

MISKOLCI EGYETEM



**MÓDSZERFEJLESZTÉSEK HETEROGÉN MÉRNÖKI
ADATHALMAZOKHOZ**

PhD ÉRTEKEZÉS TÉZISEI

KÉSZÍTETTE:

Kilik Roland

Okl. mérnök-informatikus

AKI DOKTORI FOKOZAT ELNYERÉSÉRE PÁLYÁZIK

HATVANY JÓZSEF INFORMATIKAI TUDOMÁNYOK DOKTORI ISKOLA

TERMELÉSINFORMATIKA TÉMATERÜLET

MÉRÉS- ÉS IRÁNYÍTÁSTECHNIKAI INFORMÁCIÓS RENDSZEREK TÉMACSOPORT

Doktori Iskola Vezetője

Prof. Dr. Kovács László

egyetemi tanár, DSc

Témavezetők

Prof. Dr. Szabó Norbert Péter

egyetemi tanár, DSc

Dr. Varga Attila Károly

egyetemi docens

Miskolc

2026

Kilik Roland

MÓDSZERFEJLESZTÉSEK HETEROGÉN MÉRNÖKI ADATHALMAZOKHOZ
Ph.D. értekezés tézisei

Miskolc
2026

Tartalom

Bevezetés.....	7
1. Osztályozási eljárás virtuális ügynökök mozgásának generálásához	8
1.1 Bevezetés, irodalmi áttekintés	8
1.2 Az osztályozás.....	9
1.3 Eredmények és a kialakított rendszer vizsgálata.....	10
1.3.1 A pontosság és az újdonság jellemzése.....	10
1.3.2 Az egész osztályra kiterjedő erős korreláció és annak következménye.....	12
1.4 Outlierek észlelése.....	13
1.5 Összegzés, alkalmazhatóság.....	15
2. Hisztogramszűrésen alapuló súlyozott medián zajszűrési eljárás.....	17
2.1 Bevezetés, irodalmi áttekintés	17
2.2 A súlyozott medián (WM) módszer	17
2.3 Összegzés, alkalmazhatóság.....	18
3. Súlyozott mediánszűrésen és leggyakoribb érték módszerén alapuló klaszterelemzési eljárás ..	20
3.1 Bevezetés, irodalmi áttekintés	20
3.2 A Steiner-féle MFV eljárásen alapuló klaszteranalízis.....	21
3.3 A módosított WMFV klaszterelemzési eljárás.....	22
3.4 A WMFV klaszterelemzés alkalmazása.....	23
3.5 A WMFV eljárás további alkalmazása, validációja, Hurst exponensen alapuló mérőszám bevezetése.....	24
3.6 A WMFV klaszterelemzési eljárás súlyparaméterének adaptívva tétele	26
3.7 Összegzés, alkalmazhatóság.....	27
4. Új tudományos eredmények.....	29
5. New scientific results.....	31
Összefoglalás	33
Summary	34
Hivatkozások.....	35
Saját publikációk az értekezés témakörében	41

Bevezetés

A mindennapjainkat és jövőnket is egyre inkább befolyásoló különböző rendszerek nem tudnak működni adatok nélkül, így az adatok feldolgozása, az információk kinyerése, az adathalmazok javítása talán fontosabb, mint valaha. Gondoljunk csak például arra, hogy a generatív mesterséges intelligencia megoldások is csak annyira lehetnek jók, amilyen minőségű adathalmazon tanították azokat, illetve amilyen mértékben robusztusak tudnak lenni a kieső értékekkel szemben. Nem csupán a világunkban keletkező napi adatmennyiség nő folyamatosan [1], de az adatokat felhasználó – például deep learning – megoldások száma és jelentősége is [14]. Így a személyes motiváción túl az adatok előszűrésének, osztályozásának, információkinyerésének fontossága nem csak további motivációt jelentett kutatásaimhoz, de meghatározza relevanciáját is.

Három fő területen ismertetek új tudományos eredményeket, melyek alapvetően a jel- és adatfeldolgozás témaköréhez tartoznak. Az első egy újonnan kifejlesztett osztályozási eljárás, melyet egy hisztogramalapú súlyozott medián zajsűrő eljárás, végül pedig az ezen alapuló robusztus klaszterelemzési eljárás követ. Mindegyik eljárás alapját az adott adatrendszerek előállítását követően azok manuális elemzése, heurisztikák, majd a kezdeti eredmények, észrevételek felhasználásával megalkotott szabályszerűségek adták, amelyek az iteratív módszerfejlesztések kiindulópontjai voltak. Mindhárom módszert illetően közös kiindulópont és egyben munkahipotézis volt az adatok előszűrése olyan értelemben, hogy az osztályozási eljárásnál is kiinduló feltételezésem volt, hogy jobb eredményt produkáló módszer alkotható meg, ha elérhető mennyiségek bizonyos halmazainak adott határok közötti mintáira kerülnek meghatározásra az osztályt alkotó szabályok (kvázi előszűrést alkalmazva), és ugyanígy előszűrést alkalmaztam – más, hisztogramos formában – a zajsűrésre, majd ezt alapul vevően a klaszterezésre megalkotott módszer esetében is.

1. Osztályozási eljárás virtuális ügynökök mozgásának generálásához

1.1 Bevezetés, irodalmi áttekintés

Az emberhez hasonló ágensek a mesterséges intelligencia előretörésével egyre gyakoribbá válnak. Ezen ágensek hasznossága azonban nagymértékben függ mozgásuk természetességétől.

Számos tanulmány kimutatta, hogy egy fejmozgást szintetizáló rendszer létrehozható különböző hanginformációt bemenetként használva, miközben kevés, a vizuális információkból kinyert referenciát vesz ehhez alapul [63]. Korábbi kutatások [19] megerősítették, hogy a cél nem az eredetihez minél közelebbi fejmozgások előállítása abban az értelemben, hogy a mozgás természetességének mértéke nem lehet a referencia trajektóriájához való hasonlóság mértéke. Ez a megfigyelés egybeesik a különböző alanyok neurális hálóval szintetizált fejmozgásairól kapott kezdeti minőségi visszajelzésekkel. [S1] A kezdeti vizsgálat során ugyanis egyértelműen bebizonyosodott, hogy a generált fejmozgás trajektóriájának az eredeti (lekövetett) mozgás trajektóriájához való hasonlóságával szemben az eredetihez közeli maximális fejmozgás-amplitúdó eredményez természetesebb benyomást. Ez a megállapítás vezetett a fejmozgás amplitúdójának hanginformációból történő meghatározására irányuló vizsgálataimhoz, azzal a céllal, hogy a becsült amplitúdóérték felhasználható legyen neurális hálózat által generált fejmozgások automatikus amplitúdó-korrekciójához, ráadásul azzal az elvárással, hogy a folyamat során ne legyen szükség vizuális információra.

Az automatikus mozgásgenerálásról szóló, korábbi eredményeket, illetve módszerek korlátait összehasonlító fontos eredményeket közöltek Zhou és szerzőtársai [66]. Ebben a szerzők rámutatnak, hogy a virtuális beszélő fejek élethű mozgásának generálásában az egyik fő probléma az, hogy a nagyon alapvető hanginformációs mennyiségek önmagukban nem korrelálnak kellőképpen a mozgással, így a csak hang alapján történő élethű (pl. megfelelő amplitúdójú) fejmozgás generálása nagy kihívást jelentő feladat.

A fenti szerzők hivatkoznak Hyeongwoo Kim és szerzőtársai [30] munkájára is, ahol a szerzők azt vizsgálták, hogy a mozgásgenerálás minőségének nagymértékű romlását okozza, ha a becsült/generált mozgástartomány kívül esik az adott mozgásgeneráló rendszer tanítási mintáinak mozgástartományán. Ez is megerősíti a fejmozgás amplitúdójának a hangból történő becsülésének – és természetesen a mozgásgenerálás során történő felhasználásának – fontosságát.

Az általam kifejlesztett osztályozási módszer célja az emberszerű ágensek fejmozgásainak természetességének növelése, mégpedig a valós fejmozgás mértékének pusztán a hanginformációkból történő becsülésével. Főbb jellemzői, hogy az alanyok maximális függőleges fejmozgását csak a hangmintákból határozza meg, általános alanyok esetében is működik, pontossága és hatékonysága további osztályok meghatározásával javítható, és bármely fejmozgás generáló módszer kimenetének természetességét javíthatja egy utólagos amplitúdóskálázással. Ezen fő előnyök, illetve alkalmazási lehetőség mellett a fejmozgás-tartományok és az osztályozásban használt származtatott hanginformációs mennyiségek közötti magas korrelációs érték

továbbfejlesztési lehetőségek kiindulópontja lehet [S2]. Mindemellett a módszer a későbbiekben esetlegesen alkalmas lehet annak a vizsgálatára is, hogy egy videóban szereplő beszélő valódi-e, vagy azt (annak mozgását) valamilyen generatív mesterséges intelligencia segítségével állították-e elő szintetikususan. Ennek, mint lehetséges alkalmazási lehetőségnek a jelentősége vélhetően a közeljövőben igen nagy lesz.

Módszerem megalkotását követően megvizsgáltam eljárásomat illetően az osztályszintű összefüggéseket, illetve az osztályokban található változók számát és az osztályok esetleges összevonhatóságát is. Ilyen értelemben tehát a módszer által kapott osztályszerkezeti optimalitásnak a vizsgálata is megtörtént, amely szintén bemutatásra kerül. Az értekezés tartalmaz egy újszerű outlier-detektálási eljárást is, amelynek meghatározó szerepe van az osztályozás hatékonyságának további javításában.

1.2 Az osztályozás

A doktori kutatásaim során kifejlesztett osztályozási módszer hiperterek segítségével osztja fel a jellemzőteret, ami a szakértői rendszerek jól ismert megközelítése. A figyelembe vett pontok halmazát a jellemzőtérben ekkor produkciós szabályok meghatározásával határozhatjuk meg. A szabályok ekkor valójában a konjunkciós operátorok által kifejezett feltérek metszéspontját adják meg.

Az osztályozást alkotó bemenetekhez olyan videóanyagok kerültek kiválasztásra, ahol az alany természetesen nyilvánul meg, érzékelők nélkül (szintén a természetesség, ill. a mintahalmaz könnyű bővíthetősége okán), illetve egy alany maximum 5 mintában szerepelhetett (az alanytól függő szabályszerűsége elkerülése érdekében).

Az osztályozás hangbemenetei részben közvetlenül kinyert értékek voltak (például az intenzitás gyors Fourier (FFT)-transzformációjának varianciája, autokorrelációs pitch átlaga, amplitúdó különbségek átlaga, intenzitás különbségek átlaga, keresztkorrelációs pitch átlaga, az amplitúdó varianciája, intenzitás varianciája. Másrészt más hanginformációs értékeket empirikus alapon vezettem le az előbbiekből. Így előállt alap- és a származtatott hanginformációs értékek száma összesen 29 volt. A 29 hanginformációs tulajdonsággal a teljes mintaadatbázisból osztályalkotási szabályokat készítettem, ahol az egy osztályhoz tartozó tulajdonságok értékei (jellemzően osztályonként 4-5 a rendelkezésre álló 29 tulajdonságból) egy minimum-maximum tartományban vannak. Az egy adott osztályba tartozó minták közös jellemzője az is, hogy maximális mozgásamplitúdójuk egy adott tartományon belül van. Összesen 74 osztályt határoztam meg ezzel a módszerrel

Legyen P_j egy tetszőleges tulajdonság, $j \in N$. Használjunk összehasonlító operátorokat állítások definiálására a megfelelő hipersíkot érintően. Például $P_j < 20$ azt jelenti, hogy az állítás igaz, ha $P_j < 20$. Definiáljunk egy produkciós szabályt, mint ezen állítások logikai formuláját, $f: P \rightarrow \{igaz, hamis\}$. A produkciós szabály a becsült osztályokra is utal. Ezért a szabályaink halmazát úgy formalizálhatjuk, mint (f_i, c_i) párok R halmazát, ahol a f_i egy logikai formula az audiójellemzők tartományán, és c_i a megfelelő osztály. Az osztályozási folyamat során egy minta több osztályba is tartozhat.

A 0-15 pixel, 0-25 pixel és 25 pixelnél nagyobb mozgási amplitúdók eseteire 3 fő mozgáscsoportot határoztam meg, amelyek a korábban leírt 74 osztályt tartalmazzák.

Egy alosztályba tartozás ténye becslést ad a minta maximális függőleges fejmozgására, mivel minden alosztály a három nagy mozgáscsoport egyikéhez tartozik.

Természetesen minden osztályozási módszer érzékenyen reagál a kieső értékekre. Jelen esetben egy mintát akkor tekintek outliernek, ha az bár megfelel egy adott alosztály szabályának, azonban a mozgásintervallum tekintetében kiugró (azaz a mozgás tartománya nem tartozik annak a főcsoportnak a mozgásmértékéhez, amelyhez az alosztály tartozik). A szabályokban szereplő, az adathalmaz megfigyelésével meghatározott konstansok értékeit gépi úton korrigáltam a szabálybázist alkotó mintákra azzal a kívánnalommal, hogy értékükkel a lehető legkisebb számú osztályozási hiba (outlier) legyen a halmazon [S3].

A fentiek szerint kialakított csoportok mozgásértékein egy skálázást hajtottam végre. A 0-15 pixeles főcsoportba tartozó osztályba tartozó mintákat 7 pixelre, míg a csak a 0-25 pixeles főcsoportba tartozó osztályba tartozó mintákat 13 pixelre skáláztam (az osztályalkotó minták mozgásának mediánja). A 25 pixelnél nagyobb mozgáscsoportba tartozó minták 36 pixelre lettek skálázva. Ha egy minta egy olyan osztályba tartozik, amely a 0-15 pixeles mozgásméretű főcsoportba tartozik, de emellett egy olyan osztályba is amely a 0-25 pixeles főcsoportba tartozik, akkor a szűkebb csoportot tekintettem relevánsnak (mivel értelemszerűen ez adja a pontosabb mozgásmérték becslést).

1.3 Eredmények és a kialakított rendszer vizsgálata

1.3.1 A pontosság és az újdonság jellemzése

Néhány korai fejmozgás-szintetizáló módszer 1-2 alany fejmozgás mintáit klaszterezte [11], és minden egyes klaszterhez Rejtett Markov Modellt tanítottak[21],[20]. Korábbi kutatások azonban kimutatták, hogy a mozgásgenerálás első lépéseként a képkockák szerinti klaszterezés nem a legjobb megközelítés[35]. Az általam megvalósított osztályozási megoldás – a fentiek szerint – nem ilyen módon alkalmazza az osztályozást. Fontos fejlesztési eredmény emellett az is, hogy jellemzői okán bármely fejmozgás-generálási eljárás kiegészíthető az osztályozási eljárással (végső lépésként), amennyiben a generált mozgásamplitúdók skálázhatók egy, az osztályozási módszerrel számított konstans értékkel.

Voltak más kísérletek is (bár nem ezzel a céllal) a függőleges fejmozgás amplitúdójának implicit becslésére. Ezek pontossága a generált mozgás amplitúdójának és a referenciának az arányából számítható[63],[19]. Közös jellemzőjük azonban, hogy a generált fejmozgásokat néhány alany segítségével konstruálták, és hogy a szintetizált mozgás amplitúdója és a referencia közötti szorzó egy mondatos intervallumban a bemutatott példákban is 2-10 között változott. Ez az arány akár a 60-at is elérheti [24], és Matthews és mtsai. legújabb eredménye [35] példájában 7 körül adódott. Fenti szerzők jövőbeli célja, hogy módszerük általánosítási lehetőségeit keressék, és a korpuszokon kívüli beszélőket is becslést adjanak. Egyes módszereknek a becslés kezdetétől [19] vagy a fej korábbi pozíciójából [21] származó referencia vizuális információra is szükségük van. Zhou és társai [66] egy olyan mozgásgenerálási módszert építettek fel, amely különbözik a fent említettektől, mivel csak egy vizuális képkockát használ, azonban csak az ajak (és környéke) mozgásgenerálására koncentrálnak, és egyéb korlátozásokkal rendelkeznek. Ezzel szemben az általam javasolt osztályozási

módszer különböző alanyokra épül, és nem használ vizuális információt a fejmozgás becsléséhez. A módszer általánosan alkalmazható voltához hozzájárul a fenti jellemző, miszerint a rendszer különböző alanyokra épül fel, és tesztminták által bizonyítottan a korpuszon kívüli alanyok esetében is működik azokban az esetekben, amelyekben a hangérték beleesik valamely osztályba. Kivételt képeznek ez alól a magas érzelmi töltésű esetek – ez a módszer továbbfejlesztésének egy potenciális iránya lesz. A megoldás jelen változatának legnagyobb becslési hibája – helyes osztályba sorolás esetén – a más, ismertetett módszerekkel összemérhető 8-10-es szorzót érheti el legrosszabb esetben, aláhúzendó viszont, hogy mindezt – a többi megoldással ellentétben – vizuális információ felhasználása nélkül elérve. Ezen említett maximális hibaszorzó abban az esetben áll fenn, ha a mondatban a legnagyobb amplitúdó minimális (1-2 pixel), miközben a minta csak a 0-25 pixeles mozgásintervallum-csoport egyik osztályába sorolható, miközben a 0-15 pixeles mozgásintervallum-csoport egyik osztályába sem. Az amplitúdók ez esetben ugyanis úgy kerülnek skálázásra, hogy a maximális amplitúdó érték 13-ra adódjon (a 0-25 pixeles csoport miatt). Ilyen esetek a minták kevesebb, mint 12%-ában fordulnak elő. Továbbá az esetarány tovább csökkenthető, ha több osztályt határozunk meg a 0-15 pixeles (vagy még szűkebb) tartományban – csökkentve annak az esélyét, hogy egy mindössze néhány pixeles mozgást mutató minta csak a 0-25 pixeles intervallumcsoportba kerüljön.

Egy további eredmény a területen [29] egy olyan kutatócsoporthoz köthető, akik egyrészt felhasználnak szájmozgást (mint az előző példában), illetve egyes kiválasztott alanyok fejmozgását is, és ezekből általánosítanak. Ez több szempontból természetes mozgást eredményez, azonban például a tanítóminták közötti nagy fejmozgások (és feltehetően egyéb speciális jellemzők) egyes esetekben természetellenesen nagy mozgást eredményeznek. Megjegyzendő, hogy ez javítható lenne az általam megvalósított osztályozási megoldás korrekciós fokozatként történő alkalmazásával azon módszer végső fázisaként.

A fentieket összegezve, módszeremben a generált mozgás és a referencia amplitúdójának aránya összehasonlítható a legújabb más kutatásokkal, miközben – azokkal ellentétben – nem használ vizuális információt (amely nagyban korlátozná felhasználhatóságát), az alanyok tágabb körén működőképes, továbbá a megoldás nem korlátozódik a jelenlegi állapotára, hanem új osztályok meghatározásával továbbfejleszhető.

A vizuális információ nélküli mozgásgenerálási megoldások közül egyik ismert eljárás sem mutat a fentiekkel összemérhető pontosságot ill. hibaaarányt. Ezek esetében a túlzott amplitúdó mellett szintén tipikus kimeneti hiba, hogy adott alanyok esetében nullához közeli amplitúdót generálnak[66], [33]. A mozgásgenerálás mindkét típusú hibája csökkenthető az általam javasolt osztályozási módszerrel akár ezen más eljárásokhoz csatoltan is, az adott megoldás kiegészítő fázisaként végzett automatikus amplitúdó korrekciós moduljaként alkalmazva. Emellett a vizuális információt használó megoldások kimenetének javítására is ugyanígy felhasználható.

Módszerem egyéb továbbfejlesztési lehetőségei például, hogy amennyiben egy minta a 74-ből több osztály szabályainak is megfelel, akkor eredményként megadhatjuk ezen osztályok legrövidebb mozgásintervallumának közepét. Ezzel a módosítással a 0-25 pixeles mozgásintervallum-csoportba tartozó osztályok mintáinak 14%-ánál, a 25

pixelnél nagyobb mozgásintervallum-csoportba tartozó osztályok mintáinak 42%-ánál a lehetséges mozgásintervallum tovább zsugorítható, mindkét esetben több, mint 12 pixellel. Bár ez a változat egy plusz szükséges lépést ad az amplitúdóskálázáshoz, nagyobb pontosságot eredményez. Nem célszerű azonban az osztályok összevonása (osztálypárok összevonását illetően ugyanis a párosok egyike esetében sem mutatták mérőszámok – például GAP – az összevont esetet ideálisnak, ráadásul ez értelemszerűen a pontosság csökkenését eredményezné), illetve dimenziócsökkentés vizsgált esetei sem (faktoranalízis jellemzően 1-1 osztályváltozó esetében ad 0,1 alatti specifikus variancia értéket, ill. vizsgált nemlineáris eljárások – GDA, LLC – kimenetein nem jött létre a fentiekkel összemérhető eredményű osztályozási eredmény).

1.3.2 Az egész osztályra kiterjedő erős korreláció és annak következménye

Az osztályokat alkotó mintákra felírható egy az osztályra jellemző paraméterarány, ahol az arány értéke és a mozgás közötti lineáris korreláció az osztály elemeire jellemzően 0,9 feletti. Ennek a korrelációs együttható értéknek az átlaga 42 osztályra vonatkozóan 0,87, a maximuma pedig 0,99 volt vizsgálataimban.

Ezen felüli további tapasztalat, hogy ugyanazokat a tulajdonságokat a produkciós szabályrendszer két különálló osztályának egyesítése után vizsgálva a korrelációs érték 0,03-0,1 közé esik, és az adatbázis egészét tekintve jellemzően 0,05 értékű. Ez a jellemző pedig a jelenlegi osztályozás és az azon belüli egyes osztályok érvényességét támasztja alá.

Ben Youssef és szerzőtársai [5] CCA (Canonical Correlation Analysis) segítségével 0,2 körüli korrelációt ért el a beszéd és a fejmozgás jellemzői között a kisszámú alany esetében. Egyes kutatók [24] korábban azt állították, hogy általános jelleggel csak 0,07-0,08 körüli korrelációs érték mutatható ki hang és fejmozgás között. Yehia és társai [63] ezzel szemben erős korrelációról számoltak be az intenzitás és a fejmozgás között, azonban Ben Youssef és munkatársaihoz hasonlóan csak kisszámú alany esetében, ráadásul a kapcsolat erőssége mondatról mondatra változott, és a fejmozgás becslése az intenzitásból rosszabb eredményeket mutatott, mint az intenzitás becslése a mozgásból (átlagosan 0,37 korrelációs együttható). Ezeket az eredményeket figyelembe véve, összegzésként megállapítható, hogy a beszéd- és mozgásjellemzők közötti képkockánkénti globális korreláció gyenge.

A korábbi szerzők korrelációs módszere (CCA - kanonikus korrelációs elemzés) eltér az esetemben használt Pearson-féle korrelációs mátrix számításától, ráadásul én a maximális mozgás amplitúdójával összevetésben vizsgáltam az adatokat. A gyenge „globális” (azaz alapvető hangjellemzőkre a teljes adatbázisra igaz) korrelációra vonatkozó megállapítás azonban egybevág az általam kapott eredményekkel, hiszen az én tapasztalataim is azt mutatták, hogy az adatbázis-szintű általános kapcsolat az alapvető beszédjellemzők és a mozgásamplitúdó között gyenge (jellemzően 0,05). Ezzel szemben azonban én származtatott beszédinformációk felhasználásával osztályszintű erős korrelációt találtam, ezzel kimutatva, hogy meghatározhatók olyan származtatott hangjellemzők, amelyekkel adott csoportba tartozó minták esetében ezek, és a minták mozgása között erős lineáris korrelációs kapcsolat van (legalábbis a maximális mozgásmértékre vonatkozóan).

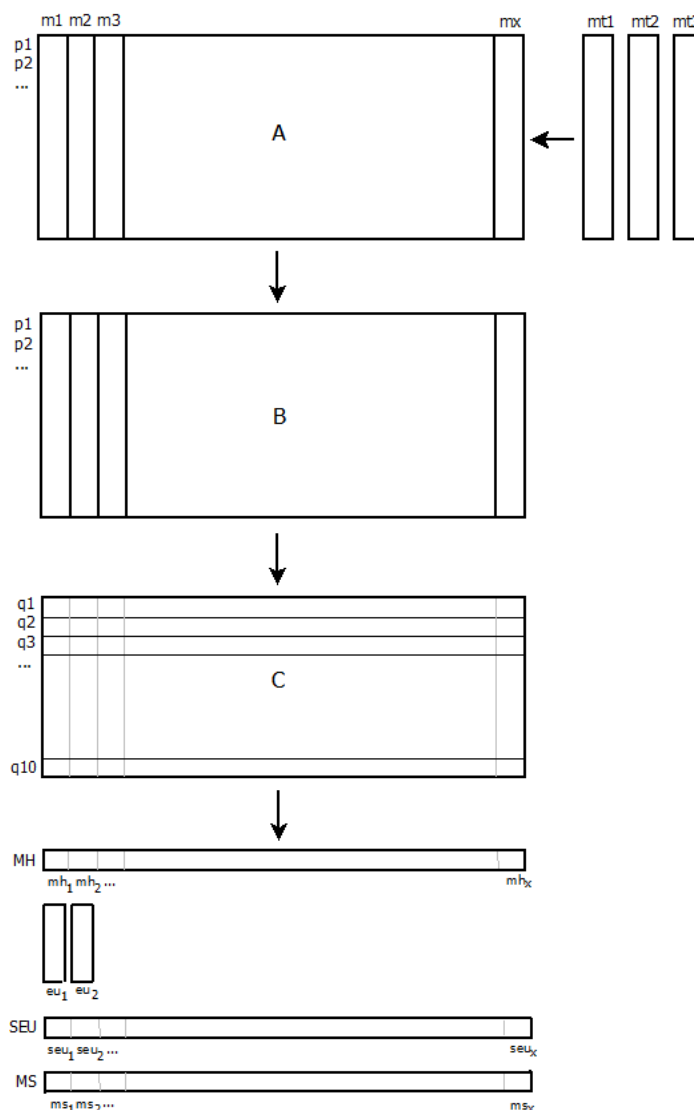
Mindezen felül, ez megerősíti az általam létrehozott/alkalmazott, származtatott hanginformációs jellemzők és az ezeken alapuló osztályok létjogosultságát is. Hozzáteve, hogy további osztályok meghatározása a rendszer jellemzőiből adódóan tovább növelné a mozgásamplitúdó-becslési pontosságot nem pusztán a korrelációs kapcsolat fent ismertetett felhasználásából adódó, de az ettől független, csak az osztályba tartozáson alapuló becslési módszer esetében is. Megjegyzendő itt, hogy ezzel ellentétben az osztályok összevonása nyilvánvalóan csökkentené azt.

1. Tézis Kidolgoztam egy kizárólag hanginformációkat felhasználó, szabályalapú osztályozási módszert, amely képes a fejmozgás mértékét kellő pontossággal becsülni. A módszer a betanításban nem szereplő alanyok mozgásának tartományát is képes becsülni. Bebizonyítottam, hogy képezhetők olyan hanginformációs paraméterek, amelyek mind a mozgásmérték osztályozását, mind pedig a hang és a fejmozgás közötti korrelációt tekintve jobb eredményt adnak, mint a hagyományosan alkalmazott hanginformációs jellemzők. Igazoltam, hogy az általam megalkotott osztályozási eljárás jelen struktúrájában megfelelő megoldást ad. A módszer az általam vizsgált esetekben jobb megoldást ad, mint a dimenziócsökkentéssel támogatott változat, illetve a gyakorlatban hagyományosan alkalmazott eljárások.

Kapcsolódó saját publikációk: [S1, S2, S3]

1.4 Outlierek észlelése

Egy minta akkor tekinthető egy osztályon belül kiesőnek (outliernek), ha hanginformációs értékei megfelelnek az adott osztály szabályának, azonban mozgást tekintve nem tartozik az osztály mozgási intervallumába. Az ilyen minták esetében a rendszer téves mozgásamplitúdó-becslést adna, ezért automatikus kizárásuk az osztályszabályban nem szereplő jellemzőik alapján indokolt. E cél elérése és az eredmények további javítása érdekében egy automatikus outlier-detektálási mechanizmust alkottam meg (1. ábra).



1. ábra. A kiugró értékek felismerésének eljárása

Az első lépésben veszünk egy A osztálymátrixot, amelynek elemei m_1, m_2, \dots . Ezt kiegészítjük a mt_1, mt_2, \dots outlier detektálás-tanulási mintákkal, amelyek szintén megfelelnek az osztály szabályának. (Ezen tanítómintákból néhány outlier, néhány nem.) Így rendre egy B mátrixot konstruálunk minden A mátrixból, azt egy mt mintával kiegészítve. Ezekben a mátrixokban a p_1, p_2, \dots hanginformációs értékek (tulajdonságok) közül minden lehetséges kombinációban 10-et választunk ki, C mátrixokat megalkotva (a ténylegesen kiválasztott 10 tulajdonság elnevezése: q_1, q_2, \dots, q_{10}).

Az 1. ábrán, amely a kiugró értékek felismerésének eljárását ismerteti, mh_1 a Mahalanobis-távolság C mátrixban a $q_1:q_{10}$ hanginformációs tulajdonságvektorra m_1 oszlop esetében, ahol m_1 az első minta. Hasonlóképpen, mh_2 ugyanez a második minta esetében, stb. eu_1 az euklideszi távolság C -ben $q_1 \dots q_{10}$ első oszlop értékei (1. minta) és a többi minta azonos $q_1:q_{10}$ indexű értékei között. eu_2 ugyanez a mátrix második oszlopának értékei esetében, stb. seu_i pedig $std(eu_i)$ és $ms_i = mh_i/seu_i$.

A C mátrixban az ms_1, ms_2 elemek, amelyeket az eredeti osztály mátrixból kaptunk, és ms_x aktuális mintából kapott elem együttesen MS vektort eredményezik. Amennyiben

ms_x elem értéke MS minimuma vagy a maximuma, akkor az adott q_1, \dots, q_{10} tulajdonságazonosítókat megtartjuk. Ahhoz, hogy egy adott q_1, \dots, q_{10} tulajdonságazonosító kombináció (például $p_1, p_3, p_4, p_8, p_{10}, p_{13}, p_{19}, p_{20}, p_{22}, p_{23}$) érvényesnek legyen tekinthető, két feltételnek kell teljesülnie. Az első, hogy több osztályra legalább 4 outlierre legyen igaz, hogy az ms_x elem (amelyet az aktuális elemhez számolunk q_1, \dots, q_{10} kombinációval), a minimum vagy a maximum értéket képviseli a saját osztálya által előállított MS vektorban. A második feltétel az, hogy ugyanaz az ms_x ne legyen egyidejűleg minimum vagy a maximum egyik ugyanazon q_1, \dots, q_{10} tulajdonságvektorral rendelkező osztályban sem, ahol az adott minta nem outlier. Ha a két fenti feltétel egyidejűleg teljesül, és így egy q_1, \dots, q_{10} kombináció érvényes, akkor minden egyes MS vektorra, amelyet ebből egy jövőbeli vizsgálati mintára és az osztály alapmintáira számítunk, a vizsgált minta outliernek tekinthető, ha az ms_x elem minimális vagy maximális MS -ben.

A módszer még azzal az – eredetinel szigorúbb – követelménnyel is, hogy minden tulajdonságkombinációnak legalább 15 outlierben kell előfordulnia (miközben nem fordulhat elő egyetlen nem kiugróban sem), minden outliert kimutatni volt képes a vizsgált esetekben. Megvizsgálásra kerültek 10-től eltérő változószámok, ill. $ms_i = (mh_i - seu_i)/mh_i$ nem pedig $ms_i = mh_i/seu_i$ vektorszámítási formula is, ezek azonban a kiindulóval szemben nem bizonyultak ideális megoldásnak.

1.5 Összegzés, alkalmazhatóság

A fentiekben a függőleges fejmozgás amplitúdójának hanginformációkból történő osztályozására szolgáló módszert mutattam be. A módszer előnye, hogy – más megközelítésekkel ellentétben – nem igényel előzetes mozgásmintákat az alanytól, és a létrehozás jellegéből adódóan alanyok tág körére is működik.

Az osztályozási megoldás megbízható becslést ad a maximális mozgás amplitúdójára pusztán a hang alapján. Ez az arány – az előtanításon alapuló módszerekkel ellentétben – tovább növelhető a tér nem lefedett részeit tartalmazó új osztályok definiálásával, illetve a becslés pontossága növelhető kisebb mozgásintervallum-osztályok megalkotásával. Azokra az esetekre, ahol az osztályozási módszer jelenleg képes becslést adni, a megoldás 87 százalékban jó becslést ad, mindennemű outlier detektálás nélkül. A módszer javításának céljával kifejlesztésre és bemutatásra került általam egy olyan outlier detektálási eljárás is, amely a fent ismertetettek szerint ezt tovább javítja. (Segítségével a helyes becslési arány akár 100 százalékra is növelhető vizsgálataim szerint.)

Az értekezésben ismertettem az általam kidolgozott eljárást, annak használatához szükséges hanginformáció kinyerését, az osztályok létrehozásának módszerét és a csoportok tulajdonságait, felhívva a figyelmet az osztályokon belül elérhető magas hang-mozgás korrelációs értékekre is, amelyek további lehetőségeket kínálnak. A fentiekben túl a módszer más megoldásokkal való összehasonlítását, hatékonyságát és az eljárás fejlesztéseit is tárgyaltam a dolgozat előbbi fejezetiben.

Összességében megállapítható, hogy a javasolt osztályozási eljárás az esetek jelentős részében javíthatja a fejmozgás-generáló módszerek kimeneteinek természetességét, azáltal, hogy lehetőséget ad az osztályozás alapján, végső fázisban történő amplitúdó

korrekcióra, mindezt olyan formában megvalósítva, amely nem igényel vizuális információt. A módszer segítségével tehát virtuális beszélő fejek, ill. robotok fejmozgás mértékei abba a tartományba skálázhatók, amelyben egy valós beszélő feje mozgott volna.

2. tézis Megalkottam a mozgásmérték osztályozási eljárás hatékonyságának növelése céljából egy a kieső értékeket hatékonyan kiszűrni képes, Mahalanobis és euklideszi távolság együttes használatával operáló outlier-detektálási eljárást.

Kapcsolódó saját publikáció: [S3]

2. Hisztogramszűrésen alapuló súlyozott medián zajszűrési eljárás

2.1 Bevezetés, irodalmi áttekintés

A Leggyakoribb Érték, vagy Most Frequent Value (MFV) módszer [46] egy nagy statisztikai hatásfokú becslési eljárás, mely robusztussága miatt jól alkalmazható zajos mérési adatrendszerek feldolgozásában. Az MFV módszert eredetileg a földtudományban fejlesztették ki, mely a geofizikai [12] [53] [48], hidrogeológiai [54] és számos más földtudományi terület [64] mellett előnyösen alkalmazható pl. a kozmológiában [64], ürgeodéziában [60], stb. Ilyen, különösen szélsőségesen zajos adatrendszerek a térinformatika területén is jelen vannak, beleértve a digitális domborzatmodellezést és a műholdas adatátvitelt. Hasonlóan elterjedt, de kevésbé robusztus és kifinomult technika a mediánszűrés [47][26].

A potenciális alkalmazási területek közül a digitális domborzati adatok zajcsökkentésére dolgoztam ki új, hisztogram-alapú robusztus szűrő módszert, amely (módosított változatában) a leggyakoribb érték módszerét is magában foglalja. Eljárásomat mind a mediánszűrés, mind az MFV eljárás által kapott eredményekkel összehasonlítottam.

Amennyiben az adatmátrixok adatpontjainak nagy százaléka kiugró zajjal szennyezett, a zajcsökkentési eljárás jobb eredményt adhat, mint a hagyományos mediánszűrés, ha a potenciálisan nagyobb eséllyel zajnak minősülő pontokat eltávolítjuk a bemeneti adathalmazból, mielőtt a mediánértéket kiszámítanánk. Ugyanezen mátrixok esetében azonban nagy valószínűséggel adódnak olyan adatrészletek, amelyeknél a szüretlen bemeneti adatok jobb eredményt adnak. Ezen két eset közötti automatikus súlyozás adja az alapját szűrési eljárásomnak [S4].

Az MFV módszer hatékonyságát az adja, hogy a leggyakoribb érték számításához szükséges súlyokat a módszer automatikusan szolgáltatja a bemenő adatokat terhelő zaj eloszlásától függetlenül (robusztusság). E miatt a legtöbb esetben különböző mértékű és eloszlású zaj esetén jobbnak bizonyul a hagyományos mediánszűrésnél.

Mivel az általam javasolt megoldás mindkét változatának megalkotásakor elsődleges cél a digitális domborzati adatok szórt zajának mozgóablakos eltávolítása volt [S6], így a fenti két bevett eljárást is ilyen adatok esetében hasonlítottam össze módszeremmel, a vizsgálatot négy különböző adatpont-arányt érintő zajterhelés esetében is bemutatva. Az összehasonlítás főként három mérőszám alapján történt, a legnagyobb hangsúlyt a L_1 norma-alapú távolságméτρika eredményekre téve.

2.2 A súlyozott medián (WM) módszer

A módszer segítségével a képmátrixokon végighaladva a mozgóablak minden egyes pozíciójában annak központi elemének korrigált értékét állítottam elő az ablak többi elemének súlyozott átlaga segítségével. Eljárásom alapváltozata elsősorban a nem nulla átlagú zajok kiküszöbölésére szolgál (amely például a mérőeszköz hibája, vagy a hosszú távú adatátvitel során jött létre).

A súlyozott átlagot futóablakkal minden egyes adatpontra kiszámítom. Ennek érdekében két független ablakszűkítési folyamat történik kezdeti lépésként, a súlyok kiszámítása

előtt. Ezeket a szűkített ablakokat az aktuális adatablaktól hozom létre, minden ablakpozícióban.

A folyamatban tehát lépésenként több ablakszűkítési folyamatot hajtottok végre, majd arányszámokat határozok meg az eredményhalmazok elemszámai és értéktartományai alapján. A nagyobb értéket az élesebb vágás jeleként értelmezve, kiválasztott szűkített adathalmazokon képezek medián, ill. átlagértékeket, melyeket köztes változóban használok fel. Ezekből skálázással és hangolásra került konstans szorzókkal részsúlyokat állítok elő.

A részsúlyokra külön megfigyeléseket is megalkottam, miszerint az egyes részsúlyok relatív nagy különbsége (például eltérő ablakszűkítések során, ugyanazon ablakpozícióban kapott átlagok nagy relatív különbsége) az azt valószínűsíti, hogy a hisztogramműveletek az aktuális ablakpozícióban torzítják az eredményt. Ezen megállapításokat felhasználtam a végső, az eredeti medián, illetve a szűkített ablakok közül kiválasztott ablak mediánja közötti végső, adott ablakpozícióbeli végső, súlyozott (korrigált) mediánérték kiszámítása során.

Módszerem zéróátlagú zajok csökkentésére hatékonyan alkalmas változatát is elkészítettem. A bemenő adatrendszert ez esetben is hisztogram alapján két lépésben előzetesen szűrtem, azonban az előbbiekhöz képest eltérő módon képezve a szűkített ablakok részsúlyokat alkotó arányszámait, illetve a súlyokban figyelembe vettem a fenti – különbségek arányára irányuló – megállapítást is, továbbá a súly értékét az is befolyásolja, hogy a szűrt és eredeti adatmátrix milyen mértékben homogén (gradiens alapján).

2.3 Összegzés, alkalmazhatóság

Az előzőekben ismertetett hisztogramalapú súlyozott medián fő célja az adatmátrixokban lévő kiugró zajok kiküszöbölése, különösen akkor, ha az adatpontok nagy százalékát szennyezi kiugró zaj.

A módszer hatékonyságát a digitális domborzati adatok zajcsökkentésére bizonyítottam több adatrendszeren, az eredeti (zajmentes) adatrendszerhez képesti RMSE, szórás és L_1 norma segítségével. Az eljárás alkalmazhatósága széleskörű, mind nulla, ill. nem nulla átlagú zajjal terhelt domborzati adatok minőségjavítására.

A vizsgált különböző zajamplitúdók és zajterhelési százalékok átlagában a súlyozott medián (WM) módszer 14-23%-kal jobb teljesítményt nyújtott a standard medián szűrési eljárásnál a különböző adatrendszereken L_1 normával mért adattávolság tekintetében a nem nulla átlagú zajok kiküszöbölésére. A módszer nulla átlagú zajok szűrésére szolgáló változata átlagosan 14,3%-kal jobban teljesített a hagyományos mediánszűrőhöz képest.

A módszer továbbfejlesztését tekintve az eljárás általános finomításán és optimalizálásán túlmenően különösen az alacsony zajexpozíciójú esetek hatékonyabb kezelése érdekében történő módosításra nyílik mód.

3. tézis Kifejlesztettem egy új, hisztogramalapú zajszűrési eljárást, amely a leggyakoribb érték módszerét (Most Frequent Value-t, röviden MFV-t) felhasználja topográfiai adatok szűrésére. A fenti módszer részben az adatok dinamikus mértékű, hisztogramalapú előszűrésével, MFV-t, illetve további súlyozásokat felhasználva a natív MFV-vel és a hagyományos mediánszűréssel összemérhető eredményt produkál. Az eljárás heterogén adatrendszeren szignifikánsan jobb zajcsökkenést produkál a hagyományos mediánszűréshez képest. Megállapítottam, hogy a hisztogramalapú szűrési eljárásban kiszámított részeredmények abszolút különbsége nagy megbízhatósággal képes előre jelezni, hogy a hisztogramalapú előszűrés jó eredménnyel alkalmazható-e az adott adatablakban vagy ellenkezőleg. Ezen megállapítást az L_1 -norma alapján számított hiba és a részeredmények különbségének számértéke közötti 0,99 értékű korrelációs együttható támasztja alá. A megfigyelést a szűrési eljárás eredményének súlyozása során felhasználtam módszerem hatékonyságának és robusztusságának növelésére, az adott ablakbeli MFV és hisztogramalapú szűrési eredmény értéke közötti súlyozás formájában.

Kapcsolódó saját publikáció: [S4]

3. Súlyozott mediánszűrésen és leggyakoribb érték módszerén alapuló klaszterelemzési eljárás

3.1 Bevezetés, irodalmi áttekintés

A mérési adatok klaszterelemzése a földtudományok különböző területein kiemelkedő fontossággal bír, elsősorban a kőzetek azonosítása és tagolása területén. Ennek megfelelően nagyméretű adatrendszeren számos ilyen irányú megoldás született a korábbiakban és a közelmúltban is, amely hatékonyan segítette a nyersanyagkutatást. Egy korai tanulmány [23] a klaszterelemzést ígéretes eszközként említette például a szénhidrogén-tároló zónák felismerésére. Napjainkban a klaszterelemzést a nem felügyelt gépi tanulás és a nagy adatelemzés matematikai technikáinak egyik csoportjaként említik [36], amelyet széles körben alkalmaznak a szénhidrogén-kutatás közzétírási problémáiban [7], [56], [61]. Zhang és munkatársai egy nem felügyelt tanuláson és földtani változások színkeveréses kiemelésén alapuló tároló-azonosítási módszert fejlesztettek ki, ahol többféle szeizmikus attribútumot klaszterelemzéssel dolgoztak fel annak érdekében, hogy kiemeljék az olaj- és gáztárolókat mutató anomáliákat [65]. Mások nem-hierarchikus klaszterelemzést használtak az áteresztőképesség előrejelzésének segítésére dolomit- és homokkőintervallumokban az Ogallah-mezőn, az USA-ban [57], a fácies meghatározására egy nyugat-afrikai homokkőformációban lévő kút esetében a formáció áteresztőképességének előrejelzése előtt [2], valamint heterogén karbonátos tározók azonosítására egy dél-iraki olajmezőn [3]. A mélyfúrású geofizikai szelvények értelmezésének újabb alkalmazásai közé tartozik az elektrofáciesek hatékony azonosítása és litológiai osztályozása [43], [62], az ún. inverziós eljárások alkalmazásához szükséges automatikus rétegvastagság-meghatározás és a szénhidrogén-tároló képződmények szonda-válaszfüggvény konstansainak becslése [50], a heterogén (és hiányos) maglaboratóriumi adatrendszernek klaszterelemzése [51], a nem-hagyományos tárolókban a magas gázpotenciálú intervallumok azonosítása és lehatárolása [28], [49], [52].

Kurrens terület a mélyfúrású geofizikai inverzió is, ahol a különböző kőzetfizikai paramétereket közvetlenül a fúróluk-szelvényekből (a mélység függvényében mért in situ adatokból) nyerik ki, ezzel hatékonyan csökkentik a költséges laboratóriumi mérések számát. A klaszterelemzés megközelítés a jövőben a nem-hagyományos szénhidrogén- és geotermikus tározókra is alkalmazható, ahol a klaszterelemzéssel végzett litológiai osztályozás és a formációhatárok felismerése jelentősen segíthet az inverziós becslések pontosságának és megbízhatóságának növelésében, így az általam kidolgozott módszernek számos potenciális egyéb alkalmazási terület mellett a földtudományi adatelemzéseket illetően is fontos alkalmazási lehetősége mutatkozik az általam vizsgált adatrendszeren túl is.

A k-középpontú klaszterelemzési módszer az adattérbeli objektumok (adatvektorok) osztályozását úgy végzi, hogy az adatmegfigyeléseket hasonlóságuk alapján előre meghatározott számú klaszterbe csoportosítja [22]. A klaszterközéppontot a klaszterelemek átlagértékeként számítja ki, ami viszonylag zajérzékeny folyamat, és csak Gauss-eloszlású adatok esetén ad optimális megoldást. Számos kísérlet történt a klaszterelemzés standard algoritmusainak módosítására, hogy robusztusabb megoldást

adjanak. A gyakorlatban ez akkor lehet indokolt, ha a Gauss-statisztikára vonatkozó feltételezések nem teljesülnek, vagy a mérési adatsorok bizonyos számú kiugró megfigyelést tartalmaznak. Például az evolúciós számítást, mint globális optimalizálási technikát alkalmazzák a varianciaarány-kritériumnak nevezett alkalmassági függvény optimalizálására a klaszterek optimális belső kohéziójának és külső elszigeteltségének elérése érdekében[9]. A kiugró értékek eliminálása fontos lépésnek bizonyul a klaszterelemzés előtt vagy közben. Egyesek erre a célra a maximum likelihood módszert alkalmazták[17], míg Filzmoser és társai adaptív outlier detektálási technikát mutattak be a feltáró geokémiai megfigyelésekhez[15]. Vizsgálataim során a kiugró értékek eliminálásának irányába indultam el, támaszkodva a hisztogramalapú zajszűrési eljárásomban tapasztaltakra.

A k -középpontú eljárás másik – azonban az outlierektől sem független – fent említett nagy problémáját, miszerint a klaszterelemzés eredménye nagymértékben függ a kezdeti centroidok kiválasztásától, a Steiner-féle MFV eljárás bizonyos mértékig önmagában is kezeli robusztussága okán. Mélyfűrési geofizikai adatok klaszterelemzése során a két módszert vizsgálva azonban olyan megfigyelést tettem, hogy bizonyos esetekben – főként, de nem kizárólag a kezdeti klaszterközéppontok zajos volta esetében – utóbbi nem megfelelő eredményt produkál. Ezen megfigyelés adta annak az alapját, hogy a középpontok megválasztása során a súlyozott medián eljárásnál alkalmazott szűrési módszeremet beilleszsem a folyamatba.

A Steiner-féle leggyakoribb érték (MFV) eljárás klaszterezésre történő alkalmazásának jellemzője, hogy az adott centroidot a klaszterelemek MFV-jeként (leggyakoribb értékeként) számolja ki (a k -means eljárásban alkalmazott átlag helyett), és a klaszterelemek középponttól való teljes eltérését súlyozott euklideszi (ún. Steiner) távolsággal határozza meg. Mindezt k számú centroidra hajtjuk végre. Nem csupán a klaszterközéppontok, hanem a klaszterek közötti távolságok is súlyozottak a kiugró adatok káros hatásainak csökkentésére. A folyamat a klaszterelemek teljesen automatikus súlyozására támaszkodik, amely nem igényel megkötést a megfigyelt változók statisztikai eloszlására vonatkozóan.

3.2 A Steiner-féle MFV eljárás alapján alapuló klaszteranalízis

Az általam főként a szénhidrogén kutatás céljára megalkotott robusztus klaszterezési eljárásban a következőkben ismertetésre kerülő alapeljárást fejlesztettem tovább. A módszert megalkotó vizsgálatokat földtudományi mélyfűrési adatokon kezdtem, így ezek segítségével ismertetem a kiindulópontot jelentő, klaszterezésre alkalmazott MFV eljárást.

Általános jellemző a mélyfűrési geofizikai adatokat tekintve, hogy az aktuális (\mathbf{D}) adatmátrix d_{il} eleme az l -edik fizikai változó értékét jelzi az i -edik mélységi koordinátánál ($l=1,2,\dots, L$, ahol L a rögzített mérési szelvények száma). Ez a módszereket ismertető mind szintetikus, mind terepi mélyfűrési példákra igaz, azzal a kiegészítéssel, hogy a több fúrásból álló alkalmazásokban az i index nem csak egy, hanem több fúrás mélységintervallumán is átfut ($i=1,2,\dots,N$). Az adatmátrix i -edik sorában a $d^{(i)}$ vektor az L -dimenziós adattérben egy objektumot képvisel, amely többféle adattípust is tartalmazhat, például nukleáris, elektromos, radioaktív, akusztikus

és egyéb típusú adatokat. A $d^{(i)}$ és a $d^{(j)}$ adatobjektumok közötti hasonlóságot általában a Minkowski-távolsággal lehet mérni

$$\delta^{(M)} = \left[\sum_{l=1}^L |d_l^{(i)} - d_l^{(j)}|^p \right]^{1/p}, \quad (27)$$

ahol $p=1$ a Manhattan-távolságot (cityblock), $p=2$ pedig az euklideszi távolságot jelenti. Ez utóbbi norma csak akkor ad optimális megoldást, ha az adatok Gauss-eloszlást követnek. Ha a megfigyelések nem Gauss-eloszlásúak, a Manhattan-távolság jobban teljesít, mivel kevésbé érzékeny a kiugró értékekre.

A robusztusabb megoldás érdekében a klaszterezésre módosított MFV eljárás helyett az adatobjektum és az adott klaszter középpontja közötti súlyozott, Steiner-távolságot számol:

$$\delta^{(St)} = \left[(\sum_{l=1}^L \phi_l)^{-1} \sum_{l=1}^L \phi_l (d_l - c^{(MFV)})^2 \right]^{1/2}, \quad (28)$$

ahol a súlyozási együtthatókat (ϕ) a Steiner-módszernél ismertetett súlyfüggvény számítás alapján kell kiszámítani. (Formálisan, ha a súlyozási együtthatók egyenlőek 1-gyel, és a centroidot a klaszterelemek átlagaként számoljuk, a fenti képlet az euklideszi távolságon alapuló K-means klaszterelemzési (CA, Cluster Analysis) módszerhez vezet).

Az csoportosítási eljárás során minden egyes iterációban minden objektumot az új klaszter alkotó legközelebbi centroidhoz rendelünk. A tényleges konfiguráció iteratív módon javul a centroidok pozícióinak újraszámításával. Ha ezek nem változnak jelentősen, a klaszterezési eljárás leáll.

3.3 A módosított WMFV klaszterelemzési eljárás

Mind a fent ismertetett, klaszterezés céljára módosított MFV, mind az általam tovább módosított MMFV (Modified MFV, MMFV) klaszterezési módszer klaszterközéppontokat számol, majd az adatokat a klaszterközéppontok köré partícionálja. Különbség viszont az MFV és MMFV módszerek illetve értékek között egyrészt, hogy az MFV esetében a kiinduló klaszterközéppont mátrix a cityblock metrikát felhasználó k-means eljárás által adott mátrix, az MMFV esetében pedig a fent bemutatott, hisztogram műveletek által az aktuális partíciókból előállított halmazok mediánjai által adott mátrix. Másrészt, míg az MFV az adatok aktuális centroidhoz való partícionálása során a partíció vektorok elemeit szorozza a módszer által kapott súlyokkal, ez a súlyozás az MMFV esetében nem történik meg.

A WMFV (Weighted MFV) módszer által adott végleges klaszterközéppontok az MFV-technika és a hisztogram-megközelítéssel segített MMFV-módszer eredményének súlyozott átlagaként adódnak megoldásomban: MFV ill. MMFV módszerek által adott klaszterközéppont értékek w_1 és w_2 súlyok által súlyozott átlagai lesznek [S7].

Az eljárásban az MMFV centroidszámítás módszere is kismértékben módosításra került a klaszterezési feladatra történő átalakítás során a korábban a szűrési módszernél ismertettekhez képest. Értéktartomány, ill. elemszám arányok alapján itt is meghatározásra kerülnek szűkített halmaz átlagok, ill. összeasonlításra kerül a csonka halmazok mérete is. Amennyiben a több részre osztott halmaz mérete nagyobb, az annak jelzése, hogy aktuális adatablak adatai nem jól strukturáltak a hisztogramalapú

szétválasztáshoz, ezért a hisztogramalapú MFV eredménynek némileg kisebb súlyt kell kapnia, míg fordított esetben ennek ellenkezően. Ezen tapasztalati megfigyelést is számításba vettem a módszer – ezen változatban még konstans, adott pontbeli értékviszonyok szerint előre meghatározott – súlyai (w_1, w_2) meghatározásában, vagyis az MFV, és a javított MFV (MMFV) értéke közötti végső súlyozásban, amelynek eredménye az aktuális klaszterközéppontot adja:

$$c_k = (w_1 * MMFV + w_2 * MFV) / (w_1 + w_2).$$

Mind az MFV, mind a hisztogramalapú MFV módszer klaszterközéppontokat számol tehát, illetve az adatokat a klaszterközéppontok köré particionálja, majd pedig a kettő közötti súlyozást valósítok meg az MFV és MMFV súlyozásával előálló WMFV (Weighted MFV) eljárásban.

3.4 A WMFV klaszterelemzés alkalmazása

A bemutatott WMFV klaszterelemzés módszert először egyetlen fúrás adataira alkalmaztam, majd az eljárást kiterjesztettem több fúrás adataira, hogy rekonstruálható legyen a klaszterek többdimenziós térbeli eloszlása, amely feltárja a vizsgált képződmények litológiai és petrofizikai jellemzőit a fúrások közötti térrészben is. Elsőként egy szintetikus, majd egy nagy in situ adatrendszer került elemzésre, amelyet egy hazai szénhidrogén mező fúrásaiban mértek (Szabó et al. 2023). A klaszterelemzés célja a fúrással harántolt miocén korú szénhidrogén-tároló ösztlet vizsgálata, a gáz- és víztároló zónák elkülönítése. A terepi eredmények pontosságát magáteresztőképesség-mérések és független fúróluk szelvényértelmezések igazolták, amelyek mindezekén túl egybeestek egy általam megalkotott, a klaszterezési módszer(ek) zajszűrő képességét jellemző gradiens metrika által adott eredményekkel [S7]. A fentiek szerint tehát az elsődleges vizsgálatok mélyfúrási adatokon történtek, sorrendben szintetikus, majd valós (terepi) 1 dimenziós, ill. terepi 2 dimenziós adatok segítségével. Az eredmények összefoglaló jellemzésekként kijelenthető, hogy a módosított MFV eljárás kimenete a legkevésbé zajos, és legjobban képes meghatározni a tározózónákat, emellett az eljárás a klaszterek és a releváns földtudományi mennyiségek (például a permeabilitás) közötti korrelációs együttható értékek szempontjából is a legjobb eredményt adta. Többszöri futtatás esetén a módosított MFV konzisztensen kisebb klaszterindex-változékonyságot mutatott, összehasonlítva a natív MFV és a k-means eljárásokkal.

4. tézis Kifejlesztettem egy leggyakoribb érték (MFV) - alapú klaszterelemzési módszert, ahol a kezdeti centroidokat hisztogram-alapú szűréssel határoztam meg. A módszert mélyfúrási geofizikai adatokra alkalmaztam és meghatároztam a klaszterek mélységi változását. A kidolgozott súlyozási eljárással hatékonyan megtisztítottam az eredményt a kieső értékektől. Ezen felül bemutattam, hogy a módszer eltérő zajszennyezettség és adathiány mértékek esetében is robusztus, megfelelő minőségi jellemzőkkel bíró csoportosítási alternatívát jelent. Az eljárást több fúrás adatrendszerének együttes feldolgozására is alkalmassá tettem. A robusztus eljárás javítja a klaszterelemzés zajérzékenységét, biztosítja a közetrétegek megfelelő felbontását, segíti a nyersanyagok azonosításának hatékonyságát és a mérési adatok földtani-geofizikai értelmezését. Kapcsolódó saját publikáció: [S7]

3.5 A WMFV eljárás további alkalmazása, validációja, Hurst exponensen alapuló mérőszám bevezetése

A korábbiakban ismertetett mélyfúrási geofizikai adatokon túl egy ezektől eltérő szerkezetű adatrendszeren is validáltam klaszterezési módszerem [S8]. A Nílus folyó vízszint ingadozásai képezték a vizsgálataim tárgyát. Ez az ember által valaha mért leghosszabb hidrológiai idősor. Az összehasonlítások eszközkészletét tágítandó a validációkban a korábbiakon túl egy új, a Hurst-indexen alapuló mérőszámot is bevezettem mindezekhez.

A validáció céljával szolgáló adatrendszer a Nílus folyó vízszintjéről i.sz. 622–1922 között gyűjtött éves rekordok formájában állt rendelkezésre, amely egyedülálló nagyméretű adathalmazt jelentett a módszerek további összehasonlításához. Az adatsor értékeit az éves vízszintek maximum és minimum értékei [59] adták.

A klaszterezési eljárások eredményeinek vizsgálatára használt három mérőszám közül az első metrika a klaszterezési eljárásokban bevett sziluett módszer volt[39]. Második mérőszámként gradiens értéket használtam (amely gyakorlatilag a klaszterindex vektorának változékonyságát méri). Ezt a súlyozott medián módszernél, illetve a klaszterezés előbb ismertetett példáiban alkalmazottak szerint számítottam, azzal a különbséggel, hogy ez esetben nem a gradiens, hanem az (1-gradiens) értéket használtam fel. Harmadik mérőszámként a Hurst-exponenst alkalmaztam [27]. Ennek számításának első lépése a kumulált adatok átlag körüli ingadozásainak terjedelmének meghatározása (amit R-rel jelölünk), majd ennek standardizálása az adatok S szórásával elosztva. Ennek szokásos jelölése RS-, ami tehát az R/S hányados.

A Hurst-exponens definíció szerint ezekből az alábbi módon számíthatjuk:

$$H = \frac{\log(RS)}{\log(N)}, \quad (32)$$

ahol N a sorozat hossza. A H értéke tehát a log-log plot meredekségét mutatja, értéke [0, 1] intervallumon változhat. Az exponens alkalmazására láthatunk egy példát Szücs és szerzőtársai cikkében[55].

A mérőszámnak azon tulajdonságát használtam fel, miszerint amennyiben a Hurst-exponens nagyobb értékű egy adott adatvektoron egy ugyanolyan méretű másik vektorhoz képest, az esetben a vektorban nagyobb mértékben található (legalább időszakos) trendek (nem pedig egy adott érték körül oszcillál). Azaz, a Hurst-exponens nagyobb értékével az adott klaszterindex vektor trendnélküli oszcillálásának kisebb mértékét mutattam ki. Ezzel együtt, az értéket nem önmagában használtam fel, hanem az adott klaszterindex vektorára kapott Hurst-exponens érték/az eredeti adatsor vektorok Hurst-exponens értékeinek maximuma hányadosaként.

A bemutatott, Hurst-exponenst tartalmazó mérőszám adta eredmények, illetve a mérőszámok jóságának vizsgálatára először mélyfúrási geofizikai adatokat tartalmazó szintetikus adatsort használtam, különböző zajszintek mellett.

A szintetikus adatrendszeren történő vizsgálatot követően a k-means, MFV és WMFV klaszterezést a nílusi adatrendszeren is futtattam. A csoportosítási eljárások összehasonlításának mérőszámait az 1. táblázat tartalmazza. Ezek a szintetikus

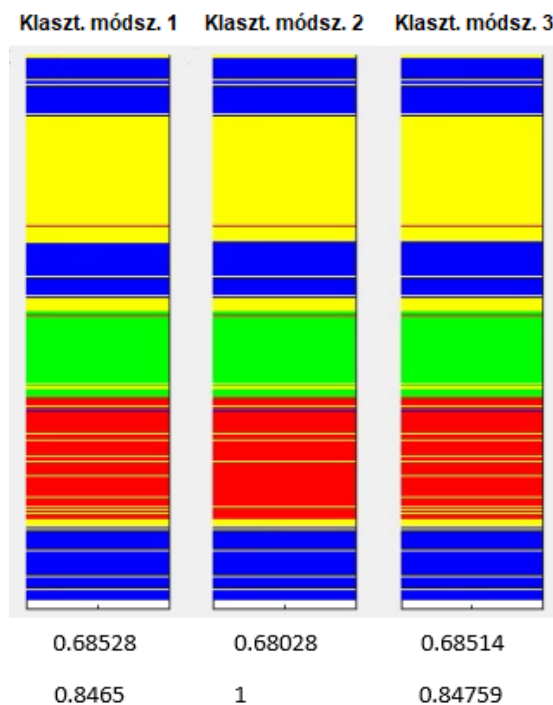
adatrendszerben alkalmazottal megegyezők voltak: első sor: sziluett átlag, második sor: sziluett átlag*(1-gradiens), harmadik sor: Hurst-exponens.

1. Táblázat Klaszterelemzés eredményeinek számszerű összehasonlítása

	k-means	MFV	WMFV
sziluett átlag	0,5082	0,4109	0,5135
sziluett átlag*(1-gradiens)	2	9	8
Hurst-exponens	1	0,7792	0,8866

Az 1. táblázatból kiolvasható, hogy a WMFV klaszterezési eljárást mutatták a mérőszámok a legjobbnak (mindhárom mérőszám esetében a harmadik oszlop értéke volt a legnagyobb).

A fenti eseten túl további futtatások esetében is vizsgáltam a mérőszámokat egyazon (kiinduló) szintetikus adatrendszeren, az előbbihez hasonlóan véletlenszerű zajjal, ezen futtatások példáját a 2. ábra tartalmazza.



2. ábra. Klaszterezést minősítő mérőszám vizsgálata (példa)

A 2. ábrán az adott klaszterezési eredmény vizuális reprezentációja alatt fentebb a sziluett érték (sziluett átlag), lentebb a Hurst-exponens adta mérőszám értéke látható. Látható, hogy az utóbbi mutatta meg helyesen, hogy melyik volt a legjobb klaszterezési kimenet.

5. tézis A Hurst-exponenst arányszámként alkalmaztam a klaszterelemzés jóságának jellemzésére. Az exponens értékét mind a klaszterindexek vektorára, mind pedig az eredeti adatsorra kiszámítva, az így kapott arány – az átlagos sziluett érték és más jósági jellemzőkhöz hasonlóan – felhasználható mérőszámként az eredmények minőségének jellemzésére. A validáláshoz használt adatrendszerre a Hurst-exponensen alapuló mérőszám összehasonlítható eredményt produkált és nem maradt el a sziluett átlagtól, mint jósági jellemzőtől. A fenti megállapításokat in situ hidrológiai adatsoron és fúróluk-geofizikai szelvények felhasználásával is alátámasztottam.

Kapcsolódó saját publikáció: [S8]

3.6 A WMFV klaszterelemzési eljárás súlyparaméterének adaptív vá tétele

Megalkottam a WMFV klaszterezési eljárás az idáig konstans $\mathbf{w}=[w_1 \ w_2]^T$ súlyainak dinamikus változatát az alábbi formában. Az eddigi két komponens (m_{e1} és m_{e2}) helyett itt négyet alkalmaztam: m_{e1}, m_{e2}, MFV , ill. a k-means eljárás adta klaszterközéppont. A négy változó közül mindet rendre az adatablak mátrixába téve, a legkisebb és második legkisebb Mahalanobis-távolságú változó értékeként meghatároztam $P1_{v1}$ és $P1_{v2}$ értékeket. Ezekből P1 értékét súlyozott átlaggal képeztem, ahol az első ($P1_{v1}$) komponens w_{p1v1} súlyát a komponens az ablak adatmátrixába helyezve a helyettesítés előtti adatablakkal való

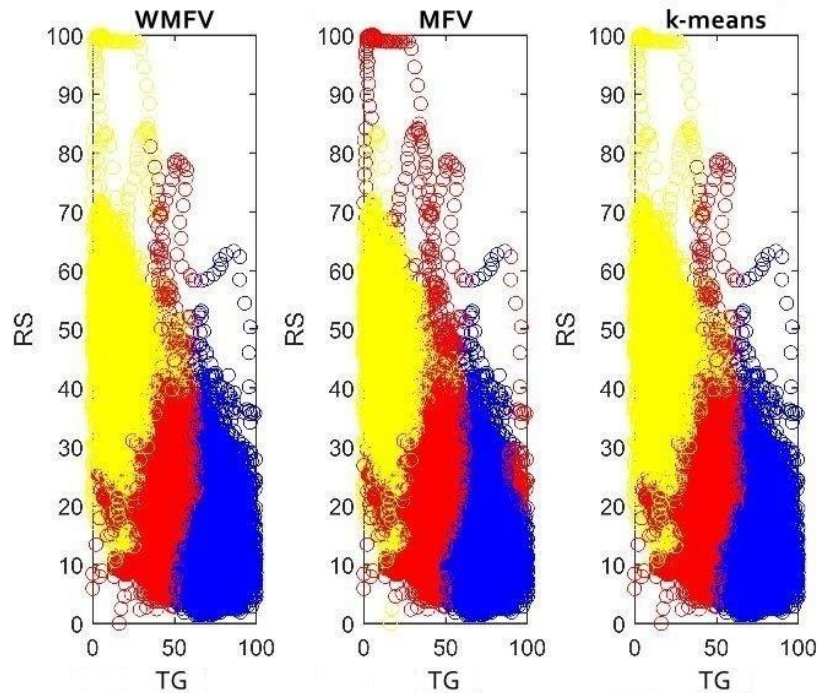
$$\text{átlag}(1/\text{Mahalanobis-távolság}) * \text{robusztus_kovariancia} \quad (32)$$

érték adta. $P1_{v2}$ súlya esetében hasonlóképp jártam el. (A robusztus kovarianciát a [38] publikáció szerint használtam.)

Ezt követően P2 értékét határoztam meg, itt mind a négy kiinduló változót az adatablak közepére téve, a maximális robusztus kovariancia értékű változó értékét véve.

Végül az adatablak középpontjának végleges értékeként P1 és P2 súlyozott átlagát számítottam ki $\mathbf{w}=[w_1 \ w_2]^T$ súlyaikkal, ahol w_1 értéke: $1/(P1 \text{ Mahalanobis-távolsága } [P1_{v1}P1_{v2}] \text{ vektortól})$, w_2 értéke hasonlóképp, $1/(P2 \text{ Mahalanobis-távolsága } [P1_{v1}P1_{v2}] \text{ vektortól})$. A fent ismertetett, az adatsoron végzett vizsgálatok összehasonlító eredményei már a WMFV módszer ezen változatával számítottak.

Az előbbieken túl egy további adatrendszeren, Debrecen környéki, nagy földtani és hidrogeológiai változékonyságot mutató mélyfúrás geofizikai adatokon is vizsgáltam módszerem [S9].



3. ábra TG-RS klaszterezési eredmények megjelenítése egy Debrecen környéki fúrás adatrendszerére esetén

A 3. ábrán látható vizuális reprezentációban megfigyelhető, hogy a WMFV módszer sokkal jobban szeparálta a 3 klaszter adatait, mint a hagyományos MFV eljárás. A k-means klaszterezés ez esetben vizuálisan hasonló eredményt produkált.

Megjegyzendő a fentiekén túl, hogy ezen az adatrendszeren módszerem, a k-means, és a tiszta MFV klaszterezés mellett a SOM (Self Organizing Map) eljárás [31] is futtatásra került. Utóbbi tanítása elsőként 900 neuronnal, 30x30-as térképmérettel, majd 4 neuronnal, 2x2-es térképmérettel történt meg, a módszerrel és a bemenetként használt adatok területével foglalkozó földtudományi szakértők által [S9].

A módszerek által kapott eredményekben mind a k-means, mind a SOM eredmények alulreprezentálták a homok arányát a szakértői jellemzések szerint, ellentétben az MFV, és WMFV eljárásokkal.

3.7 Összegzés, alkalmazhatóság

A fentiek összegzéseként kijelenthető, hogy a hisztogramalapú súlyozott medián eljárás klaszterezésre megalkotott, ismertetett változata a vizsgált egyéb módszerekkel összehasonlítva is jó eredményt produkált heterogén földtudományi adatokon. Ezt nem csupán bevett metrikákkal igazoltam, hanem egy általam megalkotott mérőszám segítségével is, amelyet külön is összehasonlító validációnak vettem alá, azt igazolandó, hogy az a vizsgált metrikákkal hasonlóan eredményesen alkalmazható. Az előbbieken túl a klaszterelemzési eljárás egy továbbfejlesztett, adaptív súlyozásos változatát is bemutattam. A módszer hatékonyan képes segíteni a nyersanyagok azonosítását robusztus klaszterezés által, továbbá a saját klaszterezési mérőszám

egyidejű (ill. önmagában is alkalmazható) használata minőségi jellemzővel minősíteni tudja a klaszterezés eredményét.

6. tézis A robusztus klaszterelemzési eljárás (WMFV) továbbfejlesztése során az alábbi eredményeket értem el:

Megállapítottam, hogy az eljárásban szereplő fő súlyok adaptív módon meghatározhatók. Megállapítottam, hogy a súlyozott medián eljárás és az MFV módszer kombinációján alapuló eljárás mind adaptív súlyos változatában mind eredeti formájában a hagyományos MFV, illetve k-means eljárásokkal összemérhető eredményt nyújt. Ezt igazoltam heterogén földtudományi adatrendszeren többféle metrika (sziluett módszer, gradiens, korreláció), ill. fúróluk-geofizikai szelvények független értelmezési eredményeinek felhasználásával. A módszert kiterjesztettem fúróluk-geofizikai adatok együttes (több fúrásra kiterjedő) értékelésére.

Kapcsolódó saját publikációk: [S7, S8, S9]

4. Új tudományos eredmények

Tudományos eredményeimet 6 tézisben foglaltam össze.

1. Tézis: Kidolgoztam egy kizárólag hanginformációkat felhasználó, szabályalapú osztályozási módszert, amely képes a fejmozgás mértékét kellő pontossággal becsülni. A módszer a betanításban nem szereplő alanyok mozgásának tartományát is képes becsülni. Bebizonyítottam, hogy képezhetők olyan hanginformációs paraméterek, amelyek mind a mozgásmérték osztályozását, mind pedig a hang és a fejmozgás közötti korrelációt tekintve jobb eredményt adnak, mint a hagyományosan alkalmazott hanginformációs jellemzők. Igazoltam, hogy az általam megalkotott osztályozási eljárás jelen struktúrájában megfelelő megoldást ad. A módszer az általam vizsgált esetekben jobb megoldást ad, mint a dimenziócsökkentéssel támogatott változat, illetve a gyakorlatban hagyományosan alkalmazott eljárások.

Kapcsolódó saját publikációk: [S1, S2, S3]

2. Tézis: Megalkottam a mozgásmérték osztályozási eljárás hatékonyságának növelése céljából egy a kieső értékeket hatékonyan kiszűrni képes, Mahalanobis és euklideszi távolság együttes használatával operáló outlier-detektálási eljárást.

Kapcsolódó saját publikáció: [S3]

3. Tézis: Kifejlesztettem egy új, hisztogramalapú zajszűrési eljárást, amely a leggyakoribb érték módszerét (Most Frequent Value-t, röviden MFV-t) felhasználja topográfiai adatok szűrésére. A fenti módszer részben az adatok dinamikus mértékű, hisztogramalapú előszűrésével, MFV-t, illetve további súlyozásokat felhasználva a natív MFV-vel és a hagyományos mediánszűréssel összemérhető eredményt produkál. Az eljárás heterogén adatrendszeren szignifikánsan jobb zajcsökkenést produkál a hagyományos mediánszűréshez képest. Megállapítottam, hogy a hisztogramalapú szűrési eljárásban kiszámított részeredmények abszolút különbsége nagy megbízhatósággal képes előre jelezni, hogy a hisztogramalapú előszűrés jó eredménnyel alkalmazható-e az adott adatablakban vagy ellenkezőleg. Ezen megállapítást az L_1 -norma alapján számított hiba és a részeredmények különbségének számértéke közötti 0,99 értékű korrelációs együttható támasztja alá. A megfigyelést a szűrési eljárás eredményének súlyozása során felhasználtam módszerem hatékonyságának és robusztusságának növelésére, az adott ablakbeli MFV és hisztogramalapú szűrési eredmény értéke közötti súlyozás formájában.

Kapcsolódó saját publikáció: [S4]

4. Tézis: Kifejlesztettem egy leggyakoribb érték (MFV) - alapú klaszterelemzési módszert, ahol a kezdeti centroidokat hisztogram-alapú szűréssel határoztam meg. A módszert mélyfúrású geofizikai adatokra alkalmaztam és meghatároztam a klaszterek mélységi változását. A kidolgozott súlyozási eljárással hatékonyan megtisztítottam az eredményt a kieső értékektől. Ezen felül bemutattam, hogy a módszer eltérő zajszennyezettség és adathiány mértékek esetében is robusztus, megfelelő minőségi jellemzőkkel bíró csoportosítási alternatívát jelent. Az eljárást több fúrás adatrendszerének együttes feldolgozására is alkalmassá tettem. A robusztus eljárás javítja a klaszterelemzés zajérzékenységét, biztosítja a kőzetrétegek megfelelő felbontását, segíti a nyersanyagok azonosításának hatékonyságát és a mérési adatok földtani-geofizikai értelmezését.

Kapcsolódó saját publikáció: [S7]

5. Tézis: A Hurst-exponenst arányszámként alkalmaztam a klaszterelemzés jóságának jellemzésére. Az exponens értékét mind a klaszterindexek vektorára, mind pedig az eredeti adatsorra kiszámítva, az így kapott arány – az átlagos sziluett érték és más jósági jellemzőkhöz hasonlóan – felhasználható mérőszámként az eredmények minőségének jellemzésére. A validáláshoz használt adatrendszerekre a Hurst-exponensen alapuló mérőszám összehasonlítható eredményt produkált és nem maradt el a sziluett átlagtól, mint jósági jellemzőtől. A fenti megállapításokat in situ hidrológiai adatsoron és fúróluk-geofizikai szelvények felhasználásával is alátámasztottam.

Kapcsolódó saját publikáció: [S8]

6. Tézis: A robusztus klaszterelemzési eljárás (WMFV) továbbfejlesztése során az alábbi eredményeket értem el:

Megállapítottam, hogy az eljárásban szereplő fő súlyok adaptív módon meghatározhatók. Megállapítottam, hogy a súlyozott medián eljárás és az MFV módszer kombinációján alapuló eljárás mind adaptív súlyos változatában mind eredeti formájában a hagyományos MFV, illetve k-means eljárásokkal összemérhető eredményt nyújt. Ezt igazoltam heterogén földtudományi adatrendszereken többféle metrika (sziluett módszer, gradiens, korreláció), ill. fúróluk-geofizikai szelvények független értelmezési eredményeinek felhasználásával. A módszert kiterjesztettem fúrólukgeofizikai adatok együttes (több fúrásra kiterjedő) értékelésére.

Kapcsolódó saját publikációk: [S7, S8, S9]

5. New scientific results

Thesis 1: I have developed a rule-based classification method utilizing exclusively acoustic information, capable of estimating head movement metrics with sufficient accuracy. The method is also capable of estimating the movement range of subjects not included in the training set. I have demonstrated that acoustic parameters can be derived which yield better results—both in terms of motion scale classification and the correlation between audio and head movement—than traditionally applied acoustic features. I have verified that the proposed classification procedure provides an appropriate solution in its current structure. In the cases examined, the method outperforms both the version supported by dimensionality reduction and the procedures conventionally used in practice.

Related publications: [S1, S2, S3]

Thesis 2: I have created an outlier detection procedure operating with the combined use of Mahalanobis and Euclidean distances, specifically designed to increase the efficiency of the motion scale classification process.

Related publication: [S3]

Thesis 3: I have developed a new histogram-based noise filtering procedure that utilizes the Most Frequent Value (MFV) method for filtering topographical data. By employing dynamic histogram-based pre-filtering, MFV, and additional weighting, the method produces results comparable to native MFV and traditional median filtering. On heterogeneous datasets, the procedure yields significantly better noise reduction compared to traditional median filtering. I have established that the absolute difference between sub-results calculated in the histogram-based filtering can predict with high reliability whether the pre-filtering will be effective in a given data window. This finding is supported by a correlation coefficient of 0.99 between the numerical value of the difference in sub-results and the error calculated based on the L_1 -norm. I utilized this observation to increase the efficiency and robustness of the method by weighting the filtering result between the MFV of the given window and the histogram-based filtering result.

Related publication: [S4]

Thesis 4: I have developed an MFV-based cluster analysis method, where initial centroids are determined via histogram-based filtering. I applied the method to borehole geophysical data and determined the depth-wise variation of clusters. Using the developed weighting procedure, I effectively cleared the results of outliers. Furthermore, I have demonstrated that the method is robust even under varying levels of noise and data gaps, providing a clustering alternative with appropriate quality characteristics. I have extended the procedure to enable the joint processing of datasets from multiple boreholes. This robust procedure improves the noise sensitivity of cluster analysis, ensures proper resolution of rock layers, and enhances the efficiency of raw material identification and the geological-geophysical interpretation of measurement data.

Related publication: [S7]

Thesis 5: I have applied the Hurst exponent as a ratio to characterize the quality of cluster analysis. By calculating the exponent for both the vector of cluster indices and the original dataset, the resulting ratio—similarly to the average silhouette value and other goodness-of-fit metrics—can be used as a measure to characterize the quality of the results. For the datasets used for validation, the Hurst exponent-based metric produced comparable results and did not underperform relative to the average silhouette value. I have supported these findings using both in situ hydrological datasets and well logs.

Related publication: [S8]

Thesis 6: During the further development of the robust cluster analysis procedure (WMFV), I achieved the following results:

I established that the primary weights in the procedure can be determined in an adaptive manner. I found that the method based on the combination of the weighted median procedure and the MFV method—both in its adaptive-weight version and its original form—provides results comparable to traditional MFV and k-means procedures. I have verified this on heterogeneous geoscientific datasets using various metrics (silhouette method, gradient, correlation) and independent interpretation results of well logs. I extended the method to the joint evaluation of multi-borehole geophysical data.

Related publications: [S7,S8,S9]

Összefoglalás

A disszertációban az adat- és jelfeldolgozás témakörében végzett módszerfejlesztési kutatásaim során elért eredményeket mutattam be. Elsőként egy virtuális ügynökök mozgásának természetességének növelésére hivatott, fejmozgás mértékek pusztán hanginformációk alapján történő becslésén alapuló osztályozási és mozgásmérték skálázási módszert ismertettem, amely akár annak megállapításában is segítség lehet, hogy egy adott videó valós, vagy generált beszélőt tartalmaz. Az osztályozási módszer megalkotása során többek között megállapításokat tettem derivált hangjellemzők és mozgásmérték közötti erős korrelációs kapcsolatokra is, amely az eljárás egyik továbbfejlesztési iránya lehet az egyéb ismertett lehetőségeken túl. Egy az osztályozási eljárás eredményeinek javítása céljával kifejlesztett egyedi outlierdetektálási eljárást is ismertettem a disszertáció ezen fejezetében.

Az értekezés második pontjában az adatok súlyozott hisztogramszűréssel való előfeldolgozásán és további súlyozásokon alapuló – főként digitális domborzati adatok kezelésére irányuló – zajszűrési eljárást mutattam be, amely az eredmények alapján jobb eredményt produkál a vizsgált esetekben a hagyományos mediánszűrésnél, de felveszi a versenyt a számos tudományterületen sikeresen alkalmazott leggyakoribb érték (MFV) robosztus módszerével is.

A disszertáció harmadik pontjában az ismertett zajszűrési eljárás alapötletét és tapasztalatait felhasználva megalkotott klaszterezési módszertant mutattam be, amely heterogén földtudományi adatokon (például terepi egy- ill. kétdimenziós mélyfúrási geofizikai adatok, valamint a Nílus folyó hosszú távú vízszintmérések) vizsgálva is bizonyította hatékonyságát. Ezen felül a klaszterezési eljárások kvantitatív minősítésére vizsgálataim szerint sikeresen alkalmazható általános mérőszámot is ismertettem munkámban.

Summary

In my dissertation, I presented the results of my methodological development research in the field of data and signal processing.

First, a classification and motion-scaling framework is introduced to enhance the perceived naturalness of virtual agents, based on the estimation of head kinematic parameters solely from acoustic input. Beyond animation, this procedure can also assist in determining whether a specific video features a real or a generated speaker.

My analysis established strong correlations between derivative acoustic features and kinematic parameters, offering a promising trajectory for future research. Furthermore, I developed a specialized outlier detection procedure to optimize classification accuracy.

The second part of the thesis introduced a noise filtering procedure—primarily aimed at handling digital terrain data—based on pre-processing data via weighted histogram filtering and additional weighting techniques. Based on the results, this method outperforms traditional median filtering in the cases examined and comparable to the Most Frequent Value (MFV) robust method, which is successfully applied across numerous scientific disciplines.

In the third section, utilizing the core concepts and experiences of the aforementioned noise filtering procedure, I presented a clustering methodology. This approach proved its efficiency when tested on heterogeneous geoscientific data, such as in situ 1D and 2D well logging datasets and long term water level measurements. Furthermore, I introduced a general metric that, according to my research, can be successfully applied to the quantitative qualification of clustering procedures.

Hivatkozások

- [1] Abdalla, H. B. (2022). A brief survey on big data: technologies, terminologies and data-intensive applications. *Journal of Big Data*, 9(1), 107.
- [2] Al-Mudhafar, W. J., & Bondarenko, M. A. (2015). Integrating K-means clustering analysis and Generalized Additive Model for efficient reservoir characterization. *77th EAGE Conference and Exhibition 2015*, (old.: 1-6.). Forrás: <https://doi.org/10.3997/2214-4609.201413024>
- [3] Al-Mudhafar, W. J., Al Lawe, E. M., & Noshi, C. I. (2019). Clustering analysis for improved characterization of carbonate reservoirs in a Southern Iraqi oil field. *Offshore Technology Conference*. Houston, USA. Forrás: <https://doi.org/10.4043/29269-MS>
- [4] Baudat, G., & Anouar, F. (2000). Generalized discriminant analysis using a kernel approach. *Neural computation*, 12(10), 2385-2404.
- [5] Ben, Y., A., S. H., & Braude, D. A. (2013). Articulatory features for speech-driven head motion synthesis. *Proceedings of Interspeech*. Lyon, Franciaország.
- [6] Calinszki, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1), 1-27.
- [7] Chen, F. F., Yang, Y. S., Pervukhina, M., Clennell, B. M., & Taylor, J. A. (2016). Clustering analysis for porous media: An application to a dolomitic limestone. *J. Pet. Sci. Eng.*, 146., 770-776. Forrás: <https://doi.org/10.1016/j.petrol.2016.07.031>
- [8] Cheng, Y., & Church, G. M. (2000).). Biclustering of expression data. *ISMB*, 8, 93-103.
- [9] Cowgill, M. C., Harvey, R. J., & Watson, L. T. (1999). A genetic algorithm approach to cluster analysis. *Computers & Mathematics with Applications*, 37., 99-108. Forrás: [https://doi.org/10.1016/S0898-1221\(99\)00090-5](https://doi.org/10.1016/S0898-1221(99)00090-5)
- [10] Davies, L. D., & Bouldin, W. D. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, 2, 224- 227.
- [11] Deng, Z., Narayanan, S., Busso, C., & U., N. (2004). Audio-based head motion synthesis for avatar-based telepresence systems. *Proceedings of the 2004 ACM SIGMM workshop on Effective telepresence*, (old.: 24-30.).
- [12] Dobróka, M., Gyulai, Á., Ormos, T., Csókás, J., & Dresen, L. (1991). Joint inversion of seismic and geoelectric data recorded in an underground coal mine. *Geophysical Prospecting*, 39 (5), 643-665. Forrás: <https://doi.org/10.1111/j.1365-2478.1991.tb00334.x>
- [13] Dobróka, T. (2021). An MFV-based image processing filter and its use in seismic tomograph. *Acta Geodaetica et Geophysica*, 56(3).

- [14] Dong, S., Wang, P., & Abbas, K. (2021). A survey on deep learning and its applications. *Computer Science Review*, 40, 100379.
- [15] Filzmoser, P., Garrett, R. G., & Reimann, C. (2005). Multivariate outlier detection in exploration geochemistry. *Computers & Geosciences*, 31., 579-587. Forrás: <https://doi.org/10.1016/j.cageo.2004.11.013>
- [16] Fujita, A., Takahashi, Y. D., & Patriota, G. A. (2014). A nonparametric method to estimate the number of clusters. *Computational Statistics & Data Analysis*, 73, 27-39.
- [17] Gallegos, M. T., & Ritter, G. (2005). A robust method for cluster analysis. *The Annals of Statistics*, 33., 347-380. Forrás: <https://doi.org/10.1214/009053604000000940>
- [18] Getz, G., E., L., & Domany, E. (2000). Coupled two-way clustering analysis of gene microarray data. *Proceedings of the National Academy of Sciences*, 97(22), 12079-12084.
- [19] Greenwood, D., Laycock, S., & Matthews, I. (2017). Predicting head pose from speech with a conditional variational autoencoder. *Interspeech 2017*, (old.: 3991-3995.).
- [20] Grimm, M., Neumann, U., Busso, C., Deng, Z., & S., N. (2007). Rigid head motion in expressive speech animation: Analysis and synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 3, 1075-1086.
- [21] Grimm, M., Neumann, U., Busso, C., Z., D., & S., N. (2005). Natural head motion synthesis driven by acoustic prosodic features. *Journal of Visualization and Computer Animation*, (3-5) 283-290.
- [22] Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: k-means clustering algorithm. *J. R. Stat. Soc. Ser. C Appl. Stat.*, 28., 100-108. Forrás: <https://doi.org/10.2307/2346830>
- [23] Hempkins, W. B. (1978). Multivariate statistical analysis in formation evaluation. *SPE California Regional Meeting*. San Francisco, USA. doi:7144-MS
- [24] Hofer, G., & Shimodaira, H. (2007). Automatic head motion prediction from speech data. *Interspeech 2007*, (old.: 722-725.).
- [25] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6), 417-441. doi:10.1037/h0071325
- [26] Huang, T. S., Yang, G. J., & Tang, G. Y. (1979). A fast two-dimensional median filtering algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27 (1), 13–18. Forrás: <https://doi.org/10.1109/TASSP.1979.1163188>

- [27] Hurst, H. (1951). Long term storage capacity of reservoirs. *Trans Am Soc Eng*, 116, 770–799.
- [28] Jarzyna, J. e. (2017). *Shale gas in Poland: Advances in natural gas emerging technologies*. (H. Al-Megren, & R. Altamimi) London: IntechOpen. Forrás: <https://doi.org/10.5772/67301>
- [29] Ji, X. e. (2022). Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. *ACM SIGGRAPH 2022 Conference Proceedings*.
- [30] Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Nießner, M., Theobalt, C. (2018). Deep video portraits. *ACM Transactions on Graphics (TOG)*, 37(4), 1-14.
- [31] Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1), 59-69.
- [32] Liu, X., Yin, J., Feng, Z., Dong, J., & Lu., W. (2007). Orthogonal neighborhood preserving embedding for face recognition. *Image Processing, 2007. ICIP 2007. IEEE nemzetközi konferencia*, 1, old.: 133-136.
- [33] Lu, Y., Chai, J., & Cao, X. (2021). Live speech portraits: real-time photorealistic talking-head animation. *ACM Transactions on Graphics (TOG)*, 40(6), 1-17.
- [34] Lüdecke, H. J., Müller-Plath, G., Wallace, M. G., & Lüning, S. (2021). Decadal and multidecadal natural variability of African rainfall. *Journal of Hydrology: Regional Studies*, 34.
- [35] Matthews, I., S., L., & Greenwood, D. (2018). Joint learning of facial expression and head pose from speech. 15, 2484-2488.
- [36] Nasraoui, O., & N’Cir, C.-E. B. (2019). *Clustering methods for big data analytics: Techniques, Toolboxes and Applications*. Springer Cham. Forrás: <https://doi.org/10.1007/978-3-319-97864-2>
- [37] O’Sullivan, P. E., Moyeed, R., Cooper, M. C., & Nicholson, M. J. (2002). Comparison between instrumental, observational and high resolution proxy sedimentary records of Late Holocene climatic change—a discussion of possibilities. *Quaternary International*, 88(1), 27–44.
- [38] Olive, D. J., & Hawkins, D. M. (2010). Robust multivariate location and dispersion. *Preprint*.
- [39] Rousseeuw P., J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53-65.
- [40] Roweis, S. T. (2002). Automatic alignment of hidden representations. *Sixteenth Annual Conference on Neural Information Processing Systems*, 15, old.: 841-848. Vancouver, Kanada.

- [41] Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika*, 66(3), 605–610.
- [42] Scott, D. W. (1992). *Multivariate density estimation: theory, practice, and visualization*. New York: John Wiley & Sons.
- [43] Sfidari, E., Kadkhodaie-Ilkhchi, A., & Najjari, S. (2012). Comparison of intelligent and statistical clustering approaches to predicting total organic carbon using intelligent systems. *J. Pet. Sci. Eng.*, 86-87., 190-205. Forrás: <https://doi.org/10.1016/j.petrol.2012>
- [44] Silva, D. V., & Tenenbaum, B. J. (2003). Global versus local methods in nonlinear dimensionality reduction. *Advances in neural information processing systems*, 120, 721-728.
- [45] Spearman, C. (1904). 'General intelligence', objectively determined and measured. *The American Journal of Psychology*, 15(2), 201–293. doi:10.2307/1412107
- [46] Steiner, F. (1991). *The Most Frequent Value. Introduction to Modern Conception Statistics*. Budapest: Akadémiai Kiadó.
- [47] Stone, D. C. (1995). Application of median filtering to noisy data. *Canadian Journal of Chemistry*, 73(10), 1573-1581. Forrás: <https://doi.org/10.1139/v95-195>
- [48] Szabó, N. P., Balogh, G. P., & Stickel, J. (2018). Most frequent value-based factor analysis of direct-push logging data. *Geophysical Prospecting*, 66. Forrás: <https://doi.org/10.1111/1365-2478.12573>
- [49] Szabó, N. P., Braun, B. A., Abdelrahman, M. M., & Dobróka, M. (2021). Improved well logs clustering algorithm for shale gas identification and formation evaluation. *Acta Geod Geophys*, 56., 711-729. Forrás: <https://doi.org/10.1007/s40328-021-00358-0>
- [50] Szabó, N. P., Dobróka, M., & Kavanda, R. (2013). Cluster analysis assisted float-encoded genetic algorithm for a more automated characterization of hydrocarbon reservoirs. *Intelligent Control and Automation*, 4., 362-370. Forrás: <https://doi.org/10.4236/ica.2013.44043>
- [51] Szabó, N. P., Nehéz, K., Hornyák, O., Piller, I., Deák, C., Hanzelik, P. P., . . . Ott, K. (2019). Cluster analysis of core measurements using heterogeneous data sources: An application to complex Miocene reservoirs. *J. Pet. Sci. Eng.*, 178., 575-585. Forrás: <https://doi.org/10.1016/j.petrol.2019.03.067>
- [52] Szabó, N. P., Valadez-Vergara, R., Tapdigli, S., Ugochukwu, A., Szabó, I., & Dobróka, M. (2021). Factor analysis of well logs for total organic carbon estimation in unconventional reservoirs. *Energies*, 14., 5978. Forrás: <https://doi.org/10.3390/en14185978>

- [53] Szabó, N., & Balogh, G. P. (2016). Most frequent value based factor analysis of engineering geophysical sounding logs. *78th EAGE Conference and Exhibition 2016*. Houten, Holland: European Association of Geoscientists and Engineers (EAGE). Forrás: <https://doi.org/10.3997/2214-4609.201600796>
- [54] Szűcs, P., & Zákányi, B. (2007). Applying most frequent value (MFV) in hydrogeological modelling. *Mérnökgeológia-Kőzetmechanika 2007*, (old.: 161-174.).
- [55] Szűcs, P., Dobróka, M., Turai, E., Szarka, L., Ilyés, C., Eid, M. H., & Szabó, N. P. (2024). Combined inversion and statistical workflow for advanced temporal analysis of the Nile River's long term water level records. *Journal of Hydrology*, *630*, 130693.
- [56] Tavakoli, V. (2018). *Geological Core Analysis: Application to Reservoir Characterization*. Springer Cham. Forrás: <https://doi.org/10.1007/978-3-319-78027-6>
- [57] Teh, W., Willhite, G. P., & Doveton, J. H. (2012). Improved reservoir characterization using petrophysical classifiers within electrofacies. *SPE Improved Oil Recovery Symposium*. Tulsa, USA. Forrás: <https://doi.org/10.2118/154341-MS>
- [58] Tibshirani, R., Walther, G., & T., H. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411-423.
- [59] Toussoun, O. (1925). *Mémoire sur l'Histoire du Nil. Mémoire Institut d'Égypte 8e*. Cairo: Le Caire Imprimerie de L'Institut Français.
- [60] Völgyesi, L., & Tóth, G. (2021). Improvement of QDaedalus measurements with continuous detection of environmental parameters. *Acta Geod Geophys*, *56*, 607-622. doi:10.1007/s40328-021-00359-z
- [61] Wang, W., Wang, Z., Leung, J. Y., Kong, C., & Jiang, Q. (2022). Petrophysical rock typing based on deep learning network and hierarchical clustering for volcanic reservoirs. *J. Pet. Sci. Eng.*, *210*, 110017. Forrás: <https://doi.org/10.1016/j.petrol.2021.110017>
- [62] Yang, H., Pan, H., Ma, H., Konaté, A. A., Y. J., & Guo, B. (2016). Performance of the synergetic wavelet transform and modified K-means clustering in lithology classification using nuclear log. *J. Pet. Sci. Eng.*, *144.*, 1-9. Forrás: <https://doi.org/10.1016/j.petrol.2>
- [63] Yehia, H. C., Kuratate, T., & E., V.-B. (2002). Linking facial animation, head motion and speech acoustics. *Journal of phonetics*, *30*(3), 555-568.

- [64] Zhang, J. (2017). Most frequent value statistics and distribution of ^7Li abundance observations. *Monthly Notices of the Royal Astronomical Society*, 468(4), 5014–5019. Forrás: <https://doi.org/10.1093/mnras/stx627>
- [65] Zhang, K., Lin, N., Fu, C., Zhang, D., Jin, X., & Zhang, C. (2019). Reservoir characterisation method with multi-component seismic data by unsupervised learning and colour feature blending. *Exploration Geophysics*, 50., 269-280. Forrás: <https://doi.org/10.1080/0812398>
- [66] Zhou, Y. H. (2020). MakeltTalk: speaker-aware talking-head animation. *ACM Transactions on Graphics (TOG)*, 39(6), 1-15.

Saját publikációk az értekezés témakörében

- [S1] Czap, L., & Kilik, R. (2015). Automatic gesture generation. *Production Systems and Information Engineering*, 7, 5–14.
- [S2] Kilik, R. (2023). Estimation of the Extent of Vertical Head Movements From Sound Alone. *Acta Marisiensis. Seria Technologica*, 20(1), 1-9.
- [S3] Roland, K. (2016.). Fejmozgás mértékek hanganyagából történő megállapításának kezdeti eredményei. *MultiScience - XXX. microCAD International Multidisciplinary Scientific Conference University of Miskolc*. Magyarország. doi:10.26649/musci.2016.064
- [S4] (Q3 kvartilis) Kilik, R. (2021). Histogram-based weighted median filtering used for noise reduction of digital elevation model data. *Acta Geodaetica et Geophysica*, 56(4), 743-764.
- [S5] Kilik, R. (2020). Applying iterative median-based filtering method for enhancing catchment area models. *Műszaki Földtudományi Közlemények*, 89(1), 357-363.
- [S6] Pecsmány, P., & Kilik, R. (2019). A leggyakoribb érték módszerének alkalmazási lehetősége a digitális domborzatmodellek zajszűrésére. *Miskolci Egyetem*.
- [S7] (Q1 kvartilis) Szabó, N. P., Kilik, R., & Dobróka, M. (2023). Robust reservoir identification by multi-well cluster analysis of wireline logging data. *Heliyon*, 9(5).
- [S8] Kilik, R., Ilyés, Cs. (2023) Klaszter és leíró statisztika elemzés közel 1500 éves vízszint adatsoron. *Új eredmények a műszaki föld-és környezettudományban*, 318-328.
- [S9] (Q1 kvartilis) Mohammed, M. A., Szabó, N. P., Kilik, R., & Szűcs, P. (2024). Examining innovative unsupervised learning techniques for automated characterization of complex groundwater systems. *Results in Engineering*, 23, 102594.