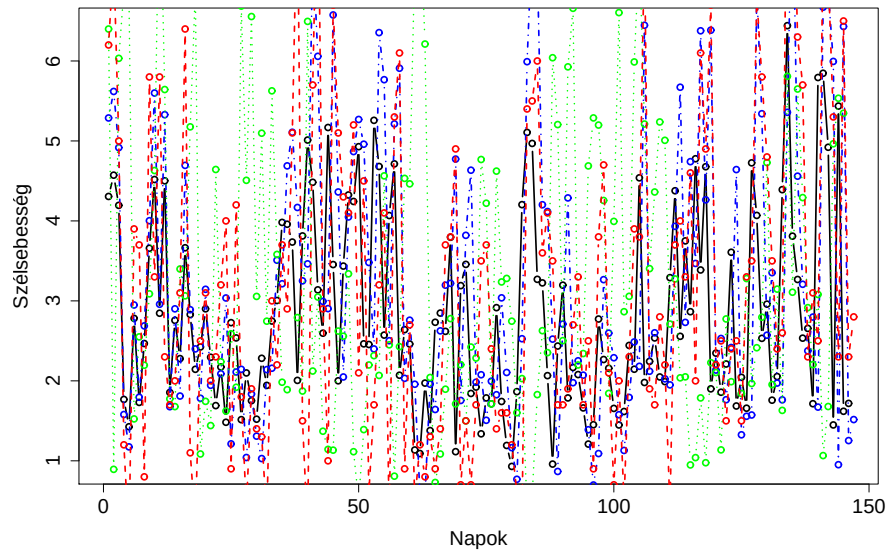
A.6. ábra. Előrejelzések Pécsre



A.7. ábra. Előrejelzések Kecskemétre



A.8. ábra. Előrejelzések Szegedre



A.9. ábra. Előrejelzések Szombathelyre

Irodalomjegyzék

- Fraley, C., Raftery, A. E., Gneiting, T. and Sloughter, J. M. (2009) EnsembleBMA: An R package for probabilistic forecasting using ensembles and Bayesian model averaging. *Technical Report* 516R, Department of Statistics, University of Washington.
- Fraley, C., Raftery, A. E. and Gneiting, T. (2010) Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Monthly Weather Review* **138**, 190–202.
- Fraley, C., Raftery, A. E., Gneiting, T. and Sloughter, J. M., Berrocal, V. J. (2011) Probabilistic weather forecasting in R. *The R Journal* **3**, 55–63.
- Giebel, G., R. Brownsword, and G. Kariniotakis (2003) The state-of-the-art in short-term prediction of wind power. Deliverable Report D1.1, Project Anemos.
- Gneiting, T. (2011) Making and evaluating point forecasts. *J. Amer. Statist. Assoc.* **106**, 746–762.
- Gneiting, T. and Raftery, A. E. (2005) Weather forecasting with ensemble methods. *Science* **310**, 248–249.
- Gneiting, T and Raftery, A. E. (2007) Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association* **102**, 359–378.
- Hágel, E. (2010) The quasi-operational LAMEPS system of the Hungarian Meteorological Service. *Időjárás* **114**, 121–133.
- Haslett and Raftery, A.E. (1989) Space-time modelling with long-memory dependence: Assessing Ireland’s wind power resource (with Discussion). *Journal of the Royal Statistical Society, series C - Applied Statistics*, **38**, 1–50.
- Horányi, A., Mile, M., Szűcs, M. (2011), Latest developments around the ALADIN operational short-range ensemble prediction system in Hungary. *Tellus A* **63** 642–651.
- McLachlan, G. J. and Krishnan, T. (1997) *The EM algorithm and extensions*. Wiley, New York.
- Raftery, A. E., Gneiting, T., Balabdaoui, F. and Polakowski, M. (2005) Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review* **133**, 1155–1174.

- Sloughter, J. M., Raftery, A. E., Gneiting, T. and Fraley, C. (2007) Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Monthly Weather Review* **135**, 3209–3220.
- Sloughter, J. M., Gneiting, T. and Raftery, A. E. (2010) Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *Journal of the American Statistical Association* **105**, 25–37.
- Tuller, S. E., and Brett, A. C. (1984) The characteristics of wind velocity that favor the fitting of a Weibull distribution in wind speed analysis. *Journal of Climate and Applied Meteorology* **23**, 124–134.
- Wilks, D. S. (2006) *Statistical Methods in the Atmospheric Sciences*. 2nd ed., Academic Press, New York.

Ábrák jegyzéke

3.1. A szélsőbesség előrejelzésének BMA sűrűségfüggvényei, az egyes csoportok sűrűségfüggvényei, az ensemble tagok, a BMA medián előrejelzés és a tényleges megfigyelések Debrecen városára a) 2010.12.30; b) 2011.03.17	18
3.2. Verification Rank Histogram	19
3.3. A szélsőbesség BMA előrejelzésének CRPS, valamint MAE és RMSE értékei (medián előrejelzés) a tanuló periódus hosszának függvényében	20
3.4. A BMA előrejelzések 67% és 90%-os lefedettsége és az előrejelző intervallumok átlagos szélessége a tanuló periódus hosszának függvényében	21
3.5. Előrejelzések Debrecenre a 2010.10.29 és 2011.03.25 közötti időszakra (BMA medián előrejelzések: fekete; a tényleges megfigyelések: szaggatott piros; az ensemble tagok mediánjai: pontozott zöld; és az ensemble tagok átlaga: szaggatott kék)	22
3.6. Pit hisztogram	23
3.7. Ensemble kontrol tagjának súlyai	23
3.8. A Gamma modell paraméterei	24
A.1. Előrejelzések Miskolcra	28
A.2. Előrejelzések Győrre	29
A.3. Előrejelzések Budapestre	29
A.4. Előrejelzések Nyíregyházára	30
A.5. Előrejelzések Nagykanizsára	30
A.6. Előrejelzések Pécsre	31
A.7. Előrejelzések Kecskemétre	31
A.8. Előrejelzések Szegedre	32
A.9. Előrejelzések Szombathelyre	32