SZENT ISTVÁN EGYETEM

Folyóvizek oldottoxigén-koncentrációjának becslése neurális hálózatokkal

Doktori (PhD) értekezés

Csábrági Anita

Gödöllő

2019

A doktori iskola megnevezése:	Műszaki Tudományi Doktori Iskola
tudományága:	Agrárműszaki tudományok
vezetője:	Prof. Dr. Farkas István egyetemi tanár, DSc SZIE, Gépészmérnöki Kar
Témavezető:	Dr. Molnár Sándor egyetemi tanár, DSc SZIE, Gépészmérnöki Kar, Mechanikai és Géptani Intézet
Társ-témavezető:	Dr. habil. Kovács József egyetemi docens ELTE, Természettudományi Kar, Általános és Alkalmazott Földtani Tanszék

....

Iskolavezető jóváhagyása Témavezetők jóváhagyása

TARTALOMJEGYZÉK

JELÖLÉSJEGYZÉK

Jelölések:

ANN	Artificial neural network, mesterséges neurális hálózatok	[-]
ANFIS	Adaptive neuro-fuzzy inference system, adaptív neuro-fuzzy következtető rendszer	[-]
ANOVA	Analysis of variance model, Varianciaanalízis	[-]
ARIMA	Autoregressive integrated moving average model, Autoregresszív integrált mozgóátlag modell	[-]
BASINS	Better Assessment Science Integrating Point and Non-point Sources	[-]
BOD	Biological oxygen demand, biológiai oxigénigény	[mg L^{-1}]
CCDA	Combined cluster and discriminant analysis, kombinált klaszter- és diszkriminancia-analízis	[-]
COD	Chemical oxygen demand, kémiai oxigénigény	[mg L^{-1}]
DO	Dissolved oxygen, oldott oxigén	[mg L^{-1}]
E-ANN	Ensemble neural network, együttes tanulást megvalósító neurális hálózat	[-]
EC	elektromos vezetőképesség	[μScm^{-1}]
EFDC	Environmental Fluid Dynamics Code	[-]
EOV	Egységes országos vetületi rendszer	[-]
fkm	Folyamkilométer	[-]
GA	Genetic algorithm, genetikai algoritmus	[-]
GRNN	Generalized regression neural network, általánosított regressziós neurális hálózat	[-]
IA	Index of agreement, Willmott-féle egyezési index	[-]
MAE	Mean absolute error, átlagos abszolút hiba	[-]
MAPE	Mean absolute percentage error, abszolút százalékos hiba	[-]
MI	Mesterséges intelligencia	[-]
MLPNN	Multilayer perceptron neural network, többrétegű perceptron neurális hálózat	[-]
MLR	Multivariate linear regression, többváltozós lineáris regresszió	[-]
MRSM	Modified response surface method, módosított válaszfelület módszere	[-]
MSE	Mean square error, átlagos négyzetes hiba	[-]
PLS2	Partial least squares regression, PLS regresszió	[-]

Q	vízhozam	$[m^3s^{-1}]$
QUASAR	QUAlity Simulation Along Rivers	[-]
QUESTOR	Quality Evaluation and Simulation Tool for River Systems	[-]
R^2	Determinációs együttható	[-]
RBFNN	Radial basis function neural network, Radiális bázis függvényes neurális hálózat	[-]
RMSE	Root mean square error, az átlagos négyzetes hiba gyöke	[-]
RNN	Recurrent neural network, visszacsatolt neurális hálózat	[-]
SVM	Support vector machine, szupport vektor gép vagy tartó vektor gép	[-]
SWAT	Soil and Water Assessment Tool	[-]
T_w	víz hőmérséklete	$[^oC]$
VIF	Variance inflation factor, variancia inflációs faktor	[-]
WASP	Water Quality Analysis Simulation Program	[-]

1. BEVEZETÉS, CÉLKITŰZÉSEK

Az értekezésem első fejezetében a téma jelentőségét fogalmazom meg, illetve a munkám célkitűzéseit mutatom be.

1.1. A téma időszerűsége és jelentősége

Az utóbbi évszázadban a növekvő ipari és mezőgazdasági tevékenység, az urbanizáció következtében természetes vizeink jelentős szennyeződéseknek, antropogén terheléseknek vannak kitéve, melyeknek során vizeink fizikailag leírható ökológiai állapota (pld. turbiditás, fényklíma, áramlási sebesség, hőmérséklet stb.) drasztikusan megváltozik. Ezen antropogén hatások megfigyelése és ellenőrzése szempontjából nagyon fontos és elengedhetetlen az adott rendszer vízminőségi változóinak ismerete, melyet a monitoring hálózatok adatai biztosítanak. A monitoring hálózat optimális működése több szempontból (pld. költséghatékonyság) vizsgálható, vagy adott esetben javítható, ha bizonyos, nehezen vagy költségesen mérhető paramétereket, más egyéb könnyen mérhető paraméterekkel becsüljük, melynek egy adekvát eszköze a mesterséges intelligencia, vagy a mesterséges neurális hálózatok nyújtotta módszerek.

Az elmúlt évtizedekben ugrásszerűen megnőtt a mesterséges neurális hálózatok alkalmazása a vízminőségi változók becslésére mind a tavakban, mind a folyókban egyaránt a neurális hálózatok előnyei miatt. A neurális hálózat alkalmazásakor csak a bemenő adatokra van szükség mindenféle előfeltétel nélkül, illetve ezek a modellek képesek a bemenő és a kimenő adatok közötti komplex kapcsolatot föltérképezni, majd a tapasztalatokat felhasználva megfelelő módon általánosítani.

A legtöbb neurális hálózatokkal való alkalmazás a természetes vizek oxigén-háztartására jellemző kémiai paramétereket vizsgálják, és ezen belül is leggyakrabban az oldottoxigén-koncentrációját (DO) becsüli, mivel ez az egyik legfontosabb vízminőségi paraméter az élővizekben, és ez az egyik meghatározó mutatója a felszíni vizek ökológiai egyensúlyának.

A folyókban lévő oldott oxigén alacsony szintjének, illetve extrém esetben az oxigénmentes állapotnak a hatását, mely fölborítja az ökoszisztémát, tömeges halpusztulást, szagokat és egyéb esztétikai kellemetlenségeket okozhat, már egy évszázada fölismerték. Az oldott oxigén szintjének dinamikája nagyon komplex; fizikai, kémiai és biológiai folyamatok kölcsönhatásait hordozza magában, melynek eredményeképpen kialakul egy dinamikus egyensúly az oxigéntermelő és az oxigénfogyasztó folyamatok között. A természetes vizek egyensúlyi állapotban öt és tizenöt mg L^{-1} közötti oldott oxigént (O_2) tartalmaznak, ez az érték függ a víz hőmérsékletétől, a víz sótartalmától és a földrajzi magasságtól. Az oldott oxigéntartalom kritikus értéke alatt a felszíni vizek ökológiai egyensúlya fölborul, amely az élővilág pusztulásához vezet.

A DO-szintet becslő neurális hálózatok alkalmazásának egyre nagyobb elterjedése ösztönözte azon törekvésemet, hogy hazai környezetre, Magyarország két legnagyobb folyójára is adaptáljam ezen modelleket úgy, hogy a kapott eredmények, összefüggések más folyók esetében is fölhasználhatóak legyenek.

1.2. Célkitűzések

Értékezésem célja, hogy a becslési eljárásokat megvalósító neurális modelleknek minél hatékonyabb alkalmazásának lehetőségeit mutassam be úgy, hogy példákat adjak arra, melyik modellt milyen módszerrel, milyen beállítással érdemes alkalmazni folyóvizek oldottoxigén-koncentrációját becsülve. Ennek bemutatására Magyarország két legnagyobb és legfontosabb folyóját választottam ki mintaterületnek, hiszen legjobb tudomásom szerint ezen folyók magyarországi szakaszának oldottoxigén-koncentrációjának becslésével még nem foglalkoztak. A kutatás során háromféle mesterséges neurális hálózatot és a többváltozós lineáris regressziót alkalmaztam.

Célom kideríteni mindkét folyó vizsgálatánál, hogy melyik modell adja a leghatékonyabb becsléseket az oldottoxigén-koncentrációra vonatkozóan.

Az oldottoxigén-koncentrációra vonatkozó időbeli előrejelzésnél célom

- megvizsgálni az antropogén hatásoktól mentes, zavartalan és zavart mintavételi pontoknál a lineáris modellhez képest hány százalékos javulást tudok elérni a neurális hálózatokkal, illetve függ-e a becslés attól, hogy a folyó egy-egy állomását milyen hatások érik.

Az oldottoxigén-koncentrációra vonatkozó térbeli előrejelzés és optimalizáció esetén célom

- a Tisza folyó összes magyarországi mintavételi pontjainak adatait földolgozva a folyó egyes térbeli szakaszait külön-külön jellemezni aszerint, hogy mennyire hatékony becslést tudok adni az adott folyószakasz oldottoxigén-koncentrációjára a különböző modellekkel.
- mindkét folyó homogén csoportosításának eredményeit fölhasználva megvizsgálni, melyik adatstruktúra lesz a leghatékonyabb a becslés szempontjából.

2.SZAKIRODALMI ÁTTEKINTÉS

Doktori értekezésemnek ebben a fejezetében bemutatom azokat a nemzetközi és hazai tudományos eredményeket, melyek a természetes vizek vízminőségét jellemző paramétereit becsülik különböző modellekkel.

2.1. Monitoring rendszerek jelentősége

Az emberiség utóbbi másfél évszázadának ipari tevékenysége, a mezőgazdaságban a műtrágyák használata, az urbanizáció hatása és a fosszilis energiahordozók fokozott használata miatt nagy változás ment végbe a föld légkörében, éghajlatában és a felszíni vizek minőségében is. A víztestek minőségét leíró paramétereinek ismerete, fokozatos ellenőrzése egyre fontosabb kérdés a klímaváltozás tükrében is. A fokozatosan végbemenő klímaváltozás, környezetszennyezés hatásai pld. a globális átlaghőmérséklet emelkedése, az óceáni áramlási rendszerek megváltozása, a biológiai sokféleség eltűnése, a csapadékeloszlás változása, az üvegházhatású gázok mennyiségének növekedése a légkörben (Padányi és Halász, 2012). Ez utóbbi hatások ellen mitigációs intézkedéseket vezettek be a világ sok országában (Kiotói Jegyzőkönyv, 1997), melynek egyik fontos eleme a fosszilis tüzelőanyagok használata során keletkezett üvegházhatású gázok kibocsátásának csökkentése (Fekete-Farkas et al., 2010) illetve ezen energiahordozók kiváltása alternatív energiaforrásokkal (Molnár S. et al., 2011c). A mitigációs intézkedések érintik mind az ipar, mind a mezőgazdaság (Molnár et al., 2011b), mind a közlekedés (Molnár et al., 2009), mind az energiahordozók felhasználásának (Molnár M. és Csábrági, 2011b) területét. Ez utóbbi igen fontos terület, hiszen az energiafogyasztás mértéke hatással van a gazdaság növekedésére is (Molnár M. et al., 2012; Molnár M. és Csábrági, 2011a).

Az utóbbi évszázadban a növekvő ipari és mezőgazdasági tevékenység, az urbanizáció hatása következtében természetes vizeink jelentős szennyeződéseknek vannak kitéve, melyek vagy pontforrás eredetűek, vagy diffúz szennyeződések. Első típusra az ipari tevékenységből származó nehézfém-szennyeződések (tiszai cianid-szennyezés, Soldán et al., 2001), a nagyobb városok kommunális szennyvíz-bevezetése, olajvezeték meghibásodása miatti talajvízszennyezés nyújtanak példát. Az utóbbi esetre példa a nitrogén és foszforvegyületek beáramlása a műtrágyák használata során, szabálytalan hulladéklerakóból talajvízbe mosódó toxinok, mint például mikroműanyagok beáramlása a vízbe (Lechner et al., 2014). Az 1980-as 1990-es évek fordulóján bekövetkezett európai környezetpolitikai szabályozások és kelet-európai gazdasági változások miatt a műtrágya-felhasználás jelentősen csökkent, amely együtt járt a folyók tápanyag koncentrációjának csökkenésével (Hatvani et al., 2015, Mander és Forsberg, 2000), ezáltal illetve a mérséklődő ipari tevékenység miatt a víz minőségének javulásával is.

Ezen antropogén hatások megfigyelése és ellenőrzése szempontjából nagyon fontos a monitoringrendszerek megfelelő kiépítése. A monitoringrendszerek egyik fő célja a víz szennyezettségének ellenőrzése és ezáltal a víz minőségének megőrzése úgy, hogy mind térben, mind időben megfelelő mérési adatokat nyújtsanak és mindezt a lehető legnagyobb költséghatékonysággal tegyék (Chilundo et al., 2008). Ezen feladatok támogatására született meg az Európai Unióban a Víz Keretirányelv (VKI, 2000/60/EC), amely jelenleg is megszabja a vízminőség-védelem célkitűzéseit és előírásait Európában. A monitoring-rendszerek létrehozásának és üzemeltetésének alapja a megfelelő mintázási stratégia kiválasztása, amely során i) meg kell határozni a monitoring-paramétereket ii) a megfelelő

mérési pontokat (térbeliség) iii) a mintavétel gyakoriságát (időbeliség) a megfelelő reprezentativitás érdekében (Füst és Geiger, 2010). Ugyanakkor a rendszeres, jól működő, megfelelő mintavételi stratégiát alkalmazó mérési, monitoring-rendszer nem elég, szükség van az adathalmaz megfelelő szintű feldolgozására, adatelemzésre. Ha a folyamatosan mért vízminőségi adatokból előrejelzéseket, illetve a paraméterek között bizonyos összefüggéseket szeretnénk kinyerni, akkor becsléseket célszerű végezni, amely a vízminőséggel foglalkozó management szakemberei számára hatékony segítség a döntéstámogató rendszerek működéséhez.

2.2. Determinisztikus és sztochasztikus modellek

A vízminőségi paraméterek becslésére, modellezésére két megközelítési módszer terjedt el: a determinisztikus és a sztochasztikus modellek alkalmazása. A vízminőség ellenőrzése, kutatása hosszú múltra tekint vissza, hiszen elsők között 1925-ben modellezték egy folyó (Ohio, USA, Heddam, 2014b) oldottoxigén-koncentrációját és a biológiai oxigénigényét klasszikus differenciálegyenleteket (SP-egyenletek) alkalmazva (Streeter és Phelps, 1925), ami nagyhatást váltott ki a kutatók körében. Később ugyanis ezek az egy-dimenziós, stacionárius egyenletek adták az alapját a determinisztikus, folyamat-vezérelt modelleknek, amelyeket a világ különböző részein fejlesztettek ki (Cox, 2003). Ilyen modellek például a BASINS, EFDC, HSPF, MIKE, QUASAL, QUESTOR, SWAT, WASP (Wang et al., 2013), melyek között vannak dinamikus és többdimenziós modellek is. Az utóbbi évtizedekben ezekhez a hagyományos vízminőségi modellekhez hidrodinamikai transzportmodelleket (Chen és Liu, 2014) és/vagy a klímaváltozás (Molnár et al., 2014a; 2014b), illetve az antropogén hatások figyelembevétele miatt légszennyezettség-modelleket is integráltak, hogy pontosabban tudják szimulálni az ökológiai folyamatokat és így pontosabb becsléseket kapjanak. Ezen kívül megoldották azt is, hogy nemcsak egyetlen pontszerű forrásra lehet alkalmazni a modelleket, hanem sokkal szélesebb vízterületre (Wang et al., 2013). Az angliai Wharfe folyó vízminőségét QUASAR (Quality Simulation Along Rivers, Whitehead et al., 1997) modellel vizsgálták (Eatherall et al., 1998), Hydrological Simulation Program FORTRAN (HSPF, US EPA, 2007) modellt alkalmaztak az Amite folyón, az USA-ban DO becslésére (Patil et al., 2012). Klasszikus és módosított SP-egyenleteket alkalmaztak a pakisztáni Ravi folyón (Haider és Ali, 2010), WASP modellel becsülték a kínai Nanfei folyó oldottoxigén-koncentrációját (Huang et al., 2017). MIKE 11 modellt (Danish Hydraulics Institute, 2001) alkalmazták a kenyai Nzoia folyó középső részének DO és biológiai oxigénigény (BOD) koncentrációjának becslésére (Kanda et al., 2015), a modellt a 2009-es évi adatokkal kalibrálták, és a 2013-as évi adatokkal validálták. Egy évvel később már ugyanezen folyónak DO koncentrációját becsülték a neurális hálózatok alapmodelljével, a Multilayer Perceptron Neural Network (MLPNN, 3.4.2 pont) modellel (Kanda et al., 2016), és megállapították, hogy a MIKE modellt nehéz volt a kenyai viszonyokra adaptálni a mérési adatok hiánya miatt. Vagyis a determinisztikus vízminőségi modellek (folyamat-alapú modellek) hátránya az, hogy nagy mennyiségű bemeneti adatra, információra van szükség (kezdeti és peremfeltételek), amit gyakran nehéz megszerezni (Najah et al., 2011; Šiljić Tomić et al., 2016), illetve a modellek implementálásához elengedhetetlen a nagyfokú tapasztalat és szakértelem, mivel ezek a modellek nagyon komplex, bonyolult rendszerek (Ji et al., 2017).

A determinisztikus modellek mellett a sztochasztikus, statisztikai modellek is megjelentek, mivel felismerték, hogy a természetes vizekben lezajló folyamatokat, a bizonytalan kezdeti és

peremfeltételeket, illetve a mérési hibákkal terhelt mérési eredményeket nehéz egy-egy konstans értékkel jellemezni. Ennek a megoldására célszerűnek tűnt valószínűségi változók, illetve sztochasztikus eljárások használata (Boano et al., 2006). Egyes statisztikai alapú modellek (pld. varianciaanalízis (ANOVA)) alkalmazási feltételei közül viszont gyakran szerepel a normális eloszlás és a lineáris kapcsolat megléte (Najah et al., 2014).

Felszíni vizek esetében az oldottoxigén-koncentrációjának dinamikája lineáris, ha csak a hőmérsékletet vesszük figyelembe. Mindazonáltal főleg folyók esetében az áramlási rendszer, a turbiditás, illetve az antropogén hatások (pl. a szennyvízbevezetés) miatt a DO dinamikája eltérhet a lineáris struktúrától (Chen és Liu, 2014), e változást viszont néhány módszer nem tudja megfelelően kezelni. Az adatsorok gyakran tartalmaznak nemlineáris mintázatokat is, és mivel a lineáris dinamikájú modell (ARIMA) önmagában nem képes föltérképezni a nemlineáris kölcsönhatásokat (Faruk et al., 2010), ezért ebben az esetben hibrid modellt, az ARIMA (autoregresszív integrált mozgó átlag modell) és az MLPNN modell hibrid változatát alkalmazták a törökországi Büyük Menderes folyó DO szintjének becslésére (Faruk et al., 2010).

2.3. Mesterséges neurális hálózatok alkalmazása

Megtapasztalva a hagyományos vízminőségi modellek, illetve a statisztikai modellek hátrányait a vízminőségi változók szimulációjával kapcsolatban, az utóbbi évtizedekben nagyon sok kutató a mesterséges neurális hálózatok (részletesebben 3.4.1 pont) különféle fajtáit választotta a vízminőségi változók becslésére. Tették ezt azért, mert a neurális hálózat alkalmazásakor csak a bemenő adatokra van szükség mindenféle előfeltétel nélkül (Chen és Liu, 2014), illetve ezek a modellek képesek a bemenő és a kimenő adatok közötti komplex, nem lineáris kapcsolatot fölismerni, föltérképezni (Najah et al., 2011), majd a tapasztalatokat felhasználva megfelelő módon általánosítani (Wen et al., 2013).

Kuo és szerzőtársai (2007), illetve Najah és szerzőtársai (2014) szerint is elsőként vízminőséget mesterséges neurális hálózattal vizsgáló alkalmazás egy németországi tó (Saidenbach) adatait használta föl, ahol az algavirágzást (phytoplankton) becsülték meg egy vízminőségi adatbázis alapján (French és Recknagel, 1994). Ezután számos alkalmazás született (Scardi, 1996; Clair and Ehrman, 1996), melyek keretében vízminőséget jellemző paramétereket mesterséges neurális hálózattal modelleztek akár mesterséges vizekben is, például szennyvíztisztító üzemben (Abyaneh, 2014; Heddam et al., 2016), illetve természetes vizekben (2.4 pont).

2.4. Tanítóhalmaz és teszthalmaz kiválasztásának módjai

A neurális hálózatok alkalmazásakor a gépi tanulás során a mintahalmazt legalább két részhalmazra bontják föl: tanítóhalmazra és teszthalmazra (bővebben a 3.4.1 pontban). A modell teljesítménye és általánosító képessége alapvetően függ e halmazok meghatározásától. A mintahalmaz tanító- és tesztelő halmazra való fölbontása háromféleképpen történhet, i) véletlenszerűen osztják föl a teljes adathalmazt tanító és tesztelő halmazra. Ebben az esetben az adott kimenetet szimulálják, modellezik a teljes rendszerre vonatkoztatva. A másik lehetőség: ii) a teljes adathalmazt különböző időintervallumra bontják föl, az egyik, a bővebb időintervallum elemei tartalmazzák a tanítóhalmazt, a szűkebb, későbbi időintervallum elemei pedig a teszthalmaz elemei lesznek, ekkor időbeli előrejelzésről beszélhetünk. A harmadik fölbontási módszer: iii) a teljes adathalmazt szomszédos, összetartozó mintavételi pontonként

választják szét tanító- és teszthalmazra, ekkor térbeli előrejelzés, térbeli jellemzés valósul meg.

Természetes vizek paramétereit mesterséges neurális hálózatokkal modellező alkalmazások közül jó néhányat a következő bekezdésben mutatok be a becsülendő paraméter, output alapján csoportosítva.

A tárgyhoz tartozó kutatások közül egy foglalkozik a teljes nitrogénkoncentráció becslésével (He et al., 2011a), az ammónium-nitrogén szint, és a teljes szerves széntartalom becslésével szintén egy tudományos közlésben foglalkoznak (Burchard-Levine et al., 2014, erről bővebben a 2.4.3 pontban). Némileg többen becsülik a természetes vizek klorofill-a tartalmát (Chen és Liu, 2015; Karul et al., 2000; Kuo et al., 2007; Palani et. al., 2008), melyek közül az első három cikkben tavak adataival vizsgálódtak, míg az utóbbi cikkben tengeri öböl adataival végeztek el becsléseket.

A kutatások zömében többnyire a természetes vizek oxigén-háztartására jellemző kémiai paramétereket vizsgálják, többek között a víz DO-szintjét, a kémia oxigénigényt (COD), illetve a BOD paramétert. Kémiai oxigénigény becslésével két tudományos cikk foglalkozik (Khalil et al., 2012; Talib és Amat, 2012), biológiai oxigénigény modellezésének lehetőségét is néhány cikkben vizsgálják (Dogan et al., 2009; Csábrági et al., 2013; Šiljić Tomić et al., 2016. utóbbiról bővebben a 2.4.3 pontban).

Van példa arra is, hogy egy folyó mindhárom paraméterét (BOD, COD, DO) egyszerre becsülik egy folyón (Emamgholizadeh et al., 2014; Najah et al., 2011). Emamgholizadeh és szerzőtársai (2014) az iráni Karoon folyó 8 mintavételi pontjának 17 éves adatait háromféle modellel: MLPNN-nel, radiális bázisfüggvényes neurális hálózattal (RBFNN), adaptív neuro-fuzzy következtető rendszerrel (ANFIS) dolgozták föl, ahol véletlenszerűen, 80%-20% arányban osztották föl a mintahalmazt tanító és teszthalmazra. A három modell közül az MLPNN modellel érték el a legjobb teljesítményt, és az RBFNN modell adta a legpontatlanabb becslést. Najah és szerzőtársai (2011) a malaysiai Johor folyó négy mintavételi pontjainak adatait modellezték háromféle modellel: MLPNN, E-ANN (együttes tanulást megvalósító NN) és SVM (szupport vektorgép), a kapott eredmények alapján az utóbbit találták a leghatékonyabb modellnek.

Adeniran és szerzőtársai (2016), Basant és szerzőtársai (2010), és Singh és szerzőtársai (2009) egyszerre modellezték egy folyó biológiai oxigénigényét és DO-szintjét MLPNN modellt használva. Míg az előbbiek a nigériai Asa folyón vizsgálódtak hat mintavételi pont adataival és 15 bementi paraméterrel úgy, hogy véletlenszerűen osztották szét a tanító és teszthalmazt, addig az utóbbi két tudományos közlemény szerzői ugyanazon indiai folyó (Gomti) nyolc állomásának tíz éves adatsorát dolgozták föl. Basant és szerzőtársai (2010) véletlenszerűen választották szét a tanítóhalmazt és teszthalmazt, Singh és szerzőtársai (2009) pedig a tíz év adatait időarányosan osztották föl az említett halmazokra, tehát ebben az esetben időbeli előrejelzés valósult meg.

A kutatók zöme viszont csak oldottoxigén-koncentrációt becsült különböző természetes vizekben más-más bemenő paraméterekkel és modellekkel (Ahmed, 2014; Akkoyunlu et al., 2011; Antanasijević et al., 2013; 2014; Ay és Kisi, 2012; Bayram és Kankal, 2015; Csábrági et al., 2017a, 2015a; He et al., 2011b; Heddam, 2014a; Ji et al., 2017; Kanda et al., 2016; Keshtegar és Heddam, 2017; Najah, et. al., 2014; Ranković et al., 2010, 2012; Šiljić Tomić et al. 2018b; Soyupak et al., 2003; Wen et al., 2013). A föltüntetett tudományos közlemények

zömében folyóvizek adataival vizsgálódtak, de akad olyan cikk is, ahol tavak, víztározók oldott-oxigéntartalmát becsülik neurális hálózatokkal (Akkoyunlu et al., 2011;. Ranković et al., 2010, 2012; Soyupak et al., 2003), sőt van példa eutróf tavakban végzett kutatásokra is, bővebben a Csábrági és szerzőtársai (2019b) cikkben. A föntebb fölsorolt kutatások nagyobb részében viszont folyóvizek oldottoxigén-koncentrációját becsülték más-más modellekkel és eltérő bemenő paraméterekkel, így a különböző adathalmaz-fölbontás szerinti csoportosítás alapján tárgyalom részletesebben az említett tudományos kutatásokat.

2.4.1. Véletlenszerű felosztást alkalmazó kutatások

A folyóvizek oldott-oxigénkoncentrációját becslő tudományos cikkek nagyobb részében a tanító- és teszthalmazokat véletlenszerűen osztották föl, ebben a pontban ezeket a kutatásokat mutatom be részletesebben.

Ahmed (2014) a bangladesi Surma folyó négy mintavételi pontjának három éves, havonta mért adatait modellezte kétféle neurális hálózattal (MLPNN, RBFNN). Az eredmények azt mutatták, hogy az RBFNN modell volt a leghatékonyabb, és abban az esetben, ha nemcsak a BOD, hanem a COD is bemenő paraméter volt. Csak GRNN modellt alkalmaztak Antanasijević és szerzőtársai (2014) a Duna szerb szakaszán lévő 17 mintavételi pont 2002-től 2010-ig mért adataira. Először 19 bemenő paraméterrel vizsgálódtak, majd különböző előfeldolgozási és bemenő paraméter kiválasztási stratégiákat (bővebben a 2.4.3 pontban) hasonlítottak össze. Ez utóbbi eredményeképpen 10 paraméter maradt, az előfeldolgozásnál pedig a min-max normalizálás volt a legjobb választás. Az érzékenységi vizsgálathoz Monte-Carlo szimulációs technikát vettek igénybe. Eredményeik szerint a hőmérséklet a legfontosabb paraméter az DO becslése során. Bayram és Kankal (2015) a török Harsit folyó 9 mintavételi pontjának egyéves adataival vizsgálódtak többváltozós lineáris regresszió (MLR) és MLPNN modell segítségével három kombinációban úgy, hogy először a hőmérséklet és a pH érték külön-külön voltak a modell bemenő paraméterei, majd együttesen alkották a bemenetet. A mintavétel 15 naponként történt. Tanító és teszthalmazra való fölbontást bár véletlenszerűen, de mégis irányítottan végezték el úgy, hogy egy-egy állomás adataiból évszakonként véletlenszerűen kivettek egy-két adatot a teszthalmazba, a többi adat maradt a tanítóhalmaz eleme. Természetesen ebben az esetben is a teljes rendszerre modellezték a kimenetet, de az irányított, véletlenszerű kiválasztással biztosították azt, hogy az adathalmaz összes struktúrája bekerüljön a tanítóhalmazba és a teszthalmazba is. Heddam (2014a) négy bemenő paraméterrel (hőmérséklet, pH, elektromos vezetőképesség, vízmélység) becsülte a DO paramétert két modellel (GRNN, MLR) a Klamath folyó (USA) egy állomásának három éves adataival. Az eredményekből pedig azt tapasztalta, hogy a GRNN modell sokkal jobb teljesítményt nyújtott. Kanda és szerzőtársai (2016) a kenyai Nzoia folyó öt mintavételi pontjának 2009-2013 között mért adataira futtatták sikeresen az MLPNN modellt úgy, hogy a 80%-20%-ban osztották föl a tanító és teszthalmazt és négy bemenő paraméterrel dolgoztak (zavarosság, hőmérséklet, pH és vezetőképesség). Keshtegar és Heddam (2017) két nemlineáris modellt az MRSM (Módosított válaszfelület módszere) és az MLPNN modelleket alkalmazták négy, az USGS (United States Geological Survey) adatbázisból vett mintavételi pontra, külön-külön, négy bemenő paraméterrel (vízhozam, pH, zavarosság és az elektromos vezetőképesség). Az eredményekből azt tapasztalták, hogy mind a négy mintavételi pontra kapott becslések pontosabbak voltak az MRSM modellekkel. Najah és szerzőtársai (2014) a malaysiai Johor folyó négy mintavételi pontjainak 1998-2007 év közötti adatait vizsgálták két modellel (ANFIS és MLPNN) úgy, hogy négy bemenő

paraméterük volt: hőmérséklet, pH, nitrát (NO3) és az ammónia (NH3-NL). Érzékenységi vizsgálatot is adtak, amelyben a nitrát fontosságát hangsúlyozták. Az eredmények alapján megállapították, hogy az ANFIS jobb teljesítményt nyújtott. Wen és szerzőtársai (2013) a kínai Heihe folyó három mintavételi pontjának 2003-2008 évek közötti adataival, 8 bemenő paraméterrel futtatták sikeresen az MLPNN-t, melynél a Bayes-féle regularizáció volt a tanítóalgoritmus. Az érzékenységi vizsgálat kapcsán megállapították, hogy a pH, NH_4 és a NO_3 volt a legfontosabb paraméter az DO paraméter becslésében.

2.4.2. Időbeli és térbeli előrejelzéseket megvalósító kutatások

A következőekben az időbeli és térbeli előrejelzéseket megvalósító tudományos munkákat ismertetem bővebben.

Antanasijević és szerzőtársai (2013) háromféle neurális hálózatot (MLPNN, általános regressziós neurális hálózat (GRNN) és visszacsatolt neurális hálózatot (RNN)) és az MLR-t hasonlították össze a Duna szerb szakaszán lévő Bezdan mintavételi hely adataival, négy bemenettel (hőmérséklet, pH, vízhozam és elektromos vezetőképesség). A 2004 és 2008 között mért adatok alkották a tanítóhalmazt, és a 2009-es év adatai alkották a teszthalmazt. Az eredményekből megállapították, hogy a legeredményesebb modell az RNN volt. Ay és Kisi (2012) kétféle neurális hálózatot (MLPNN, RBFNN) és az MLR-t hasonlították össze az coloradói Arkansas folyó két mintavételi pont (upstream, downstream) adataira, ahol négy bemenő paraméter állt rendelkezésre (hőmérséklet, pH, vízhozam és elektromos vezetőképesség). A kapott eredményekből azt tapasztalták, hogy az RBFNN modell jobb becslést adott, mint az MLPNN és az MLR, illetve a vízhozam nélküli becslés esetében jobb eredmények születtek. Végül térbeli előrejelzést is adtak, hiszen az upstream állomás hőmérsékleti és pH adataiból sikeresen becsülték a downstream állomás DO-szintjét mindhárom modellel. He és szerzőtársai (2011b) a napi DO minimumot és a napi DO változékonyságot jelezték előre MLPNN és MLR modellek segítségével a kanadai Bow folyó két mintavételi pontján. Az adatokat 15-30 percenként mérték, a 2006 és 2007 év közötti adatok alkották a tanítóhalmazt, a 2008-as év adatait a teszthalmaz tartalmazta. A napi DO minimum bemenő paraméterei a vízhozam és a hőmérséklet volt, a DO napi változékonyságot pedig a sugárzás, a hőmérséklet és a vízhozam segítségével próbálták becsülni. Mindkét esetben az MLPNN modell fölülmúlta a lineáris modellt. Ji és szerzőtársai (2017) időbeli előrejelzést adtak oldottoxigén-koncentrációjára 11 bemenő paraméterrel a kínai Wen-Rui Tang folyó 8 mintavételi pontjának 2004 és 2008 között mért adatainak segítségével. A 2008-as év adatai alkották a teszthalmazt, négyféle modellel vizsgálódtak (MLR, MLPNN, GRNN, SVM), és a kapott eredményekből az jött ki, hogy az utóbbi modell (SVM) adta a legjobb eredményt. Šiljić Tomić és szerzőtársai (2018b) a Duna szerbiai szakaszán lévő 17 mintavételi ponton mért 17 féle bemeneti adatokból az DO-szintjét becsülték MLPNN segítségével. Az adatokat 2002 és 2011 között mérték, a 2011 év adatai alkották a teszthalmazt (bővebben a 2.4.3 pontban).

A folyóvizek oldottoxigén-koncentrációnak becslésével foglalkozó tudományos közlések további részleteit a 2.1. táblázatban lehet megtekinteni, ahol a kiválasztás módja oszlopban a tanító és a teszthalmaz kiválasztásának módja van megjelenítve. A többféle alkalmazott modell közül félkövér betűtípussal van kiemelve a leghatékonyabb. A legjobb RMSE és a legjobb R^2 értéke a teszthalmazra vonatkozik. A csillaggal (*) jelölt érték MAPE értékként van megadva.

2.4.3. Mintahalmaz optimalizációja

Az ANN használata során a becslés eredményességét alapvetően meghatározza az ANN inputja, azaz a mintahalmaz felépítése. A szakirodalomban több kutató foglalkozik ANN mintahalmaz optimalizációjának lehetőségével, ami azt jelenti, hogy különféle módszerekkel úgy kell szűkíteni a mintahalmazt, hogy az optimalizált, szűkebb mintahalmazra alkalmazott modellekkel kapott becslések hatékonyabbak legyenek, mint az eredeti mintahalmazon elért eredmények.

A mintahalmaz optimalizációja, szűkítése több szempontból is megközelíthető; beszélhetünk i) a bemenő paraméterek optimalizációjáról, vagy kiválasztásáról (Antanasijević et al., 2014; Burchard-Levine et al., 2014), illetve ii) térbeli és iii) időbeli optimalizációról. Burchard-Levine és szerzőtársai (2014) egy kínai folyó két állomásának ammónium-nitrogén szintjének, és a teljes szerves széntartalmának becslésével foglalkoznak úgy, hogy hibrid modellt (GA-ANN) használva 10 bemenő paraméterből kiválasztották azt a hármat, amelyik a legfontosabb a két kimenő paraméter becslésében. GRNN modellt alkalmaztak Antanasijević és szerzőtársai (2014) a Duna szerb szakaszán lévő 17 mintavételi pont adataira, először 19 bemenő paraméterrel vizsgálódtak, majd három bemenő paraméter kiválasztási stratégiát hasonlítottak össze. Az első-két stratégia alkalmazásához nem volt szükség neurális hálózatok alkalmazására, hiszen az első a korrelációs analízisen, a második pedig a variancia inflációs faktor (VIF) meghatározásán alapult, hogy meghatározzák a bemenő változók közötti multikollinearitást (Kroll és Song, 2013). A harmadik stratégia genetikai algoritmus alkalmazásával történt a bemenő paraméterek egyéni szigma-faktorának meghatározásával. A stratégiákkal kapott különböző bementi halmazokra alkalmazott GRNN modell hatékonyságát megvizsgálva végül az első stratégia tűnt a leghatékonyabbnak. Ez alapján tíz paraméter maradt az eredeti tizenkilencből.

Šiljić Tomić et al., (2016, 2018b) cikkeiben mindhárom optimalizációra kapunk példát. Šiljić Tomić és szerzőtársai (2016) a biológiai oxigénigényt becsülték a GRNN modell segítségével a Duna szerb szakaszán mind a 17 mintavételi pontot felhasználva, 18 bemenettel, és 2002-2011 közötti időintervallumon. A 2002 és 2010 közötti adatok alkották a tanítóhalmazt, és a 2011-es adatokat használták teszthalmazként. Először térbeli, majd időbeli optimalizálást hajtottak végre, végül korrelációs analízissel döntötték el, hogy a bemenő paraméterek közül melyek azok, amelyek ténylegesen fontosak a BOD becslésében. A térbeli optimalizálás eredményeképpen két modellt kaptak, az egyik modell (GRNN-1) az első négy mintavételi pontot tartalmazza (Bezdántól Bačka Palankáig), a másik modell (GRNN-2) pedig az ötödik állomástól a 17. állomásig tart (Novi Sadtól Radujevacig). Az időbeli optimalizálásnál a 2007-2011 közötti időintervallumot választották ki. A bemeneti paraméterek optimalizálásánál a GRNN-1 modellből hat paramétert, a GRNN-2 modellből csak három paramétert lehetett kizárni.

Šiljić Tomić és szerzőtársai (2018b) a Duna szerbiai szakaszán lévő 17 mintavételi ponton mért 17 féle bemeneti adatokból az DO-szintjét becsülték MLPNN segítségével. Az adatokat 2002 és 2011 között mérték, a 2011 év adatai alkották a teszthalmazt. A Box-Behnken-féle háromfaktoros háromszintű módszert használták az adatok térbeli, időbeli optimalizálására és a megfelelő bemeneti paraméterek kiválasztására. Az eredményekből arra jutottak, hogy két modellel tudják leírni a Duna szerb folyószakaszát: az első az ún. BPNN-UP modell, mely a fölső 8 állomás adataival (Bezdántól Zemunig), 12 bemenettel és 7 év adataival vizsgálódik, a második pedig ún. BPNN-DOWN modell, mely a maradék 9 mintavételi pont (Pančevotól

Radujevacig) 6 éves adatait vizsgálja 11 bemenő paraméterrel. Mindkét esetben az utolsó, 2011-es év volt a teszthalmaz. Ezen két modell teszthalmazra kapott eredményeit összehasonlították egy előző cikkük (Antanasijević et al., 2014) eredményeivel, ahol 17 állomás 2002-2010 év adataival, és 10 bemeneti paraméterrel vizsgálódtak, és itt is 2011 év adatai alkották a teszthalmazt. Az előbbi modellel egy hajszálnyival (pár századdal) jobb eredményt kaptak az RMSE, R^2 és MAE vonatkozásában. Bár mindkét cikkben ugyanazon folyószakaszon, ugyanolyan időintervallumon, és ugyanazoknak az állomásoknak a mért adataival vizsgálódtak mégis más optimalizációs eredményt kaptak, mivel a kimenet (BOD vs. DO) illetve az alkalmazott módszerek is eltértek.

2.5. A szakirodalomi áttekintés összefoglaló értékelése

A kutatási témakörhöz kapcsolódó szakirodalom tanulmányozása során áttekintettem a determinisztikus, vízminőség-modellek kialakulását és fejlődését, illetve a statisztikus modellekkel együtt ezen modellek előnyeit és hátrányait. Ezután bemutattam a neurális hálózatok általános tulajdonságait, majd példákat adtam a neurális hálózatok egyes fajtáinak széleskörű felhasználására tavak, folyók esetleg tengerek különböző vízminőséget jellemző paramétereit becsülve. Ezen alkalmazások rámutattak arra, hogy ezek a modellek nagyon hatékony eszközök folyóvizek oldottoxigén-koncentrációjának becslésére. Ugyanis ezeknek a modelleknek csak bemenő adatokra van szükségük mindenféle előfeltétel nélkül, képesek a bemenő és a kimenő adatok közötti komplex, nem lineáris kapcsolatot fölismerni, majd a tapasztalatokat felhasználva megfelelő módon általánosítani.

A szakirodalom áttanulmányozása alapján megállapítható, hogy a neurális hálók különböző fajtáinak alkalmazása indokolt többféle víztestben, tavakban, folyókban vízminőséget jellemző paraméterek, így például az oldottoxigén-koncentrációjának becslésére. A neurális hálózatok alkalmazása lehetőséget ad arra, hogy ne csak modellezzük a becsülendő paramétert, hanem időbeli vagy térbeli előrejelzést is adhassunk a tanítóhalmaz és a teszthalmaz megfelelő kiválasztásával. Ha a mintahalmazt optimalizálni szeretnénk, akkor lehetőség van arra, hogy akár a bemenő paraméterek halmazát különféle módszerrel, illetve az adatsorok időbeli illetve térbeli homogenitását felfedezve szűkítsük az eredeti, többféle struktúrájú, nagyelemű mintahalmazt azért, hogy még hatékonyabb becslések születhessenek.

A dunai és a tiszai vizsgálatoknál példát adtam mind időbeli, mint térbeli előrejelzésre, illetve a tiszai és a dunai adatok homogén csoportjait figyelembe véve alkalmazást adtam térbeli optimalizációra is. Míg időbeli előrejelzésre számtalan példát találhatunk, térbeli előrejelzésre nem igazán van példa, a tanulmányozott tudományos munkák közül egy foglalkozott térbeli előrejelzéssel, két mintavételi pont alkalmazásával (Ay and Kisi, 2012), folyóvizek DO-szintjét becsülve (2.1. táblázat). Az általam tanulmányozott tudományos munkák közül egy sem volt, amely térbeli előrejelzést megvalósítva becsülte volna folyóvizek több mintavételi pontjának oldottoxigén-koncentrációját neurális hálózatokkal.

A szakirodalmi áttekintés eredményei a dolgozatom célkitűzéseinek véglegesítése során felhasználásra kerültek.

2.1. táblázat Folyóvizek oldottoxigén-koncentrációjának becslése a szakirodalmi áttekintés alapján

Hivatkozás	Folyó	fkm	Állomások száma	Évek száma	Kiválasztás módja	Alkalmazott modellek	Inputok száma	Legjobb RMSE (mg/L^{-1})	Legjobb R^2
Basant et al., 2010	Gomti	500	8	10	véletlen	**MLPNN** + PLS2	11	1,36	0,74
Singh et al., 2009	Gomti	500	8	10	idősoros	MLPNN	13	1,23	0,76
Bayram és Kankal, 2015	Harsit	143	9	1	irányított véletlen	**MLPNN** + MLR	2	0,94	n.a.
Emamgholizadeh et al., 2014	Karoon	n. a.	8	17	véletlen	ANFIS + **MLPNN** + RBFNN	9	3,15	0,85
Ji et al., 2017	Wen-Rui Tang	n. a.	8	5	idősoros	**SVM** + GRNN + MLPNN + MLR	11	0,97	0,86
Ay és Kisi, 2012	Arkansas	-	1	8	idősoros	MLPNN + **RBFNN** + MLR	3	0,55	0,81
Ay és Kisi, 2012	Arkansas	-	1	18	idősoros	MLPNN + **RBFNN** + MLR	3	0,22	0,97
Ay és Kisi, 2012	Arkansas	n. a.	2	n. a.	térbeli	**MLPNN** + RBFNN + MLR	2	0,66	0,85
Antanasijević et al., 2013	Duna	n. a.	1	6	idősoros	MLPNN + **RNN** + GRNN + MLR	4	0,59	0,87
Antanasijević et al., 2014	Duna	588	17	9	véletlen	GRNN	10	0,87	0,85
Šiljić Tomić, et al., 2018b	Duna	253	8	7	idősoros	MLPNN	12	0,81	0,88
Šiljić Tomić, et al., 2018b	Duna	304	9	6	idősoros	MLPNN	11	0,84	0,86
Wen et al., 2013	Heihe	n. a.	3	6	véletlen	MLPNN	8	0,46	0,94
Heddam, 2014a	Klamath	-	1	3	véletlen	**GRNN** + MLR	4	0,69	0,95
Ahmed, 2014	Surma	n. a.	4	3	véletlen	MLPNN + **RBFNN**	2	0,47	0,82
Kanda et al., 2016	Nzoia	n. a.	5	5	véletlen	MLPNN	4	0,59	0,90
Keshtegar és Heddam, 2017	n. a.	n. a.	4	7	véletlen	MLPNN + **MRSM**	4	0,64	0,82
He et al., 2011b DO_{min}	Bow	n. a.	2	3	idősoros	**MLPNN** + MLR	2	0,47	0,90
He et al., 2011b ΔDO	Bow	n. a.	2	3	idősoros	**MLPNN** + MLR	3	1,22	0,72
Adeniran et al., 2016	Asa	n. a.	6	0,5	véletlen	MLPNN	15	9,48	0,90
Najah et al., 2014	Johor	n. a.	4	10	véletlen	**ANFIS** + MLPNN	4	1,6*	0,96
Najah et al., 2011	Johor	n. a.	4	10	véletlen	E-ANN + MLPNN + **SVM**	5	0,35	0,97

3. ANYAG ÉS MÓDSZER

Értekezésemnek ebben a fejezetében bemutatom a vizsgálandó folyórendszerek homogén csoportjait meghatározó eredményeket és a kutatási céljaim megvalósításához alkalmazott modelleket és azok konfigurációjait, illetve a vizsgálatok logikai lépéseit. Ismertetem a célkitűzésemnek megfelelően a két vizsgálandó folyórendszert, az oldottoxigén-koncentráció becslésében lényeges szerepet játszó bemenő paraméterek tulajdonságait, a becsléseknél alkalmazott többváltozós lineáris regressziót, és a neurális hálózatok leggyakrabban használt háromféle fajtáját, majd megadom a modellek hatékonyságának mérésére szolgáló statisztikai mutatókat.

3.1. A Duna és a Tisza folyó

Magyarország két legnagyobb és legfontosabb folyójának oldottoxigén-koncentrációját becsültem a mért bemenő paraméterekből, mellyel legjobb tudomásom szerint ezen folyók magyarországi szakaszán még nem foglalkoztak. A két folyót más-más céllal, és más-más adatfeldolgozási technikával vizsgáltam, ami közös volt a két vizsgálatnál az a bemenő paraméterek halmaza és az alkalmazott időintervallum (1998-2003). Mindkét folyó paramétereinek mért értékeit a VITUKI Kht. (Környezetvédelmi és Vízgazdálkodási Kutató Intézet Nonprofit Kft) bocsátotta rendelkezésemre, melyek laboratóriumai az MSZ EN ISO/IEC 45001:1990 szabvány, illetve 2001-től a MSZ EN ISO/IEC 17025:2001 szabvány követelményeit teljesítették 2005-ig. A vizsgált időintervallumon belül a mérték értékek ugyanazon technikával és mérési eszközökkel lettek meghatározva. A kapott adatoknak az utolsó hat évével vizsgálódtam, vagyis ezek voltak a legfrissebb adataim.

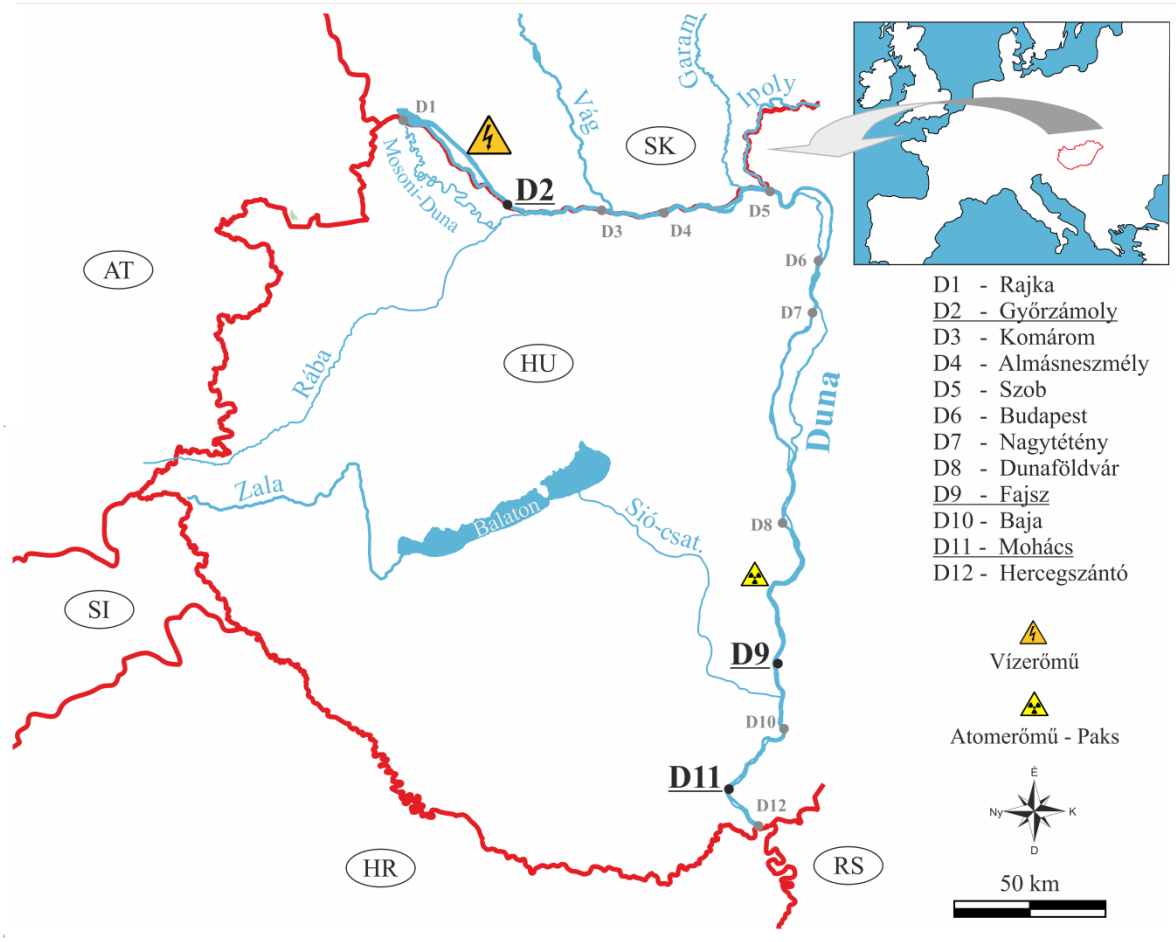
3.1.1. A Duna

A Duna Európa második leghosszabb folyója a Volga után, melynek vízgyűjtőterülete 817 000 km^2, és 2 817 km-t tesz meg a Fekete-erdőtől (Németország) a Fekete-tengerig (Románia). A magyarországi szakasz 417 km hosszú, ahol az átlagos vízhozam 2 000 m^3s^{-1}. A magyar-szlovák határnál épült a Bős-nagymarosi vízlépcsőrendszer, amely nagymértékben megváltoztatta a folyó ezen szakaszát. Ennek eredményeképpen a főmeder vízhozamának 80 százalékát elterelték a szlovák oldalra, és csak mintegy 400 m^3s^{-1} vízhozamú szakasz hagytak meg a magyar területen. A folyó 1 806 folyamkilométerénél (fkm) éri el a főmedret (Kovács et al., 2015b). Még egy további említésre méltó létesítmény is befolyásolja a Duna vizét, ez pedig a MVM Paksi Atomerőmű (1 525-1 526 fkm), melynek hűtővize a Dunába folyik, ezáltal pedig ha nem is számottevően, de növeli a folyó hőmérsékletét, ami csökkentheti a folyó oxigénbeoldhatóságát.

A Duna magyarországi szakaszának főbb természetes mellékfolyói közül négyet érdemes megemlítenünk a folyó folyási irányának sorrendjében: a Rába 27 m^3s^{-1}, a Vág 196 m^3s^{-1}, a Garam 55 m^3s^{-1} és az Ipoly 21 m^3s^{-1} vízhozammal; emellett a Sió-csatorna is igen jelentős. Ez utóbbi egy periodikusan működő csatorna vagy zsiliprendszer, amely Balaton vizét vezeti le a Dunába, vízhozama nagyban függ a tó vízhozamától. 2001 és 2005 között például az átlagos vízhozam csak 20 m^3s^{-1} volt, mert csak kisebb patakok járultak hozzá a csatorna vízhozamához, a Balaton felől nem érkezett vízmennyiség.

A Duna magyarországi szakaszán 12 mintavételi pont van (3.1. ábra). Az első, időbeli előrejelzést megvalósító vizsgálatomhoz ezen állomások közül Mohácsot (D11, 1451,7 fkm) választottam ki, mint “zavartalan” reprezentatív állomást, mivel ezt a mintavételi helyet nem

befolyásolja egyetlen egy természetes (pl. mellékfolyó) illetve más, antropogén hatás (pl. vízerőmű) sem. A D11 mintavételi pont Type 6 csoportba tartozik Sommerhäuser et al. (2003) és Liška et al. (2015) eredményeinek megfelelően. A további két vizsgált mintavételi hely, Győrzámoly (D2, 1806,2 fkm) és Fajsz (D9, 1507,6 fkm), melyek azonban "zavart" mintavételi pontoknak vehetőek, az előzőleg említett csoportosítás eredménye alapján a Type 4, illetve a Type 5 csoportba tartoznak. A D2 az első mintavételi pont azután, hogy a Bős-nagymarosi vízlépcsőrendszer alvizi csatornája újra kapcsolódik a Duna főmedrével, míg a D9-es mintavételi pont a Paksi Atomerőmű után az első mintavételi pont. A másik, térbeli előrejelzést megvalósító vizsgálathoz az összes állomás adataival dolgoztam (3.7.3 pont, 3.13. ábra).



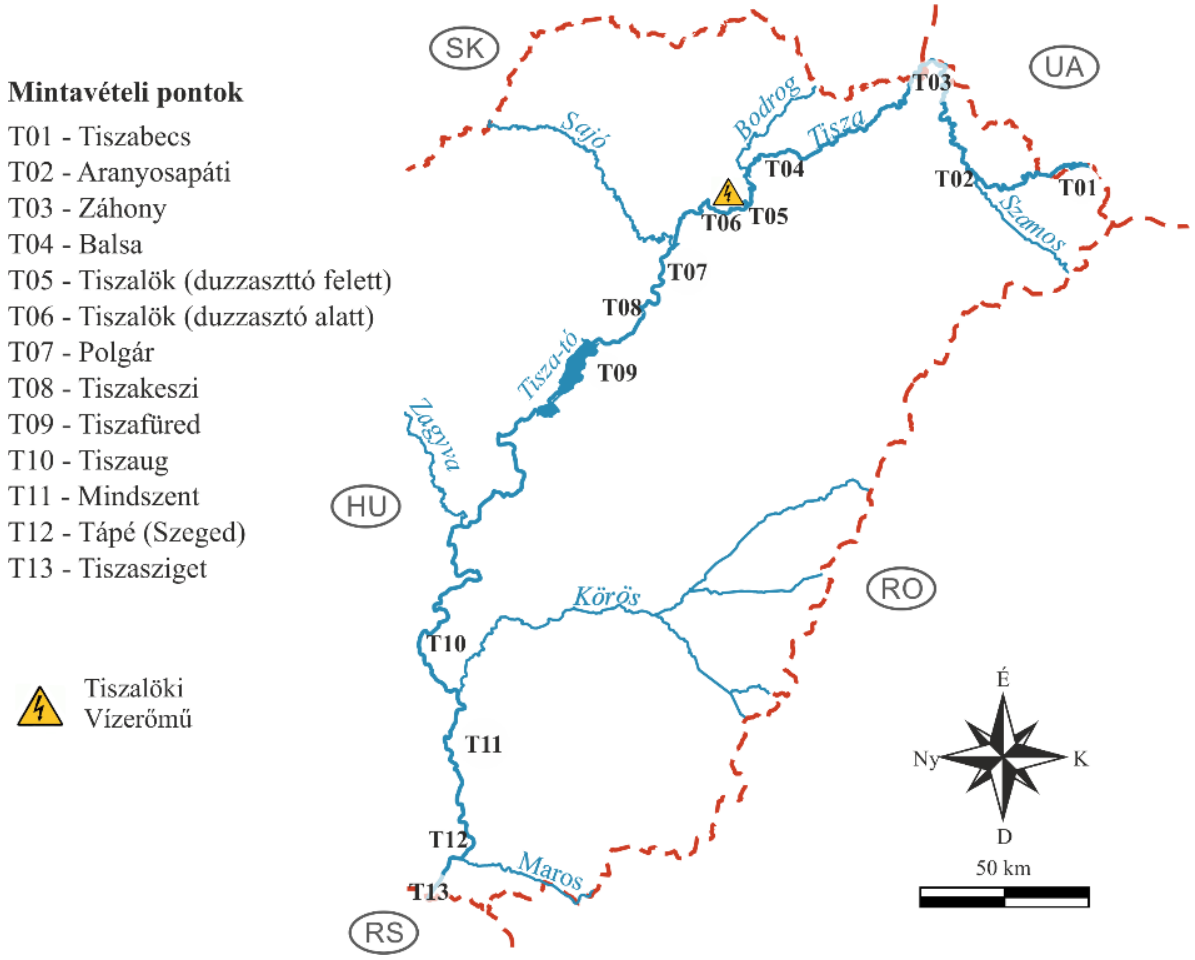
3.1. ábra Duna magyarországi szakasza (Csábrági et al., 2017a)

3.1.2. A Tisza

A Tisza a Duna leghosszabb mellékfolyója, Magyarország második legnagyobb folyója. Vízgyűjtőterülete a Kárpát-medence északi része, és egy nagyon fontos ökológiai folyosó. A Tisza Ukrajnában, a Keleti-Kárpátokban ered, és a Vajdaság közepén, Titelnél ömlik a Dunába. A Tisza vízgyűjtőterülete 157 186 km^2, melynek több mint harmada Magyarországon található (mintegy 47.000 km^2). A folyó teljes hossza 966 km (Sakan et al., 2007), mely öt országon keresztül folyik (Ukrajna, Románia, Magyarország (594,5 km), Szlovákia és Szerbia. A folyó átlagos vízhozama nagyon változató, ugyanis Tiszabecsnél a folyó első magyarországi mintavételi pontjánál 226 m^3s^{-1}, Tiszaszigetnél, az utolsó magyarországi állomásnál pedig 805 m^3s^{-1}. A folyó magyarországi mellékfolyói a folyó

folyási irányának sorrendjében, zárójelben az átlagos vízhozamuk: Szamos (118 m^3s^{-1}), Bodrog (123 m^3s^{-1}), Sajó (32 m^3s^{-1}), Zagyva (4 m^3s^{-1}), Kőrös (116 m^3s^{-1}) és a Maros (188 m^3s^{-1}). Ezek közül a Zagyva kivételével mindegyik külföldön ered. A mellékfolyók vízhozamainak értékeiből kitűnik, hogy a legnagyobb hatása a Tiszára a Szamos mellékfolyónak van, hiszen a Szamos átlagos vízhozama a torkolatnál a Tisza átlagos vízhozamának a fele. Természetesen a többi mellékfolyó is jelentős változást eredményez a folyó természetében.

Nemcsak a mellékfolyók, hanem a mesterségesen létrehozott létesítmények pld. a Tiszalöki duzzasztómű illetve a Tisza-tó is megváltoztatják a folyó vízminőségét (Kentel és Alp, 2013; Moreira és Poole, 1993). A Tisza-tó 1973-ban egy mesterségesen létrehozott tó, melyet a vízerőmű részeként létesítették, így létrehozva Magyarország második legnagyobb tavát. Manapság nagyon népszerű turisztikai célpont és természetvédelmi terület. A tó 27 km hosszú, átlagos mélysége 1,3 m, teljes területe 127 km^2. Ezen pontszerű behatásokon kívül meg kell még említeni a mezőgazdasági tevékenységből fakadó káros tápanyag-beáramlást (Mander és Forsberg, 2000; Molnár et al., 2011a), illetve a nagyobb városok (Szolnok, Szeged) környezetszennyező hatásai is számottevőek, melyek károsan befolyásolják a folyó vízminőségét. A Tisza folyó magyarországi szakaszán 13 mintavételi pont van, ezen állomások mindegyikének adatait fölhasználtam elemzéseim során (3.2. ábra, 3.1. táblázat)



3.2. ábra Tisza magyarországi szakasza (Tanos, 2017)

3.1. táblázat Tiszai mintavételi pontok jelölései, földrajzi paraméterei

Állomás kód	Állomás neve	fkm	EOVX	EOVY
T01	Tiszabecs	757	313555	931595
T02	Aranyosapáti	668,63	324874	890067
T03	Záhony	636,8	345788	881408
T04	Balsa	565	317800	836068
T05	Tiszalök (duzzasztó felett)	525,1	300124	819642
T06	Tiszalök (duzzasztó alatt)	523,1	300419	815511
T07	Polgár	487,2	287048	801740
T08	Tiszakeszi	464,1	272985	796336
T09	Tiszafüred	433,5	256591	776155
T10	Tiszaug	266,4	169753	726219
T11	Mindszent	216,2	132631	735619
T12	Tápé	177,5	101759	739083
T13	Tiszasziget	162,5	93990	731637

3.2. Vizsgált paraméterek

A kapott adatokból a következő paraméterek álltak rendelkezésre: vízhozam (m^3s^{-1}), hőmérséklet (^oC), pH, kálium-permanganátos kémiai oxigénigény (mg L^{-1}), ammónium (mg L^{-1}), nitrit (mg L^{-1}), nitrát (mg L^{-1}), ortofoszfát (μg L^{-1}), biológiai oxigénigény (mg L^{-1}), oldott oxigén (mg L^{-1}), vezetőképesség (μScm^{-1}) és klorofill-a (μg L^{-1}).

Dolgozatomban minden egyes vizsgálatnál a folyók DO-koncentrációját becsültem, mivel ez az egyik legfontosabb vízminőségi paramétere és egyben mutatója a felszíni vizek ökológiai egyensúlyának. A DO-szintjének dinamikája nagyon komplex; fizikai, kémiai és biológiai folyamatok kölcsönhatásait hordozza magában (Wang et al., 2003). Oxigén a vízbe kétféleképpen juthat: a vízben lévő növények fotoszintézése útján, vagy közvetlenül a légkörből, ezek az ún. oxigéntermelő folyamatok, vagyis oxigénforrások (Odum, 1956; Parkhill and Gulliver, 1999; Schurr and Ruchti, 1977). Másfelől azonban oxigént fogyasztanak a vízi növények és állatok légzésük folyamán, illetve az aerob mikroorganizmusok, amelyek oxigént használnak föl a dekompozíció folyamata során (Areerachakul et al., 2011), ezek az ún. oxigénfogyasztó folyamatok, vagyis oxigénnyelők. Egy folyó aktuális oxigénszintje az oxigéntermelő és az oxigénfogyasztó folyamatok közötti egyensúly függvénye (Heddam, 2014a), annak pontos értéke hűen tükrözi a természetes vizek egyensúlyát vagy annak hiányát (Ahmed, 2014). A természetes vizek DO-szintje függ a víz hőmérsékletétől, a víz sótartalmától és a földrajzi magasságtól, ha a felszíni víz egyensúlyi állapotban van, akkor nagyjából 5-15 mg L^{-1} közötti oldott oxigént tartalmaz (APHA, 1998). Ha a DO-koncentráció egy kritikus szint alá süllyed, akkor az a víz élővilágának pusztulásához vezethet.

A bemenő paraméterek megfelelő kiválasztása meghatározó jelentőségű a becslési eljárások esetén, és nagyban függ attól, hogy mi a modellek kimenete. Mivel dolgozatomban az oldottoxigén-koncentrációját szándékozom becsülni, így meg kell vizsgálni, hogy a rendelkezésemre álló paraméterek közül melyek azok a paraméterek, amelyek jelentősen befolyásolják folyóvizek DO-szintjét.

Az oxigén vízben való oldhatósága erősen hőmérsékletfüggő: ahogy a víz hőmérséklete (T_w) nő, úgy csökken a víz oxigén beoldódása, azaz a melegebb víznek kevesebb oldottoxigén-koncentrációja van, mint a hidegebbnek (Bayram et al., 2015). Vagyis a víz hőmérséklete fordítottan arányos a DO-koncentrációjával (Goldman és Horne, 1983). A természetes vizek DO-szintjének hőmérséklettől való függését sok tudományos munkában érzékenységi vizsgálattal is megerősítették (Antanasijević et al., 2014; Ranković et al., 2010; Šiljić Tomić et al., 2018a).

A pH érték egy dimenzió nélküli szám, amely az adott oldat kémhatását (savasságát, lúgosságát) jellemzi, az oldat oxóniumion-koncentrációjának a tízes alapú logaritmusának mínusz egyszeresével adható meg. A szakmai gyakorlatban azonban, a hidrogén- vagy oxóniumion-koncentráció helyett a pH értéket vizsgálják, azzal számolnak, ezt a gyakorlatot követtem én is. A savas esők illetve az ipari, és a kommunális szennyvízzel való szennyezés következtében a folyóvizek pH értéke egyre inkább a savas kémhatás felé mozdul el. Ez viszont káros a mikroorganizmusokra, melyek nagy része így elpusztul, vagyis fölerősödik a bakteriális lebomlás folyamata, ami viszont csökkenti a folyó oxigéntartalmát (Verma és Singh, 2013). Tehát a természetes vizek pH értéke jó mérőfoka a folyók, tavak egészségi állapotának. A természetes vizek oldottoxigén-koncentrációjának pH értékkel való szoros kapcsolatát számos kutatási eredményben is megerősítették (Csábrági et al., 2017a; 2015a; Ranković et al., 2010; Šiljić Tomić et al., 2018a; Wen et al., 2013). Csak e kettő bemenő paramétert (hőmérséklet és pH érték) felhasználva végeztek sikeresen DO paraméterre vonatkozó becsléseket (Bayram és Kankal, 2015).

Nagyon fontos paraméter a természetes vizek oldottoxigén-koncentrációjának becsléséhez a víz vezetőképessége (EC). Ez a tulajdonság a víz elektromos vezetőképességét méri, amely megmutatja azt, hogy a víz milyen arányban tartalmaz oldott sókat (Akkoyunlu et al., 2011). Minél kevesebb a víz vezetőképessége és a hőmérséklete, és minél nagyobb a légköri nyomás annál magasabb a víz DO-szintje (Lewis, 2006). További nagyon jelentős paraméter a vízhozam, amely nagy folyók esetében alapvető fontosságú (Kovács et al., 2015b; Tanos et al., 2015).

E négy paraméter mindezek mellett könnyen és gyorsan mérhető, (Bayram et al., 2015), így adatsoraik általában kevéssé hiányosak. Ezt a négy paramétert számosan választották folyók oldottoxigén-koncentrációjának becsléséhez (Antanasijević et al., 2013; Ay és Kisi, 2012). Előbbi a Duna szerb szakaszán vizsgálódott e négy paraméterrel, és ez motivációt adott arra, hogy én is kipróbáljam ezekkel a paraméterekkel a vizsgálataimat egy kicsit északabbra a magyarországi szakaszokon. A dunai és a tiszai vizsgálataimhoz is az alábbi paramétereket választottam: vízhozam (Q, m^3s^{-1}), hőmérséklet (T_w, ^oC), pH és az elektromos vezetőképesség (EC, μScm^{-1}), melyek könnyen mérhető, nem-specifikus, alapparaméterek. A több, mint tíz bemenő paramétert használó modellekkel szemben a kevés, és alapparamétereket használó modellek előnye, hogy a hiányzó, az oldottoxigén-koncentrációra vonatkozó adatokat könnyen meg lehet becsülni ezen paraméterek alapján (vészhelyzetben, tömeges halpusztulás esetén is). A vízminőséget tanulmányozó monitoring rendszer könnyebben fejleszthető, ha az alapparaméterek nem hiányoznak.

3.3. Adatok szűrése, előfeldolgozás, korreláció

A neurális hálózatok adatvezérelt modellek (3.4.1 pont), ezért olyan bemenő adathalmazt kell biztosítani a működésükhöz, amely minden egyes paraméter esetén rendelkezik mért értékkel. Az adatok szűrése tehát azt jelenti, hogy meg kell vizsgálni az összes bemenő paraméter

esetén rendelkezésre áll-e a mért adat. A nem mért adatokat nullával jelölték, ezeket mindenképpen ki kellett szűrnöm. Ha pedig egy helyen is hiányos volt az adathalmaz, akkor azt az elemet törölni kellett. A neurális hálózat alkalmazásánál nem feltétel, hogy ekvidisztánsak legyenek az adatok. Adathiány miatt történt az, hogy például az eredeti tiszai mintavételi pontok közül Szolnok állomást (335,4 fkm) ki kellett zárnom a vizsgálatból, mert ezen a mintavételi ponton a teljes, vizsgált időintervallumon (1998-2003) hiányoztak a víz vezetőképességére vonatkozó adatok.

A bemenő paraméterek és a becsült, vizsgálandó paraméter mért értékei között lévő lineáris kapcsolatokat is fontos feltérképezni a korrelációs együttható kiszámításával. Az *n* elemű mintában kiszámolt korrelációs együttható (*r*) szignifikancia-vizsgálatához a következő nullhipotézis - H_0: *r=0* - adható, az ellenhipotézis pedig azt jelenti, hogy a két paraméter közötti korrelációs együttható nullától különböző, tehát létezik, szignifikáns. A hipotézisvizsgálat eredménye függ a korrelációs együttható értékétől (*r*), illetve a szabadságfoktól (*f=n-2*). A szignifikancia kiszámításához a következő t-eloszlású statisztikát érdemes használni:

$$t = r\sqrt{\frac{n-2}{1-r^2}}. \tag{3.1}$$

Tehát a képlet alapján az ötszázalékos szignifikancia-szinthez (*p=0,05*), illetve a megfelelő szabadságfokokhoz tartozó t-érték ismeretében meg lehet határozni azt a korrelációs együtthatókra vonatkozó szignifikancia-határértéket ($|r|$), amelytől ha nagyobb a korrelációs együttható, akkor szignifikánsnak számít (4.2.1 és 4.3.1 pontok).

A bemeneti paraméterek általában különböző tartományokat ölelnek föl. Annak érdekében, hogy biztosítsuk, hogy minden bemenő paramétert ugyanolyan mértékben vegyük figyelembe a tanítási folyamat során, a nyers adatokon - az elsődleges adatszűrésen kívül - szinte mindig célszerű valamilyen előfeldolgozást alkalmazni (Maier és Dandy, 2000). Az így földolgozott adatok kerülnek a neurális háló bemenetére. Az előfeldolgozás legelterjedtebb formája a bemeneti adatok skálázása, normalizálása (Dogan et al., 2009). Leggyakrabban a [-1;1] és a [0;1] intervallumot szokás választani.

A másik szokásos eljárás a standardizálás, ahol valamennyi bemeneti paramétert nulla várható értékűvé és egységnyi szórásúvá transzformálunk (Altrichter et al., 2006).

A vizsgálat végén, a különböző neurális hálózatokkal meghatározott kimeneteket megfelelő utófeldolgozással – az előfeldolgozás inverzével - vissza kell transzformálni a kívánt eredménybe (3.3. ábra).

3.3. ábra A gyakorlati problémák tipikus feldolgozási folyamata

3.4. Az alkalmazott modellek

Célom, hogy a folyóvizekben mért oldottoxigén-koncentrációját, mint kimeneti változót becsülni tudjam a mért hőmérséklet, a pH érték, a vízhozam és az elektromos vezetőképesség bemeneti változókból lineáris és nemlineáris modellek (MLPNN, GRNN, RBFNN) segítségével.

Az utóbbi modellek a mesterséges neurális hálózatok rendszerébe tartoznak, amely a mesterséges intelligencia (MI) tudományterületen belül az egyik fő irányvonalnak, a számítási intelligenciának egy eleme. Mesterséges intelligenciának egy gép, program vagy egy mesterségesen létrehozott tudat által megnyilvánuló intelligenciát nevezünk, vagyis kvázi emberi módon való cselekvést, gondolkodást jelent, melyet például a Turing-teszttel (Turing, 1950) lehet ellenőrizni. A számítógép akkor állja ki a próbát, ha az emberi kérdező néhány írásos kérdés föltevése után nem képes eldönteni, hogy az írásos válaszok embertől vagy egy géptől érkeznek-e (Russel és Norvig, 2005).

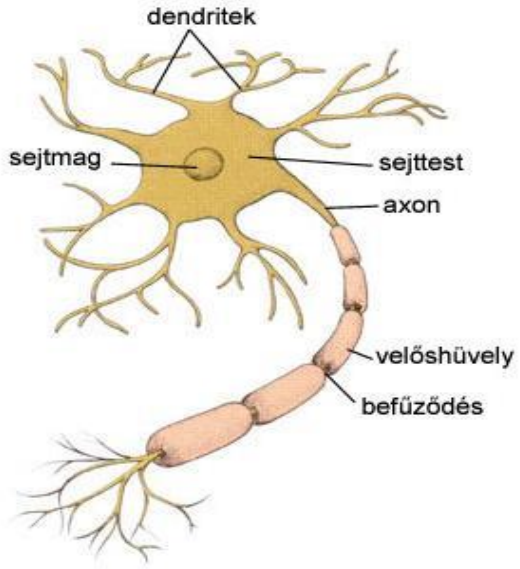
A mesterséges intelligencia a számítógép-tudomány jelentős ágát képviseli, amely intelligens viselkedéssel, gépi tanulással és gépi adaptációval foglalkozik. Két fő irányvonala: a hagyományos MI és a számítási intelligencia. Az előbbit hívják még szimbolikus MI-nak, logikai MI-nak, legfőképpen a gépi tanulás módszereivel foglalkozik, és a következő területeket foglalja magában: szakértői rendszerek, esetalapú érvelés, Bayes-statisztikán alapuló hálózatok és viselkedésalapú MI. Az utóbbi irányvonal inkább az iterációs fejlődést, tanulást helyezi előtérbe, és puha számítási technikai módszereket használ. Ide tartoznak a neurális hálózatok, a fuzzy rendszerek és az evolúciós algoritmusok (pld. genetikai algoritmusok).

3.4.1. Mesterséges neurális hálózatok

A neurális hálózatokat a természettudományi, gazdasági és a műszaki tudományok területén is nagyon gyakran alkalmazzák különböző feladatokra, mint például beszéd-, alak- és karakterfelismerésre, kép- és jelfeldolgozásra, adatbányászatban csoportosításokra, robottechnikában szabályozásokra, meteorológia területén időjárási előrejelzésekre, ipari, gazdasági, és pénzügyi folyamatok időbeli előrejelzéseinek készítésére stb. Ezeket az összetett feladatokat két típusfeladatra lehet bontani, amely az osztályozás (szeparálás) és függvényközelítés (approximáció). Dolgozatomban az utóbbi feladatra használtam a neurális hálózatokat, tehát időbeli illetve térbeli előrejelzéseket adtam, vagy csak modelleztem, becsültem a folyóvizek oldottoxigén-koncentrációját.

A mesterséges neurális hálózat az emberi idegrendszer alapján ihletett gépi tanuló architektúra, számítási matematikai modell, amely a biológia rendszerek olyan tulajdonságait vette alapul, mint a nagyszámú, de egyenként kicsi alapegységekből való felépítés (neuron, ami az idegsejt és az összes nyúlványainak együttese), valamint ezen egységek között lévő sokféle kapcsolatok (szinapszisok) és legfőképp a tanulás képessége. A neurális hálózatok biológiai analógián alapuló szimulációk, hiszen a működésük nagyon hasonlít az emberi idegrendszerbeli neuronok viselkedéséhez. Az idegrendszerbeli neuron részei a sejttest a sejtmaggal - itt történik a tényleges feldolgozása az ingerületeknek - az inputként szolgáló dendritek, valamint a sejttest feldolgozásának eredményét más neuronok felé továbbító axon, mint kimeneti elem. Az axonok ezeket a kimeneti válaszokat más neuron dendritjére továbbítják a szinaptikus bunkócskájukon keresztül (3.4. ábra).

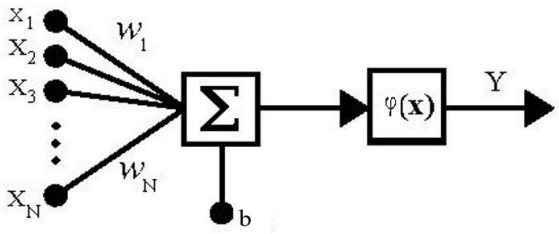
A neurális hálózat nem kívánja a jelenséget modellezni, arra törvényszerűséget megállapítani, hanem a jelenséget fekete dobozként kezeli és csak a bemenő és a kimenő adatokat tekinti (Fazekas, 2013), ezért szokás ezeket a modelleket „adatvezérelt modellnek” (data-driven model) is nevezni.



3.4. ábra Az idegsejt felépítése (Russell és Norvig, 2005)

A bemenő adatokat az $\boldsymbol{x}$ vektor, a mért kimeneti értékeket a $\boldsymbol{v}$ vektor, a becsült, számított értékeket pedig az $\boldsymbol{y}$ vektor jelölje. A jelenséget az $\big(x(1), v(1), \dots x(N), v(N)\big)$ adatok írják le, ezen input-output adatokat a statisztikában mintáknak, a természettudományokban pedig mérési, megfigyelési eredményeknek nevezik. A neurális hálózatok irodalmában ezek a minták a tanítópontok, és a tanítópontok alkotják a tanítóhalmazt. Amikor elkészül egy kívánt neurális hálózat, akkor tesztelni kell a hálózatot olyan új adatokkal, amelyeket még nem „látott" a modell, ezek a pontok lesznek a tesztpontok, és a belőlük alkotott halmaz a teszthalmaz. Nagyon fontos ezen halmazok jó megválasztása, hiszen a neurális hálózatok erőssége a jó tanulási képességben és az adaptációban rejlik (Jang et al., 1997). A modell a teszthalmaz pontjaira ad egy becslést, amit össze lehet hasonlítani a tesztpontok valódi, mért kimeneteivel, és ebből következtetést lehet adni a hálózatok hatékonyságára, általánosító képességére, vagyis az adaptációjára. A 3.6 pontban ismertetem azokat a statisztikai mutatókat, melyek jellemezni tudják a modellek hatékonyságát, általánosító képességét.

A neuron egy információ-feldolgozó egység, a neurális hálózat alapegysége, a neuronok irányított kapcsolatokkal (szinapszisokkal) vannak összekötve egymással. Ezeket a szinapszisokat számszerűsített súlyokkal (weight) tudjuk jellemezni (w_{ji}), amelyek meghatározzák a kapcsolat erősségét és előjelét, illetve kifejezik, hogy a j-edik egységtől az i-edik egység felé vezető kapcsolatról van szó (3.5. ábra). Az egy neuronból álló egységet perceptronnak nevezzük.



3.5. ábra Perceptron felépítése, ahol *b* a torzítás értéke (bias)

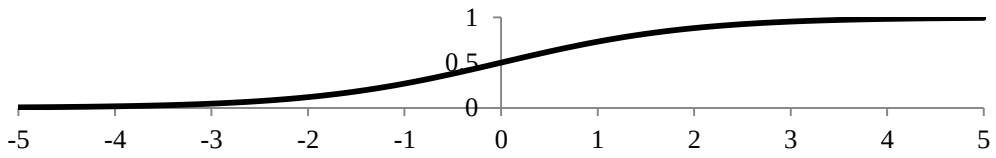
A tanítás során minden egyes perceptron a bemeneteinek egy súlyozott összegét számítja ki (Russell és Norvig, 2005), majd a kimenetét úgy kapja, hogy ezután egy ún. φ aktivációs (transzfer) függvényt alkalmaz a kapott összegre:

$$y_i = \varphi\left(\sum_{i=1}^{N} w_{ij}x_i + b\right). \tag{3.2}$$

Aktivációs függvényként leggyakrabban szigmoid függvényt, tehát *S* alakú függvényt használnak, melyek közül a logisztikus függvény képlete a következő:

$$\varphi(x) = \frac{1}{1+e^{-ax}}, \tag{3.3}$$

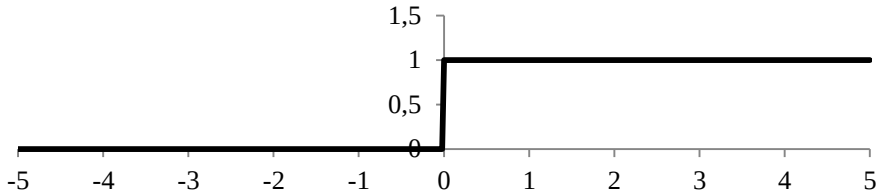
ahol a>0 konstans, $x \in \Re$, a=1 esetén a függvény gráfja az 3.6. ábrán található.



3.6. ábra Logisztikus függvény a=1 esetén

Alkalmazzák még a lépcsős függvényt is, melynek a képlete és gráfja a következő:

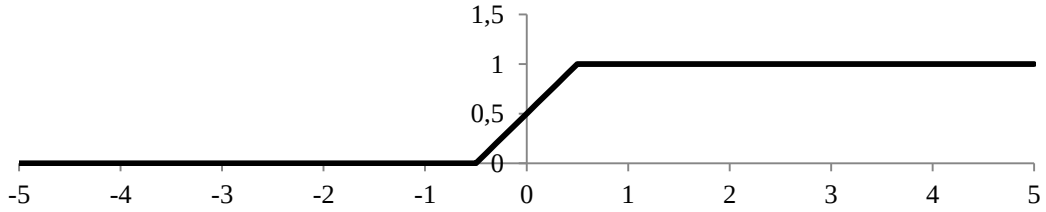
$$\varphi(x) = \begin{cases} 1 & ha\ x \geq 0 \\ 0 & ha\ x < 0 \end{cases}. \tag{3.4}$$



3.7. ábra Lépcsős függvény

A szakaszonként lineáris függvény is gyakran szerepel aktivációs függvényként, képlete és gráfja a következő:

$$\varphi(x) = \begin{cases} 0 & ha \quad x < -0{,}5 \\ x + 0{,}5 & ha \quad -0{,}5 \leq x \leq 0{,}5 \\ 1 & ha \quad x > 0{,}5 \end{cases}. \tag{3.5}$$

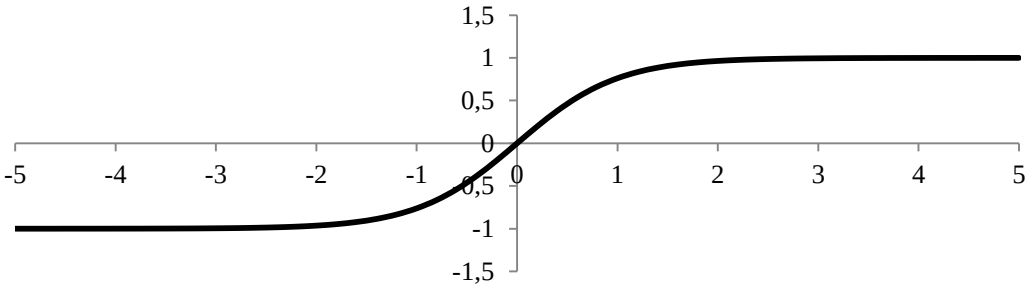


3.8. ábra Szakaszonként lineáris függvény

Ez a három függvény tekinthető valószínűségi eloszlásfüggvénynek is, hiszen mindhárom függvény monoton növekedő, jobbról folytonos, és határértéke a +∞-ben egy, -∞-ben pedig

nulla. Aktivációs függvényként alkalmazhatóak bizonyos függvények, ha ezen tulajdonságok közül egyet megváltoztatunk, vagyis ha -∞-ben nulla helyett -1 a határértéke a függvénynek. Ezen tulajdonságokat pld. a tangens hiperbolikus (3.9. ábra, (3.5)) és az előjel függvény teljesíti. Bizonyos esetekben viszont ezektől a tulajdonságoktól eltérő függvényeket is használnak, mint például a lineáris függvényt (Haykin, 1999).

$$\varphi(x) = \frac{2}{1+e^{-2x}} - 1 = \tanh(x) \tag{3.6}$$



3.9. ábra Tangens hiperbolikus függvény

A neurális hálózatok adaptív, tanuló rendszerek, melyeknek jellemzője az, hogy nem rögzített képességgel rendelkeznek, melyek csak egy adott feladat elvégzésére teszi őket alkalmassá, hanem képességeiket fejleszteni tudják, tehát alkalmazkodni tudnak a változó körülményekhez. A neurális hálózatok tanulása döntően az alábbiak közül kerül ki:

- Ellenőrzött vagy felügyelt tanítás: ebben az esetben a tanítás során be- és kimenet párokat táplálunk be a hálókba, ezek lesznek a tanítópontok, majd a kimeneten mért hiba alapján úgy módosítjuk a hálózatot, hogy a számított hiba csökkenjen, vagyis a tanítás végén a hálózatok egy adott beállítása lesz a neurális háló tudása. Ez a tanítási folyamat általában iterációs algoritmusokkal történik, a tanítópontok egyenkénti, esetenként többszöri fölhasználásával.
- Nem ellenőrzött, felügyelet nélküli tanítás: ebben az esetben nem állnak rendelkezésre az adott bemenetekhez a kívánt válaszok, tehát a hálózatnak az adott bemeneti adatokból kell valamiféle hasonlóságot, egyezőséget keresniük, és ez alapján megpróbálják kategorizálni a bemeneti adatokat úgy, hogy közben a hálózat képes önmaga módosítására, hogy minél sikeresebb legyen ez a folyamat. Emiatt nevezik ezeket a hálózatokat önszerveződő hálózatoknak, ebben az esetben is találkozhatunk iteratív tanító algoritmusokkal. Ilyen hálózatokra példa a Kohonen-háló (Kohonen, 1995).

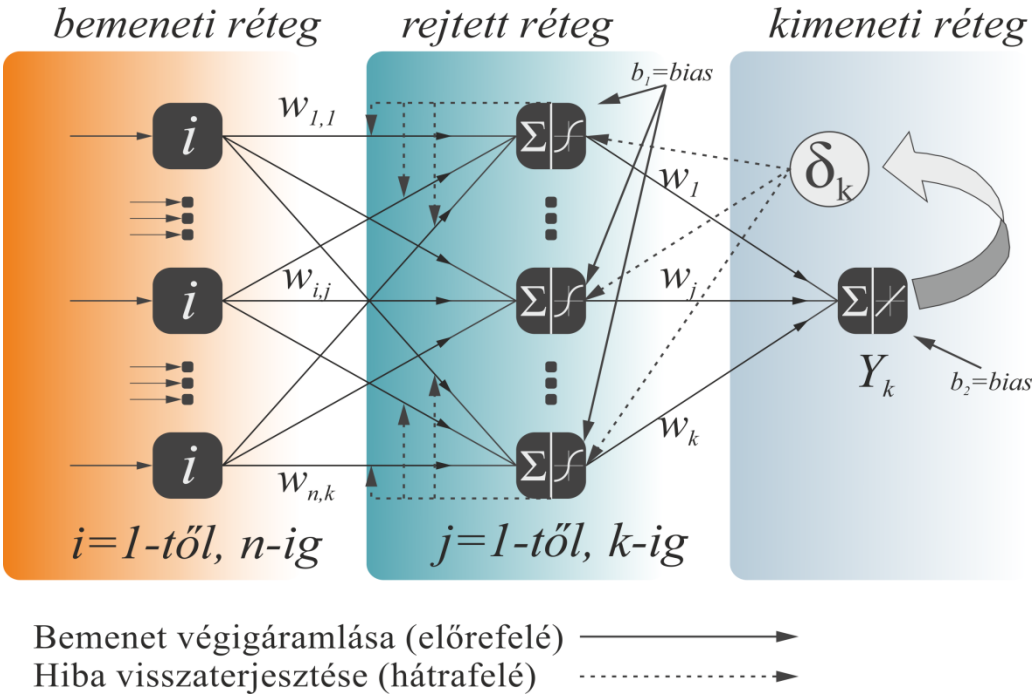
A mesterséges neuronok (node), vagy más néven perceptronok többféle hálózati struktúrában helyezkedhetnek el. A neurális hálózatok topológiáját általában irányított gráffal jellemezhetjük. A gráf csomópontjaiban helyezkednek el a neuronok, míg a kapcsolatokat az inputok és az outputok között a gráf élei mutatják, a bemenettől a kimenet felé irányítva (Altrichter et al., 2006). A neurális hálókat az egyes neuronok közötti összeköttetési mód alapján két fő csoportba sorolhatjuk: előrecsatolt hálózatokról (feedforward networks) és visszacsatolt hálózatokról (recurrent networks) beszélhetünk. Az előrecsatolt hálózat azt jelenti, hogy a jel balról jobbra áramlik, azaz egy adott rétegbeli neuron bemenete a tőle balra lévő rétegbeli neuron kimenete lesz (Fazekas, 2013), vagyis a hálózat gráfelméleti reprezentációja nem tartalmaz hurkot. A visszacsatolt neurális hálózat topológiáját reprezentáló gráf viszont rendelkezik hurokkal, erre példa az RNN modell.

Az előrecsatolt hálózati struktúrán belül az ellenőrzött, felügyelt tanítást (tanítóval történő tanítás) megvalósító háromféle neurális hálózattal vizsgálódtam (MLPNN, RBFNN, GRNN).

3.4.2. Többrétegű perceptron

A mesterséges neurális hálózat típusai közül az MLPNN modell az, amelyik a leggyakrabban használatos, a legjobban kutatott alap neurális hálózat. A modell fejlődésében az igazi áttörést az jelentette, amikor felfedezték a modell tanítására szolgáló eljárást, a hiba-visszaáramoltatási algoritmust 1986-ban (Rumelhart et al., 1986). Azóta a neurális hálózatok elmélete és alkalmazásai hatalmas fejlődést produkáltak (Fazekas, 2013). Ezt a fejlődést elősegítette az is, hogy matematikailag bebizonyították, hogy a szabványos többrétegű előrecsatolt hálózatoknak tetszőleges folytonos függvényre vonatkozó univerzális approximátor-képessége van (Hornik et al., 1989). Az univerzális approximátor-képességhez a hálózatnak legalább egy nemlineáris függvényt felhasználó réteggel kell rendelkeznie.

Az MLPNN modell úgy épül föl, hogy a neuronok, perceptronok rétegekbe szerveződnek, mely rétegek neuronjai csak a szomszédos rétegek neuronjaival vannak kapcsolatban, viszont rétegen belül illetve a távolabbi rétegek között nincs kapcsolódás. A két szomszédos réteg között teljes kapcsolat van, vagyis minden neuron a rétegen belül össze van kötve a szomszédos réteg összes neuronjával. Az MLPNN modellnek általában három rétege van: a bemeneti réteg, a rejtett réteg, amelyből több is lehet, illetve a kimeneti réteg (3.10. ábra).



3.10. ábra Az MLPNN sematikus ábrája (Csábrági et al., 2017a)

A modell tréningezése a hiba-visszaterjesztéses iterációs algoritmussal történik, melynek két lépése van. Az első lépés előrefelé, balról jobbra történik, amelynek folyamán a bemenet (x) végigáramlik a hálózaton, eléri a kimeneti réteget, ahol kiszámításra kerül a modell outputja (Y_k). Ezután a mért érték és a becsült érték különbségéből ki lehet számítani a hibát (δ_k), és elkezdődik az algoritmus második lépése, ami már visszafelé, jobbról balra történik (Emamgholizadeh et al., 2014). Ennek során a kezdőértékekről induló súlyok ($w_{i,j}$) értéke úgy fog változni, hogy minél kisebb legyen a hiba. Az algoritmusnak ez a két lépése egy körforgást ír le, melynek során az adatok egyszer végigáramolnak a rendszeren. Egy ilyen

körfolyamatot epochnak nevezünk. Az algoritmus akkor fog megállni, hogyha az elért hiba egy előre megadott hibahatár alá csökken. Ha túl kicsire választjuk a hibahatárt, akkor nagyon nagy lesz a futási idő, ez pedig problémát okozhat. Ezt a problémát „túltanításnak", „túlilleszkedésnek" hívják, amikor a tanítóhalmaz elemei „túllillesztettek" lesznek, vagyis a hálózat a tanítóhalmaz elemeinek egyéni sajátosságait tanulja be, elveszítve ezzel az általánosító-képességét (Maier és Dandy, 2000). Ez azt jelenti, hogy a tanítóhalmazra kapott hiba elkezd csökkenni, míg a független, a tanítás folyamán nem földolgozott teszthalmaz elemein elért hiba nőni kezd. Ezt a jelenséget elkerülni többféleképpen lehet:

1. kerülni kell a túl bonyolult hálózatok alkalmazását „regularizáció" (egyetlen rejtett réteg és azon belül viszonylag kevés neuron használata (Karul et al., 2000), illetve a tanítóhalmaz mintaszámának csökkentése)
2. csökkenteni kell a futási időt, az algoritmus megállítására (stopping criteria) többféle lehetőség létezik:
   a) az epochszám rögzítése (értekezésemben az epochszámot ezerre állítottam, mint pld. Dogan et al., 2009), vagyis ha már legalább ennyiszer végigáramlottak az adatok a rendszeren, akkor az algoritmus megáll.
   b) a hibahatár csökkentése (Basant et al. 2010).
   c) a teljesítmény minimális gradiensének rögzítése (Wen et al., 2013).
   d) értékelő készlet (kalibráló vagy validálóhalmaz) alkalmazása: A tanítóhalmaz egy részét elkülönítve folyik a tanítás, amelynek során a kiértékelő készletre számított hiba el kezd csökkenni, majd egy bizonyos futási idő után nőni kezd, ekkor kell megállítani a tanítást (Altrichter et al., 2006.).

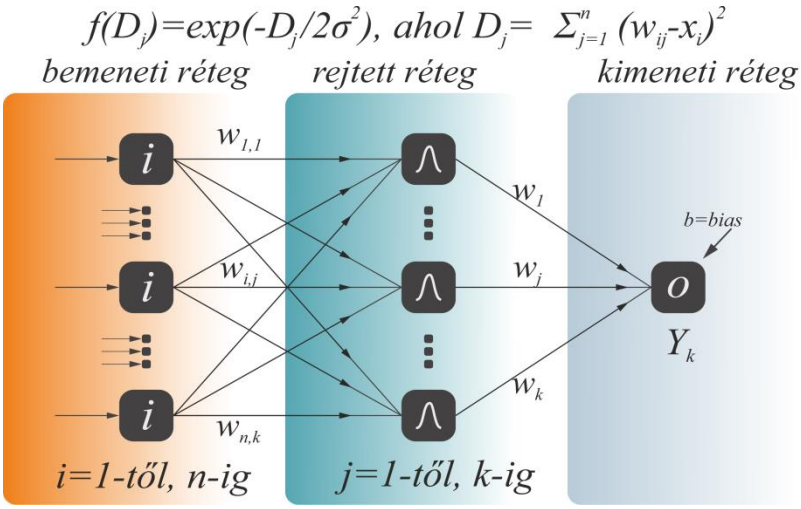
A modell tréningező algoritmusa (a hiba-visszaterjesztéses iterációs algoritmus) nagyon érzékeny a súlyok kezdeti értékeinek megválasztására (Kim és Kim, 2008). MATLAB környezetben a súlyok kezdeti értékeinek megválasztása alapértelmezésben Nguyen-Widrow algoritmussal történik (Nguyen és Widrow, 1990; Pavelka és Procházka, 2004), ami alapértelmezésben [-1;1] között vesz föl kezdeti értékeket, és mivel ezek a kezdeti súlytényezők minden futásnál más-más értéket vesznek föl, így a modell kimenete ugyanolyan beállításoknál is más-más érték lesz. Egyik lehetséges megoldás ezen probléma kiküszöbölésére a súlyok kezdeti értékének futtatás előtt való rögzítése, fixálása, amit Khalil és szerzőtársai (2012) és Palani és szerzőtársai (2008) alkalmaztak is, vagy ha nem is lesz fix a kezdeti érték, le lehet szűkíteni az alapértelmezett intervallumot (Bayram és Kankal, 2015). A másik út, amit a dolgozatomban alkalmaztam, hogy többször futtatom a hálókat ugyanazon beállításokkal. A dolgozatomban az MLPNN-vel való futtatásokat a dunai adatokra ugyanolyan beállítások mellett többször végeztem el, minden esetben hatvanszor, hogy ki tudjam küszöbölni a súlyok véletlenszerűen módszerrel meghatározott inicializálásából fakadó különbségeket (4.1.3 pont). Ezt a megoldást alkalmazták az orvostudomány területén is, amikor is csípőprotézis-beültetésen átesett emberek szűkös adathalmazát vizsgálták, és több ezerszer futtatták az MLPNN-t, hogy pontosabb eredményt kapjanak (Shaikhina és Khovanova, 2017).

A hiba visszaterjesztéses algoritmust többféle eljárás is megvalósíthatja, én ezek közül a Levenberg-Marquardt eljárást alkalmaztam (Marquardt, 1963). Egy rejtett réteget választottam és a rejtett réteg aktivációs függvénye a tangens hiperbolikus függvény volt, a

kimeneté viszont a lineáris függvény. A rejtett rétegben lévő neuronok számát minden egyes MLPNN modellel számított eredménynél megadom.

3.4.3. Radiális bázisfüggvényes neurális hálózatok

A radiális bázisfüggvényes neurális hálózatokat először Broomhead és Lowe (1988) illetve Poggio és Girosi (1990) vezették be a neurális hálózatok irodalmába. Az RBFNN modell egy ellenőrzött tanítást megvalósító előrecsatolt neurális hálózat, amely három réteget tartalmaz: egy bemeneti, egy rejtett és egy kimenő réteget (3.11. ábra). Az RBFNN hálózatnak van egy nagyon nagy előnye az MLPNN-hez képest, mégpedig az, hogy nem ragad bele egy lokális minimum környezetébe, megpróbálja a globális minimumot elérni (Haykin, 1999).



3.11. ábra Az RBFNN sematikus ábrája (Csábrági et al., 2017a)

Az RBFNN modell tréningezése két különálló, egymástól független szakaszra bontható. Az első, a k-közép módszer, ami a klaszterezés standard eljárása, melyet a bemenő adatokra alkalmazunk azért, hogy meghatározzuk a rejtett réteg radiális bázis függvényeinek középpontjait (Kim és Kim 2008), vagyis a modell rejtett rétege önszerveződő réteg. Az RBFNN rejtetett rétegében mindig radiális bázis függvények lesznek az aktivációs függvények. A radiális bázisfüggvények (RBF) $\{\varphi(\|x - x_i\|): i = 1,2, \dots, N\}$ alakú függvények, ahol $\|.\|$ az euklideszi normát jelöli. Ez az elnevezés arra utal, hogy az alapfüggvény $(\varphi(\|x - x_i\|))$ az x_i középpontból kiindulva sugár irányban változik (Fazekas, 2013). A rejtett rétegek súlyai lesznek a radiális bázis függvények középpontjai (w_{ij}), illetve az euklideszi normát kifejtve az alábbi jelölést vezetem be (Antanasijević et al., 2014 nyomán):

$$D_j = \sqrt{\sum_{i=1}^{N} \left(w_{ij} - x_i\right)^2}, \tag{3.7}$$

ahol az N a bemeneti réteg neuronszáma, x_i pedig a bemenet vektor i-edik koordinátája.

Az RBF függvények közül leggyakrabban a Gauss-féle bázisfüggvény használatos, ami az alábbi alakba írható:

$$f(D_j) = exp\left(-\frac{D}{2\sigma^2}\right), \tag{3.8}$$

ahol σ egy pozitív, rögzített paraméter, a továbbiakban szigma-faktornak nevezem.

A második szakasz már ellenőrzött tanítást valósít meg, hiszen a kimeneti réteg és a rejtett réteg közötti súlyokat illetve a torzítás értékét úgy határozza meg a tanító eljárás, hogy minél kisebb legyen az átlagos négyzetes hiba, ami a mért és a becsült kimenet értékekre vonatkozik (Gourine et al., 2012). A MATLAB környezetben a `newrb` beépített függvény használatakor nemcsak a súlyokat és a torzítás (bias) értékét, hanem a rejtett rétegben lévő neuronok számát is változtatja a modell tanítóeljárása. A tanító algoritmus indulásakor nemcsak szigma-faktort lehet megadni - ha nem adjuk meg, akkor egy alapértelmezett értékről indul, (ez általában egy) - hanem az elérendő hibahatárt is. A modell tanítóeljárása során egyesével növeli a rejtett rétegben lévő neuronok számát addig, amíg vagy eléri a kívánt hibahatárt, vagy pedig a maximális neuronok számát, ami normál esetben a mintaszámmal egyezik meg (Demuth és Beale, 2000). A kimeneti réteg a rejtett rétegben lévő RBF függvények lineáris kombinációját állítja elő, hiszen a kimeneti réteg aktivációs függvénye a lineáris függvény (Emamgholizadeh et al., 2014), azaz

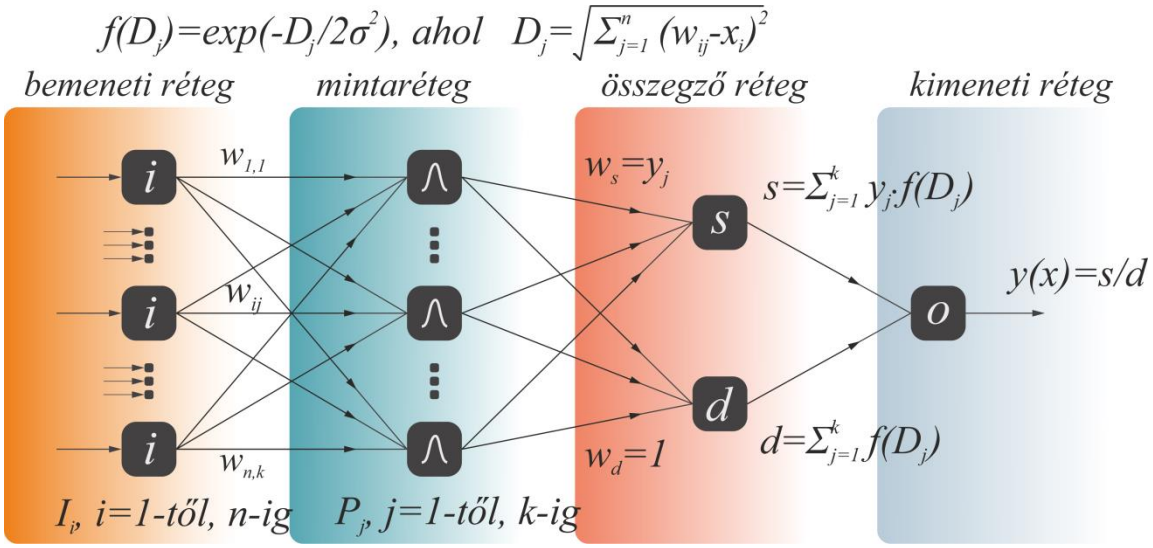
$$Y = \sum_{i=1}^{k} w_i f(D_j) + b, \tag{3.9}$$

ahol *w* a rejtett réteg és a kimeneti réteg súlyvektora, *k* a rejtett rétegben lévő neuronok száma, *b* a torzítás értéke.

A rejtett rétegben lévő neuron számát és a hozzátartozó szigma-faktort minden RBFNN modellel kapott számítási értéknél megadom.

3.4.4. Általános regressziós neurális hálózatok

Az általános regressziós neurális hálózatot először Specht vezette be 1991-ben az MLPNN modell egy lehetséges alternatívájaként (Specht, 1991). A GRNN modell tulajdonképpen egy módosított formája az RBFNN-nek (Kim és Kim, 2008). A GRNN modell egy ellenőrzött tanítást végrehajtó „egyutas" neurális hálózat, amely azt jelenti, hogy a modell tréningezéséhez nincs szükség iterációs eljárásra, hiszen a bemeneti adatok csak egyszer áramolnak végig a rendszeren és ezután már meg is születik a modell kimenete, emiatt a modell tréningezési ideje meglehetősen rövid (Wasserman, 1993). A GRNN modell egy négyrétegű, előrecsatolt neurális hálózat, egy bemeneti réteggel, egy mintaréteggel, egy összegző réteggel és egy kimeneti réteggel rendelkezik (3.12. ábra). A bemeneti rétegben lévő egységek száma megegyezik a független paraméterek számával, a mintarétegben lévő neuronok száma pedig azonos a tréningező mintaszámmal. A bemeneti és a mintaréteg közötti tanítás megegyezik az RBFNN modell bemeneti és rejtett réteg közötti tanítással, hiszen a GRNN modellnél is a Gauss-féle radiális bázisfüggvény az aktivációs függvénye a mintarétegnek. Eltérően a többi neurális hálózattól, a GRNN modellezésénél a tanító paramétereket nem kell inicializálni csak a szigma-faktort vagy más néven az összes paraméter sávszélességét kell kiszámítani (Hanna et al., 2007). Tehát a GRNN modell tanításakor meg kell adni a szigma-faktor értékét, hiszen ez az egyetlen „ismeretlen" paramétere a modellnek, melyet vagy manuálisan, vagy pedig egy iterációs eljárással érdemes meghatározni. Általánosságban a szigma-faktor optimális értéke közel van nullához (Antanasijević et al., 2014).



3.12. ábra A GRNN modell sematikus ábrája (Csábrági et al., 2017a)

Az összegző rétegben lévő neuronok száma eggyel több a kimeneti neuronok számához képest (Antanasijević et al., 2013). Mivel értekezésemben az oldottoxigén-koncentrációját becsülöm, tehát egy kimenettel dolgozom, ezért az összegző rétegben két neuron van: az egyik az S-összegző neuron, a másik D-összegző neuron. Az S-összegző neuron és a mintaréteg neuronjai közötti súlyok megfelelnek a mintahalmazhoz tartozó mért értékekkel (y_j), a D-összegző neuron és a mintaréteg neuronjai közötti súlyok értéke viszont egy. Vagyis az S-összegző neuron a mintaréteg kimenetének súlyozott összegét, a D-összegző neuron viszont a mintarétegek kimenetének „súlyozatlan” összegét állítja elő:

$$S = \sum_{j=1}^{K} y_j \, f\left(D_j\right), \tag{3.10}$$

$$D = \sum_{j=1}^{K} f\left(D_j\right), \tag{3.11}$$

ahol K a mintaréteg neuronjainak száma.

Végül a kimeneti réteg úgy állítja elő a modell becsült kimeneti értékét, hogy az S-összegző neuron kimenetét osztja a D-összegző neuron kimenetével (Hannan et al., 2010). A GRNN modellel számított eredményeknél a szigma-faktort minden esetben közlöm. Mindhárom neurális hálózatot MATLAB környezetében futtattam.

3.4.5. Többváltozós lineáris regresszió

A legegyszerűbb, leggyakrabban használt lineáris módszer a többváltozós lineáris regresszió, amely feltételezi, hogy a függő változó és a független változók közötti kapcsolat lineáris, és a független változók egymástól nem függnek, a reziduumok (hibatagok) normális eloszlást mutatnak, és nem korrelálnak sem egymással, sem a független változókkal. A becslési eljárás alapja a legkisebb négyzetek módszere (Draper és Smith, 1981; Reddy, 2011), a függő paramétert a független paraméterek lineáris függvényeként fejezzük ki:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i + \varepsilon, \tag{3.12}$$

ahol a y a függő változó, a x_i a független változó, a β_0 a regressziós konstans, a β_i-k pedig a független változók együtthatói és ε a regressziós egyenes hibatagja.

A modellbe bevont független változók számának növelésével valószínűleg a becslés jósága nem romlik, így azt hihetnénk, hogy a legjobb modellt minél több független változó bevonásával kapjuk meg. Ezzel szemben az optimális modell létrehozásához meg kell határoznunk azon független változók minimális körét, melyek még érdemi hatással vannak a függő változóra, vagyis szignifikánsak, erre ad segítséget a modell hipotézisvizsgálata.

A többváltozós statisztikai modell esetében a hipotézisvizsgálat két kérdésre keresi a választ:

1. A változók együttesen kielégítő módon magyarázzák-e a függő változót? A nullhipotézis (H0) ebben az esetben az, hogy a független változók együtthatói $(\beta_i\text{-}k)$ mind nullák, vagyis hogy a modell egészében rossz. Az ellenhipotézis (H1) pedig az, hogy a független változók együtthatói között létezik legalább egy, ami nem nulla, tehát a modellt egészében el kell fogadni az adott szignifikancia-szinten. A vizsgálatot F-próbával szokás elvégezni, és az adott F-szignifikancia érték alapján eldönthető, hogy a regressziós modell érvényes-e.
2. A független változók jó magyarázó változók-e a regressziós modellben? A nullhipotézis ebben az esetben az, hogy az adott független változó együtthatója (β_i) nulla, azaz az adott változó tetszőleges változása nem befolyásolja a függő változót. Az ellenhipotézis ellenben az, hogy az adott változó együtthatója (β_i) nullától különböző. A hipotézis tesztelésére t-próbát szokás alkalmazni minden egyes paraméterre külön-külön. A t-értékekhez tartozó p-érték (empirikus szignifikancia-szint) esetében ez azt jelenti, hogy ha a nullához közeli, 0,05 alatti p-értékek esetén - 5 %-os szignifikancia-szint esetében - a nullhipotézist elutasítjuk, az ennél nagyobb értékeknél pedig a nullhipotézist elfogadjuk.

Először a tanítóhalmazra alkalmaztam a modellt, megnéztem, hogy a modell szignifikáns-e (F-szignifikanciája), majd az egyes független paramétereknek megvizsgáltam a p-értékét, és döntöttem arról, hogy az adott paraméter szignifikáns-e a modellben. Későbbiekben csak a szignifikáns változók kerültek felhasználásra, ezekkel adtam becslést a teszthalmazra, és számoltam ki a becslés jóságát jellemző statisztikai mutatókat (4.1.1 pont). A megmaradó paraméterek listáját minden MLR eredménynél közlöm.

3.4.6. Kombinált klaszter- és diszkriminancia-analízis

„A csoportosítás (klasszifikáció) egy általánosan használt módszer a modern kutatásban. Gyakran felmerülő kérdés, hogyan lehet meghatározni a lehető legnagyobb még homogén csoportokat. Ennek szükségessége felmerülhet a mintavételi pontok számának csökkentésében, például azért, hogy azoknak a számát információveszteség nélkül csökkentsük, vagy kisebb alrendszereket vizsgálhassunk.

A csoportosítás többféleképpen végezhető el. Egyik lehetőség manuálisan, ami alatt értjük a szakmai kutatói tapasztalatot illetve ismeretet, a másik bizonyos módszerek felhasználásával, amelyek lényegében a klaszteranalízis különböző típusait jelentik. A klaszterezés során mindig nehéz szembesülni a döntéssel, hogy a csoportokat össze kell-e vonni, vagy éppen szét kell-e osztani. Ugyanakkor ez minden csoportosítás kulcsfontosságú eleme (Anderberg, 1973). Különösen fontos ez a tény, a leggyakrabban használt hierarchikus klaszteranalízis esetében (HCA; Day és Edelsbrunner, 1984).

Bármely kapott csoportosítás validálásra szorul, hiszen a kapott csoportok létezését valamilyen hipotézis vizsgálati eljárással „igazolni” kell. E célra megfelelő módszer a hazai gyakorlatban nagyon ritkán, míg a nemzetközi gyakorlatban is kevéssé gyakran használt

módszer a diszkriminancia-analízis (Borbás et al., 2014), amelynek részletes leírása megtalálható Duda és szerzőtársai (2000) és McLachlan (2004) műveiben.

A lineáris diszkriminancia-analízis (LDA) lényege, hogy az egyes megfigyelések csoportosítása után kapott halmazokat (függő változó), a mért paraméterek (független változók) legjobb lineáris kombinációjával képes megkülönböztetni. Az így kapott halmazok hipersíkokkal lesznek elválasztva. A diszkriminancia-analízis eredményeként a szeparáló hipersíkok által helyesen klasszifikált megfigyelések %-át kapjuk. Amennyiben a csoportok egymást átfedik akkor egyes megfigyeléseket több csoportba osztani nehezebb, mint csak néhány csoportba. Ennek következménye, hogy az LDA általában több megfigyelést csoportosít helyesen, ha a csoportok száma kisebb. Ez a tény a validálási folyamatot még nehezebbé, illetve problematikusabbá teszi homogén csoportok keresésekor.

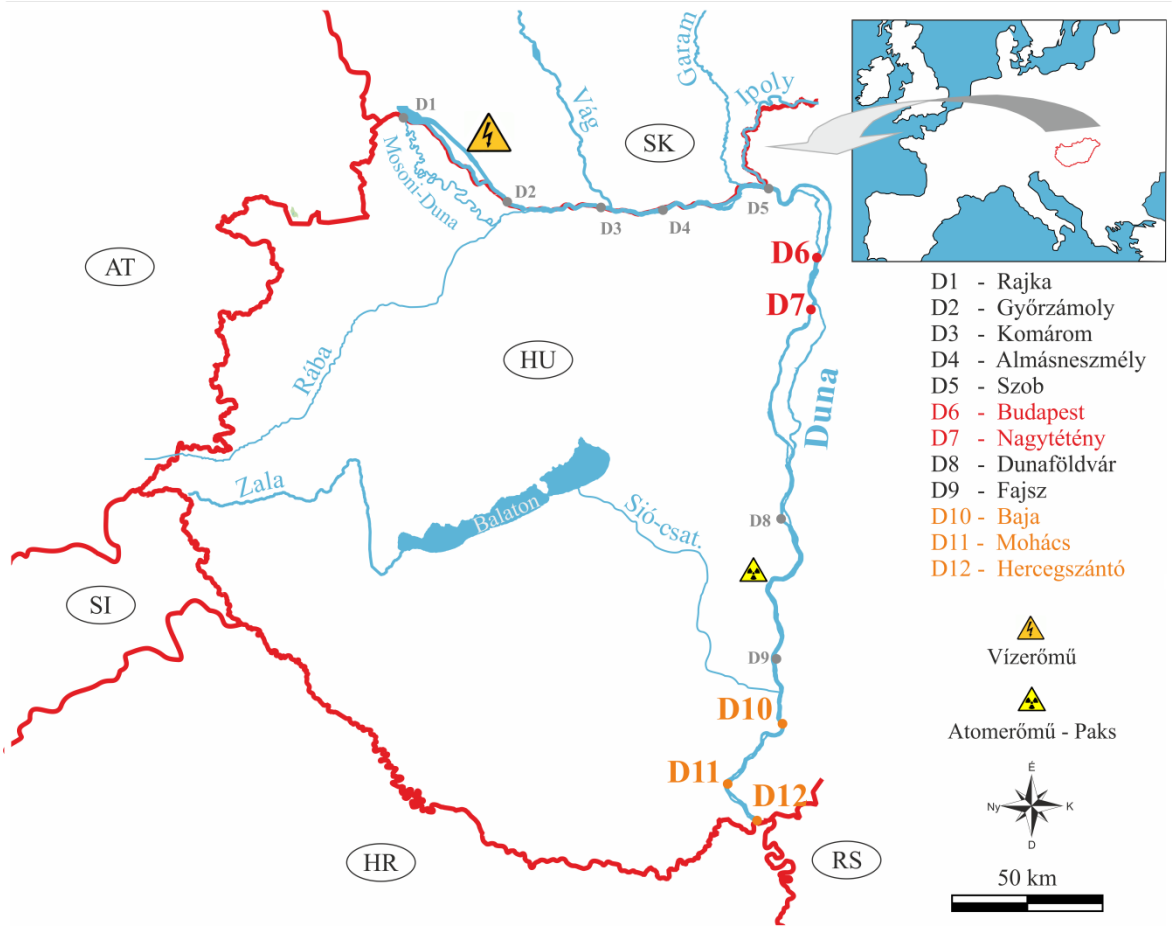
Mintavételi helyek vizsgálatakor milyen előnnyel jár, ha a mintavételi helyek homogén csoportjait meg tudjuk határozni?

- részletesebb és lokalizált információt nyerhetünk egy környezeti rendszerről (pl. egy adott rendszer kisebb léptékű egységeiben történő háttérfolyamatok keresése esetén)
- meghatározhatóak a térben elhelyezkedő homogén csoportok
- egy már meglévő és működő monitoring rendszer újrakalibrálása, ez potenciálisan költségcsökkentést jelent, információ-csökkenés nélkül

Az általános kérdés végeredményben az, hogy mi alapján válasszunk a lehetséges csoportosítások közül úgy, hogy a kapott csoportok homogének legyenek, és mindezt úgy, hogy erről objektíven dönthessünk. Erre ad választ az új módszer, ami biztosítja a homogén csoportosítást, ez pedig a Combined Cluster and Discriminant Analysis (CCDA). A módszer a mintavételi helyek homogén csoportjait találja meg, melynek elemei azonosak és nemcsak hasonlóak. (Ezekről feltételezhető, hogy ugyanazon folyamatot figyelik meg). A módszer szélesebb körben is használható, más megfigyelések ismert származási hellyel szintén vizsgálhatók CCDA segítségével, a módszer részletes leírása a Kovács et al., (2014) cikkben található.” (Kovács, 2015).

CCDA módszer alkalmazása a Dunán

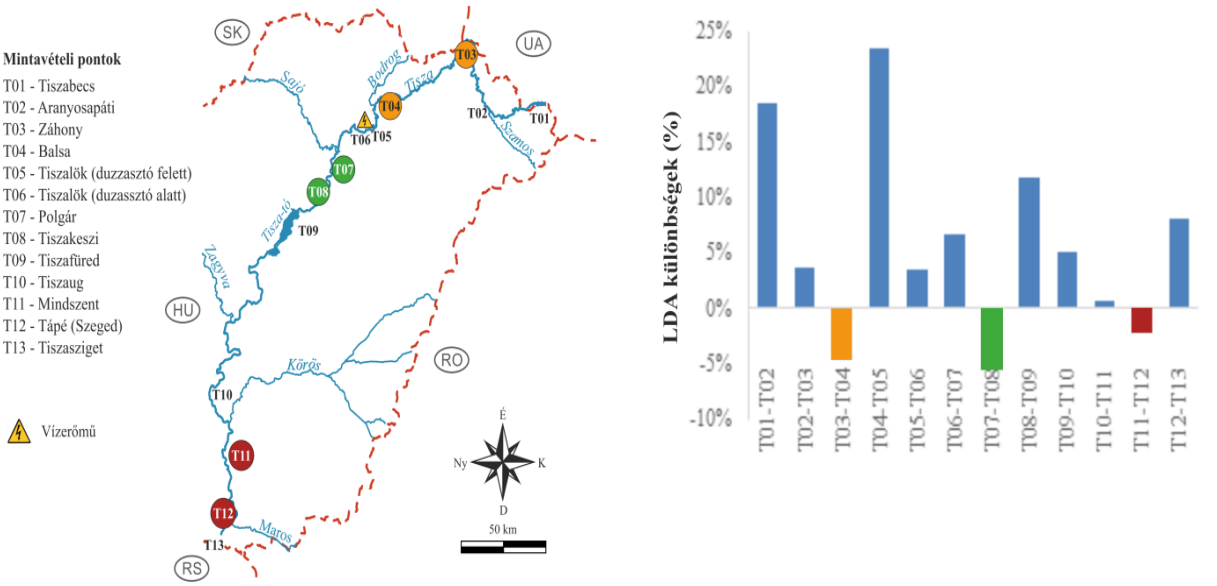
A Dunára vonatkozó CCDA vizsgálat szerint az adott 12 mintavételi pont 9 homogén csoportba bontható (Kovács et al., 2015a), melyek közül 7 darab egytagú csoport (D1, D2, D3, D4, D5, D8, D9), egy darab kéttagú (D6 és D7) és egy darab háromtagú (D10, D11 és D12) (3.13. ábra).



3.13. ábra CCDA által meghatározott homogén csoportok a Dunán (Kovács et al., 2015a)

CCDA módszer alkalmazása a Tiszán

A tiszai mintavételi pontok bemenő adatainak térbeliségét fölhasználva a bemenő adatokat csoportokba rendeztem, hogy ezáltal a modellek hatékonyságát növeljem, így térbeli optimalizációt végeztem. A kombinált klaszter- és diszkriminancia-analízis (CCDA) módszerrel való csoportokba rendezés eredményét használtam föl (Tanos et al., 2015) az 1998 és 2003 között vizsgált időszakban, mely rámutatott, hogy a Tisza vizsgált 13 mintavételi pontja 10 homogén csoportba sorolható. Ezek között három olyan csoport volt megfigyelhető, amelyek két-két mintavételi pontot tartalmaztak (3.14. ábra, és 4.12. ábra beépített térképei), a többi csoport egyelemű. Vagyis a folyó fölső szakaszán a T03-T04 alkot homogén csoportot, a középső szakaszon a T07-T08, az alsó szakaszon pedig a T11-T12 állomások tartoznak homogén csoportba. A mintavételi pontok közötti strukturális különbségek meghatározására a CCDA módszert alkalmazva a mintavételi pontok közötti különbségek is ki lettek számítva (3.2. táblázat). Az azonos színnel jelölt állomások, és a negatív különbségértékek jelölik a homogén csoportba való tartozást. E páronkénti vizsgálatok megerősítették azt az eredményt, miszerint T03-T04, T07-T08 és T11-T12 mintavételi pontok homogén csoportot alkotnak (3.14. ábra). Rámutattak arra is, hogy a két végpont (T01 és T13) között a legnagyobb a különbség (40%, 3.2. táblázat).



a) homogén csoportok jelölése b) szomszédos állomások közti különbségek

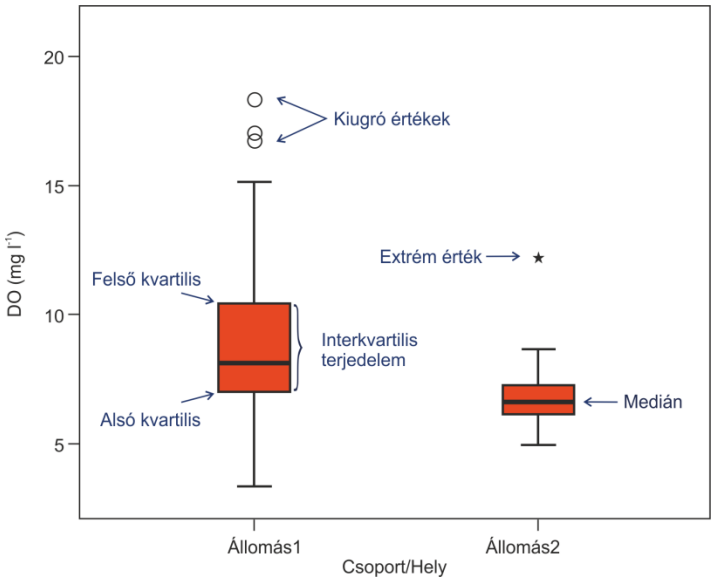
3.14. ábra CCDA módszer eredményei a Tiszán (Tanos et al., 2015)

3.2. táblázat CCDA-val kiszámított különbségek a tiszai mintavételi pontok között

CCDA	T01	T02	T03	T04	T05	T06	T07	T08	T09	T10	T11	T12	T13
T01													
T02	0,19												
T03	0,27	0,04											
T04	0,30	0,06	**-0,05**										
T05	0,33	0,25	0,24	0,23									
T06	0,31	0,23	0,21	0,25	0,04								
T07	0,34	0,23	0,23	0,23	0,10	0,07							
T08	0,33	0,23	0,22	0,21	0,10	0,05	**-0,06**						
T09	0,35	0,29	0,27	0,25	0,22	0,11	0,11	0,12					
T10	0,38	0,33	0,28	0,28	0,27	0,17	0,17	0,19	0,05				
T11	0,40	0,34	0,31	0,33	0,31	0,20	0,18	0,19	0,03	0,01			
T12	0,37	0,30	0,25	0,25	0,20	0,10	0,12	0,09	0,06	0,05	**-0,02**		
T13	0,40	0,33	0,31	0,29	0,30	0,24	0,18	0,19	0,16	0,13	0,10	0,08	

3.5. Box-and-whiskers plot diagramok

A leíró statisztikákat grafikusan is megjelenítik, az úgynevezett box–and–whiskers plot–okban (3.15. ábra), ahol a boxok (dobozok) felső határa a felső, alsó határa pedig az alsó kvartilist jelöli, ezek különbsége pedig az interkvartilis terjedelem. A fekete vízszintes vonal a dobozon belül a medián. A doboz tetejéből és aljából kiálló függőleges vonal (whisker) végpontjai a 1,5–szeres interkvartilis terjedelmet adják meg. Ha adott mért érték a 1,5–3–szoros interkvartilis terjedelmen belül van, akkor kiugró (jele: °), ha a 3–szoros interkvartilis terjedelmen is kívül esik, akkor extrém értéknek tekintendő (jele: *).



3.15. ábra Box-and-whiskers plot diagramok

3.6. A kiértékelés módszere

Az alkalmazott modellek teljesítményének értékeléséhez négy statisztikai mutatót használtam föl, melyek a következőek: átlagos négyzetes hiba négyzetgyöke (RMSE), átlagos abszolút hiba (MAE), determinációs együttható (R^2) és a Willmott-féle egyezési index (IA). Az egyes statisztikai mutatók képletei a következőek:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(v_i - y_i)^2}, \tag{3.13}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|v_i - y_i|, \tag{3.14}$$

$$R^2 = \frac{\left[\sum_{i=1}^{n}(v_i-\bar{v})(y_i-\bar{y})\right]^2}{\sum_{i=1}^{n}(v_i-\bar{v})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}, \tag{3.15}$$

$$IA = 1 - \frac{\sum_{i=1}^{n}(v_i-y_i)^2}{\sum_{i=1}^{n}[|y_i-\bar{v}|+|v_i-\bar{v}|]^2}, \tag{3.16}$$

ahol n a bemenő minták elemszáma; és v_i és y_i a i-edik mért és a becsült kimeneti adatok, a $\bar{v}$ és a $\bar{y}$ ezen mért adatok átlagát fejezi ki.

A determinációs együttható (R^2) a Pearson-féle korrelációs együttható négyzete, értéke nulla és egy közé esik. Az illeszkedés annál jobb, ha értéke minél közelebb van 1-hez. Az R^2 két változó (becsült és mért adatok) közötti kollinearitást vizsgálja, könnyen belátható, hogy ha a $v_i=Ay_i+B$ (A nem nulla és B bármilyen érték, és minden i esetén), akkor $R^2=1$. Ha az A értéke nullától eltérő, egytől (ideális érték) különböző bármilyen nagyértékű rögzített konstans minden i esetén, vagy ha B is egy, nullától (ideális érték) különböző, nagyon nagy szám,

akkor is magas R^2-et kapunk. Tehát a determinációs együttható nem érzékeny az A és a B értékeinek változására. Az IA statisztikai mutató bevezetése (Willmott, 1981) fejlődés volt a determinációs együtthatóval szemben, hiszen ezt a hátrányt sikerült kiküszöbölni az IA mutatónál. Egy másik nagy hátránya még a determinációs együtthatónak, hogy túlságosan érzékeny a kiugró, extrém adatokra, vagyis ha bekerül egy-két ilyen mért érték az adathalmazba, akkor indokolatlanul magas R^2-et kaphatunk a képletből eredő különbségnégyzetek miatt, holott a modell nem biztos, hogy elég hatékony. Sajnos ez a hátrány az IA mutatónál is tapasztalható (Legates és McCabe, 1999).

A modellek teljesítményének pontos vizsgálatához szükség van legalább egy, dimenzió nélküli, az illeszkedés jóságárára vonatkozó mutató vagy egy relatívhibát adó statisztikai mutató (pld. IA és R^2) mellett legalább egy, dimenzióval rendelkező átlagos eltérést kimutató statisztikai mérőszám (pld. RMSE, MAE) használatára is (Legates és McCabe, 1999). A legtöbb tudományos munkában az abszolút hibamutató (pld. RMSE, MAE) értékét mindig meg szokás adni és ez a mérvadó, vagyis ezen statisztikai mutató segítségével lehet a legjobban a becslési hibát szemléltetni, ez hordozza legtöbb információt a becslésről, másik előnye, hogy a mértékegysége megegyezik a kimenet dimenziójával. A különböző modellek összehasonlításakor az RMSE mutatót vettem elsődlegesen figyelembe a szakirodalmi példák alapján.

3.7. A vizsgálatok logikai lépései

3.7.1. Időbeli előrejelzés oldottoxigén-koncentrációra a Dunán

A dunai időbeli előrejelzéseket négy kombinációban végeztem el, először egyedül vizsgáltam Mohács, Fajsz és Győrzámoly adatait, mely kombinációkat a könnyebb értelmezhetőség miatt C_A, C_B, C_C jelöltem. Végül a negyedik kombinációban együtt vizsgáltam a három mintavételi pont összes adatát, ez pedig a C_D kombináció lett.

A teljes adathalmazt kétfelé bontottam: a 2003-as év adatai kerültek a teszthalmazba (26 db minta minden egyes mintavételi helyen), és a 1998-2002 év közötti adatok a tanítóhalmaz elemei lettek (128 db minta a D11 esetén, 125 db minta a D9 állomásnál és 130 db minta a D2 mintavételi pontnál). Tehát a dunai vizsgálatoknál időbeli előrejelzést valósítottam meg. Ennek pontosságát úgy ellenőriztem, hogy rendelkezésre álltak a 2003-as év pontos, mért adatai is, így pedig tudtam hibát számolni.

Mindegyik kombinációban négyféle modellel dolgoztam, a többváltozós regressziós modellt (MLR) alkalmaztam, és háromfajta neurális hálózatot: Többrétegű Perceptront (MLP), a Radiális bázis függvényes neurális hálózatot (RBFNN), illetve az Általánosított regressziós neurális hálózatot (GRNN). Ugyanannál a kombinációnál mindegyik modellnél ugyanazt a tanító és teszthalmazt alkalmaztam.

3.7.2. Térbeli előrejelzés és optimalizáció oldottoxigén-koncentrációra a Tiszán

Az alkalmazott adathalmaz, a tanítóhalmaz és a teszthalmaz kiválasztásának módszere alapján három konfigurációt alkalmaztam a Tiszai adatokon. Minden konfiguráció során arra törekedtem, hogy a tanító és a teszthalmaz közelítse a 2/3-1/3 arányt az összehasonlítás végett.

Az első konfigurációban (1. Tisza-konfiguráció, jelölése: TC1, 3.3. táblázat, 4.12a ábra beépített térképe) a Tisza folyó teljes adathalmazát véletlenszerűen választottam szét tanító és teszthalmazra, ebben az esetben a folyó teljes magyarországi szakaszára (594,5 fkm)

modelleztem az DO paramétert. A kiválasztást háromféleképpen végeztem el, a véletlenszám-generátor különböző inicializáló értékekeivel. Az első esetben (TC1-A) a 800-as alapértékkel inicializáltam, a második esetben (TC1-B) a 2000-es érték volt a véletlenszám-generátor kezdőértéke, az utolsó esetben (TC1-C) pedig 1000 volt ez az alapérték. Ezen kombinációk közül a legjobban teljesítő modell lesz tulajdonképpen a referenciamodellem, amihez a másik két konfiguráció eredményeit hasonlítom.

3.3. táblázat A tiszai konfigurációk, beállítások és albeállítások leírása

		TC1	**TC2**	**TC3**
Konfiguráció	Adathalmaz	teljes	teljes	3-3 mintavételi pont
	Az adathalmaz, a tanító- és teszthalmaz kiválasztásának módszere	véletlen	Négy szomszédos állomás a teszthalmaz elemei	Adathalmaz: két homogén állomás és folyásirányban utánuk jövő szomszédos állomás adatai
Beállítás	A módszer alkalmazása	3 kezdőérték	Négy szomszédos állomás kiválasztása	Három állomás kiválasztása
Albeállítás		nincs	nincs	A tanító- és teszthalmaz definiálása
	Esetek száma	3	4	9
	Jelölések	TC1-A, TC1-B, TC1-C	TC2-A, TC2-B, TC2-C, TC2-D	TC3-A#1, TC3-A#2, TC3-A#3, TC3-B#1, TC3-B#2, TC3-B#3, TC3-C#1, TC3-C#2, TC3-C#3

A második konfigurációban (2. Tisza-konfiguráció, jelölése: TC2, 3.3. táblázat, 4.12b ábra beépített térképe) a 13 mintavételi pontból 4 szomszédos állomás adatai kerülnek a teszthalmazba, a maradék 9 mintavételi pont adatai alkotják a tanítóhalmazt. A 13 állomást azért osztottam szét 9-4 arányban a tanító illetve a teszthalmazba, hogy tartani tudjam a kb. 2/3-1/3-os fölbontási arányt. Ennél a konfigurációnál már térbeli előrejelzést valósítottam meg a folyó egy-egy szakaszának az oldottoxigén-koncentrációjára vonatkozóan, hiszen irányítottan, mintavételi pontonként választottam ki a tanító és teszthalmazokat.

Végül a harmadik konfigurációban (3. Tisza-konfiguráció, jelölése: TC3, 3.3. táblázat, 4.12c ábra beépített térképe) 3-3 mintavételi pont adatait dolgoztam fel úgy, hogy fölhasználtam az azonosított homogén csoportok mintavételi pontjait (3.14. ábra). Mivel e homogén csoportok nagysága nem tette lehetővé, hogy tisztán homogén szakaszon dolgozzam, ezért a három kiválasztott mintavételi pontok közül kettő egy homogén csoport elemei voltak. A harmadik, folyásirányban szomszédos állomás pedig ezektől különböző tulajdonságú, inhomogén mintavételi pont volt. Ilyen lehetőség a csoportosítás alapján három volt, így harmadik a konfiguráción belül három beállítás volt lehetséges. Tehát a T03-T04-

T05 mintavétel pontok esetén (A beállítás, jelölése TC3-A) a T05 állomás volt inhomogén tulajdonságú a T03 és a T04 állomások által alkotott homogén csoporttal szemben, a T07-T08-T09 állomások esetén (B beállítás, jelölése: TC3-B) viszont a T09 volt különböző struktúrájú a T07 és a T08 mintavételi pontok alkotta homogén csoporthoz képest. Az utolsó esetben, a C beállításnál (jelölése: TC3-C, T11-T12-T13 állomások) a T13 mintavételi pont struktúrája különült el a T11 és a T12 állomások által meghatározott homogén csoporttól. Ezen beállítások (TC3-A, TC3-B és TC3-C beállítások) összes permutációjára megadom az alkalmazott modellek eredményeit, ekkor kapom meg az albeállításokat (pld. TC3-A#1). Ennél a konfigurációnál térbeli optimalizáció (2.4.3 pont) valósult meg, hiszen szűkebb mintahalmazon vizsgálódtam hatékonyabban. Ugyanakkor térbeli előrejelzést is kaptam a folyó egy-egy mintavételi pontjára az oldottoxigén-koncentrációjára vonatkozóan.

A három konfiguráción belül az egyes beállításokat a konfiguráció jele (TC1, TC2 vagy TC3) és egy kötőjel után egy betűvel (A, B, C vagy D), illetve és albeállításokat, amennyiben vannak (pld. TC3 konfiguráció esetén) betű#szám kombinációjával jelöltem (pld. TC3-B#3).

3.7.3. Térbeli előrejelzés és optimalizáció oldottoxigén-koncentrációra a Dunán

Az alkalmazott adathalmaz, a tanítóhalmaz és a teszthalmaz kiválasztásának módszere alapján kettő konfigurációt alkalmaztam a dunai adatokon. Mindkét konfiguráció során arra törekedtem, hogy a tanító és a teszthalmaz közelítse a 2:1 arányt az összehasonlítás végett. Az első konfigurációban (1. Duna-konfiguráció, jelölése: D_R) a folyó teljes adathalmazát véletlenszerűen választottam szét 2:1 arányban tanító és teszthalmazra, ebben az esetben a folyó teljes magyarországi szakaszára (417 km) modelleztem az oldottoxigén-koncentrációját. A kiválasztást a véletlenszám-generátor egy meghatározott inicializáló értékekeivel (2000) végeztem el. Ez a modell (D_R) lesz a referenciamodell és a kapott eredmények pedig a referenciaértékek.

A második konfigurációban a CCDA módszerrel a folyóra kapott csoportosítás eredményeit fölhasználva egy beállítás vizsgálata volt lehetséges (3.13. ábra), hiszen a folyón csak két darab többelemű homogén csoport van, és az egyik, a háromelemű homogén csoport adataival való vizsgálatot nem tudtam elvégezni, mert a folyásirányban szomszédos állomás már a szerb területre esik. Eszerint a D6, D7 és D8 állomások (Budapest, Nagytétény és Dunaföldvár) adataival végeztem el a vizsgálatot háromféle albeállítást alkalmazva. Először a D6 és a D8 állomások adatai (D_A), majd a D7 és a D8 állomások adatai (D_B) alkották a tanítóhalmazt, végül pedig a D6 és D7 mintavételi pontok adatai (D_C) voltak a tanítóhalmaz elemei.

Mind a Tiszára mind a Dunára megvalósított térbeli előrejelzés és optimalizáció és a referenciamodellek esetén mindegyik konfigurációban háromféle modellel dolgoztam, a többváltozós regressziós modellt (MLR) alkalmaztam, és kétfajta neurális hálózatot: a Radiális bázis függvényes neurális hálózatot (RBFNN), illetve az Általánosított regressziós neurális hálózatot (GRNN). Az MLPNN modellt nem alkalmaztam a térbeli előrejelzésre és optimalizációra, mivel a dunai adatokra megvalósított időbeli előrejelzéses vizsgálatoknál azt tapasztaltam, hogy ez a modell kevéssé ad megbízható eredményt, és nagyon időigényes az összes beállítás hatvanszori állandó futtatása (4.1.3 pont).

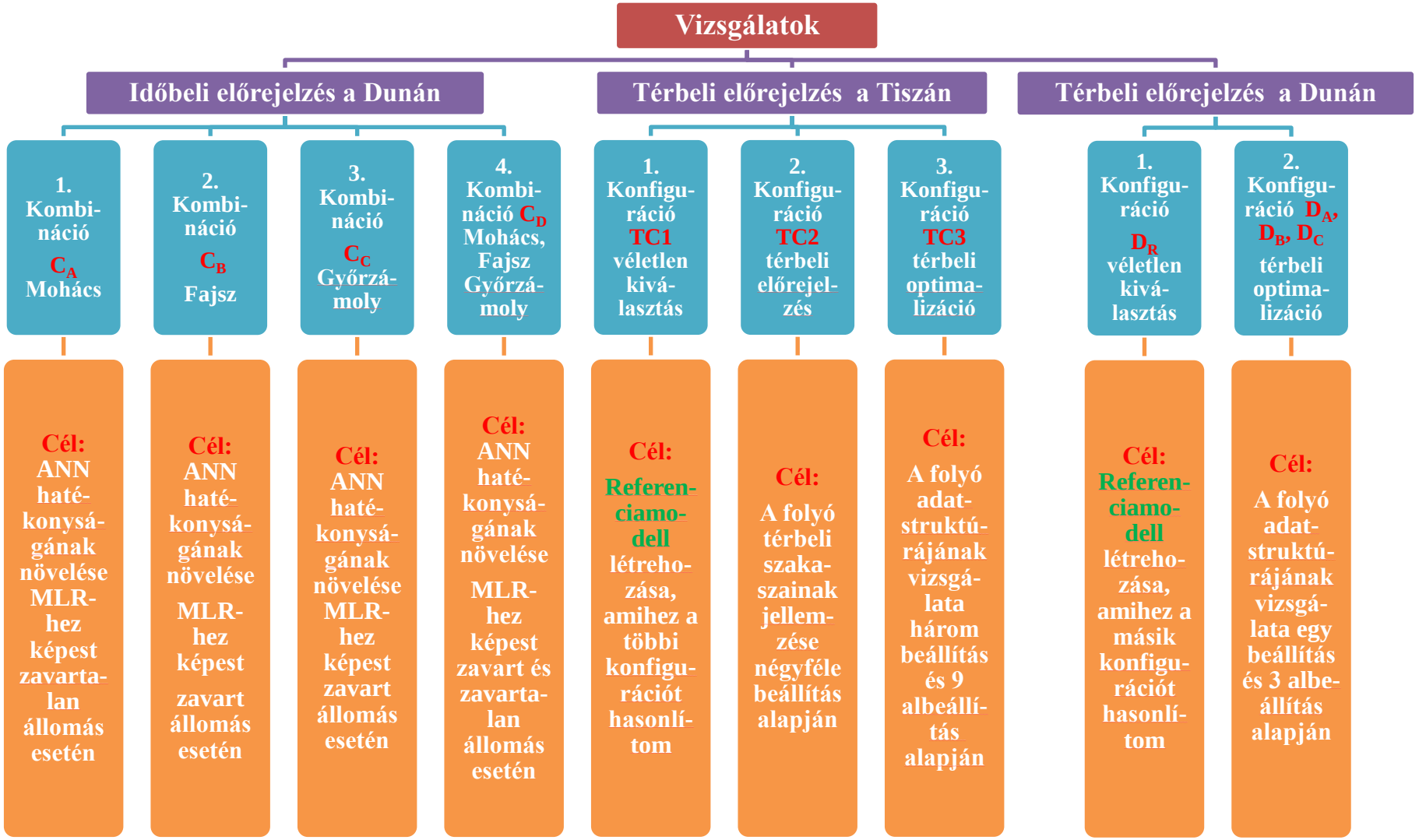
3.8. A vizsgálatok folyamatábrája

Mindkét folyó vizsgálatánál az oldottoxigén-koncentrációját becsültem ugyanazon négy paraméter segítségével. Mindkét folyónál az alkalmazott időintervallum a kapott adatok legfrissebb adatai, a 1998-2003 év közötti időszak volt.

A dunai adatokra megvalósított oldottoxigén-tartalomra vonatkozó időbeli előrejelzésnél a célkitűzésemnek megfelelően azt vizsgáltam, hogy mennyire függ a becslés attól, hogy a mintavételi pontot érik-e antropogén hatások vagy nem. Erre volt lehetőségem, hiszen a Duna magyarországi szakaszán két erőmű is található (3.1. ábra). Ennek a kérdéskörnek a vizsgálatához négy kombinációt (3.7.1 pont) alkalmaztam.

A tiszai vizsgálatokat három konfigurációban (3.7.2 pont) végeztem el, ahol az első konfiguráció tulajdonképpen egy referenciamodell, amihez a többi konfiguráció eredményét hasonlítottam. A Tisza folyó magyarországi szakaszának nagyon heterogén volta miatt (jelentősen eltérőek a bemenő paraméterek ill. az oldott oxigén is, 4.7a ábra, 4.6. táblázat) érdemes volt megvizsgálni a folyó egyes szakaszait, és ezekre jellemzéseket adni, ezért született meg a második konfiguráció. A térbeli jellemzés mellett a becslés hatékonyságán is szerettem volna javítani a tanítóhalmaz adatstruktúrájának vizsgálatával, így jött létre a harmadik konfiguráció, ahol térbeli optimalizációt is alkalmaztam.

A dunai térbeli előrejelzés és optimalizáció esetén egy referenciamodellt illetve egy olyan konfigurációt készítettem, ahol három állomás adatait vizsgáltam, melynél két állomás homogén, a harmadik pedig velük szomszédos inhomogén állomás volt. A cél itt is a tanítóhalmaz leghatékonyabb adatstruktúrájának meghatározása volt. A kutatómunkám vizsgálatainak konfigurációi, beállításai és céljai a következő folyamatábrán (3.16. ábra) figyelhetőek meg:



3.16. ábra A vizsgálatok folyamatábrája

4. EREDMÉNYEK

Ebben a fejezetben a célkitűzéseknek megfelelően a Duna és a Tisza folyó mintavételi pontjain a lineáris és a neurális hálózatokkal kapott becsléseknek és azok kiértékeléseinek, összehasonlításainak bemutatására kerül sor. Előtte ismertetem a becslési eljárásokat megvalósító modellek alkalmazási módjait, illetve szemléltetem a Duna és a Tisza folyó mintavételi pontjain mért mintahalmazok jellemzőit.

4.1. Becslési eljárásokat megvalósító modellek alkalmazásai

4.1.1. MLR alkalmazása

Minden létrehozott többváltozós lineáris modell létezését szükséges F-próbával tesztelni, amit megtettem, így például Mohács adataira, a tanítóhalmazra vonatkoztatva (C_A kombináció) kapott MLR1-el jelölt modell, amit négy független változóval hoztam létre szignifikáns, hiszen az F-teszt eredménye 8,3E-19 volt. Vagyis az így kapott MLR1 modell létezik, azonban a modellben szereplő paraméterek együtthatóit vizsgálva (p-érték) arra jutottam, hogy két paraméter (a vízhozam és a vezetőképesség) 5%-os szinten nem szignifikáns, azaz e paraméterek együtthatója nem létezik, vagyis az MLR1 modell hatékonyságát e paraméterek nem növelik. Tehát a vízhozam és a vezetőképesség paraméterek nem szignifikánsak ebben a modellben, mert p-értékük nagyobb, mint 0,05 (4.1. táblázat). Annak ellenére, hogy a tudományos gyakorlatban általában e korlátokat nem veszik figyelembe (Akkoyunlu et al., 2011; Antanasijević et al., 2013; Ay és Kisi, 2012; He et al., 2011b) szükségesnek tartottam az MLR2 modell felállítását a kérdéses paraméterek nélkül. Tehát egy másik modellt hoztam létre, amit MLR2-vel jelöltem.

4.1. táblázat MLR modellel számított együtthatók és hibák a C_A kombináció esetén

	Első eset - MLR1			Második eset - MLR2		
	Koefficiens	Standard hiba	p-érték	Koefficiens	Standard hiba	p-érték
(Konstans)	-30,25	6,23	3,54E-06	-35,46	4,75	1,2E-11
RF	0	0	0,34	-	-	-
T_w	-0,22	0,04	4,78E-09	-0,18	0,02	1,62E-18
pH	5,66	0,65	1,73E-14	5,91	0,59	1,2E-17
EC	-0,01	0	0,17	-	-	-

A második esetben, a vízhozam és a vezetőképesség nélkül az MLR2 modell is szignifikáns (3,25E-20). A két független paraméter p-értéke 0,05 alatt van, ezért ez a két paraméter szignifikáns, tehát ebben az esetben a modell, és a modell paraméterei is léteznek. A (4.1) egyenlet írja le az oldott oxigén, mint függő paraméter és a hőmérséklet és a pH, mint független paraméterek közötti kapcsolatot:

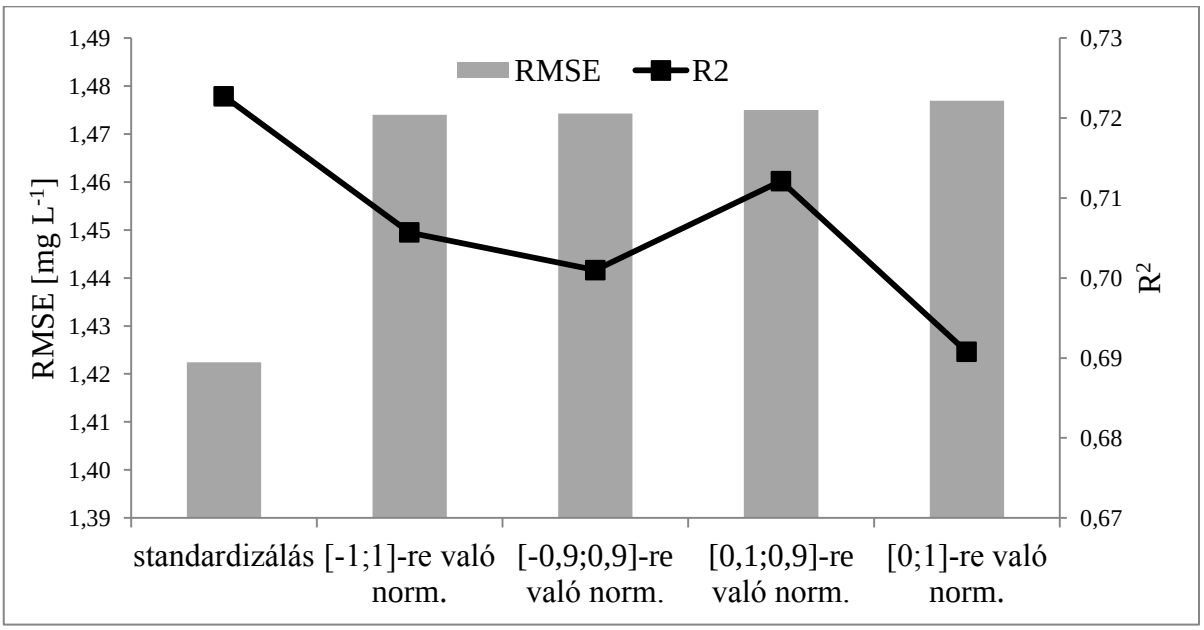
$$DO = -0{,}18 \cdot T_w + 5{,}91 \cdot pH - 35{,}46. \quad (4.1)$$

Ezt az egyenletet alkalmaztam a teszthalmazra, és az így kapott RMSE érték 2,03 mg L^{-1}, a determinációs együttható pedig 0,4 volt (4.4. táblázat, első sor). Ezt a módszert alkalmaztam az összes dunai kombinációra és a tiszai konfigurációkra is.

4.1.2. Előfeldolgozások összehasonlítása

Ahhoz, hogy a neurális hálózatok a tanítási folyamatuk során minden bemenő paramétert ugyanolyan mértékben vegyenek figyelembe a nyers adatokon szinte mindig célszerű valamilyen

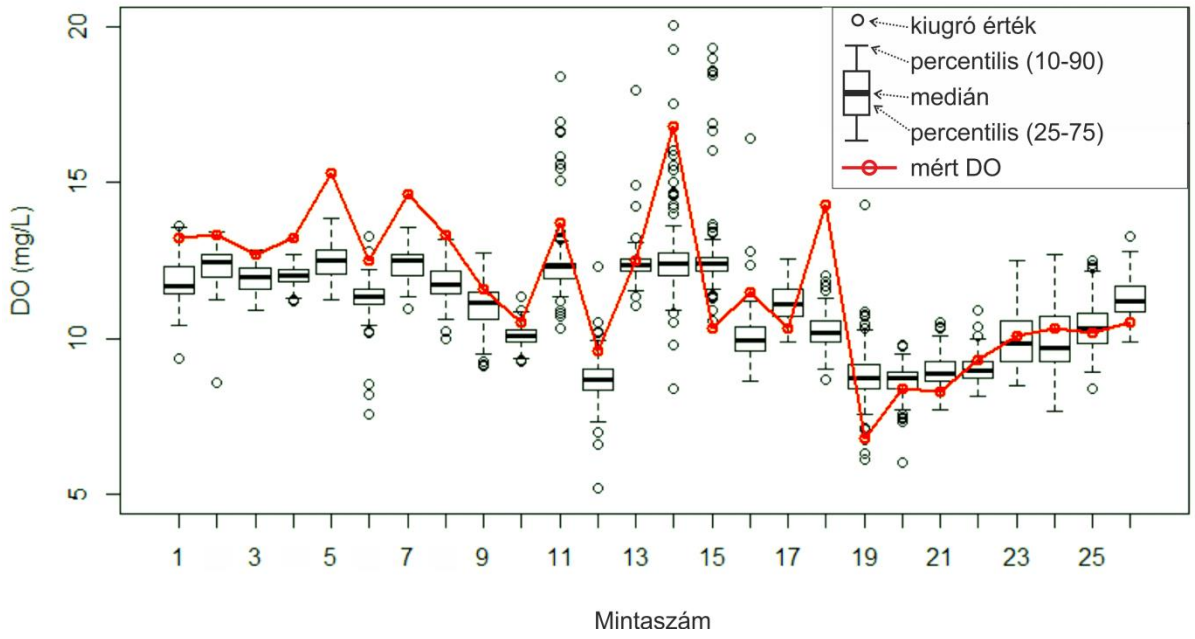
előfeldolgozást alkalmazni (3.3 pont). Mohács mintavételi pont adataival (C_A kombinációban), GRNN modellt használva ötféle előfeldolgozás eredményét ismertetem a teszthalmazra – 2003-as adatok – vonatkoztatva (4.1. ábra, 8.2. táblázat). Ezen eredményekből látszik, hogy - bár csekély különbséggel - a standardizálás volt a leghatékonyabb módszer a C_A kombinációban a különböző intervallumokra való normalizálásokkal szemben. Így minden egyes kombinációnál, konfigurációnál, ahol neurális hálózatokat alkalmaztam, standardizáltam a bemenő adatokat, és a vizsgálat végén visszatranszformáltam a kimenetet az eredeti tartományba (3.3 pont, 3.3. ábra).



4.1. ábra Előfeldolgozások összehasonlítása GRNN modellel C_A kombinációban

4.1.3. MLPNN modellel való becslés új módszere

Mint azt a MLPNN módszer leírásában bemutattam (3.4.2 pont), a modell súlytényezőinek inicializálása véletlenszerűen történik, ezáltal a modellel kapott végeredmény ugyanolyan beállítások mellett más-más érték lesz. Annak érdekében, hogy megragadjam e jelenséget ugyanazon beállításokkal futtattam az MLPNN módszert, így megvizsgálhattam, hogy ezek a véletlen inicializáló értékek milyen mértékben befolyásolják a modellel kapott eredményeket. Első közelítésben 60 db tesztfutást (ahol minden futásnál az epochszám 1000, a rejtett rétegben lévő neuronszám pedig öt volt) végeztem azonos beállításokkal a C_A kombinációra (Mohács adatai) vonatkozóan. Az eredmény rámutatott arra, hogy jelentős eltérések adódhatnak az egyes megfigyelések becslése során, melyek minden bizonnyal az inicializáló értékek különbségeiből fakadtak. Az egyes megfigyelésekre kapott becsült oldott oxigén értékeket box-and-whiskers plot diagram segítségével ábrázoltam (4.2. ábra). A legnagyobb különbségek a 14. minta becslése során voltak tapasztalhatóak, amikor a legalacsonyabb becsült érték 8,39 mg L^{-1} volt, a legnagyobb 20,01 mg L^{-1} (amíg a mért érték 17 mg L^{-1} volt) (4.2. ábra).



4.2. ábra MLPNN-nel való 60 futtatás eredménye a C_A kombinációra (Csábrági et al., 2017a)

A jelentős különbségek rámutattak arra, hogy az MLPNN hálót nem elegendő egyszer futtatni, mert ebben az esetben az eredmény félrevezető lehet. Az MLPNN háló e tulajdonsága kezelhető azzal, ha - a már eleve iterált – futásokat további iterációnak vetjük alá. Ez az ismételt iteratív megközelítés megalapozottnak bizonyult, mert ezzel a megközelítéssel kezelhetőek az MLPNN modell egy-egy futásából adódó kiugró eredmények. Tehát abban az esetben, ha nem alkalmazom az ismételt iterált futtatást véletlenszerűen fogadhatok el akár kiugróan rossz eredményeket is.

Az MLPNN módszer hatvanszor ismételt iterált futtatása (4.2. ábra) rámutatott arra, hogy a véletlenszerű inicializáló eljárás miatt, azonos beállítás mellett is jelentős különbségek lehetnek az eredményekben. A témában született korábbi tudományos művek többsége e jelenségtől eltekintettek (Dogan et al., 2009; Kuo et al., 2007; Ranković et al., 2010; Singh et al., 2009; Talib és Amat, 2012), bár Palani et al (2008) egyféle megoldást adott a problémára. Shaikhina és Khovanova (2017) viszont módszeresen alkalmazták az MLPNN iterált futtatását. Továbbiakban ezt az iterált megközelítést alkalmaztam az MLPNN háló futtatásakor, de ez a módszer újabb kérdéseket vetett föl.

Meg kellett vizsgálni, hogy ezt a plusz iterációs lépést milyen szempontok szerint célszerű elvégezni. Ezt kívántam meghatározni a következő vizsgálat során, amelyben az iterációk számát 20 és 200 között minden lépésben tízzel növeltem ugyanazon beállítások mellett. Ezáltal a teszthalmaz minden egyes megfigyelésére 20, 30, 40,…, 200 becslést hoztam létre, melyeket megfigyelésenként i) átlagoltam, illetve ii) a becslések mediánját használtam a validációhoz. Az így kialakult idősorokat (átlag és medián alapján) hasonlítottam a teszthalmaz mért idősorához és képeztem az RMSE értékeket (4.2. táblázat).

Jogosan felmerülő kérdés, hogy a két eljárás közül (átlag, medián) melyik használata javasolható? A vizsgálat eredményeinek alapján megállapítottam, hogy az átlaggal számított idősorokkal kapott RMSE értékek kissé változékonyabbak a mediánnal kapott idősorok RMSE értékeikhez képest, tehát ahogy az várható volt a medián kevésbé érzékeny az esetleges kiugró

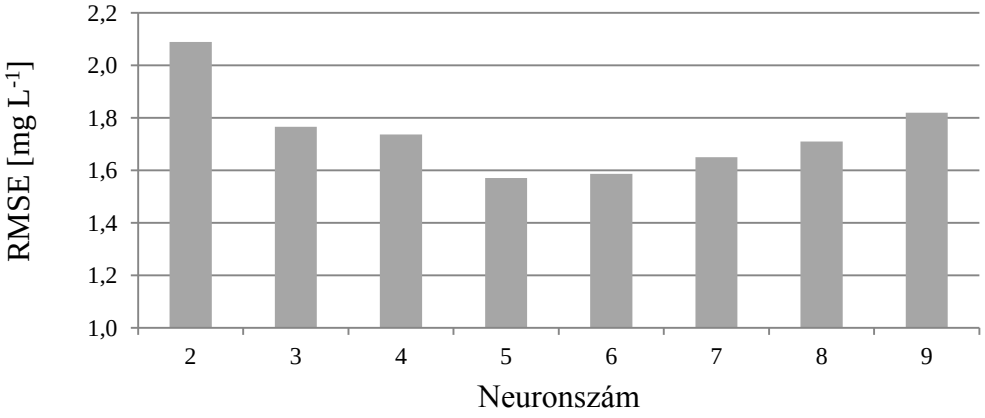
becslésekre. Mindazonáltal a két megközelítés által kapott RMSE értékeket átlagolva a mediános megközelítés nagyjából egytized mg L^{-1}-rel gyengébb eredményt hozott. Végül a doktori munkámban és a további vizsgálatok során a becslések átlagát tekintettem mérvadónak, ugyanis általánosságban az átlagot szokás használni, és jól látható, hogy nagy iterációszámnál az átlag lényegesen jobb eredményt is ad a mediános megközelítésnél.

Doktori munkámban a hatvanas iterációs lépést választottam ki, ugyanis ez az iterációszám biztosította azt, hogy a futási idő ne lépje túl a kezelhetetlen tartományt, valamint azt is, hogy az MLPNN módszer bemutatott pontatlanságait kiküszöbölje. Ennél az iterációszámnál a kapott RMSE érték 1,57 mg L^{-1} volt, amely lényegében az átlaggal kapott RMSE eredményeinek átlagával volt egyenlő (4.2. táblázat).

4.2. táblázat Átlaggal vagy mediánnal kapott idősorok RMSE értékei iterációszám függvényében

Futások száma / RMSE [mg L^{-1}]	Átlag	Medián
20	1,77	1,70
30	1,65	1,70
40	2,20	1,67
50	1,49	1,70
60	1,57	1,67
70	1,63	1,71
80	1,43	1,68
90	1,52	1,69
100	1,70	1,69
110	1,65	1,71
120	1,55	1,68
130	1,42	1,70
140	1,53	1,68
150	1,55	1,71
160	1,56	1,68
170	1,52	1,70
180	1,43	1,69
190	1,40	1,69
200	1,48	1,70
Átlag	1,58	1,69

Ezután az új módszer alkalmazásával teszteltem az MLPNN modell teszthalmazra vonatkozó teljesítményét a rejtett rétegben lévő neuronszámok függvényében a C_A kombináció esetén. Megvizsgáltam az MLPNN modell teljesítményét 2 és 9 neuron között (4.3. ábra), melynek az eredménye az volt, hogy a tesztelő halmazra vonatkozó RMSE monoton csökkent 2 és 5 között, majd monoton nőtt 5 és 9 között. A legkisebb RMSE értéket 5 neuron esetében kaptam a tesztelő halmazra. Jelen paraméterkörre az 5 neuronos beállítás adja legpontosabb MLPNN becslést. Ezért a továbbiakban az MLPNN 5 neuronos, 60-szor iterált eredményét értékeltem ki, és hasonlítottam össze a többi modell teljesítményével (4.4. ábra).

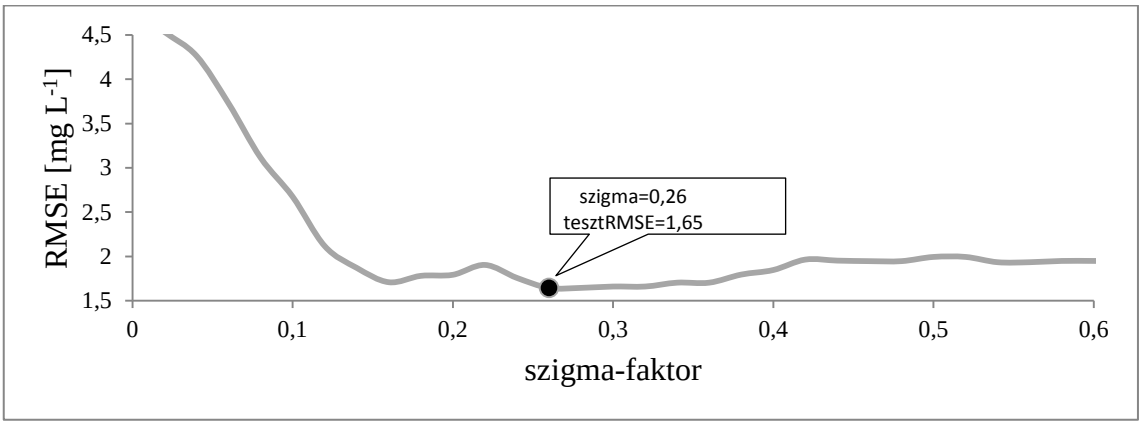


4.3. ábra MLPNN futtatása esetén a teszthalmazra kapott RMSE értékek (Csábrági et al., 2017a)

A rejtett rétegben alkalmazott neuron számot illetően Fletcher és Goss (1993) szerint a neuron számnak $2*I^{1/2}+O$ és $2*I+1$ közöttinek kell lennie, ahol az I a bemenő paraméterek száma, az O pedig a kimenő paraméterek száma. Ez jelen esetben azt jelenti, hogy az optimális neuron számnak 5 és 9 között kell lennie a rejtett rétegben, ami a C_A kombináció esetén teljesült is.

4.1.4. RBFNN és a GRNN modellel való becslés

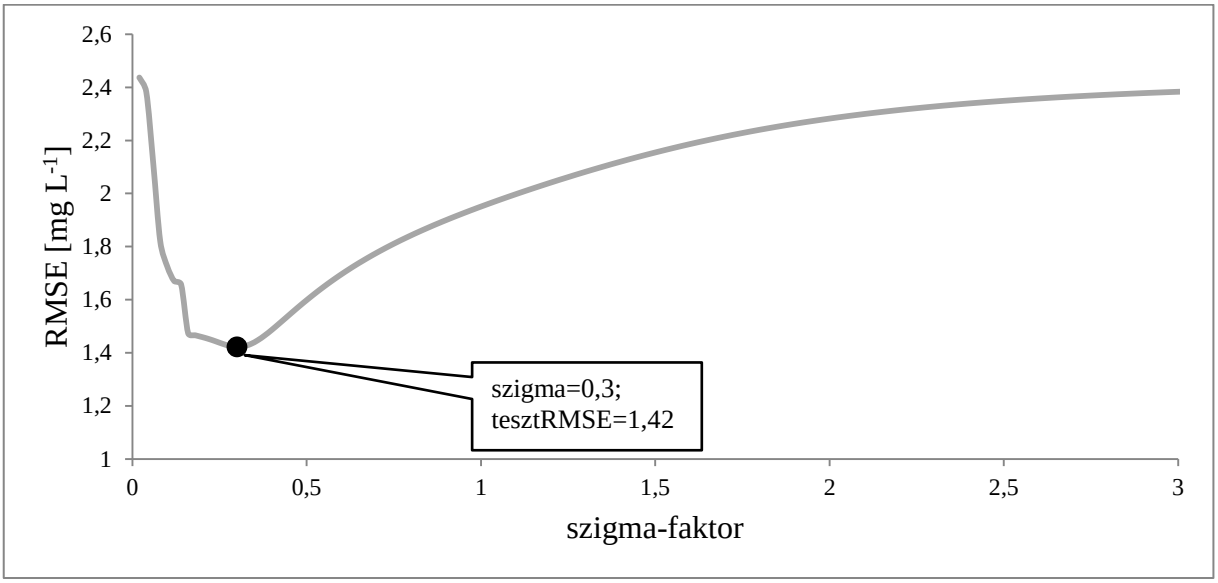
Az RBFNN háló teljesítménye azonos adathalmaz esetében kizárólag a szigma-faktor értékétől függ, hiszen a `newrb` beépített függvény addig növeli a rejtett rétegben lévő neuron számot, amíg az előre megadott MSE értékét megkapja, vagy pedig eléri a maximális neuronszámot, ami alapértelmezésben a bemenet mintaszámával egyezik meg. A tesztelő halmazra (C_A kombináció) kapott RMSE értékét néztem a szigma-faktor függvényében, amelyet 0,02 kezdőértékről indítottam egészen a 3 értékig. Azt tapasztaltam, hogy kezdetben erőteljesen csökkent az RMSE értéke, majd enyhe emelkedés után elérte a minimumot 0,26-nál, itt az RMSE értéke 1,65 mg L^{-1} (4.4. táblázat) lett. Majd ismét kicsit nőni kezdett, és 0,68 szigma-faktor érték után az RMSE konstans értéket (2,01 mg L^{-1}) vett föl (4.4. ábra) egészen a vizsgált intervallum fölső határáig. A rejtett rétegben lévő neuron szám a 0,26-os szigma-faktornál 18 volt.



4.4. ábra RBFNN futtatása esetén a teszthalmazra kapott RMSE értékek (Csábrági et al., 2017a)

A GRNN modell alkalmazásánál kizárólag az általános, minden paraméterre alkalmazható szigma-faktor értékétől függ azonos adathalmaz esetén a modell teljesítménye, így fontos feladat az optimális szigma-faktor meghatározása. A tesztelő halmaz RMSE értékét vizsgáltam a szigma-faktor függvényében 0,02-től 3-as értékig (4.5. ábra). Először erőteljes csökkenés mutatkozik egészen a minimális értékig, ahol a 0,3-as szigma-faktor értéknél a tesztelő halmazra

vonatkozó RMSE a 1,42 mg L^{-1} értéket veszi föl (4.4. táblázat), majd ezután fokozatosan, egyenletesen nő az RMSE értéke.



4.5. ábra GRNN futtatása esetén a teszthalmazra kapott RMSE értékek (Csábrági et al., 2017a)

4.2. Oldottoxigén-koncentrációra vonatkozó időbeli előrejelzés a Dunán

4.2.1. A dunai állomások mintahalmazai

A mindhárom mintavételi pontnál mért adatok leíró statisztikái (4.3. táblázat, ahol az *r* a bemenő paraméterek és a DO közötti korrelációs együttható) illetve box-and-whisker diagramok (8.1. ábra, az elkészített ábra forráskódját is tartalmazza az M3-as melléklet) rámutatnak arra, hogy a relatív szórás alapján a legváltozékonyabb paraméterek a vízhozam és a hőmérséklet, a legstabilabb paraméter viszont a pH. A becsülni kívánt oldottoxigén-tartalom közepes változékonyságot mutatott, hasonlóan a vezetőképességhez. Az is látható, hogy Mohács és Fajsz mintavételi pontok adatai között kis különbség van (Kovács et al., 2015a), míg Győrzámoly állomás adatai nagyon eltérnek az előző állomások adataitól. Mindez nem meglepő, hiszen Győrzámoly és Mohács között több mint 298 fkm van. A különbséget jól mutatja a vízhozam átlagértéke, mely Győrzámolynál majdnem 20%-kal, míg az átlaghőmérséklet közel tíz százalékkal kevesebb, mint Mohácsnál.

4.3. táblázat Mohács, Fajsz és Győrzámoly adatainak alapstatisztikája a teljes vizsgált időintervallumra (1998-2003)

Állomás	Minta-szám	Paraméter	Max	Min	Átlag	Szórás	Relatív szórás	r
D11	154	Q	5400	910	2363,19	979,18	0,41	-0,16
		T_w	25,9	0,2	12,84	7,42	0,58	-0,29
		pH	8,75	7,8	8,24	0,22	0,03	0,33
		EC	530	272	377,62	58,98	0,16	0,25
		DO	17,7	6,8	11,04	1,80	0,16	1
D9	151	Q	5310	920	2346,74	939,77	0,40	-0,08
		T_w	25,2	0,2	12,49	7,50	0,60	-0,34
		pH	8,85	7,75	8,26	0,26	0,03	0,25
		EC	525	256	371,30	57,31	0,15	0,27
		DO	15,5	7	11,20	1,60	0,14	1
D2	156	Q	5130	618	1940,58	770,25	0,40	0,16
		T_w	22,8	0	11,59	6,48	0,56	-0,58
		pH	8,9	7,02	8,13	0,27	0,03	0,08
		EC	560	294	370,56	48,71	0,13	0,42
		DO	14,03	5,76	9,82	1,54	0,16	1

Mohács mért paramétereinek sztochasztikus kapcsolatait megvizsgálva látható, hogy az oldott oxigén negatívan korrelál a vízhozammal és a hőmérséklettel. A legjelentősebb lineáris kapcsolatban az oldott oxigén a pH-val van, a korrelációs együttható (*r*) esetükben 0,33, amely figyelembe véve az adatsorok hosszát (154 adat) szignifikáns (|r|>0,16). Fajsz és Győrzámoly állomásokon mért oldott oxigén és a többi bemenő paraméterek közötti korrelációs együtthatókat tekintve az látszik, hogy mindkét állomás esetében az oldott oxigén a hőmérséklettel van a legjelentősebb lineáris kapcsolatban (-0,34 és -0,58), ugyanakkor Győrzámoly esetében az oldott oxigén és a pH közötti korrelációs együttható igen alacsony (0,08), ami nem szignifikáns, hiszen kisebb, mint 0,16.

4.2.2. Modelleredmények mind a négy kombinációra

Mind a négy kombinációra a modellteljesítmények a tanítóhalmazra vonatkozóan a mellékletben (8.1. táblázat) érhetőek el, a teszthalmazra vonatkozóan pedig a 4.4. táblázatban találhatóak. A táblázat „Jellemzők" sora tartalmazza azokat a bemenő szignifikáns paramétereket, melyek az MLR becslésben részt vettek, az MLPNN esetén a rejtett rétegben lévő neuronszámot, valamint a GRNN és az RBFNN modellnél az alkalmazott szigma-faktort, végül zárójelben az utóbbi neurális hálózati modell esetén a rejtett rétegében lévő neuronok számát. Néhány táblázat „Jellemzők" sorában is ugyanezen meghatározott tényezők szerepelnek az MLPNN modellre vonatkozó érték nélkül (4.7. táblázat 4.8. 4.9. 4.10. és 4.11. táblázat).

A C_A kombinációra vonatkoztatva mind a négy alkalmazott modell használatát ismertettem (4.1.1, 4.1.3 és 4.1.4 pont).

A C_B kombináció esetében a D9 mintavételi pontot vizsgáltam önmagában. Ekkor is két MLR modell készült. Az első a négy bemenő paraméter közül a vízhozam és az vezetőképesség paramétereket zárta ki, mert ezek a modellben nem voltak szignifikánsak. A második MLR modell e paraméterek nélkül készült, az így kapott modell minden paraméterre szignifikáns volt (4.4. táblázat). Az MLPNN módszert 4 neuron esetében szolgáltatta a legjobb eredményt. Az

RBFNN modell teljesítménye 0,12 szigma-faktor mellett volt a legjobb. Ebben az esetben a rejtett réteg neuron száma 49 volt. Végül a GRNN modell 0,3 szigma-faktor esetében volt a leghatékonyabb. A négy különböző modell közül a RBFNN modell szolgáltatta a legjobb eredményt (4.4. táblázat).

4.4. táblázat Modellteljesítmények a teszthalmazra vonatkozóan mind a négy kombináció esetén

Kombináció	Modell	MLR	MLPNN	RBFNN	GRNN
C_A	RMSE [mg L^{-1}]	2,03	1,57	1,65	1,42
	MAE [mg L^{-1}]	1,38	1,28	1,25	1,14
	R^2	0,4	0,57	0,59	0,72
	IA	0,72	0,77	0,78	0,87
	Jellemzők	T_w, pH	5	0,26 (18)	0,3
C_B	RMSE [mg L^{-1}]	1,94	1,72	1,62	1,74
	MAE [mg L^{-1}]	1,41	1,22	1,33	1,27
	R^2	0,44	0,58	0,54	0,55
	IA	0,76	0,79	0,75	0,77
	Jellemzők	T_w, pH	4	0,12 (49)	0,3
C_C	RMSE [mg L^{-1}]	1,57	1,46	1,43	1,36
	MAE [mg L^{-1}]	1,23	1,21	1,19	1,09
	R^2	0,5	0,59	0,47	0,6
	IA	0,72	0,72	0,75	0,75
	Jellemzők	Q, T_w	2	0,28 (36)	0,5
C_D	RMSE [mg L^{-1}]	1,98	1,7	1,63	1,7
	MAE [mg L^{-1}]	1,41	1,21	1,17	1,21
	R^2	0,41	0,59	0,59	0,55
	IA	0,73	0,78	0,77	0,77
	Jellemzők	Q, T_w, pH	6	0,46 (14)	0,44

A C_C kombináció esetében a D2 mintavételi pont önálló vizsgálatát végeztem el, továbbra is két MLR modellt alkalmazva. Az első MLR modell ebben az esetben a pH és az vezetőképesség paramétert zárta ki, így a második MLR modell e paraméterek nélkül készült (4.4. táblázat). Az MLPNN modell 2 neuron esetében szolgáltatta a legjobb eredményt. RBFNN modellt alkalmazva úgy tapasztaltam, hogy 0,28 szigma-faktor esetében a leghatékonyabb ez a modell. Ebben az esetben a rejtett rétegben lévő neuronok száma 36 volt. A GRNN modell 0,5 szigma-faktornál volt a leghatékonyabb. A C_C esetében a négy modell közül a GRNN módszerrel készült modell bizonyult a leghatékonyabbnak (4.4. táblázat).

Végül mindhárom mintavételi hely adatait vizsgáltam egyszerre (C_D), mint összetett rendszert. Ez esetben is két MLR modell készült. Az első MLR modell nem volt minden paraméterre szignifikáns, ezért a vezetőképesség paraméter kimaradt a második MLR modellből, amely azonban már minden paraméterre szignifikáns volt (4.4. táblázat). Az MLPNN modell alkalmazása esetén 6 neuronnal volt a leghatékonyabb a modell. RBFNN modell futtatásakor

0,46 szigma-faktort alkalmazva értem el a legjobb eredményt, ebben az esetben a rejtett rétegben lévő neuronok száma 14 volt. A GRNN modellel akkor kaptam a legjobb eredményt, ha 0,44 szigma-faktor volt a bemeneti paraméter. Az így létrehozott összetett rendszer esetében a RBFNN modell szolgáltatta a leghatékonyabb becslést.

4.2.3. Antropogén hatások befolyása a becslésekre

A következőkben a MLPNN, a GRNN és az RBFNN modellekkel a teszthalmazra kapott statisztikai mutatókat hasonlítom össze az MLR-hez viszonyítva (4.5. táblázat).

4.5. táblázat Az egyes kombinációk teszthalmazra vonatkozó RMSE és R^2 értékeinek aránya az MLR modellhez viszonyítva

RMSE	MLR	MLPNN	RBFNN	GRNN	**R^2**	MLR	MLPNN	RBFNN	GRNN
C_A	100%	77%	81%	70%	C_A	100%	143%	148%	180%
C_B	100%	89%	84%	90%	C_B	100%	133%	123%	126%
C_C	100%	93%	91%	87%	C_C	100%	117%	94%	119%
C_D	100%	86%	82%	86%	C_D	100%	135%	143%	143%

Ha mind a négy konfigurációnak a neurális hálózatokkal a teszthalmazra vonatkozó RMSE és R^2 mutatóknak MLR-hez viszonyított értékeit összevetem (4.5. táblázat), akkor azt tapasztalom, hogy a neurális hálókkal az első konfigurációban, a zavartalan pontban (Mohács, C_A) értem el a legszámottevőbb javulást a lineáris modellhez képest. A többi konfigurációban is jelentősen hatékonyabbak a neurális hálók, de nincs olyan nagymértékű teljesítményjavulás, mint hogyha csak a zavartalan pontot vizsgálnám. A zavart mintavételi pontok esetében a jelentkező antropogén hatás, zavarás megnehezíti a becslést. A D2 mintavételi pont esetében a közeli vízerőmű működtetése jelentősen manipulálja a vízhozamot (Klaver et al., 2007; Kovács et al., 2015b), amely a teljes ökológiai rendszerre befolyást gyakorol (Liang et al., 2016; Onderka és Pekárová, 2008), így a DO paraméterre is. Másrészt a D9 mintavételi pont esetében a közeli atomerőmű hűtővize – lokálisan ugyan – de befolyásolja a Duna hőháztartását és így hatással lehet a DO paraméterre (Turnpenny et al., 2010; Wetzel, 2001). Ugyanakkor az atomerőművek környezeti hatásait vizsgálva (externális költség keretében) előnyösebbnek tűnik, mint a többi fosszilis üzemanyaggal működő erőművek (Molnár M. és Csábrági, 2010; Molnár S. és Csábrági, 2010a). Mindezek megnehezítik a DO előrejelzését a D2 és D9 mintavételi pontokon. Ugyanakkor a C_D esetben egy összetett rendszert modelleztem, melyben egyszerre megtalálhatóak zavart és zavartalan mintavételi pontok, így érthető, hogy ebben az összetett rendszerben nehezebbé vált a DO előrejelzése, mint a zavartalan mintavételi pont (D11) esetén. C_D esetében az RMSE javulás, az MLR eredményéhez képest csak 14% ellentétben a C_A esetében tapasztalt 30%-kal (4.5. táblázat).

Dunai vizsgálatom eredményei azt mutatták, hogy mindhárom neurális háló, különösen a GRNN és az RBFNN hatékony eszközök folyóvizek DO-szintjének becsléséhez még abban az esetben is, ha a folyót antropogén hatások érik, bár ebben az esetben a becslés hatékonysága csekélyebb mértékben javítható.

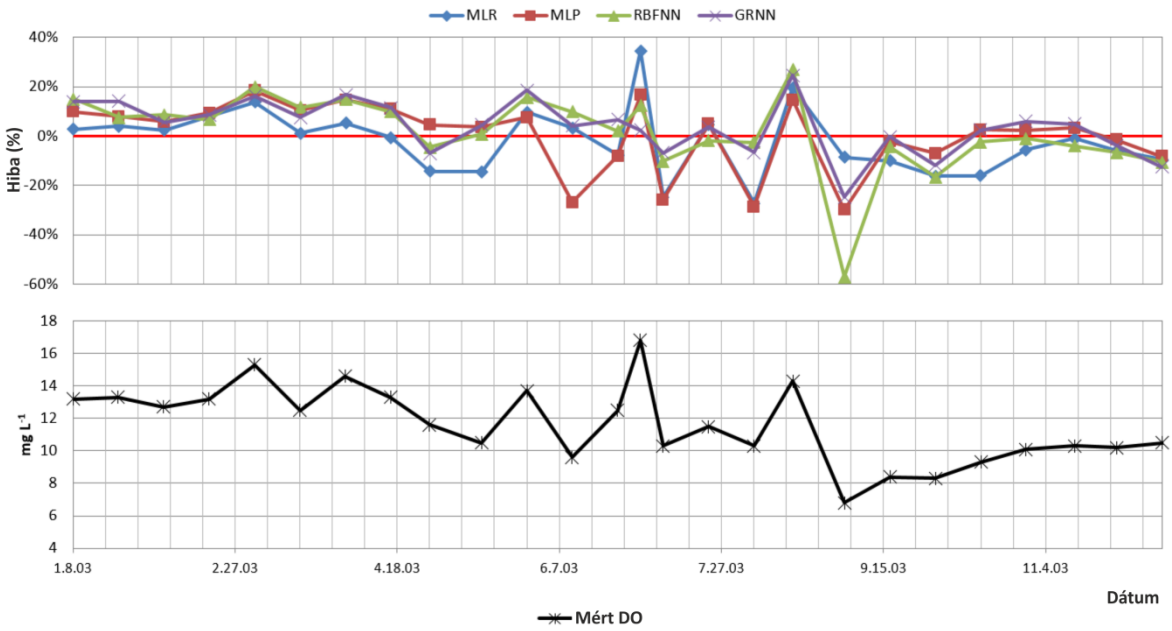
4.2.4. A leghatékonyabb modell kiválasztása a dunai vizsgálatnál

A (4.2) egyenlet alapján kiszámítottam a C_A kombinációra mind a négy alkalmazott modellre a relatív hibát (RE), amely megmutatja a százalékos hibaeloszlást (Najah et al., 2011)

$$RE(\%) = \frac{v_i - y_i}{v_i}(100\%) \tag{4.2}$$

ahol a v_i és a y_i az i-edik mért illetve becsült oldottoxigén-koncentrációjának mértéke.

A RE érték alapján megállapítható, hogy az egyes modellek hibája nem állandó az időben (4.6. ábra). Az egyes modellek közötti különbséget az adja, hogy melyik képes pontosabban megragadni a modellezett paraméter, azaz a vizsgálatomban a DO változékonyságát. A teszt halmaz alapján a legkisebb hibák azon szakaszokon várhatóak, ahol a DO varianciája kicsi (pl. ősz és tél). Ezen szakaszokon a modellek RE értékei alacsonyabbak és nincs jelentős különbségek köztük. Tavasszal és nyáron, amikor a DO koncentráció tágabb határok között változik az egyes modellek markánsan eltérő hibát produkálnak, mindamellett, hogy minden modell hibája megnő. Azonban az RE paraméter alapján is a GRNN modell bizonyult a leghatékonyabbnak ebben a kombinációban.



4.6. ábra Mind a négy modell relatív hibája a 2003-as évre vonatkozóan (Csábrági et al., 2017a)

A négy alkalmazott modell eredményeiből kitűnik (4.4. táblázat, 4.5. táblázat, és 4.6. ábra), hogy a C_A kombinációban mind tanító mind a tesztelő halmazon a GRNN modell teljesített a legjobban (8.2. ábra és 8.3. ábra).

Az egyes kombinációkra létrehozott modellek különbségei rámutattak, hogy a C_B és a C_D esetében az RBFNN modell szolgáltatta a legkisebb RMSE értéket (4.4. táblázat) a teszthalmazra vonatkoztatva. A C_C esetében pedig a GRNN modell adta a legjobb eredményt a teszthalmazra vonatkozó RMSE tekintetében. Mindegyik konfigurációnál elmondható, hogy a neurális hálózatokkal sokkal hatékonyabban lehet becsülni a DO paramétert, mint a lineáris modellel (4.5. táblázat). A három neurális háló közül a két leghatékonyabb modell a GRNN és az RBFNN, melyek jobb eredményt szolgáltattak, mint az MLPNN, amely szinte minden esetben alulmaradt a másik két modellel szemben. Egy másik előny még a GRNN és az RBFNN modellel kapcsolatban az az, hogy sokkal gyorsabb a futási idejük, szemben az MLPNN modell duplán iterált tanítási metódusával.

A négy kombinációra a teszthalmazra kapott RMSE értékek alapján megállapítható, hogy a legjobb teljesítményt a GRNN és az RBFNN modellek adták, amelyek fölülmúlták a másik neurális háló, az MLPNN teljesítményét is. A legrosszabb eredményt az MLR-rel értem el mind a négy kombináció esetén.

Megjegyezném, hogy az input paraméterekre érzékenységi vizsgálatot végeztem a Dunán C_D kombinációra mindhárom neurális hálózattal, mely azt jelezte, hogy a pH a legfontosabb a vizsgált paraméterek közül, hiszen ennek elhagyása esetén romlik legjobban a modellek teljesítménye (M4. függelék). Ez a kiegészítő vizsgálat a későbbiekben a folyó többi mintavételi pontjára is elvégzendő és egy önálló tanulmány gerincét képezheti.

4.3. Oldottoxigén-koncentrációra vonatkozó térbeli előrejelzés a Tiszán

4.3.1. A tiszai állomások mintahalmazai

A Tiszán 13 mintavételi pont van (3.2. ábra), a teljes mintaszám 1992. Az adatok 1998-2003 év közöttiek és kb. kéthetente mérték a paramétereket. Az adatok előfeldolgozásra kerültek, standardizáltam a neurális hálók használata előtt (4.1.2 pont).

A 13 mintavételi pont 1998 és 2003 között mért paramétereinek alapstatisztikáját (4.6. táblázat, ahol *r* a bemenő paraméterek és a DO közötti korrelációs együttható) és box-and-whisker diagramokat megvizsgálva (3.5 pont, 8.5. ábra) megállapítható, hogy a folyó vízhozamának átlaga megnégyszereződött a magyarországi szakaszon, vagyis Tiszasziget vízhozamának átlaga kb. négyszerese Tiszabecs vízhozam-átlagának. A Tisza pH értékének átlaga közel kéttized fokkal magasabb lett, a vezetőképességének átlaga pedig másfélszeresére nőtt a folyó folyásának irányába haladva.

A folyó 13 állomásának mért paramétereinek sztochasztikus kapcsolatait megvizsgálva az látható, hogy az oldottoxigén-koncentrációja minden mintavételi pontnál negatívan korrelál a hőmérséklettel és Tiszabecs állomást leszámítva a legjelentősebb lineáris kapcsolatban az oldott oxigén a hőmérséklettel van. A közöttük lévő korrelációs együttható (*r*) az első négy állomásnál, a folyó fölső szakaszán kevésbé hangsúlyos, 0,17 és 0,29 közé esik, ugyanakkor a Tiszalöki víztározó és az attól folyásirányban következő állomásoknál a korrelációs együttható már igen jelentőssé válik, mindenhol 0,8 fölötti értéket vesz föl. Ennek az lehet az oka, hogy a Tiszalöki vízerőmű visszaduzzasztó hatása lassítja a folyó áramlási sebességét, és így csökken a folyó turbulenciája, vagyis a turbulencia már nincs hatással a folyó DO-szintjére. Ezáltal a folyó oldottoxigén-koncentrációjára ható tényezők közül csak a hőmérséklet marad, mint egyedüli hatótényező.

Mivel minél magasabb a mintaszám, annál alacsonyabb a korrelációs együttható szignifikancia határa, ezért a korrelációs együttható szignifikancia határát elég a legkevesebb mintaszámú állomásnál megnézni. A tiszai állomások közül Aranyosapáti mintavételi pontnak a legkevesebb a mintaszáma (148). Ehhez a mintaszámhoz tartozó és 0,05 szignifikancia szint mellett a t-értékből számított korrelációs együttható akkor szignifikáns, ha $|r|>0{,}16$. Tehát úgy tűnik, hogy a Tisza folyó oldottoxigén-koncentrációjának becslésénél a hőmérséklet a legfontosabb paraméter, és minden egyes állomáson a köztük lévő korrelációs együttható szignifikáns. A többi paraméter nem minden állomáson szignifikáns.

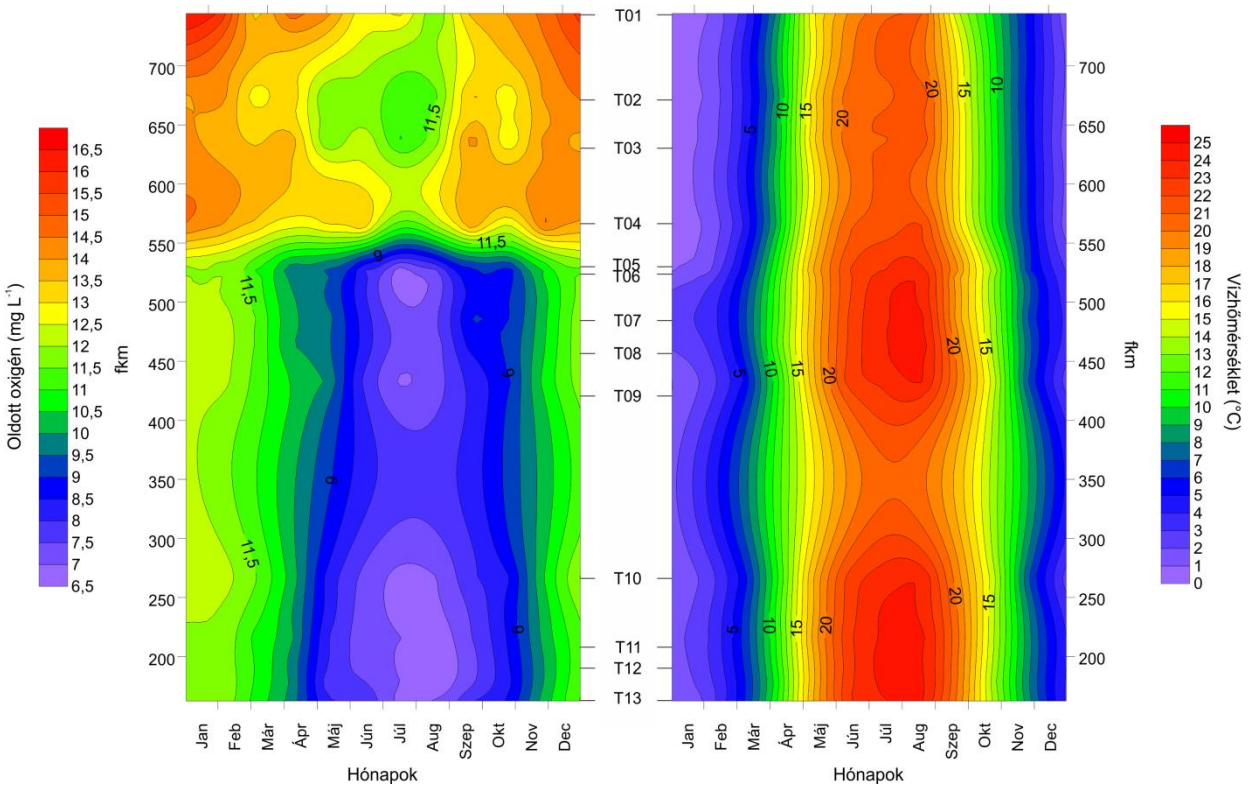
4.6. táblázat A Tiszán lévő 13 mintavételi pont alapstatisztikája a teljes vizsgált időintervallumra

Állomás	Mintaszám	Paraméter	Max	Min	Átlag	Szórás	Relatív szórás	r
T01	154	Q	3250	26,1	234,88	346	1,5	-0,32
		T_w	26,5	0	10,18	8	0,79	-0,26
		pH	8,12	7	7,74	0,21	0,03	-0,15
		EC	474	151	283,81	64,69	0,23	0,22
		DO	28,5	1,90	14,13	3,63	0,26	1
T02	148	Q	2730	48,1	383,24	440,03	1,15	-0,12
		T_w	26,1	0	11,38	7,98	0,70	-0,29
		pH	8,47	7	7,72	0,23	0,03	-0,16
		EC	670	166	351,74	107,12	0,3	0,02
		DO	23,4	6,64	12,85	2,87	0,22	1
T03	153	Q	3440	74	458,95	497,12	1,08	-0,08
		T_w	26,4	0	11,25	8,06	0,72	-0,24
		pH	8,49	7	7,75	0,21	0,03	-0,01
		EC	660	199	386,94	98,73	0,26	0,06
		DO	23,4	5,27	13,01	2,92	0,22	1
T04	152	Q	4543	52,6	479,46	599,55	1,25	-0,11
		T_w	27	0	11,44	8,19	0,72	-0,17
		pH	8,46	7	7,79	0,21	0,03	0,05
		EC	660	207	403,76	93,04	0,23	0,14
		DO	24,2	6,43	13,21	3,15	0,24	1
T05	160	Q	4543	52,6	479,46	599,55	1,25	0,09
		T_w	27	0	11,44	8,19	0,72	-0,81
		pH	8,46	7	7,79	0,21	0,03	0,002
		EC	660	207	403,76	93,04	0,23	-0,02
		DO	24,2	6,43	13,21	3,15	0,24	1
T06	152	Q	2724	10	515,43	502,33	0,97	0,12
		T_w	27,10	0	12,63	8,58	0,68	-0,8
		pH	8,73	7,41	7,84	0,21	0,03	0,04
		EC	520	191	346,65	77,30	0,22	-0,06
		DO	14,20	4,20	9,76	2,20	0,23	1

Állomás	Mintaszám	Paraméter	Max	Min	Átlag	Szórás	Relatív szórás	r
T07	152	Q	3000	62	605,51	521,06	0,86	0,13
		T_w	27	0,1	13,12	8,36	0,64	-0,81
		pH	8,5	7,48	7,85	0,19	0,02	0,15
		EC	554	198	377,36	79,69	0,21	-0,01
		DO	13,2	4,70	9,83	2	0,2	1
T08	151	Q	2260	62	594,79	482,43	0,81	0,12
		T_w	27,3	0,2	13,26	8,57	0,65	-0,82
		pH	8,47	7,4	7,86	0,19	0,02	-0,02
		EC	538	209	369,95	81,58	0,22	-0,05
		DO	13,8	5,4	9,82	2,08	0,21	1
T09	156	Q	2250	54	564,69	468,92	0,83	0,09
		T_w	27,2	0	12,72	8,77	0,69	-0,83
		pH	8,6	7,3	7,98	0,21	0,03	-0,22
		EC	660	208	373,4	84,72	0,23	0,04
		DO	13,7	4,3	9,76	2,14	0,22	1
T10	156	Q	2060	70,6	571,64	439,59	0,77	0,17
		T_w	27,6	0	12,76	8,67	0,68	-0,93
		pH	8,5	7,6	7,99	0,15	0,02	-0,37
		EC	600	227	395,02	86,28	0,22	0,03
		DO	13,5	4,5	9,51	2,20	0,23	1
T11	151	Q	2734	65,2	746,06	563,42	0,76	0,1
		T_w	28,4	0,2	13,34	8,47	0,63	-0,89
		pH	8,8	7,66	8	0,14	0,02	0,09
		EC	660	215	387,7	88,52	0,23	0,01
		DO	12,9	5	9,22	2,07	0,22	1
T12	151	Q	2734	65	752,11	566,79	0,75	0,09
		T_w	29,1	0	13,01	8,57	0,66	-0,87
		pH	8,38	7,14	7,89	0,21	0,03	-0,03
		EC	685	220	384,59	90,74	0,24	0,04
		DO	12,90	4,50	9,15	2,15	0,24	1
T13	156	Q	3220	101	938,2	668,72	0,71	0,03
		T_w	28,4	0	13,01	8,56	0,66	-0,87
		pH	8,46	7	7,96	0,2	0,03	0,04
		EC	750	243	422,25	99,09	0,23	0,12
		DO	13,20	4,8	9,35	2,03	0,22	1

A vizsgált időintervallumon belül megnéztem havonkénti felbontásban az egyes mintavételi pontokon mért oldott oxigén (4.7a ábra) illetve hőmérséklet értékek átlagát (4.7b ábra). A hőmérséklet átlagértékei jelentősen, (mintegy 3 fokkal) nőttek a folyó alsó szakasza felé, déli irányába haladva. Az is látszik, hogy a Tisza oldottoxigén-koncentrációjának átlagos szintje jelentősen, több mint 30 százalékkal csökkent a folyó alsó szakaszán a folyó fölső szakaszához képest. Észrevehető, hogy mennyire elkülönül a folyó fölső négy mintavételi pontján mért DO

koncentrációjának struktúrája a többi, délebbre lévő mintavételi pontokon mért DO koncentrációjának struktúrájától (4.7a ábra).



a) oldott oxigén b) hőmérséklet

4.7. ábra Tiszai állomások 1998 és 2003 közötti években mért paramétereinek havi átlagai

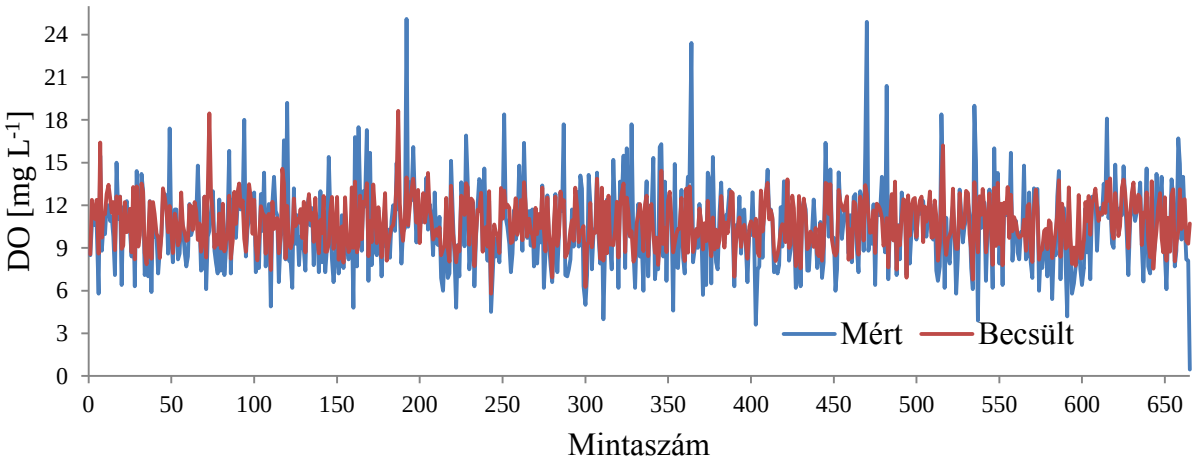
4.3.2. Referenciamodell

Az első konfigurációban a Tisza folyó vizsgálatakor a teljes adathalmazból véletlenszerűen választottam ki a tanító és teszthalmazokat 2/3-1/3 arányban. A háromféle, véletlenszerű kiválasztás (800, 2000, 1000 kezdőértékkel inicializálva) során létrejött TC1-A, TC1-B és TC1-C beállítások teszthalmazra vonatkozó eredményeit a 4.7. táblázat tartalmazza, ahol a táblázat „Jellemzők” sora - az MLPNN modell kivételével - ugyanazokat az alkalmazott modellekre vonatkozó tényezőket tartalmazza, mint a 4.4. táblázat ugyanezen sora. Az 4.8. , a 4.9. , a 4.10. és 4.11. táblázat jellemzők sorában is ugyanezen kiszámított jellemzők szerepelnek. A kapott eredmények rámutattak arra, hogy a többváltozós lineáris regressziós modelleknél egy kicsivel pontosabb becslést adnak az neurális hálózatokkal való becslések.

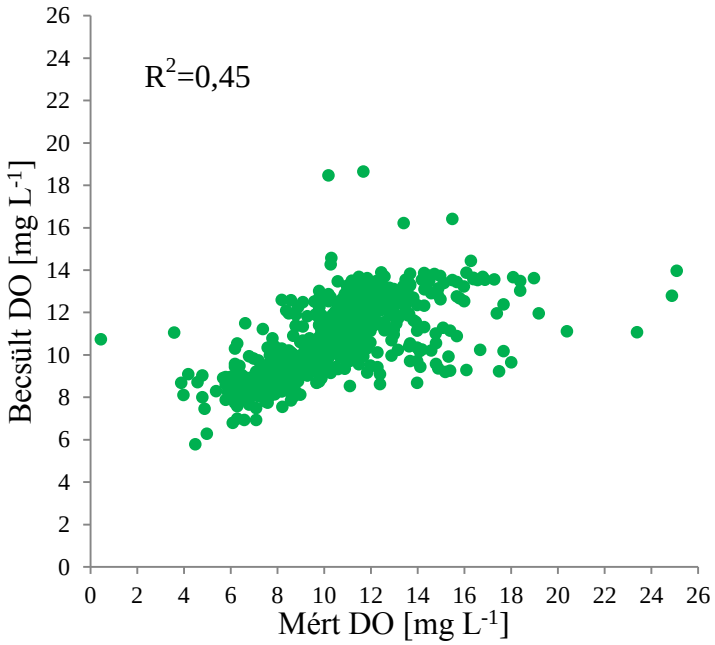
4.7. táblázat A TC1 konfiguráció eredményei a teszthalmazra vonatkozóan

Beállítások	Modell	MLR	GRNN	RBFNN
TC1-A	RMSE [mg L^{-1}]	2,40	2,33	2,38
	MAE [mg L^{-1}]	1,69	1,57	1,61
	R^2	0,35	0,39	0,39
	IA	0,72	0,76	0,77
	Jellemzők	Q, T_w, EC	0,38	0,45 (36)
TC1-B	RMSE [mg L^{-1}]	2,25	2,14	2,23
	MAE [mg L^{-1}]	1,62	1,47	1,58
	R^2	0,41	0,46	0,41
	IA	0,75	0,76	0,76
	Jellemzők	Q, T_w, EC	0,52	0,58 (15)
TC1-C	RMSE [mg L^{-1}]	2,35	2,31	2,26
	MAE [mg L^{-1}]	1,66	1,59	1,58
	R^2	0,40	0,42	0,44
	IA	0,74	0,74	0,78
	Jellemzők	Q, T_w, EC	0,54	0,52 (24)

A legjobb eredmény a TC1-B kombináció esetén GRNN modellel született szinte minden statisztikai mutató alapján, kivéve a Willmott-féle egyezési indexet (IA), mely szerint a leghatékonyabb becslés az TC1-C kombinációban az RBFNN modellel született. A 3.6 pontban leírt okok miatt a modellek összehasonlításakor az RMSE értéket veszem figyelembe a többi statisztikai mutatóval szemben, ezért az TC1-B kombináció lesz a referenciamodell (4.8. ábra és 4.9. ábra), és a kapott RMSE érték 2,14 mg L^{-1} lesz az a referenciaérték, amihez a többi konfiguráción belül kapott eredményeket hasonlítom.



4.8. ábra A TC1-B beállítás mért és a GRNN modellel becsült DO-szintjének vonaldiagramja



4.9. ábra TC1-B beállítás mért és GRNN modellel becsült DO-szintjének pontfelhő diagramja

A háromféle beállítás eredményei alapján (TC1-A, TC1-B és TC1-C) 10,69-10,73 mg L^{-1} volt az átlagos becsült oldottoxigén-koncentráció a vizsgált időszakban. Fontos kiemelni, hogy ebben a konfigurációban a teljes magyarországi folyószakaszra egyetlen becslést kaptam a létrehozott modellekkel. Ez azt feltételezi, hogy a vizsgált teljes Tisza szakasz (594,5 fkm) egységes tulajdonságokkal rendelkezik, így egyetlen modell leírhatja az azon tapasztalható oldottoxigén-koncentrációt. Azonban ez a feltevés nagy folyók hosszú szakaszaira nem teljesül (Chapman et al., 2016; Kovács et al., 2015b; Šiljić Tomić et al., 2016; 2018b), akárcsak a vizsgált Tisza szakasz esetében sem, amely nem rendelkezik homogén tulajdonságokkal (Tanos et al., 2015). A fentiek szükségessé tették, hogy a vizsgált folyót térben irányítottan osszam szét tanító és teszthalmazra, majd a különböző felosztásokra hozzak létre releváns modelleket, ez indokolta a második illetve majd a harmadik konfiguráció bevezetését.

4.3.3. Irányított kiválasztás

A TC2 konfigurációban szintén a teljes adathalmazt használtam föl a vizsgálathoz úgy, hogy mindig 9 ponttal tanítottam és négy szomszédos mintavételi ponttal teszteltem annak érdekében, hogy a folyó különböző szakaszaira is külön-külön modellt tudjak alkotni.

A következő felosztásokat alkalmaztam (3.3. táblázat): először a folyó fölső négy mintavételi pontjára teszteltem (T01, T02, T03 és T04, ezt TC2-A-val jelöltem), majd az utolsó állomás adatait megtartva a következő három állomást hozzávéve a teszthalmazhoz kaptam a következő esetet (T04, T05, T06 és T07, ezt TC2-B-vel jelöltem). Ezután megint az utolsó mintavételi ponthoz hozzávettem a következő három mintavételi pontot, és megkaptam a harmadik esetet (T07, T08, T09 és T10, ezt TC2-C-nek neveztem el), és a negyedik esetben végül az utolsó négy mintavételi pont adatai alkották a teszthalmazt (T10, T11, T12 és T13, ez pedig legyen TC2-D).

A TC2-A beállításnál kaptam a legrosszabb teljesítményt mindhárom modellel (4.8. táblázat). A legjobb eredményt - 1,64 mg L^{-1} RMSE és 1,32 mg L^{-1} MAE értéket - a GRNN modell szolgáltatta a TC2-C beállítás esetén, kivéve az R^2 és az IA értékét, amelyek egy-két századdal jobbak a lineáris modellnél. A TC2-A és a TC2-B esetben is a GRNN modell adta, igaz csekély mértékben a legjobb eredményt az RMSE tekintetében, de a többi mutató sem tért el sokkal

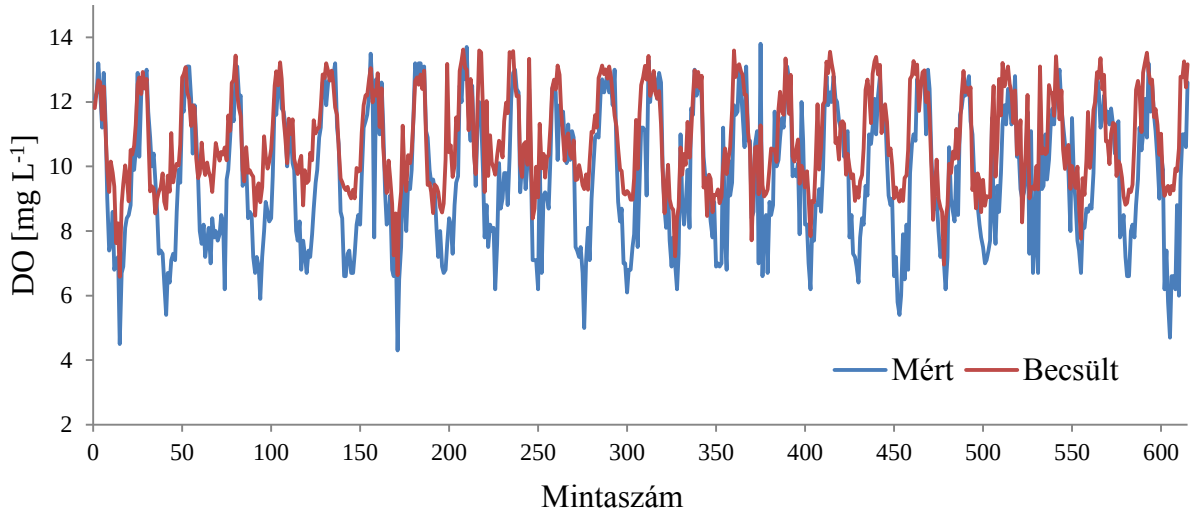
egymástól. A TC2-D beállítás esetében viszont a RBFNN modellel értem el a legjobb teljesítményt, minden statisztikai mutató ebben az esetben volt a legjobb, kivéve az R^2 értékét.

4.8. táblázat A TC2 konfiguráció eredményei a teszthalmazra vonatkozóan

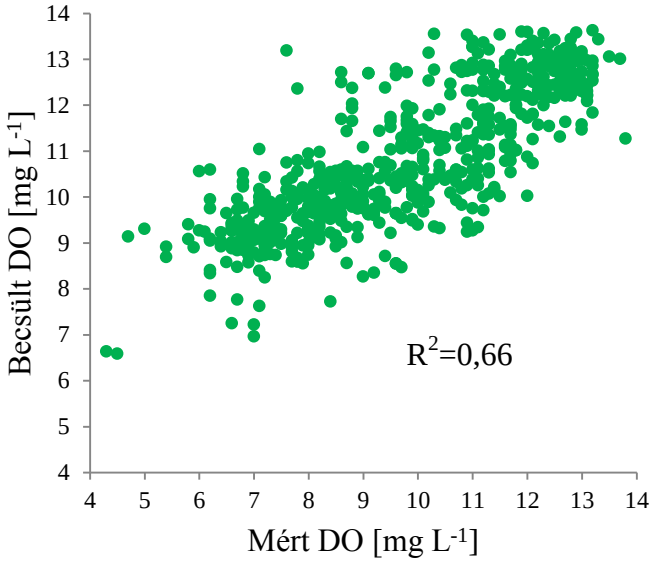
Beállítás	Modell	MLR	GRNN	RBFNN
TC2-A	RMSE [mg L^{-1}]	4,75	4,73	4,74
	MAE [mg L^{-1}]	3,64	3,72	3,76
	R^2	0,07	0,06	0,08
	IA	0,48	0,47	0,47
	Jellemzők	Q, T_w, pH, EC	1,24	0,41 (3)
TC2-B	RMSE [mg L^{-1}]	2,66	2,56	2,59
	MAE [mg L^{-1}]	1,95	1,78	1,87
	R^2	0,20	0,25	0,26
	IA	0,63	0,68	0,69
	Jellemzők	Q, T_w, pH, EC	0,3	0,35 (89)
TC2-C	RMSE [mg L^{-1}]	1,65	1,64	1,75
	MAE [mg L^{-1}]	1,38	1,32	1,36
	R^2	0,67	0,66	0,50
	IA	0,82	0,80	0,79
	Jellemzők	Q, T_w, EC	0,67	0,63 (10)
TC2-D	RMSE [mg L^{-1}]	1,71	1,81	1,69
	MAE [mg L^{-1}]	1,46	1,39	1,27
	R^2	0,70	0,61	0,60
	IA	0,81	0,79	0,83
	Jellemzők	Q, T_w, EC	0,37	0,56 (38)

A TC2 konfiguráció beállításai pontosabb becslést adtak bizonyos szakaszokra, mint a referenciamodell. A folyó fölső szakaszán rosszabb teljesítményű modellek készültek - TC2-A, TC2-B beállítások - a referenciamodellhez képest. A folyó alsóbb szakaszaira való előrejelzés esetében viszont TC2-C és TC2-D beállítások hatékonysága már felülmúlta az referenciamodell eredményeit (4.12. ábra). TC2-A esetén a teszthalmaz kimeneti tartományának terjedelme 1,9-28,5 mg L^{-1}, a tanítóhalmaz kimeneti tartományának terjedelme viszont 3,6-14,2 mg L^{-1}. Mivel a neurális hálózatok a tanítóhalmaz terjedelmén kívül értékekre nem képesek megfelelő becslést adni (Maier és Dandy, 2000), ezért a TC2-A beállítás esetében elért kedvezőtlen eredmény nem meglepő. Ezt támasztja alá az is, hogy a TC2-A és TC2-B beállítások esetében a tanító és teszthalmazokat alkotó mintavételi pontpárok átlagos CCDA-val számított különbsége (3.2. táblázat) 3%-kal magasabb - 22% versus 25% -, mint a TC2-C és TC2-D beállítások esetében. Ha csupán csak a teszt halmazok mintavételi pontjainak CCDA-val számított különbségét tekintjük a TC2-A és TC2-B beállításokban, akkor 19%-os az átlagos különbség, míg a TC2-C és TC2-D beállítások esetén csupán 10%. Ez a 3%-os illetve 9%-os különbség is alátámasztja a fenti megállapítást (3.2. táblázat).

További problémát okozott a TC2 konfiguráció esetén az a megközelítés, hogy a teszthalmaz 4 mintavételi pontja az egyes beállításoknál megelőzi a tanítóhalmaz mintavételi pontjait, tehát az a kényszerű eset áll elő, hogy az alvízi részek tulajdonságait felhasználva szeretném megbecsülni a felvízi szakaszt. Ennek hatása jelentkezik különösen hangsúlyosan a TC2-A modell eredményében. A TC2 konfiguráció koncepciója azonban mégis indokolt – annak érdekében, hogy a folyó különböző szakaszaira tudjak becslést adni – mert így javíthatóak a referenciamodell becslései, amely a TC2-C (4.10. ábra és 4.11. ábra), és TC2-D beállítás esetében teljesült is. A TC2 konfiguráció szükségszerű volt a folyó térbeli jellemzéséhez (a teljes adathalmaz alapján), azonban ennek megvalósítása nem volt megfelelően hatékony a felsőbb szakaszok esetén. A becslés hatékonyságának növelése érdekében térbeli optimalizációt hajtottam végre, hiszen rendelkezésre álltak Tanos és szerzőtársai (2015) eredményei, amelyet a TC3 konfigurációban használtam föl a 3.7.2 pont és a 3.14. ábra szerint.



4.10. ábra A TC2-C beállítás mért és a GRNN modellel becsült DO-szintjének vonaldiagramja



4.11. ábra TC2-C beállítás mért és GRNN modellel becsült DO-szintjének pontfelhő diagramja

4.3.4. Térbeli optimalizáció a Tiszán

A TC3 konfigurációban (3.3. táblázat) nem használtam föl a folyó teljes adathalmazát, hanem mindig csak három állomás adataival dolgoztam úgy, hogy két állomás egy homogén csoport

elemei voltak, a harmadik állomás pedig a folyásirányban következő állomás volt. Ebből két mintavételi pont adatai szolgáltatták a tanítóhalmaz elemeit, a harmadik állomás adatai pedig a teszthalmaz elemei voltak. Ennél a konfigurációnál is fönnáll a tanítóhalmaz és a teszthalmaz 2/3-1/3-os aránya.

Három esetet vizsgáltam (3.7.2 pont): a folyó fölső szakaszából három mintavételi pontot (T03, T04, T05, ez az „A" beállítás, 4.9. táblázat), ekkor a T03 és a T04 állomások alkották a homogén csoportot, a T05 állomás pedig ettől a csoporttól eltérő tulajdonságú, a három albeállítás jelölései: TC3-A#1, TC3-A#2 és TC3-A#3. A TC3-A#1 albeállítás esetén R^2 megbízhatatlanságát (3.6 pont) lehet tapasztalni például az MLR modellnél, ahol indokolatlanul nagy, 0,65 volt az R^2 értéke, holott az RMSE értéke viszont elég magas, 3,93 mg L^{-1} volt.

4.9. táblázat A TC3 konfiguráció „A" beállításának eredményei a teszthalmazra vonatkozóan

Albeállítás	Tanítóállomás + tesztállomás	Modell	MLR	GRNN	RBFNN
TC3-A#1	T3, T4 + T5	RMSE [mg L^{-1}]	3,93	3,70	3,58
		MAE [mg L^{-1}]	3,60	3,34	3,14
		R^2	0,65	0,42	0,34
		IA	0,93	0,51	0,54
		Jellemzők	T_w	0,76	0.59 (18)
TC3-A#2	T3, T5 +T4	RMSE [mg L^{-1}]	3,49	3,12	3,05
		MAE [mg L^{-1}]	2,48	2,21	2,24
		R^2	0,06	0,22	0,24
		IA	0,50	0,64	0,65
		Jellemzők	T_w, pH, EC	0,32	0.33 (80)
TC3-A#3	T4, T5 + T3	RMSE [mg L^{-1}]	3,35	2,93	2,97
		MAE [mg L^{-1}]	2,50	2,04	2,19
		R^2	0,06	0,24	0,22
		IA	0,50	0,64	0,64
		Jellemzők	T_w, pH, EC	0,40	0,37 (67)

A folyó középső szakaszából egy beállítással foglalkoztam (T07, T08, T09, ezt „B" beállításnak jelöltem, 4.10. táblázat), ebben az esetben a T07 és T08 állomások voltak homogének, és a T09 mintavételi pont volt más struktúrájú, a három albeállítás jelölései: TC3-B#1, TC3-B#2 és TC3-B#3.

4.10. táblázat A TC3 konfiguráció „B” beállításának eredményei a teszthalmazra vonatkozóan

Albeállítás	Tanítóállomás + tesztállomás	Modell	MLR	GRNN	RBFNN
TC3-B#1	T7, T8 + T9	RMSE [mg L^{-1}]	1,25	1,17	1,23
		MAE [mg L^{-1}]	0,98	0,88	0,94
		R^2	0,76	0,78	0,74
		IA	0,89	0,90	0,88
		Jellemzők	Q, T_w, pH, EC	0,44	0,15 (22)
TC3-B#2	T7, T9 +T8	RMSE [mg L^{-1}]	1,01	0,84	0,90
		MAE [mg L^{-1}]	0,73	0,61	0,69
		R^2	0,78	0,84	0,83
		IA	0,93	0,95	0,95
		Jellemzők	Q, T_w, pH, EC	0,36	0,1 (101)
TC3-B#3	T8, T9 + T7	RMSE [mg L^{-1}]	1,00	0,84	0,89
		MAE [mg L^{-1}]	0,76	0,59	0,68
		R^2	0,77	0,84	0,82
		IA	0,93	0,96	0,95
		Jellemzők	Q, T_w, pH, EC	0,34	0,11 (76)

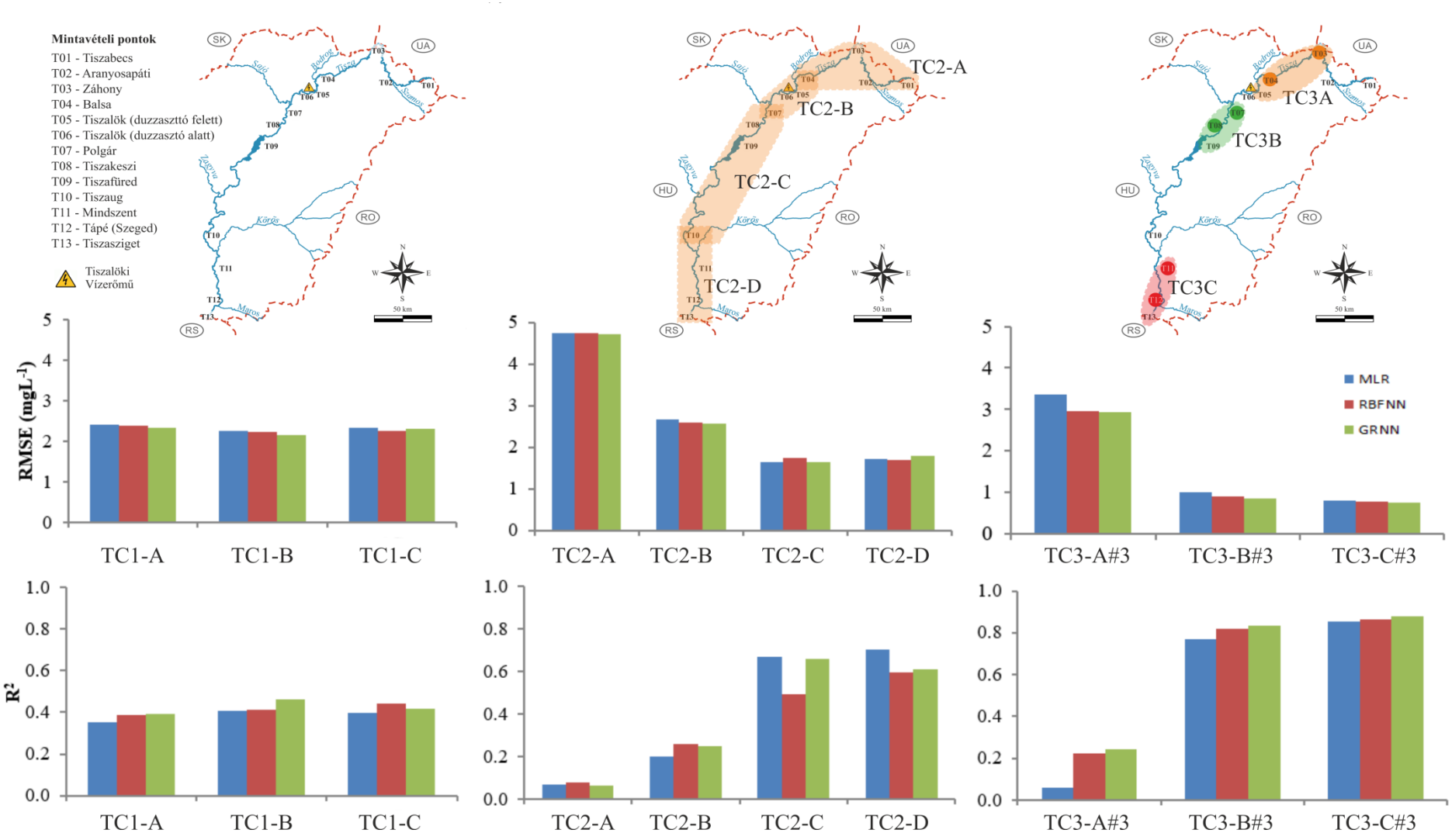
Végül a folyó utolsó három magyarországi állomását (T11, T12 és T13, ez lesz a „C” beállítás, 4.11. táblázat) vizsgáltam, ahol a T13 mintavételi pont volt más struktúrájú a T11 és T12 állomások által alkotott homogén csoporttal szemben. Mindegyik beállításnál az összes lehetséges esetet, vagyis mindhárom albeállítást kiszámítottam, jelölései: TC3-C#1, TC3-C#2 és TC3-C#3.

4.11. táblázat A TC3 konfiguráció „C” beállításának eredményei a teszthalmazra vonatkozóan

Albeállítás	Tanítóállomás + tesztállomás	Modell	MLR	GRNN	RBFNN
TC3-C#1	T11, T12 + T13	RMSE [mg L^{-1}]	0,87	0,77	0,86
		MAE [mg L^{-1}]	0,64	0,60	0,66
		R^2	0,83	0,86	0,82
		IA	0,95	0,96	0,95
		Jellemzők	Q, T_w, pH	0,56	0,08 (34)
TC3-C#2	T11, T13 +T12	RMSE [mg L^{-1}]	0,95	0,75	0,76
		MAE [mg L^{-1}]	0,67	0,53	0,56
		R^2	0,81	0,89	0,88
		IA	0,94	0,97	0,96
		Jellemzők	Q, T_w, pH	0,38	0,07 (65)
TC3-C#3	T12, T13 + T11	RMSE [mg L^{-1}]	0,80	0,74	0,77
		MAE [mg L^{-1}]	0,64	0,58	0,60
		R^2	0,85	0,88	0,86
		IA	0,96	0,96	0,96
		Jellemzők	Q, T_w, pH	0,48	0,1 (24)

4.3.5. Modelleredmények összehasonlítása

Mindhárom konfiguráció beépített térképeit és eredményeit egy összefoglaló ábrán mutatom be (4.12. ábra), amely a teszthalmazra vonatkozó RMSE és R^2 értékét jeleníti meg. A TC3 konfiguráció esetén mind a három beállításon belül a harmadik albeállításnál értem el a legjobb eredményt (TC3-A#3, TC3-B#3 és TC3-C#3), és mindegyik albeállításnál a GRNN modell szolgáltatta a leghatékonyabb becslést minden statisztikai mutató vonatkozásában (4.9. táblázat, 4.10. táblázat, és 4.11. táblázat). A „B” beállításnál a második és a harmadik albeállításnál szinte majdnem azonos RMSE értékek születtek a GRNN modellt alkalmazva, de a harmadik albeállításban négyezreddel lett kevesebb az RMSE érték. A MAE és az IA értéke is a harmadik albeállításban lett jobb, az R^2 értéke viszont mindkét esetben megegyezik. Így a három albeállítás közül ezt az esetet választottam ki a legjobb becslést adónak, és ezért a TC3-A#3, a TC3-B#3 és a TC3-C#3 albeállításoknak az elért eredményei kerültek az összefoglaló ábrára (4.12. ábra).

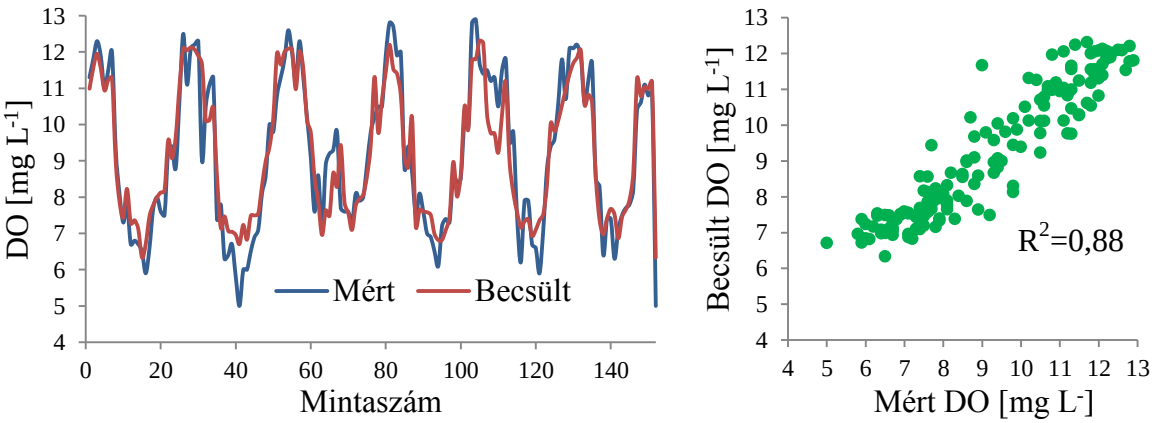


a) TC1konfiguráció b) TC2 konfiguráció c) TC3 konfiguráció

4.12. ábra Összefoglaló ábra: A tiszai konfigurációk beépített térképei és eredményei

A TC3 konfigurációban három különböző beállítást vizsgáltam (TC3-A, TC3-B és TC3-C), melyeken belül 3-3 albeállítás volt elkülöníthető. A TC3-A beállítással kapott RMSE értékek kisebbek voltak, mint 4 mg L^{-1}, amely kedvezőbb a TC2-A eredményéhez képest, de ez még mindig 37%-kal elmarad a referenciaértéktől. Ez magyarázható azzal is, hogy a T05-ös állomás a Tiszalöki duzzasztómű hatása (antropogén hatás) miatt nagyon elkülönül a másik két ponttól. A vízerőmű, mint megújuló energiaforrásokon alapuló termelési technológia mégis sokkal kedvezőbb kárköltséggel jár együtt, mint a fosszilis energiahordozóval üzemelő erőművek (Molnár S. és Csábrági, 2010b). Ez az antropogén hatás megfigyelhető a mintavételi pontok közötti CCDA-val kiszámított különbségekben, hiszen a T05 a T04 és T03 közötti CCDA-s különbségei sokkal nagyobbak (3.2. táblázat, 0.24 és 0.23), mint a TC3-B ill. TC3-C beállítások mintavételi pontjai között (3.2. táblázat, 0,11 és 0,12 illetve 0,1 és 0,08). A másik két beállítással kapott eredmények már (TC3-B és TC3-C) jóval túlszárnyalják a referenciaértéket, vagyis az alsóbb szakaszokra a TC2 konfiguráció mellett a TC3 konfiguráció esetében is hatékonyabb a becslés (4.12. ábra).

Mindhárom konfiguráció eredményét megvizsgálva a TC3 konfiguráción belül a C beállításnál (4.11. táblázat; TC3-C#1, TC3-C#2 és TC3-C#3 esetek) értem el a modellekkel a legeslegjobb eredményeket. A TC3-C beállításon belül a GRNN modell a TC3-C#3 albeállítás esetén eredményezte a legpontosabb becslést, hiszen a számított RMSE érték 0,74 mg L^{-1} volt, ez pedig 65%-kos teljesítményjavulás jelent a referenciaértékhez képest. Ekkor a T11 (Mindszent) mintavételi hely adatai alkották a teszthalmazt (4.13. ábra). A TC3-C#2 eset egyszázaddal rosszabb RMSE értéket (0,75 mg L^{-1}) eredményezett a GRNN modell vonatkozásában, viszont R^2 tekintetében szintén egy százalékkal jobb eredmény született, mint a TC3-C#3 esetben. Vagyis egy-egy mutatóban a TC3-C#3, más-más mutatóban a TC3-C#2 tűnik hatékonyabbnak. A két eset közül - a 3.6 pontban tárgyaltak szerint - az RMSE számít, tehát a TC3-C#3 tekinthető a leghatékonyabb modellnek az összes tiszai konfiguráción belül (4.3.7 pont).



a) vonaldiagram b) pontfelhő diagram

4.13. ábra TC3-C#3 beállítás mért ill. GRNN modellel becsült DO-szintjének diagramjai

4.3.6. A tanítóhalmaz leghatékonyabb adatstruktúrája tiszai adatokon

A TC3 konfiguráció „A”, „B”, és „C” beállításainak eredményei rámutattak arra, hogy azok az albeállítások adják a legrosszabb teljesítményt, amikor kizárólag a homogén csoport mintavételi pontjai szerepelnek a tanítóhalmazban, és az inhomogén mintavételi pont szerepel a teszthalmazban (4.9. táblázat, 4.10. táblázat, és 4.11. táblázat; TC3-A#1, TC3-B#1,

TC3-C#1). Vagyis a modellek hatékonysága növelhető azzal, ha a tanítóhalmaz tartalmazza a homogén csoport egyik mintavételi pontját és az inhomogén mintavételi pontot, így a tanítóhalmaz tartalmazza mind a homogén, mind az inhomogén mintavételi pont adatainak struktúráit, ezzel elősegítve a hatékonyabb becslést. Ekkor a tanítóhalmaz „kevert struktúrájú", így az ilyen eseteket „kevert struktúrának" neveztem el.

Mindemellett a TC1 konfiguráció egy jó példát mutat be a „túlságosan kevert" struktúra alkalmazásának kedvezőtlen hatására. Ugyanis a TC1 konfigurációban is egy kevert struktúrájú rendszer volt a tanítóhalmazban és teszthalmazban is, köszönhetően a véletlen kiválasztásnak. Azonban ez már annyi különböző struktúrát jelentett (jelentős különbségek vannak a vizsgált folyószakasz egyes részein, 3.2. táblázat, 3.14. ábra), amely már nem segítette elő a pontos becslést.

Megjegyzendő azonban, hogy a TC3 konfigurációban annak ellenére sikerült sokkal hatékonyabb modelleket előállítani, hogy ekkor a tanítóhalmaz mérete csupán nagyjából 300 megfigyelésből állt, ellentétben az TC1 és TC2 konfigurációval, ahol ez az érték 1300-1400 között volt. A szakirodalom szerint a megfigyelések számának növekedésével nő a becslés pontossága is (Hastie et al., 2009; Reddy, 2011, statisztikai konzisztencia) feltételezve, hogy azonos a populáció. Mindennek ellenére az eredmények rámutattak arra, hogy a csökkenő mintaszám ellenére a tanító és a teszthalmaz tudatos/irányított kijelölésével, illetve térbeli optimalizációval jelentősen javítható a modellek hatékonysága, ami nem mond ellen a statisztikai konzisztenciának, hiszen a Tisza magyarországi szakasza nem tekinthető azonos struktúrájúnak.

4.3.7. A folyó térbeli szakaszainak jellemzése

A TC2 és a TC3 konfiguráció eredményeit összevetve a referenciamodellel azt kaptam, hogy a folyó fölső szakaszán rosszabb teljesítményű modellek készültek (TC2-A, TC2-B és a TC3-A beállítások), ez pedig rámutat arra, hogy a fölsőbb folyószakasz jelentősebb változékonysága megnehezíti a becslést (Reynolds, 1984; Stanković, et al., 2012).

Ugyanakkor a TC2-C és TC2-D illetve a TC3-B és TC3-C beállítások hatékonysága, amelyek a folyó alsóbb szakaszaira adtak előrejelzést, már felülmúlták az referenciamodell eredményeit (4.12. ábra).

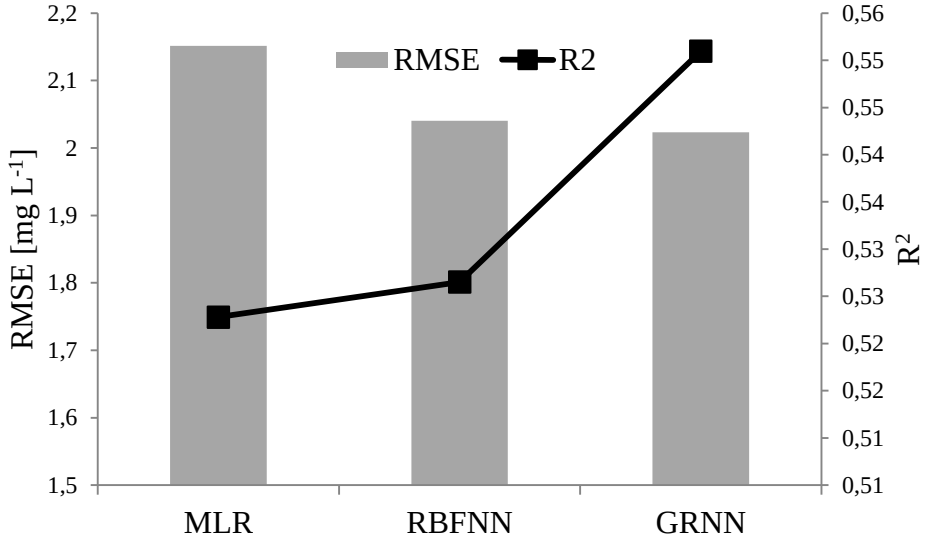
Összességében a TC2 és a TC3 konfiguráció eredményei is megerősítették, hogy a folyó alvízi szakaszain könnyebb a DO becslése, amely visszavezethető arra, hogy az áramlási sebesség lassulásával a vízminőségi paraméterek változékonysága is csökken, így egy könnyebben becsülhető rendszer keletkezik. Ennek eredményeképpen a legpontosabb becslést is a folyó alsó szakaszán értem el (TC3-C#3, 4.13. ábra) kevert struktúra alkalmazásával.

Mindazonáltal a folyóban a vízminőségi változók folyásirány szerinti változékonyságának csökkenését korábban néhány kutatás leírta (Csábrági et al., 2017b; Kovács et al., 2017; Tanos et al., 2015), amelynek következményeit tapasztaltuk a TC2 és a TC3 konfiguráció eredményeiben.

4.3.8. A tiszai konfigurációk leghatékonyabb modelljének kiválasztása

Az összehasonlítás eredményeit a teszthalmazra vonatkoztatva kell megvizsgálni, ezért mindhárom konfiguráción belül, mindhárom modell hatékonyságát jellemző kiszámított értékek a tanítóhalmazra vonatkoztatva a mellékletben érhetőek el (8.4. táblázat, 8.5. táblázat, 8.6. táblázat, 8.7. táblázat és 8.8. táblázat).

Az alkalmazott modellek hatékonyságát megvizsgálva a három konfiguráció összes RMSE és R^2 eredményét átlagolva azt az eredményt kaptam, hogy a neurális hálózatok hatékonyabb eszköznek bizonyultak, mint a többváltozós lineáris regresszió módszere (4.14. ábra). Az összehasonlítás eredménye nem meglepő, az alábbi tudományos cikkekben is a neurális modellek adtak pontosabb becsléseket a lineáris modellnél (Akkoyunlu et al., 2011; Antanasijević et al., 2013; Ay és Kisi, 2012; Chen és Liu, 2015; Csábrági et al., 2017a; 2015b; Heddam, 2014a; Ji et al., 2017). A neurális hálózatok közül a GRNN volt a leghatékonyabb modell a tiszai konfigurációk eredményei alapján.



4.14. ábra Mindhárom tiszai konfigurációban kapott RMSE és R^2 értékek átlagai

4.4. Oldottoxigén-koncentrációra vonatkozó térbeli előrejelzés a Dunán

4.4.1. A dunai állomások mintahalmazai

A Dunán 12 mintavételi pont van (3.2. ábra), a teljes mintaszám 2028. A vizsgált adatok 1998-2003 év közöttiek és kb. kéthetente mérték a paramétereket. Az adatok előfeldolgozásra kerültek, standardizáltam a neurális hálók használata előtt (4.1.2 pont).

A 12 mintavételi pont 1998 és 2003 között mért paramétereinek alapstatisztikáját (4.12. táblázat, ahol *r* a bemenő paraméterek és a DO közötti korrelációs együttható) megállapítható, hogy a négy bemenő paraméter és a kimenetet adó oldott oxigén paraméter közül a legváltozékonyabb két paraméter a hőmérséklet és a vízhozam, amit e két paraméter magas relatív szórás értéke mutat meg. A korrelációs együttható szignifikancia határát elég a legkevesebb mintaszámú állomásnál megnézni. A három dunai állomás közül Budapest mintavételi pontnak a legkevesebb a mintaszáma (136), ehhez a mintaszámhoz tartozó és 0,05 szignifikancia szint mellett a t-értékből számított korrelációs együttható akkor szignifikáns, ha $|r|>0,17$. A dunai összadathalmazt és a másik három állomás adatait megvizsgálva úgy tűnik, hogy az oldottoxigén-koncentrációja és a hőmérséklet között a legmagasabb a korrelációs együttható és minden egyes állomáson a köztük lévő korrelációs együttható szignifikáns. A többi paraméter, főleg a vízhozam nem minden állomáson szignifikáns.

4.12. táblázat Duna mintavételi pontjainak alapstatisztikája

Állomások	Minta-szám	Paraméter	Max	Min	Átlag	Szórás	Relatív szórás	r
Duna összes állomása	2028	Q	7590	237	2150,71	1041,33	0,48	0
		T_w	26,2	0	12,14	7,09	0,58	-0,40
		pH	9,2	6,9	8,21	0,28	0,03	0,21
		EC	901	244	384,16	67,07	0,17	0,30
		DO	17,7	5,76	10,49	1,73	0,16	1
D6	136	Q	6550	908	2301,15	999,14	0,43	-0,01
		T_w	24,7	0	11,99	7,17	0,60	-0,52
		pH	9,2	6,9	8,26	0,38	0,05	0,12
		EC	570	280	383,21	64,89	0,17	0,43
		DO	13	6,2	9,94	1,53	0,15	1
D7	146	Q	6550	908	2360,6	1003,71	0,43	0
		T_w	25,4	0	11,96	7,25	0,61	-0,54
		pH	9,2	7	8,28	0,33	0,04	0,13
		EC	570	280	386,27	67,57	0,17	0,46
		DO	13,4	6	10,05	1,53	0,15	1
D8	153	Q	5350	915	2299,28	898,72	0,39	-0,17
		T_w	25,3	0	12,36	7,17	0,58	-0,35
		pH	8,8	7,85	8,26	0,25	0,03	0,22
		EC	520	256	373,1	57,09	0,15	0,33
		DO	15,7	6,5	11,16	1,72	0,15	1

4.4.2. Mindkét konfiguráció eredményei

A Duna folyó oldottoxigén-koncentrációját becsültem négy bemenő paraméterrel és két konfigurációban a tanító és teszthalmaz kiválasztásának különböző módszere szerint. Az első konfigurációban a Duna folyó vizsgálatakor a teljes adathalmazból véletlenszerűen választottam ki a tanító és teszthalmazokat 2/3-1/3 arányban. Ebben az esetben a folyó teljes magyarországi szakaszára adtam becslést az oldottoxigén-koncentrációjára vonatkoztatva. A véletlenszerű kiválasztás (2000 kezdőértékkel inicializálva) során létrejött D_R beállítás teszthalmazra vonatkozó eredményeit a 4.13. táblázat tartalmazza, ahol a táblázat „jellemzők” sora tartalmazza azokat a bemenő paramétereket, melyek az MLR becslésben részt vettek, valamint a GRNN és az RBFNN modellnél alkalmazott szigma-faktort, végül zárójelben az utóbbi neurális hálózati modell esetén a rejtett rétegében lévő neuronok számát. A 4.7. táblázat, a 4.8. táblázat, a 4.9. táblázat, a 4.10. táblázat és a 4.11. táblázat jellemzők sorában is ugyanezen kiszámított jellemzők szerepelnek.

4.13. táblázat A dunai konfigurációk eredményei a teszthalmazra vonatkozóan

Beállítások	Tanítóhalmaz + Teszthalmaz	Modell	MLR	GRNN	RBFNN
D_R	2/3 + 1/3	RMSE [mg L^{-1}]	1,39	1,22	1,27
		MAE [mg L^{-1}]	1,08	0,88	0,94
		R^2	0,33	0,50	0,45
		IA	0,70	0,83	0,80
		Jellemzők	T_w, pH, EC	0,26	0,48(26)
D_A	D6, D8 + D7	RMSE [mg L^{-1}]	1,27	0,93	1,08
		MAE [mg L^{-1}]	1,01	0,69	0,89
		R^2	0,48	0,69	0,59
		IA	0,75	0,90	0,85
		Jellemzők	T_w, pH	0,26	0,42(32)
D_B	D7, D8 + D6	RMSE [mg L^{-1}]	1,41	0,96	1,01
		MAE [mg L^{-1}]	1,09	0,65	0,80
		R^2	0,37	0,66	0,65
		IA	0,73	0,89	0,87
		Jellemzők	T_w, pH	0,18	0,26(84)
D_C	D6, D7 + D8	RMSE [mg L^{-1}]	1,89	1,79	1,75
		MAE [mg L^{-1}]	1,39	1,33	1,33
		R^2	0,32	0,39	0,41
		IA	0,61	0,66	0,71
		Jellemzők	T_w, pH	0,34	0,15(71)

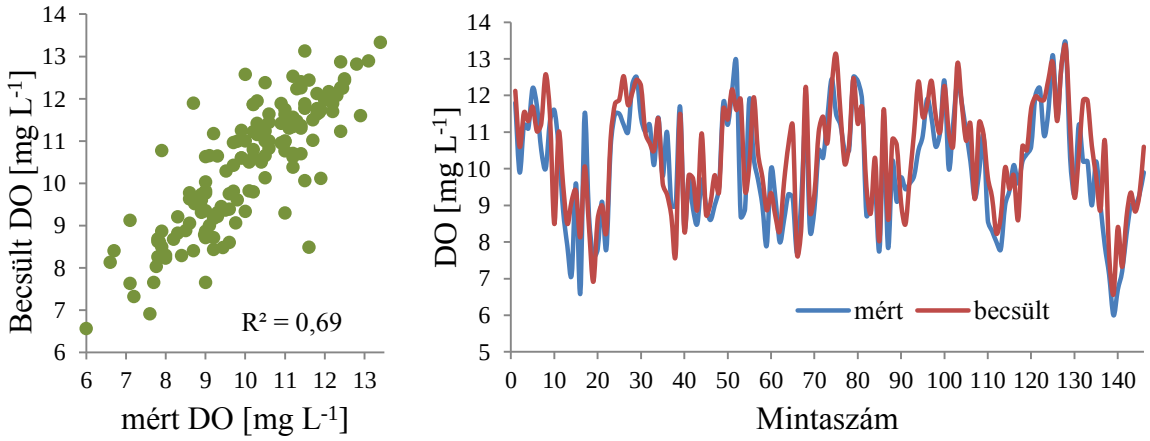
Az eredmények rámutattak arra, hogy a többváltozós lineáris regressziós modellnél jóval pontosabb becslést adtak az neurális hálózatokkal való becslések, melyek közül a GRNN modellel értem el a leghatékonyabb becslést, tehát a GRNN modellel a teszthalmazra kapott RMSE érték (1,22 mg L^{-1}) lesz a referenciaérték.

A második konfigurációban a CCDA módszerrel meghatározott dunai homogén csoportok (3.13. ábra) közül egy kételemű csoportot tudtam alkalmazni arra vonatkozóan, hogy a tanítóhalmaz leghatékonyabb adatstruktúráját meghatározzam, mivel ehhez két homogén és egy folyásirányban utánuk következő, szomszédos inhomogén állomásokra volt szükség. Ez a csoport a D6 (Budapest) és D7 (Nagytétény) állomások által alkotott homogén csoport volt, és a folyásirányban utánuk következő pont D8 (Dunaföldvár), ami már inhomogénnek tekinthető (3.7.3 pont). Így tehát egy beállítást vizsgáltam a D6, D7 és D8 állomások adataival úgy, hogy két állomás adatai alkották a tanítóhalmazt és a harmadik mintavételi pont adatai voltak a teszthalmaz elemei, hogy biztosítva legyen a 2/3-1/3 arány a tanító- és a teszthalmazra vonatkozóan. Vagyis háromféle albeállítást kellett vizsgálnom, első esetben a D6 és D8 (D_A beállítás), a második esetben D7 és D8 (D_B beállítás) mintavételi pontok voltak a tanítóhalmaz elemei, végül pedig a D6 és a D7 állomások (D_C beállítás) adatai kerültek a tanítóhalmazba.

Mindhárom beállításnál megvizsgáltam, hogy melyik modellel kaptam a legkisebb RMSE értéket a teszthalmazra vonatkozóan, vagyis melyik modell volt a leghatékonyabb. Eszerint a D_A beállítás esetén a GRNN modellel kapott 0,93 mg L^{-1} RMSE érték volt a legkisebb a másik két modellel kapott értéknél. A D_B beállítás esetén is a GRNN szolgáltatta a leghatékonyabb becslést (0,96 mg L^{-1}), viszont a D_C esetén az RBFNN modellel kaptam a legkisebb RMSE értéket (1,75 mg L^{-1}) a teszthalmazra vonatkozóan. Tehát mindhárom beállítás esetén a két neurális hálózattal hatékonyabb becslést sikerült elérnem, mint az MLR modellel.

4.4.3. A tanítóhalmaz leghatékonyabb adatstruktúrája dunai adatokon

A második konfiguráció eredményei alapján látható (4.13. táblázat), hogy a három albeállítás közül a legrosszabb eredményt a D_C albeállítás adta, és bár a három modell közül itt az RBFNN modell volt a leghatékonyabb, az általa meghatározott teszthalmazra vonatkozó RMSE érték 88% illetve 82%-kal rosszabb, mint a D_A illetve a D_B albeállítás legjobb RMSE értékei. A 4.3.6 pont szerint a D_A és a D_B albeállításoknál a tanítóhalmaz „kevert struktúrájú", hiszen egy homogén és egy inhomogén állomás adatai alkotják a tanítóhalmazt (Csábrági et al., 2019a, 2019c). A D_A és a D_B albeállításokkal kapott becslések a teszthalmazra vonatkoztatott statisztikai mutatók alapján közel hasonló értékeket mutatnak, de a D_A albeállítás és azon belül a GRNN modell adta a leghatékonyabb becslést (4.15. ábra), ahol Nagytétény állomás adatai voltak a teszthalmaz elemei. Tehát a két „kevert struktúrás" albeállítással kaptam a leghatékonyabb becsléseket a harmadik albeállítás RMSE értékéhez képest 47 és 45 százalékos javulást eredményezve.



a) pontfelhő diagram b) vonaldiagram

4.15. ábra D_A albeállítás mért ill. becsült DO-szintjének diagramjai (Csábrági et al., 2019a)

A D_A és D_B beállításoknál kapott legjobb RMSE értékek a referenciaértékhez viszonyítva 31 és 27 százalékos javulást jelentenek, tehát ennyivel hatékonyabbak a referenciamodellnél, viszont a D_C albeállításnál majdnem 43%-os teljesítménycsökkenés figyelhető meg a referenciamodellhez képest.

Mivel a mintaszám az második konfigurációban az első konfiguráció mintaszámának kevesebb, mint negyede, ezért a D_A és D_B albeállításoknál elért hatékonyságnövekedés számottevő, hiszen a mintaszámot jelentősen csökkentve (térbeli homogenitást fölhasználva) sokkal hatékonyabb becsléseket kaptam. Ugyanakkor megjegyezendő, hogy térbeli optimalizáció esetén nem mindegy, hogy hogyan csoportosítjuk egy folyó homogén és

inhomogén mintavételi pontjait a tanító- és a teszthalmazba, mert esetleg kevésbé hatékony eredmény is születhet.

4.5. Új tudományos eredmények

Kutatómunkám során folyóvizek oldottoxigén-koncentrációjának becslésén keresztül a neurális hálózatok különböző alkalmazásának bemutatásával foglalkoztam. Az eredményeim tézisekbe foglalt következtetéseit az alábbiakban foglalom össze:

1.*Antropogén hatások befolyása a becslésre*

A Duna folyó adataival időbeli előrejelzést valósítottam meg a folyó oldottoxigén-koncentrációjára vonatkozóan MLR, MLPNN, GRNN és RBFNN modellek segítségével. Az így kapott eredményekkel igazoltam, hogy az MLR modell szolgáltatta referencia-becsléshez képest a zavartalan mintavételi ponton (Mohácson) nagyobb (30%-os) javulás érhető el hatékonyság terén neurális hálózatokkal, míg a másik két zavart mintavételi pont (Fajsz és Győrzámoly) esetén csekélyebb (13-18%) javulás volt megfigyelhető. A modellek hatékonyságát a teszthalmazra kapott RMSE értékekkel mértem. Megállapítottam, hogy az antropogén hatások által zavart mintavételi pontok adataival nehezebb a lineáris modellhez képest hatékonyabb becslést adni a neurális hálózatokkal.

2.*Neurális hálózatok tanítóhalmazának leghatékonyabb adatstruktúrája*

A Tisza és a Duna folyók adataival térbeli előrejelzést, illetve térbeli optimalizációt valósítottam meg a folyók oldottoxigén-koncentrációjára vonatkozóan MLR, GRNN és RBFNN modellek segítségével úgy, hogy a folyók homogén csoportjait figyelembe véve adtam meg a tanító és a teszthalmazokat, ez a Tisza esetében három, a Duna esetében pedig egy példát jelentett. Az így kapott eredményekkel igazoltam, hogy akkor hatékonyabb a becslés a teszthalmazra vonatkozó RMSE érték alapján, ha a tanítóhalmaz „kevert struktúrájú", vagyis egyszerre tartalmazza a kételemű homogén csoport egyik mintavételi pontját és az inhomogén mintavételi pontot. Mind a négy példában az inhomogén állomás a kételemű homogén csoport folyásirányban szomszédos mintavételi pontja volt.

3.*Tisza folyó térbeli szakaszainak becsülhetősége*

A Tisza folyó adataival az oldottoxigén-koncentrációjára való térbeli előrejelzést és optimalizációt megvalósítva igazoltam, hogy a Tisza folyó magyarországi szakaszai különbözőképpen becsülhetőek. A folyó felső szakaszára nagyobb hibákkal terhelt becslések születtek. Mindez rámutatott arra, hogy adott esetben a vizsgált folyó szakaszjellegét (felső-, közép-, alsó-) is figyelembe kell venni a becslések során. A változékonyabb, nagyobb áramlású felső folyószakaszon nehezebb becsülni (pl.: Balsa, Záhony, Tiszalök-duzzasztó felett), mint az alsóbb szakaszokon (pl. Mindszent, Tápé, Tiszasziget).

4. *A leghatékonyabb modell kiválasztása oldottoxigén-koncentráció becslésére*

A vizsgálataim összes eredményével igazoltam, hogy a neurális hálózatok – főleg a GRNN és az RBFNN – hatékonyabb modellek a teszthalmazra vonatkozó RMSE értékek alapján folyóvizek oldottoxigén-koncentrációjának becslésére a többváltozós lineáris regressziónál. A neurális hálózatok és a többváltozós lineáris modellek közötti teljesítménykülönbség a Duna vizsgálatainak esetében volt nagyobb mértékű.

5. KÖVETKEZTETÉSEK ÉS JAVASLATOK

A dunai mintavételi pontok adatainak külön-külön való vizsgálatainak eredményeiből azt a tapasztalatot vontam le, hogy nagyon fontos megvizsgálni azt, milyen hatások érik a folyót az adott mintavételi pontban. A dunai vizsgálat időbeli előrejelzést megvalósító eredményeiből úgy tűnik, hogy ha nem zavartalan állomás adataival végezzük a vizsgálatot, tehát ha antropogén hatások érik a folyót (pld. erőművek, duzzasztók), akkor nehezebb pontosabb becslést adni nemcsak a lineáris, hanem a neurális hálózatokkal is. Ellenben, ha csak zavartalan mintavételi pont adataival folyik a modellezés, akkor valószínűleg pontosabb, megbízhatóbb becslések születnek.

A folyóvizek oldottoxigén-koncentrációjának becslésével foglalkozó tudományos közlések áttanulmányozása során arra a következtetésre jutottam, hogy nagyon kevés olyan tudományos kutatás, munka jelent meg, ahol a folyók mintavételi pontjait figyelembe véve osztották föl az adathalmazt tanító és teszthalmazra, vagyis térbeli előrejelzés valósult meg. Kutatómunkámban a tiszai és a dunai vizsgálat során is adok példát térbeli előrejelzésre.

A tiszai vizsgálatnál az összes mintavételi pont adataival modelleztem egyszerre két konfigurációban is, majd arra a következtetésre jutottam, hogy érdemes a folyó mintavételi pontjainak a struktúráját is megvizsgálni, hiszen ha van lehetőség közel azonos homogenitású, azonos struktúrájú mintával dolgozni, akkor a modellekkel pontosabb becslések nyerhetőek. Vagyis ha optimalizálom a bemenő adathalmazt, tehát térbeli optimalizációt végzek, akkor a tanítóhalmaz mérete kevesebb megfigyelésből, mintából fog állni, mégis hatékonyabb becsléseket kapok, ellentétben azzal a két konfigurációval, ahol a teljes tiszai mintahalmazzal történt a modellezés. A szakirodalom szerint viszont a megfigyelések számának növekedésével nő a becslés pontossága is (statisztikai konzisztencia) feltételezve persze, hogy a populáció homogén. A kapott eredmények rámutattak arra, hogy a csökkenő mintaszám ellenére a tanító- és a teszthalmaz tudatos/irányított kijelölésével, illetve a mintavételi pontok struktúrájának vizsgálatával jelentősen javítható a modellek hatékonysága, ami nem mond ellen a statisztikai konzisztenciának, hiszen a Tisza magyarországi szakasza nem tekinthető azonos struktúrájúnak.

A leghatékonyabb becslést a Tisza folyó alsó szakaszára kaptam a térbeli optimalizációval és a tanítóhalmazban „kevert struktúrát” alkalmazva, vagyis a két homogén állomás közül az egyik és a folyásirányban szomszédos inhomogén állomás alkották a tanítóhalmaz elemeit. A tanítóhalmaz leghatékonyabb adatstruktúrára vonatkozó megállapítást a Tiszán három a Dunán pedig egy példán mutattam be, és mind a négy példában a kételemű homogén csoporton kívül a folyásirányban szomszédos állomás volt az inhomogén mintavételi pont. További vizsgálataim során indokolt lehet megvizsgálni, hogy a kevert struktúra alkalmazása akkor is hatékonyabb, ha az inhomogén állomás nem szomszédos mintavételi pontja a két-, vagy többelemű homogén csoportnak.

Mind a dunai mind a tiszai vizsgálatok során kapott eredmények bebizonyították, hogy a neurális hálózatok, azon belül is a GRNN és RBFNN modellek hatékony eszközök folyóvizek oldottoxigén-koncentrációjának becslésére.

A kutatómunka további folytatása során célszerű lehet további neurális hálózatokkal, mint például tartó vektor géppel (SVM), illetve hibrid modellekkel is például adaptív neuro-fuzzy következtető rendszerrel (ANFIS) becsülni folyóvizek oldottoxigén-koncentrációját.

6. ÖSSZEFOGLALÁS

FOLYÓVIZEK OLDOTTOXIGÉN-KONCENTRÁCIÓJÁNAK BECSLÉSE NEURÁLIS HÁLÓZATOKKAL

Az emberiség utóbbi másfél évszázadának ipari tevékenysége miatt jelentős változás ment végbe a természetes vizek vízminőségében, ezért a felszíni vizek minőségét leíró paramétereinek fokozatos ellenőrzése egyre fontosabb kérdés. Az oldottoxigén-koncentrációja az egyik legmeghatározóbb mutatója a felszíni vizek szennyeződésének, így ennek a paraméternek a becslése nagyon fontos. Kutatómunkám célja bemutatni a neurális modellek minél hatékonyabb alkalmazásának lehetőségeit ezen paraméter becslésére két nagy folyóvizünk, a Duna és a Tisza adatait fölhasználva.

A kutatási témakörhöz kapcsolódó szakirodalom tanulmányozása során bemutattam a neurális hálózatok előnyeit, hátrányait, különböző fajtáit, illetve meghatároztam azokat a könnyen mérhető paramétereket, melyek legjobban befolyásolják a folyóvizek oldottoxigén-koncentrációját. Az MLR modell alkalmazásakor csak a szignifikáns, független paraméterek segítségével becsültem a vizsgált folyóvizek oldottoxigén-koncentrációját. Az MLPNN modell esetében kidolgoztam egy iterációs módszert, amellyel hatékonyabbá és megbízhatóbbá lehet tenni a modellt.

A dunai állomásokból három mintavételi pontot választottam ki aszerint, hogy ezeknél az állomásoknál a folyót milyenfajta hatások érik, így Mohács, mint zavartalan, antropogén hatásoktól mentes, referenciaállomás, Győrzámoly (Bős-nagymarosi vízlépcső) és Fajsz (MVM Paksi Atomerőmű), mint két ún. „zavart” állomás jelenik meg. A három állomást külön-külön illetve mind a hármat egyben is vizsgáltam időbeli előrejelzést megvalósítva négyféle modellel (MLPNN, GRNN, RBFNN és MLR). Jelentős javulást, 30 százalékot Mohács adataival értem el a GRNN modellel a lineáris modellel szemben, a másik két zavart mintavételi pontnál nehezebb hatékonyabb becslést adni a lineáris modellhez képest.

A tiszai állomás összes mintavételi pontjának adatait háromféle módszerrel osztottam szét tanító- illetve teszthalmazra, így az első konfigurációban egy referenciamodellt valósítottam meg véletlen kiválasztással, a másodikban térbeli előrejelzést adtam mintavételi pontok térbeli elhelyezkedése szerinti kiválasztással. A harmadik konfigurációban pedig térbeli optimalizációt valósítottam meg úgy, hogy figyelembe vettem a folyó homogén csoportjait, és három-három állomás adatait vizsgáltam egyszerre. Mindhárom konfigurációban kb. 2:1 arányban osztottam szét az adatokat a tanító és a teszthalmaz között az összehasonlítás végett, és három modellt (GRNN, RBFNN, MLR) alkalmaztam. A kapott eredményekből megállapítható, hogy a folyó magyarországi fölső szakasza nehezebben becsülhető, nagyobb hibákkal terhelt becslések születtek. Ahogy a folyásirányban egyre lejjebb haladunk úgy növekszik a modellek teljesítménye. A tiszai és a dunai térbeli optimalizációt megvalósító konfigurációkban kapott eredményekkel igazoltam, hogy ha a tanítóhalmaz „kevert struktúrájú” volt, vagyis tartalmazta az egyik homogén állomás és az inhomogén állomás adatait, akkor hatékonyabb becslés született. Térbeli optimalizációval 65%-kos javulást értem el a Tisza alsó szakaszán a referenciamodellhez képest úgy, hogy a tanítóhalmaz „kevert struktúrájú” volt. Az összes vizsgálat eredményeit megvizsgálva megállapítottam, hogy a neurális hálózatok, különösen a GRNN és az RBFNN modellek hatékonyabb eszközök folyóvizek oldottoxigén-koncentrációjának becslésére a lineáris modellel szemben.

7. SUMMARY

APPLYING ARTIFICIAL NEURAL NETWORKS FOR THE PREDICTION OF DISSOLVED OXYGEN CONCENTRATION IN DOMESTIC RIVERS

Industrial activity of the last one and a half century, increased utilisation of artificial fertilisers in the agriculture and urbanisation all contributed to the deterioration in surface water quality and reserves. Therefore the improved knowledge and monitoring of parameters describing surface water quality is becoming increasingly important. Among these parameters dissolved oxygen is one of the most important indicator of surface water pollution, therefore its estimation is highly important. The aim of my research is to present the most efficient way of applying different neural networks to estimate this parameter in the Hungarian section of two large rivers, the rivers Danube and Tisza.

Following the analysis of relevant literature sources I have assessed the typology, the advantages and disadvantages of neural network models. I have also identified those easily measurable parameters which have the largest influence on dissolved oxygen content of rivers. Throughout the application of the MLR model only the significant independent parameters were used to estimate the dissolved oxygen content of the analysed rivers. For the MLPNN model I have developed an iteration method which made the model more efficient and robust.

From the sampling stations on the Danube I have selected three sampling locations according to the types of impacts on the river. The Mohács location served as an undisturbed reference station, Győrzámoly (Bős-Nagymaros cascaded hydro) and Fajsz (MVM Paksi Atomerőmű), appeared as two „perturbed" stations. I have analysed the three stations separately and together using four different models (MLPNN, GRNN, RBFNN and MLR) implementing a temporal forecast. Using the GRNN model and data from Mohács I have achieved a significant, approx. 30% improvement over the MLR model. I have found that it would be more difficult to gain better results for the two other (perturbed) stations.

Using data from all sampling stations on Tisza I have applied three different methods to allocate data into training and test sets. In the first configuration I implemented a reference model with random allocation, in the second configuration I implemented spatial forecasting by selecting sampling points according to their geographical location. In the third configuration I implemented spatial optimisation by considering homogenous groups. In all three configurations I have distributed data in an approximately 2:1 ratio between the test and training sets and I applied three models (GRNN, RBFNN and MLR) in each case for comparative analysis. The results prove that the upper domestic section of the river is more difficult to estimate, larger error rates characterised the results. Moving downwards the river the performance of the models improves. This means that estimating parameters on the lower section of the River Tisza is easier. Also, spatial optimisation configurations of Tisza and Danube showed better performance with a „mixed structure" training set composed of one of the homogenous stations and the inhomogeneous station. The improvement on the lower section of the River Tisza was 65% over the reference model if the training set was „mixed". In all cases neural network models (specially GRNN and RBFNN) turned out to be more efficient means of forecasting dissolved oxygen content than the MLR model.

8. MELLÉKLETEK

M1: Irodalomjegyzék

1. Abyaneh, H.Z. (2014): Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters. Journal of Environmental Health Science & Engineering, Vol. 12(40). DOI:10.1186/2052-336X-12-40
2. Adeniran, K.A., Adelodun, B., Ogunshina, M. (2016): Artificial Neural Network Modelling of Biochemical Oxygen Demand and Dissolved Oxygen of Rivers: Case Study of Asa River. Journal of Environmental Research, Engineering and Management, Vol. 72, pp. 59-74.
3. Ahmed, A.A.M. (2014): Prediction of dissolved oxygen in Surma River by biochemical oxygen demand and chemical oxygen demand using the artificial neural networks (ANNs). Journal of King Saud University Engineering Sciences. DOI:10.1016/j.jksues.2014.05.001
4. Akkoyunlu, A., Altun, H., Cigizoglu, H.K. (2011): Depth-integrated estimation of dissolved oxygen in a lake. Journal of Environmental Engineering, Vol. 137, pp. 961-967.
5. Altrichter, M., Horváth, G., Pataki, B., Strausz, Gy., Takács, G., Valyon, J. (2006): Neurális hálózatok. Panem Kiadó, Budapest.
6. Anderberg, M.R. (1973): Cluster Analysis for Applications. Academic Press, New York.
7. Antanasijević, D., Pocajt, V., Povrenović, D., Perić-Grujić, A., Rictić, M. (2013): Modelling of dissolved oxygen content using artificial neural networks: Danube River, North Serbia, case study. Environmental Science and Pollution Research, Vol. 20, pp. 9006-9013.
8. Antanasijević, D., Pocajt, V., Perić-Grujić, A., Rictić, M. (2014): Modelling of dissolved oxygen in the Danube River using artificial neural networks and Monte Carlo Simulation uncertainty analysis. Journal of Hydrology, Vol. 519, pp. 1895-1907.
9. APHA (1998): Standard methods for the examination of water and wastewater, 20th ed., American Public Health Association, Washington, DC.
10. Areerachakul, S., Junsawang, P., Pomsathit, A. (2011): Prediction of dissolved oxygen using artificial neural network. International Conference on Computer Communication and Management, Vol. 5, pp. 524-528.
11. Ay, M., Kisi, O. (2012): Modeling of dissolved oxygen concentration using different neural network techniques in Foundation Creek, El Paso County, Colorado. Journal of Environmental Engineering, Vol. 138, pp. 654-662.
12. Basant, N., Gupta, S., Malik, A., Singh, K.P. (2010): Linear and nonlinear modelling for simultaneous prediction of dissolved oxygen and biochemical oxygen demand of the surface water – A case study. Chemometrics and Intelligent Laboratory System, Vol. 104, pp. 172-180.
13. Bayram, A., Kankal, M. (2015): Artificial neural network modeling of dissolved oxygen concentrations in a Turkish Watershed. Polish Journal of Environmental Studies, Vol. 24, pp. 1507-1515.

14. Bayram, A., Uzlu, E., Kankal, M., Dede, T. (2015): Modeling stream dissolved oxygen concentration using teaching-learning based optimization algorithm. Environmental Earth Sciences, Vol. 73, pp. 6565-6576.

15. Boano, F., Revelli, R., Ridolfi, R. (2006): Stochastic modeling of DO and BOD components in a stream with random inputs. Advances in Water Resources, 29, pp. 1341-1350.

16. Broomhead D., Lowe D. (1988): Multivariable functional interpolation and adaptive networks. Complex Systems, Vol. 2, pp. 321-355.

17. Burchard-Levine, A., Liu, S., Vince, F., Li, M., Ostfeld, A. (2014): A hybrid evolutionary data driven model for river water quality early warning. Journal of Environmental Management, Vol. 143, pp. 8-16.

18. Chapman, D.V., Bradley, C., Gettel, G.M., Hatvani, I.G., Hein, T., Kovács, J., Liska, I., Oliver, D.M., Tanos, P., Trásy, B., Várbíró, G. (2016): Developments in water quality monitoring and management in large river catchments using the Danube River as an example. Environmental Science & Policy, Vol. 64, pp. 141-154. DOI:10.1016/j.envsci.2016.06.015

19. Chen, W.B., Liu, W.C. (2014): Artificial neural network modeling of dissolved oxygen in reservoir. Environmental Monitoring Assessment, Vol. 186, pp. 1203-1217.

20. Chen, W.B., Liu, W.C. (2015): Water quality modeling in reservoirs using multivariate linear regression and two neural network models. Hindawi Publishing Corporation. DOI:10.1155/2015/521721

21. Chilundo, M., Kelderman, P., O'keeffe, J. (2008): Design of a water quality monitoring network for the Limpopo river Basin in Mozambique. Physics and Chemistry of the Earth, Parts A/B/C 33(8-13), pp. 655-665.

22. Clair, T.A., Ehrman, J.M. (1996): Variation in discharge and dissolved organic carbon and nitrogen export from terrestrial basins with changes in climate: a neural network approach. American Society of Limnology and Oceanography. Vol. 41(5), pp. 921-927.

23. Cox, B.A. (2003): A review of currently available in-stream water-quality models and their applicability for simulating dissolved oxygen in lowland rivers. The Science of the Total Environment, Vol. 314-316, pp. 335-377.

24. Danish Hydraulics Institute (DHI) (2001): MIKE11, User Guide & Reference Manual. Danish Hydraulics Institute, Horsholm, Denmark

25. Day, W.H.E., Edelsbrunner, H. (1984): Efficient algorithms for agglomerative hierarchical clustering methods. J. Classif, Vol. 1, pp. 7-24.

26. Demuth, H., Beale, M. (2000): Neural Network Toolbox user's guide: MATLAB. The Mathworks Inc.

27. Dogan, E., Sengorur, B., Koklu, R. (2009): Modeling biochemical oxygen demand of the Melen River in Turkey using an artificial neural network technique. Journal of Environmental Management, Vol. 90, pp. 1229–1235.

28. Duda, R.O., Hart, P.E., Stork, D.G. (2000): Pattern Classification. Wiley InterScience.

29. Draper N.R., Smith H. (1981): Applied regression analysis. Wiley, New York

30. Eatherall, A., Boorman, D.B., Williams, R.J., Kowe, R. (1998): Modelling in-stream water quality in LOIS. Science of the Total Environment, Vol. 210, pp. 499–517.

31. Emamgholizadeh, S., Kashi, H., Marofpoor, I., Zalaghi, E. (2014): Prediction of water quality parameters of Karoon River (Iran) by artificial intelligence-based models. Int. Journal of Environmental Science and Technology, Vol. 11, pp. 645-656.

32. Faruk, D.Ö. (2010): A hybrid neural network and ARIMA model for water quality time series prediction. Engineering Applications of Artificial Intelligence, Vol. 23, pp. 586-594.

33. Fazekas I. (2013): Neurális hálózatok. Debreceni Egyetem, egyetemi jegyzet

34. Fletcher, D., Goss, E. (1993): Forecasting with neural network: An application using bankruptcy data. Information & Management, Vol. 24, pp. 159-167.

35. French, M., Recknagel, F. (1994): Modeling algal blooms in freshwaters using artificial neural networks. In: Zanetti P (ed) Computer Techniques in Environmental Studies V, Vol. II, Environment Systems. Computational Mechanics Publications, Boston, pp. 87-94.

36. Füst, A., Geiger, J. (2010): Monitoringtervezés és -értékelés geostatisztikai módszerekkel I. Szakértői véleményen alapuló, „igazoló" mintázás geostatisztikai támogatása. Földtani Közlöny, 140(3), 303-312. o.

37. Goldman, C.R., Horne, A.J. (1983): Limnology. McGraw-Hill, New York

38. Gourine, B., Mahi, H., Khoudiri, A., Laksari, Y. (2012): The GRNN and RBF Neural Networks for 2D Displacement Field Modelling. Case study: GPS Auscultation Network of LNG reservoir (GL4/Z industrial complex – Arzew, Algeria). International Federation of Surveyors (FIG) working week 2012, Italy, Rome.

39. Haider, H., Ali, W. (2010): Development of dissolved oxygen model for a highly variable flow river: a case study of Ravi River in Pakistan. Environmental Monitoring and Assessment, Vol. 15, pp. 583–599. doi:10.1007/s10666-010-9240-4

40. Hanna, AM., Ural, D., Saygili, G. (2007): Neural network model liquefaction potential in soil deposits using Turkey and Taiwan earthquake data. Soil Dynamics and Earthquake Engineering, Vol. 27, pp. 521-540.

41. Hannan, S.A., Manza, R.R., Ramteke, R.J. (2010): Generalized regression neural network and radial basis function for heart disease diagnosis. International Journal of Computer Applications, Vol. 7(13), pp. 7-13.

42. Hastie, T., Tibshirani, R., Friedman, J. (2009): The elements of statistical learning: Data mining, inference, and prediction, 2nd ed., Springer

43. Hatvani, I.G., Kovács, J., Márkus, L., Clement, A., Hoffmann, R., Korponai, J. (2015): Assessing the relationship of background factors governing the water quality of an agricultural watershed with changes in catchment property (W-Hungary). Journal of Hydrology, Vol. 521, pp. 460-469.

44. Haykin S. (1999): Neural Networks: A Comprehensive Foundation. 2nd ed. Prentice-Hall. Upper Saddle River, New Jersey

45. He, B., Oki, T., Sun, F., Komori, D., Kanae, S., Wang, Y., Kim, H., Yamazaki, D. (2011a): Estimating monthly total nitrogen concentration in streams by using artificial neural network. Journal of Environmental Management, Vol. 92, pp. 172-177.

46. He, J., Chu, A., Ryan, M., Valeo, C., Zaitlin, B. (2011b): Abiotic influences on dissolved oxygen in a riverine environment. Ecological Engineering, Vol. 37, pp. 1804-1814.

47. Heddam, S., Lamda, H., Filali, S. (2016): Predicting effluent biochemical oxygen demand in a wastewater treatment plant using generalized regression neural network based approach: A comparative study. Environmental Processes, Vol. 3, pp. 153-165.

48. Heddam, S. (2014a): Generalized regression neural network-based approach for modelling hourly dissolved oxygen concentration in the Upper Klamath River, Oregon, USA. Environmental Technology, Vol. 35, pp. 1650-1657.

49. Heddam, S. (2014b): Modelling hourly dissolved oxygen concentration (DO) using dynamic evolving neural-fuzzy inference system (DENFIS)-based approach: case study of Klamath River at Miller Island Boat Ramp, OR, USA. Environmental Science and Pollution Research. 21, 9212-9227.

50. Hornik, K., Stinchcombe M., White H, (1989): Multilayer feedforward networks are universal approximators. Neural Networks, Vol. 2, pp. 359-366.

51. Huang, J., Yin, H., Chapra, S.C., Zhou, Q. (2017): Modelling dissolved oxygen depression in an Urban River in China. Water, Vol. 9:520. DOI: 103390/w9070520

52. Jang, J.S.R., Sun, C.T., Mizutani, E. (1997): Neuro-Fuzzy and soft computing. A computational approach to learning and machine intelligence. Upper Saddle River, NJ: Prentice Hall.

53. Ji, X., Shang, X., Dahlgren, R.A., Zhang, M. (2017): Prediction of dissolved oxygen concentration in hypoxic river systems using support vector machine: a case study of Wen-Rui Tang River, China. Environmental Science and Pollution Research International, Vol. 24(19), pp. 16062-16076.

54. Kanda, E.K., Kipkorir, E.C., Kosgei, J.R. (2016): Dissolved oxygen modeling using artificial neural network: A case of river Nzoia, Lake Victoria Basin, Kenya. Journal of Water Security, Vol. 2. pp. 1-7. DOI: 10.15544/jws.2016.004

55. Kanda, E.K.; Kosgei, J. R.; Kipkorir, E.C. (2015): Simulation of organic carbon loading using MIKE 11 model: a case of River Nzoia, Kenya. Water Practice and Technology, Vol. 10(2), pp. 298–304. doi:10.2166/wpt.2015.035

56. Karul, C., Soyupak, S., Cilesiz, A.F., Akbay, N., Germen, E. (2000): Case studies on the use of neural networks in eutrophication modeling. Ecological Modelling, Vol. 134, pp. 145-152.

57. Kentel, E., Alp, E. (2013): Hydropower in Turkey: Economical, social and environmental aspects and legal challenges. Environmental Science & Policy, Vol. 31, pp. 34-43.

58. Keshtegar, B., Heddam. S. (2017): Modeling daily dissolved oxygen concentration using modified response surface method and artificial neural network: a comparative study. Neural Computing and Applications. DOI: 10.1007/s00521-017-2917-8

59. Khalil, B.M., Awadallah, A.G. Karaman, H., El-Sayed, A., (2012): Application of artificial neural networks for the prediction of water quality variables in the Nile Delta. Journal of Water Resource and Protection, Vol. 4, pp. 388-394.

60. Kim S., Kim H.S. (2008): Neural networks and genetic algorithm approach for nonlinear evaporation and evapotranspiration modelling. Journal of Hydrology, Vol. 351, pp. 299-317.

61. Kiotói jegyzőkönyv (1997): Kiotói jegyzőkönyv az egyesült nemzetek éghajlatváltozási keretegyezményéhez. http://zbr.kormany.hu/download/8/72/00000/Kiot%C3%B3i%20Jegyz%C5%91k%C3%B6nyv%20HUN.pdf (letöltve: 2019.03.02)

62. Klaver, G., van Os, B., Negrel, P., Petelet-Giraud, E. (2007): Influence of hydropower dams on the composition of the suspended and riverbank sediments in the Danube. Environmental Pollution, Vol. 148, pp. 718-728.

63. Kohonen, T. (1995): Self-Organizing Maps. Springer, Berlin.

64. Kovacs J., Kovacs S., Magyar N., Tanos P., Hatvani I. G., Anda A. (2014): Classification into homogeneous groups using combined cluster and discriminant analysis. Environmental Modelling and Software, Vol. 57, pp. 52-59. DOI:10.1016/j.envsoft.2014.01.010

65. Kovács, J. (2015): Habilitációs dolgozat. Eötvös Loránd Tudományegyetem

66. Kovács, J., Kovács, S., Hatvani IG., Magyar, N., Tanos, P., Korponai, J., Blaschke, A. (2015a): Spatial optimization of monitoring networks on the examples of a river, a lake-wetland system and a sub-surface water system. Water Resources Management, Vol. 29(14), pp. 5275-5294. DOI:10.1007/s11269-015-1117-5

67. Kovács, J., Márkus, L., Szalai, J., Kovács, I.SZ. (2015b): Detection and evaluation of changes induced by the diversion of River Danube in the territorial appearance of latent effects governing shallow-groundwater fluctuations. Journal of Hydrology, Vol. 520, pp. 314-325.

68. Kovács, J., Tanos, P., Várbíró, G., Anda, A., Molnár, S., Hatvani, I.G. (2017): The role of annual periodic behavior of water quality parameters in primary production – Chlorophyll-a estimation. Ecological Indicators, Vol. 78, pp. 311-321. DOI:10.1016/j.ecolind.2017.03.002

69. Kroll, C.N., Song, P. (2013): Impact of multicollinearity on small sample hydrologic regression models. Water Resources Research, Vol. 49, pp. 3756-3769.

70. Kuo, J., Hsieh, M., Lung, W., She, N. (2007): Using artificial neural network for reservoir eutrophication prediction. Ecological Modelling, Vol. 200, pp. 171-177.

71. Lechner, A., Keckeis, H., Lumesberger-Loisl, F., Zens, B., Krusch, R., Tritthart, M., Glas, M., Schludermann, E. (2014): The Danube so colourful: A potpourri of plastic litter outnumbers fish larvae in Europe's second largest river. Environmental Pollution, Vol. 188, pp. 177-181.

72. Legates, D.R., McCabe, G.J., Jr. (1999): Evaluating the use of goodness-of-fit measures in hydrologic and hydroclimatic model validation. Water Resources Research, Vol. 35, pp. 233-241.

73. Lewis, M.E. (2006): Dissolved oxygen. Version 2.0, Chapter A6, Section 6.2., In: National Field Manual for the Collection of Water-Quality data. Book 9, US Geological Survey.

74. Liang, C., Xin, S., Dongsheng, W., Xiujing, Y., Guodong, J. (2016): The ecological benefit-loss evaluation in a riverine wetland for hydropower projects – A case study of Xiaolangdi reservoir in the Yellow river, China. Ecological Engineering, Vol. 96, pp. 34-44.

75. Liška, I., Wagner, F., Sengl, M., Deutsch, K., Slobodník, J. (2015): Joint Danube Survey 3, International Commission for the Protection of the Danube River, Vienna. (ISBN: 978-3-200-03795-3)

76. Maier, H.R., Dandy, G.C. (2000): Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environmental Modelling and Software, Vol. 15, pp. 101-124.

77. Mander, Ü., Forsberg, C. (2000): Nonpoint pollution in agricultural watersheds of endangered coastal seas. Ecological Engineering, Vol. 14, pp. 317-324.

78. Marquardt D. (1963): An algorithm for least square estimation of nonlinear parameters. Journal of the Society for Industrial and Applied Mathematics, Vol. 11, pp. 431-441.

79. McLachlan, G. (2004): Discriminant Analysis and Statistical Pattern Recognition.Wiley-InterScience.

80. Moreira, J.R., Poole, A.D. (1993): Hydropower and its constraints. United States.

81. Najah, A., El-Shafie, A., Karim, O.A., Jaafar, O., El-Shafie, A.H. (2011): An application of different artificial intelligences techniques for water quality prediction. International Journal of the Physical Sciences, Vol. 6(22), pp. 5298-5308. DOI: 10.5897/IJPS11.1180

82. Najah, A., El-Shafie, A., Karim, O.A., Jaafar, O., El-Shafie, A.H. (2014): Performance of ANFIS versus MLP-NN dissolved oxygen prediction models in water quality monitoring. Environmental Science and Pollution Research. Vol. 21(3), pp. 1658-1670.

83. Nguyen, D., Widrow, B. (1990): Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights. Proceedings of the International Joint Conference on Neural Networks. 3, Vol. 3, pp. 21 – 26. DOI:10.1109/IJCNN.1990.137819

84. Odum, H.T. (1956): Primary production in flowing waters. Limnology and Oceanography, Vol. 1, pp. 102-117.

85. Onderka, M., Pekárová, P. (2008): Retrieval of suspended particulate matter concentrations in the Danube River from Landsat ETM data. Science of the Total Environment, Vol. 397(1-3), pp. 238–243.

86. Padányi, J., Halász, L. (2012): A klímaváltozás hatásai. Nemzeti Közszolgálati Egyetem.

87. Palani, S., Liong, S., Tkalich, P. (2008): An ANN application for water quality forecasting. Marine Pollution Bulletin, Vol. 56, pp. 1586-1597.

88. Parkhill, K.L., Gulliver, J.S. (1999): Modeling the effect of light on whole-stream respiration. Ecological Modelling, Vol. 117, pp. 333-342.

89. Patil, A., Deng, Z., Malone, R.F. (2012): Temporal scale-induced uncertainty in load duration curves for instream-dissolved oxygen. Environmental Monitoring and Assessment, Vol. 185(2), pp. 1939–1949. DOI:10.1007/s10661-012-2678-x

90. Pavelka, A., Procházka, A. (2004): Algorithms for initialization of neural network weights. Sbornik prispevku 11. in: Konference MATLAB. 2, 453-459.

91. Poggio, T., Girosi, F. (1990): Regularization algorithms for learning that are equivalent to multilayer networks. Science, Vol. 247(4945), pp. 978-982.

92. Ranković, V., Radulović, J., Radojević, I., Ostojić, A., Čomić, L. (2010): Neural network modelling of dissolved oxygen in the Gruza reservoir, Serbia. Ecological Modelling, Vol. 221, pp. 1239–1244.

93. Ranković, V., Radulović, J., Radojević, I., Ostojić, A., Čomić, L. (2012): Prediction of dissolved oxygen in reservoirs using adaptive network-based fuzzy inference system. Journal of Hydroinformatics, Vol. 14, pp. 167-179.

94. Reddy T.A. (2011): Applied data analysis and modeling for energy engineers and scientists. Springer Science & Business Media, New York.

95. Reynolds, C.S. (1984): Phytoplankton periodicity: the interactions of form, function and environmental variability. Freshwater Biology, Vol. 14, pp. 111–142. https://doi.org/10.1111/j.1365-2427.1984.tb00027.x.

96. Rumelhart D.E., Hinton G.E., Williams R.J. (1986): Learning internal representation by error back propagation. in: Rumelhart, D.E. and McClelland, J.L., (Eds.), Parallel distributed processing. MIT Press, Cambridge, pp. 318-362. ISBN: 9780262680530

97. Russell, S., Norvig, P. (2005): Mesterséges intelligencia. Panem Könyvkiadó, Budapest.

98. Sakan, S., Gržetić, I., Đorđević, D. (2007): Distribution and Fractionation of Heavy Metals in the Tisa (Tisza) River Sediments. Environmental Science and Pollution Research, Vol. 14(4), pp. 229-236.

99. Scardi, M. (1996): Artificial neural networks as empirical models for estimating phytoplankton production. Marine Ecology Progress Series, Vol. 139, pp. 289-299.

100. Schurr, J.M., Ruchti, J. (1977): Dynamics of O2 and CO2 exchange, photosynthesis, and respiration in rivers from time-delayed correlation with ideal sunlight. Limnology and Oceanography, Vol. 22 (2), pp. 208-225.

101. Shaikhina T., Khovanova N.A. (2017): Handling limited datasets with neural networks in medical applications: A small-data approach. Artificial Intelligence in Medicine, Vol. 75, pp. 51-63.

102. Šiljić Tomić, A., Antanasijević, D., Ristić, M., Perić-Grujić, A., Pocajt, V. (2016): Modeling the BOD of Danube River in Serbia using spatial, temporal, and input variables optimized artificial neural networks models. Environmental Monitoring and Assessment, Vol. 188(300), pp. 1-12.

103. Šiljić Tomić, A., Antanasijević, D., Ristić, M., Perić-Grujić, A., Pocajt, V. (2018a): A linear and non-linear polynomial neural network modeling of dissolved oxygen content in surface water: Inter- and extrapolation performance with inputs' significance analysis. Science of the Total Environment. Vol. 610-611, pp. 1038-1046.

104. Šiljić Tomić, A., Antanasijević, D., Ristić, M., Perić-Grujić, A., Pocajt, V. (2018b): Application of experimental design for the optimization of artificial neural network-based water quality model: a case study of dissolved oxygen prediction. Environmental Science and Pollution Research, Vol. 25, pp. 1-11. DOI:10.1007/s11356-018-1246-5

105. Singh, K., Basant, A., Malik, A., Jain, G. (2009): Artificial neural network modeling of the river water quality – A case study. Ecological Modelling, Vol. 220, pp. 888-895.

106. Soldán, P., Pavonič, M., Bouček, J., Kokeš, J. (2001): Baia Mare Accident – Brief Ecotoxicological Report of Czech Experts. Ecotoxicology and Environmental Safety, Vol. 49, pp. 255-261. DOI:10.1006/eesa.2001.2070

107. Sommerhäuser, M, Robert, S., Birk, S., Hering, D., Moog, O., Stubauer I., Ofenböck, T. (2003): Final report for he Developing the Typology of Surface Waters and defining the relevant Reference Conditions, UNDP/GEF Danube Regional Project, Vienna. http://www.undp-drp.org/pdf/1.1_River%20Basin%20Management%20-%20Phase%201/1.1_UNDP-DRP_Typology%20of%20SW_116_fr.pdf (megtekintve: 2016.12.01.).

108. Soyupak, S., Karaer, F., Gürbüz, H., Kivrak, E., Sentürk, E., Yazici, A. (2003): A neural network-based approach for calculating dissolved oxygen profiles in reservoirs. Neural Computing and Applications, Vol. 12, pp. 166-172. DOI: 10.1007/s00521-003-0378-8

109. Specht, D. F. (1991) A general regression neural network, IEEE Transactions on Neural Networks, Vol. 2, pp. 568-576.

110. Stanković, I., Várbíró, G., Gligora Udovič, M., Borics, G., Vlahović, T. (2012): Phytoplankton functional and morpho-functional approach in large floodplain rivers. Hydrobiologia, Vol. 698, pp. 217-231. https://doi.org/10.1007/s10750-012-1148-3.

111. Streeter, H.W., Phelps, E.B. (1925): A study of the pollution and natural purification of the Ohio River. III Factors concerned in the phenomena of oxidation and reaeration. US Public Health Service. Public Health Bulletin, Vol. 146, pp. 1-75.

112. Talib, A., Amat, M.I. (2012): Prediction of chemical oxygen demand in Dondang river using artificial neural network. International Journal of Information and Education Technology, Vol. 2, pp. 259-261.

113. Tanos, P. (2017): A Tisza vízrendszerét leíró fizikai, kémiai és biológiai adatsorok vizsgálata többváltozós és idősoros adatelemző módszerekkel. PhD dolgozat. Pannon Egyetem. Festetics DI, Keszthely

114. Tanos, P., Kovács, J., Kovács, S., Anda, A., Hatvani, I.G. (2015): Optimization of the monitoring network on the River Tisza (Central Europe, Hungary) using combined cluster and discriminant analysis, taking seasonality into account. Environmental Monitoring and Assessment, Vol. 187(9), pp. 575. DOI:10.1007/s10661-015-4777-y.

115. Turing, A.M. (1950): Computing machinery and intelligence. Mind, Vol. 49, pp. 433-460.

116. Turnpenny, A.W.H., Coughlan, J., Ng, B., Crews, P., Bamber, R.N., Rowles, P. (2010): Cooling Water Options for the New Generation of Nuclear Power Stations in the UK. Env. Agency, Bristol. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/291077/scho0610bsot-e-e.pdf (megtekintve: 2016.12.03.).

117. US EPA. (2007): An approach for using load duration curves in the development of TMDLs. EPA 841-B-07-006. Washington, DC: Office of Wetlands, Oceans, and Watersheds.

118. Verma, A.K., Singh, T.N. (2013): Prediction of water quality from simple field parameters. Environmental Earth Sciences, Vol. 69, pp. 821-829.

119. Wang, H., Hondzo, M., Xu, C., Poole, V., Spacie, A. (2003): Dissolved oxygen dynamics of streams draining an urbanized and an agricultural catchment. Ecological Modelling, Vol. 160, pp. 145-161.

120. Wang, Q., Li, S., Jia, P., Qi, C., Ding, F. (2013): A review of surface water quality models. The Scientific World Journal. DOI:10.1155/2013/231768.

121. Wasserman P.D. (1993): Advanced methods in neural network. Van Nostrand Reinhold, New York. 147-158.

122. Wen, X., Fang, J., Diao, M., Zhang, C. (2013): Artificial neural network modelling of dissolved oxygen in the Heihe River, Northwestern China. Environmental Monitoring and Assessment, Vol. 185, pp. 4361-4371.

123. Wetzel, R.G. (2001): 9 - Oxygen, in: Wetzel, RG., Limnology, third ed., Academic Press, San Diego, pp. 151-168. ISBN: 9780127447605

124. Whitehead, P.G., Williams, R.J., Lewis, D.R. (1997). Quality simulation along river systems (QUASAR): model theory and development. Sciences of Total Environment, Vol. 94/95. pp. 447-456.

125. Willmott C.J. (1981): On the validation of models. Physical Geography, Vol. 2:2, pp. 184-194. DOI:10.1080/02723646.1981.10642213

M2: Az értekezés témaköréhez kapcsolódó saját publikációk

Lektorált folyóiratcikk világnyelven

1. **Csábrági, A.,** Molnár, S., Tanos, P., Kovács, J. (2017a): Application of artificial neural networks to the forecasting of dissolved oxygen content in the Hungarian section of the river Danube. Ecological Engineering, Vol. 100, pp. 63-72. (ISSN: 0925-8574) (IF: 2,74)
2. Molnár, S., Molnár, M., **Csábrági, A.** (2014a): Progress towards emission targets through the development of climate change policies and measures in Hungary. Időjárás - Quarterly Journal of the Hungarian Meterological Service, Vol. 118 (4), pp. 293-305. (IF: 0,5)
3. **Csábrági, A.,** Molnár, S., Tanos, P., Kovács, J. (2015a): Forecasting of dissolved oxygen in the river Danube using neural networks. Hungarian Agricultural Engineering, Vol. 27, pp. 38-41.
4. Borbás, E., Kovács, J., Hatvani, I.G., **Csábrági, A.,** Molnár, S. (2014): Water chemistry analysis in the sediment of the Baradla cave using geomathematical methods: Aggtelek, NE Hungary. Mechanical Engineering Letters, Vol. 11, pp. 32-43.
5. Molnár, S., Somogyi, F., **Csábrági, A.** (2011c): Comprehensive assessment of future energy needs and the role of alternative energy source. Hungarian Agricultural Engineering, Vol. 23, pp. 117-119.
6. Molnár, S., Molnár, M., **Csábrági, A.** (2011a): Impact assessment of mitigation strategies in the Hungarian agriculture. Journal of Agricultural Informatics, Vol. 2, pp. 10-17. (ISSN 2061-862X)

Lektorált cikk magyar nyelven

7. Molnár, S., **Csábrági, A.** (2010a): Externális költségek vizsgálata az erőművi kibocsátások terén EcoSense modellel. Acta Agraria Kaposváriensis, 14(3), 69-77. o.
8. **Csábrági, A.,** Molnár, S., Tanos, P., Kovács, J. (2019a): Neurális hálózatok alkalmazása ökológiai rendszerek vizsgálatában. Mezőgazdasági Technika, LX. évf. 3. sz., 2-5. o.
9. **Csábrági, A.,** Molnár, S., Tanos, P., Kovács, J., Szabó, I., Molnár, M. (2019b): Neurális hálózatok alkalmazása hazai vízminőségi vizsgálatok során. Mezőgazdasági Technika (megjelenés alatt)

Nemzetközi konferencia kiadvány

10. **Csábrági, A.,** Molnár, S., Tanos, P., Kovács, J. (2015b): Prediction of dissolved oxygen in the River Danube using linear and nonlinear models. Proceedings of the IV. International Conference of the CIGR Hungarian National Committee and the Szent István University Faculty of Mechanical Engineering. Paper NO6-3-055. 6 p. (ISBN 978-963-269-506-8)

11. Molnár, M., Fekete-Farkas, M., **Csábrági, A.** (2012): Assessment of domestic energy consumption and GDP trends. Proceedings of the International Congress Energy and Environment 2012, „Croatian Solar Energy Association", Opatija, Croatia, pp. 247-256. (ISBN 978-953-6886-18-0)

12. Fekete-Farkas, M., Molnár, M., **Csábrági, A.** (2010): Assessment of sectoral greenhouse gas mitigation options and potentials in Hungary. Proceedings of the International Congress Energy and the environment 2010, "Engineering for a low-carbon future", Opatija, Croatia, pp. 477-489. (ISBN 978-953-6886-15-9)

Magyar nyelvű konferencia kiadvány

13. Molnár, M., **Csábrági, A.** (2011a): Hazai energiafogyasztás és GDP kointegráltságának vizsgálata. Proceedings of the Erdei Ferenc VI. Tudományos Konferencia, Kecskemét, Hungary, 44-48. o. (ISBN 978-963-7294-98-3)

14. Molnár, M., **Csábrági, A.** (2010): Erőművi beruházások költségeinek vizsgálata az EcoSense modell segítségével. XII. Nemzetközi Tudományos Napok, Gyöngyös, Hungary, 233-239. o. (ISBN 978-963-9941-09-0)

Nemzetközi konferencia abstract

15. **Csábrági, A.,** Molnár, S., Tanos, P., Kovács, J., Szabó, I., Molnár, M. (2019c): Investigation of data structure in application of neural networks. Abstracts of the 21th Congress of Hungarian Geomathematicians, GeoMATES 2019, Pécs, Hungary, p. 38. (ISBN: 978-963-7068-11-9)

16. **Csábrági, A.,** Molnár, S., Tanos, P., Kovács, J. (2017b): Spatial estimation of dissolved oxygen on the River Tisza using artificial neural network. Abstracts of the 5th International Conference of CIGR Hungarian National Committee and the Szent István University, Faculty of Mechanical Engineering, Gödöllő, Hungary, p. 82. (ISBN: 978-963-269-505-1)

17. **Csábrági, A.,** Molnár, S., Kovács, J. (2013): Application of neural network on the Hungarian section of the Danube to estimate biological oxygen demand. Proceedings of the 3rd International Conference of CIGR Hungarian National Committee and Szent István University, "Engineering, Agriculture, Waste Management and Green Industry Innovation", Gödöllő, Hungary, Paper P09-2-135. (ISBN 978-963-269-371-2)

18. Molnár, S., Molnár, M., **Csábrági, A.** (2011b): Assessment of mitigation options and renewables in the agriculture: a UNFCCC case study. Proceedings of Synergy 2011 – 2nd International Conference in Agricultural Engineering of the CIGR Hungarian National Committee, "Synergy in the Technical Development of Agriculture and Food Industry", Gödöllő, Hungary, p. o5_118.pdf. (ISBN 978-963-269-249-4)

19. Molnár, M., **Csábrági, A.** (2011b): Role of non-conventional energy sources in supplying future energy needs. Bulletin of the Szent István University – Gödöllő, p. 216.

20. Molnár, S., Molnár, M., Somogyi, F., **Csábrági, A.** (2009): Mitigation through advanced technological measures – 4EM-Motor challenge project an EU – IEE Initiative. Proceedings of the International Conferences, in Agricultural Engineering Synergy & Technical Development, Gödöllő, Hungary, CD-ROM Proceedings, p6-266.pdf (ISBN 978-963-269-111-4)

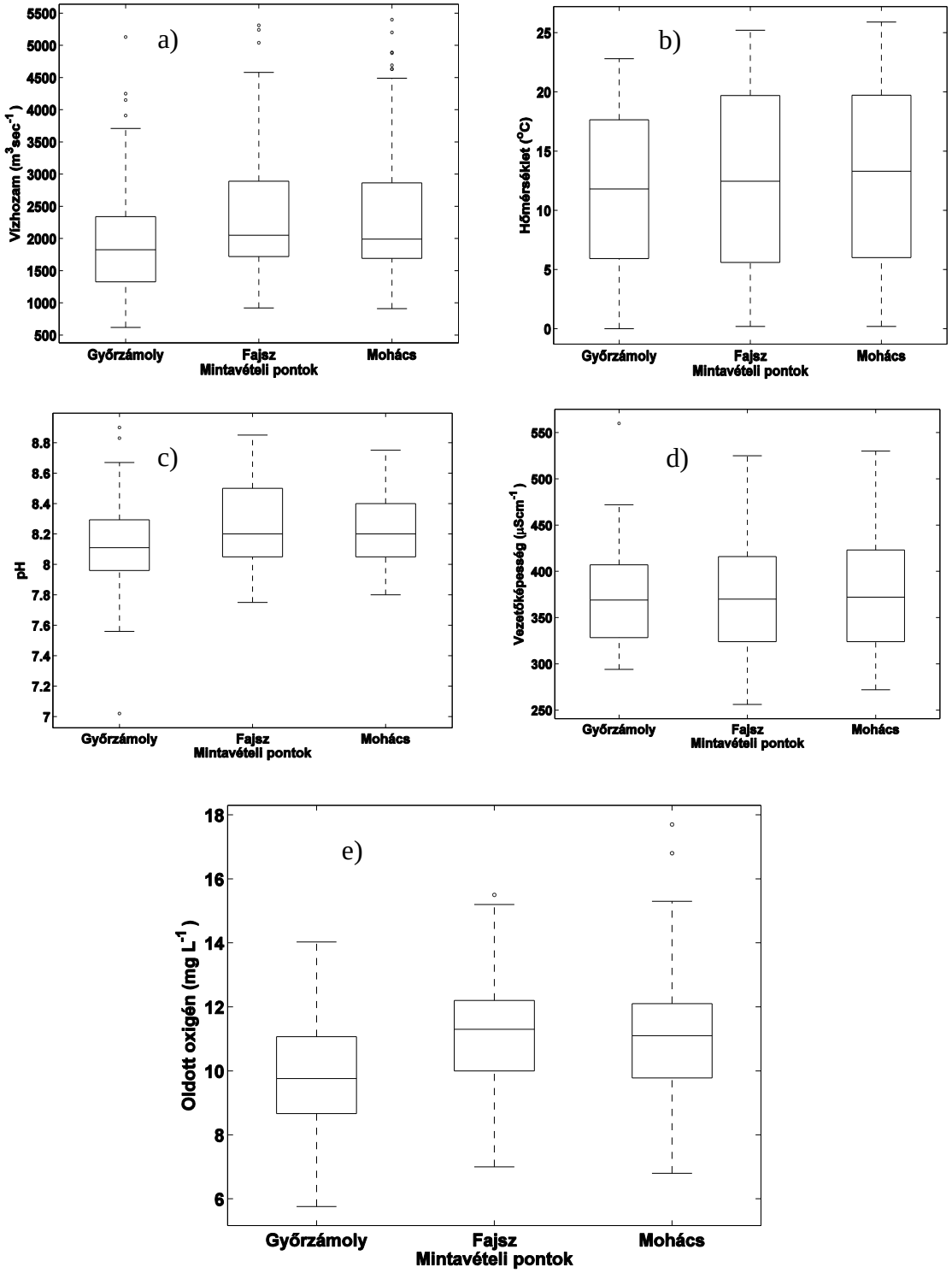
Magyar nyelvű konferencia abstract

21. Molnár, S., **Csábrági, A.** (2010b): Lehetséges hazai erőművi technológiák gazdaságossági vizsgálata a kapacitásbővítési igények tükrében. Proceedings of the 3rd International Symposium on Business Information Systems, Pécs. (ISBN 978-963-642-366-7)

Egyéb

22. Molnár, S., Molnár, M., Hatvani, I.G., Kis-Kovács, G., Bartholy, J., Pongrácz, R., Borka, Gy., Somogyi, Z., Kovacevic, T., Naárné Tóth, Zs., Takács, T., **Csábrági, A.** (2014b): 6th National Communication of Hungary to the UNFCCC, Nemzeti Fejlesztési Minisztérium, pp. 1-358.

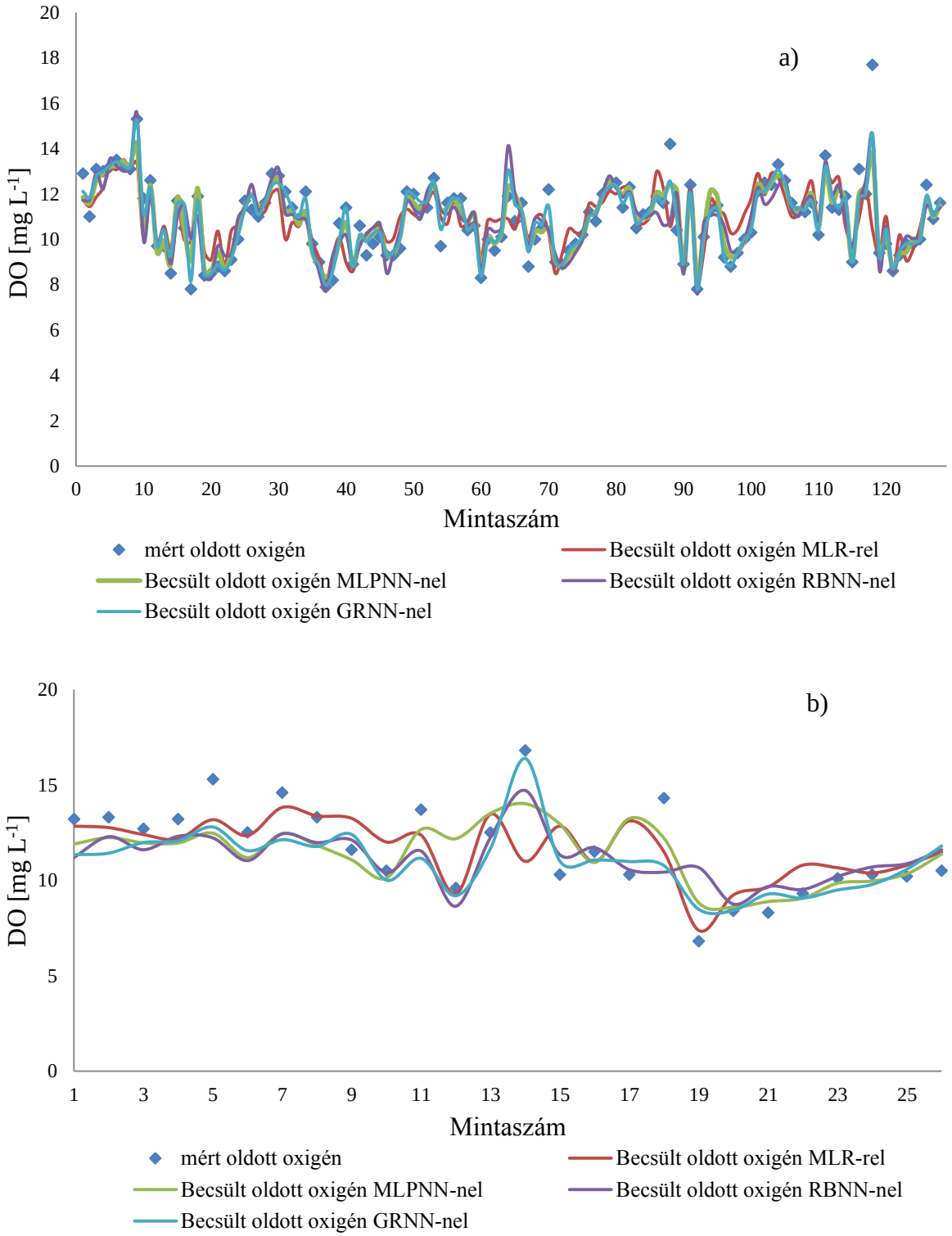
M3: A dunai vizsgálattal kapcsolatos eredmények



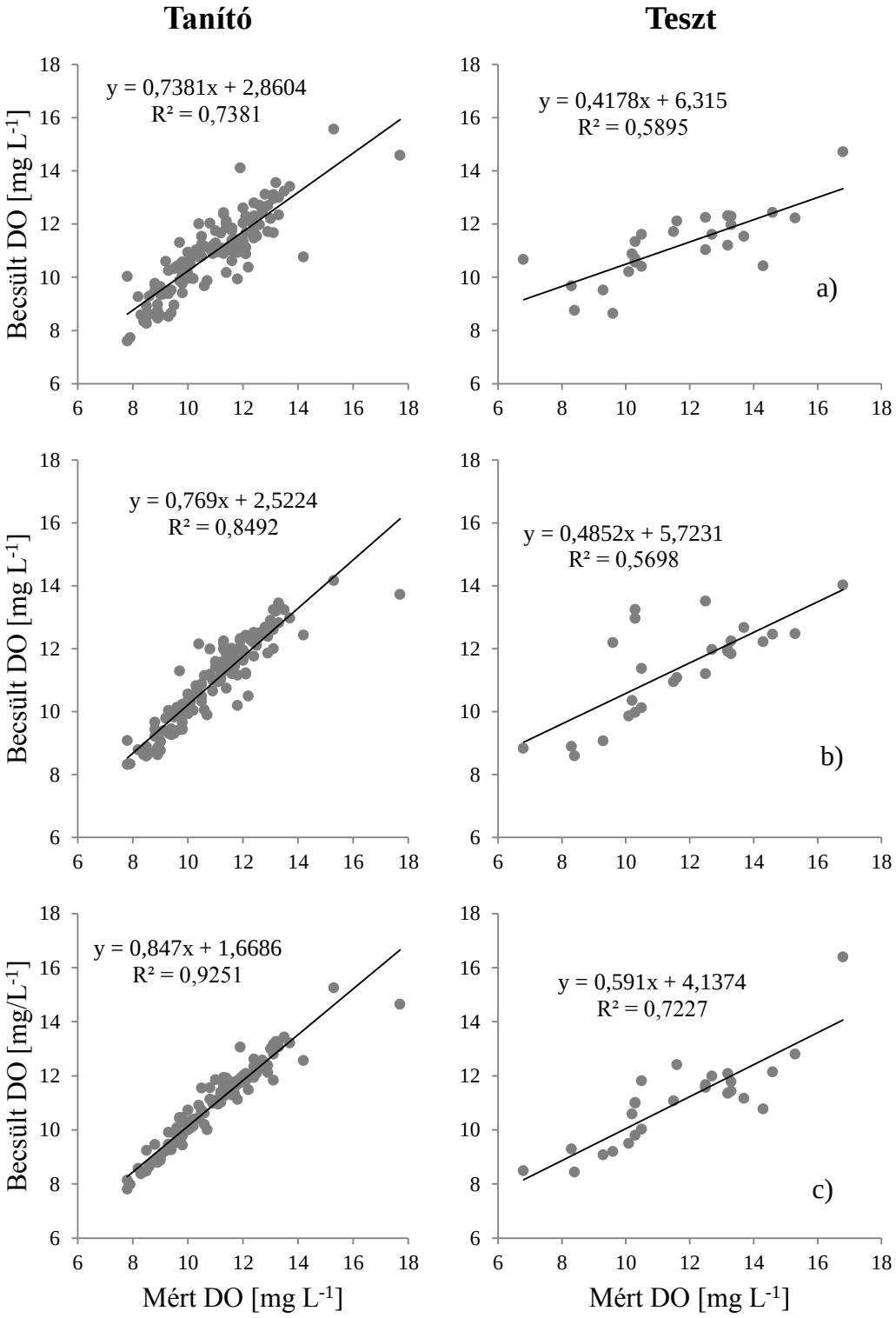
8.1. ábra A dunai mintavételi pontokon mért a) vízhozam b) hőmérséklet c) pH d) vezetőképesség és az e) oldottoxigén-koncentrációjának box-and-whisker plotjai

8.1. táblázat A vizsgált kombinációk modellteljesítményei a tanítóhalmazra vonatkoztatva

Kombináció	Modell	MLR	MLPNN	RBFNN	GRNN
C_A	RMSE [mg L^{-1}]	1,14	0,65	0,84	0,47
	MAE [mg L^{-1}]	0,79	0,43	0,62	0,27
	R^2	0,51	0,85	0,74	0,93
	IA	0,82	0,93	0,92	0,98
C_B	RMSE [mg L^{-1}]	1	0,64	0,48	0,35
	MAE [mg L^{-1}]	0,75	0,46	0,37	0,23
	R^2	0,49	0,8	0,88	0,94
	IA	0,81	0,93	0,97	0,98
C_C	RMSE [mg L^{-1}]	1,13	0,97	0,79	0,77
	MAE [mg L^{-1}]	0,84	0,74	0,6	0,56
	R^2	0,43	0,58	0,72	0,75
	IA	0,77	0,84	0,91	0,91
C_D	RMSE [mg L^{-1}]	1,31	0,96	1,11	0,92
	MAE [mg L^{-1}]	0,97	0,67	0,81	0,62
	R^2	0,36	0,66	0,54	0,7
	IA	0,72	0,88	0,83	0,89



8.2. ábra Négy modellel (MLR, MLPNN, RBFNN, GRNN) becsült és mért oldottoxigén-koncentrációk a) a tanítóhalmazra és b) a teszthalmazra vonatkoztatva C_A kombináció esetén



8.3. ábra Mindhárom neurális hálózattal – a) RBFNN, b) MLPNN, c) GRNN becsült és mért DO-szintjének pontfelhő diagramja a tanító és teszthalmazra vonatkoztatva a C_A kombináció esetén

8.2. táblázat Előfeldolgozások összehasonlítása a teszthalmazra kapott statisztikai mutatók alapján C_A kombináció esetén

Előfeldolgozás	RMSE [mg L^{-1}]	MAE [mg L^{-1}]	R^2	IA	szigma-faktor
standardizálás	1,42	1,14	0,72	0,87	0,3
[-1;1]-re való norm.	1,47	1,15	0,71	0,85	0,17
[-0,9;0,9]-re való norm.	1,47	1,15	0,70	0,85	0,15
[0,1;0,9]-re való norm.	1,48	1,16	0,71	0,85	0,07
[0;1]-re való norm.	1,48	1,15	0,69	0,86	0,08
[-0,8;0,8]-ra való norm.	1,48	1,16	0,71	0,85	0,14

A dunai állomások bemenő paramétereinek boxplotjainak (8.1. ábra) megrajzolásának MATLAB forráskódja:

```
function boxplotDunaossz(fn,p)
switch p
    case 'q'
        oszlop='B:B'
    case 'tv'
        oszlop='C:C'
    case 'pH'
        oszlop='D:D'
    case 'vez'
        oszlop='E:E'
    otherwise
        oszlop='F:F'
end
in1 = xlsread(fn,'Győrzámoly',oszlop);
in2 = xlsread(fn,'Fajsz',oszlop);
in3 = xlsread(fn,'Mohács',oszlop);

in1=in1(2:size(in1));
in2=in2(2:size(in2));
in3=in3(2:size(in3));

g1=ones(size(in1));
g2=2*ones(size(in2));
g3=3*ones(size(in3));
group=[g1;g2;g3];

boxplot([in1;in2;in3],group,'Colors','k','Symbol','w');
set(gca,'Fontsize',15,'FontWeight','bold','Linewidth',0.5);
xlabel('Mintavételi pontok');
switch p
    case 'q'
        ylabel('Vízhozam (m^{3}sec^{-1})');
    case 'tv'
        ylabel('Hőmérséklet (^{o}C)');
    case 'pH'
        ylabel('pH');
    case 'vez'
        ylabel('Vezetőképesség (\muScm^{-1})');
```

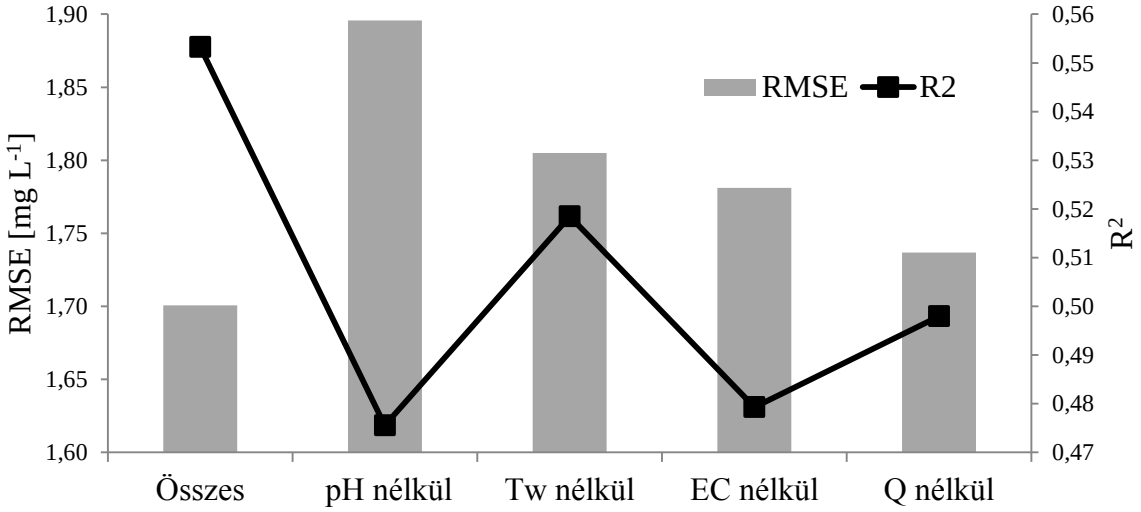
```
    otherwise
        ylabel('Oldott oxigén (mg L^{-1})');
end
text(12.5,27,'E)','Fontsize',20,'FontWeight','normal','Linewidth',0.3);
set(gca, 'XTickLabel',{'Győrzámoly','Fajsz','Mohács'});
hold on;
AQ=[quantile(in1,0.25); quantile(in2,0.25); quantile(in3,0.25)];
FQ=[quantile(in1,0.75); quantile(in2,0.75); quantile(in3,0.75)];
IQT(1:3)=0;
for i=1:3
    IQT(i)=FQ(i)-AQ(i);
end;
for i=1:size(in1)
    if in1(i)>FQ(1)+3*IQT(1) || in1(i)<AQ(1)-3*IQT(1)
        plot(1,in1(i),'r*','MarkerSize',5);
    end;
    if in1(i)>FQ(1)+1.5*IQT(1) && in1(i)<FQ(1)+3*IQT(1)
        plot(1,in1(i),'ko','MarkerSize',3);
    end;
    if in1(i)<AQ(1)-1.5*IQT(1) && in1(i)>AQ(1)-3*IQT(1)
        plot(1,in1(i),'ko','MarkerSize',3);
    end;
end;

for i=1:size(in2)
    if in2(i)>FQ(2)+3*IQT(2) || in2(i)<AQ(2)-3*IQT(2)
        plot(2,in2(i),'r*','MarkerSize',5);
    end;
    if in2(i)>FQ(2)+1.5*IQT(2) && in2(i)<FQ(2)+3*IQT(2)
        plot(2,in2(i),'ko','MarkerSize',3);
    end;
    if in2(i)<AQ(2)-1.5*IQT(2) && in2(i)>AQ(2)-3*IQT(2)
        plot(2,in2(i),'ko','MarkerSize',3);
    end;
end;
for i=1:size(in3)
    if in3(i)>FQ(3)+3*IQT(3) || in3(i)<AQ(3)-3*IQT(3)
        plot(3,in3(i),'r*','MarkerSize',5);
    end;
    if in3(i)>FQ(3)+1.5*IQT(3) && in3(i)<FQ(3)+3*IQT(3)
        plot(3,in3(i),'ko','MarkerSize',3);
    end;
    if in3(i)<AQ(3)-1.5*IQT(3) && in3(i)>AQ(3)-3*IQT(3)
        plot(3,in3(i),'ko','MarkerSize',3);
    end;
end;

print -depsc BoxplotDuna2;
hold off;
print('-dtiff','BoxplotDuna.tiff');
```

M4: Érzékenységi vizsgálat három mintavételi pont adataival a Dunán

Az érzékenységi vizsgálatot, melynek célja, hogy meghatározzam a négy bemenő paraméter közül melyik a legfontosabb az DO-szintjének becsléséhez, mindhárom neurális hálóval elvégeztem a C_D kombináció adataival. Az összes paramétert tartalmazó mintahalmazra kapott eredményekből kiindulva, minden egyes futásnál elhagytam egy-egy paramétert, és megvizsgáltam, hogy hogyan változik a modellek teljesítménye. Megvizsgáltam a tesztelő halmazra kapott RMSE értékeket, és összehasonlítottam a teljes paraméterkörrel kapott tesztelő halmazra vonatkozó RMSE értékekkel, és ez alapján fölállítottam a paraméterek rangsorát. Egyértelmű ugyanis, hogy az a paraméter a legfontosabb, amely nélkül a modell a legrosszabb teljesítményt, vagyis a legnagyobb RMSE értéket adja (8.4. ábra).



8.4. ábra GRNN modellel, a C_D kombinációra elvégzett érzékenységi vizsgálat eredményei

Az érzékenységi vizsgálatot a C_D kombinációra végeztem el mindhárom neurális hálóval (8.3. táblázat). Mindhárom neurális hálóval kapott eredmény a pH paraméter fontosságát erősítette meg, hiszen e paraméter nélkül volt a legrosszabb a teljesítménye a modelleknek. Ez az eredmény alátámasztja az MLR modellel kapott eredményeket (4.2.2 pont), mely vizsgálat során a négy paraméter közül csak a vezetőképességet lehetett kizárni, a többi paraméter szignifikáns maradt.

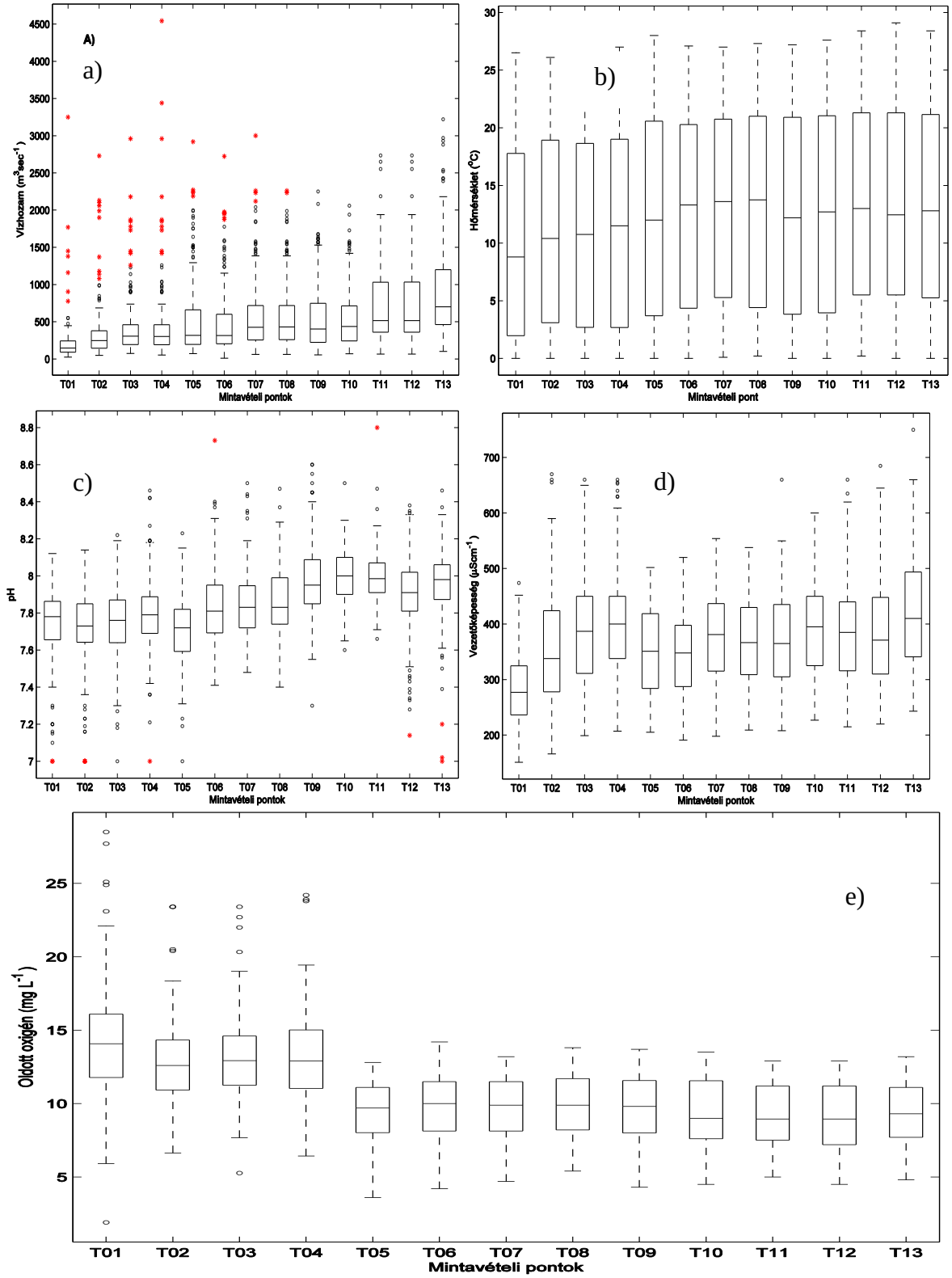
8.3. táblázat GRNN modellel elvégzett érzékenységi vizsgálat négy statisztikai mutatóval kifejezve (Csábrági et al., 2017a)

RMSE	MLPNN	RBFNN	GRNN	R^2	MLPNN	RBFNN	GRNN
Összes	1,70	1,63	1,70	Összes	0,59	0,59	0,55
pH nélkül	1,88	3,90	1,90	pH nélkül	0,43	0,29	0,48
T_w nélkül	1,84	1,81	1,81	T_w nélkül	0,51	0,45	0,52
EC nélkül	1,77	1,71	1,78	EC nélkül	0,54	0,52	0,48
Q nélkül	1,81	1,71	1,74	Q nélkül	0,46	0,51	0,50
MAE	MLP	RBNN	GRNN	IA	MLP	RBNN	GRNN
Összes	1,21	1,17	1,21	Összes	0,78	0,77	0,77
pH nélkül	1,49	1,40	1,39	pH nélkül	0,56	0,78	0,70
T_w nélkül	1,39	1,25	1,34	T_w nélkül	0,77	0,78	0,74
EC nélkül	1,38	1,18	1,28	EC nélkül	0,70	0,75	0,73
Q nélkül	1,25	1,16	1,23	Q nélkül	0,76	0,76	0,75

Mindhárom neurális hálózattal végrehajtott érzékenységi vizsgálat, melyet a C_D kombinációban végeztem el, azt mutatta, hogy a négy bemenő paraméter közül a pH érték játszik a legfontosabb szerepet az DO paraméter előrejelzésében. Ez összhangban van az MLR-rel kapott eredménnyel is, hiszen a modell a négy paraméterből három paramétert választott ki (hőmérséklet, pH, vízhozam), mint az oldottoxigén-koncentrációját legjobban befolyásoló változókat.

Az oxigén elengedhetetlen a dekomponáló baktériumok életfolyamataihoz. A dekompozíció során többek között CO_2 keletkezik, amely a pH-t a savas irányba tolja el. Mindamellett a pH csökkenése előidézhet szerves anyag dekompozíciót, melyet a DO-ból származó oxigén is elősegít. DO koncentráció csökkenés gyakran okoz pH növekedést (Akkoyunlu et al., 2011).

M5: A tiszai vizsgálattal kapcsolatos eredmények



8.5. ábra A tiszai mintavételi pontokon mért a) vízhozam b) hőmérséklet c) pH d) vezetőképesség és az e) oldottoxigén-koncentrációjának box-and-whisker plotjai

8.4. táblázat A TC1 konfiguráció eredményei a tanítóhalmazra vonatkozóan

Kombináció	Modell	MLR	GRNN	RBFNN
TC1-A	RMSE [mg L^{-1}]	2,41	1,79	2,05
	MAE [mg L^{-1}]	1,60	1,15	1,41
	R^2	0,38	0,67	0,55
	IA	0,73	0,87	0,84
	Jellemzők	Q, T_w, EC	0,38	0,45 (36)
TC1-B	RMSE [mg L^{-1}]	2,49	2,13	2,35
	MAE [mg L^{-1}]	1,70	1,43	1,64
	R^2	0,36	0,54	0,43
	IA	0,71	0,80	0,76
	Jellemzők	Q, T_w, EC	0,52	0,58 (15)
TC1-C	RMSE [mg L^{-1}]	2,44	2,11	2,19
	MAE [mg L^{-1}]	1,64	1,40	1,48
	R^2	0,36	0,54	0,48
	IA	0,72	0,80	0,80
	Jellemzők	Q, T_w, EC	0,54	0,52 (24)

8.5. táblázat A TC2 konfiguráció eredményei a tanítóhalmazra vonatkozóan

Kombináció	Modell	MLR	GRNN	RBFNN
TC2-A	RMSE [mg L^{-1}]	0,99	1,24	1,34
	MAE [mg L^{-1}]	0,75	1,01	1,07
	R^2	0,78	0,75	0,60
	IA	0,94	0,85	0,86
	Jellemzők	Q, T_w, pH, EC	1,24	0,41 (3)
TC2-B	RMSE [mg L^{-1}]	2,33	1,47	1,84
	MAE [mg L^{-1}]	1,61	0,91	1,23
	R^2	0,44	0,79	0,65
	IA	0,77	0,93	0,89
	Jellemzők	Q, T_w, pH, EC	0,3	0,35 (89)
TC2-C	RMSE [mg L^{-1}]	2,71	2,49	2,60
	MAE [mg L^{-1}]	1,94	1,76	1,82
	R^2	0,32	0,45	0,37
	IA	0,68	0,72	0,72
	Jellemzők	Q, T_w, EC	0,67	0,63 (10)
TC2-D	RMSE [mg L^{-1}]	2,71	2,07	2,38
	MAE [mg L^{-1}]	1,92	1,44	1,70
	R^2	0,27	0,60	0,44
	IA	0,64	0,83	0,77
	Jellemzők	Q, T_w, EC	0,37	0,56 (38)

8.6. táblázat A TC3 konfiguráció „A” beállításának eredményei a tanítóhalmazra vonatkozóan

Kombináció	Tanítóállomás + tesztállomás	Modell	MLR	GRNN	RBFNN
TC3-A#1	T3, T4 + T5	RMSE [mg L^{-1}]	2,97	2,47	2,33
		MAE [mg L^{-1}]	2,25	1,82	1,72
		R^2	0,04	0,37	0,41
		IA	0,27	0,63	0,76
		Jellemzők	T_w	0,76	0.59 (18)
TC3-A#2	T3, T5 +T4	RMSE [mg L^{-1}]	2,68	1,42	1,77
		MAE [mg L^{-1}]	2,01	0,97	1,30
		R^2	0,24	0,80	0,67
		IA	0,60	0,93	0,89
		Jellemzők	T_w, pH, EC	0,32	0.33 (80)
TC3-A#3	T4, T5 + T3	RMSE [mg L^{-1}]	2,77	1,77	1,97
		MAE [mg L^{-1}]	2,01	1,23	1,46
		R^2	0,27	0,73	0,63
		IA	0,63	0,89	0,88
		Jellemzők	T_w, pH, EC	0,40	0,37 (67)

8.7. táblázat A TC3 konfiguráció „B” beállításának eredményei a tanítóhalmazra vonatkozóan

Kombináció	Tanítóállomás + tesztállomás	Modell	MLR	GRNN	RBFNN
TC3-B#1	T7, T8 + T9	RMSE [mg L^{-1}]	0,94	0,58	0,79
		MAE [mg L^{-1}]	0,72	0,45	0,60
		R^2	0,79	0,92	0,85
		IA	0,94	0,98	0,96
		Jellemzők	Q, T_w, pH, EC	0,44	0,15 (22)
TC3-B#2	T7, T9 +T8	RMSE [mg L^{-1}]	1,02	0,59	0,65
		MAE [mg L^{-1}]	0,78	0,40	0,49
		R^2	0,75	0,92	0,90
		IA	0,93	0,98	0,97
		Jellemzők	Q, T_w, pH, EC	0,36	0,1 (101)
TC3-B#3	T8, T9 + T7	RMSE [mg L^{-1}]	1,03	0,55	0,70
		MAE [mg L^{-1}]	0,76	0,38	0,52
		R^2	0,76	0,93	0,89
		IA	0,93	0,98	0,97
		Jellemzők	Q, T_w, pH, EC	0,34	0,11 (76)

8.8. táblázat A TC3 konfiguráció „C” beállításának eredményei a tanítóhalmazra vonatkozóan

Kombináció	Tanítóállomás + tesztállomás	Modell	MLR	GRNN	RBFNN
TC3-C#1	T11, T12 + T13	RMSE [mg L^{-1}]	0,87	0,58	0,60
		MAE [mg L^{-1}]	0,63	0,45	0,45
		R^2	0,83	0,93	0,92
		IA	0,95	0,98	0,98
		Jellemzők	Q, T_w, pH	0,56	0,08 (34)
TC3-C#2	T11, T13 +T12	RMSE [mg L^{-1}]	0,82	0,44	0,54
		MAE [mg L^{-1}]	0,63	0,33	0,42
		R^2	0,84	0,96	0,93
		IA	0,95	0,99	0,98
		Jellemzők	Q, T_w, pH	0,38	0,07 (65)
TC3-C#3	T12, T13 + T11	RMSE [mg L^{-1}]	0,89	0,52	0,65
		MAE [mg L^{-1}]	0,64	0,39	0,51
		R^2	0,82	0,94	0,90
		IA	0,95	0,98	0,97
		Jellemzők	Q, T_w, pH	0,48	0,1 (24)

9. KÖSZÖNETNYILVÁNÍTÁS

Ezúton szeretnék köszönetet mondani témavezetőimnek: Prof. Molnár Sándornak és Dr. habil. Kovács Józsefnek, hogy lehetőséget adtak disszertációm elkészítésére, valamint hasznos tanácsaikkal irányították kutatómunkámat, és ötleteikkel, bíztatásukkal segítették dolgozatom elkészítését.

Köszönettel tartozom Dr. habil. Molnár Márknak, Dr. Hatvani István Gábornak és Dr. Tanos Péternek a doktori iskola elvégzése során a publikációk és a disszertáció megjelenésében nyújtott hasznos szakmai tanácsaikért, segítségükért és emberi támogatásukért. Külön köszönet illeti meg Dr. Tanos Pétert a dolgozat ábráinak elkészítésében nyújtott segítségéért, és Paul Thatchert a publikációk idegen nyelvű részeinek fordítása során nyújtott segítségéért.

Köszönöm a Mechanikai és Géptani Intézet és azon belül a Mérnökinformatika Központ munkatársainak - Kurucz Gabriellának és Somogyi Ferencnek - a disszertáció elkészítésében nyújtott segítségüket.

Végül, de nem utolsó sorban szeretnék köszönetet mondani a családomnak, hogy támogattak a munkám során.