LATVIJAS UNIVERSITĀTE
ĢEOGRĀFIJAS UN ZEMES ZINĀTŅU FAKULTĀTE
ĢEOGRĀFIJAS NODAĻA

MATERIĀLU ĪPAŠĪBU 3D STRUKTŪRAS ĢENERALIZĀCIJA REĢIONĀLĀ HIDROĢEOLOĢISKĀ MODELĪ AR NORMALIZĒTĀS KOMPRESIJAS DISTANCES METRIKU

MAĢISTRA DARBS

Autors: Jānis Jātnieks
Stud. apl. jj10006
Darba vadītājs:
Dr. ģeol. docents Tomas Saks

RĪGA 2012

SATURS

APZĪMĒJUMU UN SKAIDROJUMU SARAKSTS....5
IEVADS....8
1. PROBLĒMAS NOSTĀDNE UN AUTORA IEGULDĪJUMS....10
1.1. Motivācija....10
1.2. Novitāte un konteksts....11
1.3. Darba organizācija un autora ieguldījums....14
2. MĒRĶIS UN UZDEVUMI....16
3. TEORĒTISKĀ BĀZE UN LITERATŪRAS APSKATS....18
3.1. Normalizētās kompresijas distances metrika....19
3.2. Klāsteranalīze....21
3.3. Telpisko klāsteru izveide ar Voronoi teselāciju....25
3.4. MOSYS V1 modeļsistēma....27
4. MATERIĀLI UN METODES....31
4.1. Ievades dati un to pirmapstrāde....33
4.2. Normalizētas litoloģiskās struktūras urbuma griezuma serializācija....34
4.3. Eksperimentu matricu aprēķini....36
4.4. Klāsteranalīze....40
4.5. Telpisko klāsteru izveide....42
4.6. Klāsteranalīzes rezultātu iekļaušana modeļa ģeometrijā....45
4.7. Modeļa rezultātu aprēķini, izmantojot eksperimentālās ģeometrijas....51
4.8. Autokalibrācija kvartāra nogulumu slāņos....52
4.9. Citas tehniskās piezīmes....54
5. REZULTĀTI....57
5.1. Klāsteranalīzes rezultāti....57
5.2. Telpisko klāsteru kartes un tām atbilstošās ģeometrijas struktūras....59
5.3. Aprēķinu rezultāti ar eksperimentālajām ģeometrijām....64
5.5. Iegūtie uzlabojumi....66
6. REZULTĀTU ANALĪZE UN DISKUSIJA....68
7. SECINĀJUMI....71
PATEICĪBAS....72
LITERATŪRA....73
1. Pielikums – Darba gaitas automatizēšanas scenāriji....78
2. Pielikums – Telpisko klāsteru izvietojums....78

ANOTĀCIJA

Kvartāra nogulumu reprezentācija reģionālā hidroģeoloģiskā modelī ir sarežģīta problēma šo nogulumu heterogenitātes un tās atspoguļošanai nepietiekamā urbumu griezumu datu apjoma dēļ.

Lai to uzlabotu, tiek izstrādāta darba gaita, kas ļauj ģeneralizēt kvartāra nogulumu slāņiem atbilstošo filtrācijas modeļa īpašību 3D ģeometrijas struktūru, izmantojot normalizēto kompresijas distances (NKD) metriku - inovatīvu metodi neparametriskajā mašīnapmācībā, kas līdz šim brīdim ir maz lietota ģeozinātnēs.

Darba gaitā izveidoto ģeneralizāciju atbilstība tiek pārbaudīta ar modeļsistēmas aprēķiniem. Tiek iegūts rezultātu uzlabojums, salīdzinot ar agrāko reģionālā modeļa kvartāra slāņu reprezentāciju filtrācijas modeļa 3D struktūrā.

Šis NKD lietojums ir oriģināls un potenciāli noderīgs citu Zemes sistēmas daļu modelēšanai.

Atslēgas vārdi: ģeneralizācija, normalizētā kompresijas distance, litoloģiskā struktūra, telpiskā klāsterēšana, kvartāra nogulumi, hidroģeoloģiskā modelēšana.

ABSTRACT

Representation of Quaternary deposits in a regional hydrogeological model is a complex issue due to sediment heterogeneity and the small number of borehole log data available.

To improve this, a workflow is developed for generalization of the Quaternary strata in the filtration model 3D geometry structure, using the normalized compression distance (NCD) metric - a novel non-parametric machine learning method, currently underrepresented in geoscience applications.

The generalizations created with this workflow are tested by running modelling system calculations. Improved results are obtained, compared to the previous Quaternary strata representation in the filtration model 3D geometry structure of the regional model.

This use of NCD is original, and may be of interest for modelling other parts of the Earth system.

Keywords: generalization, normalized compression distance, lithological structure, spatial clustering, Quaternary sediments, hydrological modelling.

APZĪMĒJUMU UN SKAIDROJUMU SARAKSTS

BAB Baltijas artēziskais baseins.

Baltijas koordinātu sistēma Koordinātu sistēma, kas Latvijā gan literatūrā, gan ikdienas lietojumā bieži tiek jaukta ar LKS-92 ĢIS lietotāju starpā. *European Petrol Surveying Group* (EPSG) koordinātu sistēmu datu bāzē, kas kalpo par pamatu *Open Geospatial Consortium* (*OGC*) servisu (tādu kā *Web Map Service - WMS*) koordinātu sistēmu definēšanā, šī sistēma tiek apzīmēta ar kodu 25884 un ir citādi pazīstama arī kā BalticTM93 jeb ETRS89. Y ass atskaite metros sākas no ekvatora, un Latvijas teritorijā koordinātu skaitliskās vērtības pārsniedz sešus miljonus metru. Šī koordinātu sistēma tiek izmantota PUMa projekta ietvaros veidotajā Baltijas artēziskā baseina modelēšanas sistēmā MOSYS, kas tiek izmantota šī darba ietvaros.

DB Datu bāze(s).

ĢIS Ģeogrāfiskas informācijas sistēma(s).

Kompresija Domāta datu kompresija tādā izpratnē, kā to lieto datu kompresijas programmatūras izstrādes praksē – matemātikas, fizikas un datorzinātņu metožu lietišķs apvienojums programmatūrā, lai samazinātu uzglabājamo vai pārraidāmo datu apjomu. Praksē pie kompresijas algoritmu piemēriem var minēt tādus kā *LZ77, LZMA, PPMd* (kas izmantots šajā darbā) un *BZip2*. Starp ikdienā biežāk lietotajiem formātiem var minēt zip, 7z, bz2. Kā bāzes literatūra par datu kompresijas teoriju un praksi šajā darbā izmantoti galvenokārt divi avoti (Mahoney, 2011; Solomon, 2007).

labi / slikti rezultāti Lai zinātniski runātu par labiem vai sliktiem rezultātiem ir, protams, jābūt kritērijam, pēc kā tie tiek vērtēti. Ja nav minēts savādāk, tad šī darba ietvaros kā galvenais kritērijs tiek izmantota MOSYS modelēšanas sistēmas aprēķinu kvadrātiskā novirze attiecībā pret pjezometrisko ūdens līmeņu novērojumu datiem (Takčidi, 1999) konkrētajā MOSYS V1 kalibrācijas slānī. Attiecīgi ar "labiem", šī darba ietvaros, domāti tādi rezultāti, kam šī novirze ir zemāka. Šāds raksturojums tiek lietots ne tikai tāpēc, ka ir intuitīvi viegli izprotams, bet, galvenokārt tādēļ, ka novērš nepieciešamību daudzviet atkārtot pilnu kritēriju kopu pēc kuriem iegūtie rezultāti tika izvērtēti, parasti attiecībā pret citu modelēšanas aprēķinu variantu rezultātiem.

Materiālu struktūra Šī darba ietvaros ar materiālu struktūru, ja nav norādīts savādāk, domāts materiālu īpašības definējošo ģeometrijas slāņu skaits, savstarpējā secība un

biezumi kādas teritorijas (parasti telpiskā klāstera) ietvaros. Materiālu struktūra var būt domāta gan vienā dimensijā (urbuma griezums), gan divās (šī definējuma izpratnē strukturāli atšķirīgu teritoriju izplatības robežas pārskata kartē vai kartoshēmā), gan arī 3D izpratnē (strukturāli atšķirīgu teritoriju apvienojums vienā trīsdimensionālā ģeometrijas objektā, kuru veido šī definējuma izpratnē strukturāli atšķirīgas daļas).

Modeļsistēma MOSYS V1 gadījumā - kopums, kas ietver sekojošus komponentus: ģeoloģiskās uzbūves modeli, pazemes ūdens filtrācijas modeli, vielu pārneses un transformāciju ("ķīmijas") modeli, hidroloģisko modeli, datus un to organizāciju, programmatūru datu apstrādei, priekšapstrādei un rezultātu vizualizācijai, kalibrācijas metodes un aprēķinu variantus.

Plašākā nozīmē, kādā lietots darba ievadā - jebkura radniecīga vide plūsmas, cirkulācijas vai vielu pārneses modelēšanai, kas ietver lielu daļu no iepriekš uzskaitītajiem komponentiem MOSYS sistēmām un ar to palīdzību cenšas atainot dabas procesu komponentu mijiedarbību kādā plašākā teritorijā. Šī darba ietvaros, lielākoties, vienmēr attiecināts uz modeļiem, kam ir telpiskā komponente.

NKD/*NCD* Normalizētā kompresijas distance (*Normalized Compression Distance* angļu val.) - neparametriska mašīnapamācības metode, kas konceptuāli pirmo reizi aprakstīta 1998. gadā (Bennet et al., 1998), un pirmoreiz praktiski formulēta 2005. gadā (Cilibrasi and Vitanyi, 2005). Plašāks tās teorētiskais, praktiskais apraksts un liels klāsts ar eksperimentu piemēriem dots 2007. gada disertācijā, kas veltīta šai metodei (Cilibrasi, 2007).

PUMa ESF projekts „Starpnozaru zinātnieku grupas un modeļu sistēmas izveide pazemes ūdeņu pētījumiem" (2009/0212/1DP/1.1.1.2.0/09/APIA/VIAA/060), kuru realizē Latvijas Universitāte sadarbībā ar Latvijas Lauksaimniecības Universitāti, www.puma.lu.lv.

Scenārijs Šajā tekstā attiecināts uz nelielām, pārsvarā interpretētajās programmēšanas valodās sagatavotām, programmām eksperimentu automatizēšanai. Tās angļu valodā mēdz dēvēt par scripts. Šajā darbā scenāriju izstrādei lielākoties izmantotas *Python* un *R* scenārijvalodas.

Serializācija Termins, kas aizgūts no datorzinātnēm un tiek lietots, lai aprakstītu darbību, kuras rezultātā kompleksas (daudzdimensionālas, kā arī hierarhiskas) datu struktūras no reprezentācijas datora atmiņā tiek pārvērstas vienā simbolu rindā, kas ļauj tās saglabāt, pārrraidīt un pēc tam atkal rekonstruēt, balstoties

uz šo rindu. Šajā darbā vārds “serializācija” tiek lietots, lai apzīmētu urbuma griezuma, kam slāņu pārejas normalizētas pret kvartāra segas biezumu, pārveidi simbolu rindā, kur katram iežu veidam atbilst unikāls teksta simbols. Simboli tiek atkārtoti, lai reprezentētu slāņu biezumu normalizētajā urbuma kvartāra slāņu griezumā, saglabājot informāciju par slāņu biezumu attiecībām, savstarpējo novietojumu un tiem piemītošajām iežu klasēm.

IEVADS

Darbs ir strukturēts kā oriģināls lietišķs pētījums un risina reģionālajā modelēšanā aktuālu problēmu - kā reprezentatīvā veidā metodiski vispārināt kompleksus, telpiskus datus modelēšanas vajadzībām. Šajā ietvarā tiek risināta praktiska problēma reģionāla mēroga hidroģeoloģiskajā modelēšanā, PUMa projekta ietvaros izstrādātajā MOSYS V1 modelēšanas sistēmas vidē, izmantojot salīdzinoši jaunu un daudzsološu neparametriskās mašīnapmācības metodi - normalizētās kompresijas distanci (NKD, angļu val. - *normalized compression distance*). NKD iepriekš uzrādījusi labus rezultātus visplašākajā nozaru klāstā (Cilibrasi, 2007). Tomēr tās lietojums ģeozinātnēs, darba tapšanas brīdī, bija minimāls. Ir veikti eksperimenti ar NKD tālizpētes vajadzībām, kas praksē nozīmē NKD lietojumu rastra attēlu klasifikācijā (Roman-Gonzalez et al., 2011; Cerra et al. 2010; Stepinski et al., 2007), taču iztrūkst lietojumu demonstrācijas ar vektordatiem un piemēri no Zemes sistēmu modelēšanas jomām.

Šajā darbā NKD metrika tiek izmantota, lai salīdzinātu urbumu griezumu litoloģisko struktūru kvartāra nogulumu slāņos un ar klāsteranalīzes palīdzību veidotu vispārinātas, abstrahētas kvartāra nogulumu reprezentācijas reģionālā hidroģeoloģiskā modeļa MOSYS V1 filtrācijas modeļa ģeometrijas struktūrā. Darba mērķis, vispārīgā veidā, ir pārbaudīt, vai NKD var izmantot kompleksu, telpisku modeļu ģeneralizācijai uz reģionālu mērogu, jo kvartāra nogulumiem ir pārāk augsta heterogenitāte, kuru nav iespējams atspoguļot ar pieejamajiem urbumu griezumu un nogulumu izplatības karšu datiem. Plašākam darba mērķu un uzdevumu izklāstam atsevišķi ir veltīta 2. nodaļa. Darba veikšanai ir lielāks skaits uzdevumu, kuri kopumā veido vienotas darba gaitas izstrādes daļas. Maģistra darba ietvaros autors veic šīs darba gaitas tehnisko daļu izstrādi MOSYS modeļsistēmas attīstības ietvaros un ar aprēķiniem demonstrē, ka izmantotā metodika ir veiksmīgi pielietojama kvartāra nogulumu reprezentācijas uzlabošanai reģionālajā hidroģeoloģiskajā modelī.

Darba rezultāti ir tikuši aprobēti, piedaloties Eiropas ģeozinātņu savienības (*Europan Geosciences Union - EGU*) Ģenerālajā Asamblejā, 2012. gada aprīlī. Šajā starptautiskajā konferencē darba metodika un pirmie iegūtie rezultāti prezentēti divos stenda referātos (Jātnieks et al., 2012*a*; Jātnieks et al., 2012*b*) un tiem veltītajā stenda referātu diskusijas sesijā.

Darbs ietver teorētiskajiem jautājumiem veltītas nodaļas, detālu metodikas izklāstu, kas veido lielāko daļu šī darba, kā arī pārskata rezultātu apakšnodaļas, kas, papildus darba kopējiem rezultātiem, aplūko starprezultātus atsevišķām darba tehniskajām daļām.

Kopumā darba tekstā ir 28 attēli, no kuriem 25 ir autora sagatavoti, 2 tabulas un 2 pielikumi. Maģistra darbs izstrādāts ESF aktivitātes „Cilvēkresursu piesaiste zinātnei" projekta „Starpnozaru zinātnieku grupas modeļu sistēmas izveide pazemes ūdeņu pētījumiem" (2009/0212/1DP/1.1.1.2.0/09/APIA/ VIAA/060) ietvaros.

1. PROBLĒMAS NOSTĀDNE UN AUTORA IEGULDĪJUMS

1.1. Motivācija

Gandrīz nejauši 2008. gada vasarā manu interesi piesaistīja daži raksti internetā par jaunumiem datu kompresijas pētījumos. Tie bija saistīti ar Meta Mahonija PAQ8 kompresoru salīdzinoši augstajiem rezultātiem kompresijas programmu reitingu tabulās. Tādas neoficiāli apkopo lielāks skaits entuziastu, lai salīdzinātu dažādu datu kompresijas programmu efektivitāti atšķirīgu datu paraugu saspiešanā. Starp saistītajiem resursiem atradās arī saite uz Meta Mahonija 1999. gada datu kompresijas konferences tēzēm, kas kopā ar citiem darbiem (Legg and Hutter, 2007) pauda ļoti interesantu atziņu – datu saspiešanas spēja ir cieši saistīta ar inteliģences jēdzienu, tādā izpratnē, kādā to izmanto, lai mēģinātu pamatot, kas vispār ir inteliģence mākslīgā intelekta pētījumiem. Rūdija Cilibrasi 2007. gada disertācija, kas plašāk aprakstīja normalizētās kompresijas distances (NKD) izstrādes gaitu un pielietojumu, bija ļoti iespaidīga demonstrācija šai pieejai, kas pauda, ka datu kompresija varētu tikt izmantota ne tikai kā kritērijs mākslīgajai inteliģencei, bet arī praktiskām klasifikācijas vajadzībām. Šī disertācija (Cilibrasi, 2007) un ar to saistītās agrākās publikācijas (Bennet et al., 1998; Cilibrasi and Vitanyi, 2005) uzskatāmi norādīja ne tikai uz datu kompresijas algoritmu pielietojamību klasifikācijā – lietojumā, kuram tie šķietami nemaz nebija paredzēti, bet arī demonstrēja to, ka šie panākumi ir ļoti universāli un konkurētspējīgi plašā nozaru spektrā.

Mani kā ģeogrāfu, saprotams, interesēja, kādi varētu būt potenciālie lietojumi ģeogrāfijas vai kādā vismaz saistītā zinātņu nozarē šādai metrikai vai tai pietuvinātai pieejai vispār. Toreiz starp daudzajiem piemēriem no bioloģijas, lingvistikas, datorzinātnēm u.c. jomām nekas īpaši saistīts ar ģeogrāfiju, ĢIS vai Zemes zinātnēm tuvu tematiku nebija atrodams, un šis trūkums man palika prātā kā potenciāls intriģējošam pētījumam. Vēlāk parādījās *NASA* (ASV kosmosa aģentūras) (Stepinski et al., 2007) un *DLR* (Vācijas kosmosa aģentūras) (Cerra et al., 2010) eksperimentu apraksti, izmantojot normalizētās kompresijas distanci tālizpētes attēlu klasifikācijai kā neparametrisko (nevadīto) klasifikāciju, kas mani ļoti ieinteresēja. Bija pagājuši vairāki gadi, bet pārbaudot citēšanas rādītājus pirmajam rakstam, kas formāli aprakstīja NKD metriku, bija novērojama neizsīkstoša tendence pieaugt atsauču skaitam uz šo rakstu. Interese par NKD metodes pielietojamību turpina pieaugt vidēji par 100 atsaucēm katru gadu pēc Google Scholar datiem. Tāpat no šiem rezultātiem bija

redzams, ka pieaug interese par NKD pielietojumu tālizpētes vajadzībām un attēlu klasifikācijai vispār, bet citās ar ģeogrāfiju saistītajās jomās tas nebija manāms.

Tikmēr, strādājot pie MOSYS reģionālā gruntsūdens plūsmas modeļa ģeoloģiskās struktūras ģenerācijas tehniskajām problēmām, viena no tām, ar kurām saskārās ģeoloģijas darba grupa, bija pēc iespējas realitātei atbilstošākas kvartāra nogulumu reprezentācijas izveide modeļa ģeometrijā. Kvartāra nogulumu reprezentācija modeļa filtrācijas īpašību parametrizācijā ir kompleksa problēma vairāku iemeslu dēļ.

Kvartāra nogulumu heterogenitāte ir pārāk liela, lai to būtu pilnībā iespējams atspoguļot, izmantojot pieejamo informāciju no urbumu datiem un no dažāda mēroga kvartāra nogulumu kartēm. Iemesli tam ir vienkārši – urbumu datu apjoms nav pietiekams, lai atspoguļotu kvartāra heterogenitāti tikai ar urbumu griezumu datiem, savukārt, kvartāra segas kartēšanas rezultātā iegūtie nogulumu izplatības poligoni sniedz ieskatu vienīgi pirmā slāņa nogulumu izplatībā. Taču MOSYS modelēšanas sistēmā, tāpat kā vairumā hidroģeoloģiskās modelēšanas platformu, ģeoloģiskās struktūras reprezentācija tiek veidota trīs dimensiju (3D) ģeometriskajā struktūrā, atspoguļojot labākās pieejamās zināšanas par teritorijas ģeoloģisko uzbūvi, kas kalpo kā daļa no modeļa parametrizācijas pirms aprēķinu veikšanas.

Diskutējot par dažāda veida iespējamajiem risinājumiem kvartāra reprezentācijas izveidē, man ienāca prātā NKD metodika un radās pamatots jautājums: vai CompLearn.org risinājums, kurš piedāvā NKD aprēķinu kā galveno funkciju, nebūtu izmantojams kā klasifikācijas rīks problēmas risinājumam? Pēc garākas diskusijas ar darba vadītāju tika nonākts pie atziņas, ka kvartāra nogulumu reprezentācijas izveide būtībā ir ģeneralizācijas problēma un ka viens no iespējamajiem risinājumiem šai ģeneralizācijai būtu klasifikācija lielākās teritorijās ar katrai teritorijai reprezentatīvu (tipveida) urbuma griezumu. No šīs diskusijas seko šis darbs, ar to saistīto programmatūras rīku izveide, modelēšanas eksperimenti un rezultāti.

1.2. Novitāte un konteksts

Darba konteksts un iespējamā pielietojamība var izrādīties plašāka par konkrēto demonstrāciju hidroģeoloģiskā modeļa ģeometrijas struktūras ģeneralizācijā. Mūsdienās, pieaugot pieejamo datu apjomam un novērojumu periodu ilgumiem, reģionu izdalīšanas un raksturīgo īpašību ģeneralizācija ir ieguvusi jaunu nozīmi tehniskās ģeogrāfijas jomās. Tā seko divām globālām tendencēm. Viena no tām ir metodiskās iestrādnes, kas rodas, pieaugot datu ievākšanas kapacitātei ar mūsdienu datu ievākšanas un uzkrāšanas ekipējumu

producētajiem datu apjomiem. Arvien aktuālāks kļūst jautājums nevis par pašu datu vākšanu vai uzglabāšanu, bet par to apstrādi veidā, kas atspoguļotu tajos esošos rakstus, nozīmes un saistības, kas ne vienmēr ir acīmredzamas datu izmantotājam. Angļu valodas literatūrā šo paņēmienu kopumu mēdz saukt par *data mining*, kas latviski tiešā tulkojumā, ir nodēvētas par "datizraci". Datizrace ir starpdisciplināra joma, kas balansē starp statistikas, mašīnapmācības (mākslīgā intelekta) un informāciju sistēmu arvien vairāk saplūstošajām robežām. Šī virzība skar tehniskās ģeogrāfijas apakšjomas tiktāl, cik tās ir saistītas ar datu bāzēm, kas atsaucas uz to lietojumu ģeogrāfiskās informācijas sistēmās (ĢIS) un saistītajā analītiskajā programmatūrā. Otra tendence, kuras konteksts daudz vairāk ir ietekmējis šī darba tapšanu, ir plaša mēroga modelēšanas sistēmu izstrāde un šādu sistēmu izstrādei veltītie pētījumi. Mērogs ir viena no senākajām ģeogrāfijas pamatproblēmām. Mērogs ne tuvu nav aktuāls tikai kartogrāfiskajiem lietojumiem ģeogrāfijas apakšnozarēs, lai gan tieši šai kontekstā, iespējams, ir visatpazīstamākais. Mērogs reģionāla un planetāra mēroga modelēšanas sistēmās ir cieši saistīts ar izmantoto datu un modeļa struktūru reprezentativitāti – to, cik labi iespējams atspoguļot procesus vispārinātā veidā, jo plaša mēroga sistēmām vienmēr ir kāds detalizācijas līmenis, par kuru sīkāka mēroga elementus vai procesus atspoguļot nav iespējams dažādu iemeslu dēļ.

Praksē pastāv daudzveidīgas datu agregācijas metodes, kas fiziskās ģeogrāfijas nozarēs tiek lietotas nu jau gadu simtiem, piemēram, tādas pastāv temperatūras datiem – vidējās, vidējās minimālās, maksimālās; aprēķinātas laika periodiem, lielākiem kā uzkrāšanas solis no minūtēm līdz pat gadiem. Šādus datus tālāk mēdz izmantot dažādu procesu un telpisko rakstu skaidrošanai, pielietojot tos šādā ģeneralizētā veidā uz plašākiem mērogiem, piemēram, gadījumos kad novērojumu stacijas ir nevienmērīgi izkliedētas un atrodas lielos attālumos viena no otras. Tomēr aktuāls ir jautājums - kā ģeneralizēt kompleksus daudzdimensionālus datus, kuru lietojums ir aktuāls mūsdienu modelēšanas pētījumos?

Šeit nevar nepieminēt modelēšanas nozares vienam no biežāk minētajiem vēsturiskajiem pamatlicējiem, statistiķim Džordžam Boksam, piedēvēto teicienu - "Visi modeļi ir nepareizi, bet daži no tiem ir noderīgi". Tas lielā mērā būtu attiecināms uz vienu no ģeogrāfijas zinātņu klasiskajām dimensijām – **mēroga dimensiju** modelēšanas sistēmu izveidē. Plaša mēroga modeļos šis Boksa minētais "nepareizums" bieži vien ir saistāms tieši ar ģeneralizāciju. Visbiežāk to izraisa nesaderības starp dažādu datu ievākšanas un/vai agregācijas līmeni, kas tiek kombinēts ar citiem tikpat neregulāri ievāktiem un, iespējams, pēcapstrādē dažādi agregētiem datiem, jo to izkliede telpā un laikā mēdz atšķirties. Veidojot reģionāla un, vēl jo vairāk, globāla mēroga modeļus, neizbēgama ir nepieciešamība vienkāršot un vispārināt dabas veidotos rakstus uz modeļsistēmas iekšējās struktūras mērogu. Protams,

šai struktūrai ir jābūt pēc iespējas tuvākai realitātei un pēc iespējas labāk jāreprezentē pieejamās zināšanas par dabas sistēmu darbību. Tomēr kādā brīdī jebkuram telpiskam modelim tiek radīta modelēšanas sistēmas struktūra, kura tiek piepildīta ar pieejamajiem datiem, informāciju par to izkliedi telpā un likumsakarību matemātisko aprakstu. To nosaka ne tikai skaitļošanas resursu ierobežojumi, bet arī limiti mūsu izpratnē par Zemes sistēmas uzbūvi un tās komponentu mijiedarbību, kā arī novērojumu datu izkliede. Tomēr ne mazāk aktuāla ir arī pretējā problēma – **kā vispārināt** detālus un ļoti daudzveidīgus datus tā, lai šis vispārinājums būtu reprezentatīvs praktiskajam pielietojumam nepieciešamajā mērogā. Gadījumos, kad nepieciešams pielietot agregāciju vienas dimensijas ietvaros (temperatūra, mitrums, dziļums, virsmas augstuma atzīme, u.t.t.), tam ir pieejamas daudzas plaši lietotas un bieži vien samērā triviālas statistiskās metodes. Taču gadījumos, kad agregāciju nepieciešams veikt kompleksi strukturētiem daudzdimensionāliem datiem kopumā, triviāli risinājumi ir pieejami, bet nereti maz efektīvi.

Tieši šo vispārīgo problēmu ietvarā ir izstrādāts šis darbs, kura mērķis ir apvienot jaunākos lietišķos atklājumus informācijas teorijā, klasiskās hierarhiskās klāsterēšanas metodiku ģeneralizācijas mērogošanā ar aktuāliem pētījumiem reģionālajā Baltijas artēziskās baseina hidroģeoloģiskajā modelī MOSYS V1.

Šī darba praktiskais pielietojums ir piesaistīts konkrētam reģionālam hidroģeoloģiskajam modelim, taču darba organizācijas shēma un metodika ir viegli adaptējama un vispārināma uz citām modelēšanas jomām, kurām nepieciešama telpiska, 3D modeļa elementu struktūra. Šī darba rezultātā iegūto dalījumu ir plānots pielietot kvartāra uzbūves materiāla īpašību modelēšanai, izmantojot iegūtās rajonēšanas teritorijas kā teritorijas, kurās kvantitatīvi izdalītas līdzības ģeoloģiskajā uzbūvē. Taču paredzams, ka darba ietvaros izveidotā metodoloģija būtu viegli adaptējama un plaši pielietojama arī darbā ar cita veida datiem un ar nozīmi citās ģeogrāfijas studiju jomās.

Visbeidzot, gribētos pieminēt labi zināmu, bet reizēm grūti aptveramu paradoksu saistībā ar datu agregāciju. Darbs ar vispārinātiem datiem un uz tiem balstītām modelēšanas struktūras reprezentācijām var šķist brīžiem pārāk abstrakts. Skatoties modelēšanas struktūras griezumos, kas radīti šī darba gaitā, ir grūti intuitīvi novilkt robežu, kur beidzas struktūru reprezentativitāte reģionālā un lokālā mērogā. Taču šī robeža pastāv – reprezentativitāte reģionālā mērogā nenozīmē, ka paraugi, kas ievākti lokālā mērogā, liksies līdzīgi reģionāli reprezentatīvajai struktūrai, lai cik veiksmīgi arī nebūtu izraudzīti tipveida griezumi. Piemēram, 28. maija vidējā temperatūra, atspoguļota ar 30 gadu slīdošo mainīgo vērtību, kādas tiek izmantotas klimata modelēšanas eksperimentos, visdrīzāk nesakritīs ar kādā konkrētā 28. maijā veiktu novērojumu, piemēram, plkst. 12:00 vai kādā citā konkrētā laikā.

Protams, temperatūras gradientam dienas ietvaros mainoties, kādā konkrētā brīdī šī temperatūra tiešām var sakrist gan ar 30 gadu slīdošo vidējo, gan citām dažādu agregācijas darbību rezultātā iegūtām vērtībām (arī sabojāts pulkstenis ir pareizs divas reizes dienā). Taču tā būs reprezentatīva agregācijas periodā un, neskatoties uz to, izraisīs neizpratni vasaras naktī salstošam cilvēkam.

Tieši tas pats attiecas uz šī darba gaitā izveidotajām hidroģeoloģiskā modeļa 3D struktūrām kvartāra nogulumu vispārinātai reprezentācijai. Darba ietvaros izveidotās kvartāra nogulumu reprezentācijas ir atbilstošas savam mērogam, un to apliecina modelēšanas sistēmas rezultāti attiecībā pret novērojumu datiem. Taču nav pamata sagaidīt, ka, veicot urbumu kādā Latvijas teritorijas vietā, tā griezums sakritīs ar šī darba ietvaros izveidoto vispārināto ģeometriju griezumiem atbilstošajā modeļa struktūras vietā.

Šis paradokss pastāv tādēļ, ka esam adaptēti spriest par lietām un procesiem sev tveramos mērogos. Kompleksas trīsdimensionālas struktūras, kam raksturīga liela iekšēja heterogenitāte reģionālos un planetāros mērogos, ir grūti pilnībā aptvert un kvalitatīvi spriest par to reprezentativitāti. Tādēļ par šādu vispārinājumu pareizību var spriest tikai, validējot to ar aprēķiniem vai citām metodēm atbilstošajā mērogā. Lai cik abstrakti tas neliktos, šādu reprezentāciju veidošanai ir milzīga nozīme zinātnes attīstībā. Tieši tas, cik vispārīgā veidā reprezentatīvi mēs varam atveidot ap sevi esošās Zemes sistēmas daļas pietiekami pielietojamā un reizē vienkāršotā veidā, nosaka pielietojamību modeļiem, no kuriem arvien vairāk esam atkarīgi gan īstermiņa, gan ilgtermiņā plānošanas vajadzībām. Savā veidā tam ir cieša sakarība ar kompresijas pamatideju – atrast īsāku aprakstu sarežģītām lietām.

1.3. Darba organizācija un autora ieguldījums

Tā kā darbs tapis ESF projekta „Starpnozaru zinātnieku grupas un modeļu sistēmas izveide pazemes ūdeņu pētījumiem" (PUMa) pētījumu ietvaros un daudzi darbā izmantotie komponenti ir projekta darba grupas kopdarbs, maģistra darba autora ieguldījums attiecas uz tālāk nosauktajām lietām:

- Ideja par normalizētās kompresijas distances lietojumu kvartāra nogulumu struktūras ģeneralizācijā.
- Urbuma griezumu serializācijas metodes tehniskā izstrāde izmantošanai apvienojumā ar urbumu datu bāzes savienojumu un realizācija *Python* un *SQL* programmēšanas valodās, risinājuma testēšana un integrēšana klāsteranalīzes automatizēšanas skriptā. Serializācijas darbības principa idejas autors ir Juris Seņņikovs.

- Klāsteranalīzes eksperimentu automatizācijas skripta izveide *Python* valodā.
- Telpisko klāsteru izveides darba gaitas izstrāde un automatizēšana, izmantojot *R* statistikas valodu.
- Eksperimentālo kvartāra nogulumu struktūru 3D ģeometriju izveide un integrēšana MOSYS V1 Baltijas artēziskā baseina modeļsistēmas ģeometrijā. Izmantotā ģeometrijas bāze, tās izveides scenāriji un komponenti ir projekta PUMa dalībnieku kopdarbs.
- Modelēšanas sistēmas pjezometrisko līmeņu aprēķinu veikšana izveidotajām vairāk kā 162 eksperimentālajām ģeometriskajām struktūrām.
- 10 atlasītu ģeometrisko struktūru materiālu īpašību kalibrācija kvartāra nogulumu slāņos un tās gaitas rezultātu apkopošana, izmantojot MOSYS autokalibrācijas funkcionalitāti.
- Uzlaboto ģeometriju pilnas kalibrācijas gaitas veikšana labāko rezultātu ieguvei.

2. MĒRĶIS UN UZDEVUMI

Šis maģistra darbs pretendē uz oriģināla lietišķa pētījuma statusu, kura **mērķis ir izstrādāt darba gaitu, kas ļautu izveidot ģeneralizētu materiālu īpašību 3D struktūras reprezentāciju kvartāra nogulumu slāņiem reģionālā hidroģeoloģiskā modelī, izmantojot normalizēto kompresijas distances metriku**.

Šajā darbā tiek izmantots MOSYS V1 reģionālais hidroģeoloģiskais modelis Baltijas artēziskā baseina teritorijai, kurš tiek izstrādāts ESF projekta PUMa ietvaros. Modeļa izveide ir process, kas turpināsies arī pēc maģistra darba iesniegšanas. Tā versijas V1 ģeometrijas struktūra un programmatūras bāze tiek izmantotas kā atskaites punkts, uz kuru tiek balstītas šajā darbā izstrādātās ģeneralizētās ģeometrijas struktūras kvartāra nogulumiem Latvijas teritorijā.

Darba galvenā **hipotēze** ir, ka **normalizēto kompresijas distances metriku var veiksmīgi pielietot kvartāra nogulumu slāņu litoloģiskās uzbūves ziņā līdzīgu, reprezentatīvu telpisko klāsteru izdalīšanai.**

Normalizētās kompresijas distances metode ir ļoti jauna, bet daudznozaru pētījumos labi aprobēta un plašu nenoplokošu interesi izraisoša metode, kurai kopš tās pamatus aprakstošās publikācijas iznākšanas 2005. gadā (Cilibrasi and Vitanyi, 2005) katru gadu tiek veltītas ~100 jaunas publikācijas. Tā sakņojas informācijas teorijas pamatos un balstās uz universālo datu kompresijas algoritmu pielietošanu kopējas informācijas satura izvērtēšanai dažāda veida un formāta datu masīvos. Šai metrikai ir laba teorētiskā bāze (Li and Vitany, 2008; Cilibrasi, 2007; Cilibrasi and Vitanyi, 2005; Li et al., 2004; Bennet et al., 1998), un tā ir praktiski pielietojama, izmantojot CompLearn.org neparametrisko mašīnapmācības rīku komplektu, kas pieejams globālā tīmekļa adresē complearn.org. Metode ir ļoti universāla, par to labi liecina pētījumu nozaru spektrs, kas aplūkots šī darba teorētiskajā daļā.

Tiek plānots, ka darba gaitā izveidotā metodika būs viegli adaptējama darbam ar cita veida telpā izkliedētiem aprakstošiem ievades datiem, bet konkrētā darba gaita dos ieskatu ģeoloģiskās nenoteiktības kvantificēšanā un ģeneralizācijā, kā arī izstrādātās metodikas demonstrēšanā ar potenciāli nozīmīgu praktisku pielietojumu – reģionālā Baltijas artēziskās baseina mēroga hidroģeoloģiskā modeļa MOSYS V1 kvartāra nogulumu segas materiālu īpašību izlabošanā.

Lai sasniegtu šo mērķi, ir nepieciešams izvirzīt sekojošus uzdevumus:

- atlasīt eksperimentu veikšanai nepieciešamos datus un veikt datu pirmapstrādi urbumu griezumu struktūras salīdzināmības nodrošināšanai;

- izveidot urbumu griezumu transformācijas metodi tā, lai vienkāršā simboliskā formā kodētu informāciju par slāņu biezumu, litoloģijas tipu un novietojumu attiecībā pret citiem slāņiem (secību) urbuma griezumā;
- veikt literatūras studijas;
- izvēlēties programmatūras komponentus dažādu darba uzdevumu veikšanai;
- izdalīt telpiskos klāsterus;
- veikt telpisko klāsteru atbilstības pārbaudi ar hidroģeoloģiskā modeļa aprēķiniem;
- nepieciešamības gadījumā veikt inkrementālus uzlabojumus eksperimentālajai darba gaitai;
- veikt pārskata rezultātu prezentācijas materiālu sagatavošanu, un
- izdarīt secinājumus par iegūtajiem rezultātiem, to pielietojamību, trūkumiem un potenciālajiem uzlabojumiem.

Pēdējie trīs uzdevumi ir veicami paralēli tām darba tehniskajām daļām, kas producē starprezultātus un kurām atšķiras tehniskie darba komponenti.

3. TEORĒTISKĀ BĀZE UN LITERATŪRAS APSKATS

Publikācija, kas pašos pamatos formulē šajā darbā nozīmīgo vietu ieņemošo normalizētās kompresijas distances metriku (NKD), ir publicēta 2005. gadā un uz šī darba sagatavošanas brīdi citēta vairāk nekā 500 reizes (Cilibrasi and Vitanyi, 2005) pēc *Google Scholar* datiem, tomēr līdz šim vienīgais tās pielietojums ģeotelpiskā kontekstā ir attēlu (t.sk. tālizpētes) klasifikācijā (Roman-Gonzalez et al., 2011; Cerra et al., 2010; Stepinski et al., 2007). Tomēr NKD ir ļoti universāla metode (Wehner, 2007; Cilibrasi et al. 2004, Chen et al. 2004; Benedetto et al. 2002), un tās lietojums ģeozinātnēs ir potenciāli ievērojami daudzveidīgāks. Universāla mākslīgā intelekta metode datu telpiskai klāsterēšanai, izmantojot normalizētās kompresijas distanci ar līdzības dimensijas metriku, paver iespējas izstrādāt metodiku rajonēšanai pēc telpā izkliedētiem aprakstošiem paraugiem, tādējādi risinot problēmu, ar kuru, domājams, sastopas pētnieki daudzās jomās - kā kvantitatīvi aprakstīt sakarības, kuras bieži sākotnēji šķiet kvalitatīvi acīmredzamas. NKD ļauj ne tikai parādīt, ka sakarības ievades datos pastāv, bet arī piedāvā vienu universālu metriku to kopējā informācijas satura aplēsei. Reizē NKD metode piedāvā arī kvantitatīvās metodēs labi pamatotu veidu, kā kvantificēt ģeoloģisko nenoteiktību.

Ģeoloģiskā nenoteiktība ir aktuāla pētījumu tēma Zemes zinātnēs un hidroģeoloģiskajā modelēšanā gan no lietišķā, gan teorētiskā viedokļa, jo paver iespējas dot ieguvumu rezervuāru modelēšanā (Refsgaard et. al., 2011; Massonnat, 1999; Feyen and Caers, 2006). Tēma ir aktuāla arī hidroģeoloģiskajā modelēšanā Latvijā. Baltijas artēziskā baseina reģionālajā hidroģeoloģiskajā modelī, ESF projekta PUMa aktivitātes MOSYS ietvaros, tiek strādāts pie kvartāra nogulumu ģeoloģiskās uzbūves atspoguļojuma algoritmizācijas (Saks u.c., 2011) Baltijas artēziskā baseina reģionālajā hidroģeoloģiskajā modelī (Seņņikovs, 2011). Ģeoloģiskās nenoteiktības rakstura mainības kvantifikācija telpā paver iespējas uzlabojumiem kvartāra nogulumu īpašību modelēšanai un ģeneralizācijai, kas ļautu uzlabot infiltrācijas parametrizāciju hidroģeoloģiskajiem modeļiem kvartāra nogulumos. Kvartāra nogulumiem piemīt ļoti būtiska loma ūdens infiltrācijas sadalījuma ierobežošanai telpiskajā dimensijā. Darbs pie šādu heterogeēnu nogulumu reprezentējošo filtrācijas parametru izvēles un reprezentatīvas ģeneralizācijas ir sarežģīts un ietver daudz nenoteiktības (Gorokhovski, 2012).

Darba teorētiskā daļa ir konceptuāli iedalāma vairākās sadaļās. Pirmā daļa saistāma ar normalizētās distances aprēķinu pamatojumu un īpatnībām. Otrā daļa saistāma ar mēroga un slāņu saguluma heterogenitātes problēmu risināšanu hidroģeoloģiskajos modeļos. Trešā daļa attiecas uz klāsterēšanas metožu teorētisko bāzi un pielietojumu ģeneralizācijā. Ceturtā

apskata telpisko klāsteru izveidi, izmantojot Voronoi-Dirihlē teselāciju, kas citādi pazīstama kā Tīsena poligoni un ir pieskaitāma arī pie ģeogrāfiskās informācijas sistēmu (ĢIS) un analītiskās ģeometrijas pamatmetodēm. Visbeidzot tiek aplūkoti arī teorētiskie pamati, uz kuriem balstās MOSYS V1 hidroģeoloģiskās modelēšanas vide, rezultātu aprēķini un modeļsistēmas autokalibrācijas principi.

3.1. Normalizētās kompresijas distances metrika

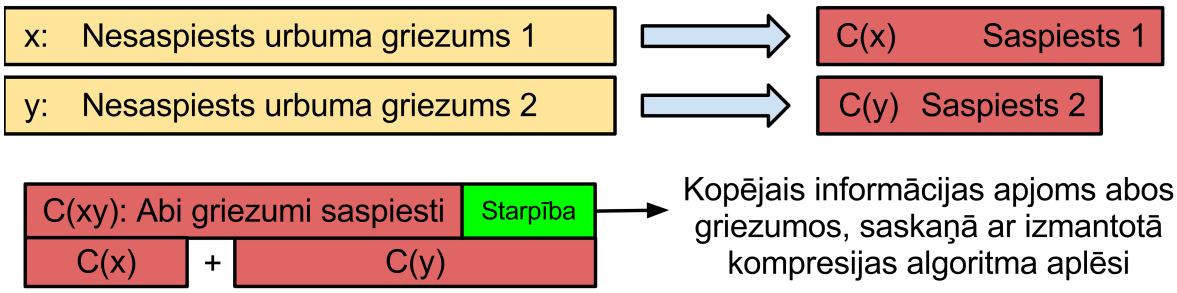
Normalizētās kompresijas distances (NKD) metrikas ieguvei tika pielietota Cilibrasi (2007) disertācijas ietvaros izstrādāto rīku kopa CompLearn.org, kas ir bezmaksas atvērtā koda programmatūra un ietver jau gatavus moduļus vairāku populāru kompresijas algoritmu izmantošanai normalizētās kompresijas distances aprēķināšanai (ibid.).

Pilns matemātisks un teorētisks skaidrojums par NKD aprēķinu ir ārpus šī maģistra darba projekta apjoma un mērķiem, un ir lieliski aplūkots pamata publikācijās par šo tēmu (Bennett et al., 1998; Li and Vitnayi, 2008; Cilibrasi and Vitanyi, 2005; Cilibrasi, 2007), bet ļoti vienkāršots skaidrojums par NKD metrikas pamatprincipiem seko šajā nodaļā.

Lai gan Šanona entropija nereti ir izejas punkts informācijas teorijai tuvu stāvošiem pētījumiem, NKD metrikas teorētiskais pamatojums sakņojas Kolmogorova kompleksitātē (Li and Vitanyi, 2008). Tā kā Kolmogorova kompleksitāte praktiski nav precīzi aprēķināma (Cilibrasi, 2007), bet ir nepieciešama to reprezentējoša vērtībā NKD metrikas ieguvei, tad Cilibrasi un Vitanji izstrādātās metodes būtība ir izmantot datu kompresijas programmas paraugu saspiešanai (Cilibrasi, 2007). NKD aprēķina formula parādīta attēlā 3.1.1. C - datu kompresijas programma, savukārt x un y domāti ievades datu paraugi. Saspiežot paraugus atsevišķi un pēc tam saspiežot tos kopā, tiek iegūta tuvināta reprezentācija Kolmogorova kompleksitātei. Šo darbību jēga sakņojas uzskatā, ka, ja divus paraugus saspiežot kopā, rezultējošo izvades datu apjoms ir proporcionāli mazāks nekā, saspiežot katru paraugu atsevišķi un rezultātu izmērus summējot, tad tie attiecīgi satur vairāk informācijas viens par otru (ilustrācija attēlā 3.1.2.).

$$\mathrm{NCD}(x,y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}$$

3.1.1. attēls. **Normalizētā kompresijas distances formula** (pēc Cilibrasi, 2007, 56. lpp.).



3.1.2. attēls. **Ilustratīvs skaidrojums normalizētās kompresijas distances darbības principam** (izveidojis autors pēc Cilibrasi 2007 apraksta, izmantojot *Google Drawing*).

Normalizējot šos rezultātus pret oriģinālu ievades datu paraugu izmēriem, ir iespējams iegūt relatīvu kopējās informācijas daudzuma tuvinājumu tajos. Šis tuvinājums, protams, ir ļoti atkarīgs no universālā bezzudumu datu kompresijas algoritma, kas tiek pielietots paraugu saspiešanai. Izstrādātais rīku komplekts *CompLearn* (pieejama adresē complearn.org) ļauj ērti veikt visas šīs darbības, izmantojot vairākas no populārākām universālo kompresijas algoritmu implementācijām (*gzip, bzip2, LZMA* un *PPMd*) (Cilibrasi, 2007). Šī pieeja seko Hatera un Mahonija idejai par to, ka spēja saspiest informāciju ir saistīta ar inteliģenci (Leg and Hutter, 2007; Mahoney, 1999). Populārāko datu kompresijas metožu apskats ir pieejams Mahonija 2011. gada elektroniskajā grāmatā, kas piedāvā lielisku pārskatu par populārākajām un veiksmīgākajām metodēm un to darbības principiem (Mahoney, 2011). Normalizētā kompresijas distance var tikt uzskatīta par universālu mākslīgā intelekta metodi, kas ļauj atrast sakarības un pasterizēt ievades datu paraugus pēc līdzībām to informācijas saturā. Lai gan vārdu salikums "mākslīgais intelekts" ir plaši interpretējams un laika gaitā priekšstati par to, kas tad "īsti" ir mākslīgais intelekts, ir būtiski mainījušies (Legg and Hutter, 2007), NKD pamatoti pretendē uz devumu mākslīgā intelekta pētījumos. Ja piekrīt Mahonija tēzei, ka informācijas kompresijas spēja ir cieši saistīta ar inteliģenci, tad NKD ir mākslīgā intelekta risinājums (Mahoney, 1999). Austrāliešu matemātiķis Haters, kurš strādā pie universāla mākslīgā intelekta definēšanas (Legg and Hutter, 2007), piedāvā materiālu atlīdzību par procentuālu uzlabojumu attiecībā pret iepriekšējo labāko rezultātu datu kompresijas programmu autoriem – Hatera balvu (Hutter, 2006).

Vairums datorlietotāju izmanto kompresijas programmas ikdienā, bet to darbības principiem un nozīmei mākslīgā intelekta kontekstā, iespējams, nevelta daudz uzmanības. Tomēr pastāv iespēja, ka šī, nu jau relatīvi ikdienišķā, daļa no mūsdienu dzīves pieder pie fundamentālām iestrādnēm pretī universālam mākslīgajam intelektam. NKD ir metrika, kas ļauj tās ērti izmantot praktisku ikdienas pētniecības problēmu risināšanai neparametriskai mašīnapamācībai.

Kā otra iespējamā problēma būtu minamas kļūdas ievades datos. Tā kā ievades materiāls ir apkopots un ievadīts manuāli, ir paredzams, ka tajā ir sastopamas kļūdas - iztrūkstoši ierakstu lauki, kļūdaini ieraksti un kļūdains pieraksts. Tomēr iepriekšējie pētījumi norāda uz to, ka NKD metrika ir samērā noturīga pret troksni ievades datos, un sintētiskos testos pat ar palielinātiem trokšņa līmeņiem ievades datos būtiski nemainīja klāsterēšanas rezultātus, kas veikti, izmantojot NKD metodi (Cebrián et al., 2007).

Kritisku skatu normalizētās kompresijas distances metodei veltī arī Cebrian et al. 2005., taču tas kliedē bažas par vēl vienu potenciālu problēmu NKD izmantošanā ģeoloģisko urbumu datu klāsterēšanā. Šis darbs īpaši vērš uzmanību uz paraugu izmēru ietekmi uz NKD vērtībām, kas ir saistīta ar aprēķina ieguvei izmantoto universālo kompresijas algoritmu parametrizāciju. Autori secina, ka NKD metrikas rezultātus būtiski kropļo paraugu izmēri, kas pārsniedz izmantotās kompresijas programmas bloka/loga izmēru tām programmām, kam šāda parametrizācija ir paredzēta (bzip2, gzip). Tajā pašā laikā autori apstiprina, ka testā izmantotais PPM tipa kompresors, kas izmanto statistisko modelēšanu datu saspiešanai, necieš no šīs limitācijas. Vēl jo vairāk - neviena no metodēm būtiski necieš no mazu ievades paraugu izmantošanas NKD aprēķinam (Cebrian et al., 2005). Šī ir ļoti noderīga atziņa, jo urbumu vertikālo griezumu apraksti, kas izmantoti kā ievades dati NKD aprēķinam šī darba ietvaros, veido 1000 baitus garus aprakstus, kas ir samērā īsi.

3.2. Klāsteranalīze

Telpiskā klāsterēšana (no angļu val. - *spatial clustering*) ir starpdisciplināra pētījumu joma ar nozīmi ģeomātikā, statistikā un nozarēs, kas saistītas ar informācijas zinātnēm, kas nodarbojas ar ļoti lielu datu apjomu glabāšanu un analīzi (Han et al., 2012). Iepazīstoties ar pieejamo pārskata literatūru par šo tematiku (Han et al., 2012; Kantardzic, 2011; Jain, 2010; Plant and Bohm, 2009; Achert, 2007; Berkhin, 2006; Lance et al., 2004; Jain et al., 1999) un pieejamo tehnisko implementāciju klāstu, ir izvēlēta hierahiskā aglomeratīvā maksimālo distanču klāsteranalīzes metode, kā realizēta *C clustering library* pakotnē (Hoon et al., 2010).

Pamatojums šai izvēlei ir daudzpusīgs, bet hierahiskajai pieejai ir vairākas priekšrocības šī darba ietvaros. Tā ir viegli uztverama un saprotama rezultātu interpretācijai, labi aprakstīta un pazīstama jau vairākus desmitus gadu. Vēl svarīgāk - tā ir deterministiska, kas nozīmē, ka vienādi parametrizēts hierahiskās klāsterēšanas rīks ar to pašu distanču matricu vienmēr dos vienādu klāsteru piederību visiem iesaistītajiem objektiem. Šis ir ļoti svarīgs apsvērums, jo ne visas klāsterēšanas metodes ir deterministiskas. Šajā pētījumā tas ir ļoti nozmīgi pie gala

rezultātu aprēķināšanas, interpretācijas un analīzes. Nedeterministiska klāsterēšanas algoritma lietojuma gadījumā, iegūstot hipotēzi noraidošus rezultātus, īpaša šaubu ēna varētu krist uz klāsteranalīzes metodes rezultātiem, no kuriem iegūti darba gaitā sagatavotie telpiskie klāsteri un no tiem atvasinātā modeļa ģeometriskā struktūra. Tāpat arī hipotēzi apstiprinošu rezultātu gadījumā nedeterministiskas klāsteranalīzes metodikas lietojums varētu būt pamats kritikai par to, ka, vairākkārtīgi atkārtojot klāsterēšanas eksperimentus ar optimizējošajiem algoritmiem, ir iespējams iegūt dažādus rezultātus. Lai to novērstu, būtu nepieciešams vai nu ļoti liels optimizācijas soļu skaits, vai arī ļoti liels skaits atkārtotu eksperimentu ar pilniem modeļsistēmas rezultātu aprēķiniem. Deterministisku rezultātu ieguve ir nozīmīga priekšrocība klasiskajiem hierahiskās klāsteranalīzes algoritmiem pretstatā k-vidējo (*k-means*) radniecīgiem klāsterēšanas paņēmieniem, kuri veic klāsteru izveidi iteratīvas optimizācijas ceļā (Han et al., 2012), un, atkarībā no norādītā iterāciju skaita, iegūtais risinājums var ne tikai nebūt deterministisks, bet prasīt arī samērā daudz datorlaika un mest šaubu ēnu uz iegūtajiem rezultātiem, kuri būtu balstīti uz šādu risinājumu.

Atšķirībā no daudzām citām metodēm, hierarhiskā algomeratīvā metode neprasa sākotnēji norādīt vēlamo klāsteru skaitu, bet ļauj izvēlēties klāsteru skaitu no dendrogrammas, kas ir pirmais un galvenais iegūstamais rezultāts šīs metodes lietojumam. Dendrogramma tiek veidota kā hierarhiska koka datu struktūra, kur katrā sazarojuma vietā ir pieejama distances vērtība starp klāsteriem, kuri savienoti šajā mezgla punktā (Hoon et al., 2010). Novērtējot to, kā mainās klāsteru skaits un struktūra pie dažāda veida datiem, eksperimenta distanču matricām un parametriem, ir iespējams izvēlēties klāsteru skaitu, pie kura dendrogrammu sadalīt fiksētā skaitā klāsteru (angļu valodas literatūrā un programmatūras dokumentācijā šo darbību mēdz dēvēt par *flattening* vai *cutting*).

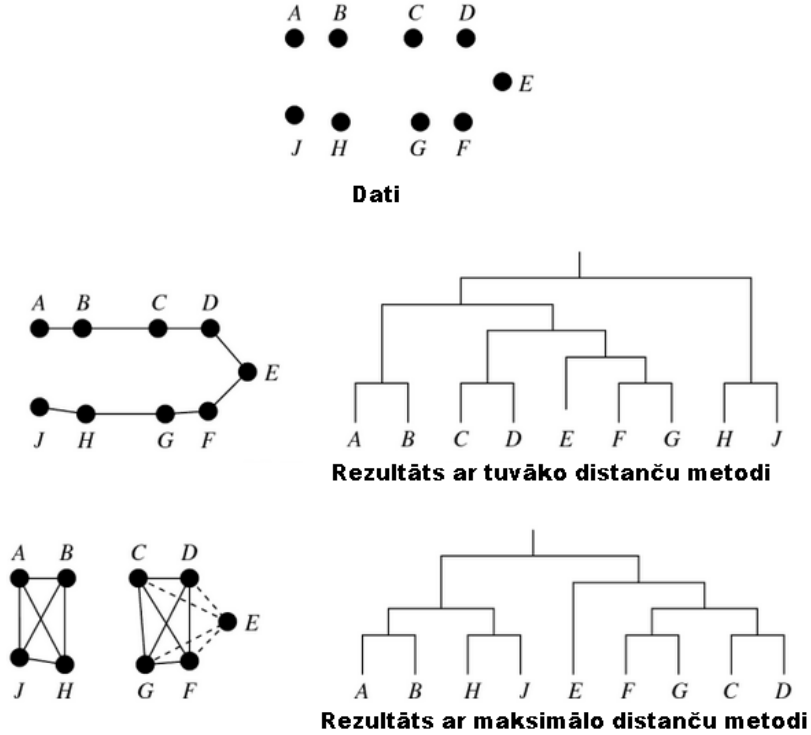
Līdz šim sagatavotie rezultāti ir balstīti uz maksimālo distanču variāciju (*complete-link*), jo ar praktiski pārbaudītajām eksperimentu distanču matricām tā veido vairāk nošķirtu klāsteru, salīdzinot ar tuvākās distances un vidējās distances variācijām.

Aglomeratīvās maksimālo distanču metodes darbības būtība īsumā ir šāda:

- katrs objekts (šajā gadījumā - urbums) tiek iekļauts savā atsevišķā klāsterī, līdz ar to sākotnējais klāsteru skaits ir vienāds ar grupējamo objektu skaitu,
- izmantojot distanču matricu, kas satur noteiktas metrikas distanču vērtības starp visiem objektiem (katrs ar katru, jeb *pair-wise* angļu val. literatūrā), tiek atrasti pēc distanču matricas kritērija tuvākie objekti un apvienoti hierarhiski augstāka līmeņa klāsterī,
- pēc katras apvienošanas šī darbība tiek iteratīvi atkārtota, ņemot iepriekšējā solī iegūtos klāsterus, līdz visi objekti ietilpst vienā klāsterī,

- katra apvienošanas darbība (kuras dēļ šo dēvē par aglomeratīvo metodi) saglabā distances kritēriju, pēc kura klāsteri tika savienoti.

Aglomeratīvā algoritma darbības rezultātā tapušo dendrogrammu jeb kokveida struktūru parasti ir iespējams vizualizēt vai nu klāsterēšanas funkcionalitāti nodrošinošajā programmatūrā, vai importēt kādā atvērtu standartu atbalstošā vizualizācijas vidē (Hoon et al., 2010). Praksē dendrogrammas analīze palīdz pieņemt lēmumus par optimālo klāsteru skaitu, kas iegūstami, saskaldot (angļu val. mēdz dēvēt par *flattening*) iegūto klāsteru modeli fiksētā skaitā klāsteru (Han et al., 2012). Vēl jo vairāk – inkrementāla klāsteru modeļa saskaldīšana dažādos skaitos klāsteru ir uzskatāma par datu struktūras izpētes metodi un ir plaši pielietota dažādās nozarēs, sākot ar bioinformātiku un beidzot ar ķīmiskām analīzēm (Kantardzic, 2011).



3.2.1. attēls. **Ilustratīvs salīdzinājums starp tuvākās un tālākās distances klāsteru aglomērācijas paņēmienu** (ilustrācija no Han et al., 2012, 463. lpp., tulkojis autors).

Iepazīstoties ar aglomeratīvo algoritmu darbības būtību, var rasties jautājumi par to, kā precīzi tiek veikta klāsteru aglomeratīvā apvienošana, jo šī daļa ir šķietami veicama vairākos veidos. Tā praksē arī ir, un teorijā, visbiežāk, izdala vismaz 3 pamata aglomeratīvās klāsterēšanas variantus, atkarībā no tā, kā tieši tie apvieno klāsterus katrā nākamajā solī. Lai gan nestandarta tehniskās variācijas par šo tēmu ir iespējamas vairāk nekā 3, tomēr literatūrā (Han et al., 2012; Kantardzic, 2011; Hoon et al., 2010) klasiski tiek aprakstīti tieši trīs pamata varianti: tuvākās distances metode (*single-link*), tālākās distances metode (*complete* jeb *maximum-link*) un vidējā (*average-link*). Kā parādīts attēlā 3.2.1., klāsteri tiek apvienoti

dažādos veidos atkarībā no tajos ietilpstošo elementu savstarpējām distancēm. Kā jau iepriekš aprakstīts, aglomeratīvie hierarhiskie algoritmi darbojas pakāpeniski (katrā nākamajā iterācijā), apvienojot iepriekšējās iterācijas izveidotos klāsterus pēc objektu skaita arvien lielākos (hierarhiski augstāka līmeņa un rupjāku ģeneralizāciju reprezentējošos) klāsteros. Līdz ar to, visiem aglomeratīvajiem algoritmiem ir vienādi pirmie un pēdējie soļi - kad visi objekti ir atsevišķi klāsteri un kad visi objekti ir apvienoti vienā klāsterī. Pārējos soļos, kad klāsterus veido iepriekšējā solī jau apvienotie objekti, pastāv dažādas versijas par to, kas tiek ņemts par kritēriju klāsteru savstarpējā tuvuma noteikšanai. Tuvākās distances metode apvieno tos klāsterus, kuros jau ietilpstošie objekti ir savstarpēji tuvāki. Tālākās distance metode, ar kuru veikti ģeneralizācijas eksperimenti šī darba ietvaros, darbojas attiecīgi pretēji – tiek apvienoti tie klāsteri, kuros distance starp tālākajiem tajos ietilpstošajiem objektiem ir vismazākā. Attiecīgi, nav grūti uzminēt, kā darbojas vidējās distances metode – apvienojot klāsterus, starp kuriem vidējais attālums starp tajos ietilpstošajiem objektiem ir mazāks. Izmantotā implementācija (Hoon et al., 2010) piedāvā vēl dažas variācijas par šo tēmu, piemēram, centroidālo metodi, kas katrā iterācijā aprēķina centroīdu katram klāsterim un nākamajā solī apvieno klāsterus, kuriem ir tuvākie šādi centroīdi. Taču šī uzlabotā metode neļauj parametrizēt algoritmu jau ar gatavu, iepriekš aprēķinātu distanču matricu, līdz ar to šo metodi nav tehniski iespējams savienot ar normalizētās kompresijas distances aprēķina programmu, kura kā rezultātu izvada urbumu līdzības matricu. Tomēr klāsterēšanas metodes izvēle, parametrizācija un pielāgošana šī darba vajadzībām nav tikai teorētisks darbs ar literatūru. Tas lielā mērā ir arī eksperimentāls darbs, kura gaitu nosaka tehnisko komponentu pieejamība un darbības īpatnības, kas pieminētas 4. nodaļā par metodēm un to tehniskajām piezīmēm.

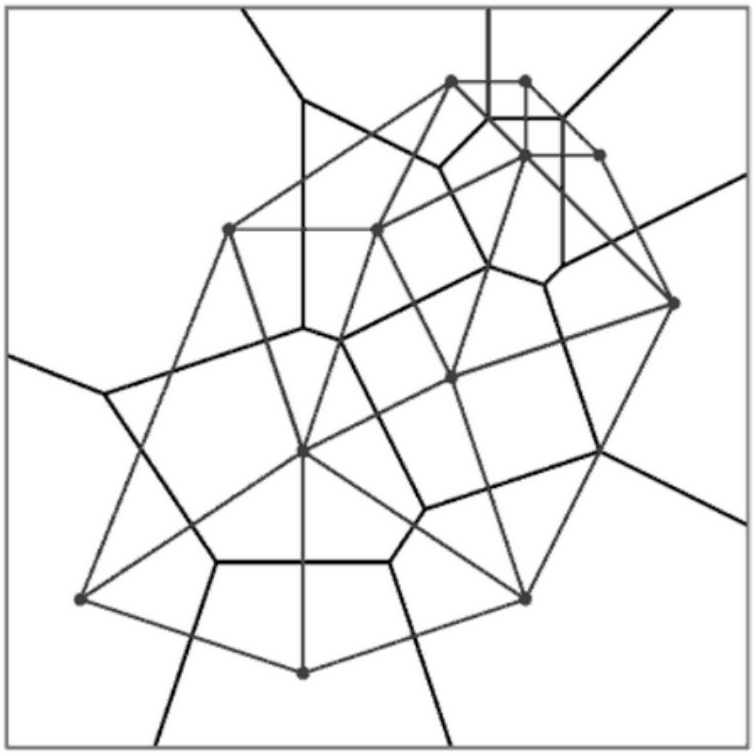
Visbeidzot, gribētos pieminēt, ka klāsteranalīze un telpiskā klāsterēšana ir piedzīvojusi daudzu jaunu metožu un uzlabojumu izstrādi pēdējo 50 gadu laikā, kurus varētu nosacīti nodalīt kā klāsteranalīzes metodikas attīstības periodu (Jain, 2010). Pēc literatūras studijām īpaši gribētos izcelt klāsterēšanas metodes, kas saistītas ar objektu blīvuma analīzi kādas konkrētas metrikas ietvaros. Pazīstamākā un arī senākā blīvuma metode ir DBSCAN (Han et al., 2012; Kantardzic, 2011; Achert, 2007), kas parasti kalpo kā piemērs blīvuma metožu apskatam literatūrā. Taču šī metode mūsdienās ir noderīga vairāk kā pamats citām metodēm un idealizēta demonstrācija blīvuma metožu darbības principiem, jo praksē ar to iegūtos rezultātus ļoti spēcīgi ietekmē parametrizācija. Tai sekojošās modernās izstrādes, kas nesen prezentētas šai tēmai veltītajā disertācijā (Achert, 2007), pieejamas klāsteranalīzes eksperimentiem veidotajā *Java* bāzētajā *ELKI* platformā (pieejama no http://elki.dbs.ifi.lmu.de/). Tomēr oriģinālajām hierarhiskās klāsteranalīzes metodēm ir

vairākas priekšrocības šī darba kontekstā: tās ir labi zināmas, plaši lietotas, tātad arī atzītas par noderīgām daudzu gadu garumā visdažādākajās nozarēs (Han et al., 2012; Kantardzic, 2011), un relatīvi vienkāršas, līdz ar to viegli saprotamas.

Ja šī darba gaitā izstrādātā metodika spētu uzrādīt vērtīgus rezultātus, izmantojot klasisko hierarhisko klāsterēšanu, tad būtu pamats uzlabot tās gaitā izmantotās metodes, t. sk. klāsteranalīzes pieeju. Tādā gadījumā modernās hierarhiskās blīvuma metodes (Achert, 2007), pēc autora domām, būtu noderīgas šādiem uzlabojumiem. Savukārt, ja iegūtie rezultāti neliecinātu par darba gaitā izstrādātās pieejas potenciālu, tad nebūtu vērts ieguldīt darbu formālai rezultātu uzlabošanai uz klāsteranalīzes metožu rēķina. Tieši tādēļ šajā darbā izmantota klasiskā hierarhiskā aglomeratīvā klāsteranalīze ar maksimālo distanču metodi.

Viens no tipiskākajiem, literatūrā minētajiem klāsteranalīzes lietojumiem ir ģeneralizācija (Han et al., 2012), kas ir klāsteranalīzes galvenais lietojums šajā darbā.

3.3. Telpisko klāsteru izveide ar Voronoi teselāciju



3.3.1 attēls. **Delonē triangulācija, pārklāta ar Voronoi poligoniem** (ilustrācija no Gallier, 2011, 274. lpp.).

Voronoi teselācija ir viena no fundamentālajām analītiskās ģeometrijas pamatmetodēm (Aurenhammer, 1991). Starp zinātniskās literatūras darbiem, kas detāli pievēršas Voronoi

teselācijai un ir plaši pieejami, gribētos izcelt divus - grāmatu, kas veltīta telpas sadalīšanas jautājumiem, izmantojot visdažādākās Voronoi teselācijai radniecīgās metodes, algoritmus un to pielietojumu (Okabe et al., 2000), kā arī apskata rakstu par Voronoi diagrammām kā fundamentālu ģeometrisko datu struktūru (Aurenhammer, 1991). Voronoi teselāciju mēdz dēvēt par Delonē triangulācijas pretstatu (ibid.)

3.3.1. attēlā Delonē triangulācijas virsotņu punkti ir nedaudz izcelti un savienoti noslēgtā triangulācijas režģī. Voronoi poligoni izvilkti, izmantojot triangulācijas tīkla virsotnes kā ģeneratorpunktus. To robežas sniedzas līdz pat attēla robežām.

Voronoi (nosaukta pēc ukraiņu tautības matemātiķa) jeb Dirihlē (pēc franču tautības matemātiķa) teselācija, izmantojot punktu koordinātas, šajā gadījumā 2D telpā kā ģeneratorpunktus, sadala šo telpu izplatības teritoriju poligonos, kur katra poligona teritorija ir tuvāk savam ģeneratorpunkta novietojumam attiecībā pret visiem pārējiem ģeneratorpunktiem. Šī darba ietvaros kā ģeneratorpunkti tiek izmantoti punkti ar urbumu novietojuma koordinātām Baltijas koordinātu sistēmā. Situācijā, kad nav zināma robeža starp dažādu objektu reprezentētās ietekmes zonām, ir pamatoti šādu robežu novilkt pa vidu starp tiem. Voronoi poligonu lietojums kopā ar hierarhisko klāsteranalīzi nav aprakstīts atsevišķā, tieši šādai metodei veltītā rakstā, bet šāds lietojums ir iepriekš minētajā, metodei veltītajā apskata rakstā (Aurenhammer, 1991, 348. lpp.).

Ir svarīgi nodalīt šī darba ietvaros izmantotos klāsteru apzīmējumus. Darba starprezultātu un rezultātu apskatā, kā arī citur tekstā, bieži tiks pieminēti divu veidu klāsteri - loģiskie un telpiskie. Telpisko klāsteru izveides solis, izmantojot Voronoi teselāciju, ir posms darba gaitā, kurā izdalās telpiskie klāsteri, par pamatu izmantojot iepriekšējā nodaļā aprakstīto loģisko klāsteru izveidi. Šis dalījums ir nepieciešams vairāku iemeslu dēļ.

Hierahiskās klāsteranalīzes rezultātā iegūtie (loģiskie) klāsteri nav tiešā veidā piesaistīti telpai tādā izpratnē, ka klāsteranalīzes algoritms tiek parametrizēts ar urbumu koordinātām tiešā veidā. Tie ir saistīti ar novietojumu telpā pastarpināti, jo urbumi, kuru līdzība pēc normalizētās kompresijas distances metrikas tiek izmantota kā ievades dati klāsteranalīzes veikšanai, atrodas noteiktās vietās ar zināmām koordinātām. Tādēļ var uzskatīt, ka to novietojums telpā tiek izmantots netieši. Hierarhiskās klāsteranalīzes rezultātā iegūtie loģiskie klāsteri, kā to prasa klāsteranalīzes teorija, ir maksimāli atšķirīgi viens no otra, un tajos ietilsptošie objekti, šajā gadījumā - urbumi, ir maksimāli līdzīgi citiem klāsterī ietvertajiem urbumiem pēc normalizētās kompresijas distances kritērija. Taču, lai iegūtu to izvietojumu telpā, kas ļautu uz telpisko klāsteru izvietojuma pamata izveidot kvartāra nogulumu reprezentāciju modelēšanas sistēmas ģeometrijā, ir nepieciešams urbumu loģisko klāsteru

piederību pārnest uz telpisko dimensiju. Šim nolūkam tiek izmantota klasiskā Voronoi teselācijas metode.

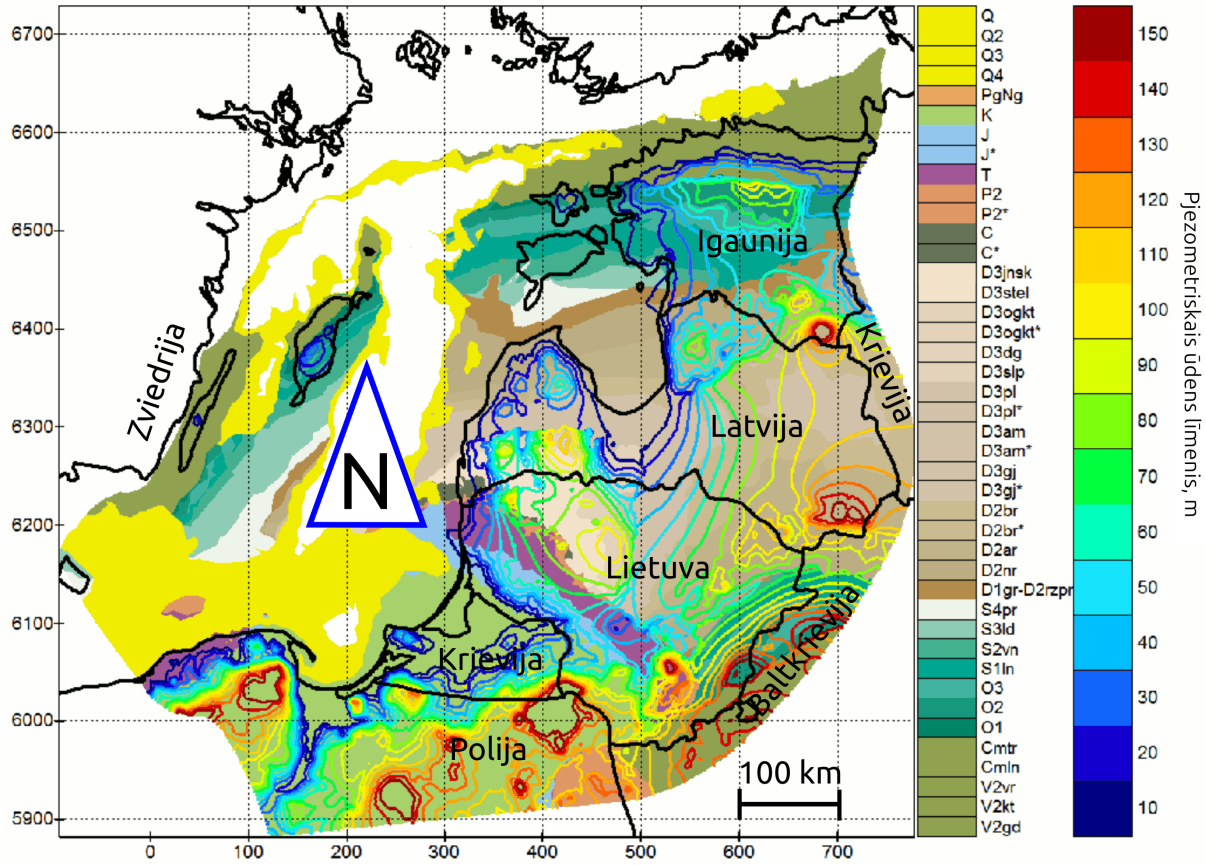
Tajā pašā laikā, pēc telpisko klāsteru izveides nav iespējams pilnībā attiekties arī no loģiskajiem klāsteriem darba tālākajā gaitā. Tieši klāsteranalīzes solī, pirms telpisko klāsteru izveides, tiek izdalīti tipveida urbumu griezumi no visiem izmantotajiem urbumiem. Pēc telpisko klāsteru ieguves, loģisko klāsteru lietojums, lielākoties, saistīts ar nepieciešamību sasaistīt katru telpisko klāsteri ar tam atbilstošo tipveida urbuma griezumu. Šo tipveida griezumu, savukārt, nosaka urbuma oriģinālā piederība loģiskajam klāsterim. Balstoties uz loģisko klāsteri, katram urbumam izdalīts atbilstošais tipveida griezums, un, izmanotojot loģiskā klāstera numuru kā atribūtu Voronoi poligoniem, katrs konkrētais telpiskais klāsteris ir nodalīts no blakusesošajiem telpiskajiem klāsteriem.

3.4. MOSYS V1 modeļsistēma

PUMa projekta ietvaros izstrādātā modelēšanas vide ir komplekss hidroģeoloģiskās modelēšanas rīku, datu un metožu kopums, kura izstrādei un saistīto metožu aprakstīšanai ir veltīti daudzi cilvēkgadi, tādēļ sistēmas visaptveroša dokumentēšana ir ārpus šī darba ietvara. Šī darba teorētiskajā daļā iekļauts konspektīvs vispārējais sistēmas teorētisko darbības principu pārskats tiem aspektiem, kas cieši saistīti ar šī darba mērķi un hipotēzes pārbaudi. Starp tiem ir kalibrācijas un filtrācijas modeļa struktūru optimizācijas principi, kvartāra nogulumu līdzšinējo reprezentāciju un vispārejie sistēmas darbības principi hidroģeoloģiskās modelēšanas teorijas ietvaros.

MOSYS V1 ir reģionāla modelēšanas sistēma, kas aptver Baltijas artēziskā baseina teritoriju, kā parādīts attēlā 3.4.1. To veido sekojoši komponenti: ģeoloģiskās uzbūves modelis (Popovs, 2011), pazemes ūdens filtrācijas modelis, vielu pārneses un transformāciju ("ķīmijas") modelis, hidroloģiskais modelis, dati un to organizācija, programmatūra datu apstrādei, priekšapstrādei un rezultātu vizualizācijai, kā arī kalibrācijas metodes un aprēķinu varianti (Seņņikovs, 2011).

Vairāki citi darbi veido plašāku pārskatu par modeļsistēmu kopumā (Bethers et al., 2011; Seņņikovs, 2011), ģeometrijas modeli (Popovs, 2011; Popovs et al., 2012), izmantotajiem datiem un ar to integrāciju saistītajām problēmām (Jātnieks u.c., 2011a; Dēliņa u.c., 2011), kā arī kalibrācijas metodiku (Virbulis et al., 2012; Stankevica, 2011; Bethers et al., 2011).



3.4.1 attēls. **Horizontālais griezums -100m dziļumā pēc absolūtajām augstumatzīmēm. Nogulumu izplatības poligoni iezīmē MOSYS modelēšanas teritoriju pa Baltijas artēziskā baseina robežu. Attēls no vizualizācijas platformas HiFiGeo.** Valstu robežas ar melnu līniju. (ilustrācija adaptēta no Seņņikovs, 2011).

Modeļsistēmas ģeometrisko modeli V1 versijai veido 42 slāņi un tā ir sintēze no urbumu datiem, ģeoloģisko virsmu kartēm, nogulumu izplatības kartēm, citu hidroģeoloģisko modeļu datiem un ģeoloģiskās interpretācijas tādos apgabalos, kur dati nav pieejami. Pamatiežu slāņu biezumu aprēķiniem tiek pielietota ģenētiskā metode, kas plašāk aprakstīta Popovs, 2011 un Popovs et al., 2012. Modeļa 3D ģeometrija tiek veidota, izmantojot neregulāro trijstūru tīklu jeb režģi, kā to dēvē MOSYS izstrādei veltītā literatūra. Katram slānim atbilstošie neregulārie trijstūru tīkli, kas veido atsevišķus modeļsistēmas slāņus un kur katram slānim uz tā virsmas tiek definēti vertikālie un horizontālie filtrācijas koeficienti, tiek savietoti vienotā 3D režģa struktūrā. Tādējādi tiek iegūts MOSYS filtrācijas modelis (Seņņikovs, 2011). MOSYS V1 versijas filtrācijas modelī visiem slāņiem tika pieņemtas katram slānim konstantas horizontālā un vertikālā filtrācijas koeficienta vērtības. Tehniski tās ir iespējams piešķirt atšķirīgas katram režģa elementam, kas šī darba ietvaros arī tiek darīts, lai veidotu uzlabotu kvartāra nogulumu ģeneralizāciju MOSYS filtrācijas modeļa 3D struktūrā.

Ar režģa elementiem, attiecībā uz šajā darbā veiktajām darbībām ar modeļsistēmas ģeometriju, ir domāti trijstūri un darbības ar tiem režģa slāņu ietvaros. Triangulācija starp slāņiem ir veikta automātiski, un to nodrošina MOSYS *Python* rīki.

MOSYS skaitliskais modelis izmanto galīgo elementu metodi ar empīrisko Darsī lamināras plūsmas vienādojumu. Izsakot šo vienādojumu katrā modeļa režģa punktā, tiek iegūta lineāro vienādojumu sistēma, kas tiek risināta ar *PARDISO Linear Solver* moduli. MOSYS hidroģeoloģiskais datormodelis ir stacionāras, nepiesātinātas pazemes ūdens plūsmas matemātiskais modelis, kas kā rezultātus iegūst pjezometrisko ūdens līmeņu telpisko sadalījumu un no tā atvasināto filtrācijas ātruma laukus (ibid.).

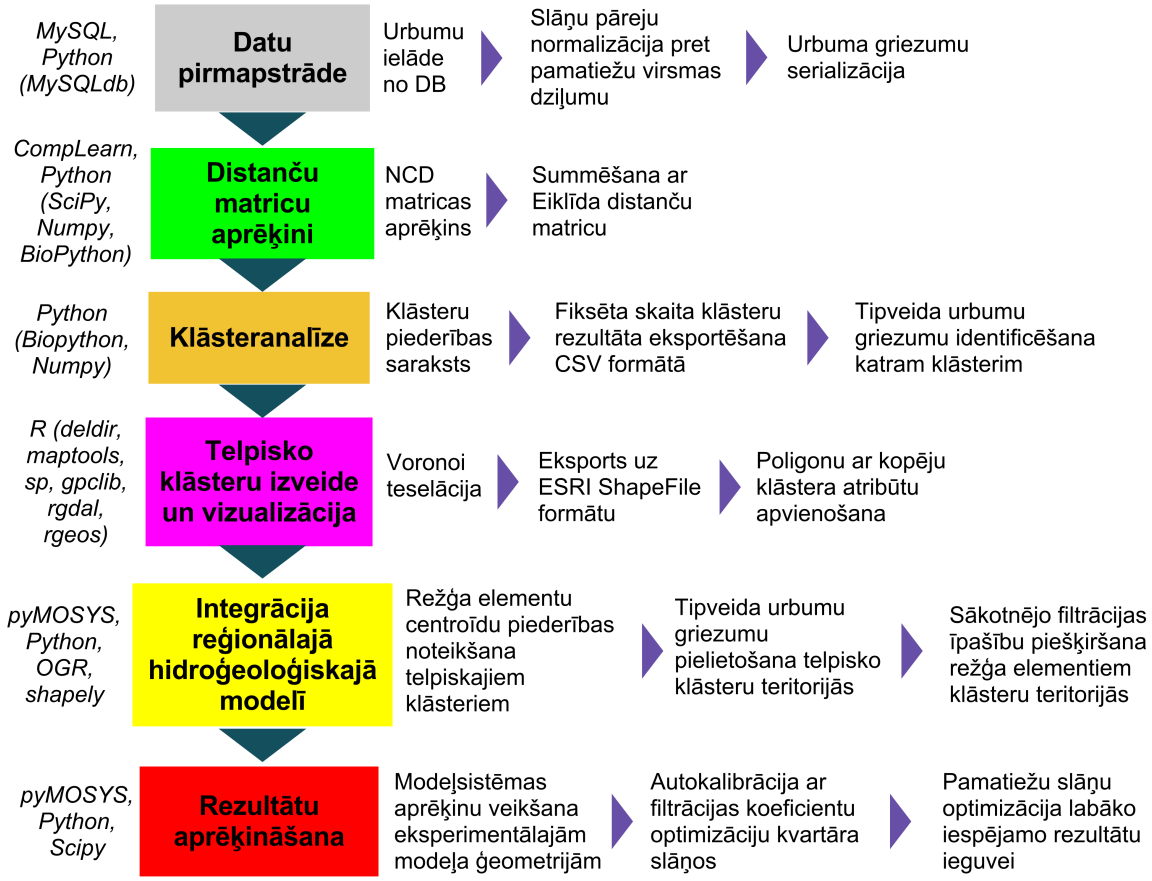
Tā kā šajā darbā tiek veidota jauna kvartāra nogulumu reprezentācija filtrācijas modeļa ģeometrijas struktūrā, tad ir nepieciešams aplūkot arī to, kāda ir MOSYS V1 oriģinālā kvartāra nogulumu reprezentācija, ar kuru tiek salīdzināti šī darba gaitā iegūtie rezultāti. Šī oriģinālā kvartāra reprezentācija ir veidota no 4 slāņiem visā BAB teritorijā (ibid.). Šos 4 slāņus veido divi pāri horizonta - sprostslāņa kombināciju. Šāds skaits kombināciju ir izvēlēts, vadoties pēc globālās slāņu iežu klašu pāreju statistikas kvartāra nogulumos (Jātnieks et al., 2012c). Latvijas un Igaunijas teritorijās šos 4 slāņus veidojošo 3 virsmu augstuma atzīmes tiek aprēķinātas, izmantojot tabulas 4.1.1. litoloģijas klašu dalījumu urbumu griezumos. Horizontu slāņus veidojošajām iežu klasēm tiek pieskaitīti smilts, grants, oļu, kūdras, augsnes un sapropeļa slāņi, savukārt, sprostslāņiem - māla, smilšmāla, mālsmilts vai aleirīta iežu klašu slāņi. Pirmā slāņa virsma sakrīt ar virsmas modeļa augstumatzīmi. Otro slāni veidojošās virsmas augstumatzīmes tiek noteiktas pēc augstumatzīmes urbuma griezuma pirmajam slānim, kas piederīgs sprostslānim pieskaitāmajām iežu klasēm. Trešā slāņa virsmas augstumatzīmes tiek noteiktas pēc urbuma griezumā otrās, horizontam pieskaitāmās klases slāņa augstumatzīmes. Savukārt, ceturtā slāņa virsmas atzīme tiek noteikta pēc otrās urbuma griezumā detektētās sprostlāņa klases slāņa augstumatzīmes, ja urbumā ir trešais slānis un šī atzīme atrodas zem tā. Pārējā modeļsistēmas teritorijā, ārpus Latvijas un Igaunijas, šo slāņu augstumatzīmes tiek sadalītas vienlīdzīgās daļās pēc kvartāra segas biezuma. Šis, pēc ģeoloģiskajiem priekšstatiem ģeneralizētais kvartāra nogulumu reprezentācijas variants, nomainīja kvartāra reprezentāciju ar vienu slāni no V0 versijas (Seņņikovs, 2011).

Visbeidzot jāpievērš uzmanība modeļsistēmas kalibrācijas metodikai. Kalibrācija ir viena no svarīgākajām hidroģeoloģiskā modeļa izstrādes gaitas sastāvdaļām (Bethers et al., 2011; Gorokhovski, 2012), un tās veikšanas principi ir nozīmīgi arī šī darba izstrādē, jo iegūtajām eksperimentālajām filtrācijas modeļa ģeometrijas struktūrām tiek vairākkārtīgi veikta kalibrācija, izmantojot MOSYS autokalibrācijas mehānismu.

MOSYS filtrācijas modelī materiālu īpašības tiek definētas ar vertikālo un horizontālo filtrācijas koeficientu, piešķirot laukiem vienādas filtrācijas koeficientu vērtības visā slānī. MOSYS V1 ir 42 slāņi filtrācijas modelī, bet kalibrācijā tiek izmantoti 25. 24 ir izdalīti pēc novērojumu urbumu stratifikācijas, kuras detalizāciju ierobežo ģeoloģisko indeksu precizitāte (daudzi horizonti novērojumu datos ir apvienoti), un 1 atbilst infiltrācijai uz virsmas (Bethers et al., 2011). Modeļa kalibrācijai ir iespējams izmantot MOSYS autokalibrācijas funkcionalitāti, kas tiek realizēta ar *Scipy L-BFGS-B* optimizācijas funkciju (Zhu et al., 1997). Kā kalibrējamie parametri tiek izmantoti vertikālie un horizontālie filtrācijas koeficienti. Šīs materiālu īpašības filtrācijas modelī tiek reizinātas ar optimizācijas algoritma piekārtotiem koeficientiem, kas ir kopēji vertikālajiem un horizontālajiem filtrācijas koeficientiem. Tas nozīmē, ka attiecība materiālu īpašībām slāņu ietvaros autokalibrācijas rezultātā saglabājas konstanta (Bethers et al., 2011). Autokalibrāciju var veikt tikai noteiktos, lietotāja izvēlētos slāņos. Pēc katras reizes, kad optimizācijas algoritms piekārto jaunus koeficientus kalibrējamajiem filtrācijas modeļa slāņiem, tiek veikts modeļsistēmas rezultātu aprēķins. Ar rezultātu šajā gadījumā tiek domāta svērtās kvadrātiskās novirzes summa pa visiem modeļsistēmas kalibrācijas slāņiem. Tā tiek aprēķināta, salīdzinot (laikā svērtos) urbumu ierīkošanas un monitoringa urbumu pjezometriskos ūdens līmeņus horizontos ar modeļa aprēķinātajiem pjezometriskajiem līmeņiem šajos horizontos. Jo mazāka novirze, jo par labāku ir uzskatāms rezultāts. Tā kā pastāv sakarība starp urbuma ierīkošanas laikā veiktajiem pazemes ūdens līmeņiem un laika dimensiju, tad, veicot aprēķinu 2000. gadam (mūsdienu scenārijam), urbumu ierīkošanas līmeņu dati tiek svērti laikā (Seņņikovs, 2011). Ņemot vērā iepriekš aprakstīto kalibrācijas veikšanas specifiku, autokalibrācija būtībā nozīmē izmaiņu veikšanu modeļsistēmas filtrācijas modeļa materiālu īpašībās, rezultātā iegūstot jaunu variantu filtrācijas modeļa ģeometrijas struktūrai, parasti ar citām materiālu īpašībām pēc optimizācijas. Šī darba izstrādē tiek izmantota optimizētā, tātad ar autokalibrāciju labākā iegūtā, modeļsistēmas filtrācijas modeļa struktūra MOSYS V1 versijai. Tajā tiek veiktas izmaiņas, lai ievietotu šī darba gaitā sagatavotās ģeometrijas iepriekšējās kvartāra reprezentācijas vietā un veiktu aprēķinus ar jaunajiem eksperimentālajiem filtrācijas modeļiem. Šī darba gaitā materiālu īpašības tiek uzdotas pa filtrācijas modeļa elementiem Latvijas teritorijā. Ir svarīgi atcerēties, ka autokalibrācijas gaitā attiecības starp elementu filtrācijas koeficientiem saglabājas nemainīgas.

4. MATERIĀLI UN METODES

Šajā darbā tiek plaši izmantota atvērtā koda programmatūra, kas padara darba gaitas reproducēšanu jebkurai ieinteresētajai pusei ievērojami pieejamāku, neatkarīgi no tā, kādām vajadzībām šāda izstrāde būtu jāatkārto. 4.1. attēls sniedz shematisku pārskatu par darba galvenajām tehniskajām daļām (krāsainie bloki), to apakšuzdevumiem (labajā pusē no krāsainajiem blokiem) un programmatūras komponentiem, kas izmantoti realizācijā (kreisajā pusē).



4.1. attēls. **Maģistra darba gaitas tehnisko daļu pārskata shēma** (sastādījis autors, izmantojot *Google Drawing* programmatūru).

Darba pašreizējā izstrāde ir pielāgota MOSYS V1 modelēšanas sistēmas specifikai, taču nepieciešamības gadījumā lielu daļu darba būtu iespējams viegli pielāgot cita veida modelēšanas risinājumam. Attiecībā uz MOSYS modeļsistēmu specifiskākā daļa ir eksperimentu rezultātā izveidoto kvartāra reprezentāciju iekļaušana filtrācijas modeļa ģeometrijas struktūrā. MOSYS V1 izmanto unikālu ģeometrijas formātu un rīku komplektu manipulācijām ar to.

Tā kā darba gaita kopumā ir kompleksa, tad šīs nodaļas sākumā tiek ieskicēti visi galvenie attēlā 4.1. parādītie darba gaitas soļi, lai sniegtu pārskatu par darba gaitas kopumu. Tālākās nodaļas apraksta katru soli sīkāk.

Vispārīgā veidā, kā galvenie ievades dati šajā darbā tiek izmantoti urbumu griezumi no DB "Urbumi" un MOSYS V1 Baltijas artēziskā baseina ģeometrijas struktūras (Popovs, 2011). Urbumu griezumi tiek pārveidoti (serializēti) un padoti savstarpējā līdzīguma aprēķinam *CompLearn ncd* utilītai (Cilibrasi, 2007) NKD aprēķina veikšanai. Rezultātā iegūtā matrica tiek izmantota kā ievade hierarhiskās aglomeratīvās maksimālo distanču klāsteranalīzes algoritmam (Hoon et al., 2010), kas tiek parametrizēts ar NKD matricu vienā variantā un ar NKD matricu, kas summēta ar normalizētu Eiklīda distanču matricu (NKD+E), otrā variantā. Šie divi matricu varianti tālāk kalpo par pamatu divām eksperimentu grupām - NKD un NKD+E klāsteranalīzes rezultātiem. Pēc klāsteranalīzes dendrogrammu reprezentācijas aprēķina tās tiek katra sadalīta (ar *cuttree* funkciju Hoon et al., 2010 implementācijā) fiksētā skaitā loģisko klāsteru, robežās no 4 - 85 klāsteriem. Katram no šiem loģiskajiem klāsteriem tiek noteikts tipveida griezums no visu tajā ietilpstošo urbumu tipveida griezumiem. Tas kopumā veido 162 klāsteranalīzes rezultātus abām eksperimentu grupām. Ņemot par pamatu katru no šiem loģisko klāsteru sadalījuma variantiem, ar Voronoi teselācijas metodi (Aurenhammer, 1991) tiek izveidoti telpiskie klāsteri Latvijas teritorijā. Savukārt, no šiem 162 telpisko klāsteru izvietojuma variantiem tiek izveidotas eksperimentālas, kvartāra nogulumu ģeneralizāciju reprezentējošas ģeoloģiskās struktūras MOSYS V1 ģeometrijas struktūrā. Visām 162 eksperimentālajām modeļsistēmas ģeometrijām tiek veikti pjezometrisko ūdens līmeņu un noviržu no novērotajiem ūdens līmeņiem baseina horizontos aprēķini. Šie rezultāti sniedz sistemātisku pārskatu par ģeneralizēto kvartāra nogulumu reprezentāciju ietekmi uz MOSYS V1 aprēķinu rezultātiem. Izmantojot šos rezultātus kā kritēriju, tika atlasītas vairākas no eksperimentālajām ģeometrijām ar labākajiem rezultātiem un tālāk kalibrētas ar hidraulisko vadītspēju optimizējošo MOSYS autokalibrācijas funkciju (Stankevica, 2011; Virbulis et al., 2012). Rezultātā 3 eksperimentālajām ģeometrijām tiek iegūta kvadrātisko noviržu summa, kas ir mazāka nekā bija iespējams iegūt, izmantojot līdzšinējo MOSYS V1 kvartāra nogulumu reprezentāciju.

4.1. Ievades dati un to pirmapstrāde

Kā galvenie izejas dati šī darba ietvaros tika izmantoti urbumu dati no Latvijas Vides Ģeoloģijas un Meteoroloģijas centra datu bāzes "Urbumi" datiem (Takčidi, 1999), kas iegādāti ESF projekta PUMa ietvaros.

Tālāk aprakstīto eksperimentālo modeļa ģeometriju izveidē izmantoti 3120 urbumu griezumu dati kvartāra nogulumos. Urbumu atlase veikta pēc vairākiem tālāk aprakstītiem kritērijiem. (I) Urbuma griezumam jāatspoguļo kvartāra nogulumu struktūra līdz pat pamatiežu virsmai - praksē tas nozīmē, ka urbumi ir atlasīti tā, lai atlasē ietilpst tikai tie urbumi, kuriem ir vēl vismaz viens slānis zem kvartāra nogulumiem. (II) Urbuma griezumā jābūt vismaz četriem, pēc tabulas 4.1.1. klasifikatora secīgi atšķirīgiem litoloģiskajiem slāņiem. Tas nenozīmē, ka visiem šiem 4 slāņiem jābūt ar 4 atšķirīgiem litoloģijas tipiem. Četras reizes pēc kārtas ir jābūt pārejām starp atšķirīgām iežu klasēm, kas izdalītas atbilstoši iežu klasifikatora tabulai (4.1.1.). Tām jāatrodas starp reljefa modeļa virsmu un pamatiežu virsmu. Urbumi tiek atlasīti ar *SQL* vaicājumu pēc aprakstītajiem kritērijiem. Tehniski tas notiek urbuma griezumu serializācijas procesa gaitā, kas aprakstīta nodaļā 4.2.

4.1.1. tabula

MOSYS iežu klasifikators un tiem atbilstošie filtrācijas koeficienti (tabula no MOSYS DB)

Klase	Iežu grupa	Horizontālais filstrācijas koeficients, m/dnn	Vertikālais filtrācijas koeficients, m/dnn
1	Smilšakmens	5	1
2	Smilšakmens, maz caurlaidīgs	1	0.5
3	Kaļķakmens, ļoti caurlaidīgs	50	20
4	Kaļķakmens	30	10
5	Kaļķakmens, maz caurlaidīgs	1	0.1
6	Dolomīts, ļoti caurlaidīgs	100	30
7	Dolomīts	30	10
8	Domerīts (maz caurlaidīgs dolomīts)	0.1	0.01
9	Māls	0.00001	0.00001
10	Smilšmāls	0.001	0.001
11	Smilšmāls-mālsmilts	0.001	0.001
12	Aleirīts-aleirolīts	0.01	0.01
13	Smilts	5	5
14	Grants, oļi	20	20
15	Ģipsis	10	10
16	Kūdra	10	10
17	Sapropelis un cita organika	0.0003	0.0003
18	Augsne	1	1
19	Cits	1	1
20	Betons	0	0

Tabulā 4.1.1. ir atspoguļots iežu iedalījums galvenajos litoloģijas tipos ar tām atbilstošajām horizontālajām un vertikālajām vadītspējām, kas ir PUMa darba grupas ekspertu viedokļu sintēze par šo vērtību pielietojamību iežu grupu ģeneralizācijai. Taču ar modeļa

autokalibrāciju veiktie aprēķini MOSYS V0 versijai parāda, ka šie koeficienti "izraudzīti ļoti veiksmīgi" (Stankevica 2011, 41. lpp). Šis klasifikators tiek izmantots urbuma griezumu iežu kodu vispārināšanai plašākās, reģionālam modelim piemērotākās iežu grupās, pēc kurām tiek piešķirtas atbilstošo slāņu filtrācijas īpašības.

Urbumu datu ielāde klāsterēšanas scenārija qclust.py ietvaros pirmapstrādē būtu iedalāma divās daļās:

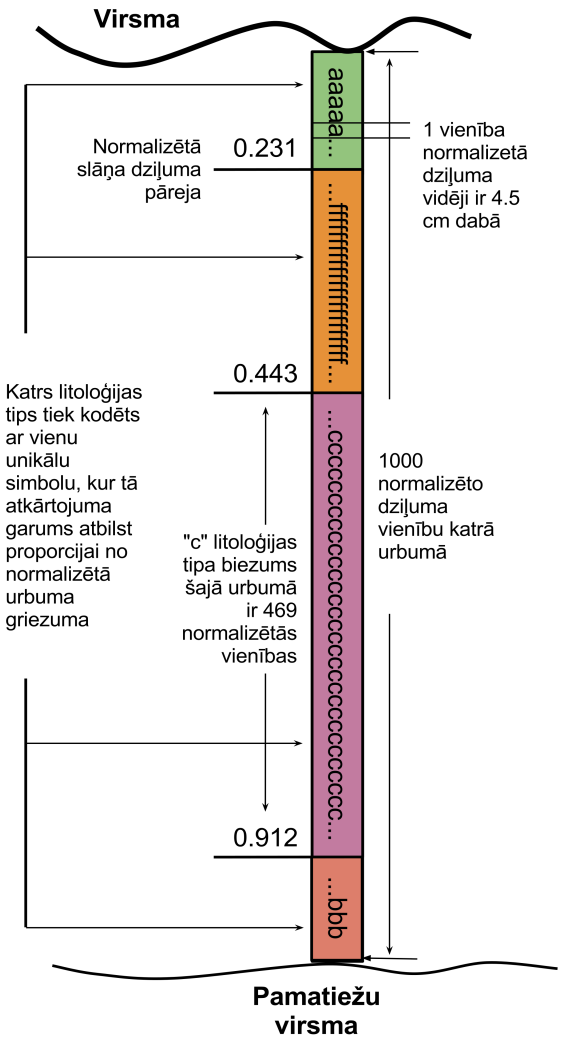
- visu urbuma kvartāra ierakstiem tiek veikts pārrēķins slāņu pāreju dziļumu vērtībām, normalizējot tās pret kvartāra segas biezumu šajā urbumā;
- urbumu numuru (datu bāzes identifikatoru) saraksts tiek ielādēts kopā ar novietojumu Baltijas koordinātu sistēmā.

Tajā pašā *Python* valodas scenārijā tika veikta urbuma griezuma ielāde katram no iepriekšējā solī izveidotā saraksta urbumiem un veikta urbuma griezuma serializācija, kuras rezultāts tika izrakstīts atsevišķā datnē uz diska. Veids, kādā notiek urbuma griezuma litoloģiskās struktūras serializācija, aprakstīts darba nākamajā nodaļā.

Otrs galvenais ievades datu masīvs ir MOSYS V1 ģeometrijas struktūra, kas tiek izmantota kā bāze, kurā tiek integrētas šī darba gaitā sagatavotās kvartāra nogulumu ģeneralizētās reprezentācijas.

4.2. Normalizētas litoloģiskās struktūras urbuma griezuma serializācija

Statistiskās konteksta modelēšanas datu kompresijas algoritmu, tādu kā šajā darbā normalizētās kompresijas distances aprēķinam izmantotais PPMd (Shkarin, 2002), priekšrocība ir spēja strādāt ar kontekstuālu informāciju, kas padara tos labi adaptētus secīguma atpazīšanai ievades datos, jo tie ir jutīgi pret kontekstuāla rakstura izmaiņām. Praktiski visi entropijas kodēšanas bezzudumu datu kompresijas algoritmi balstās uz teksta fragmentu identificēšanu ievades datos un visu sastapto fragmentu (simbolu kombināciju) uzskaiti (Mahoney, 2011). Lai padarītu urbuma griezumu salīdzināšanu piemērotāku teksta kompresijas algoritmiem, ir nepieciešams urbuma griezuma informāciju pārveidot kontekstuālā simbolu virknē, kurai jāietver šāda informācija: litoloģijas tips, slāņu biezums un savstarpējais novietojums. Datorzinātņu literatūrā par serializāciju mēdz dēvēt procedūru, kuras rezultātā kompleksas datu struktūras un objektu reprezentācijas no datora atmiņas tiek pārvērstas par secīgu simbolu virkni saglabāšanai uz datu nesēja vai pārraidei pa datortīklu (Ramanath & Datan, 2010).



4.2.1. attēls. **Pret pamatiežu virsmu normalizēta urbuma griezuma serializācijas ilustratīvs piemērs** (sagatavojis autors, izmantojot *Google Drawing*).

Urbuma numurs	Slāņa apakšas dziļums	Pret kvartāra segas biezumu normalizēts dziļums	Litoloģijas klase
11390	8	0.1	smilts
11390	12	0.151	smilts
11390	15	0.188	aleirīts
11390	24	0.301	māls
11390	28.8	0.361	aleirīts-aleirolīts
11390	56	0.703	smilšmāls
11390	58.8	0.738	grants
11390	71	0.891	smilšmāls-mālsmilts
11390	78.7	0.987	grants
11390	79.7	1	grants

4.2.2. attēls. **Urbuma griezuma piemērs ar litoloģijas klasēm un normalizētajiem slāņu pāreju dziļumiem pirms un pēc pārveides serializētā formā.** Dziļums metros, Baltijas augstumu sistēmā, normalizācija attiecīgi pret pēdējo dziļuma ieraksu (sagatavojis autors, izmantojot *LibreOffice* programmatūru un datu bāzi "Urbumi").

Lai padarītu urbumu griezumu struktūru salīdzināmu starp dažādiem urbumiem dažādās modelēšanas teritorijas vietās, urbuma slāņu dziļumu atzīmes tika normalizētas pret urbuma dziļumu kvartāra nogulumos, un slāņu pārejas tika izteiktas kā daļa no šī biezuma. Lai

tekstuāli reprezentētu normalizētos slāņu biezumus un padarītu šādus griezumus salīdzināmus, tie tika sadalīti 1000 daļās, proporcionāli pret urbuma dziļumu kvartāra nogulumos. Attiecīgi, viena veida litoloģiskais tips tika kodēts ar unikāliem teksta simboliem no ASCII kodu tabulas, atkārtojot tos tik reizes, cik nepieciešams, lai slāņa biezuma proporcija sakristu ar normalizēto biezuma vienību skaitu urbumā.

Attēlā 4.2.1. ilustrēta normalizētā urbuma griezuma serializācija darbībā, savukārt attēlā 4.2.2. demonstrēts piemērs reāla urbuma griezuma pārveidei serializētā formā. Ilustrācijā attēlotajam sintētiskā urbuma piemēram ir 4 slāņi, kas izdalīti pēc 4 atšķirīgām litoloģijas klasēm, kam piešķirti atbilstoši teksta simboli - a,b,c un f. Starp šiem slāņiem ir 3 pārejas (4. slāņa apakša ir pamatiežu virsma), kurām norādītas normalizētā dziļuma vērtības. Normalizācija šādā veidā ļauj padarīt salīdzināmus urbumus ar atšķirīgiem kvartāra nogulumu biezumiem. Līdz ar to šāds griezums ir iespējams gan urbumam, kura kvartāra nogulumos esošā daļa ir 100m bieza, gan urbumam, kur tā ir 50m bieza. Pirmajā gadījumā (100m) slāņu pārejas būs pie 23.1m, 44.3m, un 91.2m no relatīvā griezuma, bet otrajā gadījumā pie 1.55m, 22.3m un 45.6m. Šādā veidā tiek kodēti visi urbumu griezumi, kas tiek atlasīti konkrētam ģeneralizācijas eksperimentam, un katram no urbumu griezumiem piemērota šī transformācija.

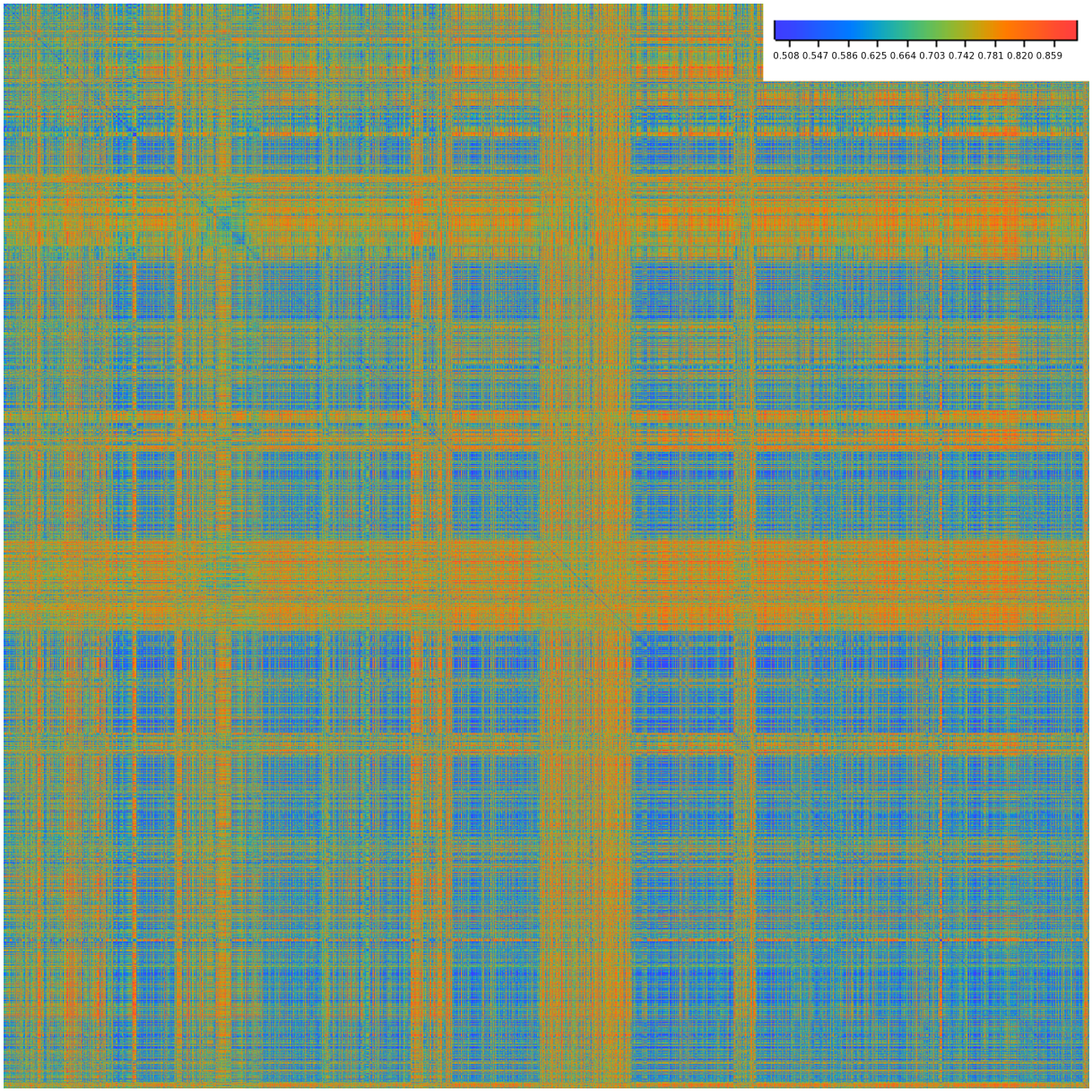
Rezultāti tika iegūti, izmantojot eksperimentu automatizācijas skriptu qclust.py (iekļauts 1. pielikumā), kas veic urbumu griezumu ielādi caur tiešu savienojumu ar urbumu datu bāzi, katru serializēto urbuma griezumu saglabājot kā atsevišķu teksta datni, kur faila nosaukumu veido urbuma numurs - datu bāzes identifikators. Katra no šīm datnēm ir 1000 baitus gara.

Šīs serializēto urbumu griezumu datnes var tālāk tikt izmantotas normalizētās kompresijas distances aprēķinam.

4.3. Eksperimentu matricu aprēķini

Normalizētās kompresijas distances vērtības tika aprēķinātas starp visiem urbumiem, izmantojot *CompLearn ncd* utilītu direktoriju režīmā (opcija -d), un rezultātā tika iegūta atdalīta teksta matrica, kuras kompakta vizualizācija aplūkojama attēlā 4.3.1. Tālāk veicamie klāsterēšanas **eksperimenti ir iedalāmi divās grupās.** Tie, kas veikti, izmantojot tikai normalizētās kompresijas distanču matricu - turpmāk grafikos un citur tekstā uz to balstītie rezultāti tiks apzīmēti ar **NKD** un klāsteru skaits ar burtu k (piemēram, NKD 12k). Otra veida eksperimenti ir balstīti uz summu NKD un Eiklīda distanču matricai, kas normalizēta pret maksimālo urbumu attālumu. Uz šo matricu summu balstītie rezultāti tiek apzīmēti ar

NKD+E un klāsteru skaits ar k burtu, iegūstot eksperimentu nosaukums tādus kā, piemēram, NKD+E 12k.



4.3.1. attēls. **Normalizētās kompresijas distanču matricas vizulizācija (angļu val. reizēm apzīmē ar terminu *heatmap*).** (izveidojis autors, izmantojot *CompLearn.org ncd* utilītu, *Python* valodas bibliotēkas *Numpy, Scipy* un *Quantum GIS* programmatūru).

Attēlā 4.3.1 attēlotā matrica aprēķināta no darbā izmantotajiem 3210 urbumu griezumiem ar PPMd kompresijas algoritmu. Katrs punkts šajā attēlā reprezentē vērtību, kura izsaka līdzību starp diviem urbumu griezumiem. Vērtības pārrēķinātas no diapazona 0-1 uz 256 krāsu pelēku toņu diapazonu ar NKD vērtību soli 0.00390625. Matrica abpus 0 vērtību diagonāles ir simetriska, tādēļ skala novietota vienā no pusēm, nezaudējot informāciju. Pēc šīs matricas ir iespējams uzskatāmāk spriest par aprēķināto NKD vērtību diapazonu. Šāda vizualizācija palīdz attēlot lielas matricas pārskatāmā veidā, jo izmēri šādas matricas attēlošanai tabulas formā prasītu daudz vairāk vietas.

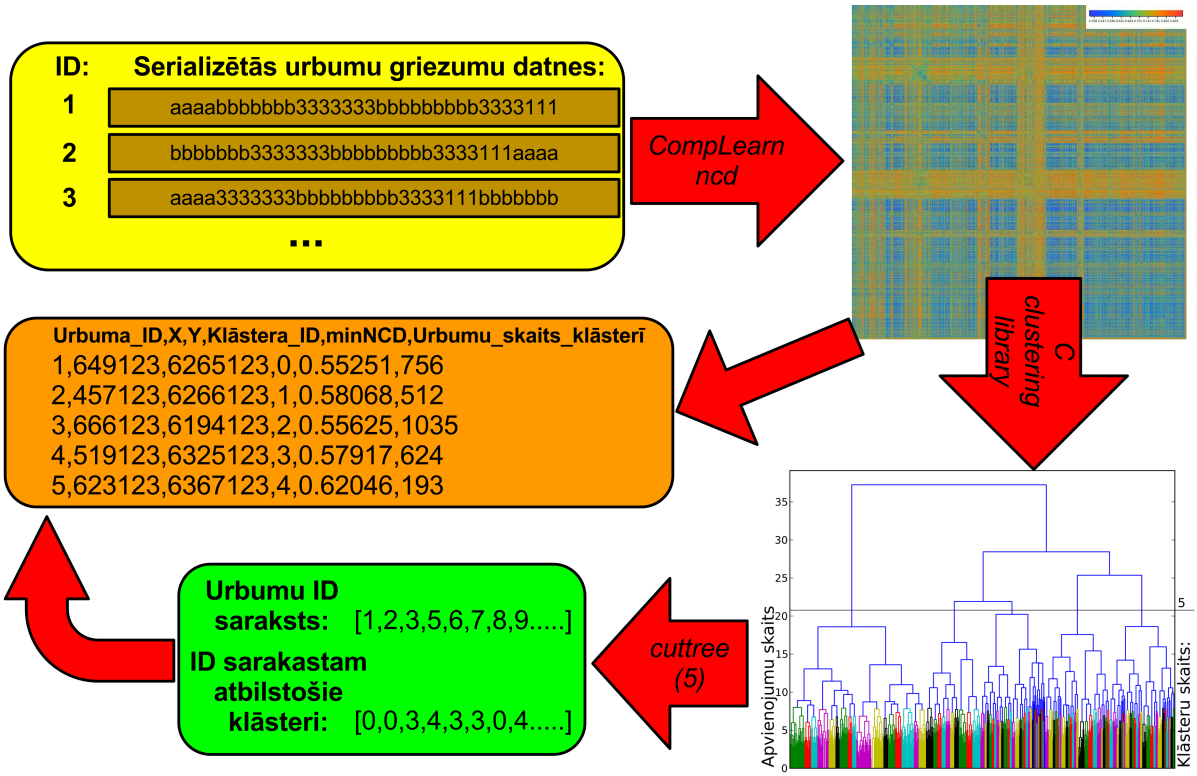
NKD+E eksperimentu grupa tika izveidota darba gaitā, iepazīstoties ar dažādiem telpisko klāsteru vizualizācijas variantiem, pirms aprēķinu veikšanas. Vizualizējot pirmos rezultātus ar NKD matricu (2. pielikums), varēja vērot ļoti saposmotu izplatību loģiskajiem klāsteriem Latvijas teritorijā. Lai piešķirtu informācijas distancei papildus ietekmi no urbumu savstarpējiem attālumiem un novērotu šādas pieejas ietekmi uz rezultātiem, tika ieviesta NKD+E eksperimentu grupa. Tās mērķis bija uzrādīt, vai informācijai par urbumu savstarpējo attālumu, papildus līdzībai NKD metrikas izteiksmē, ir kāda ietekme uz iegūstamajiem rezultātiem. Šī eksperimentu grupa kalpoja kā rezerve. Ja uz NKD matricu balstītās filtrācijas modeļa ģeometrijas struktūras nedotu labus rezultātus, bet NKD+E matrica dotu labākus, tas norādītu uz nepieciešamību kombinēt informāciju par urbumu savstarpējo novietojumu ar NKD kvantifikāciju par to litoloģiskās struktūras līdzību. Pat ja šis pārsvars būtu neliels, tā esamība, šajā gadījumā, būtu svarīgāka un varētu kalpot par pamatu uzlabotas kombinēšanas pieejas izstrādei. Šāda normalizētās Eiklīda distanču matricas summēšana ar NKD matricu, faktiski, palielina telpisko klāsteru koncentrāciju reģionos ar augstāku urbumu blīvumu, pievienojot urbumu savstarpējās litoloģiskās struktūras līdzīgumam informāciju par to attālumiem, tādējādi samazinot telpisko klāsteru saposmojumu. Attiecīgi, par vislīdzīgākajiem urbumiem pēc šīs NKD+E matricas aprēķināšanas tiktu uzskatīti tādi, kuriem ir lielāka gan litoloģiskās struktūras līdzība, gan mazāks attālums starp tiem dabā, vai arī vienai no šīm līdzībām būtu pilnībā jākompensē otras iztrūkums, lai padarītu urbumus līdzīgākus pēc šāda kritērija aglomeratīvajā klāsteranalīzē.

Lai aprēķinātu Eiklīda distanču matricu, tika ielādēts saraksts ar urbumiem un to koordinātām un padots *Scipy* bibliotēkai Eiklīda distanču matricu aprēķinam. Pēc matricas aprēķina tā tika sakārtota pēc urbuma identifikatoriem augošā secībā. Šāda pati kārtošana tika pielietota no diska ielasītajai NKD matricai, un pēc šīm darbībām tika veikta pārbaude, vai abās matricās visiem urbumiem atbilstošās kolonnas atrodas tajās pašās pozīcijās, lai nepieļautu, ka distanču informācija tiek summēta starp atšķirīgiem urbumiem, kas novestu pie nekorektiem rezultātiem. *Scipy* bibliotēka tika izmantota arī 4.3.1. attēlā redzamās matricas vizualizācijai.

Pēc šīm operācijām eksperimenta matricas tika padotas klāsterēšanas algoritmam divās kārtās: vienā klāsterēšana tika veikta, izmantojot NKD matricu vienu pašu, otrā - summējot NKD matricu ar Eiklīda distanču matricu. Pilna darba gaita ietvēra vismaz trīs distanču matricas – Eiklīda un NKD distanču matricas atsevišķi, kā arī to summu. Eiklīda distanču matrica atsevišķi tika izmantota kā kontrole testēšanas gaitā, jo tās rezultāti veido telpiski nepārtrauktus klāsterus un klāsterēšanas rezultātā iegūtie loģiskie klāsteri atbilst telpisko klāsteru skaitam. Iepriekš minētajām sakarībām neizpildoties Eiklīda distanču matricas

rezultātos, ir iespējams detektēt tehniskas kļūdas darba gaitā. Savukārt, NKD matricas rezultāti paši par sevi ļauj modelēšanas teritoriju sadalīt loģiskajos klāsteros un iepazīties ar to telpisko izvietojumu, kas veidojas, neņemot vērā urbumu savstarpējo attālumu kā papildus faktoru.

4.4. Klāsteranalīze



4.4.1. attēls. **Klāsterānalīzes darba gaitas vienkāršots shematisks piemērs** (izveidojis autors, izmantojot *Scipy, Numpy, C clustering library*, eksperimentu automatizācijas skriptu qclust.py un *Google Drawing* programmatūru).

Attēlā 4.4.1. parādīta pārsktāma darbības shēma, kas ataino klāsteranalīzes veikšanas gaitu. Serializēto urbumu griezumu datnes tika izmantotas kā ievades dati *ncd* utilītai, kas aprēķina NKD matricu. Šī matrica tika izmantota kā ievades dati hierarhiskās klāsterēšanas algoritmam, kas rada klāsteranalīzes dendrogrammu. Tā tiek pārvērsta konkrētā klāsterēšanas atrisinājumā ar fiksētu loģisko klāsteru skaitu, kas atbilst noteiktam hierarhiskam līmenim, kurā, nogriežot kokveida struktūras zarus, tika iegūts atrisinājums ar fiksētu klāsteru skaitu. Šeit ilustrētajā sintētiskajā piemērā dendrogramma tika pārvērsta atrisinājumā ar pieciem klāsteriem. Savukārt, eksperimentu automatizācijas scenārija ietvaros šī darbība tika atkārtota visiem klāsteru skaitļiem, sākot ar 4 un beidzot ar 85, ieskaitot. Izmantojot katra risinājuma klāsteru skaitu un NKD matricu, katram loģiskajam klāsterim tika noteikts reprezentatīvs tipveida griezums un izvadīts atsevišķā atdalīta teksta datnē, kuras ilustratīvs piemērs parādīts oranžajā laukumā.

Klāsterēšana tika veikta, izmantojot *C clustering library* (Hoon et al., 2010) implementāciju hierarhiskajai klāsterēšanai, kas sākotnēji radīta bioinformātikas lietojumiem, bet ir labi atbalstīta un plaši lietota daudzām citām vajadzībām, un nodrošina plašu atbalstu vairākām citām programmēšanas valodām, ne tikai C valodai, kurā tā oriģināli rakstīta. No šīs

funkciju bibliotēkas tika izmantota funkcija *treecluster* aglomeratīvajā maksimālo distanču režīmā. Rezultātā tika iegūta dendrogrammas elektroniskā reprezentācija, kas datora atmiņā organizēta kā kokveida datu struktūra (*tree*). Pati par sevi šī reprezentācija nav tik noderīga kā iespēja to sadalīt klāsterēšanas risinājumos ar iepriekš norādītu klāsteru skaitu, izmantojot funkciju *cuttree*. Tas tika darīts atkārtoti, tādējādi radot iespēju sekot izmaiņām klāsterēšanas rezultātu struktūrā un radot priekšnosacījumus rezultātu kopai, ar kuru parametrizēt reģionālo hidroģeoloģisko modeli, lai novērtētu, kā loģisko klāsteru skaits ietekmē modeļsistēmas aprēķinu rezultātus.

Jāpiebilst, ka no *Scipy* bibliotēkas tika izmantota hierarhiskās aglomeratīvās maksimālo distanču klāsteranalīzes funkcionalitāte, lai eksportētu dendrogrammas vizualizācijas SVG formātā. Tas nozīmē, ka dendrogrammas vizualizācijas saglabāšanas vajadzībām klāsteranalīze tika vēlreiz veikta ar *scipy.cluster.hierarchy.linkage* pakotni tikai tādēļ, lai vizualizētu dendrogrammu. Nepieciešamību šādi rīkoties nosaka tas, ka *C clustering library* implementācija spēj eksportēt dendrogrammu tikai tad, ja tiek ievērota specifiska darba gaita, kas nav pieejama gadījumā, ja klāsteranalīze tiek veikta ar jau iepriekš aprēķinātu matricu, kā tas tiek darīts šī darba vajadzībām. Tādēļ dendrogramma tika iegūta ar *scipy.cluster.hierarchy.linkage* pakotni, kas šādu lietojumu pieļauj. Protams, vienkāršības labad būtu iespējams izmantot klāsteranalīzei tikai *scipy.cluster.hierarchy.linkage* pakotni, taču praksē tā ir daudz lēnāka nekā Hoon et al., 2010. gada implementācija, un šī problēma atklājās darba vēlākās stadijās, tadēļ tehniski klāsteranalīzes veikšana ar *C clustering library* ir saglabāta, un *scipy.cluster.hierarchy.linkage* tiek izmantota tikai dendrogrammas vizualizācijai.

Lai panāktu šādu mērogojamu risinājumu, šī darba ietvaros tika izstrādāts eksperimentu automatizācijas scenārijs qclust.py *Python* valodā (1. pielikums), kas cikliski sadala rezultējošo dendrogrammu 4-85 loģiskajos klāsteros, no kuriem nākamajā darba gaitas solī tika izveidoti telpiskie klāsteri. Šāds klāsteru skaits ir izvēlēts kvalitatīvi, iepazīstoties ar iegūtajiem rezultātiem izstrādes gaitā, lai reprezentētu abas iespējamās galējības ar pietiekošu rezervi - ļoti rupju dalījumu mazā skaitā klāsteru, tāpat kā ļoti sīki saskaldītu teritoriju lielākā skaitā klāsteru. Loģiskie klāsteri, kas tika iegūti šajā solī, atšķiras ar to, ka tie nav piesaistīti urbumu telpiskajam izvietojumam, bet tiek iegūti tiešā veidā no klāsterēšanas algoritmam nodotās distanču matricas.

Klāsteranalīzes rezultātā tika iegūts saraksts ar katra urbuma numura piederību kādam no loģiskajiem klāsteriem atbilstoši *cuttree* funkcijas parametram, kas saskalda klāsteranalīzes dendrogrammu pieprasītajā skaitā klāsteru. Šis saraksts tiek izvadīts uz diska kā atdalīta teksta starprezultātu datne ar sekojošām kolonnām - urbuma numuru, X un Y koordinātām Baltijas

koordinātu sistēmā un urbumam atbilstošā loģiskā klāstera numuru pēdējā kolonnā. Šādas datnes tika sagatavotas visiem loģiskajiem klāsteru skaitļiem, sākot ar 4 līdz pat 85, un līdz ar to tika iegūti 85 starprezultāti katrai no divām eksperimentu matricām (NKD un NKD+E).

Tāpat vēl viena ļoti nozīmīga darbība tika veikta šajā pašā solī – tipveida urbumu izvēle starp visiem urbumiem katrā no loģiskajiem klāsteriem. Tam tika izmantota NKD matrica, no kuras katram klasterēšanas risinājumam tika atlasītas tikai tās matricas kolonnas, kuras reprezentē normalizētās kompresijas distances starp konkrētajā klāsterī ietilpstošajiem urbumiem. Visām kolonnām tika aprēķinātas vidējās vērtības, tādējādi iegūstot kvantitatīvu kritēriju, kas norāda katra urbuma vidējo līdzīgumu citiem pēc vidējā aritmētiskā. Urbuma, kuram šis vidējais ir vismazākais, griezums tika pieņemts kā materiāla struktūra visa telpiskā klāstera teritorijā. Ar materiāla struktūru šajā gadījumā tiek saprasta pret kvartāra segas biezumu **normalizētā slāņu pāreju savstarpējā biezumu attiecību secība slāņu litoloģijas klasēm,** atbilstoši tabulas 4.1.1. klasifikatoram. Lai izvadītu informāciju par tipveida griezumiem, tika izveidota atdalīta teksta datne, kas satur tipveida urbuma griezumu ierakstu rindas katram loģiskajam klāsterim, šī urbuma koordinātas un minimālo vidējo aritmētisko vērtību, pēc kuras šis urbums tika pieņemts kā tipveida griezums konkrētajam klāsterim.

4.5. Telpisko klāsteru izveide

Lai no loģiskajiem klāsteriem iegūtu telpiskos klāsterus, tika izmantota viena no klasiskām analītiskās ģeometrijas metodēm – Voronoi teselācija, kas sīkāk aprakstīta šai metodei veltītajā teorētiskās daļas apakšnodaļā. Voronoi poligonu izveidei kā ievades dati tika izmantoti katra urbuma novietojums koordinātu sistēmas plaknē. Šajā kontekstā tos var dēvēt arī par ģeneratorpunktiem Voronoi teselācijas izveidei.

Urbumu novietojumi tika ielādēti caur datu bāzes savienojumu datu ielādes un pirmapstrādes solī, kas aprakstīts nodaļā 4.1. Eksperimenta automatizācijas scenārijs qclust.py, kas veic klāsteranalīzi loģisko klāsteru ieguvei, izraksta uz diska datni ar urbumu numuriem, to koordinātām Baltijas koordinātu sistēmā un loģiskā klāstera numuru katram no analīzē iesaistītajiem urbumiem (piemērs 4.1.1. attēlā). Iegūtais starprezultāts tika izmantots kā ievade šajā darba solī, kur no urbumu un loģisko klāsteru saraksta tika izveidoti telpiskie klāsteri Latvijas teritorijai.

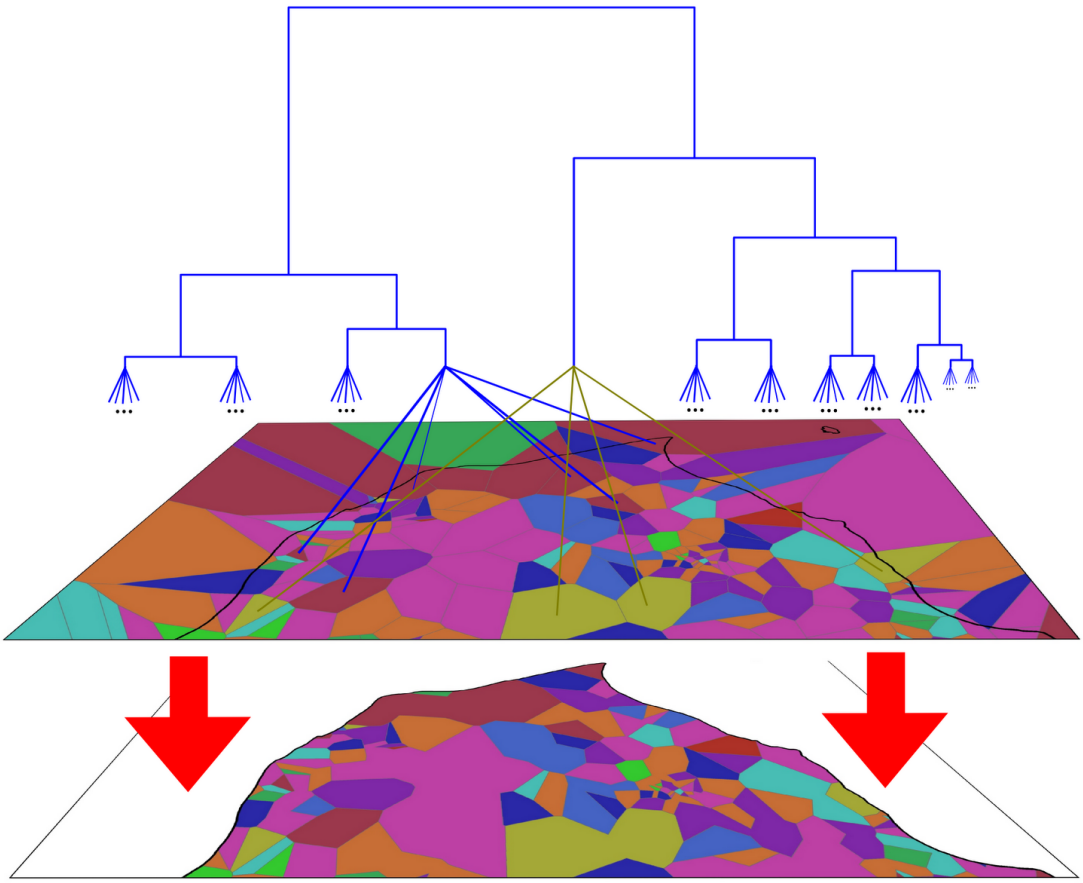
Atdalītā teksta datne ar urbumu sarakstu tika ielādēta datora atmiņā kā tabula, un urbumu koordinātas tika izmantotas kā ievades dati *deldir* bibliotēkas tāda paša nosaukuma funkcijai, kas veic Voronoi teselāciju no urbumu novietojuma ģeneratorpunktiem. Iegūtais

rezultāts tika pārveidots *R* telpisko datu formātā kādu realizē *sp* pakotne. Rezultātā tika izveidoti Voronoi poligoni telpiskā formātā, kur katram poligonam kā atribūts tika piešķirts tā loģiskā klāstera numurs. Lai iegūtu rezultātu, kur robežas pastāv tikai starp tādiem poligoniem, kuri pieder atšķirīgiem loģiskajiem klāsteriem, tālāk tika veikta telpiskās analīzes "izšķīdināšanas" darbība (*Dissolve*), izmantojot *maptools* pakotnes funkciju *unionSpatialPolygons*, kas ir vispārēja telpisko poligonu agregāciju operācija, kurai par parametru tiek uzdoti klāsteru piederības poligonu identifikatori. To sakritība tika izmantota par kritēriju blakus esošu poligonu apvienošanai vienā poligonā, dodot rezultātu, kādu ĢIS lietojumprogrammatūrā parasti iegūst ar *Dissolve* telpiskās analīzes funkciju. No iegūto rezultātu poligonu slāņa tika izgriezti poligoni Latvijas teritorijā, izmantojot Envirotech GIS Latvija 9.0 Latvijas robežas vektordatu slāni un *rgeos* pakotnes funkciju *gIntersect*. Tāpat kā *unionSpatialPolygons* funkcija, *gIntersect* nosaukums sakņojas matemātiskajā kopu teorijā, kur *Union* no angļu valodas tulkojas kā kopu apvienojums un *Intersection* - kā šķēlums. Kopu teorija ir daudzu telpiskās analīzes metožu pamatos, taču ikdienas ĢIS lietojumprogrammās šīs darbības ir pierasts dēvēt par izšķīdināšanas (*Dissolve*) un izgriešanas (*Clip*) darbībām, lai gan tās ir funkcionāli ekvivalentas šī darba vajadzībām. Tā kā *R* valoda ir orientēta uz statistisko datu apstrādi, tad tās ietvaros izmantotā terminoloģija mēdz būt vairāk radniecīga matemātikas zinātnēm.

Šo darbību rezultātā iegūtie telpiskie klāsteri tika izvadīti *ESRI Shapefile* formāta datnē, kas satur telpisko klāsteru robežas Latvijas teritorijas ietvaros un katram telpiskajam klāsterim atbilstošā loģiskā klāstera numuru. Šis numurs ir nepieciešams, lai tālākajā darba gaitā saistītu katru telpisko klāsteri ar tam atbilstošo tipveida urbuma griezumu, kas tika izdalīts klāsteranalīzes gaitā un ir atšķirīgs katram loģiskajam klāsterim, kā aprakstīts iepriekšējā nodaļā. *Shapefile* rakstīšana tika veikta, izmantojot pakotni *rgdal*, kas ļauj izmantot GDAL/OGR bibliotēku telpisko formātu lasīšanai (GDAL, 2012) un rakstīšanai *R* valodas scenārijos.

Visa šī darba daļa tehniski tika realizēta *R* statistikas valodā. Tās izvēle šīs darba daļas realizācijai bija saistīta ar salīdzinoši labo atbalstu tieši Voronoi teselācijas izveides un saistīto operāciju automatizēšanai. Šī automatizēšana ļauj augstāk minētās darbības atkārtot tūkstošiem reižu, kas atvieglo testēšanu darba gaitā un ļauj novērtēt iegūtos rezultātus ar maksimāli daudzveidīgiem parametriem. *R* valoda ļauj ne tikai veiksmīgi un ļoti ātri veikt aprēķinus telpiskiem datiem, bet arī eksportēt iegūtos rezultātus uz ĢIS formātiem, tādiem kā plaši lietotais *ESRI Shapefile,* uz kuru tika eksportēti šajā solī iegūtie rezultāti. Šie rezultāti tika izmantoti par pamatu ģeneralizētās litoloģiskās struktūras integrēšanai reģionālajā hidroģeoloģiskajā modelī MOSYS V1 kā aprakstīts nākamajā nodaļā. *ESRI Shapefile*

izmantošana ļāva arī darba gaitā viegli pārskatīt ar darba automatizācijas scenārijiem iegūtos starprezultātus un vizualizēt telpisko klāsteru slāņus ĢIS vidē.

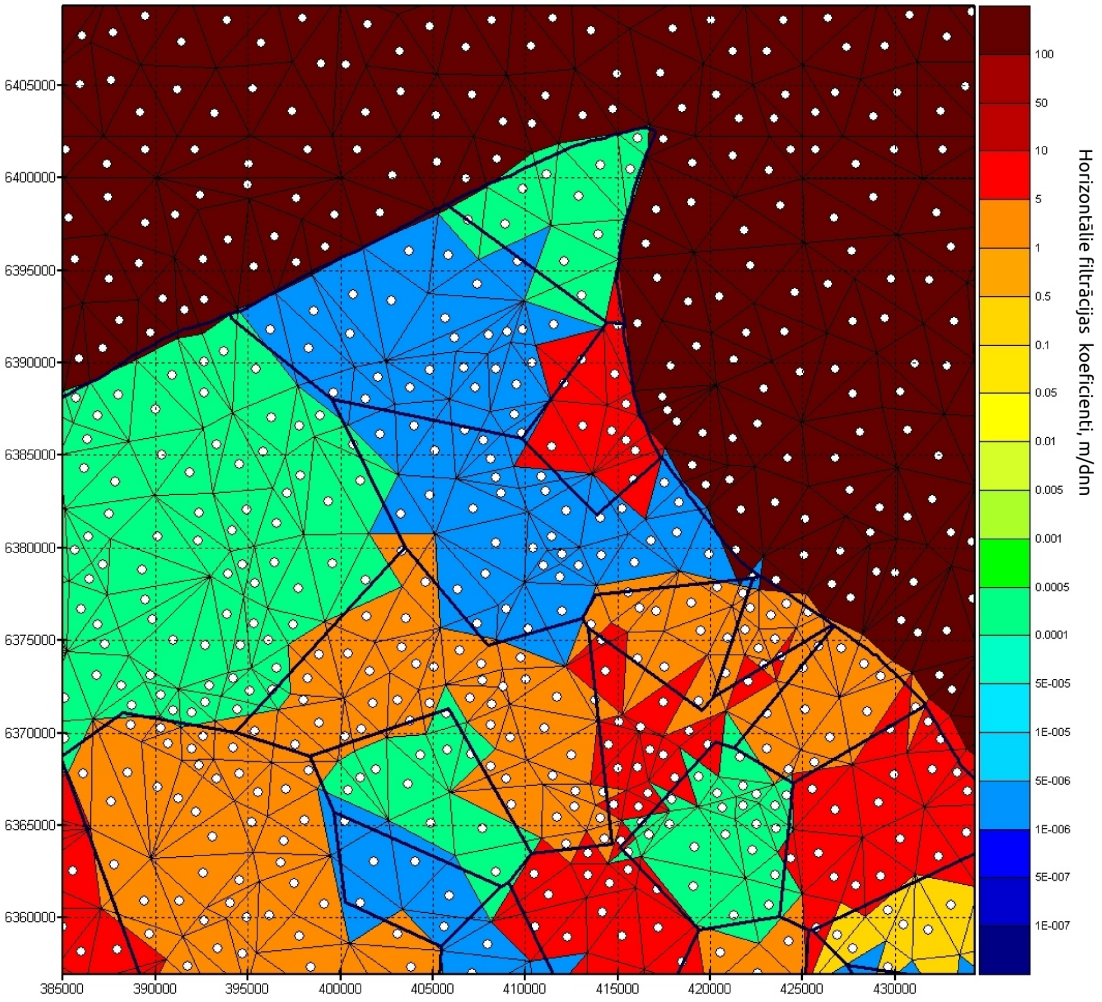


4.5.1. attēls. **Telpisko klāsteru izveides metodikas ilustrācija** (izveidojis autors, izmantojot *Scipy, Numpy, C clustering library*, eksperimentu automatizācijas scenārijus qclust.py, ddac.r un *Inkscape* vektografikas redaktoru).

Attēlā 4.5.1. ilustrēta telpisko klāsteru izveides loģika, ilustrējot to diviem no 12 loģiskajiem klāsteriem Kurzemes Z daļā, taču process notiek tādā pašā veidā visai Latvijas teritorijai un eksperimentiem ar visiem loģisko klāsteru skaita variantiem, sākot ar 4 un beidzot ar 85 loģiskajiem klāsteriem. Katram urbumam tuvāko teritoriju reprezentējošais poligons, kura robežas aprēķinātas ar Voronoi metodi, iegūst sev atbilstošā loģiskā klāstera atribūtu no klāsteru piederības rindas (ilustrēts 4.4.1. attēlā), kas iegūta no klāsteranalīzes funkciju bibliotēkas *C clustering library*. Šie atribūti tika izmantoti kā kritērijs, pēc kura blakusesoši Voronoi poligoni ar kopējiem loģiskajiem klāsteru identifikatoriem tika apvienoti lielākos poligonos, kas veido darba galvenos starprezultātus – telpiskos klāsterus. Telpisko klāsteru teritorijā konkrētajā hierarhiskās detalizācijas līmenī ir atšķirīgi tipiskie urbumu griezumi, un telpiskie klāsteri reprezentē šo griezumu pielietošanas teritoriju modelēšanas sistēmas ģeometrijā. Šīs ģeometrijas izveide tika veikta nākamaja solī.

4.6. Klāsteranalīzes rezultātu iekļaušana modeļa ģeometrijā

Iepriekšējā solī iegūto rezultātu iekļaušanai modeļa ģeometrijā ir izstrādāta tālāk aprakstītā darba gaita. Tehniski darba daļu, kas realizē šo darba gaitas soli Clusters_25AVGen.py skriptā, ir salīdzinoši daudz, taču būtībā tās ir iedalāmas divās lielākās daļās. Pirmajā solī tiek noteikta modeļsistēmas ģeometrijas elementu piederība telpisko klāsteru teritorijā. Klāsterēšanas teritorijā ietilpstošie modeļsistēmas 3D ģeometrijas elementi tika sakārtoti atbilstoši telpiskā klāstera tipveida urbuma griezuma struktūrai modeļa ģeometrijā. Tad tiem tika piešķirtas materiāla īpašības no ģeneralizētā iežu klasifikatora tabulā 4.1.1., kas sagatavota MOSYS V1 hidroģeoloģiskā modeļa vajadzībām.



4.6.1. attēls. **Modeļa režģis 2D skatījumā ar trijstūru režģa elementu centroīdiem un telpisko klāsteru robežām** (izveidojis autors, izmantojot HiFiGeo vizualizācijas vidi, NKD 12 klāsteru eksperimenta modeļa ģeometriju un telpisko klāsteru starprezultātu *Shapefile* formātā)**.**

Lai izveidotu ģeometrisko struktūru, kurā esošie modeļa ģeometrijas slāņi kvartāra nogulumos tiktu aizvietoti ar klāsteranalīzes rezultātā iegūto griezuma struktūru, vispirms bija nepieciešams noteikt, kuri modeļa ģeometrijas elementi pieder atbilstošajiem telpiskajiem

klāsteriem un kādi ir to tipveida urbumu numuri. Šim mērķim tika izmantota *Shapely* funkciju bibliotēka (Gillies, 2012) telpiskās analīzes darbībām apvienojumā ar *OGR* (GDAL, 2012) bibliotēku ĢIS vektordatu formātu lasīšanai.

4.6.1. attēls demonstrē telpisko klāsteru izvietojumu un modeļsistēmas elementu vertikālās hidrauliskās vadītspējas (piešķirtas pēc klasifikatora 4.1.1.) vērtības pirmajam slānim. Šīs vērtības piešķirtas saskaņā ar klāsteru tipveida griezumiem un elementu piederību klāsteriem, kas noteikta pēc to centroīdu novietojuma. Attēlā centroīdi ir baltā krāsā, trijstūru elementu malas - melnā ar šauru līniju, un klāsteru malas - ar treknu, tumši zilas krāsas līniju. Trijstūra elementu krāsas pēc horizontālās hidrauliskās vadītspējas vērtībām atbilst logaritmiskajai skalai attēla labajā pusē.

Izmantojot iepriekšējā solī *R* vidē sagatavotos telpiskās klāsterēšanas rezultātus *ESRI Shapefile* formātā, katram telpisku klāsteri reprezentējošajam poligonam tiek noteikti tajā ietilpstošie modeļa režģa elementu centroīdi. Tam tika izmantota *Shapely* šķēlumu metode (*intersects*) kompleksām objektu kolekcijām (Gillies, 2012). Šajā gadījuma kā kompleksā objektu kolekcija tiek izmantots multipoligonu objekts no telpisko klāsteru rezultāta *Shapefile* datnes. Telpiskie klāsteri ar vienādiem loģiskā klāstera atribūtiem, kas nozīmē, ka tiem ir kopējs vienāds tipveida urbuma griezums, tiek ielādēti kā multipoligoni, veidojot multipoligonu skaitu, atbilstošu tipveida urbumu griezumu skaitam. Tad šie multipoligoni tiek šķelti (izmantojot jau minēto *Shapely intersects* metodi, kas ir radniecīga ĢIS *Clip* jeb izgriešanas funkcijām) ar elementu centroīdu novietojuma punktu slāni. Šīs telpiskās analīzes operācijas rezultātā *Shapely* kā rezultātu izvada tos elementu centroīdus, kuriem ir kopējs tipveida urbuma griezums (kas atvasināts no viena veida loģiskā klāstera). Rezultātā tiek iegūta norāde uz vēlamo materiālu struktūru (kuru reprezentē konkrēts urbuma griezuma numurs) katra režģa elementa teritorijā. Informācija par to, kuru urbumu normalizētie griezumi izmantojami kā materiālu struktūra Latvijas teritorijā esošajiem modeļa režģa elementiem, ir pamats klāsteranalīzes rezultātu iekļaušanai MOSYS V1 ģeometrijas struktūrā.

Neregulārā trijstūru tīkla (režģa) izmantošana modeļsistēmas ģeometrijas izveidē nosaka to, ka katrs punkts jebkurā no ģeometrijas slāņiem var būt kopējs vairākiem trijstūriem. Blakus esošiem trijstūru režģa elementiem, savukārt, var atšķirties iepriekš noteiktie atbilstošie tipveida griezumi. Tādējādi, ja blakus esošu trijstūra elementu teritorijās pēc iepriekšējā solī veiktās telpiskās analīzes ir noteikti atšķirīgi tipveida urbumu griezumi šo režģa elementu teritorijā esošās materiālu struktūras definēšanai, tad ir nepieciešams mehānisms, kā šādus konfliktus risināt uz pārejas starp klāsteru robežām. Situāciju sarežģī ne tikai tas, ka viens punkts var ietilpt lielā skaitā trijstūru, bet arī tas, ka tehniski katram no šiem trijstūriem var būt atšķirīgs normalizētais tipveida griezums ar atšķirīgu kvartāra nogulumu

slāņu skaitu. Organogramma, kas atspoguļota attēlā 4.6.2., ataino šī darba ietvaros izstrādāto algoritmu, kas rēķina režģa punktu augstumatzīmes eksperimentālo modeļsistēmas ģeometriju izveidei. Tas ietver konfliktsituāciju risināšanu, kad punkts pieder režģa elementiem ar atšķirīgiem tipveida griezumiem. Tāpat algoritms saglabā augstumatzīmes arī ārpus Latvijas teritorijas, kur netiek izmantota šī darba ietvaros izstrādātā kvartāra nogulumu reprezentācijas ģeneralizācija kvartāra nogulumu slāņos.

Pēc iepriekšējā solī iegūtajiem klāsteru centroīdu piederības un to tipveida griezumu numuriem, tālāk tika izveidota kvartāra reprezentācija, balstoties uz MOSYS V1 ģeoloģiskās struktūras ģeometriju. Par bāzi tika ņemta modeļsistēmas MOSYS V1 ģeometrijas struktūra ar autokalibrācijas gaitā optimizētajām materiālu īpašībām, kur kvartāra reprezentāciju atveido iepriekšējā, uz 4 slāņiem balstītā ģeneralizācija, kā aprakstīts nodaļā 3.4.

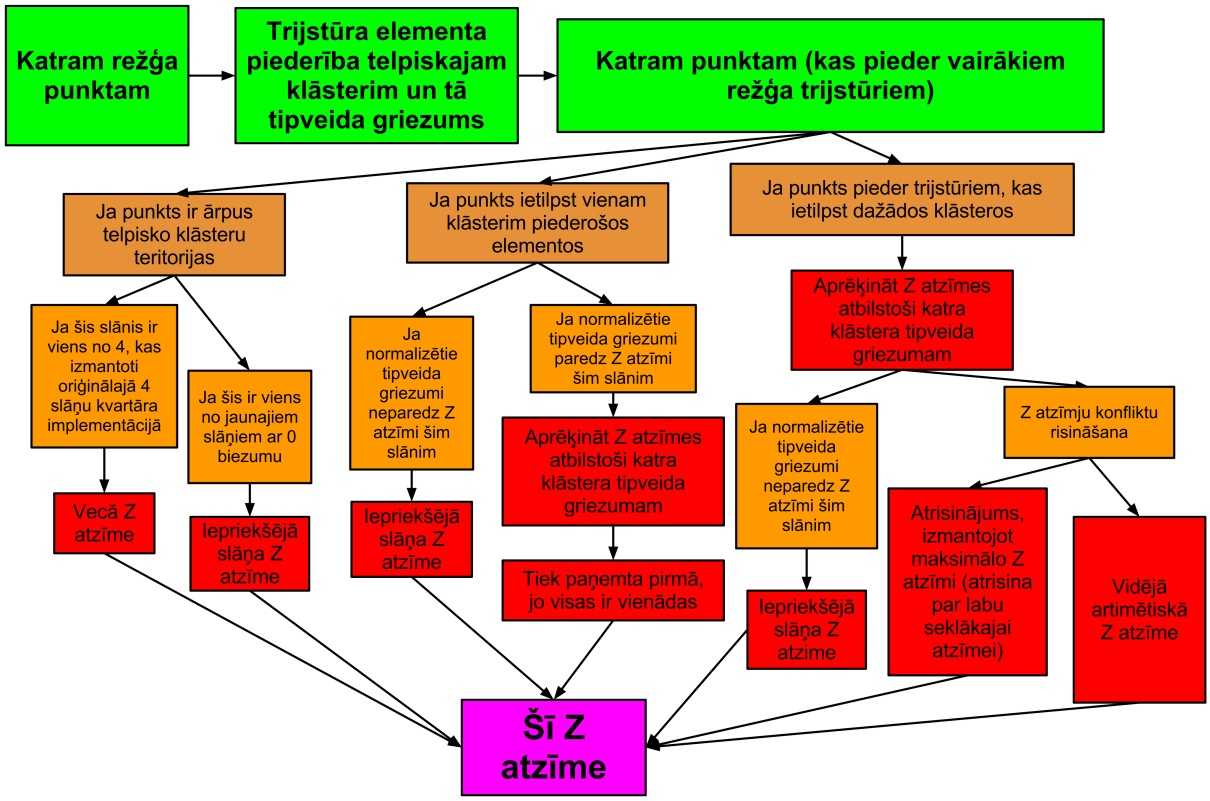
Šajā ģeometrijas struktūrā tika ievietots papildus slāņu skaits, vadoties pēc tā, cik slāņu ir tipveida griezumam ar vislielāko slāņu skaitu. Šie slāņi tika ievietoti ģeoloģiskajā struktūrā bez biezuma un kalpo kā tehniskais ietvars. Tiem tiks aprēķinātas jaunās absolūtās augstumu atzīmes režģa punktiem tad, ja to paredz konkrētā klāstera tipveida griezums (ja šis ietvara slānis eksistē griezumā). Piemēram, ja struktūrā tiek ielikti 4 papildus ietvara slāņi un ir tikai daži telpiskie klāsteri, kuru tipveida griezumi nosaka šādu slāņu skaitu, tad pārējo klāsteru teritorijā secīgi dziļākajiem slāņiem augstumatzīmes netiks aprēķinātas. Kvartāra nogulumus reprezentējošo klāsteru skaits var mainīties starp klāsteriem atkarībā no tā, cik slāņu ir to attiecīgajos tipveida griezumos. Tātad, ja pēc litoloģijas klasifikatora (4.1.1. tabula) klasēm izdalīto maksimālo slāņu skaits ir 8 un MOSYS V1 struktūrā oriģināli kvartāra nogulumiežus reprezentēja 4 slāņi, tad, veidojot jauno struktūru, tajā tika ievietoti vēl papildus 4 slāņi. To sākotnējais biezumus ir 0, jo to režģa punktu augstumatzīmes šajos slāņos sākotnēji tiek definētas vienādas ar subkvartāra virsmas augstumatzīmēm. Pārējie pamatiežu slāņi tika pārkopēti uz jauno struktūru bez izmaiņām, atskaitot jaunas slāņu identifikācijas piešķiršanu. Šis solis ir tehnisks un saistīts ar modeļa ģeometrijas formāta īpašībām, bet neatstāj ietekmi uz aprēķinu veikšanai būtiskajām materiālu īpašībām. Tās tiek pilnībā pārņemtas no oriģinālās MOSYS V1 ģeometrijas struktūras, un pēc butības mainās tikai struktūrā ietilpstošo sāņu skaits un numerācija. Tā kā jaunajiem slāņiem nav biezuma, tad, izmantojot šādu ģeometriju aprēķiniem, iegūtajiem rezultātiem būtu jābūt identiskiem ar oriģinālo MOSYS V1 ģeometrijas struktūru iegūtiem rezultātiem. Tātad, struktūrā veiktās izmaiņas šajā solī ir tikai tehniskas un rada "vietu" jaunās, kvartāra nogulumus reprezentājošās materiālu struktūras iekļaušanai modeļsistēmas filtrācijas modeļa ģeometrijā.

Tāpat pirms režģa punktu augstumatzīmju aprēķinu veikšanas ir jāveic vēl viena darbība. Tā kā jaunā kvartāra ģeometrijas reprezentācija tiek veidota, balstoties uz tipveida

urbumu griezumiem, kuru litoloģiskā struktūra ir izteikta pret kvartāra nogulumu segu normalizētās dziļumu vērtībās (kā iepriekš ilustrēts attēlos 4.2.1. un 4.2.2.), tad, veidojot kvartāra nogulumu reprezentāciju jaunajos kvartāra nogulumu slāņos, modeļsistēmas ģeometrijā ir jāveic pāreja no normalizētajām dziļumu vērtībām uz absolūtajām augstumatzīmēm. Šim nolūkam ir jāzina kvartāra segas biezums katrā modeļsistēmas režģa punktā.

Lai iegūtu kvartāra segas biezumu, virsmas reljefa modeļa slānis un kvartāra virsmas slānis tika izvadīti atsevišķās datnēs, kas satur visu režģa punktu absolūto augstumatzīmju sarakstus kā kolonnas atdalīta teksta formā. Šo apmaiņas formātu MOSYS izstrādātāji dēvē par Z datnēm, un tas ir vispārpieņemts veids, kādā tiek veikta liela daļa darbību ar MOSYS ģeometrijām. Atņemot no virsmas modeļa Z datnes (topo.z, slānis 0) vērtībām subkvartāra virsmas slāņa (subQ.z, slānis 4) Z datni, tika iegūta Z datne (biezums.z), kas satur kvartāra segas biezumu katrā modeļa ģeometrijas režģa punktā. Šīs kvartāra segas biezuma vērtības tika paturētas atmiņā kā daļskaitļu rinda, kas tika pielietota katra režģa punkta augstumazīmes aprēķinam katrā slānī, kura nepieciešamību noteica atbilstošais(-ie) tipveida griezums(-i).

Attēlā 4.6.2. atspoguļotais algoritms veido ciklu pa modeļa ģeometrijas punktiem katrā no modeļsistēmas ģeometrijas slāņiem.



4.6.2. attēls. **Organogramma režģa punktu absolūto augstumu atzīmju aprēķinam kvartāra nogulumu slāņiem modeļsistēmas MOSYS V1 ģeometrijā** (izstrādājis un izveidojis autors, izmantojot Google Drawing programmatūru).

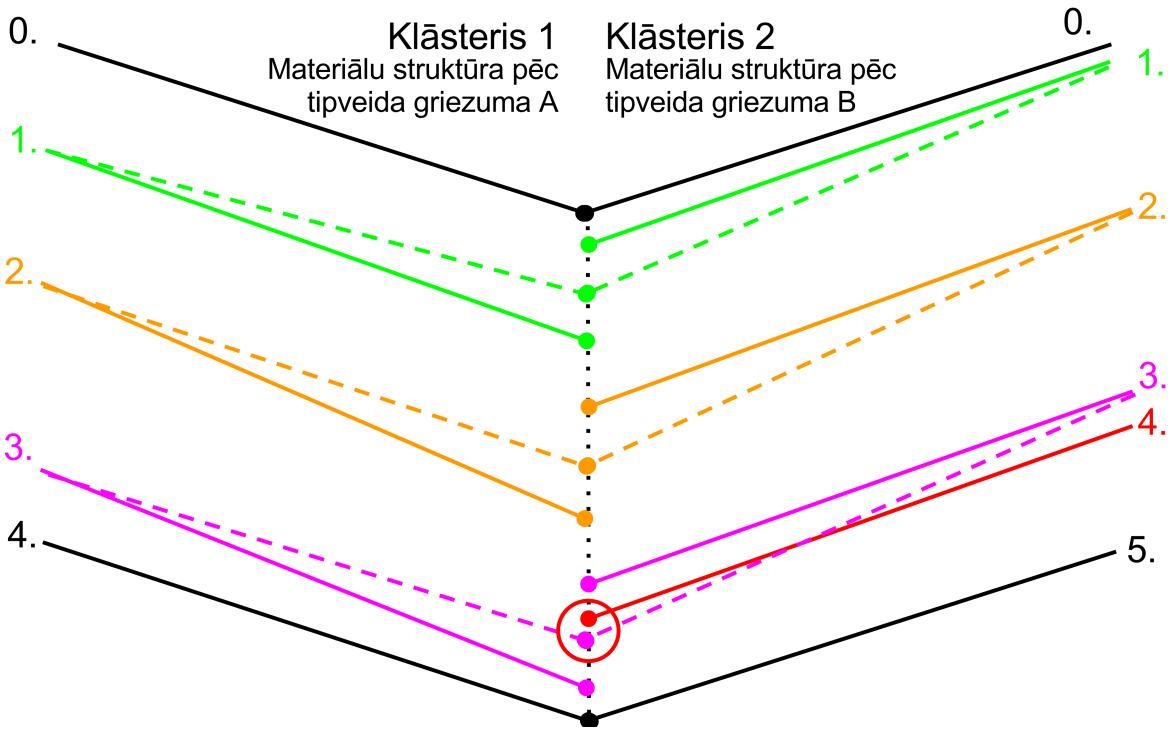
Pēc šī (4.6.2. attēlā ilustrētā) algoritma tika aprēķinātas augstumatzīmes visiem modeļa režģa ģeometrijas punktiem. Šiem punktiem ir vienāda numerācija visos ģeometrijas slāņos, un tie pieder pie vienādi numurētiem režģa trijstūru elementiem, kuriem iepriekš ir piesaistīta informācija par atbilstošo tipveida griezumu. Lai gan aprēķina gaita punktu augstumiem, lielākoties, ir atspoguļota ilustrācijā (4.6.2.), tomēr daži aspekti jāpaskaidro sīkāk.

Ar punkta Z jeb augstumatzīmes aprēķinu šajā ilustācijā tiek saprasts pret kvartāra segas biezumu normalizētā slāņa dziļuma pārrēķins uz absolūto augstumatzīmi metros virs jūras līmeņa, balstoties uz tipveida griezuma struktūru. Katram punktam katrā slānī tika noteikta piederība trijstūra elementiem, kuriem katram tika noteikts tipveida urbuma griezums.

Normalizēto tipveida griezumu iekļaušanu ģeometrijā sarezģī ne tikai konfliktu risināšanas nepieciešamība uz telpisko klāsteru robežām, bet arī tas, ka tiem atbilstošo materiāla struktūru reprezentējošo tipveida urbumu slāņu skaits var atšķirties. Tā, piemēram, pastāv iespēja, ka režģī ir punkti, kas kopīgi trim (un bieži vien vairākiem) dažādos klāsteros ietilpsošiem trijstūriem. Piemēram, lielākam skaitam trijstūru piederošs punkts ir atrodams attēlā 4.6.1. pie X ass koordinātu atzīmes 405 000 uz dienvidiem no Y ass atzīmes 6 360 000. Katram no tiem jāaprēķina augstumatzīme pēc trim atšķirīgiem tipveida griezumiem. Kā atspoguļots 4.6.2. organogrammā, punkta augstumatzīmju kandidātvērtības tiek aprēķinātas pēc katra no normalizētajiem tipveida griezumiem, un konflikti ir atrisināmi divos veidos - augstumatzīmju vērtībām aprēķinot vidējot artismētisko vērtību (tālāk dēvēta par vidējošanu) vai visos gadījumos vienkārši pieņemot seklāko kandidātvērtību. **Šajā darbā** aplūkotajām ģeometrijām **izmantota vidējošanas** stratēģija, taču sagatavotais *Python* valodas scenārijs pieļauj tehniski pārslēgt arī seklākās augstumatzīmes režīmu visu konfliktu risināšanai. Seklākās augstumatzīmes režīms ir vienkāršāks un darba gaitā tika izstrādāts pirmais. Tā izmantošana nozīmē konflikta atrisināšanu par labu detālākai ģeometrijai kopumā. Ja vairākiem slāņiem pēc kārtas starpklāsteru konflikti atrisināti, izmantojot seklāko slāņu augstumatzīmes, tas nozīmē, ka uz klāsteru robežām tika dota priekšroka tipveida griezumiem, kuros ir vairāk slāņu.

Savukārt, atrisinot augstumatzīmju kandididātvērtību konfliktus ar vidējošanu, pastāv iespēja nonākt topoloģiskā pretrunā ar slāņu secību. Lai to ilustrētu, iztēlosimies divdimensiju šķērsgriezumu modeļsistēmas ģeometrijas struktūrā, kur ir redzama pāreja starp diviem klāsteriem (attēls 4.6.3.). To teritorijā tiek pieņemta materiālu struktūra no diviem tipveida griezumiem, bet slāņu skaits tajos ir atšķirīgs. Pieņemsim, ka pēdējais sānis ir subkvartāra virsma, kuru neskar kvartāra nogulumu reprezentācijas slāņu augstumatzīmju aprēķins, tāpat kā augstumatzīmes virsmas modelī (0. slānī). Lai būtu vieglāk atšķirt slāņus pēc secības, kas nosaka to, kurām augstumatzīmēm būtu jāveic vidējošana, 1., 2. un 3. slānis abās klāsteru

robežas pusēs iekrāsots vienādās krāsās. Tādās pašās krāsās iekrāsotas un ar pārtrauktām līnijām savienotas ir pēc vidējošanas aprēķinātās augstumatzīmes abpus klāsteru robežas. Kā redzams, tad gadījumos, ja tipveida griezuma struktūrā ir atšķirīgs skaits slāņu, var iegūt rezultātus, kur 3. slāņa vērtība tiek aprēķināta **zem** 4. slāņa vērtības. Lai izvairītos no šādām situācijām, topoloģisko konfliktu rašanās tika speciāli kontrolēta, ģeometrijas izveides gaitā izmantojot vidējošanas konfliktu risināšanas pieeju. Ja šāda situācija tiek konstatēta, kādam no secīgi dziļākiem slāņiem iegūstot vidējoto vērtību virs secīgi seklākiem slāņiem, tad algoritms pārslēdzas uz seklākās augstumatzīmes pārņemšanu.



4.6.3. attēls. **Ilustratīvs piemērs augstumatzīmju kandidātvērtību konfliktu risināšanai ar vidējošanu un piemēru topoloģiskai problēmai, atšķirīgu slāņu skaitam** (izveidojis autors, izmantojot *Google Drawing*).

Pēc tam, kad visiem režģa punktiem visos kvartāra nogulumu slāņos ir aprēķinātas augstumatzīmes, lai pabeigtu normalizēto tipveida griezumu materiālu struktūras īpašību pārnesi uz modeļa ģeometrijas kvartāra reprezentāciju, kvartāra nogulumu slāņiem tika piešķirtas materiālu īpašības. Kā jau minēts metožu pirmajā nodaļā, tad tabulā 4.1.1. atspoguļotais iežu klasifikators satur filtrācijas koeficientu vērtības ģeneralizētajām iežu klasēm, pēc kurām tika piešķirtas materiālu īpašības modeļsistēmas slāņos oriģinālajā MOSYS kvartāra nogulumu ģeneralizācijā ar 4 slāņiem. Lai iekļautu klāsteriem atbilstošās materiālu struktūras pēc tipveida urbumu griezumiem jaunajā ģeometrijā, materiālu īpašības vairs nevar piešķirt visam slānim vienādas, bet tās tika piešķirtas katram režģa elementam katrā ģeometrijas struktūras kvartāra nogulumu slānī. Tas tika darīts divos soļos.

Pirmajā solī no ģeometrijas, kurai iepriekš aprakstītajā veidā tika aprēķinātas režģa punktu augstumazīmes Latvijas teritorijā, tika ielasītas filtrācijas koeficientu vērtības visos kvartāra reprezentāciju veidojošos slāņos, iegūstot katra elementa (trijstūra) numuram katrā slānī atbilstošo horizontālo un vertikālo filtrācijas koeficientu vērtību. Tad cikliski klāsterēšanas apgabalā ietilpstošajiem trijstūriem katrā no kvartāra nogulumu slāņiem tika piešķirti jaunie filtrācijas koeficienti pēc atbilstošā tipveida griezuma slāņa secības. PyMOSYS rīki piedāvā no ģeometrijas struktūras ielasīt trijstūru elementu atribūtus (t.sk. vertikalās un horizontālās filtrācijas koeficientu vērtības) kā vārdnīcu (*Python dictionary* objekts). Tas ļauj mainīt vērtības tikai tiem trijstūra elementiem, kuri atrodas klāsterēšanas teritorijā, tādējādi neveicot liekas cikliskas darbības pa daudz lielāku skaitu režģa elementu, kuri atrodas ārpus Latvijas teritorijas. Tiem tiek saglabāta līdzšinējā MOSYS V1 kvartāra reprezentācija un tās filtrācijas koeficienti.

Visu šajā nodaļā iepriekš aprakstīto darbību kopumu realizē šī darba ietvaros sagatavotais autora *Python* valodas scenārijs Clusters_25AVGen.py. Šī programma, izmantojot telpisko klāsteru *Shapefile* slāni ar klāsteru identifikatoriem poligonu atribūtu laukos un DB savienojumu tipveida urbumu griezumu ielādei, izveido MOSYS ģeometrijas struktūras, kas ir tālāk izmantojamas modeļsistēmas aprēķinu veikšanai. Šajā nodaļā aprakstītās darbības tiek atkārtotas katram no 162 klāsteranalīzes rezultātiem, šīs darba daļas nobeigumā iegūstot 162 MOSYS filtrācijas modeļa ģeometrijas struktūras ar dažādām kvartāra nogulumu ģeneralizācijas versijām Latvijas teritorijā. Tādējādi tika radīta iespēja veikt modeļsistēmas rezultātu aprēķinus, lai veiktu šī maģistra darba hipotēzes pārbaudi.

4.7. Modeļa rezultātu aprēķini, izmantojot eksperimentālās ģeometrijas

Lai novērtētu, kā modeļsistēmas aprēķinus ietekmē šajā darba gaitā izveidotās kvartāra nogulumu reprezentācijas, un demonstrētu šo ietekmi sistemātiskā veidā, tika veikti modeļsistēmas aprēķini 162 pēc iepriekšējās nodaļas apraksta izveidotajām modeļsistēmas materiālu struktūras ģeometrijām. Katra no šīm ģeometrijām ir materiālu īpašību 3D struktūras ģeneralizācija kvartāra nogulumu slāņos.

Aprēķini tika uzstādīti ar tiem pašiem robežnosacījumiem, kā veicot aprēķinus ar MOSYS V1 oriģinālo ģeometrijas struktūru (Seņņikovs, 2011). Ar aprēķinu veikšanu šajā gadījumā tiek saprasts pirmais solis autokalibrācijas gaitā, veicot aprēķinu stacionārā režīmā mūsdienu scenārijam (ibid.), kura rezultātā tiek iegūta kvadrātisko noviržu summa pa visiem modeļsistēmas ģeometrijas slāņiem. Šādi aprēķini tiek veikti visām 162 šī darba ietvaros

sagatavotajām eksperimentālajām modeļsistēmas ģeometrijas struktūrām un attiecīgi tiek iegūts tāds pats skaits rezultātu. Tātad, 81 rezultāts, katrai no divu veidu (**NKD un NKD+E)** matricām, uz kurām balstīti tālākie klāsteranalīzes rezultāti, pēc kuriem, savukārt, sagatavotas eksperimentālās ģeometrijas.

4.8. Autokalibrācija kvartāra nogulumu slāņos

Kā jau aprakstīts nodaļā 4.6., tad visas eksperimentu ģeometrijas tika izveidotas, balstoties uz optimizēto MOSYS V1 ģeometrijas struktūru, kas nozīmē, ka pirms ģeneralizētās kvartāra nogulumu reprezentācijas iekļaušanas modeļsistēmas ģeometrijā materiālu īpašības tika kalibrētas, lai pēc iespējas samazinātu summāro kvadrātisko novirzi visos modeļsistēmas slāņos attiecībā pret pjezometriskajiem līmeņiem monitoringa urbumos un attiecībā pret ūdensguves urbumu ierīkošanas līmeņu mērījumiem. Kā jau iepriekš minēts, kalibrācija ir hidroģeoloģisko modeļu izveides neatņemama sastāvdaļa (Gorokhovski, 2012; Bethers et al., 2011). Eksperimentālo struktūru izveides gaitā Latvijas teritorijā, kur tika veikta ģeneralizācija kvartāra nogulumu reprezentācijai, modeļa ģeometrijā tika iekļautas materiālu īpašības, izmantojot MOSYS iežu klasifikatoru (tabula 4.1.1.). Tas nozīmē, ka jaunā kvartāra nogulumu reprezentācija attiecībā pret pārējo daļu modeļa ģeometrijas būtu uzskatāma par nekalibrētu. Lai novērtētu ģeneralizēto kvartāra reprezentāciju ietekmi uz modeļsistēmas aprēķinu rezultātiem, tika veikta kalibrācija arī jaunizveidotajām ģeometrijām. Ideālā gadījumā šo jaunizveidoto ģeometrijas daļu būtu vēlams kalibrēt nošķirti no pārējas modeļa ģeometrijas, taču pašreizējā MOSYS izstrādes stadijā nav tehniskās infrastruktūras, lai to darītu, un šāda mehānisma izstrāde ir ārpus darba ietvariem. Tādēļ vienīgais risinājums ir veikt atkārtotu filtrācijas koeficientu lauku kalibrāciju jaunizveidotajām ģeometrijas struktūrām visā teritorijā.

Pilna Baltijas artēziskā baseina modeļa MOSYS ģeometrijas struktūras kalibrācija ir laikietilpīgs process, kas prasa ļoti ievērojamus mašīnlaika resursus un operatīvās atmiņas apjomu. Tādēļ nepieciešams atlasīt mazāku skaitu eksperimentālo ģeometrijas struktūru, kurām veikt kalibrāciju ar materiālu īpašību optimizējošo algoritmu, kā aprakstīts MOSYS kalibrācijai veltītajos literatūras avotos (Virbulis et al., 2012; Stankevica, 2011; Bethers et al., 2011). Lai atlasītu šīs ģeometrijas struktūras, par kritēriju tika izmantota summārā kvadrātiskā novirze visos modeļsistēmas slāņos, kas aprēķināta visām 162 eksperimentālajām ģeometrijas struktūrām, kā aprakstīts iepriekšējā nodaļā un kā atainots rezultātu nodaļas attēlos 5.3.1. un 5.3.2. No abās distanču matricās (NKD un NKD+E) balstītajām eksperimentālajām modeļsistēmas ģeometrijas struktūrām tika atlasītas 5 struktūras katrai: 3 labākās, 1 viduvēja un 1 vissliktākā. Tātad, kopā 10 ģeometrijas struktūras. Galvenais mērķis šādai atlasei ir

pārliecināties, ka modeļsistēmas ģeometrijas struktūras ar labākajiem rezultātiem saglabā savu pozitīvo rezultātu pārsvaru pār pārējām divām ģeometrijas struktūrām ar sliktākiem rezultātiem, jo nevar izslēgt iespēju, ka kāda no ģeometrijas struktūrām ar viduvējiem rezultātiem pēc kalibrācijas var dot daudz labākus rezultātus. Diemžēl, pilnībā šo iespēju nevar izslēgt, neveicot kalibrāciju visām eksperimentālajām modeļsistēmas ģeometrijas struktūrām, taču šādi mašīnlaika resursi darba autoram un PUMa darba grupai nebija pieejami. Savukārt, ja pēc sākotnējā aprēķina ģeometrijas struktūras ar labākajiem rezultātiem arī pēc kalibrācijas modeļsistēmas kvartāra slāņos uzrādītu labākus rezultātus nekā pārējas, tas norādītu, ka saglabājas sakarība starp sākotnējo aprēķinu un rezultātiem, kas iegūstami pēc kvartāra nogulumu slāņu kalibrācijas.

Šīm atlasītajām ģeometrijas struktūrām tika veikta kalibrācija kvartāra nogulumu slāņos, izmantojot MOSYS autokalibrācijas funkcionalitāti (ibid.). Šo ģeometrijas struktūru autokalibrācijas gaitas ir atspoguļotas rezultātu nodaļā, attēlā 5.4.1. Aprēķinu rezultātu jutīgums pret kalibrāciju modeļsistēmas kvartāra nogulumu slāņiem daļēji ļauj spriest arī par konkrētās ģeometrijas kvartāra nogulumu reprezentācijas atbilstību, jo maz ticams, ka kvartāra nogulumu ģeometrijas struktūras, kas vāji reprezentē situāciju dabā, būs iespējams kalibrācijas gaitā uzlabot vairāk nekā tādas, kas ir tuvākas situācijai dabā.

Jāņem vērā, ka ar autokalibrāciju iegūtie rezultāti nav deterministiski: vairākkārtīgi uzsākot kalibrāciju ar atšķirīgu parametrizāciju, piemēram, mainot slāņus, kuri tiek kalibrēti, ir iespējams iegūt atšķirīgas rezultātus. Respektīvi, nav garantijas, ka kāds konkrētais rezultāts ir labākais iespējamais (Bethers et al., 2011). Iegūtā minimālā kvadrātisko noviržu summa var būt lokālais minimums. Tomēr ir iespējams novērot, kā eksperimentālās ģeometrijas struktūras reaģē uz kalibrāciju. Praksē ir iespējams atkārtoti uzsākt autokalibrāciju ar atšķirīgu parametrizāciju, norādot dažādas slāņu grupas, kurām tiek veikta kalibrācija, tādējādi mazinot iespējamību, ka autokalibrācijas gaitā iegūtais labākais rezultāts ir lokālais minimums. Tas arī tiek darīts struktūrām, kuras uzrāda labākos rezultātus kvartāra slāņu autokalibrācijas gaitā (5.4.1. attēls).

Rezultātā tika atlasītas eksperimentālās ģeometrijas struktūras ar vislabākajiem rezultātiem un tām tika turpināta autokalibrācija, kalibrējot pamatiežu slāņus, lai tālāk uzlabotu ar tām iegūtos rezultātus. Kalibrācijas gaita tika salīdzināta ar MOSYS V1 bāzes ģeometriju, kuras mazāka summārā kvadrātiskā novirze ir 814 metri.

4.9. Citas tehniskās piezīmes

Šajā nodaļā ir apkopotas dažas sīkākas darba tehniskās gaitas piezīmes ārpus pārējām metožu nodaļām.

Lai veiktu NKD matricas aprēķinu, tika izmantots Škarina *PPMd* (Shkarin, 2002) datu kompresijas algoritms, kas realizēts šāda paša nosaukuma programmā. Starp pieejamajām kompresijas metodēm šī ir viena no labākajām realizācijām labi pazīstamajam PPM (*prediction by partial matching* – prognozēšana ar daļēju atpazīšanu) algoritmam, kuru 1984. gadā pirmoreiz apraksta Jaunzēlandiešu zinātnieki Klīrijs un Vittens. Šī ir viena no nozīmīgākajām publikācijām datu kompresijas nozarē (Cleary and Witten, 1984), jo šim darbam bija ļoti liela praktiskā un teorētiskā nozīme. Tā pavēra ceļu modernajiem statistiskajiem datu kompresijas algoritmiem, no kuriem šodien ir izaugušas tādas eksprimentālās programmas kā PAQ saimes varianti, kas ir vieni no labākajiem šī brīža eksperimentālajiem datu kompresoriem (Mahoney, 2011). Ja PAQ saimes programmas ir vairāk zināmas kompresijas entuziastu aprindās, tad RAR kompresijas formāts un to veidojošās programmas ir plašāk pazīstamas. NKD aprēķinus ar PAQ un RAR metodēm *ncd* utilīta nekādā formā pagaidām nepiedāvā.

Darba gaitā tika sastapta tehniska, ar skaitļošanas platformu saistīta, limitācija attiecībā uz *PPMd* kompresijas algoritma izmantošanu NKD metrikas aprēķiniem. *Complearn-mod-ppmd* modulis, kas nodrošina *PPMd* algoritma izmantošanu *CompLearn* rīku ietvaros, nav pieejams uz Ubuntu amd64 platformas. Tas ierobežo urbumu skaitu, kurus var izmantot NKD matricas aprēķinam ar *PPMd* kompresoru. Bez šīs limitācijas, izmantojot, piemēram, *gzip* datu kompresijas programmu, būtu iespējams palielināt izmantoto urbumu skaitu līdz vairāk nekā 6000 urbumiem, šajā darbā 3120 izmantoto vietā. Tomēr priekšroka tika dota *PPMd* kompresijas metodei, tās teorētiskā pamatojuma dēļ, kas balstās uz statistisko konteksta modelēšanu. Tā seko principiem, kas ir aprakstīta Cleary and Witten, 1984, publikācijā.

Urbumu vertikālo griezumu atlase, pirmapstrāde un normalizācija tika veikta *MySQL* datu bāzu vadības sistēmas vidē, kas tiek izsaukta no *Python* scenārijvalodas programmas, apvienojot *SQL* un *Python* valodu priekšrocības darba automatizēšanai un ievades datu sagatavošanai tālākai apstrādei ar CompLearn rīkiem.

Darbā tika izmantota CompLearn rīku komplekta versija 1.1.7. un utilīta *ncd* NKD aprēķinam ar *PPMd* algoritmu, izmantojot libcomplearn-mod-ppmd-1.0.7 pakotni Ubuntu GNU/Linux 11.10 operāciju sistēmas vidē. NKD aprēķina rezultāti, lai iegūtu distanču

matricu delimitēta teksta veidā satur atšķirīguma (*dissimilarity*) matricu visiem urbumu pāriem.

Telpisko klāsteru izveides darba daļas automatizācija tika veikta R statistikas valodā, jo tā piedāvā ērtu un efektīvu veidu ātram Voronoi teselācijas aprēķinam un eksportam *ESRI Shapefile* formātā, izmantojot funkciju bibliotēkas *deldir, sp, maptools, rgeos, gcplib* un *rgdal*. Dažādu papildus funkciju nodrošināšanai šī darba ietvaros sagatavotais *R* valodas scenārijs izmanto vairākas citas funkciju bibliotēkas, kas nav kritiski svarīgas rezultātu sasniegšanai, bet padara darba automatizēšanu ērtāku (piemēram, *base, tseries, lattice*). Izvēle krita uz *R* statistikas valodu, jo, darba autoram par lielu pārsteigumu, bija grūti piemeklēt atvērtu telpiskās analīzes automatizēšanas vidi ārpus *ESRI ArcGIS* saimes rīkiem, kas spētu nodrošināt vienuviet visas nepieciešamās funkcijas (pamata Voronoi poligonu izveidi, poligonu izgriešanu (*Clip*) un izšķīdināšanu (*Dissolve*), vēlams arī rezultātu automātisku saglabāšanu rastra formātā). Tika izskatīti sekojoši varianti - *ESRI ArcGIS*, *SEXTANTE* (*SAIG Kosmo* ietvarā), *Quantum GIS* un *R. Quantum GIS* telpiskās analīzes darbības ar vektordatiem piedāvā veikt, izmantojot *fTools* pakotni, kurai gatavas saskarnes funkciju izmantošanai scenārijos ārpus *fTools* nav. *SEXTANTE* rīki nodrošina daudzsološu darba gaitas modeļu izveidi līdzīgi *ArcGIS ModelBuilder*, bet *SEXTANTE* versijai 0.5 pietrūka Voronoi poligonu izveides funckijas. Savukārt, *ESRI ArcGIS* rīku izmantošana būtu bijusi iespējama, bet nevēlama nepieņemami lēnā ātruma dēļ, kādā *ArcGIS* ģeoprocesors veic Voronoi poligonu aprēķinu. Ar šo problēmu autors bija saskāries jau iepriekš, citam lietojumam, kuram bija nepieciešams Voronoi poligonu aprēķins no vairākiem tūkstošiem ģeneratorpunktu (Jātnieks u.c., 2011*b*), tādēļ izvēle tika izdarīta par labu *R* rīkiem. Šī izvēle bija veiksmīga, jo *R* pakotne *deldir* uz tās pašas datortehnikas spēja veikt 3120 Voronoi poligonu izveidi aptuveni pusminūtes laikā, ļaujot pabeigt telpisko klāsteru izveidi visiem 162 variantiem 96 minūtēs kopā ar izgriešanas un izšķīdināšanas darbībām, kā arī rezultātu izvadi uz diska un rezultātu kontrolei paredzētu pārskata kartoshēmu izveidi PNG formātā.

Visas darbības ar telpiskajiem datiem un modeļsistēmas ģeometriju veiktas, izmantojot BalticTM93 jeb Latvijas ĢIS lietotāju starpā plaši lietoto ETRS89 koordinātu sistēmu, kas agrāk Latvijā bieži tikusi kļūdaini dēvēta par LKS-92. Mūsdienas tā ir universāli identificējama tīmekļa servisos, piemēram, plaši lietotajā OGC WMS (*Web Map Service*), ar EPSG (*European Petroleum Suvey Group*) datu bāzes kodu 25884. Tās lietojumu nosaka modelēšanas projekta specifika un tas, ka modelēšanas teritorija aptver Baltijas valstis un pieguļošās teritorijas no Polijas, Krievijas, Zviedrijas un Baltkrievijas.

Modeļsistēmas aprēķinu vide tika uzstādīta uz 2 datoriem, kopsummā nodrošinot 12 Intel Core 2 klases procesoru kodolus aprēķinu veikšanai. Ņemot vērā modeļsistēmas

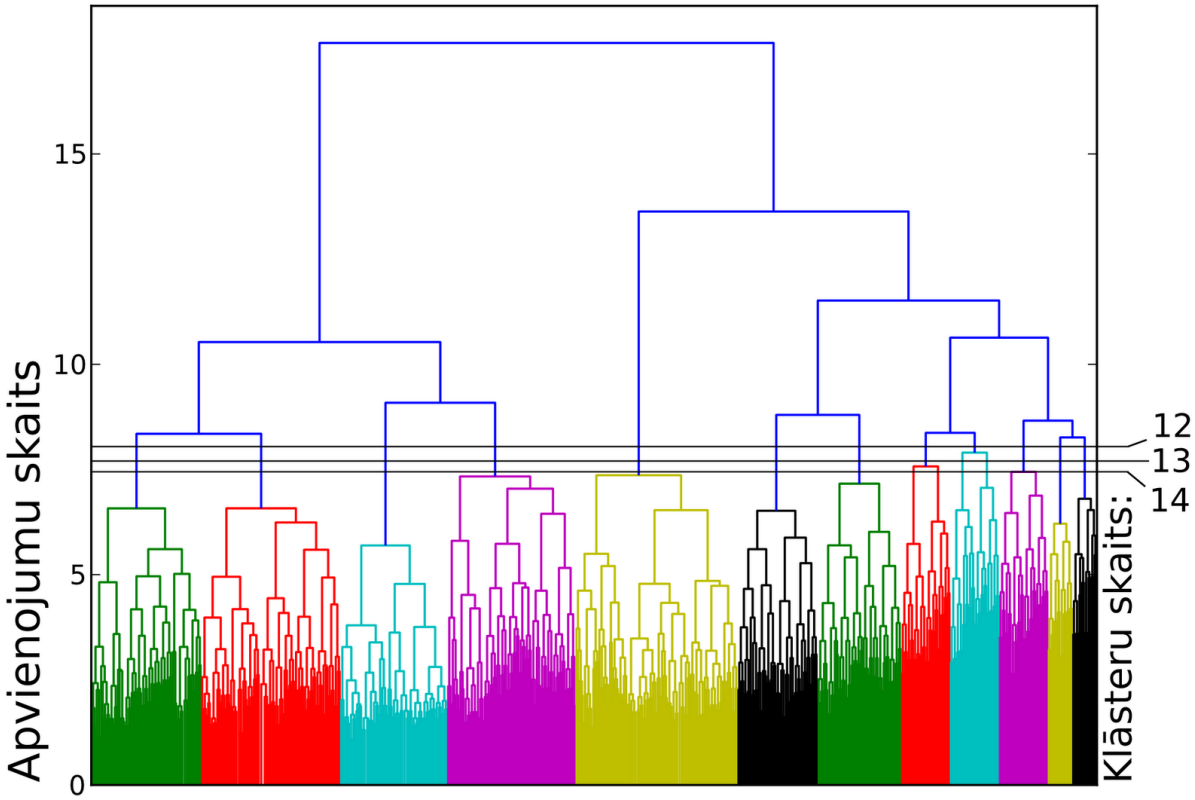
MOSYS V1 ģeometriskās struktūras kompleksitāti un detalizācijas pakāpi, aprēķinus bija iespējams veikt tikai ar datoriem, kuriem ir vismaz 8 GB operatīvās atmiņas (*RAM*) un 64 bitu Windows operāciju sistēmas. Darba aprēķinu daļa tika realizēta uz sekojošas konfigurācijas datoriem: 2 x X7350 @ 2.93Ghz, 32GB RAM un Q9450 @ 3.73 Ghz, 8GB RAM. Abām mašīnām aprēķinu laiks ir līdzīgs, vienai iterācijai parasti prasot 5-6 minūtes mašīnlaika, kas nozīmē 9-10 iterācijas stundā. Kopumā šī darba rezultātu aprēķini ar iepriekšminētajiem skaitļošanas resursiem notika apmēram mēnesi, kopsummā veicot modeļsistēmas rezultātu aprēķinus vairāk kā 9000 reižu. Lielākā daļa šo aprēķinu sastāda autokalibrācijas iterācijas. PyMOSYS darba vidi, kurā tiek veikta modeļsistēmas struktūras ģenerācija un aprēķini, veido lielāks skaits komponentu, no kuriem lielāko daļu nodrošina *Python(x,y)* zinātniskās skaitļošanas distribūcijas versija 2.7.2.1. Šī programmatūras pakete satur vairumu populāru zinātniskās programmēšanas rīku *Python* programmēšanas valodai, t.sk. *Numpy* un *Scipy* bibliotēkas, ģeodatu apmaiņai neaizstājamos GDAL/OGR rīkus (GDAL, 2012). Papildus *Python(x,y)* iekļautajām funkciju bibliotēkām, šajā darbā tika izmantota *shapely* bibliotēka*,* lai noteiktu modeļa ģeometrijas režģa elementu piederību telpiskajiem klāsteriem, un *BioPython,* lai izsauktu klāsteranalīzes algoritmu kolekciju *C clustering library 3.0* (Hoon et al., 2010).

Aprēķinu veikšanai tika uzstādīta *PARDISO Linear System Solver* programmatūra ar 3 mēnešu izmēģinājuma licencēm uz aprēķiniem izmantotajiem datoriem. Tā tiek izmantota MOSYS modelēšanas sistēmas aprēķinu veikšanai gan lokālajiem aprēķiniem, gan uz skaitļošanas klāsteriem, izmantojot OpenMPI paralēlās skaitļošanas vadībai (Gabriel et al., 2004).

5. REZULTĀTI

Šajā nodaļā apkopoti gan visu darba gaitu starprezultāti, gan arī kopējie kvartāra uzbūves klāsteranalīzes rezultāti, kas iegūti, izmantojot normalizētās kompresijas distances metriku kā galveno kritēriju urbumu griezumu litoloģiskās struktūras līdzībai.

5.1. Klāsteranalīzes rezultāti



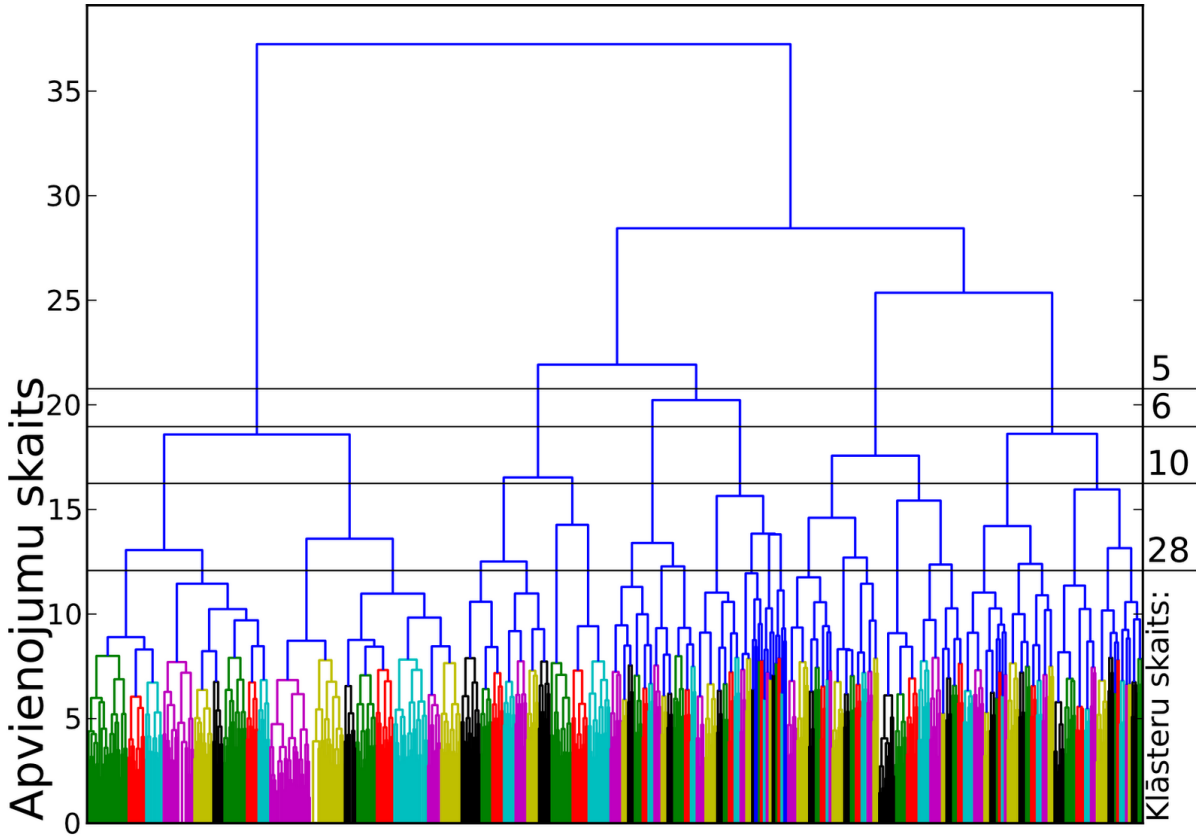
5.1.1. attēls. **Hierariskās aglomeratīvās maksimālo distanču klāsteranalīzes dendrogramma, iegūta izmantojot NKD matricu** (izveidojis autors ar *scipy.cluster.hierarchy.linkage* pakotni un *Inkscape* vektorgrafikas redaktoru).

Attēls 5.1.1. atspoguļo vienu no galvenajiem šī darba rezultātiem - hierarhiskās maksimālo distanču klāsterēšanas metodes dendrogrammu NKD matricai. Šai dendrogrammai ir nozīmīga loma tālakājā darba rezultātu izklāstā, jo tā ļauj pārskatāmi spriest par datu iekšējo struktūru, un tieši NKD matricas klāsterēšanas rezultāti ir tie, pēc kuriem izdevās izveidot modeļsisēmas aprēķinu novirzi samazinošo kvartāra nogulumu reprezentāciju, kas ir aprakstīta nākamajās nodaļās. Kā varēja sasgaidīt, izmantojot hierarhisko metožu dendrogrammas (Han et al., 2012), ir iespējams daļēji spriest par datiem atbilstošo ģeneralizācijas līmeni. Labajā pusē ar līnijām iezīmētas vietas, kurās, dendrogrammu sadalot (ar *cuttree* funkciju) 12 un 13 loģiskajos klāsteros un šo rezultātu izmantojot kvartāra

nogulumu ģeneralizētu modeļsistēmas ģeometrijas struktūru izveidē, ir iespējams iegūt labākos aprēķinu rezultātus. Kā redzams, izdalot vairāk nekā 14 klāsterus, datu iekšējā struktūra sāk kļūt daudz sadrumstalotāka. Klāsteru skaitam, kas lielāks par 12, dendrogrammas (attēls 5.1.1.) zari iekrāsoti atšķirīgās krāsās. No šīs dendrogrammas vienas pašas nav iespējams iepriekš noteikt, kāds varētu būt reprezentatīvais ģeneralizācijas līmenis konkrētajiem datiem, tomēr ir iespējams kvalitatīvi spriest, ka sadrumstalotība augstākajiem abstrakcijas līmeņiem sāk strauji pieaugt virs 12-14 klāsteriem, un tas norāda uz to, ka dotajiem datiem globāli reprezentatīva ģeneralizācija varētu būt atrodama šajā klāsteru skaita reģionā. To tālāk parāda arī aprēķinu rezultāti, kas ir izklāstīti nodaļā 5.3.

Šī ir ļoti noderīga informācija, jo gadījumā, ja dendrogrammu dalītu mazākā skaitā loģisko klāsteru skaitu, lai samazinātu rezultējošo eksperimentālo ģeometriju skaitu un, līdz ar to, arī veicamo aprēķinu skaitu, tad šajā gadījumā dendrogramma būtu ļoti noderīga loģisko klāsteru skaita parametrizācijai.

Pēc dendrogrammas NKD+E eksperimentu matricai, kas atspoguļota attēlā 5.1.2., varētu spriest, ka optimālais klāsteru skaits varētu svārstīties ap 28 klāsteriem, taču to neapstiprina rezultāti attēlā 5.3.1. Tas, iespējams, ir saistīts ar NKD+E eksperimentu grupas kopumā sliktajiem rezultātiem, kas sīkāk aplūkoti pie rezultātu analīzes.

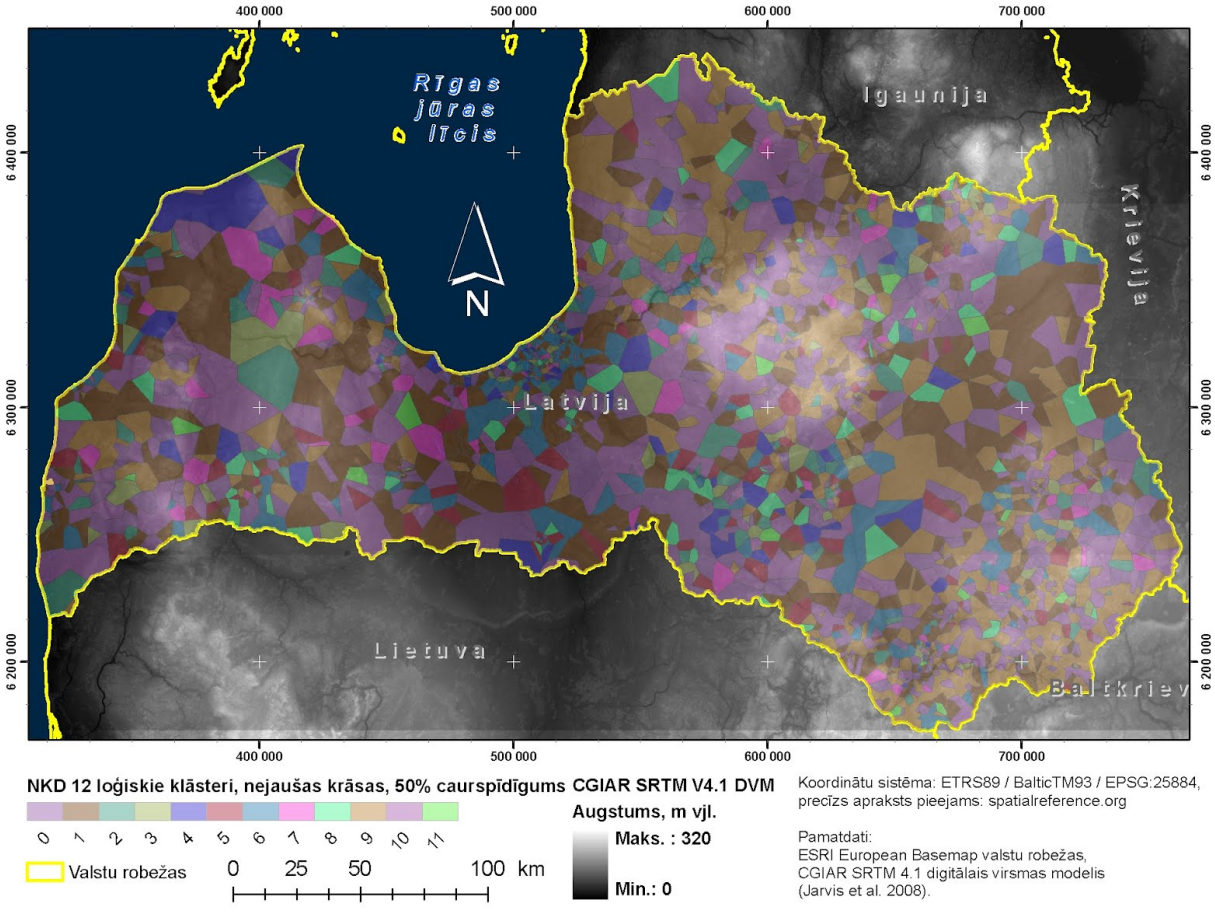


5.1.2. attēls. **Hierariskās aglomeratīvās maksimālo distanču klāsteranalīzes dendrogramma, iegūta izmantojot NKD+E matricu** (izveidojis autors ar *scipy.cluster.hierarchy.linkage* pakotni un *Inkscape* vektorgrafikas redaktoru).

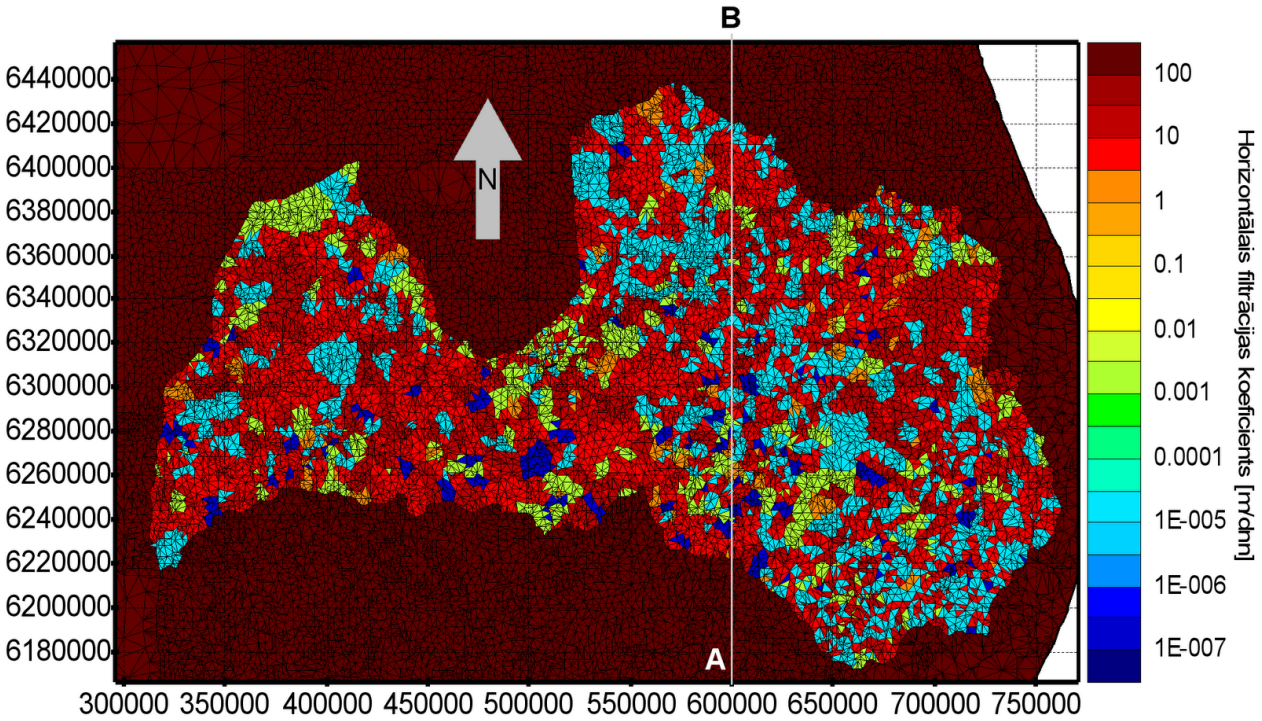
Kā jau iepriekš aprakstīts metožu sadaļā, telpisko klāsteru izveide ir nodalīta no loģisko klāsteru ieguves, kas tika veikta, izmantojot hierarhisko maksimālo distanču metodi ar abām eksperimentu matricām - **NKD** un **NKD+E** atbilstošajām dendrogrammām**.**

5.2. Telpisko klāsteru kartes un tām atbilstošās ģeometrijas struktūras

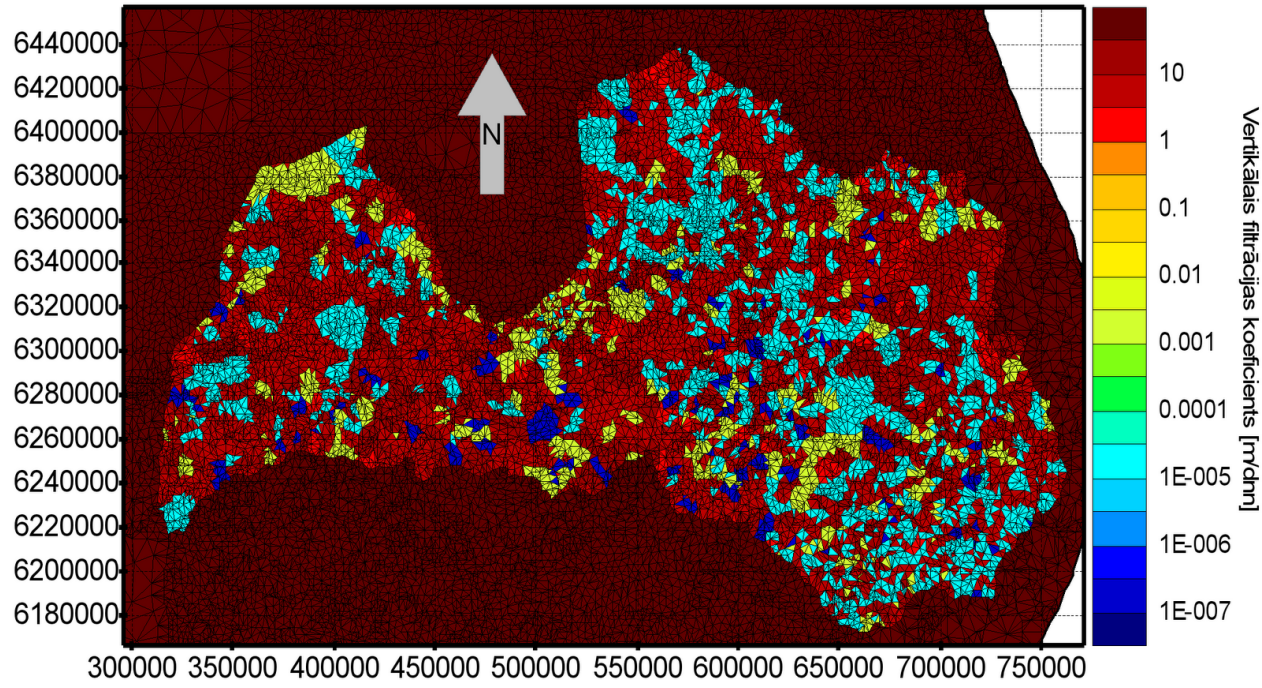
Kopsummā, izmantojot divas eksperimentu matricas (NKD un NKD+E), tika iegūti 162 telpisko klāsteru rezultāti, uz kuriem tālāk tika balstīta eksperimentālo materiālu īpašību ģeometriju izveide. Šajā nodaļā aplūkoti starprezultāti struktūrām, kas darba beigās uzrādīja labākos aprēķinu rezultātus.



5.2.1. attēls. **Telpiskie klāsteri, eksperiments NKD 12k** (izveidojis autors, izmantojot *R* statistikas valodu Voroni teselācijas ieguvei, *Python* eksperimentu automatizācijas scenāriju klāsteranalīzes rezultātu iegūšanai un *ESRI ArcGIS* vidi noformēšanai, kā arī attēlā minētos pamatdatus).



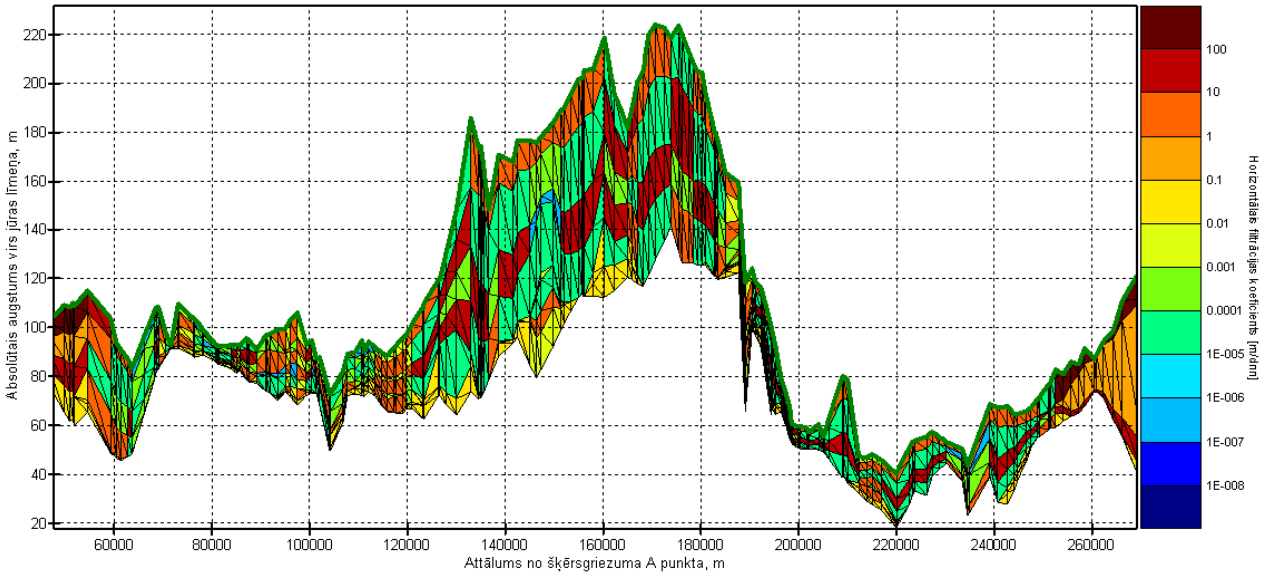
5.2.2. attēls. **Horizontālie filtrācijas koeficienti filtrācijas modeļa ģeometrijas režģa pirmajā slānī, eksperiments NKD 12k.** Profila līnija AB izmantota 5.2.4 un 5.2.5. šķērsgriezumu ieguvei (izveidojis autors, izmantojot *HiFiGeo* vizualizācijas vidi un NKD 12k modeļa ģeometrijas struktūru).



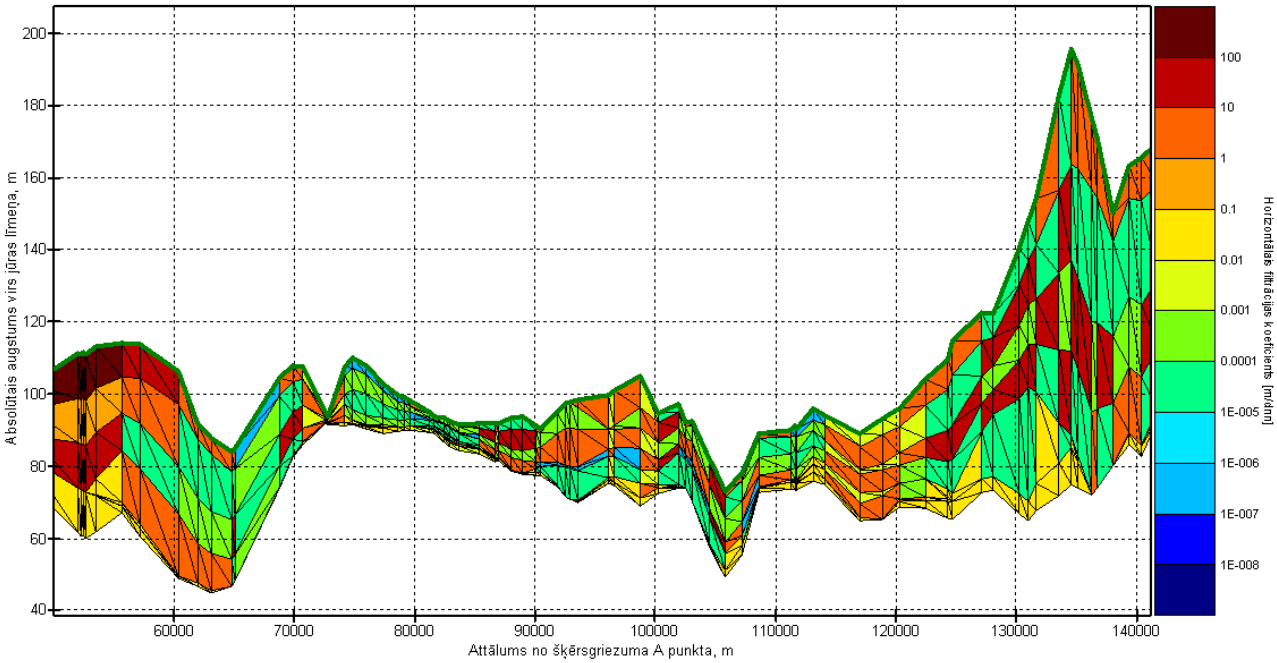
5.2.3. attēls. **Vertikālie filtrācijas koeficienti filtrācijas modeļa ģeometrijas režģa pirmajā slānī, eksperiments NKD 12k** (izveidojis autors, izmantojot *HiFiGeo* vizualizācijas vidi un NKD 12k modeļa ģeometrijas struktūru).

Attēli 5.2.2 un 5.2.3. sniedz ieskatu modeļsistēmas ģeometrijas režģī, kas balstīts uz 5.2.1. attēla telpisko klāsteru sadalījuma vizualizāciju, izmantojot HiFiGeo programmatūru. Attēlā 5.2.1. atspoguļotais telpisko klāsteru izvietojums faktiski parāda dažādu tipveida griezumu teritoriju savstarpējās robežas (telpiskos klāsterus). Savukārt, attēlos 5.2.2. un 5.2.3. vizualizēti filtrācijas koeficienti, kas piešķirti režģa elementiem pēc tiem atbilstošo tipveida

griezumu pirmo slāņu iežu klasēm. Atšķirīgiem tipveida griezumiem var būt kopīgas iežu klases pirmajā slānī. Tas nozīmē, ka filtrācijas īpašību ziņā atšķirīgo teritoriju robežas 5.2.2. un 5.2.3. attēlos var vizuāli nesakrist ar telpisko klāsteru robežām attēlā 5.2.1.

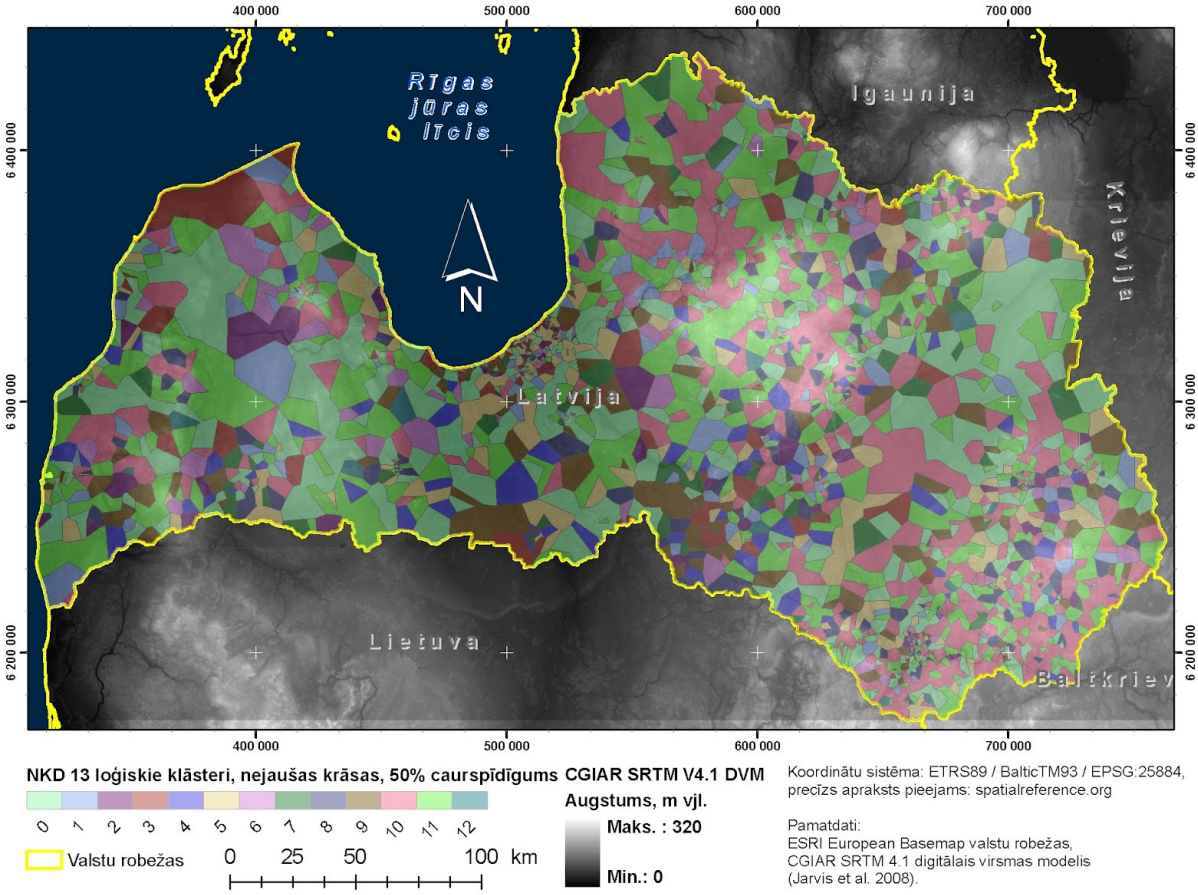


5.2.4. attēls. **Horizontālie filtrācijas koeficienti kvartāra nogulumu slāņiem, šķērsgriezumā AB. Eksperimentālā modeļsistēmas ģeometrijas struktūra NKD 12k** (izveidojis autors, izmantojot *HiFiGeo* vizualizācijas vidi un NKD 12k modeļa ģeometrijas struktūru).



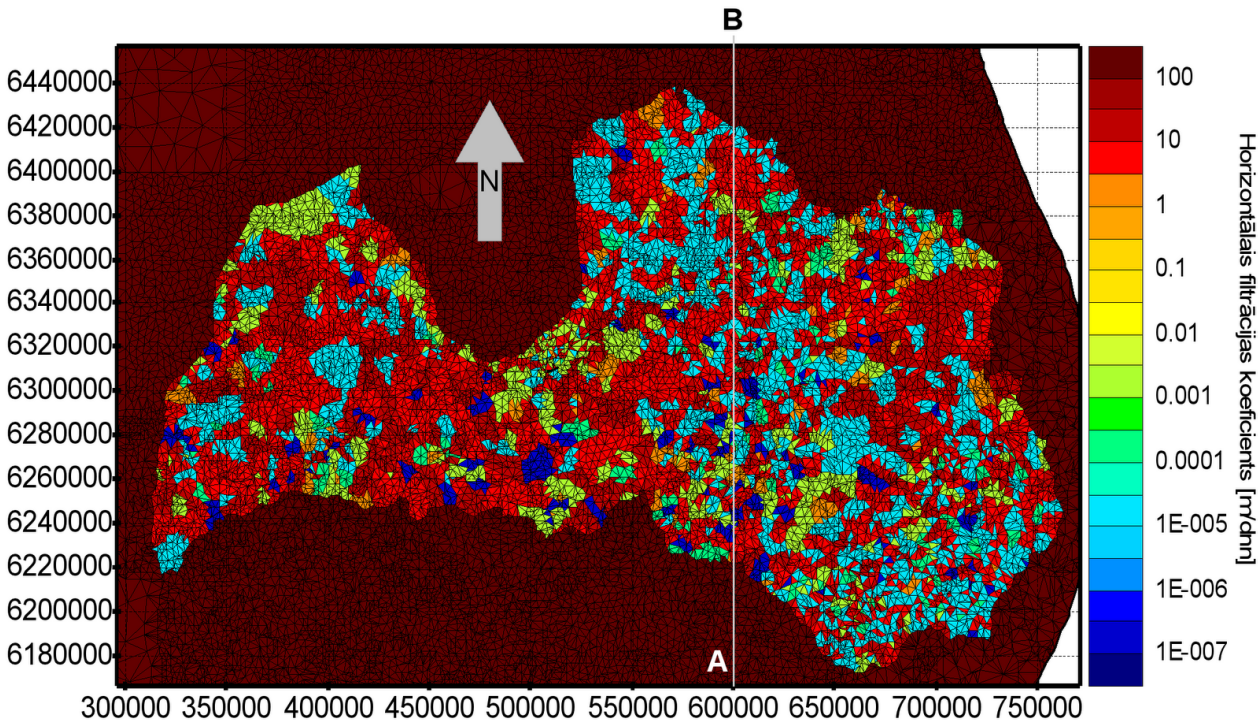
5.2.5. attēls. **Horizontālie filtrācijas koeficienti kvartāra nogulumu slāņiem, šķērsgriezuma AB tuvinājumā. Eksperimentālā modeļsistēmas ģeometrijas struktūra NKD 12k** (izveidojis autors, izmantojot *HiFiGeo* vizualizācijas vidi un NKD 12k modeļa ģeometrijas struktūru).

Attēlos 5.2.4. un 5.2.5. atspoguļotie modeļsistēmas ģeometrijas šķērsgriezumi rāda vertikālos filtrācijas koeficientus pa modeļsistēmas režģa elementiem, sniedzot ieskatu iegūtās ģeometrijas kompleksitātē.

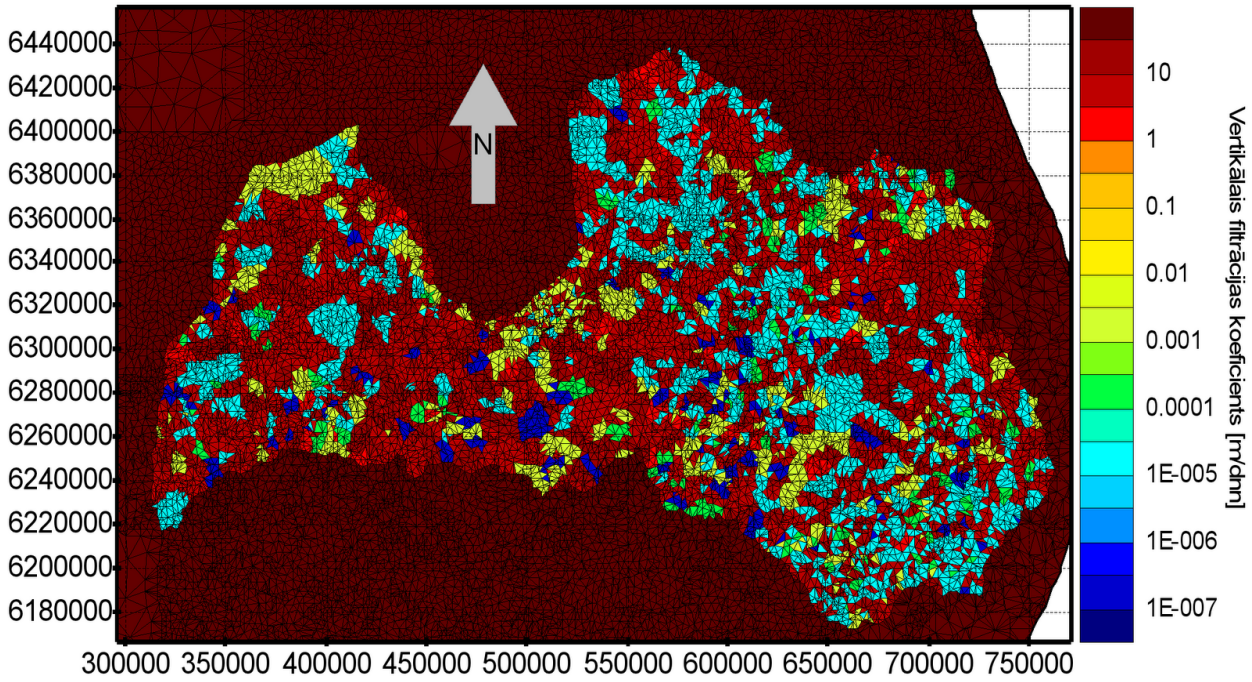


5.2.6. attēls. **Telpiskie klāsteri, eksperiments NKD 13k** (izveidojis autors, izmantojot *R* statistikas valodu Voroni teselācijas ieguvei, *Python* eksperimentu automatizācijas scenāriju klāsteranalīzes rezultātu iegūšanai un *ESRI ArcGIS* vidi noformēšanai, kā arī attēlā minētos pamatdatus).

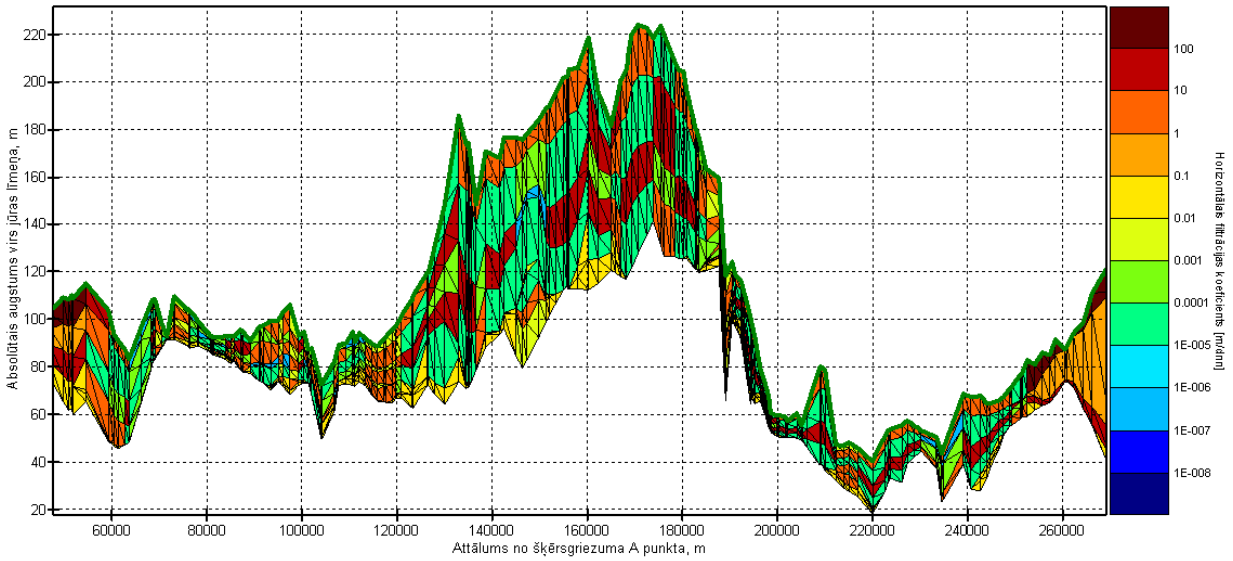
Tieši tāpat kā attēli 5.2.1 - 5.2.5. paskaidro starprezultātu pāreju no telpiskajiem klāsteriem uz modeļsistēmas filtrācijas modeļa ģeometrijas struktūras reprezentāciju eksperimentam NKD 12k, attēli 5.2.6. - 5.2.10. ilustrē to pašu eksperimentam NKD 13k. Šādus attēlus būtu iespējams sagatavot par visām 162 eksperimentālajām ģeometrijām, taču šeit tie ir sagatavoti tām divām ģeometrijām, kurām tika iegūti labāki aprēķinu rezultāti, salīdzinot ar MOSYS V1 optimizēto ģeometriju. Ar pārējo eksperimentu telpisko klāsteru rezultātiem ir iespējams iepazīties šī darba 2. pielikumā gan *Shapefile* formātā, gan PNG formāta kartoshēmu veidā, kuras tika izmantotas rezultātu kontrolei *R* scenārija izstrādes laikā.



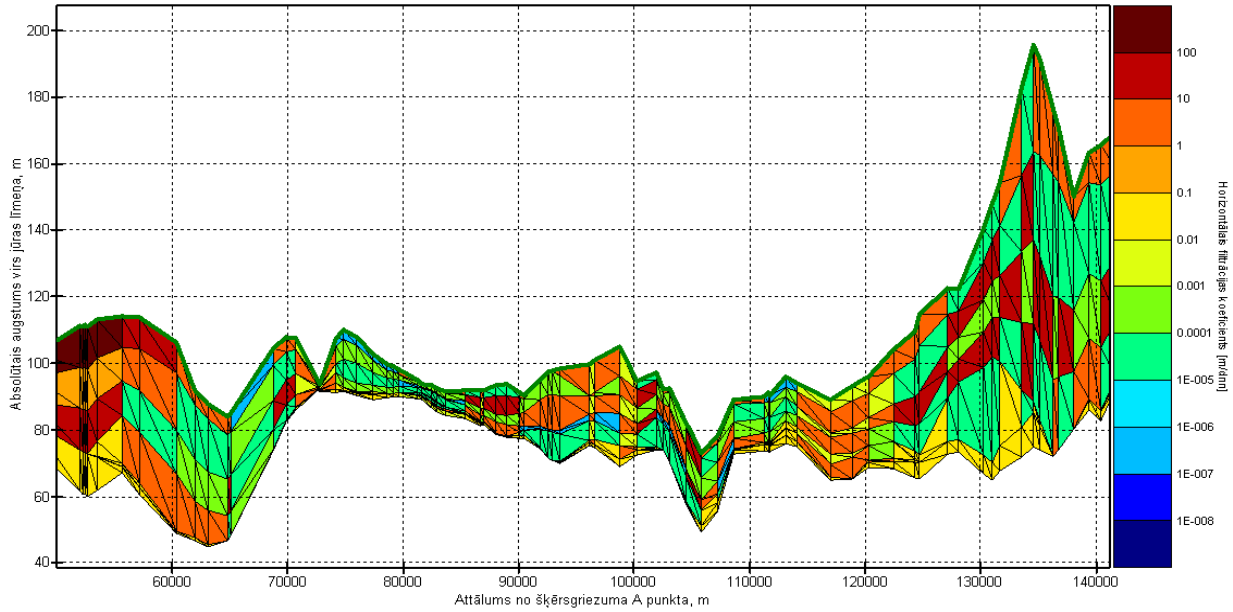
5.2.7. attēls. **Horizontālie filtrācijas koeficienti filtrācijas modeļa ģeometrijas režģa pirmajā slānī, eksperiments NKD 13k.** Profila līnija AB izmantota 5.2.9 un 5.2.10. šķērsgriezumu ieguvei (izveidojis autors, izmantojot *HiFiGeo* vizualizācijas vidi un NKD 13k modeļa ģeometrijas struktūru).



5.2.8. attēls. **Vertikālie filtrācijas koeficienti filtrācijas modeļa ģeometrijas režģa pirmajā slānī, eksperiments NKD 13k** (izveidojis autors, izmantojot *HiFiGeo* vizualizācijas vidi un NKD 13k modeļa ģeometrijas struktūru).



5.2.9. attēls. **Horizontālie filtrācijas koeficienti kvartāra nogulumu slāņiem, šķērsgriezumā AB. Eksperimentālā modeļsistēmas ģeometrijas struktūra NKD 13k** (izveidojis autors, izmantojot *HiFiGeo* vizualizācijas vidi un NKD 13k modeļa ģeometrijas struktūru).



5.2.10. attēls. **Horizontālie filtrācijas koeficienti kvartāra nogulumu slāņiem, šķērsgriezuma AB tuvinājumā. Eksperimentālā modeļsistēmas ģeometrijas struktūra NKD 13k** (izveidojis autors, izmantojot *HiFiGeo* vizualizācijas vidi un NKD 13k modeļa ģeometrijas struktūru).

5.3. Aprēķinu rezultāti ar eksperimentālajām ģeometrijām

5.3.1. attēls. **Loģisko klāsteru skaits pret summāro kvadrātisko novirzi. Ģeometrijas struktūras,**

kas veidotas, balstoties uz NKD matricu ar hierarhisko maksimālo distanču klāsteranalīzi (izveidojis autors, izmantojot *LibreOffice Calc* programmatūru).

5.3.2. attēls. **Loģisko klāsteru skaits pret summāro kvadrātisko novirzi. Ģeometrijas, kas veidotas, balstoties uz NKD+E matricu ar hierarhisko maksimālo distanču klāsteranalīzi** (izveidojis autors, izmantojot *LibreOffice Calc* programmatūru).

Attēli 5.3.1. un 5.3.2. atspoguļo modeļsistēmas aprēķinu rezultātus ar visām šī darba ietvaros sagatavotajām eksperimentālajām kvartāra nogulumu ģeometrijām, kas integrētas MOSYS V1 ar autokalibrāciju jau iepriekš optimizētajā bāzes ģeometrijā. Salīdzinot abas eksperimentu grupas, uz NKD matricu balstītie rezultāti ir labāki nekā uz NKD+E matricu balstītie. NKD matricas rezultātiem ir zemāka summāro kvadrātisko noviržu summa, mediāna un vidējais aritmētiskais, kā parādīts tabulā 5.3.1. Tas uzskatāmi parāda, ka kopsummā šī eksperimentu grupa dod labākus rezultātus, jo to summārās kvadrātiskās novirzes ir mazākas.

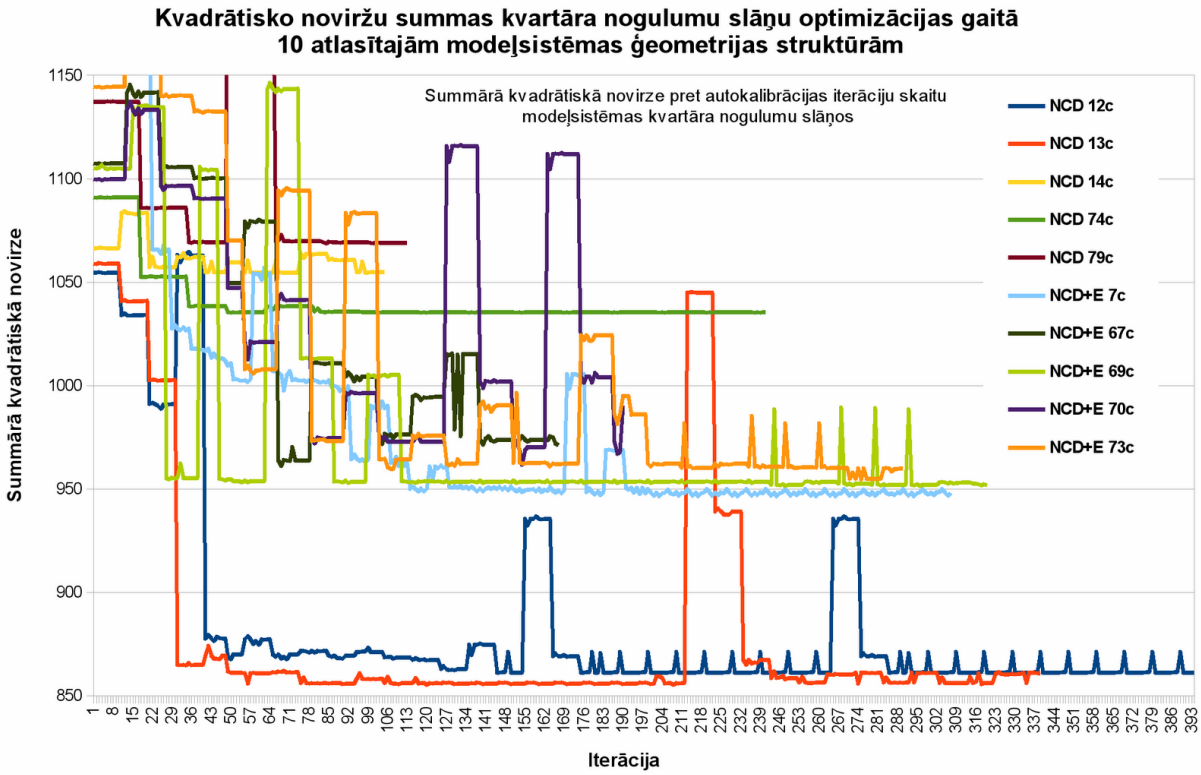
5.3.1. tabula.

Statistika par NKD un NKD+E eksperimentu grupu rezultātiem kopumā

	NKD eksperimentiem (summārās kvadrātiskās novirzes)	**NKD+E eksperimenti** (summārās kvadrātiskās novirzes)
Vidējais aritmētiskais	1098	1226
Summa	90053	99374
Mediāna	1095	1147

5.4. Filtrācijas modeļa ģeometrijas struktūru optimizācija kvartāra slāņos

Attēlā 5.4.1. parādīta 10 atlasīto ģeometriju autokalibrācijas gaita. Šīs ģeometrijas tika atlasītas, vadoties pēc sākotnējiem aprēķinu rezultātiem, kas atspoguļoti attēlā 5.2.1. (uz NKD matricu bāzētajiem eksperimentiem) un attēlā 5.2.3. (uz NKD+E matricu balstītajiem eksperimentiem). No katras eksperimentu matricas tika ņemtas 5 ģeometrijas: 3 labākās, 1 no ģeometrijām ar vidējiem rezultātiem un 1 ar sliktāko rezultātu konkrētajai eksperimentu grupai.



5.4.1. attēls. **Summārās kvadrātiskās novirzes pret iterāciju skaitu 10 atlasītajam modeļsistēmas ģeometrijām kvartāra slāņu kalibrācijas gaitā** (izveidojis autors, izmantojot *pyMOSYS* autokalibrācijas funkciju un *LibreOffice Calc* programmatūru).

Optimizācijas funkcijas gaita nav deterministiska un liecina tikai par labāko tās gaitā sasniegto rezultātu katrai ģeometrijai. Tomēr, kā redzams attēlā 5.4.1., katra no izvēlētajām ģeometrijām atšķirīgi reaģē uz optimizāciju. Tas sniedz ieskatu, kā materiālu īpašību reizināšana ar algoritma piekārtotajiem koeficientiem, kalibrācijas uzdevumā iekļautajiem kvartāra nogulumu slāņiem, maina summāro kvadrātisko novirzi visa aprēķina rezultātā. NKD 12k un NKD 13k modeļa ģeometrijas struktūras uzrāda labākus rezultātus nekā pārējās.

5.5. Iegūtie uzlabojumi

Ģeometrijas NKD 12k un NKD 13k, kas pēc kvartāra slāņu optimizācijas uzrādīja labākos rezultātus, tika optimizētas tālāk, pakļaujot autokalibrācijas mehānismam arī pamatiežu slāņus, lai pēc iespējas vairāk samazinātu summāro kvadrātisko novirzi, kas šajā darbā tika izmantota par galveno kritēriju, pēc kura tika vērtētas modeļsistēmas eksperimentālās kvartāra ģeometriju struktūras. Iegūtie uzlabojumi tika izvērtēti attiecībā pret rezultātiem, kas iegūti, izmantojot oriģinālo MOSYS V1 filtrācijas modeļa ģeometrijas struktūru. Tās summārā kvadrātiskā novirze visos kalibrācijas slāņos ir 814 metri (Virbulis et al., 2012).

Mazākās iegūtās vērtības summārajai kvadrātiskajai novirzei pēc autokalibrācijas ir **762 m eksperimentam NKD 13k un 732m eksperimentam NKD 12k. NKD 12k** ģeometrijas rezultāts sastāda **10% uzlabojumu** attiecībā pret oriģinālo MOSYS V1 optimizēto ģeometriju, uz kuru tika balstītas visas šī darba gaitā sagatavotās materiālu ģeometrijas struktūras. Otras labākās ģeometrijas struktūras (NKD 13k) rezultāts atbilst 6% uzlabojumam.

6. REZULTĀTU ANALĪZE UN DISKUSIJA

Aplūkojot darba gaitā iegūtos starprezultātus un gala rezultātus optimizētajām ģeometrijām, ir nepieciešams atskatīties uz iegūto rezultātu kopsummu.

Darba rezultātā ir iegūtas kvartāra nogulumu reprezentācijas 3D materiālu struktūras 162 variantos, kuri iekļauti Baltijas artēziskā baseina hidroģeoloģiskā modeļa MOSYS V1 filtrācijas modeļa ģeometrijas struktūrā.

Kā rāda attēli 5.3.1., 5.3.2 un tabula 5.3.1., tad uz **NKD** matricas balstītās ģeneralizēto kvartāra nogulumu ģeometrijas struktūras dod labākus rezultātus nekā uz **NKD+E** matricas balstītās struktūras pirms atkārtotas modeļa kalibrācijas. Pēc **NKD** matricas dendrogrammas būtu bijis iespējams identificēt to, ka ģeometrijas NKD 12k, NKD 13k vai ģeomterijas ar tām līdzīgu loģisko klāsteru skaitu varētu dot veikmīgi ģeneralizētas reprezentācijas. Tas rāda, ka nākotnē, izmantojot šādu darba gaitu filtrācijas modeļa ģeometrijas struktūru izveidē, būtu iespējams ievērojami samazināt veidojamo eksperimentālo ģeometriju skaitu. Taču šeit jānorāda, ka to pašu nevar teikt par NKD+E eksperimentu grupu. Tā tika izveidota telpisko klāsteru izveides solī, kad uz NKD matricu balstītie telpiskie klāsteri vizuāli radīja pārāk lielas sadrumstalotības efektu. Lai ļoti vienkāršā veidā ietvertu urbumu blīvuma ietekemi kādā konkrētā reģionā un novērtētu tā ietekmi uz rezultātiem, tika izveidota NKD+E eksperimentu grupa. Iegūtie rezultāti pilnībā noraida šāda lietojuma nepieciešamību. Neviena no NKD+E eksperimentu grupas filtrācijas modeļa ģeometrijas struktūrām neuzrādīja NKD 12k un NKD 13k ģeometrijām tuvus rezultātus ne pirms, ne pēc kalibrācijas. Šādi rezultāti varētu norādīt uz to, ka urbumu attālumu "informācijas distances" un Eiklīda distanču dimensijas nav savienojamas šādā lietojumā. Iegūtie rezultāti rāda, ka uz NKD+E matricu balstītos eksperimentus šādā formā var neveikt vispār, jo tie dod sliktākus rezultātus nekā uz NKD matricu balstītie. Tā kā izveidotā darba gaita, faktiski, ir paredzēta litoloģiskās struktūras datu agregācijai uz reģionāli reprezentatīvu mērogu, tad šī informācija ir nozīmīga, jo norāda, ka tieši NKD matrica, kas apkopo informāciju par visu urbumu griezumu savstarpējo līdzību, ir labāk pielietojama šādai ģeneralizācijai.

Sagatavoto eksperimentālo 162 ģeometriju skaits, ar kurām veikti aprēķini, demonstrē, ka rezultāti ir iegūti sistemātiskā veidā un nav nejauši. Summārās kvadrātiskās novirzes aprēķins, ar visām 162 darba gaitā izveidotajām ģeometrijas struktūrām pirms atkārtotas kalibrācijas, norāda ne tikai uz labākajām, bet arī uz sliktākajām modeļsistēmas ģeometrijas struktūrām, kas iegūtas ar normalizēto kompresijas distanci kā ģeneralizācijas kritēriju. Tas ir nozīmīgi, jo tās kalpo arī kā negatīvā kontrole. Mainot kvartāra reprezentāciju modeļsistēmas

filtrācijas modelī, ir iespējams iegūt arī sliktus rezultātus, kas mazina iespēju, ka iegūtie summārie uzlabojumi ir nejauši. Tāpat tas norāda, ka uzlabot kvartāra nogulumu reprezentāciju MOSYS V1 filtrācijas modeļa ģeometrijā ir svarīgi labāku rezultātu ieguvei.

Darba gaitā tika atlasītas 2 perspektīvākās ģeometrijas struktūras (NKD 12k, NKD 13k), balstoties uz to rezultātiem kvartāra slāņu autokalibrācijas gaitā. Tām tika veikta filtrācijas koeficientu kalibrācija pamatiežu slāņos, lai tālāk uzlabotu iegūstamos rezultātus. Šīs divas filtrācijas modeļa ģeometrijas struktūras ieguva mazāku kvadrātisko noviržu summu nekā oriģinālā MOSYS V1 ģeometrijas struktūra, apstiprinot darbā izvirzīto hipotēzi. Ģeometrijas struktūras **NKD 12k** aprēķinu rezultāti rāda, ka labākā, darba gaitā iegūtā, ģeneralizētā kvatāra nogulumu reprezentācija samazināja kvadrātiskās novirzes summu modelēšanas sistēmas kalibrācijas slāņos par **10%**.

Tāpat jānorāda uz tehniskajām limitācijām attiecībā uz labāko iegūto rezultātu. Tā kā visas eksperimentālās ģeometrijas struktūras tika ievietotas optimizetajā MOSYS V1 materiālu struktūrā, tad būtu bijis vēlams jaunizveidotās kvartāra ģeometrijas reprezentācijas kalibrēt nošķirti no pārējās modeļa teritorijas. Tas tehniski nebija iespējams, bet pastāv varbūtība, ka, veicot kalibrāciju nošķirti jaunizveidotajām ģeometrijas daļām un pārējam modelim, varētu iegūt labākus rezultātus.

Aplūkojot šķērsgriezumus attēlos 5.2.4., 5.2.5., 5.2.9. un 5.2.10., var gūt ieskatu modeļa ģeometrijas struktūras heterogenitātē no materiālu īpašību viedokļa. Šāds materiālu īpašību izplatības saposmojums 3 dimensijās ir ievērojami kompleksāks par agrāko kvartāra nogulumu reprezentāciju ar četriem, materiālu īpašību ziņā homogēniem, slāņiem. Tas ietekmē slāņu kalibrāciju. Kalibrējot modeļsistēmas ģeometrijas struktūras slāņus, kuru ietvaros materiālu īpašības ir šādā veidā saposmotas, kalibrācijas algoritms, tāpat kā gadījumā ar homogēniem slāņiem, reizina visas slāņa materiālu īpašības ar vienu koeficientu. Tas nozīmē, ka proporcijas starp blakus esošu elementu materiālu īpašībām paliek nemainīgas (Bethers et al., 2011). Tādējādi, kalibrējot šādi saposmotus slāņus, kalibrācijas mehānismam ir mazākas iespējas kombinēt slāņus labāku rezultātu ieguvei, reizinot to materiālu īpašības ar piekārtotajiem koeficientiem, jo katrs slānis var būt saposmots neatkarīgi no pārējiem. Pastāv iespēja, ka, kalibrācijas laikā iegūstot uzlabotus rezultātus viena klāstera ietvaros, ir iespējams izraisīt negaidītus blakusefektus citā vietā. Tādējādi, pēc autora domām, šādi saposmots filtrācijas modelis, kas dod labus aprēķinu rezultātus, visdrīzak liecina par to, ka tajā ietvertā kvartāra nogulumu reprezentācija ir atbilstoša savam mērogam.

Tāpat, salīdzinot iegūto labāko kvartāra nogulumu slāņu reprezentācijas ar oriģinālo, kuru veidoja četri, materiālu īpašību ziņā homogēni, slāņi, var secināt, ka vertikālais saposmojums telpiskajiem klāsteriem raksturīgajos tipveida griezumos ar krasām robežām

starp tiem labāk reprezentē kvartāra nogulumu uzbūvi reģionālā mērogā nekā horizontālais saposmojums, kā tas ir oriģinālajā MOSYS V1 filtrācijas modelī.

Plašākā mērogā skatoties, šajā darbā izmantotā pieeja filtācijas modeļa ģeometrijas struktūras īpašību parametrizācijā ir radniecīga citu autoru darbiem, kuri nodarbojas ar litoloģiskās heterogenitātes izraisītās nenoteiktības kvantifikāciju modelēšanas problēmu risināšanai (Paradis et al., 2011; Quental et al., 2012; Ouellon et al., 2008; Caers, 2001; Masonnat, 1999). Šeit būtu jānodala divas pieejas: ģeostatistiskā pieeja (Quental et al., 2012; Ouellon et al., 2008; Caers, 2001; Masonnat, 1999) un otra, kuru varētu dēvēt par klasifikācijas-agregācijas pieeju. Šis darbs ir veltīts jaunai agregācijas pieejas darba gaitas izpētei, kas ir konceptuāli radniecīga Paradis et al. 2011. gadā aprakstītajai.

Uz ģeostatistiskajām metodēm balstītie risinājumu mēdz ļoti atšķirties pēc to pieejas un nav viegli iedalāmi sīkāk. Tomēr nereti to metodikas var izmantot modeļu ģeometrijas granularitātes sabiezināšanai, balstoties uz pieejamo ievades informāciju, lai gan tas, protams, nav obligāti. Savukārt, klasifikācija piedāvā sadalīt interesējošo teritoriju pēc vairākām reprezentatīvām klasēm un, lai to darītu, atrast potenciāli būtisko lielā datu apjomā. Līdz ar to, šāda pieeja ir pievilcīga tieši reģionālajā hidroģeoloģiskajā modelēšanā, jo ir svarīgi nošķirt būtiskāko no lokāli izkliedēta "trokšņa". Ja pieejamais urbumu griezumu datu apjoms ir pārāk neliels, lai pietiekami atbilstoši realitātei atspoguļotu situāciju dabā, tad iespēja nošķirt lokālo no reģionāli reprezentatīvā ir nozīmīga priekšrocība šādam NKD metrikas lietojumam.

Iegūti rezultāti paši par sevi nepierāda**,** ka darbā izstrādāto metodiku varēs vispārināt jebkuras stratificētas vides ģeneralizācijai Zemes sistēmu modelēšanā. Tomēr šis darbs demonstrē, ka materiālu īpašību 3D struktūras ģeneralizācija, kā kritēriju izmantojot normalizētās kompresijas distanci, ļauj uzlabot reģionālā hidroģeoloģiskā modeļa kvartāra slāņu reprezentāciju un būtu interesanti veikt līdzīgus eksperimentus citam līdzīgam lietojumam. Darbā izmantotais urbumu griezumu serializācijas paņēmiens ir adaptējams līdzīgu eksperimentu veikšanai cita veida stratificētas vides ģeneralizēcijai modelēšanas pētījumos (ja vien šī stratificētā vide ir stabila modelēšanas vajadzībām noderīgā laika periodā).

7. SECINĀJUMI

Ir izpildīti darbam izvirzītie uzdevumi un **sasniegts darba mērķis. Ir izstrādāta darba gaita,** kas ļauj izmantot NKD un hierarhisko klāsteranalīzi kvartāra nogulumu reprezentācija ģeneralizēšanai materiālu īpašību 3D struktūrā, kas veido MOSYS modeļsistēmas filtrācijas modeli.

Iegūtie rezultāti **apstiprina** darba sākumā izteikto **hipotēzi**, ka normalizēto kompresijas distanci var izmantot materiālu īpašību 3D struktūras ģeneralizācijā. Ar šajā darbā izstrādāto metodiku ir iegūta ģeneralizēta kvartāra nogulumu slāņu reprezentācija reģionālā hidroģeoloģiskajā modelī MOSYS V1, kas samazina summāro kvadrātisko novirzi sistēmas kalibrācijas slāņos par 10% attiecībā pret līdzšinējo kvartāra nogulumu reprezentāciju filtrācijas modeļa ģeometrijā.

Normalizētās Eiklīda distanču matricas kombinēšana ar NKD matricu (NKD+E eksperimentu grupa) pasliktina aprēķinu rezultātus. Tas norāda, ka tās viena otru nepapildina noderīgā veidā. Savukārt, NKD metrika, lietota atsevišķi, ir pietiekama, lai, no reģionālā hidroģeoloģiskā modeļa viedokļa, veiksmīgi klasificētu kvartāra nogulumu uzbūves ziņā līdzīgas teritorijas.

Telpisko klāsteru veidotais vertikālais saposmojums filtrācijas modeļa ģeometrijas struktūrā ir reprezentatīvāks nekā tikai horizontālais saposmojums, kā agrākajā modeļsistēmas kvartāra slāņu reprezentācijā.

PATEICĪBAS

Šī darba izstrāde nebūtu bijusi iespējama bez daudzu cilvēku tehniskā, konsultatīvā un morālā atbalsta.

Vislielākā pateicība manai ģimenei par neizsīkstošo atbalstu un pacietību ikdienā, kas ļāva man veltīt darbam nepieciešamo laiku, lai izstrādātu to līdz šeit apkopotajam veselumam. Paldies Ievai un Ancei par neaizstājamo darbu pie pareizrakstības kļūdu labošanas un Mārim par mēnesi ziedotā datorlaika.

Tāpat izsaku pateicību ESF projektam „Starpnozaru zinātnieku grupas un modeļu sistēmas izveide pazemes ūdeņu pētījumiem" (2009/0212/1DP/1.1.1.2.0/09/APIA/VIAA/060), kas finansiāli atbalstīja šī pētījuma tapšanu un aprobāciju Eiropas lielākajā ģeozinātņu konferencē. Paldies tā pieteikuma autoriem Andim Kalvānam un darba vadītājam Tomasam Sakam gan par daudzajiem vērtīgajiem padomiem un konsultatīvo atbalstu, gan par iespēju strādāt šajā unikālajā pētījumu projektā, risinot sabiedrībai un zinātnei aktuālas problēmas un iegūstot nenovērtējamu pieredzi starpdisciplinārā Eiropas mēroga pētījumu projektā.

Sirsnīgs paldies PUMa komandas kolēģiem, bez kuru palīdzības šis darba nebūtu tapis tā pašreizējā formā. Jurim Seņņikovam, kurš diskusijā pie kafijas tases ieminējās par pamata principiem idejai, pēc kuras es vēlāk realizēju urbumu griezumu serializācijas metodiku; Andrejam Timuhinam, kurš ieguldīja vairākus vēlus vakarus uzlabotā MOSYS kalibrācijas scenārija ātrākai pabeigšanai. Tāpat vēlos pateikties autokalibrācijas sistēmas līdzautorei Ilzei Klintij par daudzajiem praktiskajiem ieteikumiem darbā ar autokalibrācijas scenārijiem un par pacietību, palīdzot man nokomplektēt modeļsistēmas bāzes versiju ģeometrijas uzlabojumu testēšanai. Paldies Konrādam Popovam par ilgo darbu pie esošās ģeometrijas bāzes pamatiežu slāņiem, uz kuru balstīti šajā darbā aprakstītie eksperimenti.

Nobeigumā es vēlētos pieminēt, ka šī darba izstrāde nebūtu bijusi iespējama bez daudzajiem atvērtā koda programmatūras komponentiem, kuri veido daļu no šī darba tehniskās gaitas. Nav grūti iztēloties, ka vēl tikai 10 gadus atpakaļ šāda darba izstrādi būtiski apgrūtinātu pamata metodes realizējošo programmatūras komponentu pieejamība.

LITERATŪRA

Publicētā

Aurenhammer, F. 1991. Voronoi Diagrams - A Survey of a Fundamental Data Structure. *ACM Computing Surveys*. 23(3), 345-405.

Benedetto, D., Caglioti, E., Loreto, V. 2002. Language trees and zipping. *Physical Review Letters*. 88(4), 2-5.

Bennett, C.H., Gacs, P., Li, M., Vitanyi, P., Zurek, W. 1998. Information Distance. *IEEE Transactions on Information Theory*. 44(4), 1407-1423.

Caers, J. 2001. Geostatistical reservoir modelling using statistical pattern recognition. *Journal of Petroleum Science and Engineering*. 29(3/4), 177–188.

Cebrian, M., Alfonseca, M., Ortega, A. 2005. Common Pitfalls Using the Normalized Compression Distance: What to Watch Out for in a Compressor. *Communications in Information & Systems*. 5(4), 367-384.

Cebrian, M., Alfonseca, M., Ortega, A. 2007. The Normalized Compression Distance Is Resistant to Noise. *IEEE Transactions of Information Theory*. 53(5), 1895 – 1900.

Cerra, D., Mallet, A., Gueguen, L., Datcu, M. 2010. Complexity Based Analysis of Earth Observation Imagery: an Assessment. *IEEE Geoscience and Remote Sensing Letters*. 7(1), 8-12.

Cilibrasi, R. 2007. *Statistical Inference Through Data Compression*. ILLC Dissertation Series DS-2007-01, Institute for Logic, Language and Computation, Universiteit van Amsterdam.

Cilibrasi, R., Vitanyi, P. 2005. Clustering by Compression. *IEEE Transactions of Information Theory*. 51(4), 1523 - 1545.

Cilibrasi, R., Vitanyi, P.M.B., Wolf, R. 2004. Algorithmic clustering of music based on string compression. *Computer Music Journal*. 28(4), 49-67.

Cleary, J.G., Witten, I.H., 1984. Data compression using adaptive coding and partial string matching. *IEEE Transactions on Communications*. 32(4), 396–402.

Dēliņa, A., Saks, T., Jātnieks, J. un Popovs, K. 2011. Baltijas artēziskā baseina ģeoloģiskā uzbūve hidroģeoloģiskajam modelim - pieejamo datu implementācija un problēmas. *Latvijas Universitātes 69. zinātniskā konference. Ģeoloģijas sekcijas apakšsekcija "Baltijas artēziskā baseina pazemes ūdeņi." Referātu tēzes.* Rīga, Latvijas Universitāte, 30.-32.

Feyen, L., Caers, J. 2006. Quantifying geological uncertainty for flow and transport modeling in multi-modal heterogeneous formations. *Advances in Water Resources*. 29(6), 912–929.

Gabriel, E., Fagg, G.E., Bosilca, G., Angskun, T., Dongarra, J.H., Squyres, J.M., Sahay V., Kambadur, P., Barrett, B., Lumsdaine, A., Castain, R.H., Daniel, D.J., Graham, R.L. and Woodall, T.S. 2004. Open MPI: Goals, Concept, and Design of a Next Generation MPI Implementation. *Proceedings, 11th European PVM/MPI Users' Group Meeting*. Budapest, Hungary, 97-104.

Gallier, J. 2011. *Geometric Methods and Applications: For Computer Science and Engineering*. New York [etc], Springer-Verlag, 274.

Gorokhovski, V. 2012. *Effective Parameters of Hydrogeological Models*. Heidelberg [etc], Springer.

Han, J., Kamber, M., Pei, J. 2012. *Data Mining: Concepts and Techniques.* 3rd edn. Waltham (Massachusetts), Morgan Kaufmann Publishers.

Jain, A.K., Murty, M.N., Flynn, P.J. 1999. Data Clustering: A Review. *ACM Computing Surveys*. 31(3), 264-323.

Jain, A.K. 2010. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*. 31(8), 651-666.

Jatnieks, J., Popovs, K., Klints, I., Timuhins, A., Kalvans, A., Delina, A., Saks, T. 2012*a*. Scalable Generalization of Hydraulic Conductivity in Quaternary Strata for Use in a Regional Groundwater Model. *Geophysical Research Abstracts*. EGU General Assembly 2012, Vienna, Austria, EGU2012-6550-1.

Jatnieks, J., Saks, T., Delina, A., Popovs, K. 2012*b*. Lithological Uncertainty Expressed by Normalized Compression Distance, *Geophysical Research Abstracts*. EGU General Assembly 2012, Vienna, Austria, EGU2012-540.

Jātnieks, J., Popovs, K., Ukass, J., Karušs, J., Pērkone, E., Raga, B., Babre, A., Saks, T. 2011*a*. Ģeotelpisko datu sagatavošana PUMa projekta ietvaros. *Latvijas Universitātes 69. zinātniskā konference. Ģeogrāfija, Ģeoloģija, Vides zinātne. Referātu tēzes.* Rīga, LU Akadēmiskais apgāds, 305.-306.

Jātnieks, J., Popovs, K., Ukass, J., Saks, T., Dēliņa, A. 2012*c*. Useful statistics for describing hydraulic conductivity of the Quaternary strata from the Latvian borehole log data. *The 70th Scientific Conference of the University of Latvia, Session of Geology, Section "Groundwater in Sedimentary Basins". Abstract Book*. Riga, University of Latvia, 35.-36.

Kantardzic, M. 2011. *Data mining : concepts, models, methods, and algorithms.* 2nd edn. Hoboken, New Jersey, John Wiley & Sons Ltd.

Lance, P., Haque, E., Liu, H. 2004. Subspace Clustering for High Dimensional Data: A Review. *ACM SIGKDD Explorations Newsletter.* 6(1), 90-105.

Legg, S., Hutter, M. 2007. Universal Intelligence: A Definition of Machine Intelligence. *Minds and Machines*. 17(4), 391-444.

Li, M., Chen, X., Li, X., Ma, B., Vitanyi, P. 2004. The similarity metric. *IEEE Transactions of Information Theory*. 50(12), 3250 – 3264.

Li, M., Vitanyi, P. 2008. *An Introduction to Kolmogorov Complexity and its Applications*. 3rd edn. New York, Springer-Verlag.

Massonnat, G.J. 1999. Breaking of a Paradigm: Geology Can Provide 3D Complex Probability Fields for Stochastic Facies Modelling. *SPE annual technical conference and exhibition: Houston TX, 3-6 October 1999. Volume Omega: Formation evaluation and reservoir geology*. Houston, Texas, 329-342.

Okabe, A., Boots, B., Sugihara, K., Chiu, S.N. 2000. *Spatial tessellations: concepts and applications of Voronoi diagram*. Chichester, John Wiley and Sons.

Ouellon, T., Lefebvre, R., Marcotte, D., Boutin, A., Blais, V., Parent, M. 2008. Hydraulic conductivity heterogeneity of a local deltaic aquifer system from the kriged 3D distribution of hydrofacies from borehole logs, Valcartier, Canada. *Journal of Hydrology*. 351(1/2), 71–86.

Paradis, D., Gloaguen, E., Lefebvre, R., Rivera, A. 2011. Tying geophysics to hydrogeology: a learning machine approach to characterise heterogeneous granular aquifers. *GeoHydro2011 Proceedings Papers.* Quebec, Canada.

Plant, C., Bohm, C. 2009. Novel Trends in Clustering. In: Furtado, P.N. (ed.). *Evolving Application Domains of Data Warehousing and Mining: Trends and Solutions*. Hershey, IGI Global, 185-211.

Quental, P., Almeida, J.A., Simões, M. 2012. Construction of high-resolution stochastic geological models and optimal upscaling to a simplified layer-type hydrogeological model. *Advances in Water Resources*. 39, 18-32.

Ramanath, S., Dathan, B. 2010. *Object-Oriented Analysis and Design*. London [etc.], Springer, 105.

Refsgaard, J.C., Christensen, S., Sonnenborg, T.O., Seifert, D., Højberg, A.L., Troldborg, L. 2011. Review of strategies for handling geological uncertainty in groundwater flow and transport modelling. *Advances in Water Resources*. 36, 36-50.

Roman-Gonzalez, A., Veganzones, M.A., Grana, M., Datcu, M. 2011. A novel data compression technique for remote sensing data mining. *Proceedings of the ESA-EUSC-JRC 2011*. Ispra, Italy, 101-104.

Saks, T., Kalvāns, A., Seņņikovs, J., Timuhins, A., Dēliņa, A. 2011. Kvartāra nogulumu ģeoloģiskās uzbūves atspoguļojuma algoritmizācija reģionālā hidroģeoloģiskā modelī.

Latvijas Universitātes 69. zinātniskā konference. Ģeogrāfija, Ģeoloģija, Vides zinātne. Referātu tēzes. Rīga, LU Akadēmiskais apgāds, 369.–370.

Shkarin, D. 2002. PPM: one step to practicality. *Proceedings of the Data Compression Conference 2002*, 202- 211.

Solomon, D. 2007. *Data Compression The Complete Reference*. 4th edn. London, Springer-Verlag.

Stepinski, T., Vilata, R., Ghosh, S. 2007. Machine Learning Tools for Automatic Mapping of Martian Landforms. *IEEE Intelligent Systems*. 22(6), 100-106.

Virbulis, J., Klints, I., Timuhins, A., Sennikovs, J., Bethers, U. 2012. Calibration of the hydrogeological model of the Baltic Artesian Basin. *Geophysical Research Abstracts*. EGU General Assembly 2012, Vienna, Austria, EGU2012-10003-1.

Wehner, S. 2007. Analyzing worms and network traffic using compression. *Journal of Computer Security*. 15(3), 303-320.

Zhu, C., Byrd, R.H., Nocedal. J. 1997. Algorithm 778: L-BFGS-B, FORTRAN routines for large scale bound constrained optimization. *ACM Transactions on Mathematical Software.* 23(4), 550 - 560.

Nepublicētā

Achert, E. 2007. Hierarchical Subspace Clustering: Dissertation. Munich, Faculty of Mathematics, Computer Science and Statistics, Ludwig Maximilian University of Munich.

Berkhin, P. 2006. *A Survey of Clustering Data Mining Techniques*. Accrue Software Inc. Sk. 27.05.2012. Pieejams: http://csis.pace.edu/~ctappert/dps/d861-12/session3-p2.pdf
Atsauce tekstā (Berkhin, 2006).

Bethers, U., Saks, T., Sennikovs, J., Timuhins, A., Virbulis, J. 2011. Hydrogeological model of the Baltic Artesian Basin (submitted for *Hydrogeology Journal*).

GDAL. 2012. *GDAL - Geospatial Data Abstraction Library: Version 1.9.0*. Open Source Geospatial Foundation. Sk. 27.05.2012. Pieejams: www.gdal.org
Atsauce tekstā (GDAL, 2012).

Gillies, S. 2012. *The Shapely User Manual Revision 1.2.14*. github.com. Sk. 27.05.2012. Pieejams: http://toblerity.github.com/shapely/manual.html
Atsauce tekstā (Gillies, 2102).

Hoon, M., Imoto, S., Miyano, S. 2010. *The C Clustering Library*. The University of Tokyo, Institute of Medical Science, Human Genome Center. Sk. 27.05.2012. Pieejams: http://bonsai.hgc.jp/~mdehoon/software/cluster/cluster.pdf
Atsauce tekstā (Hoon et al., 2010).

Hutter, M. 2006. Hutter 50'000€ Prize for Compressing Human Knowledge. Sk. 27.05.2012. Pieejams: http://prize.hutter1.net/

Atsauce tekstā (Hutter, 2006).

Mahoney, M. 2011. *Data Compression Explained*. Dell Inc. Sk. 12.12.2011. Pieejams: http://mattmahoney.net/dc/dce.html

Atsauce tekstā (Mahoney, 2011).

Mahoney, M. 1999. *Text Compression as a Test for Artificial Intelligence*. Florida Institute of Technology. Sk. 12.12.2011. Pieejams: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.48.7038&rep=rep1&type=ps Atsauce tekstā (Mahoney, 1999).

Jarvis, A., Reuter, H.I., Nelson, A. & Guevara, E. 2008. Hole-filled SRTM for the globe Version 4. CGIAR-CSI SRTM 90m Database. Sk. 10.11.2011. Pieejams: http://www.cgiar-csi.org/data/elevation/item/45-srtm-90m-digital-elevation-database-v4

Atsauce tekstā (Jarvis et al., 2008).

Jātnieks, J. 2011*b*. Kvartāra datubāzes urbumu telpiskās informācijas piesaiste, izmantojot ģeokodēšanas elementus pa ierobežotiem Voronoi poligoniem: studiju darbs LĢIS/TMAI. Rīga, LU Ģeogrāfijas un Zemes zinātņu fakultāte, Latvijas Universitāte.

Popovs, K. 2011. Geoloģiskās struktūras trīsdimensiju modeļa algoritmizēta izveide Latvijas teritorijai: maģistra darbs. Rīga, LU Ģeogrāfijas un Zemes zinātņu fakultāte, Latvijas Universitāte.

Popovs, K., Saks, T., Ukass, J., Jatnieks, J. 2012. 3D geological model of the Latvia sedimentary cover through genetic algorithms (submitted for *Estonian Journal of Earth Sciences*).

Seņņikovs, J. 2011. *Baltijas artēziskā baseina hidroģeoloģiskā modeļa versija V1: attīstība otrajā projekta gadā.* Latvijas Universitāte, Sk. 27.05.2012. Pieejams: http://www.puma.lu.lv/fileadmin/user_upload/lu_portal/projekti/puma/V1prezentacija/V1_SennikovsJ.pdf

Atsauce tekstā (Seņņikovs, 2011).

Stankevica, I. 2011. Baltijas artēziskā baseina hidroģeoloģiskā modeļa kalibrācija: bakalaura darbs. Rīga, LU Fizikas un matemātikas fakultāte, Latvijas Universitāte.

Takčidi, E. 1999. Datu bāzes "Urbumi" dokumentācija, Valsts ģeoloģijas dienests, Rīga.

1. Pielikums – Darba gaitas automatizēšanas scenāriji

Pielikuma saturs ir noderīgs elektroniskā formā un ir lejuplādējams no adreses:
http://home.lu.lv/~sg30022/Mag_1_Pielikums_scenarijprogrammas.7z

---

2. Pielikums – Telpisko klāsteru izvietojums

Pielikuma saturs ir noderīgs elektroniskā formā un ir lejuplādējams no adreses:
http://home.lu.lv/~sg30022/Mag_2_Pielikums_Telpiskie_klasteri.7z