A MESTERSÉGES INTELLIGENCIÁRA ÉPÜLŐ ELJÁRÁSOK ELMÉLETI ALAPJAI

MOLNÁR SÁNDOR

ÖSSZEFOGLALÁS

A közlemény a mesterséges intelligenciával, gépi tanulással, egyéb osztályozási eljárásokkal kapcsolatos főbb eljárásokat tekinti át, amelyeket a precíziós mezőgazdaság területén is alkalmazhatnak. Bemutatja a gépi tanulás és adatbányászat alapvető fogalmait, kitérve a mély tanulás kérdésére. A neurális hálózatok elméleti felépítését és alapelemeit áttekintve kitér néhány, a főáramú gyakorlatban alkalmazott megközelítésre, így többek között az általános regressziós neurális hálózatokra (GRNN) a többrétegű perceptron (MLPNN) hálózatokra és a radiális bázisfüggvényes neurális hálózatokra (RBFNN) sor. Emellett foglalkozik az adatbányászatban használható néhány osztályozási eljárással, bemutatva a wavelet elemzés alkalmazási lehetőségeit, továbbá egy új eljárást, a hierarchikus klaszteranalízist is.

SUMMARY

Molnár, S: THEORETICAL BACKGROUND OF ARTIFICIAL INTELLIGENCE BASED PROCESSES

This article provides a review of the mainstream methods of artificial intelligence, machine learning and classification algorithms which are applicable in precision agriculture. The basic concepts of data mining, machine learning and deep learning are presented. The theoretical structure and building elements of artificial neural networks are then introduced and some mainstream approaches are reviewed, including general regression neural networks (GRNN), multilayer perceptron networks (MLPNN) and radial base function neural networks (RBFNN). Following these some classification methods used in data mining are presented in the article including wavelet analysis and the new method of hierarchical cluster analysis.

BEVEZETÉS

A precíziós mezőgazdaság megvalósításával bizonyítottan sokszorosával növelhető a termelés hatékonysága. Mindez természetesen jelentős gazdasági előnyökkel is jár, emellett nem kevésbé fontos a természetvédelmi-, környezetvédelmi- és általában pozitív hatása.

A precíziós mezőgazdaság bevezetéséhez elengedhetetlenül fontos a gazdálkodási tevékenység körülményeinek mélyreható ismerete. Ezen a területen is számos esetben alkalmazhatóak a gépi intelligenciára alapozódó eljárások és megoldások, amelyek jövőbeli jelentőségét nehéz túlbecsülni. Az ún. mesterséges intelligencia olyan gyűjtőfogalom, amely számos eltérő megközelítést ölel fel, többek között a gépi tanulás, a mély tanulás, az adatbányászat, a big data témaköreit, amelyek az iparban, a mezőgazdaságban, a gyógyászatban, a reklámiparban, és az élet szinte minden területén egyre elterjedtebbek. Mivel ezek a területek szorosan kapcsolódnak egymáshoz, átfedik egymást, épp ezért érdemes áttekinteni, mit is takarnak ezek az elnevezések, még ha nincsenek is kiforrott, a szakma által egységesen és következetesen használt definíciók.

ALAPFOGALMAK

A modern számítógépes hálózati eszközök segítségével ma már a korszerű termelőeszközök nagyszámú adatot mérnek, és az olcsó háttértároló eszközöknek köszönhetően ezeket az adatokat rögzítik és tárolják is. Gyakran olcsóbb, gyorsabb és egyszerűbb új, nagyobb kapacitású tárolóeszközöket beszerezni, mint a meglevő adatokat átvizsgálni és a szükségteleneket törölni. Szigorúan véve a sok adat elnevezésére szolgál a *big data* kifejezés. Az adatokon túl ide lehet érteni a hozzájuk kapcsolódó feladatokat; a méréssel, rögzítéssel, továbbítással, tárolással, feldolgozással, vizualizációval, és nem utolsó sorban az értelmezéssel, előrejelzésekkel kapcsolatos technikákat, eljárásokat.

A számítógépek képesek tanulásra, vagyis a megismert adatok alapján helyes előrejelzéseket képesek adni. A *gépi tanulást* olyan esetekben használják, amikor explicit algoritmusok programozása nem nyújt megfelelő megoldást, például az adatok nagy mennyisége miatt, vagy azért mert nehéz egyértelműen megfogalmazni az adott problémát (például képfelismerés). Néhány jellemző alkalmazási terület lehet például az email-szűrés, az optikai karakterfelismerés, a számítógépes biztonság területe (vírusok, támadások) és a számítógépes látás.

A mély tanulás (deep learning) olyan gépi tanulási módszer, amelynek során sok példával tanítanak egy sok rétegből álló neuronhálózatot. Tipikus példa erre a képfelismerés: a rendszernek rengeteg képet mutatnak meg, amin például kutya, vagy egy fontosabb alkalmazási területen rákos daganat, van, majd rengeteg olyat is, amelyeken nem kutya van. Ezek után a rendszer képes lesz eldönteni, hogy egy újonnan látott képen kutya van-e vagy sem. Nem adja meg viszont a “kutyaság” definícióját, tehát amolyan fekete dobozként működik, amibe nem látunk bele. Ez a működés közel áll az emberi látás, a felismerés, az intelligencia működéséhez. A módszer nagyon hatékony, rendkívül számolásigényes, és nagyon-nagyon sok adatra van szüksége. Ahhoz, hogy az ilyen típusú tanulási eljárások hatékonyak legyenek nagy teljesítményű grafikus processzorra (GPU) van szükség. A 2017-

es év nagy szenzációját jelentő AlphaGo és AlphaZero programok is ilyen mély tanulást használnak.

Az *adatbányászat* fő területe jellemzően az adatokban levő mintázatok, struktúrák és adatok közötti összefüggések keresése, az adatokból érdekes, értékes és értelmes információ kinyerése. Az elnevezés némileg megtévesztő, mert az adatbányászat célja nem az adatok előállítása (kibányászása), nem is az adatokban való keresgélés, hogy megtaláljuk a néhány fontos adatot; hanem a szerkezet, a szabályszerűségek feltárása (*Molnár*, 2019).

A továbbiakban elsősorban erre, az összefüggések keresésére, koncentrálunk a gépi tanulás fő eszköztárának bemutatása során.

NEURÁLIS HÁLÓZATOK ÉS GÉPI TANULÁS

A neurális hálózatokat a természettudományi, gazdasági és a műszaki tudományok területén egyaránt gyakran alkalmazzák különböző feladatokra, mint például beszéd-, alak- és karakterfelismerésre, kép- és jelfeldolgozásra, adatbányászatban csoportosításokra, robottechnikában szabályozásokra, a 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).

A mesterséges neurális hálózat az emberi idegrendszer alapján ihletett gépi tanuló architektúra, amely a biológia rendszerek olyan tulajdonságait vette alapul, mint a nagyszámú, alapegységekből (neuron) való felépítés, valamint ezen egységek között lévő nagyszámú kapcsolatok (szinapszisok). A neurális hálózat nem kívánja az adott jelenséget modellezni, arra törvényszerűséget megállapítani, hanem azt fekete dobozként kezeli, csak a bemenő és kimenő adatokat tekinti. Emiatt szokták ezeket a modelleket „adatvezérelt modellnek” (data-driven model) is nevezni.

A bemenő adatokat az **x** vektor, a mért kimeneti értékeket a **v** vektor, a becsült, számított értékeket pedig az **y** vektor jelöli. A jelenséget az 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, amelyek a tanítóhalmazt alkotják. A létrehozott neurális hálózatot minden esetben tesztelni szükséges, ehhez szükségesek a

1. ábra *A perceptron felépítése, ahol b a torzítás értéke (bias)*

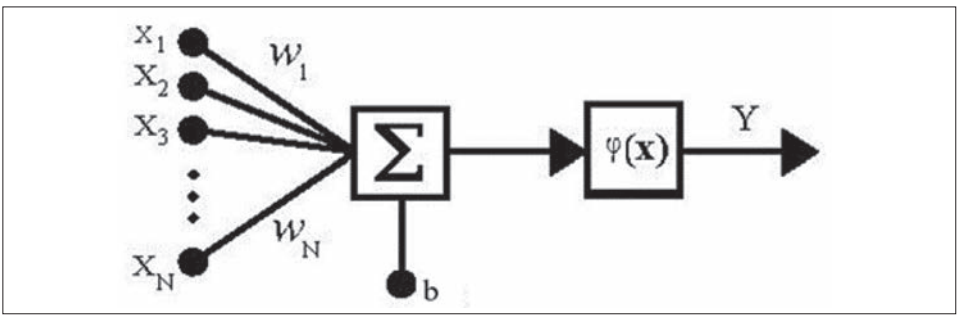

Figure 1. Structure of a perceptron, where b is the extent of bias

tesztpontok, azaz a teszthalmaz. E halmazok meghatározása alapjaiban befolyásolja a később létrehozott neurális hálózat hatékonyságát (*Jang mtsai,*1997).

A neuron egy információ-feldolgozó egység, a neurális hálózat alapegysége, amelyek irányított kapcsolatokkal (szinapszisokkal) vannak összekötve egymással. Ezeket a szinapszisokat számszerűsített súlyokkal (weight) tudjuk kifejezni (wji), amelyek meghatározzák a kapcsolat erősségét és előjelét, illetve kifejezik, hogy a j-edik egységtől az i-edik egység felé vezető kapcsolatról van szó (*1. ábra*). Az egy neuronból álló egységet perceptronnak nevezzük.

Tanítás során minden egyes perceptron a bemeneteinek egy súlyozott összegét számítjuk ki (*Russell és Norvig*, 2005), majd a kimenet meghatározására a ϕ aktivációs (transzfer) függvényt alkalmazzuk a kapott összegre:

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

Aktivációs függvényként leggyakrabban szigmoid, tehát S alakú, függvényt használunk, amelyek közül a logisztikus függvény képlete a következő:

$$\varphi(x) = \frac{1}{1+e^{-ax}} \qquad x \in \Re$$

A neurális hálózatok adaptív, tanuló rendszerek, amelyek jellemzője, hogy nem rögzített képességgel rendelkeznek, amelyek csak egy adott feladat elvégzésére teszi azokat alkalmassá, hanem képességeiket fejleszteni tudják, tehát alkalmazkodni tudnak a változó körülményekhez.

TÖBBRÉTEGŰ PERCEPTRON (MLPNN) HÁLÓZATOK

Az MLPNN modellben a neuronok, perceptronok rétegekbe szerveződnek, amely rétegek neuronjai csak a szomszédos rétegek neuronjaival vannak kapcsolatban, viszont a 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, illetve a kimeneti réteg (*2. ábra*). A modell a hiba-visszaterjesztéses iterációs algoritmuson alapszik. A hiba visszaterjesztéses algoritmust többféle eljárás is megvalósíthatja, például a Levenberg-Marquardt eljárás (*Marquardt*, 1963).

RADIÁLIS BÁZISFÜGGVÉNYES NEURÁLIS HÁLÓZATOK (RBFNN)

Az RBFNN modell egy ellenőrzött tanítást megvalósító előrecsatolt neurális hálózat, amely három réteget tartalmaz: egy bemenő, egy rejtett és egy kimenő réteget. Az RBFNN nagy előnye az MLPNN-hez képest, hogy nem egy lokális minimum környezetébe koncentrálódik, hanem a globális minimumot keresi. Az RBFNN modell tanító szakasza két különálló, egymástól független, részre bontható (*3. ábra*). 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, vagyis a modell rejtett rétege önszerveződő réteg. Az RBFNN rejtetett rétegében mindig radiális bázis

2. ábra *A többrétegű perceptron neurális hálózat sematikus ábrája*

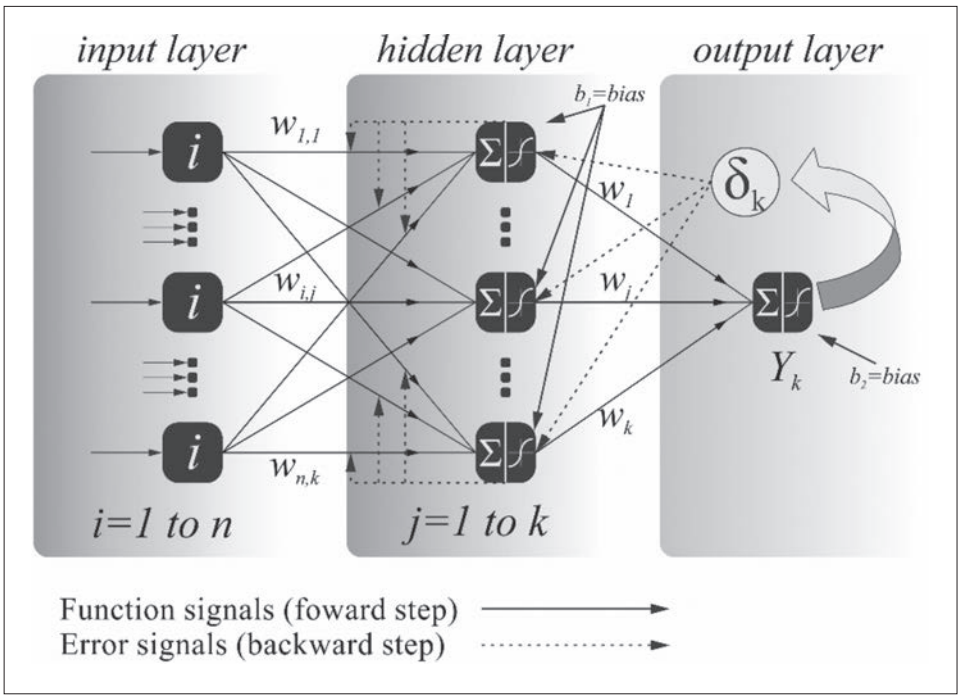

Figure 2. The scheme of a multilayer perceptron neural network

függvények lesznek az aktivációs függvények. 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.

ÁLTALÁNOS REGRESSZIÓS NEURÁLIS HÁLÓZATOK (GRNN)

Az általános regressziós neurális hálózatot először Specht vezette be 1991-ben az MLPNN modell egy lehetséges alternatívájaként. A GRNN modell az RBFNN egy módosított formája. A GRNN modell ellenőrzött tanítást végrehajtó „egy utas" neurális háló, azaz a modell tanító folyamatához nincs szükség iterációs eljárásra, a bemeneti adatok csak egyszer áramolnak végig a rendszeren, emiatt a modell tréningezési ideje meglehetősen rövid. A GRNN modell egy négyrétegű, előrecsatolt neurális hálózat, amely egy bemeneti réteggel, egy mintaréteggel, egy összegző réteggel és egy kimeneti réteggel rendelkezik. 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 tréningezéssel, hiszen a GRNN modellnél is a Gauss-féle radiális bázis függvény a mintaréteg aktivációs függvénye. A GRNN modell használata során csupán a szigma-faktor értékét szükséges megadni (*4. ábra*).

3. ábra *A radiális bázis függvényes neurális hálózatok sematikus ábrája*

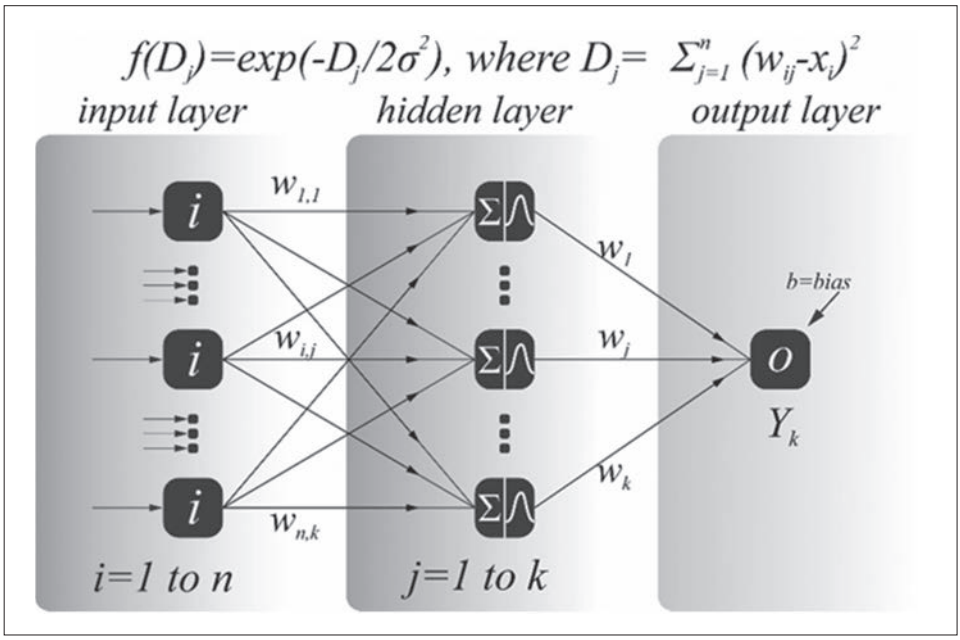

Figure 3. The scheme of radial base function neural networks

EGYÉB KÖVETKEZETETÉSELMÉLETI MEGKÖZELÍTÉSEK

Bizonyos objektumok csoportosításának (klasszifikációjának) igénye gyakran felmerülő probléma a különböző tudományterületeken. Egy általánosan használt módszer a modern kutatásokban a klaszteranalízis (CA; *Everitt és mtsai*, 2011). Alkalmazható többek között különböző fajokra a biológiában, azonosan viselkedő területek kijelölésére a távérzékelésben és mintavételi pontok csoportosítására a föld- és környezettudományokban. A CA során az egyik legfontosabb kérdés, hogy hogyan lehet meghatározni a legnagyobb csoportokat, azaz azon objektumok körét, melyek hasonló tulajdonságokkal rendelkeznek. Mindez lehetővé teheti például a mintavételi pontok számának csökkentését minimális információveszteség mellett, vagy akár anélkül.

Az klaszteranalízis egyik leggyakoribb típusa a hierarchikus klaszterezés (HCA). Alkalmazása során a kiindulópontban minden objektum külön csoportba tartozik, majd a további lépések során minden esetben összevonjuk a két legközelebbi csoportot. A csoportok összevonását egészen addig folytatjuk, amíg az összes objektum egyetlen csoportba nem kerül. A létrehozott klaszter alapvető módon attól függ, hogy az egyes objektumok közötti távolságot milyen módszerrel határozzuk meg. A föltudományi kutatásokban például gyakori a négyzetes euklideszi távolság alkalmazása, illetve a Ward-módszerrel létrehozott HCA, mert a csoportokon belüli variancia így minimalizálható (*Hatvani* és *mtsai,* 2011; *Kovács* és *mtsai,* 2012a, 2012b). A klaszteranalízis eredménye dendrogramon ábrázolható. Attól függően, hogy mely távolságon belül tartoznak az egyes objektumok

4. ábra *Az általános regressziós neurális hálózatok sematikus ábrája*

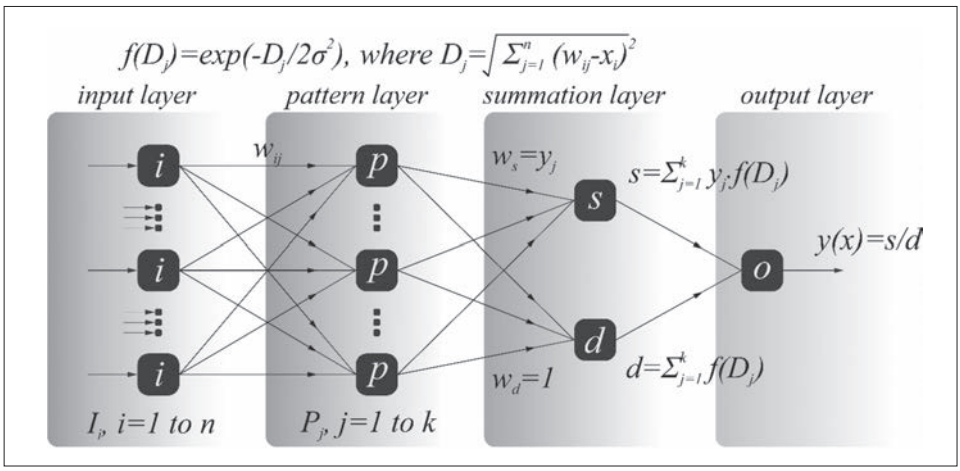

Figure 4. The scheme of general regression neural networks

azonos csoportba különböző csoportosításokat határozhatunk meg, azaz a csoportosítás attól függ, hogy a dendrogramot milyen transzformált távolságnál „vágjuk el" (*5. ábra*).

Klaszterezés során mindig nehéz eldönteni, hogy az egyes csoportokat mely távolságon belül szükséges összevonni, mindazonáltal e döntés az adott csoportosítás alapját képezi.

Számos kutatás foglalkozott a különböző csoportosítási technikák javításával. Az ökológiában *McKenna* (2003) adott erre jó példát, amikor ökológiai közösségeket tanulmányozott. *Rowan és mtsai,* (2012) pedig kockázatbecslésen alapuló módszert dolgozott ki Anglia, Skócia és Észak-Írország tavainak csoportosítására.

5. ábra *Példa a hierarchikus klaszteranalízis eredményére*
A szaggatott vonal öt csoportnál „vágja el" a dendrogramot

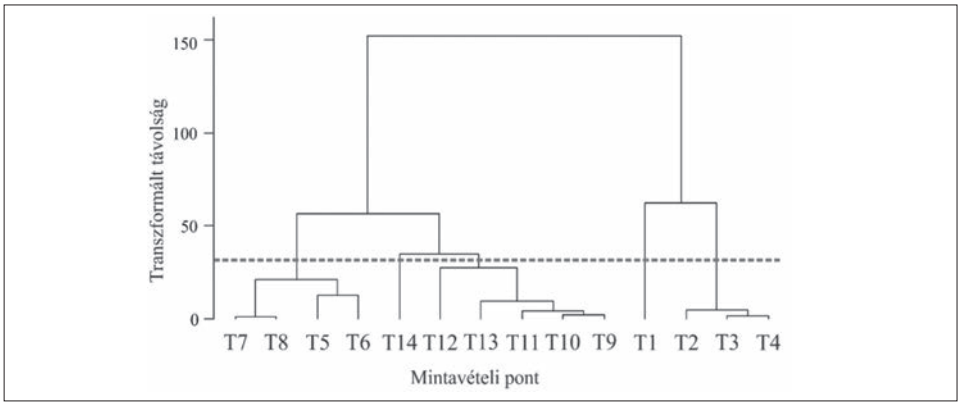

Figure 5. Example of hierarchical cluster analysis

the dotted line cuts the dendogram at 5 groups

Yang (2012) pedig többjelölős csoportosításon (multi-label classification) alapuló modellt alkotott a fenntartható ártérmenedzsment érdekében.

Fontos kérdés továbbá, hogy az egyes paraméterek milyen mértékben befolyásolják az adott csoportok létrejöttét. E kérdés megválaszolásában jelentős segítséget nyújt a Wilks' λ statisztika (*Wilks*, 1932), ami az adott paraméterre vonatkozóan a csoportokon belüli és a teljes eltérések négyzetösszegeinek hányadosa:

$$\lambda = \frac{\Sigma_i \Sigma_j (x_{ij} - \overline{x_i})^2}{\Sigma_i \Sigma_j (x_{ij} - \overline{x})^2}$$

ahol az x_{ij} az *i*-edik csoport *j*-edik eleme, $\overline{x}_i$ az *i*-edik csoport és $\overline{x}$ az összes adat átlaga.

Ha a kapott λ érték egyenlő 1-gyel (λ=1), akkor a csoportok átlagai nem különböznek, tehát a vizsgált paraméter nem befolyásolta a csoportok alakulását. Ha a kapott λ érték egyenlő 0-val (λ=0) akkor a paraméter maximálisan befolyásolta a csoportok alakulását (*Hatvani* és *mtsai,* 2011; *Kovács* és *mtsai,* 2012a, 2012b). Eredményként felállítható a paramétereknek egy Wilks' λ statisztika szerinti sorrendje, amiből az adott paraméter csoportosításban betöltött szerepe eldönthető.

Bármely kapott csoportosítás azonban validálásra szorul, mert azok létezését valamilyen hipotézisvizsgálati eljárással is igazolni kell. Erre a célra megfelelő módszer a diszkriminancia analízis. A Fischer-féle lineáris diszkriminancia analízis során (LDA) az eredeti adatok olyan lineáris kombinációját alkotjuk meg, ahol a csoportokon belüli változékonyság minimális, míg a csoportok közötti különbségek maximálisak.

Matematikai értelemben azokat az a_i, vektorokat keressük, ahol az maximálisan eleget tesz a normalitás és a korrelálatlanság feltételének a transzformált térben.

$$\frac{a_i{}^T S_K a_i}{a_i{}^T S_B a_i}$$

A képletben az S_K a csoportok közötti, az S_B pedig a csoporton belüli kovariancia mátrix:

$$S_K = \frac{1}{n}\sum_{i=1}^{k} n_i (\overline{x_i} - \overline{x})(\overline{x_i} - \overline{x})^T,$$

$$S_B = \frac{1}{(n-k)}\sum_{i=1}^{k} n_i \hat{\Sigma}_i,$$

ahol k a csoportok száma, n_i a megfigyelések száma az i-edik csoportban, n a csoportonkénti megfigyelések számának összege, $\overline{x}_i$ a mintaátlag az i-edik csoportban, $\overline{x}$ az összes megfigyelés átlaga és $\hat{\Sigma}_i$ az i-edik csoport kovariancia mátrixa.

Az a_i vektorok alapján számított lineáris diszkrimináló síkokkal a megfigyeléseket *k* csoportba soroljuk. A diszkriminancia analízis által készített és az eredeti csoportbeosztás alapján a helyesen kategorizált esetek százalékos aránya kiszámítható (*Kovács és Erős*, 2017). A diszkriminancia analízis a klaszteranalízissel készített csoportbeosztás ellenőrzésére is hatékonyan használható (*Kovács* és *mtsai,* 2012a, *Hatvani* és *mtsai,* 2014). Amennyiben azonban a csoportok egymást átfedik, az

egyes megfigyelések több csoportba osztása nehezebb. Ennek következménye, hogy az LDA általában nagyobb százalékban csoportosít helyesen, ha a csoportok száma kisebb. Ezt a tényt a validálási folyamat során szem előtt kell tartani.

Különböző objektumok csoportosítása esetén általában a cél egy optimális csoportosítás meghatározása. A legtöbb esetben azonban nehéz kiválasztani a lehetséges csoportosítások közül, hogy melyik az optimális. *Davis és Bouldin* (1979) vagy *Dunn* (1973) módszere ennek a feladatnak a meghatározását célozzák. Az említett megoldásoktól azonban eltér a kombinált klaszter- és diszkriminancia analízis (CCDA, *Kovács* és *mtsai,* 2014), mert – amint látható lesz a későbbiekben – az optimális csoportszámot objektív módon határozza meg. Az optimálisnak tekintett csoportok hasonló objektumokat tartalmazhatnak, de ez nem feltétlenül jelenti azt, hogy homogének is. A CCDA további tulajdonsága, hogy nem csupán egy optimális csoportosítás meghatározását teszi lehetővé (*Kovács és mtsai*, 2012b, 2015; *Kovács és Erőss*, 2017), hanem homogén csoportok keresését is.

A CCDA tehát két korábban bemutatott módszert kapcsol össze, a HCA-t és LDA-t. Az előbbi létrehozza a csoportokat, illetve azok hierarchiáját, ahol a legelemibb szinten minden mintavételi pont saját csoportot alkot, míg a legdurvább felosztásban minden mintavételi pont egyetlen csoporthoz tartozik. Az LDA pedig olyan síkokat ad meg, amelyek a megjelölt csoportokat optimálisan választják el és eredményként a síkok által helyesen klasszifikált megfigyelések százalékos arányát adja.

A CCDA 3 fő lépésből áll:

I. Az alapcsoportosítás létrehozása HCA-val.

II. Magciklus, amelyben az előzetes (alapcsoportosítás) és *k* darab véletlenszerű csoportosítás helyessége kerül meghatározásra LDA segítségével.

III. Az eredmények kiértékelése a magciklus alapján, ahol a csoportok további alcsoportokra történő bontásáról kell döntést hozni a homogén csoportok elérése érdekében.

Az I-III lépések ismétlése mindaddig szükséges, amíg további bontás már nem javasolt a harmadik lépésben.

A CCDA használata előtt természetesen szükséges az adatok előkészítése. Fontos megjegyezni, hogy a CCDA nemcsak normális eloszlás esetén biztosít kielégítő eredményeket, hanem másféle eloszlások esetén is, mindaddig, amíg az adatoknak a normális eloszlástól való eltérését az eloszlás ferdesége, és nem a kiugró értékek okozzák. Mindezek mellett fontos megjegyezni, hogy nem lehetnek hiányzó adatok.

Legyen például N a mintavételi helyek száma. Első lépésként a mintavételi helyek $SL_1,\ldots,SL_N$ alapcsoportosítása szükséges. Egy ilyen alapcsoportosítás *N* különböző csoportosításból áll $GR_1,\ldots,GR_N$. Ezek rekurzívan következőképpen nyerhetők:

$GR_N=\{\{SL_1\},\ldots,\{SL_N\}\}$ jelentse, hogy az *N* különböző mintavételi hely *N* különféle csoportot alkosson. Az *i* (*N-1*- től,…,*1*-ig) esetén, a GR_i csoportosítás a GR_i+1 csoportosításból úgy nyerhető, hogy pontosan két csoportot olvaszt egybe a GR_i+1 csoportosításból, míg a többi csoport megmarad GR_i csoportosításban is. Természetesen a két egybeolvasztott csoport az adott lépésben egymáshoz a lehető „legközelebbi" kell, hogy legyen.

Így a GR_i csoportosítás mindig *i* csoportot tartalmaz. A GR_1 csoportban minden mintavételi hely egy csoporthoz tartozik, azaz $GR_1=\{SL_1,\ldots,SL_N\}$. Egy ilyen csoportosítást hoz létre a HCA a mért paraméterek mintavételi pontonkénti átlagaira. A távolságszámítás során *Ward* (1963) módszerét alkalmaztuk, de természetesen valamely más módszer is alkalmazható. HCA használata esetén, $GR_1,\ldots GR_N$ csoportosítások az így kapott dendrogram különféle távolságoknál történő elvágásával kaphatók.

A II. lépésben, minden így kapott csoportosításra $GR_2,\ldots,GR_N$-re az úgynevezett magciklust kell futtatni. A GR_1-es csoportosításra a magciklust nincs értelme futtatni, mivel minden mintavételi hely egy csoporthoz tartozik. A magciklus alapvető ötlete azon alapul, hogy összehasonlítja, hogy a csoportokhoz tartozó megfigyelések milyen mértékben különíthetők el LDA segítségével egy véletlenszerű beosztáshoz képest, illetve, hogy az előbbi szignifikánsan jobb-e az utóbbinál. Mindez azonban a nem homogén csoportok jelenlétére utal a vizsgált csoportosításban.

A III. lépésben az eredményeket értékelve, a csoportok tovább bontásáról kell döntést hoznunk. Jelölje például i^* azt a csoportosítást, amelyre d_i maximális. Az ehhez tartozó GR_{i^*} csoportosítás tekintendő optimálisnak. Mindez azonban nem jelenti azt, hogy a GR_{i^*} csoportosítás homogén. Csak ha $i^*=1$, azaz $GR_{i^*} = GR_1 =\{SL_1,\ldots,SL_N\}$ az optimális csoportosítás az $SL_1,\ldots, SL_N$ mintavételi helyeken, és ez tekinthető homogénnek. Ebben az esetben a GR_{i^*} csoportosítás csoportjait alcsoportoknak nevezzük (sub-groups, $SG_1,\ldots,SG_{i^*}$). Ezen alcsoportok iteratív vizsgálata szükséges a fenti három lépés segítségével, mindaddig, amíg homogén csoportokat nem találunk. Ez azt jelenti, hogy először az SG_1 alcsoporthoz keresünk egy alapcsoportosítást, majd ennek csoportosításait vizsgáljuk a magciklus segítségével, melynek eredményei alapján további bontásról dönthetünk a harmadik lépésben, ha ez szükséges. $SG_2,\ldots,SG_{i^*}$ alcsoportokat hasonlóképpen vizsgáljuk (CCDA, 2014).

A Combined Cluster and Discriminant Analysis (CCDA; *Kovács* és *mtsai,* 2014) során tehát nem csupán hasonló csoportokat kerestünk, hanem homogéneket is, amelyek elemei azonos tulajdonságokkal rendelkeznek. A CCDA módszer ennek alapján lehetővé teszi, hogy objektív alapon dönthessünk a csoportok homogenitásáról.

A periociditás vizsgálatra az átlagképzésnél kifinomultabb eljárás a Lomb-Scargle-módszer (L-S; *Lomb,* 1976; *Scargle*, 1982; *6. ábra*). Az L-S-módszer szignifikancia szintet rendel egy adott periódus meglétéhez, így pontosabb képet ad esetünkben például az éves periodicitásról is. Az L-S-módszer azonban, miután azonosította az éves periódussal rendelkező komponenseket egy adott idősorban, arról már nem ad felvilágosítást, hogy az adott periódus az egész vizsgált időszakban jelen volt-e vagy sem. Ennek oka, hogy az L-S módszer időben nem lokalizált.

A periodicitás vizsgálat wavelet-traszformációval is lehetséges. A waveletspektrum-becslés legnagyobb előnye az, hogy idő-frekvencia felbontást tesz lehetővé. A módszer lényege egy dekomponálási eljárás (Fourier-transzformációval), amelynek során a vizsgált jelet trigonometrikus (szinusz és koszinusz) függvényekre bontjuk. Az oszcilláló komponensek állandó változékonysága megköveteli a spektrumbecslő eljárás nagyfokú adaptivitását, ezt a követelményt

6. ábra *Példa a periodicitás vizsgálat módszereihez*
A havi alapstatisztikák segítségével megjelenített éves periodicitás
A) A Lomb-Scargle-módszer grafikus eredménye; B) Amely szerint a kiválasztott idősorban az éves periodicitás jelenléte szignifikáns.

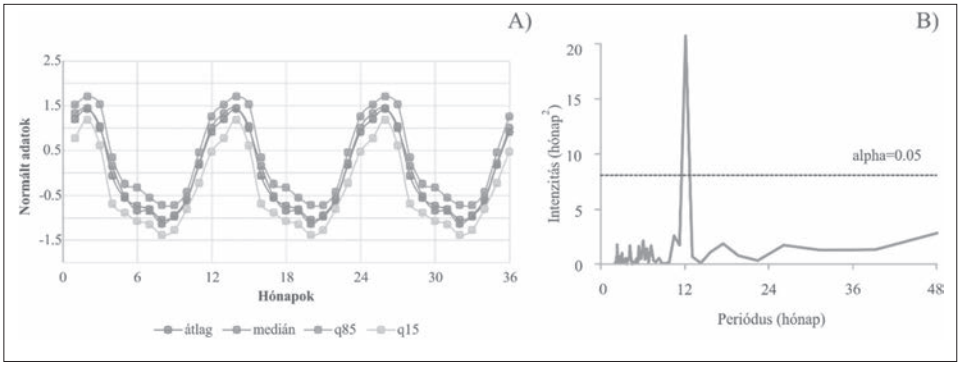

Figure 6. Example for periodicity assessment

Annual periocidity displayed on the basis of monthly periodicity.
A) Graphical results of the Lomb-Scargle method, B) Shows that the presence of annual periodicity is significant in the examined time series.

elégíti ki az alkalmazott wavelet-transzformáció. A wavelet-módszer tehát időben és frekvenciában lokalizált, azaz idő-frekvencia felbontást eredményez, így ezzel lehetővé válik, hogy a jelnek időben változó periodikus jellegzetességeit megragadjuk.

A wavelet analízis alkalmazkodóképességét a hosszú szinusz hullámok alkalmazása helyett, sok rövid ”wavelet” használata eredményezi. A wavelet-transzformáció definícióját, az adatok és a wavelet függvények konvolúciójaként ként értelmezhetjük:

$$W_n^X(s) = \sqrt{\frac{\delta t}{S}} \sum_{n'=1}^{N} X_n' \Psi_0^* \left[\frac{(n'-n)\delta t}{S} \right]$$

ahol a csillag (Ψ^*) a komplex konjugáltat jelöli, X_n az eredeti idősor, és a skála, Ψ a waveletfüggvény és δt a felbontás mértéke.

Az adaptivitás a skálázási eljárásban jelentkezik: a fő waveletből (mother wavelet) sorozatos skálázással, azaz nyújtással és összenyomással származtatja az eljárás a daugther waveleteket. A transzformáció dilatációs függvényét felül és alul áteresztő szűrők hierarchiájával reprezentálhatjuk. Ezeknek a szűréseknek a sorozatán keresztül a jel egyre nagyobb felbontású komponensekre bomlik. Eredetileg a wavelet-transzformáció éppen az ilyen többszörös felbontást szolgálta, vagyis a jeleknek a skálatartományban (scaling space) történő dekomponálását annak érdekében, hogy feltárható legyen a jelek esetleges önhasonló struktúrája (self-similarity structure). Ebben az esetben éppen a skálaegyüttható szolgáltatják a felbontás végeredményét (*Kovács,* 2015).

Annak érdekében, hogy PSD becslést lehessen végezni a wavelet-transzformációval, speciális waveletet (pl. Morlet-wavelet) célszerű választani és megfelelő transzformációkkal származtatni kell a skálából a frekvenciatengelyt.

A Morlet-wavelet egyesíti a trigonometrikus függvényeknek azt az előnyét, hogy oszcillálnak az exponenciális függvény gyors lefutásával, ami a lokalizáltságot biztosítja. (*Kovács és mtsai*, 2010)

Ahhoz, hogy a spektrális összetevők szignifikanciájának kérdését felvethessük, a nullhipotézishez tartozó "háttérspektrum"-ot kell megválasztanunk. A legtöbb természeti folyamatban a háttér jól reprezentálható fehér vagy vörös zajjal (a fehér zaj spektruma teljes, minden frekvencia-összetevőt tartalmaz, a vörös zajban az alacsony frekvenciás komponensek vannak hangsúlyozva). Földtani folyamatokra reális választás a vörös zaj, ekkor a Fourier-teljesítményspektrum eloszlása $\chi 2$, és mivel a lokális wavelet spektrum megegyezik az átlagos Fourier-spektrummal, a lokális wavelet spektrum konfidencia intervalluma ebből számítható (*Kovács,* 2007).

A módszer alkalmazása az R szoftverben, a dplR csomag ad lehetőséget. Mivel a wavelet spektrum két független változó függvénye (idő és frekvencia), általában valamilyen háromdimenziós vizualizálási technika alkalmazásával mutatható be az eredménye. A különböző programok izovonalas ábrát használnak erre a célra (*7. ábra*). A színskála a periodicitás meglétének valószínűségét jelöli: a meleg színek felé növekszik az adott periódus valószínűsége, illetve 5%-os szignifikancia szinten a vastagított, fekete vonallal lehatárolt terület fogadható el periodikusnak.

Az ábrán az A) A Morlet-anyawavelet sematikus ábrája; B) A wavelet spektrumanalízis (WSA) kimeneti eredménye Szolnokon, az NO_3-N változó esetében; C). A panel felső ábrája az adott változó újramintavételezett adatsorát ábrázolja. Az alsó, izovonalas ábra maga PSD grafikon 5%-os szignifikancia szinten, vörös zajhoz hasonlítva a vastag fekete kontúrral határolt terület. A sraffozott terület a COI-t jelöli, míg a vízszintes vonal jelöli az éves periódus szintjét.

7. ábra *Éves, 12 hónapos periódus, azonosítása Lomb-Scargle-periodogram segítségével a NO_3-N paraméter esetében*

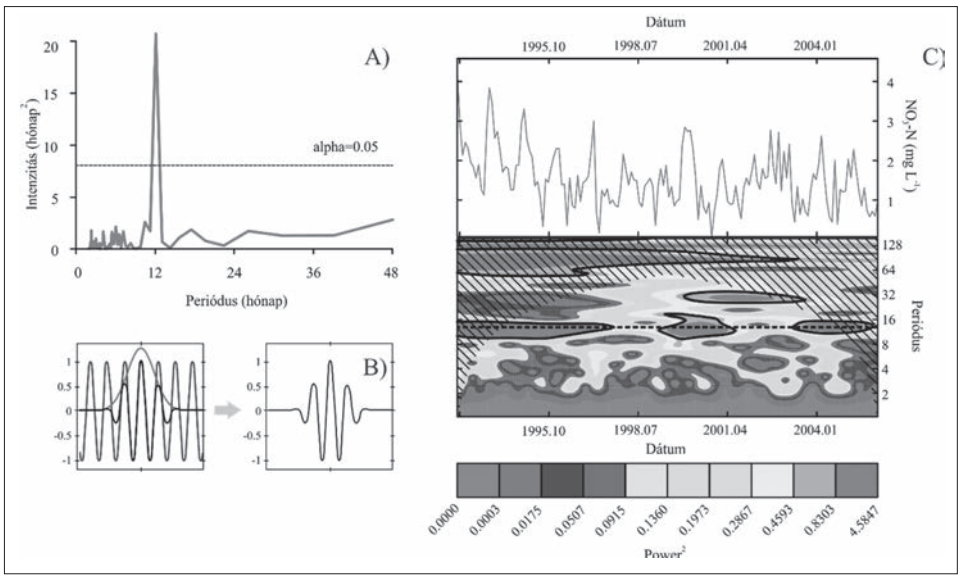

Figure 7. Identification of annual periodicity using a Lomb-Scargle periodogram for NO_3-N parameter

8. ábra *Példa a harmadfokú spline interpolációra egy megfelelő (A) és egy hibás (B) illesztés esetében*

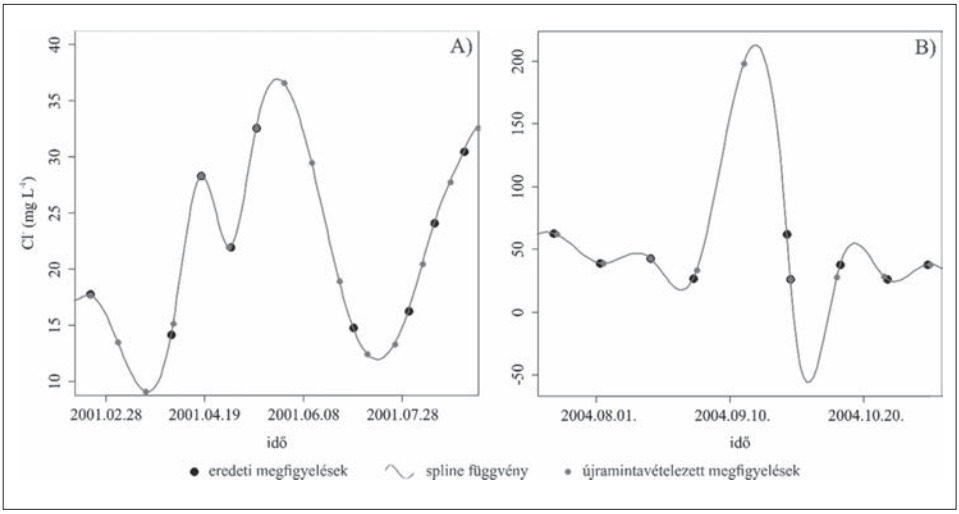

Figure 8. Example for a third degree spline interpolation for a correct (A) and incorrect (B) fit

Számos periodicitás vizsgálat megköveteli az időben ekvidisztáns idősort, azonban a leggyakrabban a rendelkezésre álló adatok ezt a feltételt nem teljesítik. Emiatt tehát legtöbbször szükséges az idősorok ekvidisztáns újramintavételezése. Ilyen esetekben különböző interpolációs módszerek alkalmazhatóak, amelyek segítségével minden időpillanatra becsülhetőek értékek, például harmadfokú spline interpoláció segítségével (*8. ábra*).

Bizonyos esetekben azonban az újramintavételezéssel, illetve az adatpótlás során, torzíthatjuk az adatokat, például abban az esetben, ha az eredeti megfigyelések között jelentős adathiány van. Ekkor az illesztett polinom „elhagyhatja" a mért adatok nagyságrendjét és az újramintavételezett megfigyelés kiugró értéket fog eredményezni. Ezért a spline illesztést minden mintavételi ponton, az összes változó esetében ellenőrizni kell. Az újramintavételelezés általában kevesebb hibával terhelt abban az esetben, ha az eredeti megfigyelések átlagos időközei közel vannak az újramintavételezett ekvidisztáns adatsor időközéhez.

A *főkomponens-analízis* (PCA) egy általánosan elterjedt többváltozós dimenziócsökkentő eljárás. Vízminőségi idősorok elemzése során is gyakran használt, amely alkalmas az adatokat alakító háttérfolyamatok azonosítására is (*Tabachnick mtsai,* 2001). A többváltozós adatelemző módszerek többségének, így a főkomponens-analízisnek is alkalmazhatósági feltétele, hogy a megfigyelések száma legyen nagyobb, mint a változók száma (*Füstös és mtsai*, 1986).

A PCA során a megfigyelt, korreláló változóinkon ortogonális transzformációt hajtunk végre, amelynek eredményeképpen korrelálatlan, új változókat kapunk. Ezen új változók (főkomponensek, PC_i) az eredeti változók (X) lineáris kombinációjaként számíthatók:

- $a_1, \ldots, a_k$ vektorok normáltak: $\|a_i\|_2 = 1 \; \forall \; i \in \{1, \ldots, k\}$
- $PC_1 = Xa_1$ legnagyobb mintavarianciával rendelkező változó.

- $PC_2 = Xa_2$ legnagyobb mintavarianciával rendelkező változó, amely az első főkomponenssel (PC_1) korrelálatlan.
- $PC_3 = Xa_3$ legnagyobb mintavarianciával rendelkező változó, amely az első két főkomponenssel ($PC_1, PC2$) korrelálatlan.
- stb.

A probléma megoldására létrehozhatunk egy A, ortonormált mátrixot, amelynek oszlopai a kovariancia mátrix normált sajátérték vektorait tartalmazza. Például, ha $A = [a_1, \ldots, a_k]$ akkor a_i az i-edik legnagyobb sajátértékhez tartozó sajátérték vektor. XA oszlopai tartalmazzák a főkomponenseket ($XA = [Xa_1, \ldots, Xa_k] = [PC_1, \ldots, PC_k]$).

Mivel az A mátrixban a sajátérték vektorok csökkenő sorrendben követik egymást, ezért a főkomponensek rendre a variancia egyre kisebb hányadát magyarázzák. A számított főkomponensek közül az első néhányat megtartva a változók száma csökkenthető viszonylag kis információveszteség mellett. Mivel az egyes változók gyakran különböző nagyságrendűek, ezért normált adatok használata javasolt. Az eredeti $X_1, \ldots, X_p$ változók és a $PC_1, \ldots, PC_p$ főkomponensek közötti $a_1, \ldots, a_p$ korrelációs együtthatók a főkomponens súlyok.

Amennyiben nem ismerjük a kovariancia mátrixot (és az esetek döntő részében nem ismerjük), akkor a tapasztalati szórásmátrix (vagy normált esetben a tapasztalati korrelációs mátrix) sajátértékeit és sajátvektorait használjuk fel. Ezeknek a becsült sajátvektoroknak a segítségével számoljuk ki a mintaelemek (becsült) főkomponenseit.

A kapott súlyok klaszteranalízissel csoportosíthatók (*Xu és mtsai,* 2009), de ebben az esetben minden főkomponens azonos súllyal kerül felhasználásra. Amennyiben csak az első néhány főkomponens alapján osztályozunk, akkor az adatok teljes varianciájának csupán egy részét vesszük figyelembe (*Vega és mtsai,* 1998).

A félvariogram a geostatisztika alapfüggvénye, mely a következőképpen írható le (*Füst,* 2004; *Molnár* és *Füst,* 2002; *Molnár* és *mtsai.* 2010). Jelölje $Z(x)$ valamely paraméter egy adott x helyen, $Z(x+h)$ pedig-től h (térbeli vagy időbeli) távolságra lévő értékét. A $Z(x)$ és $Z(x+h)$ értékek különbségének szórásnégyzete:

$$D^2[Z(x) - Z(x+h)] = D^2[Z(x)] + D^2[Z(x+h)] - 2COV[Z(x), Z(x+h)]$$

Azonos populációba (stacionárius idősorból vagy mezőből származó) tartozó minták esetén a fenti szórásnégyzet nem függ x-től, így

$$D^2[Z(x) - Z(x+h)] = 2D^2[Z(x)] - 2COV[Z(x), Z(x+h)] = 2\gamma(h)$$

A függvényt a paraméter variogramjának, a $\gamma(h)$ függvényt pedig félvariogramjának nevezzük.

Diszkrét minták esetén, ha az adatpárok száma N, az empirikus félvariogram számítására a Matheron-féle összefüggés szolgál:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(x_i) - Z(x_i + h)]^2$$

Röghatásnak (C_0) nevezzük a félvariogram $\gamma(h)$ határértékét 0-ban. Ennek nagysága egyrészt a vizsgált paraméter h távolságon bekövetkezett, mérhető nagyságú változásával, másrészt azzal magyarázható, hogy a mérés, illetve az elemzés során a mérési mód és műszer pontatlansága miatt véletlen jellegű hibát követünk el.

Hatástávolság (a) alatt azt a távolságot értjük, amelyen belül a minta még

hatást gyakorol környezetére. Ezen a távolságon túl a minták korrelálatlanok. A félvariogramon a hatástávolságot annak a pontnak az abszcisszája jelenti, amelynél a félvariogram függvény értéke állandósul. Ehhez a ponthoz tartozó érték az ordinátán a küszöbszint, amely végeredményben a röghatás (C_0) és a redukált küszöbszint (C_0) összege (*Kovács és mtsai,* 2012a). A variogram vizsgálatok eredményei felhasználhatók a térbeli és időbeli mintavételezés gyakoriság becsléséhez (*Füst,* 1997; *Füst,* 2011; *Füst és Geiger,* 2010; *Hatvani és mtsai*, 2012; 2014; *Kovács és mtsai,* 2005; 2011; 2012a).

IRODALOMJEGYZÉK

CCDA (2014): A CCDA szoftver 1.1 verziója és dokumentációja elérhető a http://cran.r-project.org/web/packages/ccda/ címen. Programozási nyelv: R, a program mérete: 9.11 KB

Davies, D.L. - Bouldin, D.W. (1979): A cluster separation measure. Pattern Analysis and Machine Intelligence, IEEE Transactions on Pattern Analysis and Machine Intelligence, 1. 224-227.

Dunn, J.C. (1973): A fuzzy relative of the ISODATA process and its use in detecting compact well–separated clusters. Journal of Cybernetics, 3. 32-57.

Everitt, B.S. – Landau, S. – Leese, M. – Stahl, D. (2011): Cluster Analysis, Fifth Edition, Wiley, UK, 346

Füst, A.(1997): Geostatisztika. Eötvös Kiadó, Budapest.

Füst, A. (2004): Short Course of Geostatistics (manuscript in Hungarian). Szent István Egyetem, Gépészmérnöki Kar, Informatika Tanszék, Gödöllő.

Füst, A. (2011): A természeti folyamatok monitoring hálózatainak tervezése és működtetése. . Bányászati Kohászati Lapok 144. 19-25.

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

Füstös, L. – Meszéna, G. – Simonné, M.N. (1986): Többváltozós adatelemzés. Akadémiai Kiadó, Budapest.

Hatvani, I.G. - Clement, A. - Kovács, J. - Székely Kovács, I. – Korponai, J. (2014): Assessing water-quality data: The relationship between the water quality amelioration of Lake Balaton and the construction of its mitigation wetland. J. Great Lakes Res., 40. 115-125.

Hatvani, I.G. - Kovács, J. – Korponai, J. (2012): Mintavételezési gyakoriság optimalizálása variogram függvénnyel a Kis–Balaton Vízvédelmi Rendszer példáján. Természetvédelmi Közlemények 18. 202-210.

Hatvani, I.G. - Kovács, J. - Székely Kovács, I. – Jakusch, P. – Korponai, J. (2011): Analysis of long-term water quality changes in the Kis-Balaton Water Protection System with time series- cluster analysis and Wilks' lambda distribution. Ecological Engineering, 37. 629-635.

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.

Kovács, J. (2007): Modern geomatematikai módszerek alkalmazása hidrogeológiai feladatok megoldására. Doktori (PhD) értekezés, Szegedi Tudományegyetem, Szeged.

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

Kovács, J. – Erőss, A. (2017): Statistically optimal grouping using combined cluster and discriminant analysis (CCDA) on a geochemical database of thermal karst waters in Budapest. Appl. Geochem., 84. 76-86.

Kovács, J. - Hatvani, I.G. - Korponai, J. - Székely Kovács, I. (2010): Morlet wavelet and autocorrelation analysis of long-term data series of the Kis-Balaton water protection system (KBWPS). Ecological Engineering, 36. 1469-1477.

Kovács, J. - Hatvani, I.G. - Kovács, I.S. - Jakusch, P. - Tanos, P. - Korponai, J. (2011): Key question of sampling frequency estimation during system calibration, on the example of the Kis-Balaton Water Protection System's data series. Georgikon for Agriculture: A Multidisciplinary J. Agricult. Sci., 14. 53-67.

Kovács, J. - Kovács, S. - Hatvani, I.G. - Magyar, N. - Tanos, P. - Korponai, J. - Blaschke, A. (2015): Spatial optimization of monitoring networks on the examples of a river, a lake-wetland system and a sub-surface water system. Water Res. Manag., 29. 5275-5294.

Kovács, J. - Kovács, S. - Magyar, N. - Tanos, P. - Hatvani, I.G. - Anda, A. (2014): Classification into homogeneous groups using combined cluster and discriminant analysis. Environ. Modelling and Software, 57. 52-59.

Kovács, J. – Nagy, M. – Czauner, B. - Székely Kovács, I. – Borsodi, A.K. - Hatvani, I.G. (2012b): Delimiting sub-areas in water bodies using multivariate data analysis on the example of Lake Balaton (W Hungary). J. Environ. Manag., 110. 151-158.

Kovács, J. - Reskóné Nagy, M. - Kovácsné Székely, I. (2005): Mintavételezés gyakoriságának vizsgálata tér–statisztikai függvénnyel a Velencei–tó példáján. Hidrológiai Közlöny 85. 68-71.

Kovács, J. - Tanos, P. - Korponai, J. - Székely Kovács, I. - Godár, K. - Godár-Sőregi, K. - Hatvani, I. G. (2012a): Analysis of water quality data for scientists. In: Voudouris, K., Voutsa, D. (Eds.), Water quality monitoring and assessment. InTech, Rijeka, 65-94.

Lomb, N.R. (1976): Least-squares frequency analysis of unequally spaced data. Astrophysics and Space Science, 39. 447-462.

Marquardt, D. (1963): An Algorithm for Least-Squares Estimation of Nonlinear Parameters. SIAM Journal on Applied Mathematics, 11. 431-441.

McKenna Jr, J.E. (2003): An enhanced cluster analysis program with bootstrap significance testing for ecological community analysis. Environmental Modelling & Software, 18. 205-220.

Molnár, S. - Füst*, A.* (2002): Környezetinformatikai modellek I. Szent István Egyetem, Gödöllő.

Molnár, S. - Füst*, A. - Szidarovszky, F. - Molnár, M.* (2010): Környezetinformatikai modellek II. Szent István Egyetem, Gödöllő.

Molnár, M. (2019): Big Data in Geosciences –Challenges and Novelties, GEOMates 2019

Rowan, J.- Greig, S.- Armstrong, C.- Smith, D.- Tierney, D. (2012): Development of a classification and decision–support tool for assessing lake hydromorphology. Environmental Modelling & Software, 36. 86-98.

Russel, S. – Norvig, P. (2005): Mesterséges Intelligencia: Modern megközelítésben, Panem

Scargle, J.D. (1982): Studies in astronomical time series analysis. II–Statistical aspects of spectral analysis of unevenly spaced data. Astrophys. J., 263. 835-853.

Tabachnick, B. G. – Fidell, L. S.- Osterlind, S. J. (2001): Using multivariate statistics. Allyn & Bacon.

Vega, M. – Pardo, R. – Barrado, E. – Debán, L. (1998): Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis. Water Research 32. 3581-3592.

Wilks, S.S. (1932): Certain generalizations in the analysis of variance. Biometrika, 24. 471-494.

Xu, H. - Yang, L. - Zhao, G. - Jiao, J. – Yin S. - Liu Z.(2009): Anthropogenic Impact on Surface Water Quality in Taihu Lake Region, China. Pedosphere 19. 765-778.

Yang, Q.- Shao, J.- Scholz, M.- Boehm, C.- Plant, C. (2012): Multilabel classification models for sustainable flood retention basins. Environ. Modelling & Software, 32. 27-36.

Érkezett: 2019. július

Szerző címe: *Molnár S.*
Szent István Egyetem, Gépészmérnöki Kar
Author's address: Szent István University, Faculty of Mechanical Engineering
H-2100 Gödöllő, Páter Károly utca 4.
Molnar.Sandor@gek.szie.hu