

P81 B 163792

Oklevél sorszáma: BME-0378/2021
Intézményi azonosító szám: FI23344
MKKR szintje: 6. szint
EKKR szintje: 6. szint



OKLEVÉL

Ezen oklevél tanúsítja, hogy

Markovics Dávid

(született: Markovics Dávid, Magyarország, Veszprém, 1997. október 9.) a

Budapesti Műszaki és Gazdaságtudományi Egyetem

energetikai mérnöki alapképzési szakán tanulmányi kötelezettségeinek eleget tett,
alapfokozatot és

energetikai mérnök

szakképzettséget szerzett. A képzés ideje 7 félév.

Oklevelének minősítése: kiváló

Budapest, 2021. január 6.



dékan

MARKOVICS DÁVID
SZAKDOLGOZAT

BUDAPESTI MŰSZAKI ÉS GAZDASÁGTUDOMÁNYI EGYETEM
GÉPÉSZMÉRNÖKI KAR
ENERGETIKAI GÉPEK ÉS RENDSZEREK TANSZÉK



SZAKDOLGOZATOK

BUDAPESTI MŰSZAKI ÉS GAZDASÁGTUDOMÁNYI EGYETEM
GÉPÉSZMÉRNÖKI KAR
ENERGETIKAI GÉPEK ÉS RENDSZEREK TANSZÉK

MARKOVICS DÁVID
SZAKDOLGOZAT

Különböző gépi tanulási algoritmusok optimalizálása és összehasonlítása napelemparkok statisztikai alapú termelés előrejelzésére

Konzulens:
Mayer Martin János
doktorandusz

Budapest, 2020

Szerzői jog © Markovics Dávid, 2020.

Ide kell befűzni az eredeti feladatkiírási lapot!

NYILATKOZATOK

Elfogadási nyilatkozat

Ezen szakdolgozat a Budapesti Műszaki és Gazdaságtudományi Egyetem Gépészmérnöki Kara által a Diplomatervezési és Szakdolgozat feladatokra előírt valamennyi tartalmi és formai követelménynek, továbbá a feladatkiírásban előírtaknak maradéktalanul eleget tesz. E szakdolgozatot a nyilvános bírálatra és nyilvános előadásra alkalmasnak tartom.

A beadás időpontja:

témavezető

Nyilatkozat az önálló munkáról

Alulírott, *Markovics Dávid* (CNEJCK), a Budapesti Műszaki és Gazdaságtudományi Egyetem hallgatója, büntetőjogi és fegyelmi felelősségem tudatában kijelentem és sajátkezű aláírással igazolom, hogy ezt a szakdolgozatot meg nem engedett segítség nélkül, saját magam készítettem, és dolgozatomban csak a megadott forrásokat használtam fel. Minden olyan részt, melyet szó szerint vagy azonos értelemben, de átfogalmazva más forrásból átvettem, egyértelműen, a hatályos előírásoknak megfelelően, a forrás megadásával megjelöltem.

Budapest, 2020 december 11.

szigorló hallgató

TARTALOMJEGYZÉK

1. Bevezetés.....	1
1.1. Célkitűzés, problémafelvetés	1
1.2. Áttekintés	3
2. Szakirodalmi áttekintés/Előzmények	5
2.1. Fizikai modell	5
2.2. Statisztikai módszerek.....	5
2.2.1. Regresszív modellek.....	6
2.2.2. Lineáris regresszió	7
2.2.3. Artificial Neural Network	7
2.2.4. k-Nearest Neighbours (k-NN)	7
2.2.5. Support Vector Machines	8
2.2.6. Decision Trees	8
2.2.7. Ensemble	9
2.2.7.1. AdaBoosting regression	9
2.2.7.2. Gradient Boosting	10
2.2.7.3. Random Forests.....	10
2.3. Hibrid módszerek	11
3. Kiértékelési módszerek.....	12
3.1. Célfüggvény problémája.....	12
3.2. Bevezetett mérőszámok	15
3.2.1. Gradiens Index	15
3.2.2. Terület index.....	15
3.3. Verifikáció	16
4. Modellalkotás	17
4.1. Adatsorok értelmezése	17
4.2. Karbantartások felderítése	18
4.3. Saját modell.....	18
4.3.1. A vizsgált intervallum.....	19
4.3.2. Kritériumok	19
4.3.3. Értékadás.....	20
4.4. Python algoritmusok optimalizálása	21
4.4.1. Módszerek.....	21
4.4.2. Prediktorok	21
5. Eredmények, összehasonlítás, kiértékelés	22
5.1. Idősoros csúsztatás	22
5.2. Átlagoló hatás.....	23

5.3. Eredmények a különböző módszerekre	24
5.4. Mix módszerek	25
5.5. Terület és gradiens index kapcsolata	26
5.6. Le és felszabályozás szerinti súlyozás	28
5.7. Korreláció	28
5.8. Gradiens index, szórás	30
5.9. Az időhorizont hatása	32
5.10. Pénzügyi értékelés	33
6. Összefoglalás, eredmények értékelése	36
6.1. Összefoglalás	36
6.2. Javaslatok, következtetések, tanulságok	36
7. Felhasznált források	38
8. Summary	40

ELŐSZÓ

A fotovoltaikus (PV) energiatermelés világszerte intenzíven fejlődik, egyre nagyobb részesedést vállal a villamosenergia termelésből. A klímavédelmi célok megvalósításában a hazai Nemzeti Energia és Klíma Teroben is jelentős és egyre erősebb szereppel bír ez a technológia, így igen aktuális, erősen kutatott téma az időjárásfüggő tulajdonságából adódó stochasztikus termelésingadozások csökkentése, kiegyensúlyozása az ellátásbiztonság veszélyeztetése nélkül.

A dolgozatomban a termelés előrejelzési módszerek közül a statisztikai alapú, gépi tanulási algoritmusokat vizsgáltam, illetve egy saját eljárást készítettem, amelyek 2019-es negyedórás sugárzás, hőmérséklet, szélsebesség adatok alapján adnak becslést egy 500 kW alatti napelempark day-ahead és intraday termelésére. A statisztikai módszereknek sajátossága, hogy gyakorlatilag nem kell hozzá semmilyen ismeret az adott napelempark elhelyezkedéséről, technológiájáról, sem az energiatermelés fizikai hátteréről. Ez egy nagyon erős tulajdonság, másfelől viszont szükség van megfelelő mennyiségű megbízható adatra a korábbi üzemről, amely tehát azt jelenti, hogy újonnan épülő parkok esetében nem használhatóak tisztán statisztikai módszerek.

A munkám során nem csak magukkal az algoritmusokkal foglalkoztam, hanem igyekeztem a témát a hazai, illetve nemzetközi szabályozási rendszerek vizsgálatával körüljárni, és az eredményeket ennek megfelelően praktikusán értékelni. Céлом volt továbbá, hogy olyan módszereket találjak, amelyek alkalmazásával az egyes eljárásokkal kapott menetrendek tovább javíthatóak.

* * *

Köszönetnyilvánítás

Köszönöm Mayer Martin Jánosnak a közel 10 hónapnyi konzulensi munkáját, amely nélkül ez a dolgozat nem valósulhatott volna meg.

Köszönöm az Országos Meteorológiai Szolgálatnak az időjárás, és a Magyar Villamos Műveknek a termelés adatokat, amelyek segítségével értékes eredményeket tudtam elérni.

Köszönöm a Pannon Green Powernek, hogy szakmai gyakorlatom alatt betekintést nyerhettem a napelemparkok menetrendezésének valós működésébe, kihívásaiba.

Budapest, 2020 december 11.

Markovics Dávid

JELÖLÉSEK JEGYZÉKE

A táblázatban a többször előforduló jelölések magyarázata, rövidítések feloldása található. Az egyes mennyiségek jelölése – ahol lehetséges – megegyezik hazai és a nemzetközi szakirodalomban elfogadott jelölésekkel. A ritkán alkalmazott jelölések magyarázata első előfordulási helyüknél található.

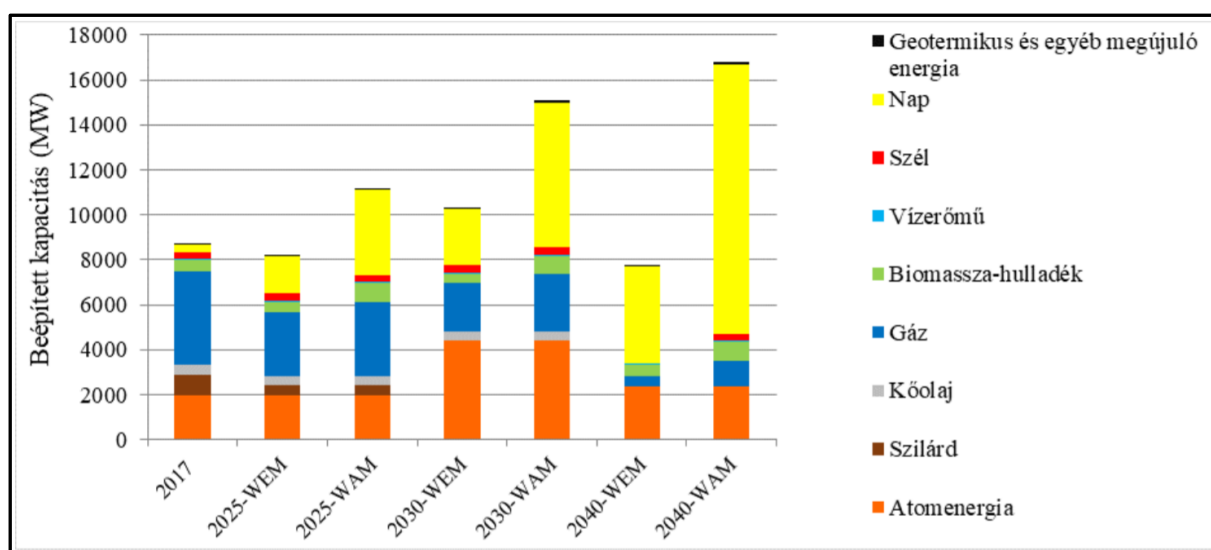
Indexek, rövidítések, idegen kifejezések	
Jelölés	Megnevezés, értelmezés
<i>adam</i>	a Python scikit-learn könyvtárk MLP neurális hálójának egyik típusa
<i>i</i>	általános futóindex (egész szám)
<i>j</i>	általános futóindex (egész szám)
mix2	az MLP_adam és SVM algoritmusok eredményeinek átlagolásával kapott menetrend
mix3	a mix2 menetrend a napi üzemidő két szélén alkalmazott csökkentő korrekcióval
sep	Terület index alapon szeparált algoritmus
ABR	Adaptive Boosting Regression
DTR	Decision Tree Regression
GBR	Gradient Boosting Regression
GI	gradiens index
k-NN (knn)	k-Nearest Neighbour
LR	Lineáris regresszió
MAE	Mean Absolute Error – átlagos abszolút hiba
MAVIR	Magyar Villamosenergia-ipari Átviteli Rendszerirányító Zrt.
MLP	Multi Layer Perception
MVM	Magyar Villamos Művek Zrt.
OMSZ	Országos Meteorológiai Szolgálat
rMAE	relative Mean Absolute Error – átlagos termelésre fajlagosított átlagos abszolút hiba
RMSE	Root Mean Square Error – a hibák négyzetes középértéke
rRMSE	relative Root Mean Square Error – átlagos termelésre fajlagosított átlagos abszolút hiba
SK	a saját készítésű modell, amelyet a 4.3. pont tartalmaz
SVM	Support Vector Machines
TI	terület index
wMAE	weighted Mean Absolute Error – az eltérések irányának függvényében súlyozott átlagos abszolút hiba

1. BEVEZETÉS

1.1. Célkitűzés, problémafelvetés

A munkám célja a fotovoltaiikus energiatermelés előrejelzési módszereinek ismertetése, ezek elemzése, olyan összefüggések után kutatva, amelyek alkalmazásával a metrendezés pontossága növelhető, illetve amelyek a megfelelő módszerek kiválasztását segíthetik. Ennek értelmében nem feltétlen arra törekedtem, hogy a legjobb megoldást a végtelenségig finomítsam, hanem arra, hogy különböző elgondolások mentén végzett módosítások, kombinációk hatásait vizsgáljam.

A téma fontosságát megalapozza az a tény, hogy a világon a PV termelés abszolút nagysága, és ezáltal részesedése az energiamixből folyamatosan nő. Az IAE statisztikái [1] alapján a 2017-es évben a világ PV termelése 443 554 GWh volt, amely a világ teljes, kb. 26 000 000 GWh-ás éves villamosenergia termeléséhez képest még elenyésző részesedést jelent. Azonban hosszú távon, a trendek alapján arra lehet következtetni, hogy a világ országainak energiamixének jelentős pillérei lesznek a beépített PV kapacitás, és a hozzá tartozó, az ingadozó termelést kiegyenlítő tartalék erőművek, energiatárolók. A trendek kapcsán érdemes hivatkozni a magyar Nemzeti Energiastratégia 2030 c. könyv 2012-es kiadásában foglalt scenáriókra [2], illetve a 2020 elején kiadott Nemzeti Energia- és Klíma Terv [3] frissített jövőképeire. Utóbbi két szélsőséges esetet tartalmaz, a WEM, amely csak a jelenlegi intézkedéseken alapul, és a WAM, amelyik további kiegészítő intézkedéseket is figyelembe vesz.



1. Ábra: az áramtermelési kapacitás-összetétel várható alakulása a WEM és a WAM forgatókönyv szerint

Mivel a kormány a PV központú scenáriókat tűzte ki céljául [4], melyhez kiegészítő intézkedések szükségesek, a WAM esetet érdemesebb vizsgálni, amely 2030-ra 6 000, 2040-re pedig 12 000 MW beépített fotovoltaiikus kapacitást jósol. Ahhoz, hogy ezt értelmezni tudjuk, a következőkben megemlítenék néhány adatot a jelenlegi helyzetről.

A MAVIR Zrt. adatai [5], alapján 2020 október 4-én új csúcs született Magyarországon az 50 kW feletti fotovoltaiikus termelést illetően. Az említett nap déli csúcsidejében a termelés elérte az 1034 megawattot, ezzel megdöntve az április 16-i [6] rekordot. Ebben az őszi negyedórán a hazai villamosenergia termelés 26,27, míg importtal együtt 22,61%-át szolgáltatta az ipari szintű napelemtermelés. A tavaszi és az új őszi rekord körülményeit a következő táblázatban foglaltam össze.

1. Táblázat: 2020 április 16. és október 4. összehasonlítása

	Rekord	Beépített teljesítmény	Nettó arány	Bruttó arány	Kiserőművek össz-teljesítménye
Ápr. 16.	942,8 MW	1016 MW	22 %	19,3 %	460,77 MW
Okt. 4.	1034 MW	1221,9 MW	26,27 %	22,61 %	508 MW

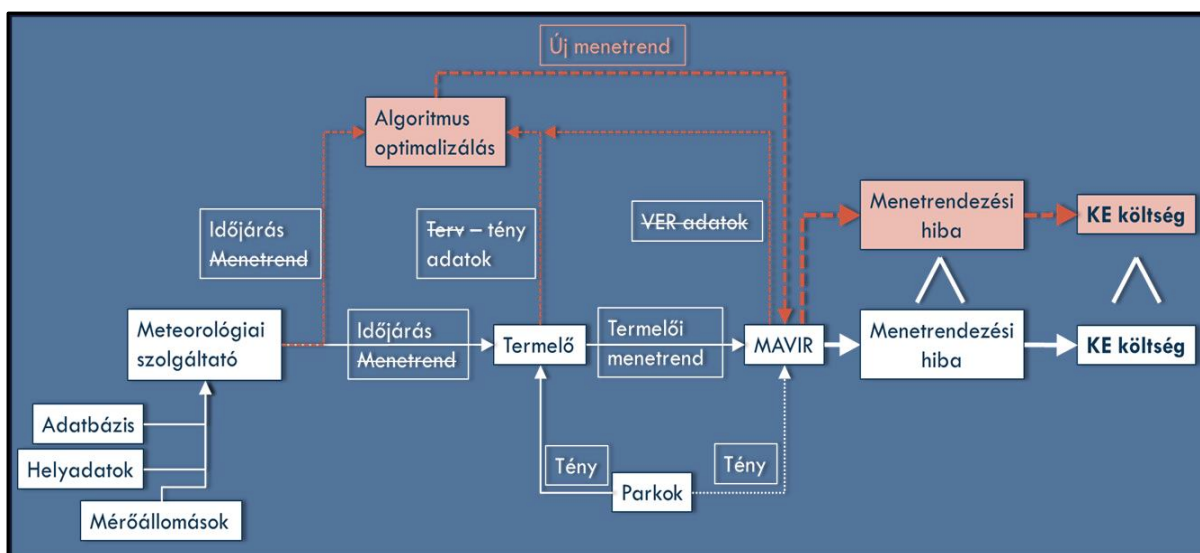
A részesedéshez egyébként hozzátartozik az a tény [7] is, hogy a COVID'19 járvány alatti villamosenergia fogyasztás alacsonyabb volt az áprilisi hónapban, mint a korábbi évek hasonló időszakában, amely egyébként elsősorban az import csökkenésében nyilvánult meg rendszer szinten.

A másik érdekes fejlemény, amelyet még meg kell említeni a várható jövőbeni hazai PV kapacitás elhelyezéséhez a villamosenergia-rendszerben és ennek értékeléséhez, az az energiamix 2020 április 5-i, dél körüli alakulása. Ugyanis ekkor a korábban említett 16-i csúcshoz hasonlóan magas PV termelés volt, a nem túl meleg, de erősen napos tavaszi napon, amely miatt még a Paksi Atomerőművet is le kellett terhelni. A MAVIR Zrt. piacműködtetési igazgatójának értékelése alapján [7], a jelentős naperőmű termelés mellett a rossz menetrendezés okozta ezt a kényes rendszerállapotot. A korábban említett Nemzeti Energiastratégiában a hazai villamosenergia termelés másik jelentős szereplője a paksi atomenergia lesz, amelyre jellemző, hogy nem képes gyors termelés változtatásokra, így, annak tükrében, hogy már nagyságrendileg 1000 MW-ot sem tudott a rendszer úgy menetrendezni és kiszabályozni, hogy a paksi blokkokra ne legyen hatással, elgondolkodtató, hogy ennek többszörösét hogyan lehet majd kezelni.

Utóbbi felvetésre nincsen egyértelmű megoldás, lehet, hogy a tervben lévő nagy felbontású és megbízható meteorológiai előrejelzések fejlesztése [3], az akkumulátoros VPP-k (virtual power plant), a szivattyús-tározós erőművek fogják megoldani az ingadozó termelés szabályozását, illetve lehet abban is bízni, hogy a térben szeparáltan elhelyezkedő telepeket érintő stochasztikus hatások (pl. felhőzettség) kiegyenlítik egymást, nem pedig erősítik egymás tranzienseit. Ennek a dolgozatnak a témája egy, az előzőektől különböző szegmens, az előrejelzési módszerek vizsgálata lesz.

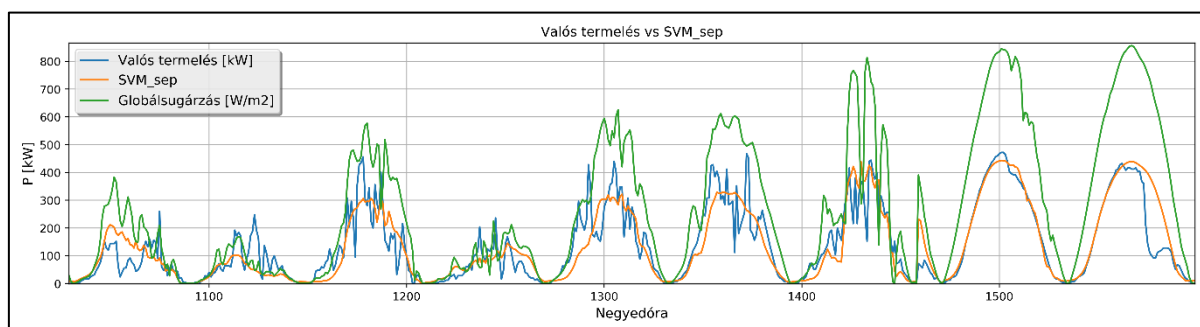
1.2. Áttekintés

A menetrendfejlesztés alábbi ábrán szemléltetett rendszerszintű elhelyezkedésének téglavörös színnel kiemelt részei képzik ennek a dolgozatnak a témáját, míg az áthúzott kifejezések olyan lehetőségeket jeleznek, amelyek jelen munkának nem képezik szerves részét, azonban más körülmények között vizsgálható aspektusok lennének.



2. Ábra: menetrendfejlesztés áttekintése, a vizsgált vonatkozások kiemelésével

Egyszerűen megfogalmazva a menetrendfejlesztés célja az, hogy az erős bizonytalansággal bíró időjáráselőrejelzésből a lehető legkisebb hibával becsüljük meg a nap-elemparkok termelését. Itt fontos leszögezni, hogy a cél nem a tökéletes menetrend, hanem annak a középútnak az megtalálása, amelyik a legkevésbé érzékeny a kiszámíthatatlan hatásokra. Szemléltetésképpen a következő diagramon idősorosan ábrázoltam a jósolt sugárzást, az egyik saját menetrendet és az ezekhez tartozó valós termelést.



3. Ábra: a menetrendezés célja (valós termelés, menetrend, sugárzás)

A következőkben a napelemparkok termelés előrejelzésére vonatkozó módszereket fogom ismertetni, majd az értékelési szempontokat, illetve azt az eljárást, amelyet felépítettem egy valós parkra optimalizálva. Majd a Python scikit learn könyvtárának gépi tanulási algoritmusainak eredményeit fogom bemutatni, összegezve a tanulságokat, amelyet a kutatási munkám során tapasztaltam.

2. SZAKIRODALMI ÁTTEKINTÉS/ELŐZMÉNYEK

2.1. Fizikai modell

A termelés előrejelzési módszerek legkézenfekvőbb típusa a fizikai modell [8] alkotáson alapul. Ennek lényege az, hogy a Nap – napelem - hálózat rendszerben lejátszódó fizikai folyamatokat matematikailag megfogalmazva a bemenő időjárási adatokból a modell megadja a várható termelési értéket. Az ilyen módszereket indirektnek is szokás nevezni, amely arra utal, hogy a modell elsősorban a napsugárzás várható alakulását fogja megadni, amelyből aztán a fizikai rendszer modellje alapján következtethető a teljesítmény.

A fizikai modell megalkotásához nincs szükség korábbi üzemre és abból származó adatokra, ezáltal nem tartalmazza a napelemparkok, és a helyi adottságok tapasztalati jellegű tulajdonságait. Tehát az ilyen predikciós eljárásokhoz kiemelt fontosságú az adott napelemtípus karakterisztikájának minél pontosabb gyári meghatározása. Szükség van továbbá az elhelyezkedésből adódó geometriai jellemzők (óraszög, deklináció, zenit- és azimutszög) meghatározására, illetve ismerni kell a környező árnyékoló hatásokat.

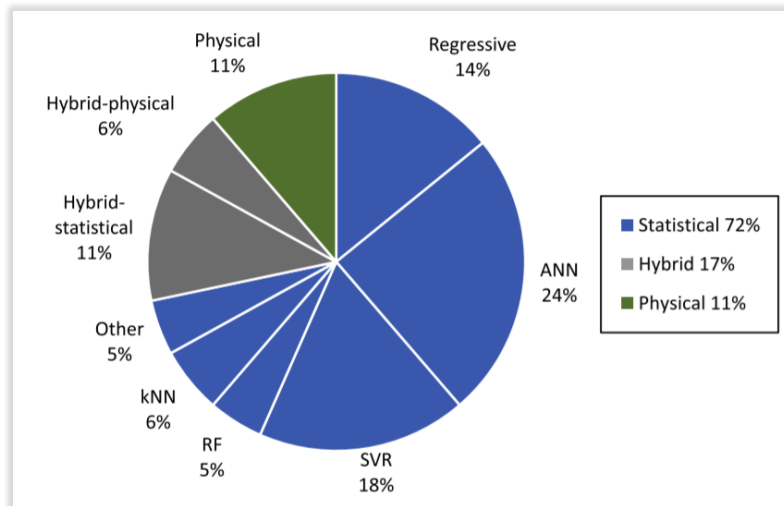
Mindezek alapján meg kell határozni az adott napelemparkhoz tartozó *napsugárzás modellt*, azaz a nappályáját, amely alapján meg lehet adni a várható horizontális sugárzásnak a napelemek síkjába eső részét. Továbbá fontos elkészíteni a *cellahőmérséklet modellt*, mivel a teljesítmény szempontjából fontos tényező a berendezések melegezése, az energiaátalakítás hatásfokára gyakorolt jelentős hatása miatt. A hálózati csatlakozásig hátralévő veszteségek mértékére a napelem belső *áramköri modellje* és a DC-AC átalakításhoz kapcsolódó *inverter modell* alapján következtethetünk.

Mivel a fizikai modelleknek egy bemenete van, a meteorológiai szolgálat által biztosított adatsor a várható környezeti jellemzőkről, jobban ki vannak téve ennek pontosságának, mint azok a módszerek, amelyek ezeket az előrejelzéseket a korábbi statisztikák alapján értékelik, és ezáltal esetleg konzervatívabb becsléseket adnak.

2.2. Statisztikai módszerek

A statisztikai módszerekhez nincs szükség semmilyen információra a berendezés elhelyezkedéséről sem a benne lejátszódó pontos fizikai folyamatokról. Az ilyen módszerek a múltbeli adatok alapján fogalmazzák meg explicit összefüggéseket az időjárás és termelés között, amelyek segítségével majd az meteorológiai időjárás előrejelzésből meg lehet becsülni a napelempark várható termelését. Ezek alapján érzékelhető, hogy az ilyen eljárásokhoz szükséges, hogy a park már megfelelő ideje üzemeljen, és erről az időszakról hiteles adatok legyenek. Ugyanakkor nagy előnye, hogy a mérési, vagy meteorológiai előrejelzéssel kapcsolatos hibákat ki lehet szűrni. A témában született

legtöbb publikáció az ilyen statisztikai módszerekkel foglalkozott, ebből következik a relevanciája. Az egyes típusok arányait a következő ábra foglalja össze egy 2016-ban készült kutatás [9] nyomán:



4. Ábra: a kutatások megoszlása az egyes módszerek szerint

A következőkben a teljesség igénye nélkül néhány átfogó modellt, illetve azokat a statisztikai módszereket fogom ismertetni, amelyek gyakran előfordulnak a vonatkozó irodalomban.

2.2.1. REGRESSZÍV MODELLEK

A regresszív eljárások a függő (pl. sugárzás) és néhány független (prediktor) változó kapcsolatát írják le a napelempark teljesítményével. Az idősorok elemzése szempontjából megkülönböztethetők a lineáris, nem-lineáris és stacionárius (egy állandó érték körül fluktuáló), vagy tranzienst állapotos.

A *lineáris stacionárius* modellek közé tartoznak az Auto-Regresszív (AR) eljárások, amelyek a kimenetet a prediktorok lineáris kombinációjaként adják meg, illetve az egyszerű Moving Average (MA) módszerek, amelyeket akkor használunk, amikor az adatok állandó varianciával helyezkednek el egy egyensúlyi érték körül. A kettő keverékéből kapjuk az Auto-Regressive Moving Average (ARMA) modelleket, amelyek mind a késleltetett múltbeli adatokat, mind a hibákat figyelembe veszik.

Az AR modelleket kiegészítve exogén adatokkal kapjuk az ARX családot, amelyet bővítve a mozgó átlagokkal adódik az ARMAX rövidítéssel elnevezett eljárás, amely bevezeti az externális változókat ugyanabba az idősoros analízisbe.

A *lineáris tranzienst (nem-stacionárius)* modellek kategóriájába sorolandóak az Auto-Regressive Integrated Moving Average (ARIMA) eljárások, amelyek az AR és MA

komponensek úniójaként állíthatóak elő, illetve a Seasonal ARIMA módszerek, amelyek bevezetik a szezonális függést a rendszerbe.

Léteznek továbbá *nem-lineáris stacionárius* módszerek, ilyen módszer például a Non-linear AR-exogenous (NARMAX) modell.

2.2.2. LINEÁRIS REGRESSZIÓ

Az elemzés során használt első és legkézenfekvőbb módszer a lineáris regresszió, amely lineárisan független prediktorok értékének lineáris kombinációjaként ad becslést, a változók együtthatóit a legkisebb négyzetes hibák elvén optimalizálva. A lineáris regresszió sokszor referenciaként szolgál, különböző helyszínek, szolgáltatók, vagy módszerek gyors összehasonlítására, értékelésére alkalmas.

2.2.3. ARTIFICIAL NEURAL NETWORK

Ahogy az 3. ábra is mutatja, a mesterséges neurális hálókat (ANN) alkalmazó módszerek a leggyakrabban előfordulóak a fotovoltaiikus teljesítmény előrejelzési kutatások közül. Az ilyen módszereket a neuronok működése inspirálta, a lényege az, hogy a neuronok egy csoportja összekapcsolva egy neurális hálót alkot (NN). A kapcsolatokat numerikusan súlyokkal látjuk el, melyek végső értéküket a tanulási fázisban nyerik el. Az ezen az elven alapuló eljárások széles körben alkalmazhatóak, ahol nagy mennyiségű adatot kell kezelni. Különböző topológiák léteznek, amelyek alapján csoportosítani lehet az ARR eljárásokat. Ezeket legfőképp a rejtett rétegek száma alapján szokás osztályozni, léteznek egy (simple perception) és többrétegű (multi-layer) megoldások. A neurális hálók közül a Multi-Layer Perceptron (MLP) típusú modellek emelhetők ki, melyek a különböző függvények általános közelítéseként széles körben alkalmazhatónak bizonyultak. Ugyanakkor számos másik módszer is létezik, ilyen pl. az Adaptive Neuro-Fuzzy (ANFIS), vagy az Adaptive Resonance Theory (ART).

2.2.4. K-NEAREST NEIGHBOURS (K-NN)

A k-NN az egyik legegyszerűbb gépi tanulási módszer, alapvetően minta felismerésen alapul, a lényege az, hogy az adott pillanatbeli állapotot összehasonlítva a korábbi adatokkal, amelyeket a tanuláshoz használunk, kiszámítja az euklidészi távolságokat, és az első k darab „szomszéd” alapján következteti a kimenetet.

2.2.5. SUPPORT VECTOR MACHINES

Az SVM egy ellenőrzött módszer, amelyet elsőként Vapnik és Lerner (1963) mutattak be, majd a később Cortes és Vapnik (1995) fejlesztett tovább, hogy használható legyen osztályozási problémák megoldására.

Regressziós problémák kapcsán használva a Support Vector Regression machines (SVR) néven ismertek. A többi módszerhez képest erősségük, hogy kiemelkedő általánosító képességgel bírnak, illetve jól kezelik a nem-lineáris rendszereket. Több lineáris regresszió együtteseként funkcionálnak átalakított prediktorokkal, úgy, hogy közben egyszerű, nem túl komplex rendszert alkotva jól illeszkednek az adatokra.

Három domináns paraméter van, amelyek finomhangolása elengedhetetlen a kívánt teljesítményhez. Az első az ϵ , amely a pontosság szempontjából használatos, a második az érték (cost) paraméter (C), amely a pontosság és a komplexitás közti kompromisszumért felelős, a harmadik pedig a γ , amely a magok működését szabályozza, illetve ez felelős a prediktorok magasabb rendű tulajdonság térbe való átalakításáért.

2.2.6. DECISION TREES

A döntéshozó fák módszere egy egyszerű, a mindennapokban is gyakran használt algoritmus, amely a következő két eljárás alapjaként is szolgál. Röviden összefoglalva lényege az, hogy a döntési folyamat a kiindulási kérdéstől egészen az eredményig a prediktorok függvényében elágazások sorozatán keresztül halad át. Minden elágazásban egy-egy kérdést fogalmazunk meg, amelyre két válaszlehetőséget adva az összes lehetőséget le kell tudnunk fedni. Továbbá fontos, hogy a kérdések olyanok legyenek, amelyek szűkítik a potenciális kimenetek számát. A döntéshozó fákra egy egyszerű példa az intervallum felezéses módszer, amellyel általános iskolában tanítanak, hogy kitaláljuk, milyen számra gondolt a másik egy 1-től 100-ig tartó skálán, úgy, hogy csak igennel és nemmel válaszolhat a kérdésre. Valamint, majd később a 3. ábrán látható hazai kiegyenlítő energia elszámolási rendszer is jól illusztrálja a fák felépítését.

Azok a gépi tanulási módszerek, amelyek ilyen fa típusú felépítést használnak, a betanítás fázisában a prediktorokat értékük szerint csoportosítják, amelyek így egy levelet alkotnak, kimenetként pedig a levélhez kapcsolódó prediktorokhoz tartozó értékek átlagát adják. A megfogalmazott problémától (törzs) az elágazásokban feltette kérdéseken (ágak) keresztül jutunk el a megfelelő levélhez és annak kimeneti értékéhez. A levelek pontos kiosztása egy optimalizációs folyamat eredményeként, hibaminimalizálással adódik. Többváltozós esetekhez komplexebb struktúra tartozik.

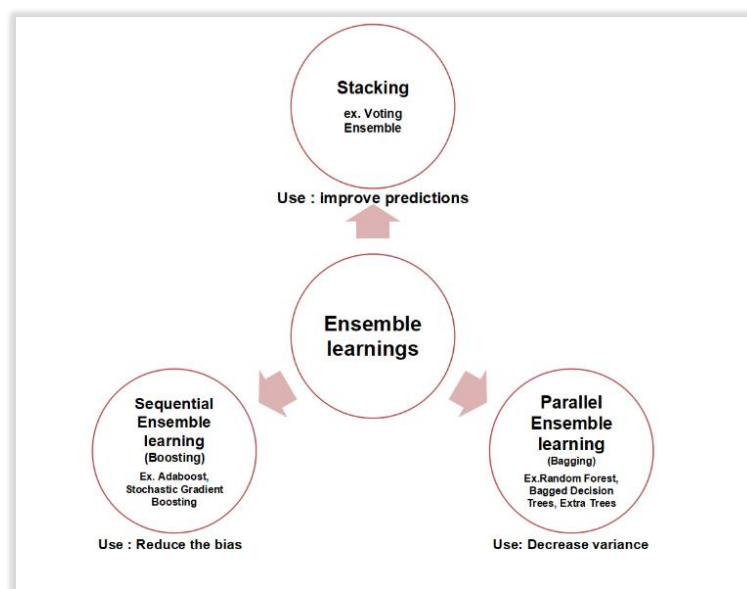
Az ilyen eljárások előnyei kapcsán többek között megemlítenéd [10] az egyszerűsége, az, hogy különböző adattípusokra használható, valamint az adatsorok zavaró hatásait is jól kezeli. Ezzel szemben hajlamos túl komplex rendszereket alkotni, amely

nem általánosítja megfelelően az adatokat, illetve van, hogy a rendszer stabilitása nem megfelelő, ilyenkor az adatokban bekövetkező kis változás is egészen más fát eredményez. Ezt hívják varianciának, amelyet a *bagging* és a *boosting* módszerekkel orvosolhatunk.

2.2.7. ENSEMBLE

An Ensemble tanulási módszerek közös tulajdonsága, hogy több különböző fa alapú algoritmust kombinálnak, ezáltal jobb előrejelzést adnak, mint az egyedi megoldások. Emögött a kijelentés mögött az az elv áll, hogy egy csoport gyenge tanuló összeállva erős tanulási képességgel bír, ezáltal javítható a modell pontossága.

Az Ensemble technikáknak három fő csoportja van, amelyet a következő ábra [11] szemléltet:



5. Ábra: Ensemble módszerek csoportjai

Tehát a fő típusai a *stacking*, *boosting* és a *bagging*, melyek közül utóbbi kettő kapcsán fogom bemutatni a releváns módszereket.

2.2.7.1. AdaBoosting regression

Az AdaBoosting, azaz Adaptive Boosting volt az első sikeres [12] alkalmazása annak az általános felvetésnek, amelyre ezen módszerek elvi szinten épülnek, miszerint különböző közepesen, vagy erősen pontatlan módszerek kombinációjával megfelelő predikciót kaphatunk. Az AdaBoosting gyenge tanulói alapvetően olyan döntéshozó fák, amelyek csak egy választással bírnak, ezért csak „tuskó”-nak nevezzük őket. Az algoritmus súlyozza a megfigyeléseket úgy, hogy a nehezebben klasszifikálható elemeket

nagyobb, míg a már így is pontosakat kisebb értékekkel látja el. Ezt a folyamatot az iterálás során újra és újra elvégezve a nehezen klasszifikálható elemek egyre növekvő súlyokat kapnak, amíg az algoritmus meg nem találja azt a modellt, amely megfelelően azonosítja ezeket a mintákat is. Maga a predikció a gyenge tanulók többségi szavazatának pontossággal súlyozott átlagaként áll elő.

Mostanra az ezt a módszert már nem csak fa típusú, hanem bármilyen algoritmus, akár neurális hálók javítására lehet használni.

2.2.7.2. *Gradient Boosting*

A Gradient Boosting, az az AdaBoosting általánosításaként jött létre Breiman [13] munkásságának eredményeként, aki megalkotta az arcing ígét, amely mozaikszó lényegében az adaptív újrásúlyozást és kombinálást jelenti. Az arcing algoritmusban minden lépés egy súlyozott minimalizálásból és az osztályozók, valamint a súlyozott prediktorok újraszámításából áll. Ezt a módszert Friedman fejlesztette tovább, tőle származik a gradient boosting machines elnevezés, amelyre későbbiekben csak gradient boostingként hivatkoznak. A kapott modell három részből áll:

- ¶ Hibafüggvény, amely a probléma típusának függvényében meghatározandó, lehetek standard (pl. regresszió esetén négyzetes hibák összege, míg klasszifikációra logaritmikus hiba), de egyedi esetekre definiálható saját célfüggvény.
- ¶ Gyenge tanuló, amely alapesetben döntéshozó fa algoritmus.
- ¶ Additív modell – egyszerre egy fát adunk hozzá a modellhez úgy, hogy a már meglévőket a továbbiakban nem módosítjuk.

2.2.7.3. *Random Forests*

A módszert elsőként Breiman (2001) fejlesztette ki, maga az eljárás döntéshozó és regressziós fák együttes használatán alapul, az egyes egyedi fák átlagos előrejelzését adja eredményül. Gyakori megoldás ugyanarra az esetre több, különböző fát illeszteni, mivel ezek eredményeinek átlagolásával lehetséges a variancia csökkentése, az enyhén növekvő befolyásoltság ellenére, ezáltal összességében a modell javítható.

A Random Forest elnevezést a struktúrája alapján kapta, melynek lényege az [14], hogy az egyes döntéshozó fák a kérdések csak egy véletlenszerűen kiválasztott csoportjára válaszolnak, a tanító adatsorok szintén véletlenszerűen kiválasztott része alapján. Ezáltal növelhető az erdő diverzitása, ami a predikciót összességében erősebbé teszi. Mivel az egyedi fák alapvetően ugyanarról az adatsorról tanítva, a zavaró tényezők, vagy rossz adatok alapján ugyanazokat a hibás következtetéseket vonhatják le ezt a nagyfokú korrelációt lehet csökkenteni a fenti módszerrel.

2.3. Hibrid módszerek

A hibrid módszereket ugyanaz az elv ihlette, mint az Ensemble-t, mely szerint a különböző eljárások kombinált használata javítja a teljesítményt, mivel így az egyes területek erősségei kiegészítik egymást. Kétféle megközelítés létezik a témában, az egyik a hibrid-statisztikai, a másik pedig a hibrid fizikai. Előbbi kettő vagy több különböző statisztikai modell összekapcsolásán alapul, utóbbi pedig a fizikai és statisztikai modellek fúzióját jelenti. Számos kutatás jelent meg ebben a kategóriában is, amelyek az általuk használt kombinációkkal rendre jobb teljesítményt nyújtottak, mint az egyedülálló technikák.

3. KIÉRTÉKELÉSI MÓDSZEREK

3.1. Célfüggvény problémája

A predikciós eljárás pontosságára többféle módszert lehet alkalmazni. A gépi tanulási módszereknél a két leggyakrabban használt mérőszám a MAE (mean absolute error) és az RMSE (root mean square error) [15].

Az MAE érték az abszolút hibák átlagos értékét jelenti, a következő képlettel számítható:

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$$

Ahol:

- ¶ N az elvégzett becslések elemszáma, a vizsgált napok karbantartásmentes, nullánál nagyobb sugárzási intenzitású negyedóráinak száma,
- ¶ y_i az i -edik negyedóra jósolt termelés
- ¶ $\hat{y}_i$ az i -edik negyedóra valós termelése

A képletből látható, hogy a MAE egyenlő súllyal vesz figyelembe minden hibát, nem tesz különbséget nagy-kicsi és negatív-pozitív irányú eltérésekkel szemben.

Ezzel szemben a RMSE a becslés átlagos négyzetes hibáját adja meg, a következő képlettel számítható, az előzőekkel megegyező jelölésekkel:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}$$

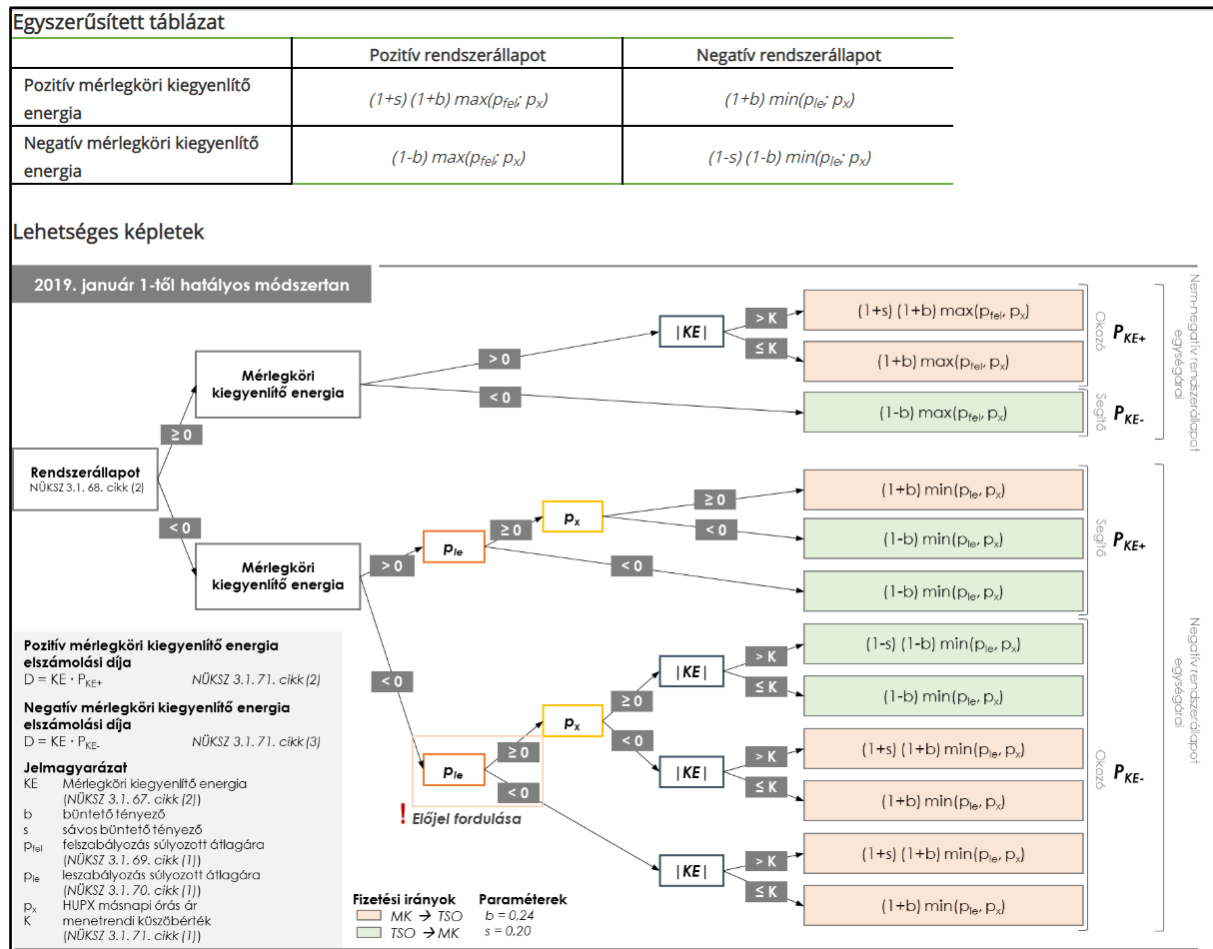
Azáltal, hogy az egyes eltérések négyzetre vannak emelve, az RMSE jobban bünteti a nagyobb hibákat, mint a kisebbeket. Hasonlóan az MAE-hoz az előjelek között nem tesz különbséget ez a módszer sem.

A kiértékelés során az ismertett hibáknak az átlagos termelésre vonatkoztatott értékeit használtam, melyekre relatív hiba (rMAE) és rRMSE-rént fogok hivatkozni az alábbi képletek alapján számítva:

$$rMAE = \frac{MAE}{\bar{y}} \quad rRMSE = \frac{RMSE}{\bar{y}}$$

Lényeges kérdés, hogy melyik módszer számít a legértékesebbnek. Erre alapvetően nincs egyértelmű válasz, mindig a cél alapján kell dönteni. A napelemparkok termelés

előrejelzésének pontossága kapcsán a legjobb mérőszám a forintban fizetett kiegyenlítő energia költsége. Ennek számítására a MAVIR honlapján [16] található leírásokat és a megértést segítő táblázatokat.



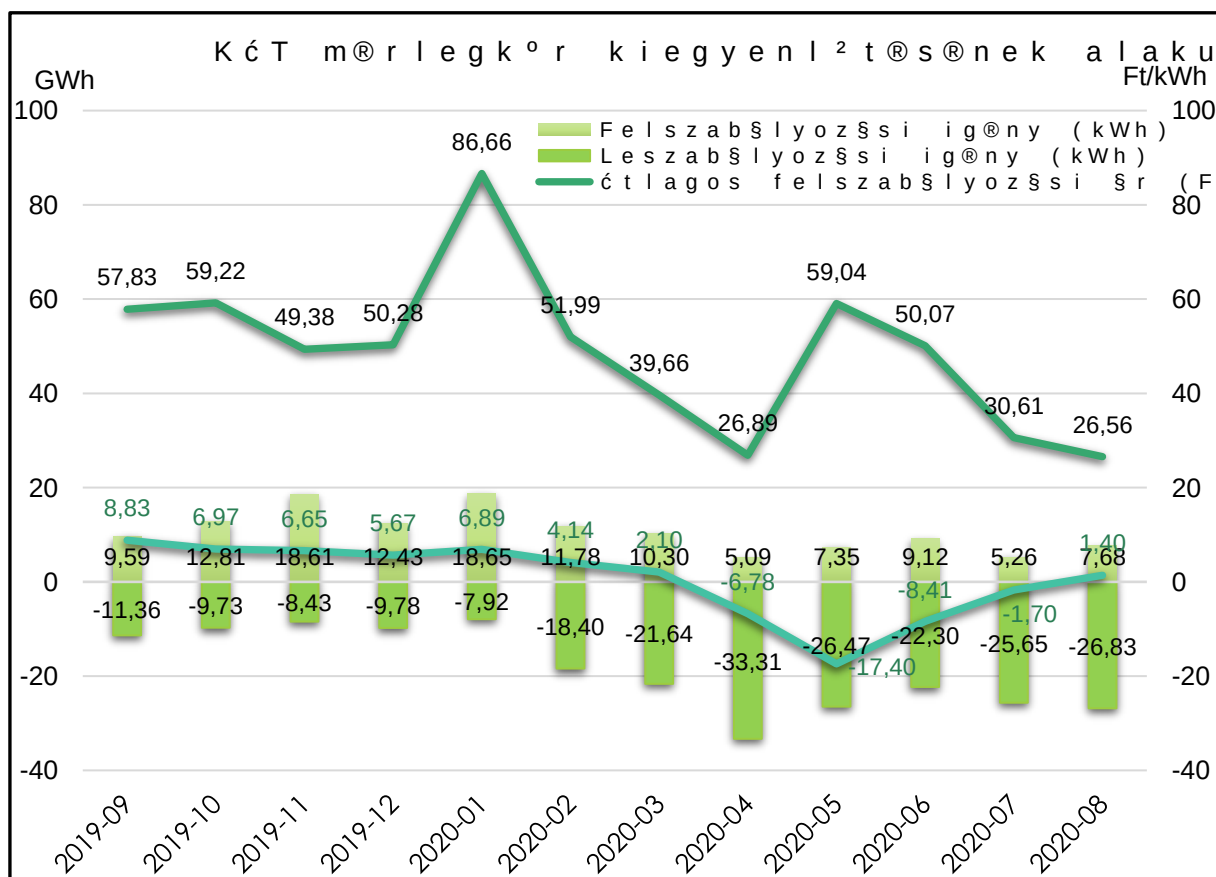
6. Ábra: kiegyenlítő energia egységköltség számításának módszertana

A fenti ábra által szemléltetett módszertan 2019. január 1-től érvényes. A kiegyenlítő energia egységárának meghatározásához több szempontot is figyelembe kell venni. Az elszámolás megkülönbözteti mérlegkör menetrendtől való eltérése okozta aktivált kiegyenlítő szabályozási energiának a segítő oldal rendszerállapotához viszonyított irányát, az eltérés nagyságát, és a segítő oldal adott irányra vonatkozó egységárait. Továbbá megkülönbözteti a büntető (b) és a sávós büntető (s) tényezőt, amelyek közül utóbbi a kiugró (egy adott küszöbérték feletti) értékeket súlyosbítja.

Ezen ábra mentén látható továbbá az, hogy két különböző időpontban indukált azonos irányú és nagyságú menetrendezési hiba termelői költsége nem csak a piaci árak függvényében változik, hanem nagyban függ a mérlegkör többi engedélyesétől és a teljes rendszer állapotától. Hiszen, ha például egy KÁT-os napelempark üzemeltető a

menetrendezési hibája *okozó*, azaz hozzájárul a teljes mérlegkör állapotának adott irányú hibájához, és a mérlegkör hasonlóképpen okozza a teljes rendszer hibáját, akkor a hónap végén erre a negyedórára fizetni kell majd, míg, ha az okozott mérlegköri eltérés *segítő* a teljes rendszerállapotra nézve, akkor a rendszerirányító fizet a mérlegkörnek, amely a havi elszámolásban utóbbi szempontjából negatív előjellel szerepel majd.

Mindezek tükrében a Ft alapú előrejelzéshez becsülni kellene a teljes hazai rendszer menetrendeltérésének irányát, a vonatkozó mérlegkör eltérésének irányát, és még ezek mellett az egységköltséghez szükséges hazai és nemzetközi piacon aktivált kiegyenlítő energia egységköltség várható értékét. A rendszer összetettségét és az adatok elérhetőségét figyelembe véve egyelőre nem reális Ft-ra optimalizálni, főleg annak tükrében, hogy az elszámolást havi összesítéssel kapják az engedélyesek, így tehát új módszerre váltva egy ideig visszacsatolás nélkül kell menetrendezni.



7. Ábra: KÁT mérlegkör kiegyenlítésének alakulása 2020. augusztusig

A MAVIR adatpublikációjából [17] látható, hogy a felszabályozási árak átlagosan jóval alacsonyabbak, mint a leszabályozásé. Emiatt a termelők rendre alulmenetrendezik az erőműveiket, amely a szabályozási igények mértékéből látszik. Az új rend-

szerre átváltozva az első negyedévben a leszállítási igény elkezdett nőni, majd áprilisban tetőzött, az árak azonban igazán csak májusban szálltak el, amikor a piac reagált a tudatos alul menetrendezésre. A két irány közti árrés csökkentésével a pénzügyi és fizikai optimum közelíthető lenne egymáshoz.

Az eljárások pénzügyi optimalizálása ennek a dolgozatnak nem tartalma, a korábban említett két (RMSE, MAE) módszert fogom alkalmazni.

3.2. Bevezetett mérőszámok

Az előrejelzések fejlesztése során felmerült az igény arra, hogy valamilyen számszerűsíthető mennyiség alapján el lehessen dönteni egy adott nappól, hogy menetrendezési szempontból mekkora kihívás lesz, milyen várható hibával lehet majd megjósolni a termelést. Erre a célra két olyan mérőszámot vezettem be, amelyek használatára a vonatkozó szakirodalomban nem találtam példát. A következőkben ezeket fogom bemutatni.

3.2.1. GRADIENS INDEX

A gradiens index a következő képlettel számítható egy adott napra a valós termelés értékelésére:

$$GI_i = \frac{B \tilde{u}_i - \tilde{u}_{i-1} \cdot S_i}{B \tilde{u}_i}$$

ahol $\tilde{u}_i$ az adott nap i -edik napsütöttségi negyedórájának teljesítménye, de lehetne akár globálsugárzás is, ha különböző szolgáltatói előrejelzéseket hasonlítanánk össze.

Ez a mutatószám azt hivatott reprezentálni, hogy egy adott nap termelése mennyire ingadozó, amelyből következtethetünk arra, hogy mekkora lesz a relatív hibája az előrejelzéseknek. Mivel az időjáráselőrejelzés, vagy a szolgáltatótól függően konkrétan a termeléselőrejelzés általában már egy átlagolt, hibára optimalizált adatsor, jóval kevésbé fluktuál, ezért az időjárás előrejelzéseket és a valós termelést nem érdemes GI alapján összehasonlítani, csak azonos célú adatsorok esetén érdemes ezt a mérőszámot alkalmazni.

3.2.2. TERÜLET INDEX

A terület index a következő képlettel számítható egy adott napra a várható termelés ingadozás értékelésére:

$$TI_i = \frac{B \tilde{u}_i}{\tilde{u}_{i-1} \cdot S_i}$$

ahol $\tilde{U}_i$ az adott nap i -edik napsütötte negyedórájának jósolt felületi globálsugárzás értéke W/m^2 -ben, $\tilde{U}_{a,j}$ a napelempark maximális teljesítménye, $\tilde{\epsilon}_j$ pedig az adott nap vizsgált negyedóráinak száma. Ennél a mutatószámnál a nyári magasabb sugárzásértékeket kompenzálja a napsütötte órák számának növekedése.

Ezt a mutatószámot arra használtam, hogy az átlagoló jellegű előrejelzésből kiszűrjem a felhős napokat. Egy határérték alatt felhős, felette pedig napos kategóriába sorolva, a két típusra külön-külön betaníthatóak az algoritmusok, ezáltal elérhető, hogy a menetrendezési szempontból ideális napok ne legyenek a túlzott átlagolás miatt alulbecsülve. Különböző méretű parkok esetén a termelést érdemes normálni egy fix értékre.

3.3. Verifikáció

Ahogy a 3.1.-es pontban korábban kifejtettem, a különböző eljárások minőségére, a becslések jóságának meghatározása nem egyértelmű. Emellett még kevésbé triviális különböző lokációkra, PV technológiára és időjárás szolgáltatóra vonatkozó kutatások eredményeinek összehasonlítása, mivel erre vonatkozóan nem létezik széles körben elfogadott és használt viszonyítási alap [18]. Ennek oka többek között az is, hogy különböző régiókban eltérő lehet a szabályozási környezet mind a napelemparkok által termelt energia átvételére, mind a kiegyenlítőenergia elszámolási rendszerére vonatkozóan. Így tehát a világ különböző részein más-más célfüggvényre optimalizálnak a témában.

Ebben a kérdésben egy olyan vizsgálatot végeztem el, amely a menetrendezési hibákat az eltérés irányától függő súlyokkal veszi figyelembe. Ez a hibaszámítás azt hivatott bemutatni, hogy a különböző területeken jellemző szimmetrikus (pl. USA) vagy aszimmetrikus (pl. Magyarország Spanyolország, Portugália) árazás milyen hatással van a menetrendezés céljára [19], illetve mely módszerek azok, amelyek jól, és melyek azok, amelyek rosszul reagálnak a fel- és leszabályozási ár közti rés növekedésére. Az árkülönbség érthető, hiszen felszabályozás esetén a rendelkezésre állás mellett az energia díja is kifizetésre kerül, míg leszabályozás esetén utóbbival nem kell számolni, azonban ez a többletköltség amerikai mintára beépíthető a szimmetrikus árba is. A következő képlettel számítottam a súlyozott rMAE, azaz wMAE értékeket:

$$wMAE = \frac{\sum_{i=1}^Q \sum_{j=1}^R \frac{\rho}{\tilde{U}_{a,j}} \left(\tilde{U}_i - \tilde{U}_{a,j} \right)}{\sum_{i=1}^Q \sum_{j=1}^R \frac{\rho}{\tilde{U}_{a,j}}}$$

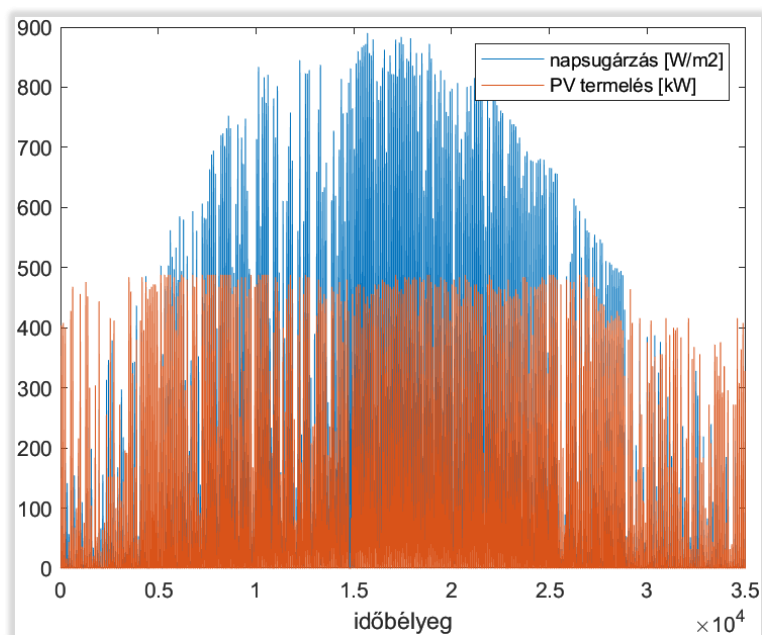
Ahol:

- ¶ $\tilde{U}_{a,j}$ az átlagos valós termelés a vizsgált időszakban,
- ¶ $\rho_{f,i}$ és $\rho_{s,i}$ a le- és felszabályozási árakat reprezentáló súlyok,
- ¶ Q db alulmenetrendezett és R db felülmenetrendezett negyedóra futó változói.

4. MODELLALKOTÁS

4.1. Adatsorok értelmezése

Az önálló alkotómunka során az MVM Zrt. egyik dunántúli 488 kW-os napelemparkjának 2019 évi intrahour negyedórás termelési adataival, és az Országos Meteorológiai Szolgálat (OMSZ) által, szintén negyedórás bontásban biztosított, várható sugárzást, hőmérsékletet és szélsébséget tartalmazó adatsorok alapján dolgoztam. Az OMSZ-től 00-24 és 24-48 típusú előrejelzéseket kaptam, amely kódokat úgy kell értelmezni, hogy az előrejelzés kézbesítése a 00 pillanatban történik, a vonatkozó időtartam pedig az ettől számított időértékek közti sáv. Tehát előbbi az intraday (aznapi), utóbbi pedig a day-ahead (másnapi) előrejelzésre használható. A másnapi menetrendet délelőtt 10:00-ig kell leadni, amelyen aznapi módosításokat lehet végezni a 2,5 órával későbbi negyedórától kezdve. Ezek az adatok tehát egy másnapi és egy egyszeri aznapi módosításra adnak lehetőséget az éjjel frissített előrejelzés alapján. A munkám során a 00-24 adatsorral kezdtem el foglalkozni, az időjárási adatok pontosságát nem próbáltam meg javítani, tehát végig direkt módszerrel dolgoztam.



8. Ábra: napsugárzás és PV termelés az év során

Ahhoz, hogy az adatsorokat össze lehessen fűzni, az időbélyegek egységesítése az első lépés. Az OMSZ a statisztikáit az egyesített koordinált világidőben [20] (UTC - Universal Time Coordinated) adja meg, amelyhez képest Magyarország télen az UTC+1, nyáron pedig az UTC+2 időzónát használja. Ennek értelmében össze kell hangolni a két adatsort, én ezt úgy csináltam, hogy az időjárás előrejelzést egy órával későbbre csúsztattam, azaz UTC+1-re hoztam, a termelési adatok pedig az óráátállítások

figyelembevételével így megfelelően voltak megadva. Ezek a módosítások egyébként szerencsére nem okoznak problémát, mivel mindig március és október utolsó vasárnapjának hajnalán, 2 és 3 óra között történik, amikor nincsen termelés. Viszont fontos ezt a rendezést tudatosan és visszakövethetően elvégezni, különben elcsúszhat az adatsor, ami 35040 időegységnél nem biztos, hogy észrevehető.

Az azonos számú minták miatt a január 1-re vonatkozó day-ahead menetrendet az intraday-jel helyettesítettem.

4.2. Karbantartások felderítése

A modell felépítéséhez már szükséges azonosítani a karbantartások miatti leállásokat, és ezeket mindenképpen el kell távolítani a vizsgált, vagy vizsgálandó időszakok közül. Az ilyen üzemhiányos állapotokat nem annyira könnyű automatikusan felismertetni a programmal, ezért érdemes lehet vizuálisan áttekinteni az adatsorokat anomáliák után kutatva. A leállások jellegzetesége természetesen a 0 termelés, ugyanakkor az általam vizsgált adatsorban több esetben is bekerültek 4, esetleg 8 kW-os termelés értékek a nullák közé, így ezek észlelése már kevésbé triviális. Érdemes vizsgálni a karbantartásos időszak széléit, mivel az ilyen munkálatokat várhatóan nem pontban valamelyik negyedóra kezdetén fejezik be, ezért lesznek olyan időszakok, amelyben részben rendes üzem, részben pedig leállás volt, ezért az átlagos termelés ezekben az esetekben is torzító hatással van az elemzésekre.

4.3. Saját modell

A saját modelletem MATLAB fejlesztői környezetben készítettem el, a korábban említett adatsorok alapján. A statisztikai módszerek közül egy olyat választottam, amely az előrejelzést az adott negyedóra termelésére hasonló pontok keresésével, majd azok kiértékelésével ad becslést. Mindezt úgy valósítottam meg, hogy az egy évnyi adatsort felosztottam két részre, az év minden öt napjából az első négyet feltöltöttem adatokkal, az utolsót pedig üresen hagytam, majd ezutóbbi negyedórára adtam becslést. Habár az időjárásra közvetlenül nincs hatással, hogy a hét melyik napjáról van szó, közvetetten el lehet képzelni az emberi tevékenységtől függő olyan hatásokat (pl.: szmog), amelyek ciklikusan befolyásolhatják a napelemparkok termelését, ezért választottam az ötödik napokat, mivel így a hétköznapok és hétvégék egyenletes eloszlásúak.

Ahhoz, hogy az adott negyedórával foglalkozzon a programkód, a hozzátartozó átlagos sugárzási teljesítménynek legalább 2 W/m^2 -nek kell lennie, ezáltal kiszűrtem az éjszakai, illetve a napkelte, vagy -nyugta széléin lévő átmeneti időszakot, amikor a napelempark épphogy, vagy egyáltalán nem termel. Továbbá kritérium még, hogy a

negyedóra, amelyre a predikciós folyamatot elvégezzük, ne legyen az előző pontban említett karbantartási folyamatok által zavart.

4.3.1. A VIZSGÁLT INTERVALLUM

Ha már megvannak a negyedórák, amelyekre becslést szeretnénk adni, el kell dönteni azt, hogy milyen intervallumba tartozó alapadatokat vizsgáljunk ehhez. A megoldásom ebben a kérdésben két értéket különböztet meg, egy kisebb (I1) és egy nagyobb (I2) határt. Az *i*-edik vizsgálat mindig azzal a lépéssel kezdődik, hogy az teljes éves adatsort a *circshift* paranccsal a tartomány közepére (17520-as indexű helyre) forgatja a kód, és az így kapott mátrix (17520-I1):(17520+I1) soraiban lévő elemeket fogja a hasonlóság szempontjából vizsgálni és értékelni. Röviden összefoglalva az adott negyedórához a I1 negyedórás környezetében lévő értékeket használjuk fel első körben. A második tartományhatár akkor jön képbe, ha az *i*-edik vizsgálatnál nem találunk legalább 3 olyan hasonló pontot, amely a kritériumoknak megfelel. Ekkor a vizsgált tartományt kibővítjük a tágabb tartomány és a szűkebb tartomány különbségére. Az optimalizálás során az I1 értékre 2000-re adódott, a tágabb tartomány értékére 12000-et szoktam használni, utóbbi jelentősen nem befolyásolja a teljes becslés hibáját, ha csak ritkán fordul elő, azonban I1 csökkentésével folyamatosan nő a jelentősége.

4.3.2. KRITÉRIUMOK

A hasonlóság vizsgálatának feltétele, hogy az *i*-edik és a *j*-edik elem egymáshoz képest egy bizonyos tartományon belüli környezeti értékekkel rendelkezzen. Erre a három vizsgált paraméter esetében különböző módszereket és tűrési tartományokat is kipróbáltam, ezek közül a legjobbnak (rMAE, rRMSE szempontjából is) bizonyulót fogom bemutatni.

A napsugárzás szempontjából relatív 7,5 %-os, abszolút korlátként pedig 10 W/m²-es értékeket határoztam meg. előbbi a nagyobb, utóbbi pedig a felkelő, vagy nyugvó nap esetén mért alacsony sugárzási értékek vizsgálatakor domináns. A szélesség kapcsán szintén abszolút és relatív korlátokat állítottam be, 50%, és 2 m/s-os határokkal. A hőmérsékletre vonatkozó feltétel pedig pusztán annyi, hogy a két érték különbsége 8 °C-nál kisebb legyen. Amennyiben tehát a *j*-edik elem esnek a feltételeknek megfelel, hasonlóknak minősül, melynek pontosságát a következő lépésben kerül kiértékelésre.

4.3.3. ÉRTÉKADÁS

A hasonlóan minősített értéket először osztályozzuk a sugárzás, szél, hőmérséklet és tendencia szempontjából, a következő képletekkel:

$$\begin{aligned}
 {}^{\circ}Y^{\circ}Y \quad \rho &= \frac{{}^{\circ}Y_i \quad {}^{\circ}Y_j}{\dot{I} \quad \dot{A} \quad \dot{\varnothing} \quad \dot{H} \quad \dot{Y} \quad \dot{H} \quad \dot{P} \quad \pi} \quad \mathcal{O} \dot{I} \quad \acute{o} \quad {}^{\circ}\Omega^{\circ}\Omega \acute{\omega} \breve{\Omega} \grave{o} \quad \acute{e} \quad \grave{i} \\
 {}^{\circ}O^{\circ}Y \quad \rho &= \frac{{}^{\circ}Y_i \quad {}^{\circ}Y_j}{\psi} \quad \mathcal{O} \Omega \acute{\omega} \breve{\Omega} \grave{o} \quad \acute{e} \quad \grave{i} \\
 {}^{\circ}Y \acute{\alpha}^{\circ}Y \rho &= \frac{{}^{\circ}Y \acute{\alpha} \quad {}^{\circ}Y \grave{\alpha}}{\dot{I} \quad \dot{A} \quad \dot{\varnothing} \quad \dot{H} \quad \dot{Y} \quad \dot{H} \quad \dot{P} \quad \pi} \quad \mathcal{O} \dot{I} \quad \acute{\alpha} \quad \Omega \quad \grave{\alpha} \quad {}^{\circ}\Omega \acute{\omega} \breve{\Omega} \grave{o} \quad \acute{e} \quad \grave{i} \\
 {}^{\circ}YY \quad \rho &= \frac{{}^{\circ}Y_{i\circ} \quad {}^{\circ}Y_{j\circ} \quad {}^{\circ}Y_{i\circ} \quad {}^{\circ}Y_{j\circ}}{\varsigma \quad \mathcal{O} \dot{I} \quad \dot{A} \quad \dot{\varnothing} \quad \dot{Y}_{i\circ} \quad \dot{H} \quad \dot{Y}_{j\circ} \quad \dot{H} \quad \dot{Y}_{i\circ} \quad \dot{H} \quad \dot{Y}_{j\circ} \quad \dot{H} \quad \dot{P} \quad \pi} \quad \mathcal{O} \dot{o} \quad \Omega \quad \acute{e} \quad \Omega^{\circ}\Omega \acute{\omega} \breve{\Omega} \grave{o} \quad \acute{e} \quad \grave{i}
 \end{aligned}$$

Ahol:

- | | | | |
|---|---|---|--|
| ¶ | ST | – | sugárzási tényező |
| ¶ | S | – | az adott negyedóra átlagos sugárzási értéke |
| ¶ | i, j | – | a szokásos indexelési elv alapján (i-edik negyedóra értéke, j-edik negyedóra értéke) |
| ¶ | sugfaktor | – | (0,1] intervallumon tetszőlegesen megválasztott érték, amivel a sugárzásra vonatkozó hasonlóság értékét súlyozni lehet |
| ¶ | HT | – | hőmérséklet tényező |
| ¶ | T | – | az adott negyedóra hőmérséklete |
| ¶ | hfaktor | – | (0,1] intervallumon tetszőlegesen megválasztott érték, amivel a hőmérsékletre vonatkozó hasonlóság értékét súlyozni lehet |
| ¶ | SzT | – | szél tényező |
| ¶ | szelfaktor | – | (0,1] intervallumon tetszőlegesen megválasztott érték, amivel a szélsőségekre vonatkozó hasonlóság értékét súlyozni lehet |
| ¶ | TT | – | tendenciatényező |
| ¶ | tendfaktor | – | (0,1] intervallumon tetszőlegesen megválasztott érték, amivel a vizsgált negyedóra megelőző, és követő időszavok sugárzásait összehasonlító értéket súlyozni lehet |
| ¶ | ${}^{\circ}Y_{i\circ} \quad \dot{H} \quad \dot{Y}_{j\circ}$ | – | annak a hét negyedórának a sugárzásértékei, amelynek közepén a vizsgált (i-edik) negyedóra található |

Ha megállapítottuk a négy súlytényezőt mind az m darab hasonlóan minősített j-edik elemre, a következő lépésben elvégezzük a becslést az alábbi formula segítségével:

$$\omega_i = \frac{B_{\mathcal{O}U_0}^{\mathcal{Z}} \quad {}^{\circ}Y \quad \mathcal{O} \quad {}^{\circ}Y \quad {}^{\circ}Y \quad {}^{\circ}Y \quad {}^{\circ}Y \quad \acute{\alpha} \quad {}^{\circ}Y}{B_{\mathcal{O}U_0}^{\mathcal{Z}} \quad {}^{\circ}Y \quad {}^{\circ}Y \quad {}^{\circ}Y \quad {}^{\circ}Y \quad {}^{\circ}Y \quad \acute{\alpha} \quad {}^{\circ}Y}$$

Összességében tehát egy súlyozott átlagot számolunk, amely súlytényezőit egyrészt egyenként tetszőlegesen hangolhatunk, hogy mennyire „büntesse” az eltérést, illetve a t paramétert növelve a hasonlóság szempontrendszer alapján jobb értékeket súlyát növelhetjük. A paraméterek megválasztására nincs direkt képlet, mivel sok változója van, ezért csak parciális optimumkereséssel lehet pontosítani a becslést.

4.4. Python algoritmusok optimalizálása

A kutatási munkám második felében átváltottam Python programozási nyelvre, és ebben folytattam az elemzéseket. Az adatbázisok kezelését pandas, az algoritmusokat pedig a scikit learn könyvtár segítségével vizsgáltam Jupyter Notebook fejlesztői környezetben.

4.4.1. MÓDSZEREK

A scikit learn könyvtár algoritmusai közül a következőkkel dolgoztam:

- ¶ Linear Regression (LR),
- ¶ Decision Tree Regressor (DTR),
- ¶ AdaBoosting Regression (ABR),
- ¶ Gradient Boosting Regression (GBR),
- ¶ k-Nearest Neighbour (knn),
- ¶ Support Vector Machines (SVM),
- ¶ Multi Layer Perception (MLP).

A gépi tanuláshoz az alapadatokat 80-20%-ban osztottam tanuló és teszt adatsorokra, melyek előállíthatóak véletlenszerű, vagy tudatos mintavételezéssel is. Előbbi pontosabb átlagos értékelést nyújt nagy számú (100-200) futtatás esetén, azonban a módszerek egyidejű vizsgálatakor időtakarékoság és a saját készítésű módszerrel való összehasonlíthatóság miatt a korábban ismertetett ötödik napos kiválasztási módszert alkalmaztam, és a kiértékelést is ennek mentén fogom bemutatni. Az adatokat a pontosabb becslés miatt skáláztam a StandardScaler függvénnyel.

4.4.2. PREDIKTOROK

A bemenő adatok kiválasztásánál többféle módszerrel is próbálkoztam, a legjobb eredményeket úgy kaptam, hogy az adott i -edik negyedórára vonatkozó sugárzási érték mellett az egy órával későbbi levegő hőmérsékletet, az adott nap sorszámát, az $[i-2, i+2]$ intervallumban lévő további negyedórák és az $i-1, i+1$ -edik negyedórák átlagát vettem figyelembe.

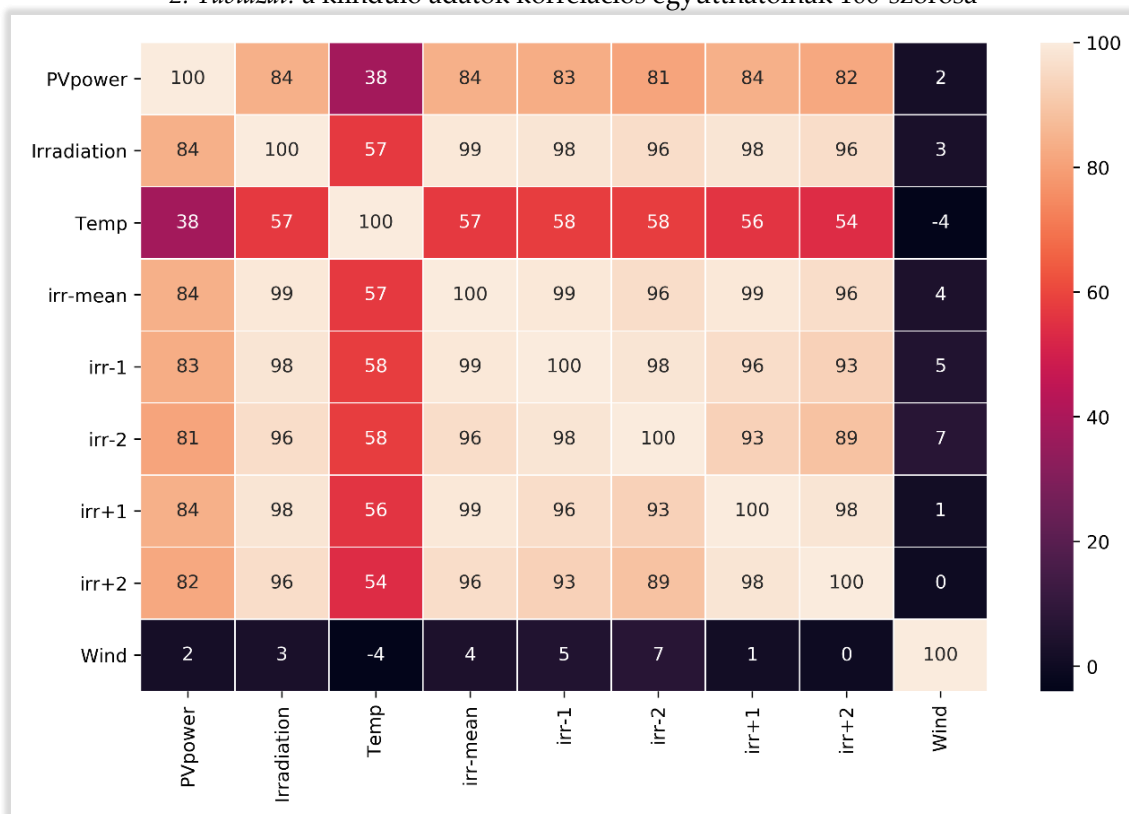
5. EREDMÉNYEK, ÖSSZEHASONLÍTÁS, KIÉRTÉKEELÉS

5.1. Idősoros csúsztatás

A saját modell készítése során azt tapasztaltam, hogy a későbbi hőmérséklet adatok használata pontosabb becslést eredmény ad. Statisztikai módszer lévén, alapvetően nincs szükség értelmezni az optimalizálás során kapott paraméter értékeket, ugyanakkor ebben az esetben érdemes lehet megfontolni a háttérben lévő folyamatokat. A fizikai indoklásom erre vonatkozóan azon alapul, hogy a hőmérséklet értékeket az OMSZ a levegőre adja meg, azonban a levegőt a beeső sugárzás elnyelt része által felmelegített földfelszín, vagy egyéb felszíni tereptárgyak (pl. napelemek) melegítik fel. Ebből következően a felszín és a levegő hőmérséklete között a közegek hővezetési tulajdonságainak megfelelően van egy időbeni késleltettség. Mivel a napelemek termelése a cellák hőmérsékletétől függ, nem a levegőétől, nem triviális, hogy az aktuális értékekkel közelíthető a legjobban a pillanatnyi cellahőmérséklet. Összegezve a tapasztalataim alapján a következtetésem az, hogy az adott negyedóra cellahőmérsékletének hatását jobban jellemzi az egy órával későbbre várt levegőhőmérséklet, mint az azonos időbélégű.

A következő táblázat az így kapott prediktorok egymáshoz viszonyított korrelációs együtthatóját mutatja be.

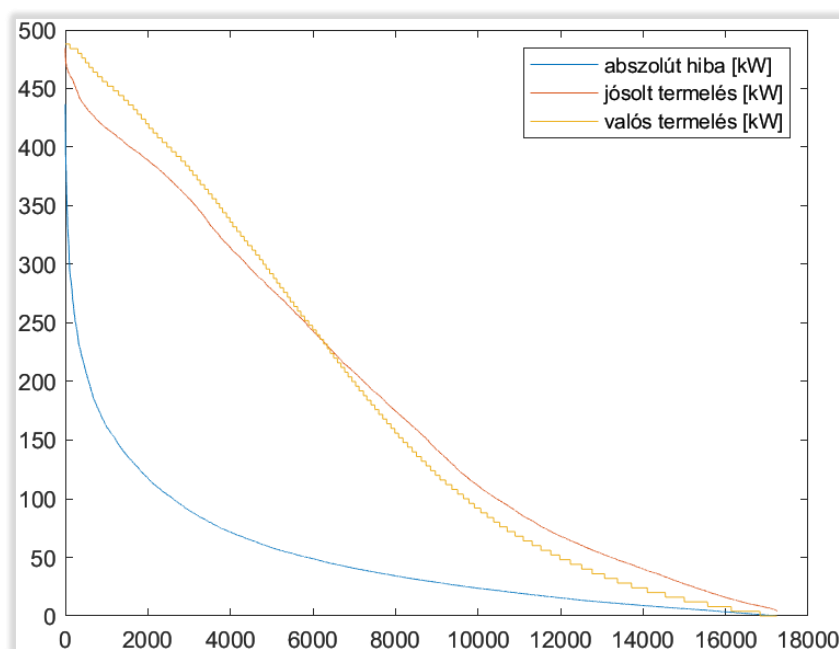
2. Táblázat: a kiinduló adatok korrelációs együtthatóinak 100-szorosa



Ebből érdekességképpen még megemlítendő, hogy az i -edik negyedóra termelésére nagyobb magyarázóerővel bír az azt megelőző, illetve követő negyedórák sugárzásának átlaga, mint maga az i -edik időszakra vonatkozó előrejelzés.

5.2. Átlagoló hatás

A következő ábrán a valós és jósolt termelési tartamdiagramokat, valamint az abszolút hibákat csökkenő sorrendbe rendezve ábrázoltam a saját készítésű módszeremre.



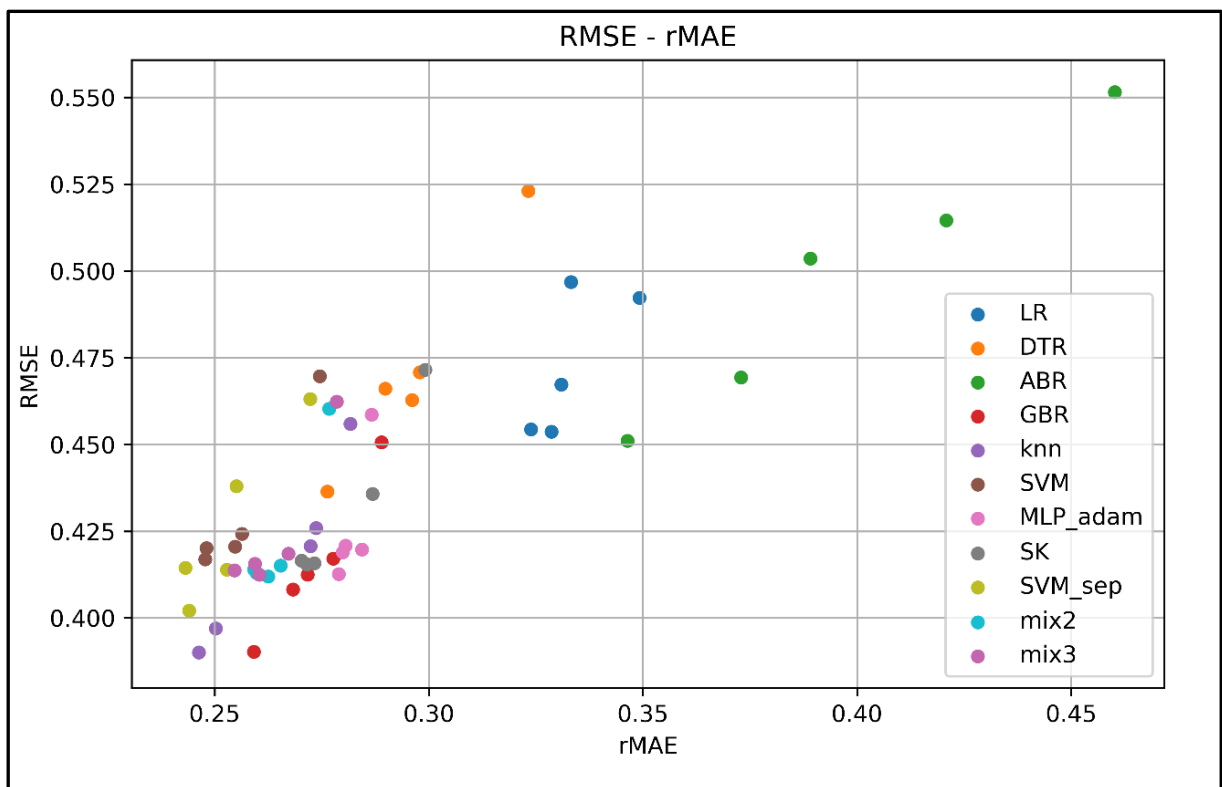
9. Ábra: tartamdiagramok

Az ábra jól szemlélteti, hogy összességében a módszer az átlagolás miatt konzervatív becslést ad, nagy teljesítmények esetén alul, kisebbeknél pedig felülbecsüli a valós termelést. Az is látható, hogy vélhetően hasonló okokból a maximális teljesítményt jóval ritkábban éri el, mint a valós esetben. Ennek kapcsán elvégeztem egy elemzést, amely során a jósolt termelés szerint csökkenő sorrendbe rendezett predikciós mátrixban a becsült értékeket kicseréltem a hasonlóan csökkenő sorrendbe rendezett valós termelésekkel, majd a negyedórák indexe szerint visszarendeztem az adatokat. Ez a módszer nem egy konstruktív modell, valós becslést nem lehet így adni, ugyanakkor lehetne olyan eljárásokat létrehozni, amely az átlagos termelés alatti értékeket lefelé, a felettieket pedig felfelé súlyozza, a konkrét értékek függvényében. Ennek a módszernek az idealizált változata az, amit a valós termelések ismeretében meg tudtam valósítani, így ennek értékelése indokolt.

A tapasztalat az, hogy az átlagos abszolút hiba és ezáltal a relatív hiba jelentősen csökkent, ugyanakkor ezzel egyidejűleg az RMSE, ami a nagyobb hibákra érzékeny, megnőtt. Tehát ennek a korrekciónak nincs egyértelműen pozitív, vagy negatív hatása.

5.3. Eredmények a különböző módszerekre

A módszereket az említett módon relatív hiba (rMAE) és rRMSE szerint értékeltem. A következő diagramon ábrázoltam az eredményeket az öt vizsgált tanuló-teszt adatsorra vonatkozóan.



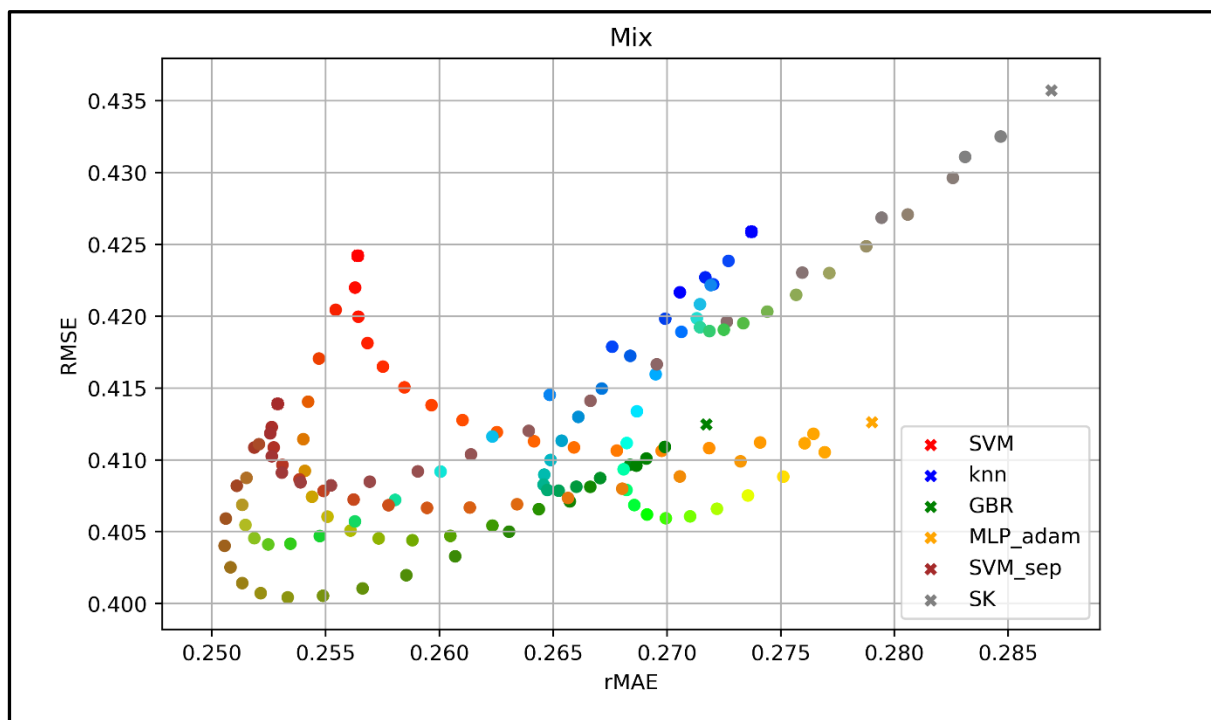
10. Ábra: intraday eredmények

Láthatóan a két hibaszámítást figyelembe véve nincs egyértelműen legjobb módszer, azonban vannak olyanok, amelyek kiemelkedőek. Ezek név szerint az SVM, ennek terület index alapján szeparált verziója kifejezetten rMAE szempontjából, a knn mindkettő, GBR pedig inkább rRMSE irányából számítanak jónak. Az MLP, illetve a saját modellem (SK) a középmezőnyben, a lineáris regresszió az ABR és a DTR algoritmusok pedig nem bizonyultak pontosnak, ezért utóbbiakkal kevesebbet foglalkoztam.

Ezen az ábrán a mix2 nevű módszer az MLP_adam és az SVM_sep átlagolásából származik, míg a mix3 ennek továbbfejlesztése az adott napi termelési görbék felfutó és leívelő ágának korrekciójával.

5.4. Mix módszerek

Az előző pontban kiemelt legjobb módszerekből súlyozott átlagolással is készítettem hibrid-statisztikai előrejelzéseket. A súlyokat az egyikről fokozatosan a másik irányába áthelyezve színátmenetesen ábrázoltam a következő diagramon.



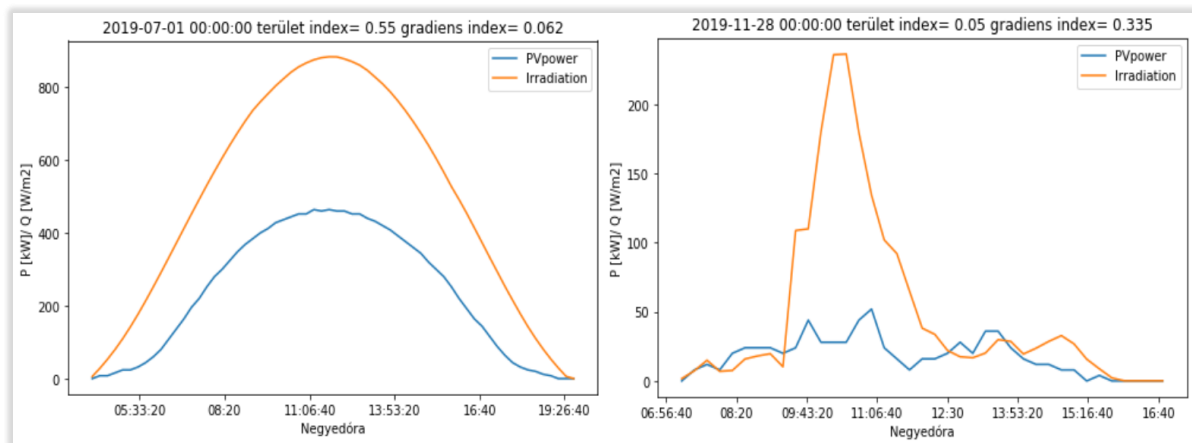
11. Ábra: módszerek kombinálásának eredményei

A diagramon látható, hogy a módszerek rendre javíthatóak más algoritmusokkal való kombinálással, azonban ezek egytől-egyig Pareto hatékonynak bizonyultak, az átmenet középső (lényegi) részén a két hibatípus csak egymás rovására javítható. Megfigyelhető továbbá, hogy az algoritmusok kombinálása az ábrázolt módszerek között minden esetben javította a menetrend pontosságát, még a leggyengébb algoritmus is tudta javítani a legerősebbet (SK-SVM_sep). Ez a megállapítás meglehetősen fontos, és tovább erősíti azt a gondolatot, hogy a menetrendezés célja nem a fluktuációk követése, hanem a középérték megtalálása, amelyet segíti a zavaró hatások diverz szűrése.

A kombinálás eredményeként láthatóan az SVM_sep és a GBR átlagolása bizonyult a legjobbnak, a továbbiakban azonban ezzel a vegyes megoldással nem foguk tovább foglalkozni, mivel alapvetően ennek a munkának a célja maguknak a módszereknek, és különböző praktikáknak a vizsgálata, nem pedig a menetrend végtelenségig történő fejlesztése.

5.5. Terület és gradiens index kapcsolata

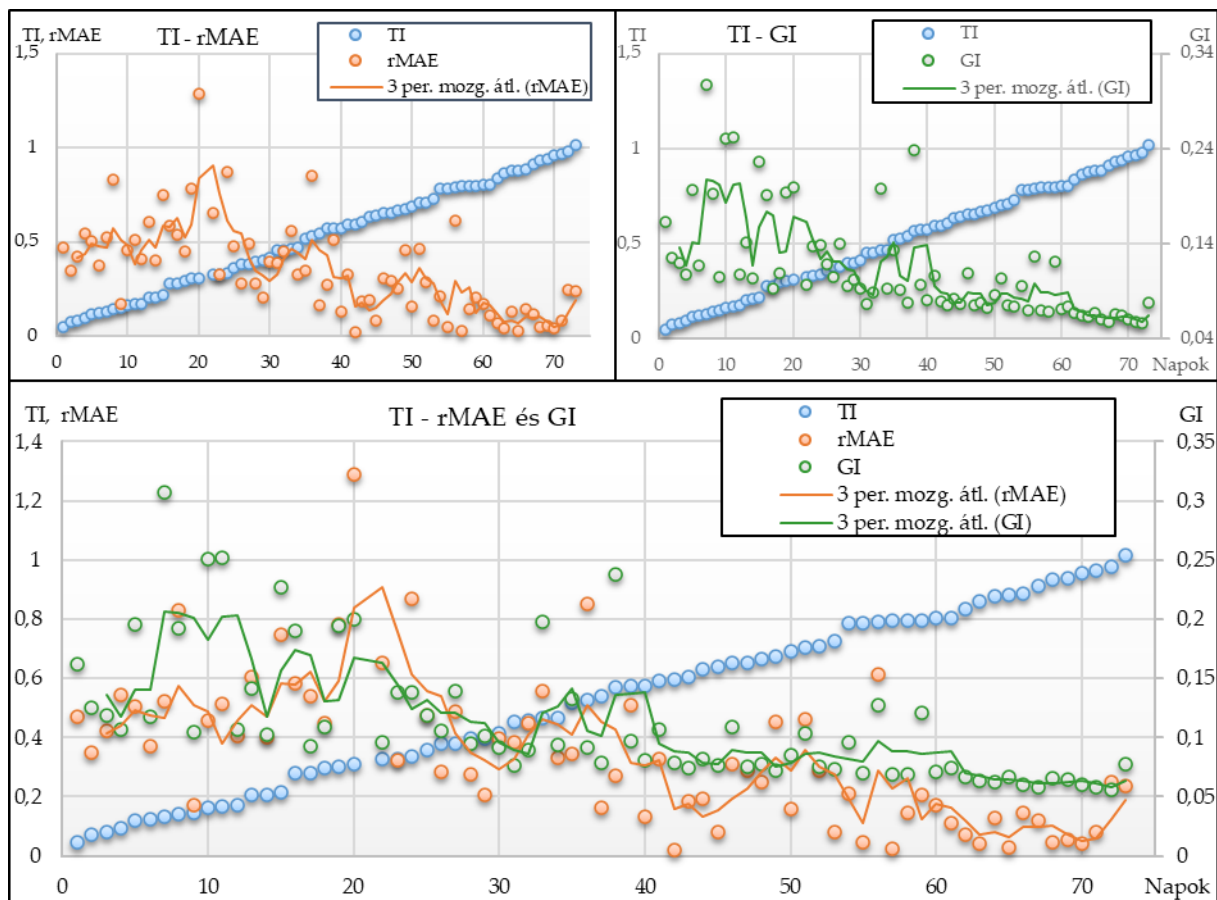
Ezt a két mutatószámot a korábban ismertetett módon a céllal alkalmaztam, hogy az időjáráselőrejelzés alapján el lehessen dönteni a következő napról, hogy menetrendezés szempontjából milyen lesz, és ennek megfelelően lehessen előrejelzési módszert választani. Az alábbi diagramon két teljesen eltérő napot sugárzási előrejelzését és valós termelését ábrázoltam. Az első egy ideális nyári nap a diagram és az



12. Ábra: menetrendezés szempontjából jó és rossz nap bemutatása

alacsony GI alapján, amelyet a magas terület indexű előrejelzésből sejteni lehetett. Ezzel szemben a második egy olyan napot ábrázol, amelyen az alacsony termelés még relatíve nagy ingadozásokkal is terhelt, tehát menetrendezési szempontból a lehető legrosszabb. A diagram fölötti számokból látszik, hogy a nagyon alacsony terület indexből lehetett erre számítani.

A két mutatószám, illetve a menetrendezési hiba közti kapcsolatot az alábbi diagram csoporton szemléltettem, amelyen terület index szerinti rendezésben láthatóak, az összetartozó gradiens index és az rMAE értékek. GI és rMAE esetén a kis, TI-nél pedig a nagy értékek hivatottak reprezentálni a rosszul menetrendezhető napokat.

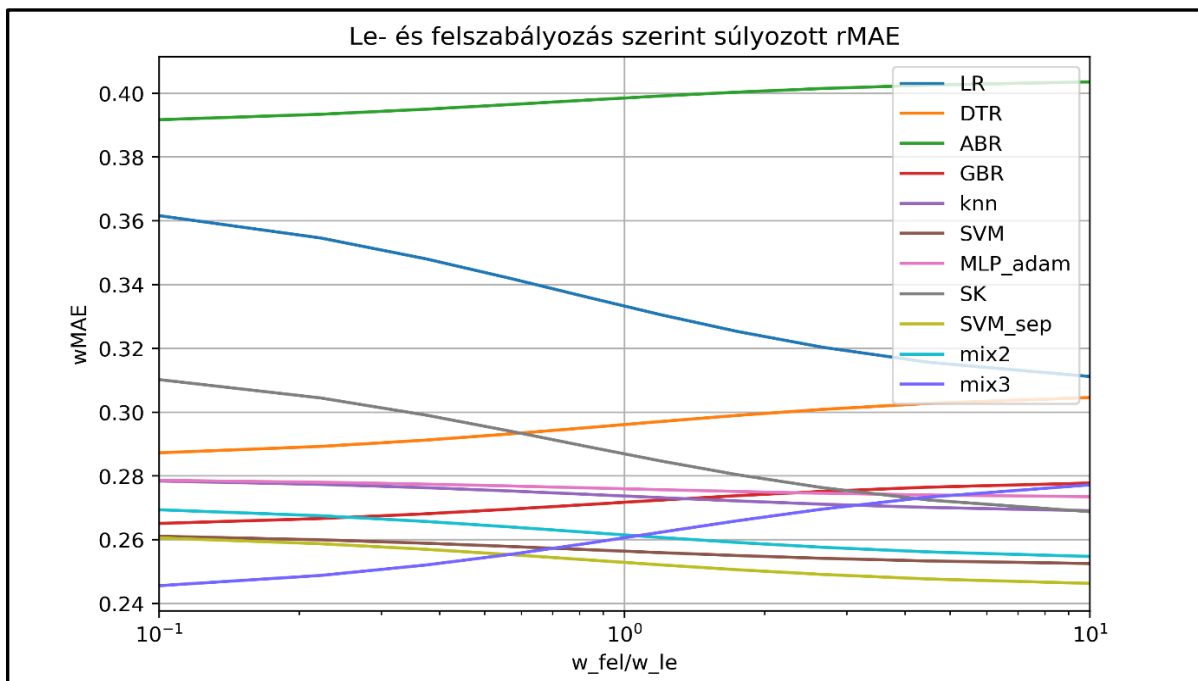


13. Ábra: TI szerint növekvő sorba rendezve a TI, GI és rMAE trendjeinek bemutatása

Vizuálisan megállapítható, hogy a két index trendje ellentétes, ahogyan vártuk. A mutatószámok nem tudják jól lekövetni a karbantartás miatt csonkított napokat, illetve a kis termelésű, erősen felhős napokat így ezek okozhatnak kiugró értékeket. Azonban arra a célra, hogy az erősen változékony napokat kiszűrjük, megfelelően jól működik a TI. Ez azért fontos, mivel ezáltal előre el tudjuk becsülni, hogy mennyire lesz nehezen menetrendezhető az adott nap, ami lehetővé teszi, hogy az előrejelzés konzervativitását (átlag felé torzítását) kezelni tudjuk. Így a különböző típusokat kiválogatva külön optimalizált algoritmusokat tudunk használni. Erre a gondolatmenetre hivatkozok terület index alapú szeparálásként.

5.6. Le és felszabályozás szerinti súlyozás

Az eredményeket piaci szempontból vizsgálva, a fel és leszabályozási árak arányában a következő hibagörbéket produkálták az egyes módszerek.



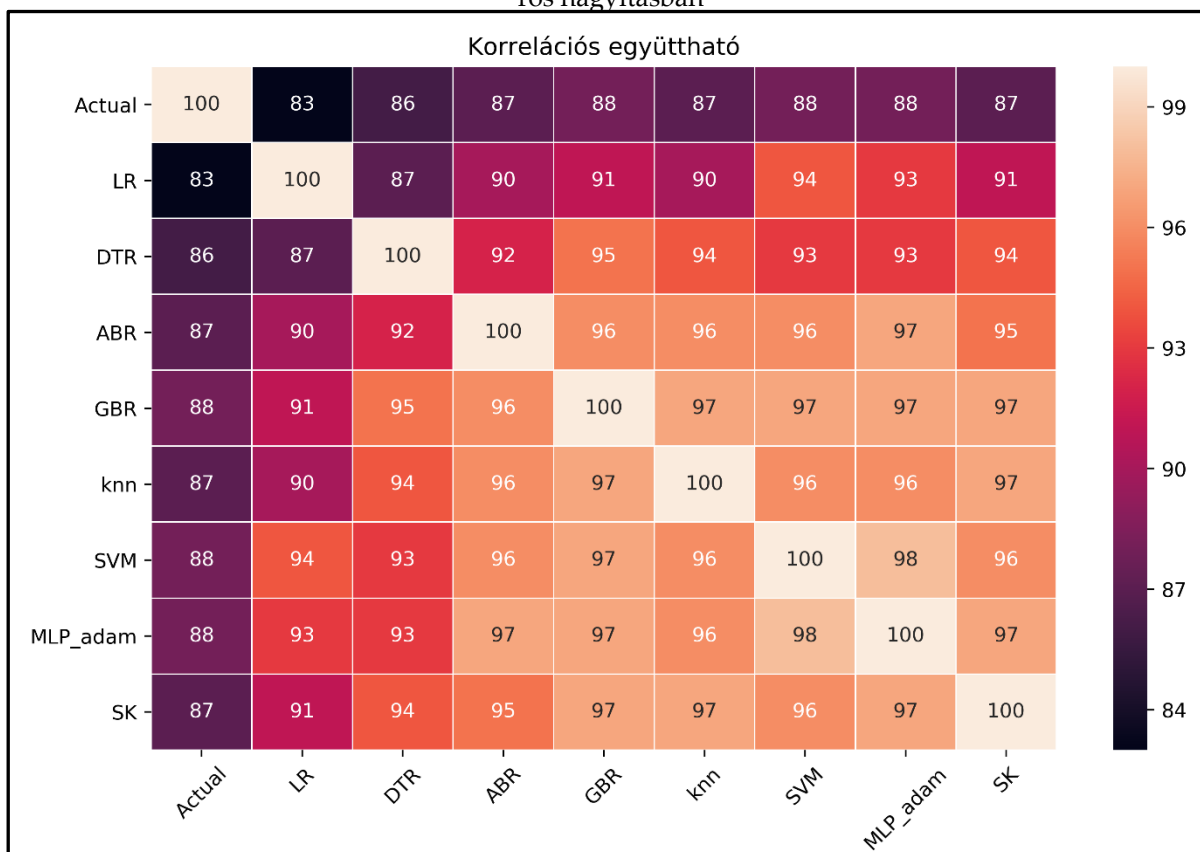
14. Ábra: le- és felszabályozás szerint súlyozott rMAE

Az x tengelyen a felirat szerint, a le- és felszabályozás súlyainak (egységárának) aránya látható, míg az y tengelyen a kapott súlyozott rMAE, azaz wMAE (weighted MAE) értékeket ábrázoltam. A pl. az USA-ban használt szimmetrikus árazási rendszer az $x=1$ helyen található, míg a hazai rendszerre jellemző 3-5, de akár 10-15-szörös arány ettől jobbra látható. Amit ebből a diagramból le lehet szűrni, az az, hogy a módszerek közül vannak olyanok, amelyek a hazai szinten alkalmazott alulmenetrendezés támogató elszámolási rendszerben működnek jobban, sőt, a többség egyébként ilyen, azonban az ABR, DTR, mix3 és GBR módszerek esetében ez a trend ellentétes. Itt visszacsatolnék a korábban az 5.4-es pontban ismertetett kombinált módszerekre, amelyeknél a GBR volt az egyik legjobb társmódszer, mivel ezt a hazai környezetben árnyalja az előzőekben realizált felülmenetrendezés támogató jellege.

5.7. Korreláció

Alapvetően a korrelációra kisebb hangsúlyt fektettem az eljárások fejlesztése során, azonban elvégeztem néhány kimutatást az értékeléshez. Az alábbi táblázatban összefoglaltam az egyes módszerek és a valós termelés korrelációs tényezőit. Az 1. táblázatra visszacsatolva, a valós termelés és a globálsugárzás közti korrelációs együttható 0.842 volt a teljes adatsorra vonatkoztatva, jelen értékek pedig az ötödik negyedóraakra vonatkoznak.

3. Táblázat: a valós termelés és az egyes módszerek közti korrelációs együtthatók bemutatása 100-szoros nagyításban



Az első szembetűnő információ a színek alapján az, hogy a magyarázóerő az rMAE, rRMSE kapcsán megállapított sorrendet adja vissza, a referenciaként használt LR az, amelyik ebből a szempontból gyengébb eredményt adott. A második észrevétel, hogy a valós termeléshez képest jóval erősebb a kapcsolat az egyes módszerek között, ami viszonylag triviális, hiszen mind valamilyen szempontrendszer szerinti optimális középértéket próbál adni. A harmadik pedig az, hogy habár a MLP és az SVM a két fő hibaszámítás alapján éppen ellentétes szempontok szerint minősült jónak, a kettő közti magyarázóerő nagyobb, mint a köztes megoldásokkal vett.

5.8. Gradiens index, szórás

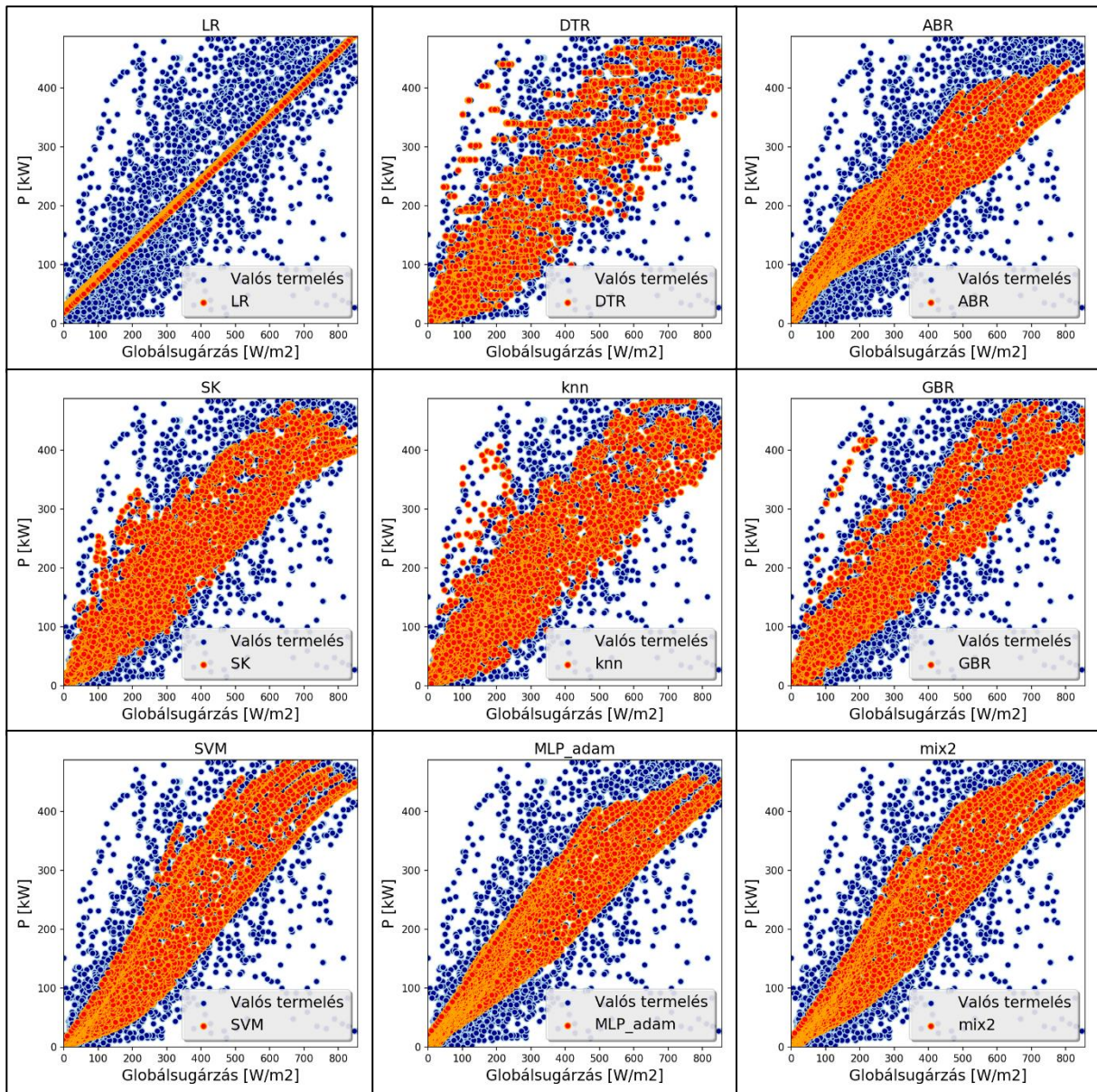
Az alap koncepcióm az, hogy az előrejelzések közül nem az lesz a legpontosabb, amelyik megpróbálja a fluktuációkat minél inkább követni, hanem az, amelyik igyekszik az ingadozások középértékét megtalálni. Ennek az állításnak az igazolásához a módszereket összehasonlítottam a teljes évre vonatkozóan az előrejelzett termelés gradiens indexe, szórása és a menetrend hibái alapján.

4. Táblázat: gradiens index hatása a hibákra

	Valós	LR	DTR	ABR	GBR	knn	SVM	MLP	SK	SVM_sep	mix2	mix3
gi	0,157	0,076	0,163	0,084	0,085	0,108	0,085	0,081	0,099	0,081	0,081	0,097
rMAE		0,333	0,296	0,344	0,272	0,274	0,256	0,274	0,287	0,253	0,261	0,285
rRMSE		0,497	0,463	0,451	0,412	0,426	0,424	0,416	0,436	0,414	0,415	0,433
mean	174,2	177,8	177,5	182,4	178,3	173,9	178,6	177,0	177,6	176,8	177,8	175,4
std	155,3	134,3	149,0	114,0	141,1	141,5	150,2	131,8	137,6	148,2	140,7	143,4

Az előrejelzés kapcsán a szórás az egyik könnyedén mérhető jellemző, amivel kapcsolatban összefüggések után lehet kutatni. Az std az átlagtól (mean) való átlagos eltérést adja meg, ebből adódóan ez az érték a szélsőségek gyakoriságával ellentétes jellegű. A mi esetünkben nem egy átlagtól, hanem a napi várható középértéktől való ingadozás mértéke az, ami releváns, erre azonban nem egyértelmű, hogy milyen mérőszámot érdemes alkalmazni, ezért vezettem be a gradiens indexet, amely egyszerűen számítható, ugyanakkor jelentéstartalmát validálni kell. Egyelőre annyi látszik, hogy a pontosság egy bizonyos fokához szükséges a megfelelő sávban lévő GI, azonban utóbbi nem okozza feltétlenül előbbi teljesülését. A pontosabb értékeléshez további módszerek, előrejelzése, illetve más napelemparkok vizsgálata lenne szükséges.

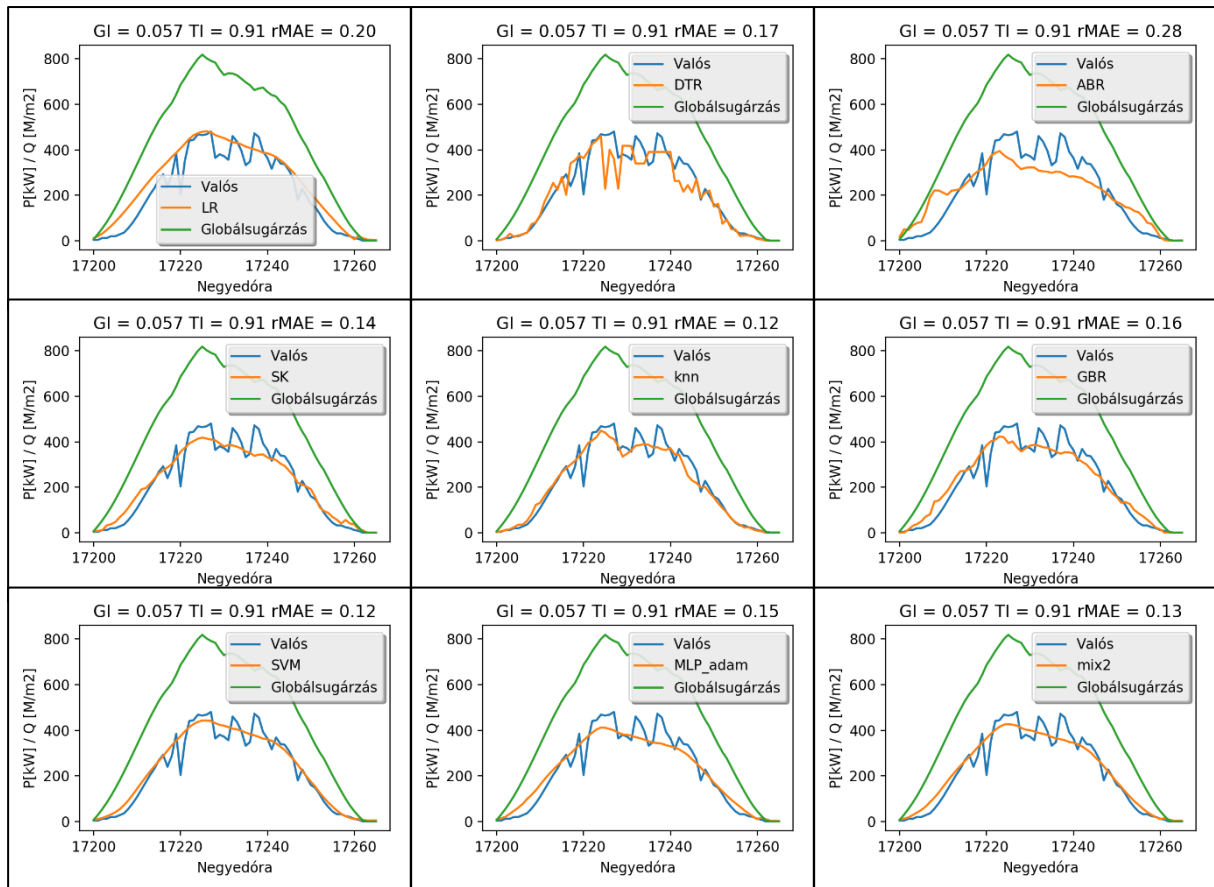
A módszerek szórási diagramjain jól látszanak az elvi különbségek. A diagramokon kék pontokkal jelöltem a valós termelést a hozzá tartozó $[W/m^2]$ mértékegységű globálsugárzás függvényében. A piros pontok pedig szintén a sugárzás függvényében a jósolt fotovoltikus teljesítményt mutatják be. A bemenő adattípusok (Irr-1, Irr-2, stb.) növelésével ezek a diagramok egyre komplexebbé válnak, ezért ezeket az ábrákat egyetlen prediktorral, a jósolt sugárzási értékkel generáltam, mivel így alkalmasabbak a módszerek összehasonlítására.



15. Ábra: szórási diagramok (LR, DTR, ABR, SK, knn, GBR, SVM, MLP, mix2)

A diagramok alapján az a következtetés, hogy a sugárzáshoz képesti szórás és a pontosság között nincs szoros kapcsolat, hiszen láthatóan a két véglet, a LR és a DTR egyik sem erős módszer, míg például a k-nn, GBR annak ellenére tudnak éppen rRMSE-ben eredményesebbek lenni (amely a kiugró hibákat jobban bünteti), hogy láthatóan szélesebb tartományban jósolnak egy adott sugárzási értékhez.

A szórási diagramokhoz hasonlóan látványosak a napi sugárzási előrejelzés, menetrend és termelés görbék alakulásai a különböző esetekre. Mivel a menetrendek viszonylag hasonlóak, egy relatíve szűk sávban mozognak, egyben nem vizualizálhatóak, ezért a szórási diagramokkal megegyező elrendezésben ábrázoltam, ezt már az összes prediktor alapján becslve.



16. Ábra: egy június végi nap sugárzás előrejelzése, valós termelése és menetrendrendjei különböző módszerekkel

A napi diagramokon jól bemutatják az egyes módszerek erősségeit, gyengeségeit. Például a DTR a hajnali, illetve esti órákat jól eltalálja, azonban nap közben túlságosan változékony becslést ad, megpróbálja az összes fluktuációt lekövetni. Ezzel szemben az LR, éMLP és még hatékonyabban az SVM arra törekszik, hogy folytonosabb becslést adjon, amely sokkal kevésbé ingadozik. A két elv között található a knn, a saját modellem, és a GBR, amelyek habár enyhén, de megpróbálják követni a változékony idő miatti ingadozásokat. Az ábrák feliratában a vizsgált nap gradiens és terület indexe, illetve az adott módszer előrejelzési hibája látható. A kimaradt eljárások ábrázolását nem tartottam fontosnak, mivel jellegre megegyeznek az alapszámításokkal szórás és napi menetrend tekintetében is.

5.9. Az időhorizont hatása

A következőkben az intraday és day-ahead előrejelzések pontossága közti eltéréseket fogom bemutatni. A menetrendjavítás legtriviálisabb, illetve ebből adódóan leginkább

szorgalmazott területe az időhorizont, tehát a menetrend leadása és a vonatkozó negyedórák közti idő csökkentése. A jelen esetben alkalmazott kétféle előrejelzés megfelel a hazai szabályozásnak, hiszen a másnapi menetrendet délelőtt 10:00-ig, az aznapi módosításokat pedig minimum 2,5 órával későbbre lehet leadni. Ennek értelmében a most vizsgált esetről már a mostani körülmények között is lehetne közelebbi intraday menetrendezést végezni, nyilván, ehhez olyan szolgáltatói szerződésre van szükség, amely megengedi a napi többszöri lekérdezést. A horizont csökkentésére ezen túl a piacműködtetés nemzetközi szintű fejlesztése ad lehetőséget.

Az eljárások és a kód ugyanaz a két esetre, a teljes évre való összehasonlíthatóság kedvéért a január 1-i hiányzó másnapi menetrendet az aznapival pótoltam. A két előrejelzés közti eltéréseket fogom ismertetni, a jellegre azonos trendeket nem.

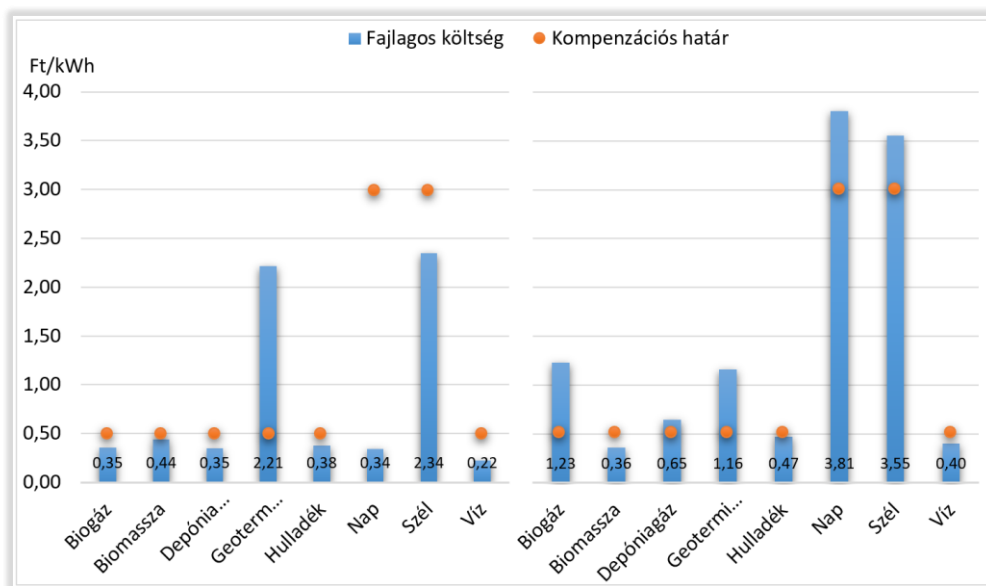
5. Táblázat: aznapi és másnapi menetrendek átlagos hibái

		LR	DTR	ABR	GBR	knn	SVM	MLP	SK	SVM_sep	mix2	mix3
ID	rMAE	0,333	0,297	0,420	0,273	0,265	0,256	0,284	0,280	0,254	0,266	0,264
	rRMSE	0,473	0,472	0,518	0,416	0,418	0,430	0,427	0,431	0,426	0,423	0,425
DA	rMAE	0,348	0,318	0,419	0,289	0,288	0,273	0,303	0,301	0,267	0,284	0,282
	rRMSE	0,489	0,498	0,521	0,437	0,448	0,455	0,450	0,455	0,451	0,446	0,448

Az öt időszakra vonatkoztatott hibák átlaga jellegre hasonlóan alakul, értékben egy kb. 5 %-os relatív javulás érhető el az ID menetrend módosítással.

5.10. Pénzügyi értékelés

Ahhoz, hogy a néhány százaléknyi, ezreléknyi pontosság különbséget értékelni lehessen, szükség van valamilyen becslésre a fajlagos költségeket illetően. A MAVIR Zrt. KÁT és prémium támogatás statisztikai adatai c. havi adatpublikációinak KAT_KE_fajlagos munkalapján minden hónapban közli az egyes primerenergia típusokra vonatkozó fajlagos kiegyenlítőenergia költséget. A 2020. április 1. óta érvényben lévő elszámolási rendszer a kiegyenlítőenergia költségeket fokozatosan a termelőkre terheli át, az átlagosan éves szinten kb. 3 Ft/kWh egységköltséget jelenleg még pótdíjcsökkentésként levonják a havi szaldóból, azonban ezt pár éven belül fokozatosan nullára csökkentik majd. Továbbá, ahogyan a következő diagramon is látszik, a fajlagos költségek megoszlása erősen ingadozó, éves szinten a 3 Ft/kWh csökkentéssel elvileg közel nullára adódna a pótdíj, azonban a havi eltérések miatt jellemzően lesznek nullás, és ráfizetéses hónapok. Mivel a negatív szaldó nem vihető tovább, a 3 Ft-os csökkentés nem lesz minden hónapban teljesen kihasználva. Ez a 3 Ft/kWh egyébként a teljes KÁT-os nap-elemkapacitásra vonatkozik, tehát ezt az értéket már jelentősen befolyásolja a portfólió szintű menetrendezés javító hatása.

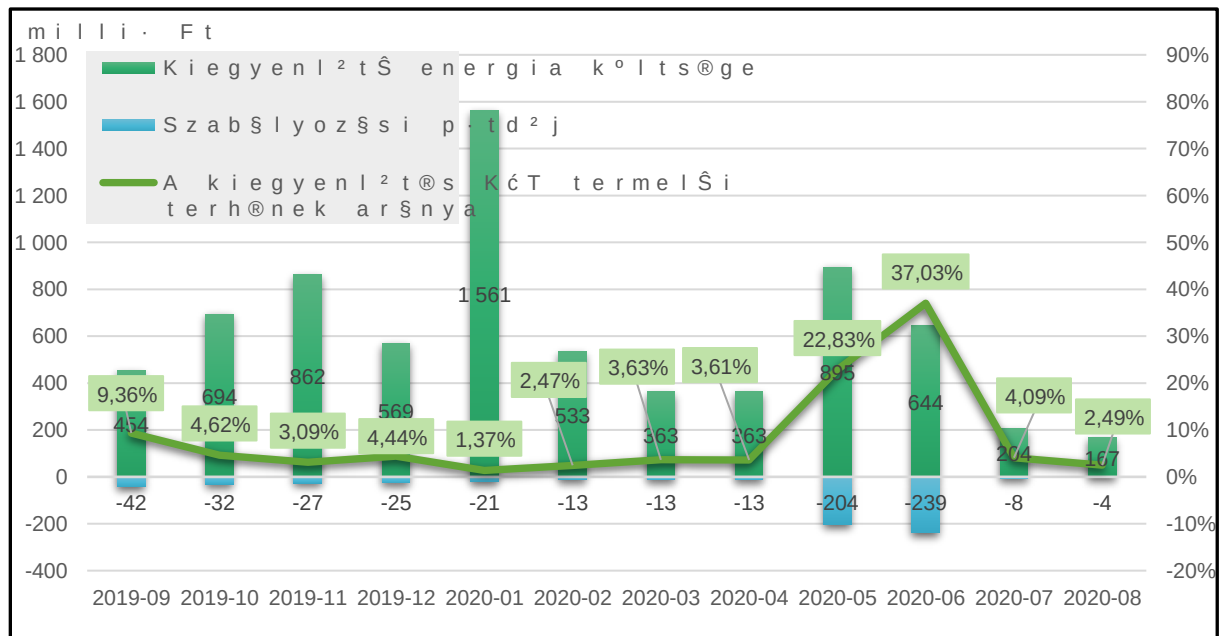


17. Ábra: fajlagos kiegyenlítőenergia költségek 2020 augusztus és május esetén

Ha tehát abból indulunk ki, hogy a kiegyenlítőenergia megtermelt kWh energiára vonatkoztatva átlagosan 3 Ft, akkor a 10% menetrendezési hiba csökkentés, 0,3 Ft/kWh költségcsökkentést jelent, jelenleg még a rendszerirányítónak, de a pótdíjcsökkentési tényező redukálásával majd egyre inkább a termelőknek. Természetesen ez a gondolatmenet nem veszi figyelembe az kiegyenlítőenergia egységár ingadozását. Tovább számolva a saját erőmű példára vonatkozóan, 488 kW beépített termelésre és 18,38%-os kihasználási aránnyal a következtetésem az, hogy 10% javulás az átlagos szintről évente, beépített kW-onként 483 Ft költségcsökkentést jelent. Azaz, egy 50 MW-os erőművi portfólióra 24 150 000 Ft, míg a teljes hazai kapacitásra 589 743 000 Ft körüli nagyságrenddel lehet számolni. Ugyanakkor az első fejezetben bemutatott jövőkép alapján arra lehet következtetni, hogy a növekvő napelem termelés kiegyenlítésére a jelenlegi tartalék kapacitás nem lesz elég, olyan erőművekre is szükség lesz, amelyek drágábban üzemelnek, ez pedig várhatóan a kiegyenlítő energia fajlagos költségét növelni fogja.

Az előbbi számítás kapcsán még egy adatsort érdemes bemutatni, amely szintén a MAVIR Zrt. említett adatpublikációjában található, a *KAT_KE_potdij* munkalapon. A 18. ábra bemutatja havi bontásban a 2019 szeptember és 2020 augusztus közti időszak KÁT kiegyenlítő energia költségeinek, illetve a termelői terhek alakulását. A diagramon zöld oszlopokkal jelezték a mérlegkör teljes költségét, amely a 3.1-es pontban körül járt árazással megállapított egységárai és igénybe vett szabályozási energiák negyedóránkénti szorzatainak előjeles összegeként áll elő. Ezzel részletesebben a *KAT_KE_elszamolas* munkalap foglalkozik. Ennyibe kerül tehát a mérlegkör kiegyenlítése, amelyből az engedélyesek a saját hibájukkal arányosan veszik ki a részüket, amelyből a pótdíjcsökkentést levonva megkapjuk a ténylegesen befizetendő összeget.

Ezt ábrázolják a kék oszlopok, míg a zöld vonaldiagramm azt mutatja be, hogy a költségek milyen arányban terhelik a mérlegköri engedélyeseket.



18. Ábra: KÁT (validált) kiegyenlítő energia és szabályozási pótdíj alakulása

Az ábrán látható az új rendszer hatása, május és június hónapokban több, mint 200 millió Ft pótdíjat kellett fizetnie az engedélyeseknek. A napelemparkok menetrendezési pontosságának javításából adódó költségcsökkentésre pontosabban akkor lehet majd becslést adni, ha a jelenlegi rendszerben legalább egy évnyi adat áll majd rendelkezésre.

6. ÖSSZEFOGLALÁS, EREDMÉNYEK ÉRTÉKELÉSE

6.1. Összefoglalás

A kutatási munkám során felépítettem egy előrejelzési módszert napelemparkok termelésének időjárási adatok alapján történő becslésére, amely során alaposabban megismertem a MATLAB fejlesztési környezetet. Az eljárásom hasonló pontok keresésén alapul, a napsugárzás, hőmérséklet és szélsősebesség alapján kiszűri a vizsgált negyed-órához hasonló korábbi elemeket, majd a hasonlóság mértéke alapján jóslatot ad a termelésre.

A kutatási munkám második felében a Python programozási nyelvvel kezdtem el foglalkozni, azzal a céllal, hogy az évtizedek óta fejlesztett gépi tanulási algoritmusokat alaposabban megismerjem, és ezek használatát elsajátítsam.

A vizsgálathoz kétféle időjáráselőrejelzést használtam az OMSZ jóvoltából, ezek közül az egyik day-ahead adatsor volt, azaz olyan becslés, amelynél a kiadás és az időszak kezdete között 24 óra különbség van, míg a másik intraday változatnál ez a különbség nulla. Jelenleg még mindkét előrejelzés típus egyaránt fontos, ezért a módszereimet mindegyikre teszteltem a fejlesztés során folyamatosan.

6.2. Javaslatok, következtetések, tanulságok

A saját modellem elkészítése során több különböző paraméter változtatásával próbálkoztam optimalizálni az előrejelzést. Ezek közül kiemelném a hőmérséklet értékek „csúsztatásának” pozitív hatását. Ehhez kapcsolódóan a tanulság az, hogy az említett természeti tényezők egy órával későbbi értékét figyelembevéve az előrejelzés pontossága növelhető.

Megfigyeléseim alapján kijelenthetem, hogy azok az előrejelzések, amelyek jobbnak bizonyultak a betanulás alatt, egytől egyig konzervatív becslést adtak, ami azt jelenti, hogy a jóslott és a valós tartamdiagramokat összevetve az eljárások mindig az átlagos termelés irányába „csaltak”. Ezzel kapcsolatban elvégeztem egy korrekciós eljárást, ami a két tartamdiagramot azonossá tette, azonban ez vegyes eredményeket adott, az átlagos abszolút hibát csökkentette, ugyanakkor az RMSE értékére negatív hatással bírt.

Az előrejelzés javítható még szeparálással, azaz a napokat felhőzöttség alapján külön optimalizált algoritmus használatával. Az általam erre a célra bevezetett módszer sikeresnek bizonyult. Hasznosnak tartanám egy olyan értékelési kör beépítését, amely gépi tanulóval klasszifikálná a napokat nem túl sok, maximum 3-4 típusra, amelyekre majd külön optimalizált algoritmusokkal lehetne menetrendezni.

A módszerek eredményeinek összehasonlításánál azt tapasztaltam, hogy vannak módszerek, amelyek rRMSE, és vannak, amelyek rMAE alapján jobbak inkább. A legjobb négy algoritmusnak az SVM, MLP, GBR és a k-NN adódott. A módszerekből képezhetőek olyan hibrid eljárások, amelyek javítják a pontosságot, azonban ezeknek sem létezik egyértelmű optimuma, rendre Pareto görbéket produkáltak.

Tapasztalataim szerint a kiértékelés problémája az egyik legfontosabb kérdés a témában. Ameddig a piaci cél, a költségcsökkentés nem fogalmazható meg egyértelmű módon a menetrendezés pontosságára vonatkozóan, addig nem lehet eldönteni, hogy melyik a legjobb módszer. Ugyanígy fontos kérdés a menetrendfejlesztés beárazása, azaz, hogy mit ér adott százaléknyi javulás. Ennek értékére lehet közelítő becslést adni az 5.10-es pont alapján, amelynek validáláshoz azonban a rendszer működéséről több adat kell.

Véleményem szerint a módszerek optimalizálásának egy adott időjárás előrejelzés esetén véges határa van, amelyen túl csak az időhorizont csökkentésével és az előrejelzés pontosításával javítható a menetrend.

Amennyiben a lehető legjobb menetrendet szeretnénk elérni, azt a módszerek argumentumainak és prediktorainak megfelelő beállításával, eljárások kombinálásával, terület index alapú szeparálással, és a hajnali, esti szakaszok korrekciójával érdemes optimalizálni a célfüggvény ismeretében. Habár más adatsorokkal nem volt alkalmam dolgozni ezen kutatás keretein belül, azt gondolom, hogy a korábban említett kiemelkedő módszerek vélhetően más parkok esetében is jobban funkcionálnának, azonban a paramétereket esetleg kell meghatározni, különböző előrejelzésekre, napelem típusokra és méretre nem lehet általánosítani. Erre az adott napelempark – szolgáltató párosra egy SVM – GBR – knn kombinációt javasolnék terület index alapú szeparálással az aktuális hazai fel és leszállási árak költségének arányában wMAE-re optimalizálva.

Jövőbeni célként megfogalmaztam, hogy kidolgoznék egy olyan módszert, amely nem csak negyedórás, hanem hosszabb távú hasonlóságot vizsgálna a forrásjegyzékben megjelölt cikkhez [21] hasonlóan. A kutatást a jövőben nem csak egy, hanem több napelempark adatai alapján szeretném folytatni, a portfólió szintű menetrendezés hatásának vizsgálatára. Továbbá szeretnék különböző időjárás előrejelzéseket összehasonlítani, összefüggések után kutatva a maximális pontosság és az előrejelzés jellege között.

7. FELHASZNÁLT FORRÁSOK

1. IEA (2017): *Data – Renewables and waste*, [Link](#), dátum: 2020.10.24.
2. Nemzeti Fejlesztési Minisztérium (2012): *Nemzeti Energiastratégia 2030*. [Link](#), dátum: 2020.10.24.
3. Innovációs és Technológiai Minisztérium (2020): *Nemzeti Energia- és Klíma Terv (2030-40)*
4. Dr. Kaderják Péter (Országgyűlés Fenntartható Fejlődés Bizottsága, ITM) (2019): *Tájékoztató a Magyar Kormány energia- és klímapolitikai irányairól, céljairól, tevékenységéről*, [Link](#), 2020.10.24
5. MAVIR Zrt. (2020): *Megdőlt a tavaszi napenergia rekord*, [Link](#), dátum: 2020.10.24.
6. MAVIR Zrt. (2020): *Újabb rekordok születhetnek a folyamatos napos időben* [Link](#), dátum: 2020.10.24.
7. Herczeg Sándor (2020): *Fogyasztás és termelés hazai és régiós változása a pandémiás időszakban*. REKK Workshop 2020. április 22. A koronavírus hatása az energiapiacokra
8. Boncz Norbert (2019): *Napelemek matematikai és statisztikai alapú termelés előrejelzésének összehasonlítása*. Szakdolgozat – BME Energetikai Gépek és Rendszerek Tanszék
9. J. Antonanzas, N. Osorio, R. Escoba, R. Urraca, F.J. Martinez-de-Pison, F. Antonanzas-Torres (2016): *Review of photovoltaic power forecasting*. Solar Energy 2016/ 78-111, DOI: 10.1016/j.solener.2016.06.069
10. Rajesh S. Brid: *Decision trees*. [medium.com](#), [Link](#), dátum: 2020.05.21.
11. Ashish Patel (2019): *Ensemble learning - The heart of Machine learning*. [medium.com](#), [Link](#), dátum: 2020.10.24.
12. Jason Brownlee (2016): *A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning*. [medium.com](#), [Link](#), 2020.10.24.
13. Leo Breiman (1997): *Prediction games and arcing algorithms*, Statistics Department University of California Berkeley, CA. (4720), [Link](#), 2020.10.24.
14. William Koehrsen (2017): *Random Forests simple explanation*. [medium.com](#), [Link](#), dátum: 2020.10.24.
15. JJ: *MAE and RMSE — Which Metric is Better*. [medium.com](#), [Link](#), dátum: 2020.10.24.
16. MAVIR Zrt.: *Kiegyenlítő energia elszámolás*, [Link](#), dátum: 2020.10.24.
17. MAVIR Zrt. (2020): *KAT_es_premium_tamogatas_statisztikai_adatai_202008*
18. Dazhi Yang et al. (2020): *Verification of deterministic solar forecasts*. Solar Energy 2020/20-37, DOI: 10.1016/j.solener.2020.04.019 [Link](#)
19. J. Antonanzas et al. (2020): *Influence of electricity market structures on deterministic solar forecasting verification*, Solar Energy 2020/44-46, DOI: 10.1016/j.solener.2020.04.017 [Link](#)

20. Wikipédia: Egyezményes koordinált világidő, [Link](#), dátum: 2020.05.21.
21. Dazhi Yang, Elynn Wu, Jan Kleissl (2019): *Operational solar forecasting for the real-time market*. International Journal of Forecasting 2019/1499-1519, DOI: 10.1016/j.ijforecast.2019.03.009

8. SUMMARY

Photovoltaic (PV) energy production is under great development, the installed capacity is rising exponentially worldwide, gaining higher share from the total portfolio of sources. Also according to Hungarian energy strategy, PV technology plays a really important role in accomplishing the various climate protection purposes, this is one of the reasons why the topic of balancing and decreasing the fluctuating production of weather-based energy sources, without risking the safety of supply, is so actual and widely researched.

This research is focused on the statistical forecasting method, some machine learning algorithms are tested and compared with each other and a self-made one. The predictions are based on the 15 minutes resolution day-ahead or intraday meteorological forecast data (irradiation, wind and temperature), paired with the real production statistics of an under 500 kW PV park. The nightly and under upkeep intervals are removed from the observations. The procedure developed in Matlab searches for intervals with similar weather in the same period of the year as the observed date and predicts by the weighted average of the collected quarters. The results of this are compared with the optimized built-in algorithms of Python's scikit library (LR, DT, SVM, k-NN, ABR, GBR, MLP), some mixtures and one method, that predicts after classifying the days to clear and cloudy ones, with the help of an introduced index number.

The field of comparison and evaluation is heavily researched, since there is no generally accepted and current method for verification of schedules, made for PV parks with different location size, and meteorological data supply. According to this, the methods are only compared to each other relatively. Also the aims are undefined, since the transmission system operator observes the PV production on the level of balance group, while the licensees want to reduce the monthly billed cost of balancing power, which has a complex calculating methodology with many drivers, hence it is impossible to create an explicit formula to convert deviation to money. Based on the connecting researches, I used mean absolute error and root mean square error normalized with the average production for evaluation, and weighted mean absolute error for observing the algorithms in different regulation systems.

The best algorithms happened to be SVM, MLP, GBR and k-NN. The bias can be decreased by adding the data of nearby quarters to the predictors. By taking the weighted average of the results of different methods, the scores can be increased, but I have not found an ultimate best one for each error, all the mixtures ended up a Pareto curve.

The field is open to various kind of researches, and the methods can be further developed. I see the highest potential in the decrease of time horizon and simplifying the evaluation to fixate the aims.