PROGRAMMI VALDKONDLIKU TEADUS- JA ARENDUSTEGEVUSE TUGEVDAMINE (RITA)
TEGEVUSE 1 TEENUSE OSUTAMISE LEPING nr 7.8-3/18/17

Kaugseire andmete kasutuselevõtt avalike teenuste väljatöötamisel ja arendamisel

Lisa 3

DOI: 10.23673/re-259

Põllumajandusmaade kasutuse seire

LÕPPARUANNE

Dokumendi koostasid: Kaupo Voormansik, Mihkel Järveoja, Marharyta Domnich, Indrek Sünter, Tanel Tamm, Mait Lang, Valentina Sagris, Tõnu Oja ja Kalev Sepp

Tartu 2020

TARTU ÜLIKOOL
Eesti Maaülikool Estonian University of Life Sciences
TAL TECH
KAPPA ZETA
Eesti Teadusagentuur Estonian Research Council
Euroopa Liit Euroopa Regionaalarengu Fond
Eesti tuleviku heaks
KESKKONNAMINISTEERIUM
SISEMINISTEERIUM
MAJANDUS- JA KOMMUNIKATSIOONI-MINISTEERIUM
MAAELUMINISTEERIUM

Metoodika

Põllukultuuride tuvastamise metoodika lõppversioon on kirjeldatud lisatud dokumendis „D4.8 Põllukultuuride tuvastusmudeli kirjeldus koos täpsushinnangutega, 4. iteratsioon (lõpparuanne) – sügis 2020“ lk 46-94.

Põldude Sentinel-1 ja -2 aegridadest eksete eemaldamise metoodika on kirjeldatud lisatud dokumendis „D4.5 Lisa 1. Põllukultuuride tunnusvektorite aegridade rühmitamine klasterdamise abil ja eksete eemaldamine õpetusandmetest – Metoodika kirjeldus“ lk 36-45.

Mullakaardi andmete rakendamise metoodika on toodud lisatud dokumendis „D4.1 Kirjanduse ülevaade ja esialgne põllukultuuride tuvastusmudeli kirjeldus“ lk 6-35.

Testimine ja veahinnangud

Testimise metoodika, testimise tulemused ja veahinnangud on analüüsitud lisatud dokumendis „D4.8 Põllukultuuride tuvastusmudeli kirjeldus koos täpsushinnangutega, 4. iteratsioon (lõpparuanne) – sügis 2020“ lk 16-38.

Hinnangud ja soovitused

Hinnangud ja soovitused põllukultuuride tuvastusmetoodika Eestis rakendamiseks on toodud lisatud dokumendis „D4.8 Põllukultuuride tuvastusmudeli kirjeldus koos täpsushinnangutega, 4. iteratsioon (lõpparuanne) – sügis 2020“ lk 38-40.

Prototüüptarkvara koos metoodika testimiseks vajaliku näidisandmestikuga on vabalt kättesaadav siit: https://bitbucket.org/kappazeta/rita-evaluator/src/master/

Kasutusalad

Väljatöötatud satelliitseire-põhist põllukultuuride tuvastamise metoodikat saab rakendada järgmisteks kasutusjuhtudeks:

1. Põllumaade pindalatoetuste kontroll Eesti Põllumajanduse Registrite ja Informatsiooni Ametis (PRIAs).
2. Riiklikuks objektiivseks põllumajandusmaade statistikaks, et leida kui mitmel hektaril ja põllul igat põllukultuuri Eestis igal aastal kasvatati.
3. Täppispõllunduse rakendustes tehnoloogilise komponendina, kus edasisteks teenusteks ja analüüsideks on vaja teada, mis põllukultuuri igal aastal igal põllul kasvatati.

Maksumusehinnang

Enne täpsema maksumuse hinnangu juurde minekut peaks PRIA, kui põllukultuuride tuvastamise info peamine kasutaja, mõtlema kuidas ta soovib sateliitseire monitooringule üle mina, sest lahenduse kogukulu (s.h. PRIA oma töötajate tööaeg mitme aasta peale kokku) ja saadud tulemuse kvaliteet sõltub tugevalt valitud lähenemisest. Teiste Euroopa riikide praktikaid uurides leiame peamiselt kaks lähenemist:

A. Osta satelliitseire info sisse hooajapõhise teenusena. Teenusepakkuja tarnib makseagentuurile igal aastal kõigi valitud põldude kohta niitmise tuvastamise ja põllukultuuride määramise tulemused vastavalt kokkulepitud graafikule (nt. juuni lõpus, juuli lõpus ja septembri alguses). Tulemused tehakse kättesaadavaks veebikaardil, *.SHP failide ning masinloetava API kaudu. Antud lähenemise on valinud nt Taani ja Hollandi põllumajandustoetuste makseagentuurid.
B. Arendada välja monitooringuks vajalik infosüsteem makseagentuuri sees. Teenusepakkujalt ei osteta tulemust vaid arendustöö tundi. Seda teed näib minevat nt Poola makseagentuur.

Mõlemal lähenemisel on oma plussid ja miinused, vt Tabel 1 ja Tabel 2.

Tabel 1. Monitooringule üleminek teenuste ostmise kaudu, plussid ja miinused.

Teenuste ostmise plussid +	Teenuste ostmise miinused -
1. Pakkujal on huvi pakkuda parimaid lahendusi. 2. Uuendused jõuavad PRIAni jooksvalt. 3. Pakkuja vahetamine on lihtne. 4. Teenuste globaalne konkurents langetab hinda ja tõstab kvaliteeti. 5. PRIA saab rohkem tegeleda küsimusega „Mida kasulikku saab kaugseirest tulnud teabega teha?“, mitte „Kuidas kaugseirega kasulikku teavet kätte saada?“	1. PRIAl puudub otsene kontroll teenuste arenduse üle. Saab öelda, mida on vaja teha, aga mitte nii väga kuidas seda saavutada.

Tabel 2. Monitooringule üleminek majasiseste infosüsteemide arendamise kaudu, plussid ja miinused.

Infosüsteemi arendus +	Infosüsteemi arendus -
1. Esineb teoreetiline võimalus, et PRIA saab sõltumata välistest osapooltest süsteemi kasutada ja edasi arendada. 2. Rohkem sõltumatust (aga ka rohkem kohustusi ise kõigi väljakutsetega hakkama saada ning oma infosüsteemid kaasajas hoida).	1. Pakkujad, kes pakuvad sarnast teenust globaalselt, ei soovi ega saa oma parimaid lahendusi PRIAle maha müüa. 2. Pakkuja lahenduste uuendused ei jõua PRIA käsutusse jooksvalt. 3. Süsteemi keerukuse tõttu on arendaja vahetamine äärmiselt raske.

	4. Tehniliste küsimuste lahendamisele kuluv ressurss tuleb sisu arvelt. 5. Maailmaga sammu pidamine on kallis, sest kõik uued asjad tuleb oma majas uuesti arendada, ka vanade süsteemide käimas hoidmine on arvestatav aja- ja rahakulu.

Vaadates alternatiivsete lahenduste plusse ja miinuseid veel kõrgemalt, Eesti riigi ja eriti Majandus- ja Kommunikatsiooniministeeriumi poliitikate ja programmide vaatenurgast ja eeldades, et teenusepakkujaks on Eesti ettevõte on siin veel üks oluline aspekt. Eesti ettevõtte majandusedu saavutamine ja eksport on oluliselt lihtsamini saavutatavad, kui monitooringu lahendused ja töötlusahel on üles ehitatud teenusena. Sama teenust teistesse Euroopa riikidesse (nt. Lätti või Soome) müüa on oluliselt lihtsam, kui hakata igas riigis uut infosüsteemi välja arendama. Eesti IT-ettevõtete lisandväärtus töötaja kohta on oluliselt väiksem kui Soome IT-ettevõtetel, sest Eesti IT-ettevõtted müüvad enamasti (riigiasutustele) oma arendustöö tundi, mitte ei müü üle maailma kasutatavaid teenuseid.

Põllukultuuride tuvastamise operatiivrakenduse välja arendamiseks PRIA majasisese infosüsteemina tuleb:

1. Viia läbi detailne analüüs ja tarkvara-arenduse kavandamine. Selgitamaks välja kui suures ulatuses saab kasutada PRIA olemasolevaid infosüsteeme, kui palju tuleb olemasolevaid infosüsteeme uuendada ning kui palju tuleb arendada välja uusi komponente. Plaan korrektselt dokumenteerida. Töömahu hinnang on 240-400 h, 70 € tunnihinna korral on see 16 800 – 28 000 eurot.
2. Uuendada Sentinel-1 ja -2 tunnuskomplekti aegridade arvutamise töötlusahelat. Töömahu hinnang on 960-1440 h, 70 € tunnihinna korral on see 67 200 – 100 800 eurot.
3. Vastavalt RITA projekti raames välja töötatud metoodikale arendada välja tuvastusmudeli tarkvara koos juurdekuuluva äriloogikaga. Töömahu hinnang on 1120-1760 h, 70 € tunnihinna korral on see 78 400 – 123 200 eurot.
4. Töötada välja rakenduse kasutajaliides ja uuendada masinliideseid PRIA olemasolevate infosüsteemidega ühendumiseks. Töömahu hinnang on 640-960 h, 70 € tunnihinna korral on see 44 800 – 67 200 eurot.
5. Saadud lahendust testida, vigu parandada ja siluda. Arvestades SATIKAS infosüsteemi arendamise kogemust kulub selleks arvatavasti 1-2 aastat, sest kõik erijuhud ja ka ESA andmete tehniliste tingimuste muudatused ei pruugi kohe avalduda. Töömahu hinnang on 960-1600 h, 70 € tunnihinna korral on see 67 200 – 112 000 eurot.

Kokkuvõttes kuluks järelikult põllukultuuride tuvastamise operatiivrakenduse PRIA majasisese infosüsteemina välja arendamiseks 274 400 – 431 200 eurot.

Ostes põllukultuuride tuvastamise infot hooajapõhise teenusena maksaks see OÜ KappaZeta hinnangul 36 000 eurot aastas, millele lisanduks ühekordne integratsiooni kulu hinnaga 100 000 eurot. Hind teenusena tuleb oluliselt odavam paljuski seetõttu, et on võimalik kasutada juba olemasolevaid OÜ KappaZeta satelliitandmete töötlusahela tarkvarakomponente ja neid ei tule otsast peale uuesti välja arendada.

Andmed

Projekti käigus loodud peamine andmekogu on Eesti põldude tunnuste aegridade komplekt. Aegread on 2018 ja 2019 vegetatsiooniperioodist, mõlema aasta puhul enam kui 112 000 põllu kohta. Aegread sisaldavad Sentinel-1 ja -2 parameetreid, ilmaandmeid (sademesummad, temperatuurid) ja täiendavaid ruumiandmeid (mullatüüp ja normeeritud asukoha-koordinaadid). Andmed on esitatud CSV-tabelite ja Postgre SQL andmebaasi-väljavõtetena ning leitavad Tartu Ülikooli Raamatukogu DataDOI süsteemist aadressilt: http://dx.doi.org/10.23673/re-261

Andmed on avalikud ja mõeldud tasuta kasutamiseks, jagatud *Creative Commons* 4.0 litsentsiga.

Täpsem info andmete kohta: Mihkel Järveoja, e-post: mihkel.jarveoja@kappazeta.ee

D4.1 Kirjanduse ülevaade ja esialgne põllukultuuride tuvastusmudeli kirjeldus

Koostasid: OÜ KappaZeta, Tartu Ülikool ja Eesti Maaülikool

Projekt RITA1/02-52 „Kaugseire andmete kasutuselevõtt avalike teenuste väljatöötamisel ja arendamisel“

Tartu 2019

Sisukord

1 Põllukultuuride tuvastamine masinõppe abil kasutades satelliitmõõtmistel põhinevaid tunnuseid – kirjanduse ülevaade

Ghazaryan jt. (2018) leidsid, et Landsat-8 OLI multispektraalsetele andmetele Sentinel-1 mikrolainealas tehtud mõõtmiste lisamine parandab põlluviljade klassifitseerimise täpsust. Uuring tehti Ukrainas veidi alla 170 põllul ajavahemikus 2014-2016. Põllukultuurid jagati klassidesse: taliviljad, taliraps, mais, päevalill, soja, teised. Autorid kasutasid NDVI aegridade kirjeldamiseks mudelit, kus aegridu lähendati kahe erineva harmoonilise võnkumise abil. Saadud lähendite amplituudi ja faasi andmeid kasutati kirjeldavate tunnustena lisaks kolme kasvuperioodi heleduste keskmistele ja radaripiltide perioodi keskmistele. Mikrolaineala andmete lisamine andis klassifitseerimise täpsuses 2–5% lisaks multispektraalsetele andmetele. Üldine klassifitseerimise täpsus küündis 85-88 protsendini. Keksuvise perioodi piltide kasutamisel saadi üldiselt kevadiste või suviste perioodi piltidega võrreldes suurem täpsus, mida võib ka eeldada.

Roy ja Yan (2018) kasutasid sarnaselt Ghazaryan jt. (2018) tööga aegridade modelleerimist harmooniliste võnkumiste mudelite abil. Meetodit testiti USA-s kuuel 5000 × 5000 piksliga (30 m) katsealal. Põllukultuure kirjeldati NDVI aegridadega 2010 aastal, mis saadi Landsat-5 TM ja Landsat-7 ETM+ atmosfäärikorrektsiooniga piltidelt. Iga piksli aegrea kirjeldamise jaoks kasutati kas kaht või kolme harmoonilise võnkumise komponenti sõltuvalt vaatluste (piltide arvust), mis pidi olema vastavalt vähemalt 15 või 21. Sõltuvalt testalast varieerus klasside (sh mets, vesi jms) arv viiest neljateistkümneni. Fourier ridade kasutamine võib olla üks võimalustest aegridade andmeid siluda enne klassifitseerimise ja tuvastamise rakendamist. Eeldades, et põllukultuuri kirjeldava tunnuse ajaline käik on lähendatav erineva sagedusega harmooniliste võnkumiste kaalutud keskmisena, võib niiviisi saada klassifitseerimisalgoritmi jaoks paremini silutud aegread võrreldes lineaarse interpoleerimisega. Silumismudeli parameetrite lähendamine eeldab samas, et piltidelt arvutatavate tunnuste aegrida on pigem tihedalt vaatustega sisustatud. Vegetatsiooniperiooni alguses ja lõpus on seetõttu lähendid suurema veaga.

Kanjir jt (2018), kasutasid Sloveenias aegridade analüüsi meetodit BFAST Monitor niitude hooldamise ja põlluharimise anomaaliate tuvastamiseks. Uuringu empiiriline andmestik saadi kolmelt 7 km^2 suuruselt alalt. Meetod kasutab eeldatavalt korrapäraselt muutuvat tunnuse aegrida näidisena ja otsib uuel perioodil sellest hälbeid. Vaatlusühikuna kasutati polügonide keskmisi NDVI väärtuseid, mis saadi Sentinel-2 MSI piltide aegreast. Kuigi uuringu empiiriline andmestik oli pigem väike võrreldes teiste siintoodutega, siis selliste põllumaade jälgimiseks, kus võib eeldada kirjeldavate tunnuste perioodilist muutumist, tasub BFAST Monitor meetodit ühe testina kasutada. Kirjeldavate tunnuste eeldatava aegrea võib koostada näiteks varasemate satelliidipiltide ja kiirguslevi mudeli abil.

Conrad jt (2017) kasutasid otsustuspuude metsa (*random forest*) ja Landsat-5 TM pilte ning koostasid Fergana orus 377 278 ha põldude kohta 2010., 2011. ja 2012. aasta temaatilised

kaardid, kus eristati klasse puuvill, nisu jt teraviljad, viljapuuaiad, sööt, riis, teised ja kalatiigid. Eesmärk oli uurida põldude mitmekesisust sõltuvalt reljeefist ja kaugusest niisutussüsteemidest. Selle idee taustal tasub arvatavasti ka Eestis kaaluda mullakaardilt ja maapinna kõrgusmudelilt saadavate piirkondlike tunnuste kaasamist kirjeldavate tunnustena või mõnede põllukultuuride kohaliku esinemistõenäosuse hinnangute saamiseks.

Cai jt (2018) uurisid süvaõppe abil (*deep neural networks)* maisi ja sojaoa põldude eristamist Illinoisi osariigis USA-s kasutades selleks Landsat-5 TM, Landsat-7 ETM+ ja Landsat-8 OLI pilte ajavahemikust 2000-2015. Kirjeldavateks tunnusteks olid spektraalsed heleduskoefitsiendid kui ka multispektraalsed indeksid (NDVI, GCVI, EVI ja LSWI). Empiiriline andmestik (*USDA's Cropland Data Layer*) oli 13 959 põllu kohta. Kuigi empiirilise andmestiku puhul polnud tegemist päris kohtvaatlustega vaid satelliidipiltidel põhineva klassifikatsiooniga, siis selle täpsuseks hinnatakse siiski üle 95%. See tähendab, et kuna kontrollandmestik, millega süvaõppe abil saadud hinnanguid võrreldi, oli tegelikult juba satelliidipiltidel põhinev hinnang, siis on tulemuste usalduspiirid arvatavasti laiemad, kui autorid järeldustes välja toovad. Uurimisala katsid WRS2 koordinaatsüsteemi järgi kaks satelliidipildi kaadrit, millel oli omavahel ülekate, aastane piltide arv oli 80-90. Kirjeldavad tunnused keskmistati põldude kaupa. Aegread siluti Savitzky–Golay filtriga. Klassifitseerimisalgoritmi rakendati kahe viimase aasta andmetel ja treenimiseks kasutati varasemaid andmeid. Autorid näitasid, et kasvuperioodi alguses on nende kahe põllukultuuri eristamise täpsuseks 70%. 90%-line täpsus saavutatakse, kui kasvuperioodi algusest on möödunud umbes 75 päeva. Arvestades kasvuperioodi alguseks Eestis püsivalt üle +5°C õhutemperatuuri saabumise ajana 20. aprilli, vastaks meil sellele ajale juuni lõpp või juuli algus.

Bargiel (2017) pakub välja uue fenoloogiamustrite klassifitseerimise meetodi (*phenological sequence patterns, PSP-classification)*, mis põhineb kuueastmelisel fenoloogilise skaala väärtuste klassifitseerimisel näidiste järgi. PSP meetod eeldab kasvuperioodi katvat tihedat kirjeldavate tunnuste aegrida, mida võimaldavad ainult radarmõõtmised. Iga põllukultuuri näidiste järgi arvutatakse sihtpikslitele kuue fenoloogilise seisundi tõenäosused. Seejärel otsitakse läbi kuue fenoloogiaseisundi hinnangu kõige suurema tõenäosusega põllukultuuri. Meetod sisaldab iga sihtpiksli ümbruses kaheksa naaberpiksliga arvestamist peale klassi kuuluvuse tõenäosuste arvutamist. Kasutati 99 kahepolaarset (VV & VH) Sentinel-1A ülesvõtet ajavahemikust 13. oktoober 2014 kuni 8. oktoober 2016. Rohkem kui kahesajal põllul tehtud uuring näitas, et PSP meetodiga on võimalik üheksa põllukultuuri (rohumaa, kartul, mais, raps, suhkrupeet, oder, talioder, rukis, talinisu) eristamisel pikslite kaupa saada sõltuvalt kultuurist (40)70-95 % täpsus. Kaera ja odra äratundmine oli väiksema täpsusega ja neid pakuti teisteks teraviljadeks (rukis, talinisu, talioder).

Griffiths jt (2019) kasutasid Sentinel-2 MSI ja Landsat-8 OLI ühildatud piltide andmebaasi. Piltide aegrida siluti ja täideti andmeteta augud, et saada tervet Saksamaad kattev 10-päevase ja kuuajase intervalliga ning aastaaja (kevad, suvi sügis, talv) kohta iseloomulikud kogu riiki katvad satelliidipiltide kihid. Aegreas kasutati 2015. aasta kasvuperioodi lõpu ja 2016. aasta kasvuperioodi pilte, et eelmise aasta piltide abil eristada paremini talivilju.

Klassifitseerimismeetodiks oli otsustuspuude mets. Kirjeldavateks tunnusteks olid satelliidipiltidelt spektraalsed heledused. Klassifitseerimisskeemis olid järgmised klassid (rohumaa, taliviljad, mais, taliraps, suviteraviljad, suhkrupeet, kartul, viinapuuaiad, segametsad, okasmetsad, ehitised ja vaba pinnaga vesi). Näidiste andmed pärinesid 2006. aastast kolmelt liidumaalt Mecklenburg-Vorpommern (MV), Brandenburg (BB) and ja Bayern (BY). Iga klassi jaoks kasutati 1500 näidist treeninguks ja 1000 näidist valideerimiseks. Näidiste suurus oli pindala järgi üle ühe hektari. Üldine klassifitseerimistäpsus oli 10-päevaste komposiitide kasutamisel veidi üle 80%, kuuajaliste ja aastaaja keskmiste komposiitide kasutamisel täpsus kahanes.

Inglada jt. (2015) uurisid 12 suurel testalal, mis asusid erinevate kasvutingimustega geograafilistes piirkondades, põllukultuuride kaardi koostamist kasutades SPOT-4 HRV-IR ja Landsat-8 OLI piltide aegridu. Piksli naabruse info kasutamiseks testiti segmenteerimist, kuid leiti, et algoritmi parameetrite stabiilsete lähendite saamine oli keeruline ja eelistati servaefekti arvestavaid ruumilisi filtreid. Aegridade koostamiseks kasutati kvaliteetsete vaatluste vahel interpoolimist. Kirjeldavate tunnustena kasutati atmosfäärialuseid heleduskoefitsiente, NDVI ja NDWI indeksit ja piksli üldist heleduse väärtust (mitmemõõtmelise Eukleidilise kauguse põhimõttel). Närvivõrke ei kasutatud. Parimaks algoritmiks osutus otsustuspuude mets. Enamikel testaladel, kus põllukultuurid olid lapiti selgelt eristatavad ja põldudel ei kasvanud puittaimestikku, saavutati üle 80 %-line täpsus.

Matton jt. (2015) kasutasid samu testalasid nagu Inglada jt. (2015) ning uurisid võimalusi automatiseeritud ja väheste näidiste arvuga töötava põllukultuuride tuvastamise süsteemi loomiseks. Kasutati SPOT-4 HRV-IR kanalitele vastavaid spektraalseid heledusi ka Landsat-8 OLI piltidelt. Aegridade silumist ei kasutatud. Testiti kahte töötlusskeemi, mille aluseks oli: 1) k-means klasterdamine (100 klastrit) ja saadud klastrite tuvastamine ja 2) põllukultuuride teadaolevat pindalalist jaotust arvestav suurima tõepära (*maximum likelihood*) meetod. Kirjeldavate tunnustena kasutati multispektraalsetelt satelliidipiltidelt pikslite kaupa ja pildisegmentide kaupa arvutatud spektraalsete heleduste väärtuseid NDVI indeksi aegrea kuju kirjeldavatel olulistel hetkedel. Nendeks hetkedeks olid NDVI indeksi väärtuse maksimum, miinimum, suurim tõus, väikseim tõus, mida ei seotud nende ilmnemise ajaga. Andmetöötluseks valiti lõpuks välja punase ja NIR kanali heledus NDVI miinimumi hetkel ning heledus rohelises, punases ja NIR kanalis NDVI maksimumväärtuse hetkel. SWIR kanal ei osutunud informatiivseks. Suurima tõepära meetod ehk õpetava valimiga klassifitseerimisskeem osutus paremaks kui ainult *k-means* klasterdamisel põhinev lahendus.

Valero jt (2016) uuringus, mille eesmärgiks oli dünaamilise põldude maski (*boole* mask*)* koostamine, kasutati samu testalasid, mis Matton jt. (2015) ja Inglada jt. (2015) töös. Maski koostamist alustatakse vegetatsiooniperioodi alguses. Maski uuendatakse iga kord, kui tekib kasutatav ülesvõte. Vegetatsiooniperioodi lõpuks identifitseerib mask need põllud, kuhu vähemalt üks põllukultuur on sel aastal istutatud või külvatud. Rohumaad, mitmeaastased põllukultuurid ja puittaimestikuga alad jäetakse välja. Kirjeldavate tunnustena kasutatakse NDVI, NDWI ja üldise heleduse aegridu, kusjuures NDWI ja heledus on abitunnused ning

peamine analüüs tehakse NDVI põhjal. Täienevate aegridade kirjeldamiseks kasutatakse seitsetteist ajaga seotud muutujat, mille hulgas on NDVI suurima ja vähima väärtuse tuvastatud kuupäevad, suurim tõus jne. Eraldi vaadeldakse pikslipõhist klassifitseerimist ning objektipõhist klassifitseerimist, kus põldude piirid on mingil viisil teada (näiteks segmenteerimise põhjal). Klassifitseerimiseks kasutati otsustuspuude metsa. Binaarse maski täpsuseks kasvuperioodi lõpul saadi üldiselt 90% kõikidel testaladel.

Vuolo jt. (2018) uurisid, kui palju annab põllukultuuride tuvastamise täpsusele juurde mitme kuupäeva Sentinel-2 MSI ülesvõtete kasutamine. Uuritavaks perioodiks oli 2016-2017. Eristati põlde, kus kasvasid porgandid, mais, sibulad, kartul, kõrvits, soja, suhkrupeet, päevalill ja taliviljad. Pikslite klassifitseerimiseks kasutati otsustuspuude metsa, mille hüperparameetreid ei optimeeritud peale selle, et harunemiseks kasutatavate tunnuste arvuks võeti ruutjuur kirjeldavate tunnuste arvust. Iga järgneva pildi lisandumisel lisandus seega ka uus komplekt kirjeldavaid tunnuseid, milleks olid Sentinel-2 MSI ülesvõtetelt saadud maapinnalähedased spektraalsed peegeldustegurid. Selgus, et ühe pildi kasutamisel, kui kasvuperiood algas, oli klassifitseerimise täpsus ca 50%. Klassifitseerimise täpsus kasvas koos piltide lisandumisega ja suve keskel juulis saavutati suurim 91-95%-line üldine täpsus. Augustikuised pildid täpsust enam ei suurendanud.

Jakimow jt. (2018) võrdlesid suurt hulka multispektraalseid vegetatsiooniindekseid (NDVI, EVI, EVI2, SAVI, nüü, GEMI, NDMI, NBR, NBR2, MIRBI, BAIM), muuhulgas tuttmütsiindekseid (TCB, TCG, TCW) ja lisaks ülesvõtete ajaga seotud tunnused (DTIME-1, DTIME+1, DOY) Brasiilia karjamaade majandamise seireks. Ajaga seotud tunnusteks oli päevade arv kahe kasutatava ülesvõtte vahel või päeva number aastas. Uurimisalaks oli üks WRS-2 kaadrite jaotusskeemi kaardileht (~180x180 km ala). Kirjeldavad tunnused saadi Landsat-8 OLI ülesvõtete seeriatest. Klassifikatsiooni moodustasid põletatud karjamaa, pinnasetöötlusega karjamaa, üle põletatud võsa, karjamaa, võsa, infrastruktuur. Tulemused näitasid, et peale tuttmütsiindeksite komplektis oleva heleduse ei olnud ühelgi teisel indeksil erilisi eeliseid. Spektraalsel infol põhinevad tunnused olid olulisemad ajal põhinevatest tunnustest. Viimased varieerusid rohkem ja seetõttu oli ka nende normeeritud olulisus väiksem. Kuna uuritavas protsessis oli tegemist maa-ala ülepõletamisega, mille tulemusena tekib väga selgelt eristuv, nähtavas ja NIR spektriosas madala ja SWIR spektriosas pigem kõrge peegeldusteguriga (sõltuvalt pinna niiskusest) signatuur, siis on ka arusaadav, miks üldine heledus oli veidi suurema informatiivsusega, kui teised kasutatud kirjeldavad tunnused.

Copernicuse Sentineli satelliidipiltide kasutamine Euroopa Liidu põllumajandustoetuste järelevalveks (sh kultuuride määramiseks) on suure tähtsusega rakendus. Viimastel aastatel on teema olulisusest aru saadud ja tehtud mitmeid rakendusuuringuid.

Estrada jt. (2017) rakendasid Sentinel-2 NDVI aegridu, et tuvastada erineva niisutusrežiimiga põlde, millele on Hispaanias erinev toetusskeem. Lisaks uuringus rakendatud NDVI indeksile, soovitasid autorid põllukultuuride kirjeldamiseks kasutada ka NDWI (*i k Normalized Difference Water Index*) ja SAVI (*i k Soil Adjusted Vegetation Index*) indekseid. Sõltuvalt eristatavate klasside arvust saavutati klassifitseerimise kogutäpsuseks 85-95%.

Filizzola jt. (2018) uuris Landsat-5 TM ja Landsat-7 ETM+ NDVI ajaliste käikute põhjal põllumajandusmaa kasutusotstarbe klassifitseerimise võimalusi ja saavutas CORINE 2012 andmekogu (maakattekaardi) suhtes 82% klassifitseerimise kogutäpsuse.

Khaliq jt. (2018) võrdlesid Sentinel-2 MSI aegridade ja üksikpiltide põhist põllukultuuride tuvastamist otsustuspuude metsa meetodil. Eristati ainult kuut erinevat põllukultuuride klassi. Leiti, et aegrea kasutamine suurendab täpsust oluliselt – klassifitseerimise kogutäpsus 91% vs 65-83% erinevate üksikpiltide põhjal.

Schmedtmann jt. (2015) rakendasid Landsat-7 ETM+ piltide aegrida ja SVM meetodit. Portugali testalal 12 põllukultuuri klassiga tehtud töös saavutati ainult 68% klassifitseerimise kogutäpsus, sest paljude kultuuride spektraalsed signatuurid olid väga sarnased. Pakuti välja lähenemine täpsuse tõstmiseks soovitus jätta madalama usaldusväärsusega põllud treeningvalimist välja.

Sitokonstantinou jt. (2018) katsetas objekti- ehk põllupõhist kultuuride tuvastamist Sentinel-2 aegridadel. Võrdlevalt rakendati nii SVM kui ka otsustuspuude metsa meetodit Hispaania testalal, eristati 9 klassi. SVMi abil saadi veidi täpsemad tulemused kui otsustuspuude metsaga. Sisendandmete tunnuskomplektist osutusid kõige kasulikumaks lähisinfrapuna kanal (8), punase serv (5-7), lühilaine infrapuna (11 ja 12), NDVI ja PSRI (*i k Plant Senescence Reflectance Index*). Ajamomentidest olid täpsuse lisamisel kõige kasulikumad mai ja juuli pildid.

Veloso jt. (2017) uurisid nisu, maisi, rapsi, sojaoa ja päevalille põldude satelliitseire parameetrite ajalisi käike Lõuna-Prantsusmaa testalal. Uuritavateks parameetriteks olid Sentinel-2 MSI NDVI, Sentinel-1 VH/VV tagasihajumise suhe ja VH ning VV kalibreeritud tagasihajumised eraldi. Käikude uurimise järgi püüti leida üldistusi kultuuride ja nende erinevate fenoloogiliste faaside „silma järgi“ eristamiseks.

2 Kirjanduse analüüsi kokkuvõte

Põllukultuuride tuvastamiseks kasutatakse nii pikslipõhiseid lahendusi kui ka homogeensete segmentide analüüsi. Valdavalt saadakse kirjeldavad tunnused (*feature variables*) multispektraalsete piltide aegreast, mikrolaineala andmete lisamine parandab üldist täpsust suhteliselt vähe. Samas on välja pakutud ka radarmõõtmiste tiheda aegrea eeliseid kasutav kuueastmelisel fenoloogiaskaalal põhinev meetod, mis andis samuti üsna häid tulemusi, kuid empiiriline andmestik oli siiski pigem tagasihoidlik (kakssada põldu). Multispektraalsete satelliidipiltide pikslite aegrea väärtuseid töödeldakse kas harmooniliste võnkumiste kombinatsioone kasutades või siis mingi näidistele tugineva klassifitseerimismeetodi abil. Enamlevinum neist oli otsustuspuude mets, närvivõrkude kasutamise kohta suurtele aladel oli vähe näiteid. Kirjeldavate tunnustena kasutatakse enamasti NDVI indeksit või siis atmosfäärialuseid heleduskoefitsiente NDVI indeksi aegrea olulise muutusega hetkedel. Operatiivse kaardistamise ülesande puhul on selgunud, et kasvuperioodi alguses on klassifitseerimise täpsus 50%, suurim täpsus (80-95%) saavutatakse suve keskpaigaks ning

augustikuised ülesvõtted olulist infot konkreetse aasta puhul juurde ei anna. Eelneva aasta sügise piltide kasutamine aitab taliviljade tuvastamise täpsust parandada.

Enamus näiteid oli suhteliselt väikese eristatavate põllukultuuride klasside arvuga (reeglina <10). Mida rohkem eristatavaid klasse, seda madalam oli tavaliselt klassifitseerimise kogutäpsus. Eesti näitel 25 kultuuri eristamise soov näib pigem ambitsioonikas ning tuleb olla valmis, et sarnased kultuurid hakkavad omavahel segamini minema ja klasside arvu tuleb täpsuse tõstmise huvides vähendada.

Sõltuvalt rakendatud seadetest olid saavutatud täpsused otsustuspuude metsa, SVMi ja närvivõrkudega suhteliselt sarnased. Närvivõrkude valiku kasuks räägib nende järjest laiem levik, arengupotentsiaal ja tugi vabavaralistes tarkvarapakettides (nt Keras/TensorFlow). Süvaõppe närvivõrkudel on mitmeid edasiarendusi, mis lubavad tulevikus täpsust veelgi tõsta. Nt sidumnärvivõrgud (*i k CNN – Convolutional Neural Networks*) suudavad arvestada objektide suuruse ja kuju, põllusisese muutlikkuse ning ümbruskonnainfoga, mis on seda kasulikum, mida heterogeensem ja detailide rohkem on uuritav ala satelliidipiltidel.

3 Viited

Bargiel, D. 2017. A new method for crop classification combining time series of radar images and crop phenology information. Remote Sensing of Environment 198, 369–383.

Cai, Y., Guan, K., Peng, J., Wang, S., Seifert, C., Wardlow, B., Li, Z. 2018. A high-performance and in-season classification system of field-level crop types using time-series Landsat data and a machine learning approach. Remote Sensing of Environment 210, 35–47.

Conrad, C., Löw, F., Lamers, J.P.A. 2017. Mapping and assessing crop diversity in the irrigated Fergana Valley, Uzbekistan. Applied Geography 86 102-117.

Estrada, J., Sánchez, H., Hernanz, L., Checa, M. & Roman, D., 2017. Enabling the Use of Sentinel-2 and LiDAR Data for Common Agriculture Policy Funds Assignment. ISPRS International Journal of Geo-Information, 6(8), p.255.

Filizzola, C., Corrado, R., Falconieri, A., Faruolo, M., Genzano, N., Lisi, M., Mazzeo, G., Paciello, R., Pergola, N. & Tramutoli, V., 2018. On the use of temporal vegetation indices in support of eligibility controls for EU aids in agriculture. International journal of remote sensing, 39(14), 4572-4598.

Ghazaryan, G., Dubovyk, O., Löw, F., Lavreniuk, M., Andrii Kolotii, A., Schellberg, J., Kussul, N. 2018. A rule-based approach for crop identification using multi-temporal and multi-

sensor phenological metrics, European Journal of Remote Sensing, 51:1, 511-524. DOI: 10.1080/22797254.2018.1455540

Griffiths, P., Nendel, C., Hostert, P. 2019. Intra-annual reflectance composites from Sentinel-2 and Landsat for national-scale crop and land cover mapping. Remote Sensing of Environment 220 (2019) 135–151.

Inglada, J., Arias, M., Tardy, B., Hagolle, O., Valero, S., Morin, D., Dedieu, G., Sepulcre Canto, G., Bontemps, S., Defourny, P., Koetz, B. 2015. Assessment of an operational system for crop type map production using high temporal and spatial resolution satellite optical imagery. Remote Sens. 7, 12356-12379. doi:10.3390/rs70912356

Jakimow, B., Patrick Griffiths, P., van der Linden, S., Hostert, P. 2018. Mapping pasture management in the Brazilian Amazon from dense Landsat time series. Remote Sensing of Environment 205, 453–468.

Kanjir, U., Đurić N., Tatjana Veljanovski, T. 2018. Sentinel-2 Based Temporal Detection of Agricultural Land Use Anomalies in Support of Common Agricultural Policy Monitoring. ISPRS Int. J. Geo-Inf. 7, 405. doi:10.3390/ijgi7100405

Khaliq, A., Peroni, L. & Chiaberge, M., 2018, June. Land cover and crop classification using multitemporal Sentinel-2 images based on crops phenological cycle. In 2018 IEEE Workshop on Environmental, Energy, and Structural Monitoring Systems (EESMS) (pp. 1-5). IEEE.

Matton, N., Sepulcre Canto, G., Waldner, F., Valero, S., Morin, D., Inglada, J., Arias, M., Bontemps, S., Koetz, B., Defourny, P. 2015. An Automated Method for Annual Cropland Mapping along the Season for Various Globally-Distributed Agrosystems Using High Spatial and Temporal Resolution Time Series. Remote Sens. 7, 13208-13232; doi:10.3390/rs71013208

Roy, D.P., Yan, L. 2018. Robust Landsat-based crop time series modelling. Remote Sensing of Environment, https://doi.org/10.1016/j.rse.2018.06.038 (artikkel on avaldamisel).

Schmedtmann, J. & Campagnolo, M., 2015. Reliable crop identification with satellite imagery in the context of common agriculture policy subsidy control. Remote Sensing, 7(7), 9325-9346.

Sitokonstantinou, V., Papoutsis, I., Kontoes, C., Lafarga Arnal, A., Armesto Andrés, A. & Garraza Zurbano, J., 2018. Scalable parcel-based crop identification scheme using Sentinel-2

data time-series for the monitoring of the Common Agricultural Policy. Remote Sensing, 10(6), p.911.

Valero, S., Morin, D., Inglada, J., Sepulcre, G., Arias, M., Hagolle, O., Dedieu, G., Sophie Bontemps, S., Defourny, P., Koetz, B. 2016. Production of a Dynamic Cropland Mask by Processing Remote Sensing Image Series at High Temporal and Spatial Resolutions. Remote Sens. 2016, 8, 55. doi:10.3390/rs8010055

Veloso, A., Mermoz, S., Bouvet, A., Le Toan, T., Planells, M., Dejoux, J.F. & Ceschia, E., 2017. Understanding the temporal behavior of crops using Sentinel-1 and Sentinel-2-like data for agricultural applications. Remote sensing of environment, 199, 415-426.

Vuolo, F., Neuwirth, M., Immitzer, M., Atzberger, C., Ng, W.-T. 2018. How much does multi-temporal Sentinel-2 data improve crop type classification? Int J Appl Earth Obs Geoinformation 72 (2018) 122–130.

4 Põllukultuuride tuvastusmudeli kirjeldus

4.1 Sissejuhatus

Käesolev dokument kirjeldab üldisel tasemel põllukultuuride tuvastusmudeli, eristatavad põllukultuurid ja sisendandmed. Tuvastusmudel ja selle põhjal tulevikus arendatav tarkvara on vajalik põllumajandustoetuste haldamiseks Põllumajanduse Registrite ja Informatsiooni Ametis (PRIAs) ja Eesti põllumajandusmaa kohta riikliku statistika koostamiseks. Vaatlusühikuks on põllud vastavalt PRIA defineeritud Eesti põldude vektorkihile. Iga üksiku piksli kultuuri määramine ja pidevate põllukultuuri kaartide loomine pole käesoleva projekti raames vajalik.

Dokument on mõeldud eelkõige projektimeeskonnale siseseks kasutamiseks ning on oluliseks sisendiks järgnevate lähteandmete ettevalmistus- ja mudeliarendustööde juures (tulemid D4.2-D4.8). Detailne tehniline andmemudel on toodud lisas 1.

Vastav uurimis- ja arendustöö on tehtud ja käesolev dokument on valminud projekti RITA1/02-52 „Kaugseire andmete kasutuselevõtt avalike teenuste väljatöötamisel ja arendamisel“ raames.

4.2 Eristatavad põllukultuurid

Esialgne eristatavate põllukultuuride (kultuurigruppide) nimekiri on:

1. Aedmaasikas
2. Astelpaju
3. Heintaimed, kõrrelised
4. Heintaimed, liblikõielised
5. Kaer
6. Kanep
7. Kartul
8. Mais
9. Mustkesa
10. Peakapsas
11. Põldhernes
12. Põlduba
13. Porgand
14. Punapeet
15. Rukis
16. Suvinisu ja speltanisu
17. Suvioder
18. Suviraps ja -rüps
19. Suvitritikale
20. Sööti jäetud maa

21. Talinisu
22. Talioder
23. Taliraps ja -rüps
24. Talitritikale
25. Tatar

NB! Tegelikult, andmebaasi läks nii (tähestiku järjekorras):

Rita_kood Rita_grupp

1. Aedmaasikas
2. Astelpaju
3. Heintaimed, kõrrelised
4. Heintaimed, liblikõielised
5. Kaer
6. Kanep
7. Kartul
8. Mais
9. Mustkesa
10. Muu
11. Peakapsas
12. Porgand
13. Punapeet
14. Põldhernes
15. Põlduba
16. Rukis
17. Suvinisu ja speltanisu
18. Suvioder
19. Suviraps ja -rüps
20. Suvitritikale
21. Sööti jäetud maa
22. Talinisu
23. Talioder
24. Taliraps ja -rüps
25. Talitritikale
26. Tatar

Täpne vastavus PRIA olemasoleva kultuuride nimekirja ja põllukultuuride koodidega on toodud projekti *SharePoint*i kataloogis failis „*kultuuride loend.xlsx*“.

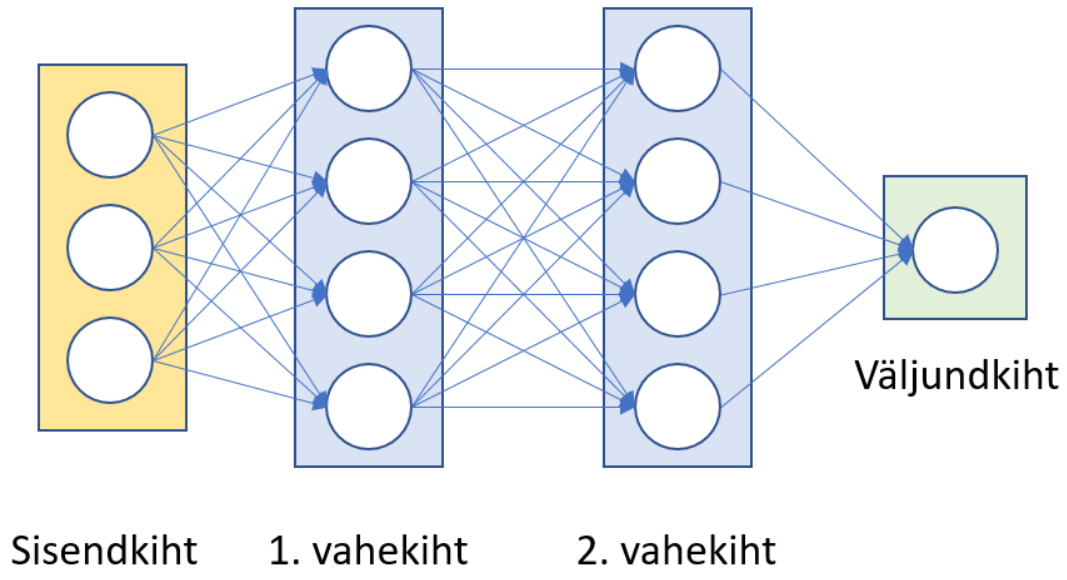
Kui uurimistöö käigus selgub, et otstarbekas on eristatavate kultuuride nimekirja muuta (osa gruppe liita või vastupidi eristust suurendada detailsemaid gruppe juurde tehes), siis nimekirja uuendatakse vastavalt. Võib juhtuda ka, et külvipinna järgi vähem levinud kultuuride (nt peakapsas, porgand, punapeet või aedmaasikas) ei tule kokku piisavat arvu näidispõlde, mille

põhjal metoodikat arendada ning seetõttu tuleb eristatavate kultuurigruppide nimekirja koondada.

4.3 Tuvastusmudeli kontseptsioon ja arhitektuur

Tuvastusmudeliks on süvaõppe mudel (varasemalt tuntud ka kui tehislikud närvivõrgud), mis töötab satelliidipiltidelt arvutatud parameetrite väärtusi kasutades põldude aegridadel. Süvaõppe mudelid on viimase kümnendi jooksul kiiresti arenenud ja mitmesuguste sisend-väljund klassifitseerimisülesannete juures ennast tõestanud kui ühed täpsemad lahendused. Süvaõppe mudeli kasuks räägib ka asjaolu, et saavutatavad täpsused on seda paremad, mida suuremad on õpetusandmetes näidistena kasutatavad andmekomplektid. Põllukultuuride tuvastamisel on Eestis kasutada igal aastal üle 100 000 põllu andmed.

Kuna ruumi-, aja- ja järjestusinfo ei ole praeguse ülesande puhul esmatähtsad, pole vaja kasutada keerukaid sidum- ja rekursiivseid närvivõrke *(i k CNN – convolutional neural networks ja RNN – recurrent neural networks*). Välja pakutud arhitektuur on suhteliselt lihtne ja koosneb sisendkihist ja *n* vahekihist, mis on omavahel täielikult ühendatud (*i k fully connected layers*). Viimasel kihil kasutatakse aktivatsioonifunktsioonina *softmax*-i, mis annab igale põllule tõenäosused kõigi võimalike nimekirjas olevate kultuuride kohta (25) nii, et tõenäosuste summa on 100%. Mudeli täpne kihtide ja neuronite arv igal kihil selgub arendustöö käigus iteratiivselt. Täielikult ühendatud närvivõrgu põhimõtteskeemi vt Joonis 1. Joonisel toodud närvivõrgus on sisendkihil kolm neuronit ja kaks vahekihti, kus mõlemal on neli neuronit.



Joonis 1. Närvivõrgu põhimõtteskeem.

4.4 Lähteandmed ja tunnuskomplekt

Sisendkihi tunnuskomplekt tuletatakse kolmest allikast:

1. Sentinel-1 IW režiimi ja Sentinel-2 MSI satelliidipildid iga aasta 1. aprillist kuni 31. oktoobrini,
2. muud olemasolevad ruumiandmed,
3. ilmaandmed.

Sentinel-1 parameetritest tuleks kaasata tunnuskomplekti:

- VV-kanali 6-päeva koherentsus.
- VH-kanali 6-päeva koherentsus.
- VV-kanali 12-päeva koherentsus
- VH-kanali 12-päeva koherentsus
- VV-kanali tagasihajumine.
- VH-kanali tagasihajumine.
- VH/VV suhe.

Sentinel-2 MSI parameetritest tuleks kaasata tunnuskomplekti:

- Kõik 10 ja 20 m lahutusega üksik-kanalid (B2, B3, B4, B8, B5, B6, B7, B8, B8a, B11 ja B12).
- Järgmised vegetatsiooniindeksid: TC_Wetness, TC_Vegetation, TC_Brightness, Misra_Yellow_Vegetation, PSRI, WRI, NDWI, NDVIre, NDVI.

Vegetatsiooniindeksid arvutatakse järgmiste valemitega:

NDVI=(B8-B4)/(B8+B4) (1)

NDVIre=(B8-B6)/(B8+B6) (2)

NDWI=(B8-B11)/(B8+B11) (3)

WRI=(B2+B3)/(B8+B11) (4)

PSRI=(B4-B2)/B6 (5)

Misra_Yellow_Vegetation=0.723*B3-0.597*B4+0.206*B6-0.278*B8 (6)

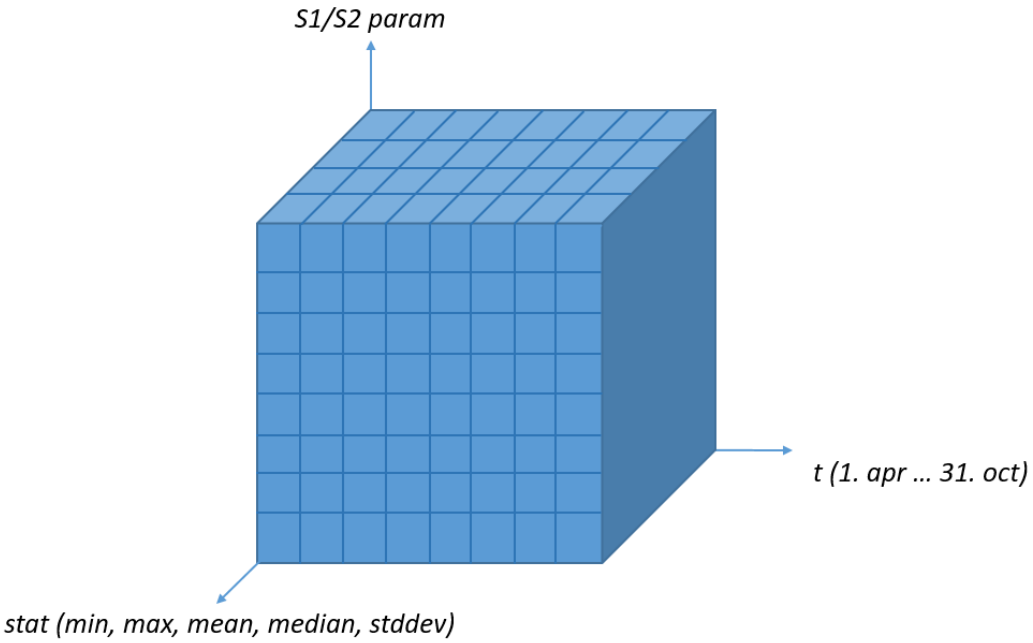
TC_Brightness=0.304*B2+0.279*B3+0.474*B4+0.558*B8+0.508*B11+0.186*B12 (7)

TC_Vegetation=−0.285*B2-0.244*B3-0.544*B4+0.724*B8+0.0840*B11-0.180*B12 (8)

TC_Wetness=0.151*B2+0.197*B3+0.328*B4+0.341*B8-0.711*B11-0.457*B12 (9)

Sentinel-2 MSI aegridade puhul peab arvestama, et olenevalt pilvkattest võib olla põldude lõikes kasutatavate andmepunktide arv vegetatsiooni perioodi kohta väga erinev. Seetõttu võib olla vajalik täiendavate piirangute kehtestamine aegridade täielikkusele (nt. vähemalt 10

andmepunkti vegetatsiooniperioodi kohta ja vähemalt üks pilt igast kuust vahemikus aprillist augustini), et vältida mudeli õpetamist puudulike sisendandmetega.



Joonis 2. Satelliitseire parameetritest moodustatud tunnuskomplekt 3-mõõtmelise tensorina.

Kokkuvõttes moodustub Sentinel-1 ja -2 parameetritest tunnuskomplekt, mida võib vaadelda kolmemõõtmelise tensorina (Joonis 2). Esimeses mõõtmes muutub ajateljel andmete kuupäev 1. aprillist 31. oktoobrini, teisel teljel S1 või S2 parameeter (nt S1 VV-kanali 6-päeva koherentsus, S2 NDVI, jne) ning kolmandal teljel antud parameetri põhjal arvutatud põllupõhise pikslite komplekti statistik (miinimum, maksimum, keskväärtus, mediaan ja standardhälve).

Olemasolevatest ruumiandmetest tuleks tunnuskomplekti kaasata:

- Mullatüüp – pindala alusel suurima osakaaluga põllul esinev mullatüüp.
- Normeeritud (0 ja 1 vahel) asukohakoordinaadid geograafiliste iseärasuste arvestamiseks.

Mullatüüp on Eesti mullastiku kaardi järgi tuletatud üldistatud mullatüüp. Kokku eristatakse põllukultuuride tuvastamisel 23 erinevat mullatüüpi. Täpsemalt on eristatavate mullatüüpide tuletamine Eesti mullastiku kaardist kirjeldatud lisas 2.

Ilmaandmetest tuleks tunnuskomplekti kaasata:

- Päeva keskmine temperatuur (^{o}C).
- Sademesummad (mm).

Päeva keskmine temperatuur tuleks esitada rasterandmetena 1 km ruumilise ja 1 päevase ajalise lahutusega.

Sademesummad tuleks esitada rasterandmetena 1 km ruumilise ja 3 h ajalise lahutusega. Summad võiks olla ajastatud kahes variandis 02:00-05:00 UTC, 05:00-08:00 UTC, jne ning 13:00-16:00 UTC, 16:00-19:00 UTC, jne, sest Sentinel-1 ülelennud Eestist on vastavalt umbes 4:45 UTC ja 16:00 UTC, aga kui seda on väga tülikas teha, siis võib ka lihtsalt ühes faasis kõik hoida: 00:00-03:00 UTC, 03:00-06:00 UTC, 06:00-09:00 UTC jne.

Aegread peavad olema arvutatud igal aastal alates 1. aprillist kuni 31. oktoobrini. Kõik tunnuskomplekti andmepunktid tuleb enne mudelile ette andmist arvutada ühtlaselt 1 päevase sammuga, kasutades vaheväärtuste arvutamiseks lineaarset interpoleerimist. Taliviljade täpsema eristuse huvides võib kaaluda ka eelmise sügise aegridade lisamist tunnuskomplekti.

Kõigi Sentinel-1 ja -2 parameetrite puhul tuleb arvutada iga põllugeomeetria kohta järgmised statistikud: 1) miinimum, 2) maksimum, 3) mediaan, 4) aritmeetiline keskmine, 5) standardhälve.

Statistikute arvutamisel tuleb kasutada võimalikult puhast põllukultuuri pikslite komplekti. Selleks kasutatakse vastavaid PRIA vektorandmeid, et eristada põllugeomeetriast rohealad (hekid, puudesalud, kivihunnikud, kraavid jms) ja kaasata arvutustesse ainult põllukultuuri esindavad pikslid. Väga soovitav on kasutada ka põllugeomeetriate sisse poole puhverdamist, et vältida segupikslite kasutamist statistikute arvutamisel.

Detaile andmemudeli kirjeldus on toodud dokumendi lõpus osas 0.

4.5 Mudeli väljundi kirjeldus

Mudeli väljundiks on *softmax*-vektor, mis annab iga põllu kohta tõenäosused kõigi nimekirjas olevate kultuuride kohta. Mudeli väljundi põhjal arvutatakse veamaatriksid ja muud vigade diagnoosimiseks ja silumiseks vajalikud arvnäitajad ja graafikud.

Mudeli tulemused arvutatakse kolmel korral aastas:

- Juuni lõpu seisuga.
- Augusti alguse seisuga.
- Hooaja lõpu (oktoobri lõpu) seisuga.

Täpsete ajamomentide kuupäevade valik selgub projekti käigus – uurime, kuidas analüüsi kaasatud aegridade pikkused tulemuste usaldusväärsust mõjutavad. Lähtume põhimõttest - nii vara kui võimalik, aga siiski piisavalt hilja, et tulemuste usaldusväärsus oleks kõrge. Viimane sõna kuupäevade valikul on uuringu tellijal, s.o. PRIAl.

4.6 Arendustöö metoodika

Arendustöö metoodika on iteratiivne. Peale igat mudeli sobitamist analüüsitakse vigu ja koostatakse nimekiri täiustusideedest järgmise iteratsiooni jaoks.

Õpetusandmetena kasutatakse PRIA taotluste andmeid, kus on märgitud iga põllu kohta, mis kultuur seal kasvama peaks. PRIA andmetel on enamasti taotlusel märgitud kultuur õige ja vigu on alla 5%. Järelikult on nad kasutatavad mudeli õpetusandmetena pärast tunnuskomplekti eksete analüüsi – näidised, mille tunnuskomplekt erineb klassi keskmisest väga palju (nt. 1,5 standardhälbe võrra) on tõenäoliselt teine kultuur ja seda õpetusandmetena ei kasutata.

Täiendavate õpetus- ja testandmetena kasutatakse ka PRIA inspektorite, RITA projekti välitööde ja droonimõõtmiste andmeid. 2018 aastast oli PRIA inspektorite andmeid rohkel kui 4000 põllu kohta, kus on välitöödega tõestatult teada, mis kultuur igal põllul kasvas. Kuidas kombineeritakse eksetest puhastatud taotluste, PRIA inspektorite, RITA välitööde ja droonimõõtmiste andmeid pole veel praeguseks määratud ja see otsustatakse töö käigus peale põhjalikku andmetega tutvumist.

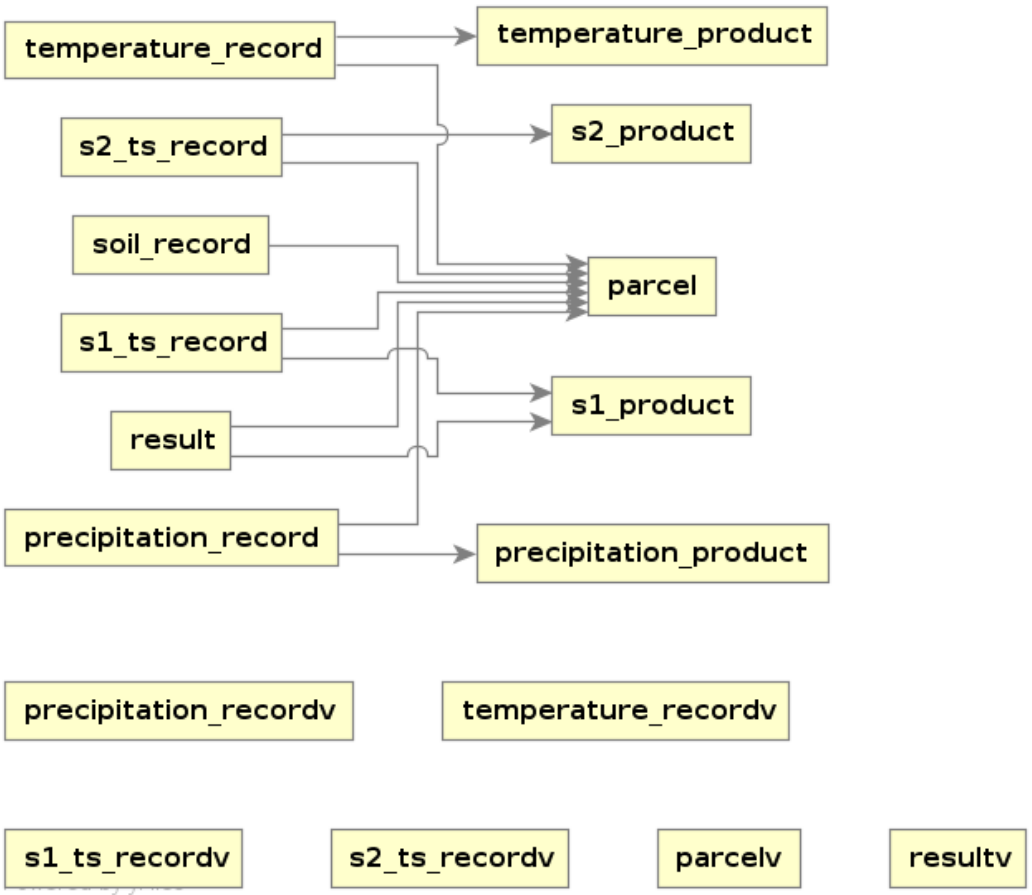
Järgides süvaõppe välja kujunenud praktikaid jagatakse õpetusandmed treening-, valideerimis-, ja testkoguks. Treeningkogu järgi toimub mudeli sobitamine, valideerimiskogu järgi õige mudeli parameetrite komplekti valimine ja isoleeritud testkogu järgi täpsuse hindamine. Kuna andmekogu on suur (>100 000 põldu), siis valideerimis- ja testkogu osakaalud ei pea olema väga suured. Soovitav on jagada õpetusandmed treening-, valideerimis-, ja testkoguks 90%/5%/5% osades.

Lisa 1 - Andmemudeli kirjeldus

Andmemudelis on kolm skeemat:

- public – hallatakse tabeleid, millel ei ole raster tüüpi veerge;
- precipitation – hallatakse sajusummade rastertabeleid;
- temperature – hallatakse temperatuurisummade rastertabeleid.

Vaadetest ja tabelite omavahelistest seostest annab ülevaate Joonis 3. Vaateid kasutatakse selleks, et hoida salvestusruumi optimaalsena, lihtsustada andmete kasutust ja vähendada väliste liideste sõltuvust tabelites tehtavatest andmemudeli muudatustest.



Joonis 3. Public skeema tabelite omavahelised seosed ja vaadete nimetused. v-lõpulised nimetused tähistavad andmebaasi vaateid. Noole ots tähistab seda tabelit, millest välisvõti pärineb.

Vaadete ja tabelite veergude informatsioon koos seoste ülevaatega on illustreeritud järgmisel joonisel:

s2_ts_recordv **parcelv_narrow**

s1_ts_recordv **parcelv_wide**

resultv_narrow **geometry_columns**

resultv_wide **geography_columns**

precipitation_recordv

temperature_recordv

soil_recordv

spatial_ref_sys

srid	INTEGER
auth_name	CHARACTER VARYING(256)
auth_srid	INTEGER
srtext	CHARACTER VARYING(2048)
proj4text	CHARACTER VARYING(2048)

result

id	SERIAL
parcel_id	INTEGER
s1_product_id	INTEGER
estimation_1_aedmaasikas	REAL
estimation_2_heintaimed_korrelised	REAL
estimation_3_liblikoieliste_segud_alla_80	REAL
estimation_4_liblikoielised_yle_80	REAL
estimation_5_kanep	REAL
estimation_6_kartul	REAL
estimation_7_poldhernes	REAL
estimation_8_polduba	REAL
estimation_9_muu_kaunvili	REAL
estimation_10_mais	REAL
estimation_11_astelpaju	REAL
estimation_12_marjapoosad_viljapuud	REAL
estimation_13_mustkesa	REAL
estimation_14_peakapsas	REAL
estimation_15_porgand	REAL
estimation_16_punapeet	REAL
estimation_17_suviraps_ryps	REAL
estimation_18_suvinisu_speltanisu	REAL
estimation_19_suvioder	REAL
estimation_20_kaer	REAL
estimation_21_tatar	REAL
estimation_22_taliraps_ryps	REAL
estimation_23_rukis	REAL
estimation_24_talinisu	REAL
estimation_25_talioder	REAL
estimation_26_talitritikale	REAL
estimation_27_muu_teravili	REAL
estimation_28_muu	REAL
estimation_101_aedmaasikas	REAL
estimation_102_heintaimed_korrelised	REAL
estimation_103_heintaimed_liblikoielised	REAL
estimation_104_kanep	REAL
estimation_105_kartul	REAL
estimation_106_kaunviljad	REAL
estimation_107_mais	REAL
estimation_108_marjapoosad_viljapuud_astelpaju	REAL
estimation_109_mustkesa	REAL
estimation_110_peakapsas	REAL
estimation_111_ryhvelkultuurid	REAL
estimation_112_suviraps_ryps	REAL
estimation_113_suviteraviljad	REAL
estimation_114_taliraps_ryps	REAL
estimation_115_taliteraviljad	REAL
estimation_116_muu	REAL
created_at	TIMESTAMP(6) WITH TIME ZONE
updated_at	TIMESTAMP(6) WITH TIME ZONE

s1_product

id	SERIAL
start_date	TIMESTAMP(6) WITH TIME ZONE
stop_date	TIMESTAMP(6) WITH TIME ZONE
filename	TEXT
relative_orbit_number	INTEGER
created_at	TIMESTAMP(6) WITH TIME ZONE
updated_at	TIMESTAMP(6) WITH TIME ZONE

result_s1_product_id_fkey

s2_ts_record

id	SERIAL
parcel_id	INTEGER
s2_product_id	INTEGER
b2_mean	REAL
b3_mean	REAL
b4_mean	REAL
b5_mean	REAL
b6_mean	REAL
b7_mean	REAL
b8_mean	REAL
b8a_mean	REAL
b11_mean	REAL
b12_mean	REAL
ndvi_mean	REAL
ndwi_mean	REAL
ndvire_mean	REAL
tc_wetness_mean	REAL
tc_vegetation_mean	REAL
tc_brightness_mean	REAL
misra_yellow_vegetation_mean	REAL
psri_mean	REAL
wri_mean	REAL
b2_median	REAL
b3_median	REAL
b4_median	REAL
b5_median	REAL
b6_median	REAL
b7_median	REAL
b8_median	REAL
b8a_median	REAL
b11_median	REAL
b12_median	REAL
ndvi_median	REAL
ndwi_median	REAL
ndvire_median	REAL
tc_wetness_median	REAL
tc_vegetation_median	REAL
tc_brightness_median	REAL
misra_yellow_vegetation_median	REAL
psri_median	REAL
wri_median	REAL
b2_min	REAL
b3_min	REAL
b4_min	REAL
b5_min	REAL
b6_min	REAL
b7_min	REAL
b8_min	REAL
b8a_min	REAL
b11_min	REAL
b12_min	REAL
ndvi_min	REAL
ndwi_min	REAL
ndvire_min	REAL
tc_wetness_min	REAL
tc_vegetation_min	REAL
tc_brightness_min	REAL
misra_yellow_vegetation_min	REAL
psri_min	REAL
wri_min	REAL
b2_max	REAL
b3_max	REAL
b4_max	REAL
b5_max	REAL
b6_max	REAL
b7_max	REAL
b8_max	REAL
b8a_max	REAL
b11_max	REAL
b12_max	REAL
ndvi_max	REAL
ndwi_max	REAL
ndvire_max	REAL
tc_wetness_max	REAL
tc_vegetation_max	REAL
tc_brightness_max	REAL
misra_yellow_vegetation_max	REAL
psri_max	REAL
wri_max	REAL
b2_stdv	REAL
b3_stdv	REAL
b4_stdv	REAL
b5_stdv	REAL
b6_stdv	REAL
b7_stdv	REAL
b8_stdv	REAL
b8a_stdv	REAL
b11_stdv	REAL
b12_stdv	REAL
ndvi_stdv	REAL
ndwi_stdv	REAL
ndvire_stdv	REAL
tc_wetness_stdv	REAL
tc_vegetation_stdv	REAL
tc_brightness_stdv	REAL
misra_yellow_vegetation_stdv	REAL
psri_stdv	REAL
wri_stdv	REAL
created_at	TIMESTAMP(6) WITH TIME ZONE
updated_at	TIMESTAMP(6) WITH TIME ZONE

s1_ts_record_s1_product_id_fkey

precipitation_product

id	SERIAL
start_date	TIMESTAMP(6) WITH TIME ZONE
stop_date	TIMESTAMP(6) WITH TIME ZONE
filename	TEXT
created_at	TIMESTAMP(6) WITH TIME ZONE
updated_at	TIMESTAMP(6) WITH TIME ZONE

precipitation_record_precipitation_product_id_fkey

precipitation_record

id	SERIAL
parcel_id	INTEGER
precipitation_product_id	INTEGER
precipitation_sum_mean	REAL
created_at	TIMESTAMP(6) WITH TIME ZONE
updated_at	TIMESTAMP(6) WITH TIME ZONE

precipitation_record_parcel_id_fkey

result_parcel_id_fkey

s1_ts_record

id	SERIAL
parcel_id	INTEGER
s1_product_id	INTEGER
cohvh_mean	REAL
cohvv_mean	REAL
s0vh_mean	REAL
s0vv_mean	REAL
vhvv_mean	REAL
cohvh_median	REAL
cohvv_median	REAL
s0vh_median	REAL
s0vv_median	REAL
vhvv_median	REAL
cohvh_min	REAL
cohvv_min	REAL
s0vh_min	REAL
s0vv_min	REAL
vhvv_min	REAL
cohvh_max	REAL
cohvv_max	REAL
s0vh_max	REAL
s0vv_max	REAL
vhvv_max	REAL
cohvh_stdv	REAL
cohvv_stdv	REAL
s0vh_stdv	REAL
s0vv_stdv	REAL
vhvv_stdv	REAL
created_at	TIMESTAMP(6) WITH TIME ZONE
updated_at	TIMESTAMP(6) WITH TIME ZONE

parcel

id	SERIAL
f_id	TEXT
x_norm_loc	DOUBLE PRECISION
y_norm_loc	DOUBLE PRECISION
datasource	TEXT
created_at	TIMESTAMP(6) WITH TIME ZONE
updated_at	TIMESTAMP(6) WITH TIME ZONE
geometry_valid	BOOLEAN
geom	geometry

parcel_crop

id	SERIAL
parcel_id	INTEGER
crop_id	SMALLINT
created_at	TIMESTAMP(6) WITH TIME ZONE
updated_at	TIMESTAMP(6) WITH TIME ZONE
clusteranalyses_verified	BOOLEAN

parcel_crop_parcel_id_fkey

s1_ts_record_parcel_id_fkey

parcel_crop_crop_id_fkey

crop

id	SMALLSERIAL
code	TEXT
name	TEXT
type	TEXT
created_at	TIMESTAMP(6) WITH TIME ZONE
updated_at	TIMESTAMP(6) WITH TIME ZONE

soil_record_parcel_id_fkey

temperature_record_parcel_id_fkey

s2_ts_record_parcel_id_fkey

soil_record

id	SERIAL
parcel_id	INTEGER
soil_type	TEXT
soil_type2	TEXT
soil_type3	TEXT
soil_type4	TEXT
created_at	TIMESTAMP(6) WITH TIME ZONE
updated_at	TIMESTAMP(6) WITH TIME ZONE

temperature_record

id	SERIAL
parcel_id	INTEGER
temperature_product_id	INTEGER
temperature_mean	REAL
created_at	TIMESTAMP(6) WITH TIME ZONE
updated_at	TIMESTAMP(6) WITH TIME ZONE

temperature_record_temperature_product_id_fkey

temperature_product

id	SERIAL
stop_date	TIMESTAMP(6) WITH TIME ZONE
filename	TEXT
created_at	TIMESTAMP(6) WITH TIME ZONE
updated_at	TIMESTAMP(6) WITH TIME ZONE

s2_product

id	SERIAL
start_date	TIMESTAMP(6) WITH TIME ZONE
stop_date	TIMESTAMP(6) WITH TIME ZONE
filename	TEXT
relative_orbit_number	INTEGER
created_at	TIMESTAMP(6) WITH TIME ZONE
updated_at	TIMESTAMP(6) WITH TIME ZONE

s2_ts_record_s2_product_id_fkey

Järgnevas tabelis on toodud tabelite ja nende veergude selgitused

Tabeli/veeru nimi	Selgitus
parcel	Põldude ruumiandmed, mille kohta põllukultuuride hinnanguid antakse.
created_at	Loomise aeg
f_id	Välise süsteemi identifikaator.
geom	Põllu MULTIPOLOYGON tüüpi ruumikuju EPSG:3301 koordinaatsüsteemis.
geometry_valid	Ruumikuju sobilikkust kajastav veerg. Väärtus on false kui ruumikuju ei ole sobilik hinnangu andmiseks.
id	Identifikaator.
updated_at	Muutmise aeg
precipitation_product	Sademesummade produktide nimistu, mille alusel on aegridu arvutatud.
created_at	Loomise aeg.
filename	Sademesummade produkti failinimi.
id	Identifikaator.
start_date	Sademesummade arvestamise algusaeg.
stop_date	Sademesummade arvestamise lõpuaeg.
updated_at	Muutmise aeg.
precipitation_record	Põldude sademesummade statistilised näitajad.
created_at	Loomise aeg.
id	Identifikaator.
parcel_id	Viide põllu id-le tabelis parcel.
precipitation_product_id	Viide sademesummade produkti tabeli ID-le.
precipitation_sum_mean	Sademesummade rastri keskmine väärtus põllu kohta.
updated_at	Muutmise aeg.
result	Põllukultuuride hinnangute tulemused
created_at	Loomise aeg
estimation_aedmaasikas	Aedmaasika esinemise tõenäosuse hinnang.
estimation_astelpaju	Astelpaju esinemise tõenäosuse hinnang.

estimation_heintaimed_korrelised	Heintaimed, kõrrelised esinemise tõenäosuse hinnang.
estimation_heintaimed_libilikoielised	Heintaimed, liblikõielised esinemise tõenäosuse hinnang.
estimation_kaer	Kaer esinemise tõenäosuse hinnang.
estimation_kanep	Kanep esinemise tõenäosuse hinnang.
estimation_kartul	Kartul esinemise tõenäosuse hinnang.
estimation_mais	Mais esinemise tõenäosuse hinnang.
estimation_mustkesa	Mustkesa esinemise tõenäosuse hinnang.
estimation_peakapsas	Peakapsas esinemise tõenäosuse hinnang.
estimation_poldhernes	Põldhernes esinemise tõenäosuse hinnang.
estimation_polduba	Põlduba esinemise tõenäosuse hinnang.
estimation_porgand	Porgand esinemise tõenäosuse hinnang.
estimation_punapeet	Punapeet esinemise tõenäosuse hinnang.
estimation_rukis	Rukis esinemise tõenäosuse hinnang.
estimation_sooti_jaetud_maa	Sööti jäetud maa esinemise tõenäosuse hinnang.
estimation_suvinisu_ja_speltanisu	Suvinisu ja speltanisu esinemise tõenäosuse hinnang.
estimation_suvioder	Suvioder esinemise tõenäosuse hinnang.
estimation_suviraps_ja_rups	Suviraps ja -rüps esinemise tõenäosuse hinnang.
estimation_suvitritikale	Suvitritikale esinemise tõenäosuse hinnang.
estimation_talinisu	Talinisu esinemise tõenäosuse hinnang.
estimation_talioder	Talioder esinemise tõenäosuse hinnang.
estimation_taliraps_ja_rups	Taliraps ja -rüps esinemise tõenäosuse hinnang.
estimation_talitritikale	Talitritikale esinemise tõenäosuse hinnang.
estimation_tatar	Tatar esinemise tõenäosuse hinnang.
id	Identifikaator
parcel_id	Põllu identifikaator parcel tabelis.
s1_product_id	Sentinel-1 produkti identifikaator s1_product tabelis.
updated_at	Muutmise aeg
s1_product	Sentinel-1 produktide nimistu, mille alusel on aegridu arvutatud.
created_at	Loomise aeg.
filename	Sentinel-1 produkti nimi.

id	Identifikaator.
relative_orbit_number	Satelliidi ülelennu orbiidi suhteline number.
start_date	Satelliidipildi algusaeg.
stop_date	Satelliidipildi lõpuaeg.
updated_at	Muutmise aeg.
s1_ts_record	Põldude Sentinel-1 koherentsuse ja sigma0 statistilised näitajad.
cohvh_max	Põllu koherentsrastri VH kanali maksimum väärtus.
cohvh_mean	Koherentsrastri keskmine väärtus põllu kohta VH kanalis.
cohvh_median	Põllu koherentsrastri VH kanali mediaan väärtus.
cohvh_min	Põllu koherentsrastri VH kanali miinimum väärtus.
cohvh_stdv	Põllu koherentsrastri VH kanali standardhälbe väärtus.
cohvv_max	Põllu koherentsrastri VV kanali maksimum väärtus.
cohvv_mean	Koherentsrastri keskmine väärtus põllu kohta VV kanalis.
cohvv_median	Põllu koherentsrastri VV kanali mediaan väärtus.
cohvv_min	Põllu koherentsrastri VV kanali miinimum väärtus.
cohvv_stdv	Põllu koherentsrastri VV kanali standardhälbe väärtus.
created_at	Loomise aeg.
id	Identifikaator.
parcel_id	Viide põllu id-le tabelis parcel.
s0vh_max	Põllu sigma0 rastri VH kanali maksimum väärtus.
s0vh_mean	Sigma0 rastri VH kanali keskmine väärtus põllu kohta.
s0vh_median	Põllu sigma0 rastri VH kanali mediaan väärtus.
s0vh_min	Põllu sigma0 rastri VH kanali miinimum väärtus.
s0vh_stdv	Põllu sigma0 rastri VH kanali standardhälbe väärtus.
s0vv_max	Põllu sigma0 rastri VV kanali maksimum väärtus.
s0vv_mean	Sigma0 rastri VV kanali keskmine väärtus põllu kohta.
s0vv_median	Põllu sigma0 rastri VV kanali mediaan väärtus.
s0vv_min	Põllu sigma0 rastri VV kanali miinimum väärtus.
s0vv_stdv	Põllu sigma0 rastri VV kanali standardhälbe väärtus.
s1_product_id	Viide Sentinel-1 produkti tabeli ID-le.
updated_at	Muutmise aeg.

vhvv_max	Põllu sigma0 rastri VH ja VV kanalite suhte (VH/VV) maksimum väärtus.
vhvv_mean	Sigma0 rastri VH ja VV kanalite suhte (VH/VV) keskmine väärtus põllu kohta.
vhvv_median	Põllu sigma0 rastri VH ja VV kanalite suhte (VH/VV) mediaan väärtus.
vhvv_min	Põllu sigma0 rastri VH ja VV kanalite suhte (VH/VV) miinimum väärtus.
vhvv_stdv	Põllu sigma0 rastri VH ja VV kanalite suhte (VH/VV) standardhälbe väärtus.
s2_product	Sentinel-2 produktide nimistu, mille alusel on aegridu arvutatud.
created_at	Loomise aeg.
filename	Sentinel-2 produkti nimi.
id	Identifikaator.
relative_orbit_number	Satelliidi ülelennu orbiidi suhteline number.
start_date	Satelliidipildi algusaeg.
stop_date	Satelliidipildi lõpuaeg.
updated_at	Muutmise aeg.
s2_ts_record	Põldude Sentinel-2 kanalite ning indeksite statistilised näitajad.
b11_max	Satelliidi pildist arvutatud põllu b11 kanali maksimum väärtus.
b11_mean	Satelliidi pildist arvutatud põllu keskmine b11 kanali väärtus.
b11_median	Satelliidi pildist arvutatud põllu b11 kanali mediaan väärtus.
b11_min	Satelliidi pildist arvutatud põllu b11 kanali miinimum väärtus.
b11_stdv	Satelliidi pildist arvutatud põllu b11 kanali standardhälbe väärtus.
b12_max	Satelliidi pildist arvutatud põllu b12 kanali maksimum väärtus.
b12_mean	Satelliidi pildist arvutatud põllu keskmine b12 kanali väärtus.
b12_median	Satelliidi pildist arvutatud põllu b12 kanali mediaan väärtus.

b12_min	Satelliidi pildist arvutatud põllu b12 kanali miinimum väärtus.
b12_stdv	Satelliidi pildist arvutatud põllu b12 kanali standardhälbe väärtus.
b2_max	Satelliidi pildist arvutatud põllu b2 kanali maksimum väärtus.
b2_mean	Satelliidi pildist arvutatud põllu keskmine b2 kanali väärtus.
b2_median	Satelliidi pildist arvutatud põllu b2 kanali mediaan väärtus.
b2_min	Satelliidi pildist arvutatud põllu b2 kanali miinimum väärtus.
b2_stdv	Satelliidi pildist arvutatud põllu b2 kanali standardhälbe väärtus.
b3_max	Satelliidi pildist arvutatud põllu b3 kanali maksimum väärtus.
b3_mean	Satelliidi pildist arvutatud põllu keskmine b3 kanali väärtus.
b3_median	Satelliidi pildist arvutatud põllu b3 kanali mediaan väärtus.
b3_min	Satelliidi pildist arvutatud põllu b3 kanali miinimum väärtus.
b3_stdv	Satelliidi pildist arvutatud põllu b3 kanali standardhälbe väärtus.
b4_max	Satelliidi pildist arvutatud põllu b4 kanali maksimum väärtus.
b4_mean	Satelliidi pildist arvutatud põllu keskmine b4 kanali väärtus.
b4_median	Satelliidi pildist arvutatud põllu b4 kanali mediaan väärtus.
b4_min	Satelliidi pildist arvutatud põllu b4 kanali miinimum väärtus.
b4_stdv	Satelliidi pildist arvutatud põllu b4 kanali standardhälbe väärtus.
b5_max	Satelliidi pildist arvutatud põllu b5 kanali maksimum väärtus.
b5_mean	Satelliidi pildist arvutatud põllu keskmine b5 kanali väärtus.
b5_median	Satelliidi pildist arvutatud põllu b5 kanali mediaan väärtus.
b5_min	Satelliidi pildist arvutatud põllu b5 kanali miinimum väärtus.
b5_stdv	Satelliidi pildist arvutatud põllu b5 kanali standardhälbe väärtus.
b6_max	Satelliidi pildist arvutatud põllu b6 kanali maksimum väärtus.
b6_mean	Satelliidi pildist arvutatud põllu keskmine b6 kanali väärtus.

b6_median	Satelliidi pildist arvutatud põllu b6 kanali mediaan väärtus.
b6_min	Satelliidi pildist arvutatud põllu b6 kanali miinimum väärtus.
b6_stdv	Satelliidi pildist arvutatud põllu b6 kanali standardhälbe väärtus.
b7_max	Satelliidi pildist arvutatud põllu b7 kanali maksimum väärtus.
b7_mean	Satelliidi pildist arvutatud põllu keskmine b7 kanali väärtus.
b7_median	Satelliidi pildist arvutatud põllu b7 kanali mediaan väärtus.
b7_min	Satelliidi pildist arvutatud põllu b7 kanali miinimum väärtus.
b7_stdv	Satelliidi pildist arvutatud põllu b7 kanali standardhälbe väärtus.
b8_max	Satelliidi pildist arvutatud põllu b8 kanali maksimum väärtus.
b8_mean	Satelliidi pildist arvutatud põllu keskmine b8 kanali väärtus.
b8_median	Satelliidi pildist arvutatud põllu b8 kanali mediaan väärtus.
b8_min	Satelliidi pildist arvutatud põllu b8 kanali miinimum väärtus.
b8_stdv	Satelliidi pildist arvutatud põllu b8 kanali standardhälbe väärtus.
b8a_max	Satelliidi pildist arvutatud põllu b8a kanali maksimum väärtus.
b8a_mean	Satelliidi pildist arvutatud põllu keskmine b8a kanali väärtus.
b8a_median	Satelliidi pildist arvutatud põllu b8a kanali mediaan väärtus.
b8a_min	Satelliidi pildist arvutatud põllu b8a kanali miinimum väärtus.
b8a_stdv	Satelliidi pildist arvutatud põllu b8a kanali standardhälbe väärtus.
created_at	Loomise aeg.
id	Identifikaator.
misra_yellow_vegetation_max	Satelliidi pildist arvutatud põllu MISRA yellow vegetation maksimum väärtus.
misra_yellow_vegetation_mean	Satelliidi pildist arvutatud põllu keskmine MISRA yellow vegetation väärtus.
misra_yellow_vegetation_median	Satelliidi pildist arvutatud põllu MISRA yellow vegetation mediaan väärtus.

misra_yellow_vegetation_min	Satelliidi pildist arvutatud põllu MISRA yellow vegetation miinimum väärtus.
misra_yellow_vegetation_stdv	Satelliidi pildist arvutatud põllu MISRA yellow vegetation standardhälbe väärtus.
ndvi_max	Satelliidi pildist arvutatud põllu NDVI maksimum väärtus.
ndvi_mean	Satelliidi pildist arvutatud põllu keskmine NDVI väärtus.
ndvi_median	Satelliidi pildist arvutatud põllu NDVI mediaan väärtus.
ndvi_min	Satelliidi pildist arvutatud põllu NDVI miinimum väärtus.
ndvi_stdv	Satelliidi pildist arvutatud põllu NDVI standardhälbe väärtus.
ndvire_max	Satelliidi pildist arvutatud põllu NDVIRE maksimum väärtus.
ndvire_mean	Satelliidi pildist arvutatud põllu keskmine NDVIRE väärtus.
ndvire_median	Satelliidi pildist arvutatud põllu NDVIRE mediaan väärtus.
ndvire_min	Satelliidi pildist arvutatud põllu NDVIRE miinimum väärtus.
ndvire_stdv	Satelliidi pildist arvutatud põllu NDVIRE standardhälbe väärtus.
ndwi_max	Satelliidi pildist arvutatud põllu NDWI maksimum väärtus.
ndwi_mean	Satelliidi pildist arvutatud põllu keskmine NDWI väärtus.
ndwi_median	Satelliidi pildist arvutatud põllu NDWI mediaan väärtus.
ndwi_min	Satelliidi pildist arvutatud põllu NDWI miinimum väärtus.
ndwi_stdv	Satelliidi pildist arvutatud põllu NDWI standardhälbe väärtus.
parcel_id	Viide põllu id-le tabelis parcel.
psri_max	Satelliidi pildist arvutatud põllu PSRI maksimum väärtus.
psri_mean	Satelliidi pildist arvutatud põllu keskmine PSRI väärtus.
psri_median	Satelliidi pildist arvutatud põllu PSRI mediaan väärtus.
psri_min	Satelliidi pildist arvutatud põllu PSRI miinimum väärtus.
psri_stdv	Satelliidi pildist arvutatud põllu PSRI standardhälbe väärtus.
s2_product_id	Viide Sentinel-2 produkti tabeli ID-le.
tc_brightness_max	Satelliidi pildist arvutatud põllu TC brightness maksimum väärtus.
tc_brightness_mean	Satelliidi pildist arvutatud põllu keskmine TC brightness väärtus.

tc_brightness_median	Satelliidi pildist arvutatud põllu TC brightness mediaan väärtus.
tc_brightness_min	Satelliidi pildist arvutatud põllu TC brightness miinimum väärtus.
tc_brightness_stdv	Satelliidi pildist arvutatud põllu TC brightness standardhälbe väärtus.
tc_vegetation_max	Satelliidi pildist arvutatud põllu TC vegetation maksimum väärtus.
tc_vegetation_mean	Satelliidi pildist arvutatud põllu keskmine TC vegetation väärtus.
tc_vegetation_median	Satelliidi pildist arvutatud põllu TC vegetation mediaan väärtus.
tc_vegetation_min	Satelliidi pildist arvutatud põllu TC vegetation miinimum väärtus.
tc_vegetation_stdv	Satelliidi pildist arvutatud põllu TC vegetation standardhälbe väärtus.
tc_wetness_max	Satelliidi pildist arvutatud põllu TC wetness maksimum väärtus.
tc_wetness_mean	Satelliidi pildist arvutatud põllu keskmine TC wetness väärtus.
tc_wetness_median	Satelliidi pildist arvutatud põllu TC wetness mediaan väärtus.
tc_wetness_min	Satelliidi pildist arvutatud põllu TC wetness miinimum väärtus.
tc_wetness_stdv	Satelliidi pildist arvutatud põllu TC wetness standardhälbe väärtus.
updated_at	Muutmise aeg.
wri_max	Satelliidi pildist arvutatud põllu WRI maksimum väärtus.
wri_mean	Satelliidi pildist arvutatud põllu keskmine WRI väärtus.
wri_median	Satelliidi pildist arvutatud põllu WRI mediaan väärtus.
wri_min	Satelliidi pildist arvutatud põllu WRI miinimum väärtus.
wri_stdv	Satelliidi pildist arvutatud põllu WRI standard hälbeväärtus.
soil_record	Põldude mullastiku andmeid sisaldav tabel.
created_at	Loomise aeg.
id	Identifikaator.

parcel_id	Viide põllu id-le tabelis parcel.
soil_type	Pindala alusel suurima osakaaluga põllul esinev mullatüüp.
updated_at	Muutmise aeg.
temperature_product	Päevakeskmiste temperatuuride produktide nimistu, mille alusel on aegridu arvutatud.
created_at	Loomise aeg.
filename	Päevakeskmise temperatuuri produkti failinimi.
id	Identifikaator.
stop_date	Päevakeskmise temperatuuri arvestamise lõpuaeg.
updated_at	Muutmise aeg.
temperature_record	Põldude päevakeskmiste temperatuuride statistilised näitajad.
created_at	Loomise aeg.
id	Identifikaator.
parcel_id	Viide põllu id-le tabelis parcel.
temperature_mean	Päevakeskmise temperatuuri rastri keskmine väärtus põllu kohta.
temperature_product_id	Viide päevakeskmise temperatuuri produkti tabeli ID-le.
updated_at	Muutmise aeg.

Lisa 2 - Eesti 1:10000 digitaalse mullakaardi tüpoloogia üldistamine masinõppe jaoks

Mullakaardi andmestik on alla laetav Eesti Maa-ameti kodulehelt: https://geoportaal.maaamet.ee/est/Andmed-ja-kaardid/Mullastiku-kaart-p33.html

Mullakaardi andmestikku saab siduda erinevate klassifikatsioonieeskirjadega läbi mullašifri ja rühmitada võttes aluseks nende paiknemise muldade ordineeritud maatriksil (Joonis 4), kus on litoloogiline-geneetiline telg ning üldistatud niiskuse telg (Kõlli jt 2004). Selle maatriksi võtsid aluseks muldade rühmitamisel Lang jt (2018) puistute liigilise koosseisu hindamise uuringus, kus nii mullakaarti ja multispektraalseid satelliidipilte kasutati näidistel põhinevas masinõppeprotseduuris.

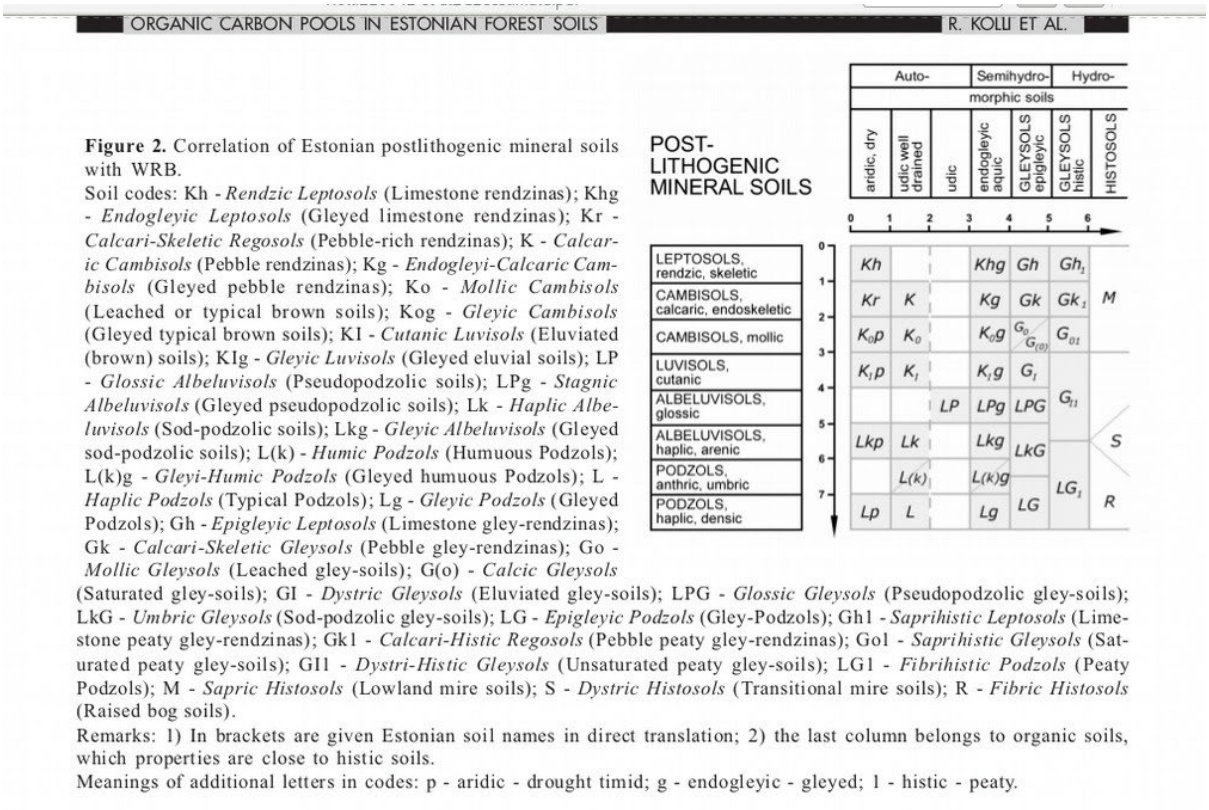


Figure 2. Correlation of Estonian postlithogenic mineral soils with WRB.
Soil codes: Kh - *Rendzic Leptosols* (Limestone rendzinas); Khg - *Endogleyic Leptosols* (Gleyed limestone rendzinas); Kr - *Calcari-Skeletic Regosols* (Pebble-rich rendzinas); K - *Calcaric Cambisols* (Pebble rendzinas); Kg - *Endogleyi-Calcaric Cambisols* (Gleyed pebble rendzinas); Ko - *Mollic Cambisols* (Leached or typical brown soils); Kog - *Gleyic Cambisols* (Gleyed typical brown soils); KI - *Cutanic Luvisols* (Eluviated (brown) soils); KIg - *Gleyic Luvisols* (Gleyed eluvial soils); LP - *Glossic Albeluvisols* (Pseudopodzolic soils); LPg - *Stagnic Albeluvisols* (Gleyed pseudopodzolic soils); Lk - *Haplic Albeluvisols* (Sod-podzolic soils); Lkg - *Gleyic Albeluvisols* (Gleyed sod-podzolic soils); L(k) - *Humic Podzols* (Humuous Podzols); L(k)g - *Gleyi-Humic Podzols* (Gleyed humuous Podzols); L - *Haplic Podzols* (Typical Podzols); Lg - *Gleyic Podzols* (Gleyed Podzols); Gh - *Epigleyic Leptosols* (Limestone gley-rendzinas); Gk - *Calcari-Skeletic Gleysols* (Pebble gley-rendzinas); Go - *Mollic Gleysols* (Leached gley-soils); G(o) - *Calcic Gleysols* (Saturated gley-soils); GI - *Dystric Gleysols* (Eluviated gley-soils); LPG - *Glossic Gleysols* (Pseudopodzolic gley-soils); LkG - *Umbric Gleysols* (Sod-podzolic gley-soils); LG - *Epigleyic Podzols* (Gley-Podzols); Gh1 - *Saprihistic Leptosols* (Limestone peaty gley-rendzinas); Gk1 - *Calcari-Histic Regosols* (Pebble peaty gley-rendzinas); Go1 - *Saprihistic Gleysols* (Saturated peaty gley-soils); GI1 - *Dystri-Histic Gleysols* (Unsaturated peaty gley-soils); LG1 - *Fibrihistic Podzols* (Peaty Podzols); M - *Sapric Histosols* (Lowland mire soils); S - *Dystric Histosols* (Transitional mire soils); R - *Fibric Histosols* (Raised bog soils).
Remarks: 1) In brackets are given Estonian soil names in direct translation; 2) the last column belongs to organic soils, which properties are close to histic soils.
Meanings of additional letters in codes: p - aridic - drought timid; g - endogleyic - gleyed; 1 - histic - peaty.

Joonis 4. Muldade maatrikstabel võetuna Kõlli jt (2004) artiklist.

Muldi litoloogilis-geneetilise/niiskuse maatriksi abil liigitades jagati mullad 23 erineva koodiga klassi: 0, 10, 11, 13, 14, 16, 21, 31, 37, 42, 43, 44, 45, 48, 51, 53, 57, 61, 63, 64, 73, 77 ja 109. Täpne vastavus saadud mullaklassi (mullatüübi) ja esialgse šifri vahel on toodud jagatud SharePointi kaustas failis *„Mudeli_definitsioon/Mullakaardi_kasutamine/SIF1_LOEND.DBF*“. Teatmikufailis *SIF1_LOEND.DBF* on järgmised väljad:

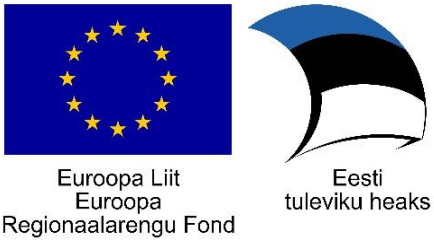
- Sif1 - väljas sif1 mullakaardi andmebaasis olev märgijada,
- Arv - vastava märgijadaga kirjete arv,
- Siffer - eeldatav mullašiffer,
- Klass_mtrx - muldade ordinatsioonimaatriksi järgi omistatud üldine klass,
- Litgen_y ja veerez_x mulla koordinaadid maatriksil.

Faili märgistik on 1250 (Eastern European Windows). Mulla eeldatava šifri ja mulla koordinaatide väärtustes võib kohati olla ebatäpsusi, kuid üldiselt selgus pistelisel kontrollil hea kooskõla mullakaardi andmebaasi algse tüübikoodiga. Vajadusel on kasutajal lihtne viia sisse parandusi tõstes klasse kokku nende mullatüüpide osas, kus praegune eristus ei oma tähtsust klassifitseeritava tunnuse (põllukultuur) jaoks, või siis lisades kohtadesse, kus muldade alamtüübid mõjutavalt selgelt põllukultuuri valikut.

Viited

Kõlli, R., Asi, E., Köster, T. 2004. Organic Carbon Pools in Estonian Forest Soils. Baltic Forestry, 10 (1): 19-26.

Lang, M., Kaha, M., Laarmann, D., Sims, A. 2018. Construction of tree species composition map of Estonia using multispectral satellite images, soil map and a random forest algorithm. – Forestry Studies | Metsanduslikud Uurimused 68, 5–24.



D4.5 Lisa 1. Põllukultuuride tunnusvektorite aegridade rühmitamine klasterdamise abil ja eksete eemaldamine õpetusandmetest

Metoodika kirjeldus

Koostasid: Mait Lang, Tartu Ülikool
Mihkel Järveoja, OÜ KappaZeta

Projekt RITA1/02-52 „Kaugseire andmete kasutuselevõtt avalike teenuste väljatöötamisel ja arendamisel“

Tartu 2020

Sisukord

1. Sissejuhatus

Kõikide PRIA põldude kohta on taotlusel märgitud, mis seal kasvama peaks. Enamasti on see õige, aga PRIA hinnangul võib seal olla kuni 5% vigu. Selleks, et põldude andmekogu saaks kasutada mudeli arenduseks, on vaja näidiste seast võimalikult palju vigadega kirjeid (ekseid) eemaldada.

Üldiselt on põllukultuurid üle välja homogeensed ehk pikslite väärtused konkreetses spektraalkanalis või radarmõõtmiste põhjal arvutatud tunnuse korral põllu piires oluliselt ei varieeru. Eeldatavalt võiks olla kaugseiretunnuste aegridade põhjal võimalik rühmitada põlde nii, et samasse rühma satuvad mingi kindla kultuuriga põllud. Teisalt ei ole aegread siledad mõõtemääramatuse, vaatesuuna, valgustatuse muutuva geomeetria ja pilvemaski vigade tõttu. Samuti on paljude põllukultuuride signatuurid ja nende muutus ajas omavahel sarnased, suuremaid erinevusi võib leida hoopis sama kultuuriga erinevate põldude vahel. Siiski on aegridade klasterdamine hea võimalus otsida suurest andmestikust võimalikke erindeid.

Andmete klasterdamiseks leiab erinevaid meetodeid. Käesolevas töös kasutati andmeanalüüsiks R-keskkonnas paketti kml (Genolini jt, 2015), mille loomisel on lähtutud just aegridade erisustest. Paketis sisalduvad ka graafilised töövahendid andmestikust esmase ülevaate saamiseks.

Klasterdamise metoodika ja töötlusahela pani kokku Mait Lang (Tartu Ülikool). Klastritest eksete leidmise tööriistad ja põhimõtted Mihkel Järveoja (KappaZeta OÜ).

2. Andmestik ja selle ettevalmistamine

Andmestik koosnes tabelitest *parcelv_narrow.csv (ja parcelv_wide)*, *s1_ts_recordv.csv, s2_ts_recordv.csv*, milles olid põllukultuuride andmed, Sentinel-1 SAR aegread ning Sentinel-2 MSI aegread. Tabelite omavaheliseks sidumiseks oli väli *parcel_id*. Põllukultuurid olid antud nii numbrilise ID-koodiga kui ka sõnalise seletusega. Edasiseks tööks kasutati numbrilist ID-tähist (*crop_id*).

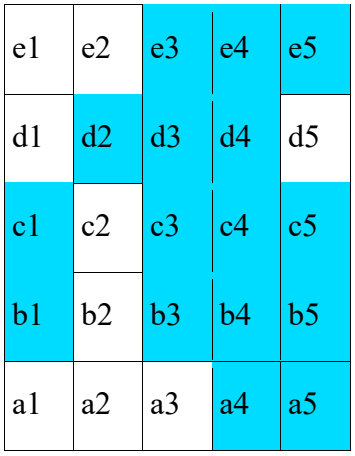
Mainitud .csv failid eksporditi andmebaasist järgmiste käskuga:

```
\COPY (SELECT parcel_id, parcel_f_id, crop_id, crop_code, crop_name, x_norm_loc, y_norm_loc FROM parcelv_narrow) TO '.../parcelv_narrow.csv' DELIMITER ',' CSV HEADER;

\COPY (SELECT * FROM s1_ts_recordv) TO '.../s1_ts_recordv.csv' DELIMITER ',' CSV HEADER;

\COPY (SELECT * FROM s2_ts_recordv) TO '.../s2_ts_recordv.csv' DELIMITER ',' CSV HEADER;
```

Põldude tabelist *parcelv* võeti väljad *parcel_id, crop_id, x_norm_loc, y_norm_loc*, millest viimased kaks on põllu suhtelised koordinaadid. Sentinel-2 MSI andmete tabelist võeti väljad *parcel_id, s2_product_id, s2_product_start_date, b2_mean, b4_mean, b8_mean, b11_mean*

ja *ndvi_mean*. Sentinel-1 SAR tabelist võeti väljad *parcel_id, s1_product_id, s1_product_stop_date, cohvv_mean, vhvv_mean.*

Eesti ala rühmitati suhteliste koordinaatide (*x_norm_loc, y_norm_loc)* järgi esmalt 25 osaks, millest ruutude ühendamise järel põldude arvu ühtlustamiseks jäid alles kohati laiendatud d2, c1, e4, c3, d3, b1, a5, d4, b3, b5, c4, b4, a4, c5, e3, e5 (Joonis 5).



Joonis 5. Eesti ala rühmitamine suhteliste koordinaatide järgi ja ruutude ühendamise järel alles jäänud ruudud (sinised).

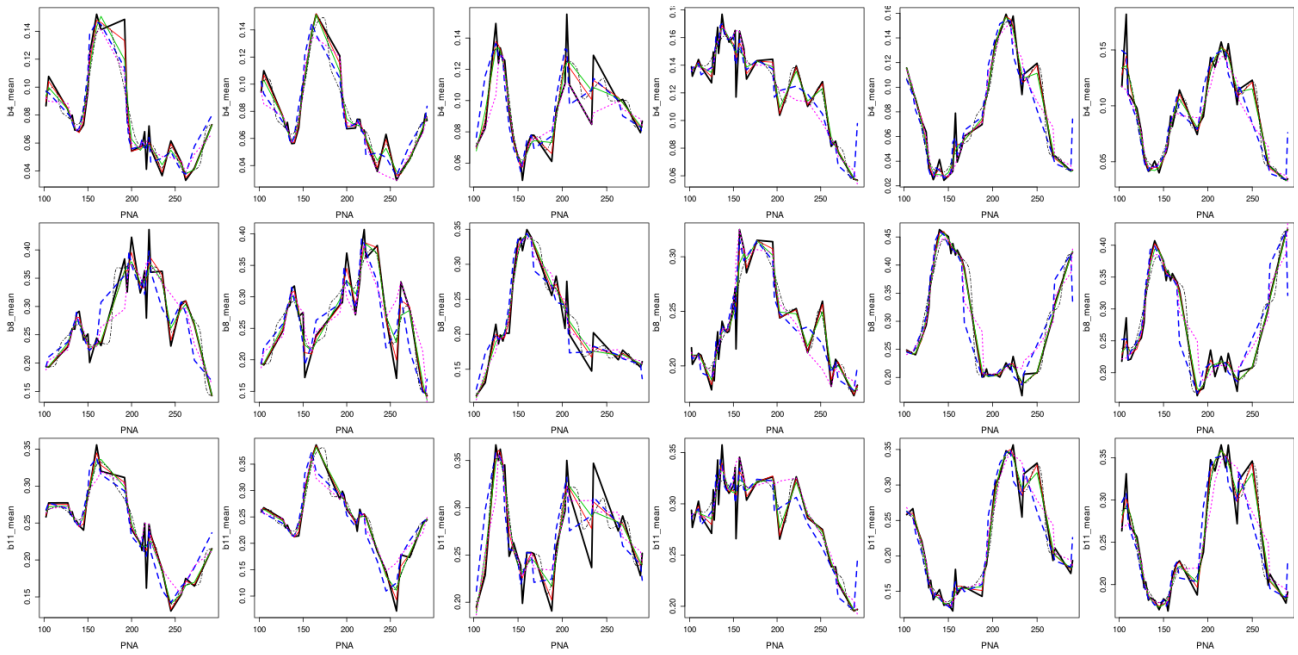
Ettevalmistuse viimaseks osaks oli aegridade silumine ja täitmine nii, et peamise kultuuri kasvuperioodi iga päeva kohta oleks olemas iga tunnuse lugem igal põllul. Klasterdamise protseduurile saab dokumentatsiooni kohaselt anda ette puuduvate väärtuste leidmise meetodi. Siiski eelistati klasterdamise töö ajal aegridade täitmise asemel nende eelnevat põhiandmetabelis silumist, sest katsetamise käigus tulnuks muidu iga kord kulutada aega andmeridade silumisele. Kasvuperioodiks võeti päevade vahemik 120–250 (30. aprill – 7. september), mis teeb kokku 131 päeva. Kaugseiremõõtmiste kuupäevad teisendati päeva numbriks aastas. Aegridade täitmiseks kasutati kml paketist protseduuri *imputation* nii, et rea keskel puuduvad väärtused interpooliti lineaarselt, aga otstes puuduvate asemele kasutati lähimat väärtust. See on ka protseduuri vaikimisi seadistus. Aegridade silumine oli üsna ressursinõudlik protseduur.

Põldude ja kaugseireandmete aegridade kokkuliitmisel tekkis tabel *pollud*, kus iga rida oli üks põld.

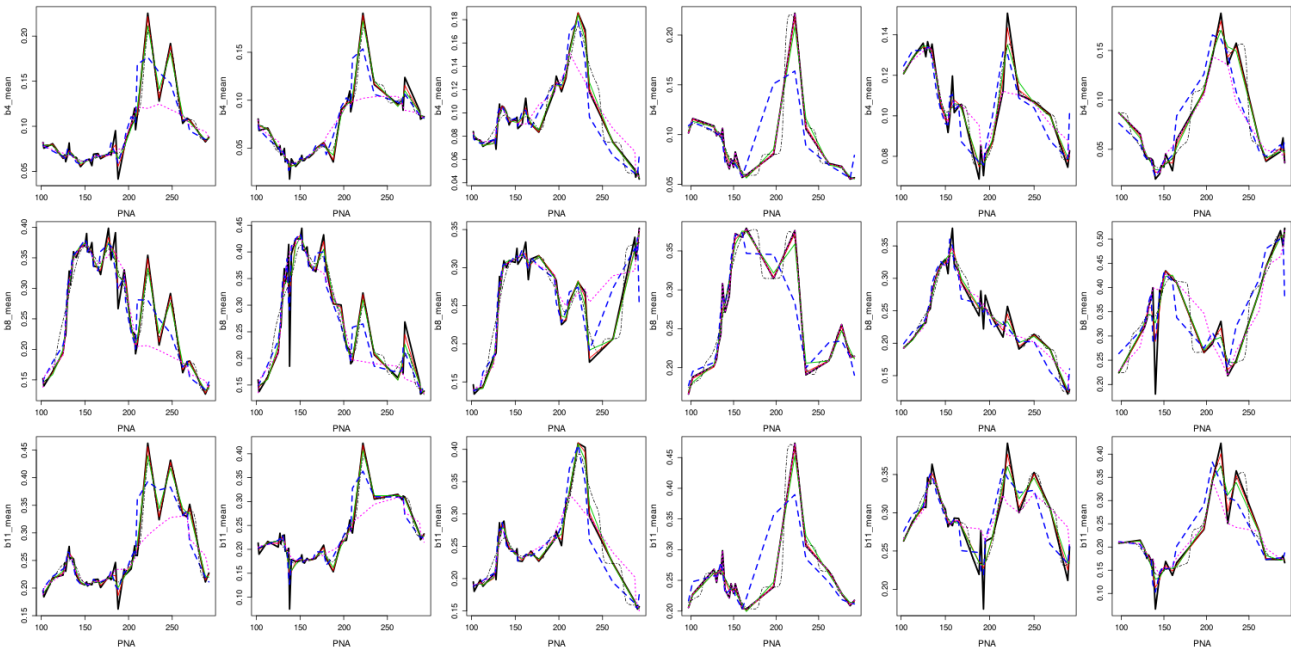
Radartunnuste aegridade silumiseks kasutati kerneli (R-i funktsioon *filter*) abil keskmistamist kaaludega (0.025, 0.04,0.085,0.2,0.3,0.2,0.085,0.04,0.025). Kaalude jaotus leiti eksperimenteerides nii, et säilitada võimalikult palju andmeridade omapära, aga samas kõrvaldada lokaalset müra.

Käesolevas töös ei leitud head lahendust Sentinel-2 MSI andmeridade silumiseks (joonis 6) ja silumist ei rakendatud. Sentinel-2 MSI aegridades võib hõredates osades üksik vaatlus olla nii

info (näiteks niitmine või viljakoristus) või siis hoopis pilvemaski viga. Joonistel (6 ja 7) toodud näited on üsna tüüpilised Sentinel-2 MSI aegread, millest paistab nii põllukultuuri üldine kasvukäik kui ka lokaalne variatsioon. Spektri lähiinfrapunases kanalis (MSI-8) on sama põllukultuuri heleduste käigud pigem sarnased, aga punases (MSI-4) ning keskmises infrapunases (MSI-11) esineb olulisi erinevusi. Praktilise rakenduse poolt vaadates võiks olla tegemist signaaliga, mille alusel saame tuvastada ekseid. Spektraalsete heleduste kasvukäikude tõlgendamisel tuleb siiski arvestada, et seda mõjutavad: mulla spektraalne heledus (sõltub oluliselt pealispinna niiskusest), roheliste taimelehtede katvus, taimeosade optilised omadused ja paiknemine ruumis, õite ja viljade hulk ning vaatesuuna- ja valgustatuse geomeetria. Sama liiki põllukultuuri puhul võib olla kasutatud erinevaid agrotehnikaid.



Joonis 6. Mõned näited aegridadest. Joonisel on kaks kartulipõldu, kaks suviodra põldu ja kaks taliodra põldu. Jämeda joonega on toodud interpoolitud aegrida, teised jooned näitavad võimalusi aegrea juhuslike võngete silumiseks.



Joonis 7. Talinisu põldude spektraalsete aegridade näiteid. Jämeda joonega on toodud interpoolitud aegrida, teised jooned näitavad võimalusi aegrea juhuslike võngete silumiseks.

3. Klasterdamine

Põldude klasterdamist tehti eelnevalt rakendatud ruumiühikute (ruutude ja nende ühendite) piires. Nii on võimalik arvestada lokaalseid erinevusi fenoloogias. Kuigi iga põllu kohta on arvutatud väga rikkalik Sentinel-1 ja -2 parameetrite tunnuskomplekt, siis suur osa tunnustes sisalduvast infost siiski kattub ja tõenäoliselt on võimalik hõlmata väga suur osa (>95%) kogu põllukultuuride eristamiseks vajalikust infost valides oskuslikult ainult osa tunnuskomplektist. Sel viisil väheneb oluliselt aeg ja arvutusvõimsus, mida klasterdamine nõuab. Antud juhul kasutati klasterdamiseks korraga nelja tunnust: Sentinel-2 MSI kanal 11 (*b11_mean*), Sentinel-2 MSI andmetest arvutatud NDVI indeks (*ndvi_mean*), Sentinel-1 SAR andmetest arvutatud polarisatsioonide suhe (*vhvv_mean*) ja koherents (*cohvv_mean*).

Klastrite arvu valimine on kogu protseduuri kõige subjektiivsem osa. Intuitiivselt võiks olla eelistatud põllukultuuride arvuga sama klastrite arv (detailne klassifikatsioon (*narrow*): 28, jäme klassifikatioon (*wide*): 16), kuid sisendtunnustes olev müra ning erinevate põllukultuuride tihti pigem sarnased signatuurid annavad alust klastrite arvu kahandada. See võimaldab töötlusprotseduuri kiirendada ja samas ikkagi välja tuua need objektid, mis oma tunnuste poolest eeldatavasse gruppi ei satu. Mõningase eksperimenteerimise järel valiti klastrite arvuks 10. Klasterdamiseks kasutati protseduuri *cld*() vaikimisi seadetega, klastrite arvuks määrati 10. *Cld*() kasutab klastrite tsentrite leidmiseks *k-means* algoritmi erinevaid modifikatsioone, vaikimisi on meetod *kmeans++* (Genolini jt, 2015). Klastrite eristamise eesmärgiks on minimeerida klastrite sisemist varieeruvust ja maksimeerida klastrite vahelist erinevust. Objektide määramisel klastritesse kasutatakse vaikimisi eukleidilist kaugust. Kuna

k-means sisaldab juhuslikku komponenti klastrite algseisu leidmisel, siis salvestatakse iga tellitud klastrite arvu kohta 20 erinevat klasterdamise tulemust. Klasterdamise tulemused järjestatakse indeksite järgi, mis näitavad klastrite sisemise ja klastrite vahelise variatsiooni suhet.

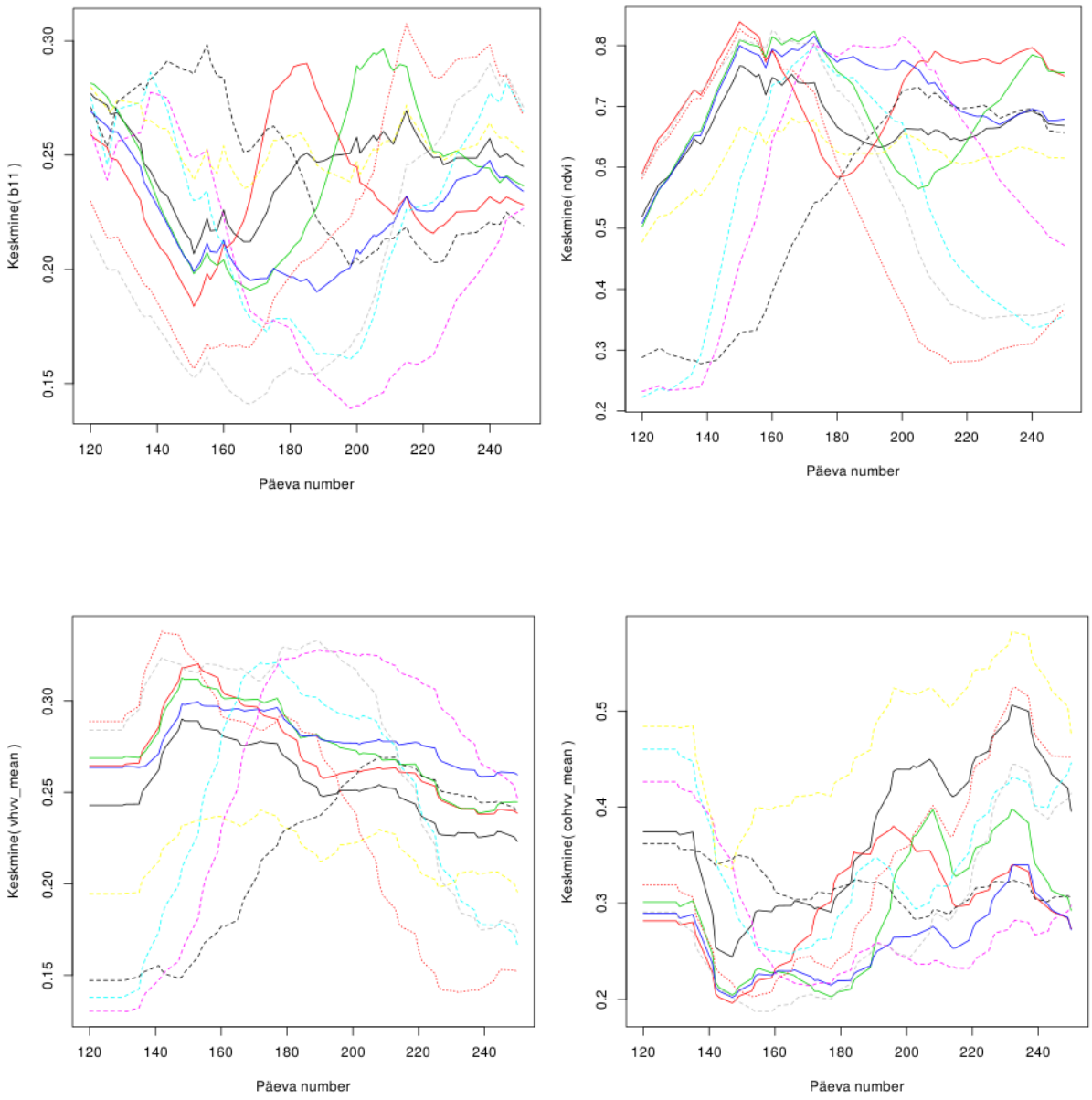
Klasterdamise tulemusena kirjutati iga ruudu kohta samanimelisse kausta iga tunnuse jaoks:

1) põllu kuuluvus klastrisse (*_*klastrid-kultuurid.loend*);

2) klastrite tunnusvektorite keskväärtuste tabel (**.klastrikeskmine*);

3) põllukultuuride ja neile omistatud klastrite kokkuvõte (**_klastrid-kultuurid.risttabel*) ja

4) joonised klastrite tunnusvektoritest (Joonis 8).



Joonis 8. Klastrite signatuurid ruudus d2 olevatele põldudele. Klastri tunnusvektoritesse kandub edasi põldude aegridade andmetes olev müra. Aegridade silumine on aidanud radartunnuste signatuurid muuta palju selgemaks.

4. Klasterdamise prototüüptarkvara kirjeldus

Lisaks klasterdamise metoodikale koostati prototüüptarkvara põllukultuuride tunnusvektorite aegridade klasterdamiseks R-keskkonnas paketi kml3d (Genolini jt, 2015) abil.

4.1 Eeldused

Arvutisse on paigaldatud R ning paketid *kml* ja *kml3d.* Graafilist liidest pole vaja, kõiki toiminguid ja etappe saab käivitada Linuxi terminalist.

```
R> install.packages("kml")
R> install.packages("kml3d")
```

4.2 Klasterdamise etapid

Klasterdamise protseduur koosneb järgmistest etappidest.

- 001.import.R abil loetakse CSV-failidest sisse põldude tabel ning kaugseireandmed.
  ```
  R> source('001.import.R')
  ```
- 002.lisa.ryhmitusruut.R lisab põldude tabelisse geograafilise asukoha järgi rühmitava tunnuse *ruut_id.*
  ```
  R> source('002.lisa.ryhmitusruut.R')
  ```
- 003.ettevalmistus.klasterdamiseks.R tõstab kaugseireandmed klasterdamise protseduuri jaoks sobivalt põldude tabelisse ning silub selle käigus ka radariandmetest arvutatud tunnuste müra. Kõige ajakulukam etapp.
  ```
  R> source('003.ettevalmistus.klasterdamiseks.R')
  ```
- 009.eralda_ruudud.R tõstab põldude andmestiku geograafiliste üksuste kaupa eraldi tööruumidesse, et klasterdamise tööd saaks käivitada rööbiti.
  ```
  R> source('009.eralda_ruudud.R')
  ```
- 010.yhe_tunnuse_jargi.ruut.R teeb ühe tunnuse järgi klasterdamise ning kirjutab välja tulemused. Skriptina terminalilt käivitades tuleb anda kaasa töödeldava ruudu tähis, mille järgi laaditakse tööruum. Seda kasutati ettevalmistavas katses.
- 011.mitme_tunnuse_jargi.ruut.R teeb mitme tunnuse järgi klasterdamise ning kirjutab välja tulemused. Skriptina terminalilt käivitades tuleb anda kaasa töödeldava ruudu tähis, mille järgi laaditakse tööruum. Saab paralleelselt käivitada .sh skriptidega eri ruutude peal.
  ```
  bash klasterdamine_osa1.sh
  bash klasterdamine_osa2.sh
  ```

Tulemused kirjutatakse geograafilise rühmitamistunnuse (*ruut_id*) kaupa kaustadesse.

5. Eksete leidmine klasterdatud andmetest

Geograafiliste ruutude kaupa klasterdatud andmestikust eksete leidmiseks tehti Pythoni lühiprogramm ja töötati välja eksete leidmise metoodika, mille eesmärgiks on iga kultuurigrupi piires eemaldada kõige väiksemahulisemad klastrid (ja nendesse klastridesse määratud põllud), mis moodustavad kokku kuni 10% kogu selle kultuurigrupi põldudest.

Väikeste klastrite eemaldamise eelduseks on oletus, et enamus taotlustel märgitud põllukultuure on siiski õiged ja suurematesse klastritesse kuuluvad põllud ei ole eksed.

Järgnevalt kirjeldatud samme korratakse tsüklina iga ruudu ja selle kohta käiva **_klastrid-kultuurid.loend* faili puhul:

1. Luuakse risttabel põllukultuuridest ja neile omistatud klastritest. Risttabel on ülesehituselt nagu klasterdamise kõrvalproduktina tekkiv **_klastrid-kultuurid.risttabel* (vt Joonis 9).

klastrid

kultuurid	1	2	3	4	5	6	7	8	9	10	
1	0	2	1	0	0	0	2	0	1	1	
2	988	865	770	5	437	1	71	20	6	22	sum_crop
3	128	134	265	4	221	3	103	20	3	14	
4	32	74	40	3	47	3	69	9	17	25	
5	0	0	0	0	0	0	5	0	0	0	
6	0	0	0	0	0	0	6	2	1	0	
7	0	0	0	41	2	15	5	9	26	0	
8	0	0	0	2	0	11	10	28	3	1	
9	1	0	0	3	3	0	2	2	0	0	
10	0	0	0	0	1	0	31	1	0	0	
11	0	5	0	0	2	0	0	0	0	0	
12	1	3	1	0	2	0	0	0	0	0	
13	5	4	6	1	3	1	8	1	25	1	
17	1	0	0	1	1	3	50	127	3	0	
18	0	2	2	169	1	163	24	74	36	8	
19	1	4	3	393	4	146	27	27	125	18	
20	2	2	5	118	11	164	63	72	37	3	
21	0	1	2	1	1	0	8	5	0	0	
22	9	28	0	4	0	0	1	1	0	21	
23	5	6	4	2	0	6	1	0	3	7	
24	1	2	6	7	0	4	0	1	42	157	
25	0	0	0	0	0	0	0	0	3	8	
26	1	1	1	0	0	0	0	0	1	1	
27	0	0	0	1	0	1	0	3	0	0	
28	0	33	4	0	1	0	3	0	2	0	

Joonis 9. Kultuuride ja klastrite risttabel ühe ruudu kohta. Punase kasti sees on illustreeritud kultuurigrupi põldude summa definitsioon. Arvud risttabelis näitavad põldude arvu.

2. kultuurigruppide ja klastrite risttabelist leitakse rea- ehk kultuurigrupi põldude summa (*sum_crop*) (vt Joonis 9);
3. leitakse arvuline piirväärtus, mida võib igast kultuurigrupist eemaldada ehk ekseks määrata. Antud juhul 10%, *ehk limit = 0.1*sum_crop*. Kui see piirväärtus on alla 1 põllu, siis ei saa ühtegi klastrit ega seal olevaid põlde ekseteks määrata;
4. antud kultuurigrupi klastrid sorteeritakse kasvavalt ja hakatakse neid alates väikseimast eksete hulka lisama, seni kuni määratud limiiti pole ületatud. Klastrite numbrid, mis mahuvad limiidi sisse, kogutakse kokku, leitakse nendesse kuuluvate põldude id-d ja määratakse ekseteks.
5. Ekseks määratud põllud kirjutatakse .csv faili, kust need loetakse RITA andmebaasi tabelisse *parcel_crop* veergu *clusteranalyses_verified = FALSE*.

Viited

Genolini, C., Alacoque, X., Sentenac, M., Arnaud, C. 2015. kml and kml3d: R packages to cluster longitudinal data. Journal of Statistical Software, 65 (4).

Euroopa Liit
Euroopa
Regionaalarengu Fond

Eesti
tuleviku heaks

D4.8 Põllukultuuride tuvastusmudeli kirjeldus koos täpsushinnangutega

4. iteratsioon (lõpparuanne) – sügis 2020

Koostasid: OÜ KappaZeta, Tartu Ülikool ja Eesti Maaülikool

Projekt RITA1/02-52 „Kaugseire andmete kasutuselevõtt avalike teenuste väljatöötamisel ja arendamisel“

Tartu 2020

Sisukord

Sissejuhatus

Uurimis- ja arendustöö on tehtud ja käesolev aruanne on valminud projekti RITA1/02-52 „Kaugseire andmete kasutuselevõtt avalike teenuste väljatöötamisel ja arendamisel“ raames.

Projekti eesmärgiks oli huvigruppide kokkuleppel valitud valdkondades läbi viia kaugseire andmete kasutamisvõimaluste analüüs ja uute, avaliku sektori asutuste tööd tõhusamaks muutvate rakenduste ning andmehalduse prototüüpide väljatöötamine ning nende kasutamise testimine pilootaladel.W

Üheks valdkonnaks oli põllmajandusmaade kasutuse seire Põllumajanduse Registrite ja Informatsiooni Ameti (PRIA) ning Maaeluministeeriumi vastutusalas (tööpakett nr 4, „Maa“). Selle tööpaketi uuringute ja arendustöö kitsamaks eesmärgiks oli Eesti oludesse sobiva kaugseire-põhise põllukultuuride tuvastamise metoodika välja töötamine.

Tööpaketi 4 („Maa“) raames on varem valminud järgmised tulemid, mida selles aruandes enam põhjalikult ei käsitleta:

a) teemakohase teaduskirjanduse ülevaade (tulem D4.1),
b) andmemudeli kirjeldus (tulem D4.1),
c) Eesti mullakaardi tüpoloogia üldistamine masinõppe jaoks (tulem D4.1),
d) toorandmestik (tulemid D4.2 ja D4.3)
e) metoodika kirjelduse varasemad iteratsioonid (tulemid D4.4, D4.5, D4.6)
f) avalik prototüüptarkvara testandmestikul hinnangute andmiseks, mis on alla laetav aadressilt: https://bitbucket.org/kappazeta/rita-evaluator/src/master/ (tulemid D4.7 ja D4.10)

Käesolev aruanne koos lisadega käsitleb järgmisi teemasid:

i) 2019.–2020. loodud põllukultuuride tuvastusmudeli arhitektuuri,
ii) sisendiks kasutatud 2018. a ja 2019. a suve satelliidipiltide ja Eesti põldude pealt arvutatud andmekomplekti,
iii) kultuurigruppide klassifikatsiooni,
iv) sisendandmetest eksete eemaldamise metoodikat,
v) mudelite tulemuste võrdlemise ja täpsuse hindamise metoodikat,
vi) andmetele tehtud eeltöötlust viimaks andmed masinõppe mudeli jaoks sobivale kujule,
vii) mudelite tulemusi ja täpsushinnanguid,
viii) tunnuste olulisuse hinnangut
ix) ja valideerimiseks läbiviidud katsetust alternatiivse masinõppe meetodiga (otsustuspuudemets).

Mudeli vigade analüüsi järel antakse kokkuvõttes soovitusi mudeli edasi arendamiseks ja täpsuse tõstmiseks.

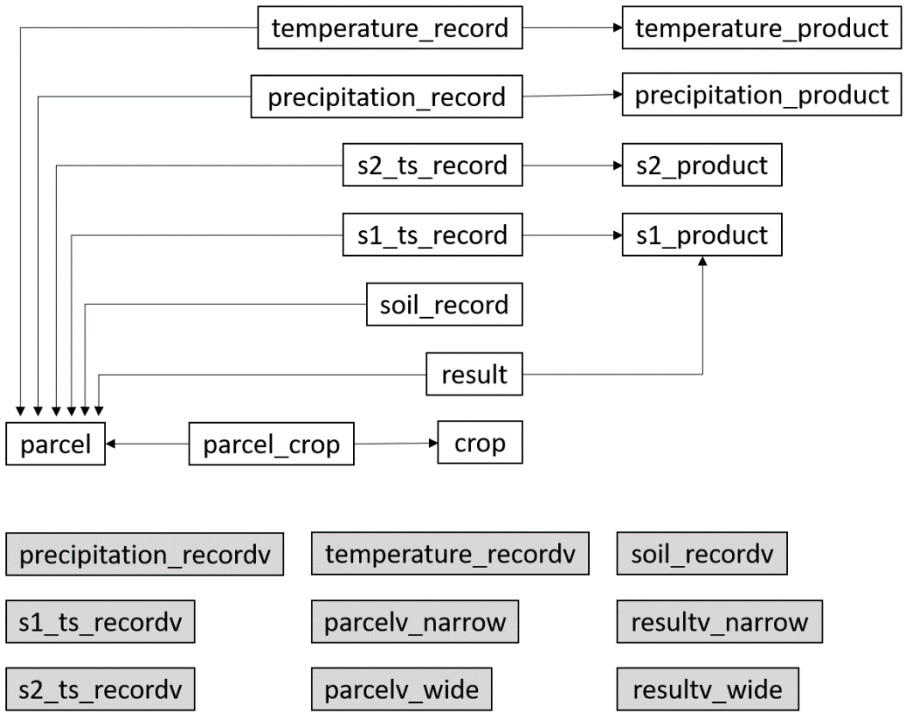
Dokument on mõeldud avalikuks kasutamiseks, et soodustada edasist arendustööd põllukultuuride automaatsel tuvastamisel ja anda suuniseid võimaliku operatiivsüsteemi loomiseks.

2. Lähteandmete ja tuvastusmudeli kirjeldus

2.1 Andmemudeli kirjeldus

Põhjalikum andmemudeli kirjeldus on toodud tulemis D4.1. Kuna uurimustöö jooksul on see mõnevõrra muutunud, siis toome siin uuesti välja andmebaasi lõpliku struktuuri.

Mõlema aasta andmed on hoiustatud erinevates, ent identse struktuuriga PostgreSQL andmebaasides. Andmebaasitabelite ja vaadete omavahelistest seostest annab ülevaate Joonis 10.



Joonis 10. Public skeema tabelite omavahelised seosed ja vaadete nimetused. Halli taustaga kastid tähistavad andmebaasi vaateid. Noole ots tähistab seda tabelit, millest välisvõti pärineb.

Võrreldes esialgse struktuuriga, lisasime tabeli „crop“, mis sisaldab erinevate klassifikatsioonide klasse ning tabeli „parcel_crop“, mis seob klasside info konkreetsete põldudega. Selline täiendus oli vajalik, et võimaldada kahe erineva klassifikatsiooni samaaegset kasutust ja loob võimaluse kasutada isegi rohkem kui kahte klassifikatsiooni, jättes ülejäänud andmebaasi struktuuri muutmata.

Eri klassifikatsioonide kasutuselevõtu tõttu tuli luua ka uusi andmebaasivaateid („*_narrow“, „*_wide“), et lihtustada andmete pärimist andmetöötluse järgmistes etappides.

Tabelis „parcel“ loodi uus veerg „datasource“, mis kirjeldab antud põllu sildi kvaliteeti:

- ’y’: vastab PRIA välikontrolli andmetele
- ’g’: geomeetria muudetud PRIA välikontrolli alusel
- ’n’: PRIA välikontrolli pole tehtud
- ’e’: ekse, ehk kultuur varieerub põllu piires PRIA välikontrolli andmetel liiga palju

Klasteranalüüsil leitud eksed (vt pt 3.2) on andmebaasi märgistatud tabelis „parcel_crop“, kus veerg *clusteranalyses_verified=False*. Mudeli sobitamisel kasutasime ainult põlde, mis polnud määratud ekseteks.

2.2 Lähteandmete kirjeldus

Tensori loomisel kaasatakse valiidsete põldude (piisavalt suured põllud, mille kohta on piisavalt „puhtaid“ Sentinel-1 SAR (S1) ja Sentinel-2 MSI (S2) piksleid peale põldude ruumikujude sisse poole puhverdamist) kõigi S1 ja S2 parameetrite pikslikomplektide ruumiliste mediaanide aegread, temperatuuri aegread, igale põllule vastav domineeriv mullatüüp, põldude normeeritud koordinaadid ja sademesummade aegread.

Eeltöötlusse kaasati järgnev tunnuskomplekt:

1) S1 VV-kanali 6-päeva koherentsus.
2) S1 VH-kanali 6-päeva koherentsus.
3) S1 VV-kanali tagasihajumine.
4) S1 VH-kanali tagasihajumine.
5) S1 VH/VV tagasihajumise suhe.
6) S2 sinine kanal (B2).
7) S2 roheline kanal (B3)
8) S2 punane kanal (B4)
9) S2 lähisinfrapuna kanal (B5)
10) S2 lähisinfrapuna kanal (B6)
11) S2 lähisinfrapuna kanal (B7)
12) S2 lähisinfrapuna kanal (B8)
13) S2 lähisinfrapuna kanal (B8a)
14) S2 lühilainelise infrapuna kanal (B11)
15) S2 lühilainelise infrapuna kanal (B12)
16) S2 NDVI indeks
17) S2 NDWI indeks
18) S2 NDVIre indeks
19) S2 TC_Wetness indeks
20) S2 TC_Vegetation indeks
21) S2 TC_Birghtness indeks
22) S2 Misra_Yellow_Vegetation indeks
23) S2 PSRI indeks
24) S2 WRI indeks

25) Mullatüüp
26) Sentinel-1 satelliidi ülelennule eelneva 3 tunni sademesumma (mm)
27) Sentinel-1 satelliidi ülelennule eelneva 24 tunni sademesumma (mm)
28) Päevakeskmine temperatuur (°C)
29) Põllu normeeritud asukoha koordinaadid

Vastav päring andmekomplekti valikuks oli:

include_columns = {
's1_ts': ['cohvh_median', 'cohvv_median', 'vhvv_median', 's0vh_median', 's0vv_median'],
's2_ts': ['b2_median', 'b3_median', 'b4_median', 'b5_median', 'b6_median', 'b7_median',
'b8_median', 'b8a_median', 'b11_median', 'b12_median', 'ndvi_median', 'ndwi_median',
'ndvire_median', 'tc_wetness_median', 'tc_vegetation_median', 'tc_brightness_median',
'misra_yellow_vegetation_median', 'psri_median', 'wri_median'],
'soil': ['soil_type'],
'precipitation': [],
'temperature': ['temperature_mean'],
'parcel': ['x_norm_loc', 'y_norm_loc']
}

Andmekomplektid koostati kahe aasta kohta: 2018 ja 2019. Kuna nende andmekomplektide tunnused ja parameetrid on identsed, siis järgnevalt kirjeldatakse neid kui ühte ja kuupäevade puhul kasutatakse üldistavat terminit „aasta“ .

Kõik aegrea tüüpi parameetrid on interpoleeritud 1-päevase sammuga vahemikku 2018-04-01…2018-09-01 ja 2019-04-01…2019-09-01. Kokku on aegreale tensoris reserveeritud 215 päeva (aasta-04-01…aasta-10-31), ent reaalselt on siiski kasutuses aasta-04-01…aasta-09-01 kuupäevadega piiratud vahemik ja ülejäänud päevad täidetakse tühjade väärtustega. Aegread piirati 1. septembriga mitmel põhjusel:

1. kui antud metoodikat kasutada operatiivsüsteemis näiteks taotluste kontrollimisel, siis hilisem ajaperiood pole enam eriti oluline,
2. enamus kultuure on selleks ajaks koristatud ning iseloomulikud signatuurid olemas, mistõttu pole mõtet järgnevate kuud andmetega mudeli mahtu ja arvutusvõimsuse vajadust kasvatada,

Temperatuuri väärtused on skaleeritud vahemikku 0–1, mis vastab tegelikule temperatuuride vahemikule: -20…+35 °C.

Sademesummade väärtused on skaleeritud vahemikku 0–1, mis vastab 3 h summade puhul tegelikule vahemikule 0–45 mm ja 24 h summade puhul vahemikule 0–135 mm.

Tingituna satelliidipiltide kaadripiiride ülekatvusest on S1 ja S2 parameetrite aegridade ajaline tihedus ebaühtlane. Sisendandmete eeltöötluse käigus eemaldati ülekattealadelt täiendavad aegreapunktid, et saada geograafilisest asukohast sõltumatu ühtlane S1 ja S2 aegridade ajaline tihedus.

Välja jäeti ka põllud, mille Sentinel-1 või Sentinel-2 aegreapunkte oli aasta peale vähem kui 10. Aastal 2018 oli selliseid põlde kokku 239, 2019. aastal 92.

2.3 Kultuurigruppide klassifikatsioon

Kasutasime kahte erinevat klassifikatsiooni, mis lepiti arenduse käigus PRIA-ga kokku: üldine (*wide*) ja detailne (*narrow*). Üldine jaotus koosneb 16-st kultuurigrupist ja detailsem 28-st (Joonis 11).

kood	nimi	tüüp
101	Aedmaasikas	wide
102	Heintaimed, kõrrelised	wide
103	Heintaimed, liblikõielised	wide
104	Kanep	wide
105	Kartul	wide
106	Kaunviljad	wide
107	Mais	wide
108	Marjapõõsad, viljapuud ja astelpaju	wide
109	Mustkesa	wide
110	Peakapsas	wide
111	Rühvelkultuurid	wide
112	Suviraps ja -rüps	wide
113	Suviteraviljad	wide
114	Taliraps ja -rüps	wide
115	Taliteraviljad	wide
116	Muu	wide

(16 rows)

kood	nimi	tüüp
1	Aedmaasikas	narrow
2	Heintaimed, kõrrelised	narrow
3	Liblikõieliste segud (alla 80%)	narrow
4	Liblikõielised (üle 80%)	narrow
5	Kanep	narrow
6	Kartul	narrow
7	Põldhernes	narrow
8	Põlduba	narrow
9	Muu kaunvili	narrow
10	Mais	narrow
11	Astelpaju	narrow
12	Marjapõõsad ja viljapuud	narrow
13	Mustkesa	narrow
14	Peakapsas	narrow
15	Porgand	narrow
16	Punapeet	narrow
17	Suviraps ja -rüps	narrow
18	Suvinisu ja speltanisu	narrow
19	Suvioder	narrow
20	Kaer	narrow
21	Tatar	narrow
22	Taliraps ja -rüps	narrow
23	Rukis	narrow
24	Talinisu	narrow
25	Talioder	narrow
26	Talitritikale	narrow
27	Muu teravili	narrow
28	Muu	narrow

(28 rows)

Joonis 11. Kultuuride üldine (vasakul) ja detailne (paremal) klassifikatsioon.

2.4 Andmestiku jaotamine kogudesse

Õpetusandmed jagati treening-, valideerimis-, ja testandmestikuks. Treeningandmestiku järgi toimus mudeli sobitamine, valideerimisandmestiku järgi mudeli parameetrite komplekti valimine ning isoleeritud ja fikseeritud testandmestiku järgi täpsuse hindamine.

Kogu andmestik klasside kaupa on kirjeldatud Tabel 3 ja Tabel 4.

Nagu tabelitest näha, on põldude arv eri klassides väga erinevad – mõnes klassis võib olla üle 80 000 isendi, samas kui teises alla 100. Suur enamus kõigist meie andmekogus olevatest põldudest kuulub klassi 2, ehk kõrreliste heintaimede hulka. Sellisel kujul andmete kasutamine närvivõrkudes võib viia tasakaalust väljas klassifikatsioonini (*imbalanced classification problem*), kus isendite jaotus klasside lõikes on kallutatud, ning seetõttu ka klassifikatsiooni tulemus.

Seda probleemi saab lahendada kolmel moel:

1) Koguda rohkem andmeid alaesindatud klassidesse.

2) Valida mudeli sobitamisel õige meetrik, mille põhjal otsuseid teha. Üldõigsus (*accuracy*) ei ole tasakaalust väljas andmekogu puhul väga adekvaatne, vaid pigem isegi eksitav.
3) Andmestiku ümberjaotamine (*resampling*), sealhulgas andmete sünteetiline rikastamine ehk laiendamine (*data augmentation*).

Mudeli arenduse alguses (varasemad iteratsioonid) eirasime klasside vahelise tasakaalu probleemi ning jagasime kogu andmestiku treening-, valideerimis- ja testandmestikuks 90%/5%/5% osades.

Lõplike mudelite puhul on aga seda probleemi lahendada püütud, katsetades eelpool mainitud võimalustest kahte viimast. Rohkem õpetusandmeid meil koguda polnud võimalik. Meetrikute teemat käsitletakse lähemalt peatükis 4.1.

Tabel 3. Põldude arv erinevates klassides detailse klassifikatsiooni puhul. Lisaks on lisatud kui suure osa (%) moodustavad selle klassi näidised kogu andmestikust.

Kultuur		2018+2019	
Kood	Nimi	kõik näidised	%
1	Aedmaasikas	334	0,15
2	Heintaimed, kõrrelised	95328	42,47
3	Liblikõieliste segud (alla 80%)	23761	10,59
4	Liblikõielised (üle 80%)	8168	3,64
5	Kanep	640	0,29
6	Kartul	797	0,36
7	Põldhernes	6050	2,70
8	Põlduba	2765	1,23
9	Muu kaunvili	433	0,19
10	Mais	1428	0,64
11	Astelpaju	407	0,18
12	Marjapõõsad ja viljapuud	585	0,26
13	Mustkesa	636	0,28
14	Peakapsas	80	0,04
15	Porgand	77	0,03
16	Punapeet	71	0,03
17	Suviraps ja -rüps	5600	2,50
18	Suvinisu ja speltanisu	13040	5,81
19	Suvioder	22188	9,89
20	Kaer	10423	4,64
21	Tatar	702	0,31
22	Taliraps ja -rüps	6177	2,75
23	Rukis	3882	1,73
24	Talinisu	15647	6,97
25	Talioder	2105	0,94
26	Talitritikale	747	0,33
27	Muu teravili	208	0,09
28	Muu	2169	0,97

Summa	224448

Tabel 4. Põldude arv erinevates klassides üldise klassifikatsiooni puhul. Lisaks on lisatud kui suure osa (%) moodustavad selle klassi näidised kogu andmestikust.

Kultuur		2018+2019	
Kood	Nimi	kõik näidised	%
101	Aedmaasikas	334	0,15
102	Heintaimed, kõrrelised	95334	42,50
103	Heintaimed, liblikõielised	31615	14,10
104	Kanep	640	0,29
105	Kartul	796	0,35
106	Kaunviljad	9235	4,12
107	Mais	1429	0,64
108	Marjapõõsad, viljapuud ja astelpaju	988	0,44
109	Mustkesa	634	0,28
110	Peakapsas	80	0,04
111	Rühvelkultuurid	176	0,08
112	Suviraps ja -rüps	5597	2,50
113	Suviteraviljad	46713	20,83
114	Taliraps ja -rüps	6183	2,76
115	Taliteraviljad	22308	9,95
116	Muu	2229	0,99
Summa		224291	

2.4.1 Klasside tasakaalustamine ja andmete rikastamine

Klasside tasakaalustamiseks saab arvukamaid klasse kärpida näidiste kustutamisega (*under-sampling*) ja vähemarvukamaid suurendada olemasolevate näidiste kopeerimisega (*over-sampling*). Määrasime klassi näidiste maksimumhulgaks 12 000, ehk klassides, kus oli rohkem põlde, neid lihtsalt treeningandmestikku ei kaasatud. Väiksematesse klassidesse lõime juurde tehislikke aegridu, võttes olemasolevate näidiste aegread ja lisades neile juhuslikku müra. Sellisel viisil kasvatasime väikeste klasside näidiste arvu treeningkogus võrdseks suuremate klasside omale (umbes 9000 põldu igas klassis, väiksed erinevused tulenevad sünteetiliste

näidiste loomise loogikast, mille käigus tekitatakse korrutisi olemasolevatest näidistest) ja viisime klassid enam-vähem tasakaalu. Seejärel jaotasime andmestiku treening-, valideerimis- ja testkoguks 75/20/5% osades. Valideerimis- ja testandmestikku sünteetilisi näidiseid ei kaastatud. Kui mingis klassis oli põlde vähem kui 150, siis need jagati treening-, valideerimis- ja testandmestikuks 40%/30%/30% osades. Sellist lisatingimust rakendati eesmärgil, et ka väiksemates kultuurigruppides satuks valideerimis- ja testkogusse piisavalt põlde.

Lõplikke mudeli sobitamisel kasutatud andmekogusid vt Tabel 5 ja Tabel 6.

Tabel 5. Erinevate õpetuskogude suurused detailse klassifikatsiooni puhul. Arvukamaid klasse (üle 12 000 näidise) on kärbitud kõigis kogudes ja väiksemate klasside treeningkogudesse on loodud sünteetilisi näidiseid.

Kultuur		2018+2019		
Kood	Nimi	train	val	test
1	Aedmaasikas	8820	67	16
2	Heintaimed, kõrrelised	9000	2400	600
3	Liblikõieliste segud (alla 80%)	9000	2400	600
4	Liblikõielised (üle 80%)	6126	1633	408
5	Kanep	8640	128	32
6	Kartul	8985	159	39
7	Põldhernes	9078	1210	302
8	Põlduba	8296	553	138
9	Muu kaunvili	8802	87	21
10	Mais	8568	285	71
11	Astelpaju	8874	81	20
12	Marjapõõsad ja viljapuud	8780	117	29
13	Mustkesa	8622	127	31
14	Peakapsas	8992	24	24
15	Porgand	8711	23	23
16	Punapeet	8149	21	21
17	Suviraps ja -rüps	8400	1120	280
18	Suvinisu ja speltanisu	9000	2400	600
19	Suvioder	9000	2400	600
20	Kaer	7818	2084	521

21	Tatar	8432	140	35
22	Taliraps ja -rüps	9266	1235	308
23	Rukis	8733	776	194
24	Talinisu	9000	2400	600
25	Talioder	7895	421	105
26	Talitritikale	8976	149	37
27	Muu teravili	8892	41	10
28	Muu	8135	434	108
Summa		240 990	22 915	5773
Kogu andmestik		252 536		

Tabel 6. Erinevate õpetuskogude suurused üldise klassifikatsiooni puhul. Arvukamaid klasse (üle 12 000 näidise) on kärbitud kõigis kogudes ja väiksemate klasside treeningkogudesse on loodud sünteetilisi näidiseid.

Kultuur		2018+2019		
Kood	Nimi	train	val	test
101	Aedmaasikas	8316	67	16
102	Heintaimed, kõrrelised	9000	2400	600
103	Heintaimed, liblikõielised	9000	2400	600
104	Kanep	7680	128	32
105	Kartul	7774	159	39
106	Kaunviljad	6927	1847	461
107	Mais	7511	286	71
108	Marjapõõsad, viljapuud ja astelpaju	8151	197	49
109	Mustkesa	7632	127	31
110	Peakapsas	9632	24	24
111	Rühvelkultuurid	7980	35	8
112	Suviraps ja -rüps	8398	1119	279
113	Suviteraviljad	9000	2400	600
114	Taliraps ja -rüps	9274	1236	309
115	Taliteraviljad	9000	2400	600
116	Muu	8360	446	111

Summa	133 635	15 271	3830
Kogu andmestik	1526		

2.5 Tuvastusmudeli kirjeldus

Põllukultuuride tuvastusmudeliks on sügavõppe mudel, mis töötab satelliidipiltidelt arvutatud parameetrite põldude aegridadel (vt täpsemalt dokumendist D4.1 Kirjanduse ülevaade ja esialgne põllukultuuride tuvastusmudeli kirjeldus, lk 11-17).

3 Andmetöötlus ja mudeli sobitamine

3.1 Toorandmete töötlus

S1 aegread arvutati PRIA SATIKA tarkvaraga ja S2 aegread Calvaluse töötlusahelaga ESTHUBi keskkonnas.

S2 arvutustulemusi filtreeriti järgmiste tingimuste põhjal:

- *numPasses* > 0 (ülelendude arv antud ajavahemikus (ühe päeva jooksul))
- *ndvi_numValid* > 5 (kasutatavate pikslite hulk põllu piires)
- *cloud_flag_arithMean* < 0.5 (pilve osakaal põllu geomeetria ulatuses)

Et vähendada pilvede poolt rikutud aegrea punktide hulka ja muid anomaaliaid:

1) kustutati aegreapunktid, kus *ndvi_mean* < 0,12. Piirväärtuse määramisel võeti aluseks 9 päeva kõigi põldude *ndvi_mean* väärtuste histogrammid. Valitud päevade hulgas oli nii poolpilves kui selgeid päevi ja vähemalt üks pilt igast kuust vahemikus aprill-oktoober. Histogrammide jaotuse ja S2 piltide reaalsete pikslivääruste võrdlemisel leiti, et *ndvi_mean* väärtused alla 0,12 on tõenäoliselt pilvede poolt mõjutatud. Samas tuleb teadvustada, et selle piirväärtuse rakendamisega võivad kevadisest perioodist olla kaduma läinud ka mõned pilvevabade küntud põldude aegreapunktid, kus NDVI väärtused pistelisel kontrollil olid väga madalad.
2) kustutati aegreapunktid, kus *ndvi_stdv* > 0,17. Piirväärtuse määramisel võeti aluseks 9 päeva kõigi põldude *ndvi_stdv* väärtuste histogrammid. Valitud päevade hulgas oli nii poolpilves kui selgeid päevi ja vähemalt üks pilt igast kuust vahemikus aprill-oktoober.
3) S1 aegreapunktide hulgast kustutati need, mille *vhvv_median* > 1. Nii kõrge VH/VV tagasihajumise suhe viitab tehisobjekti (nt elektriliin) tagasipeegelduse küllastusele antud põllu piires ja järelikult sellised andmed põllukultuuri kirjeldamiseks ei sobi.

Normeeritud põldude asukoht on arvutatud järgmise ümbriku järgi (ühikuks on meetrid):

	X	Y
MAX	740000	6615000
MIN	370000	6377000
vahe	370000	238000

Asukoht_X=(point_X - 370000)/370000

Asukoht_Y=(point_Y - 6377000)/238000

3.2 Sisendandmetest eksete eemaldamine

Kõikide PRIA põldude kohta on taotlusel märgitud, mis seal kasvama peaks. Enamasti on see õige, aga PRIA hinnangul on taotlejate esitatud andmetes <5% vigu. Selleks, et põldude andmekogu saaks kasutada mudeli arenduseks, on vaja sealt võimalikult palju vigadega kirjeid (eksed) eemaldada. Eksete eemaldamine tehti satelliidandmete tunnuskomplekti põhjal, kasutades järgmisi parameetreid:

- Sentinel-2 MSI kanal 11 (b11_mean)
- Sentinel-2 MSI andmetest arvutatud NDVI indeks (ndvi_mean)
- Sentinel-1 SAR andmetest arvutatud polarisatsioonide suhe (vhvv_mean)
- Sentinel-1 SAR andmetest arvutatud koherents (cohvv_mean).

Põllud jagati geograafiliselt 16 rühma, mille sees klasterdati üleval nimetatud parameetrite aegridade alusel 10 klastrisse. Seejärel määrati ekseteks ja eemaldati õpetusandmetest iga kultuurigrupi piires väikseimad klastrid (ja klastritesse kuuluvad põllud), mis moodustasid kokku kuni 10% kogu selle kultuurigrupi põldudest. Täpsem eksete eemaldamise metoodika on leitav eraldi dokumendist (Põllukultuuride D4.5 Lisa 1, 2020).

Antud metoodikaga eemaldatud põldude arvud aastate ja klassifikatsiooni kaupa:

Aasta	Detailne klassifikatsioon (*narrow*)	Jäme klassifikatsioon (*wide*)
2018	7098	7078
2019	7131	7350

Eksete eemaldamise tõhusust kontrolliti enne ja pärast eemaldamist õpetatud mudelite (ainult S1 ja S2 parameetridega detailse klassifikatsiooni peal) eksimismaatrikiste võrdlemise teel (vt metoodikat pt 9). Võrreldi nii täpsust valideerimiskogul (Tabel 7) kui ka vähemolulistest kultuuridest puhastatud eksimismaatrikseid Kappa analüüsi teel (Tabel 8).

Tabel 7. Valideerimiskogu täpsushinnangud enne ja pärast eksete eemaldamist. Mudeleid rakendati tingumusega „sampling_thresholds = [150, 5000]“, mis tähendab, et igast kultuurigrupist võeti õpetuskogusse maksimaalselt 5000 põldu.

Aasta	2018		2019	
Mudel	eksed sees	eksed eemaldatud	eksed sees	eksed eemaldatud
Val acc	0,7996	0,8137	0,8392	0,8619
Val loss	0,7941	0,6947	0,5889	0,5244

Tabel 8. Kärbitud eksimismaatriksite erinevuse hindamine Kappa meetodil. Kui kahest eksimismaatriksist arvutatud Z-i väärtus on suurem kui 1,96, võib järeldada ($p < 0,05$), et võrreldud eksimismaatriksid on tõepoolest oluliselt erinevad ja see pole tingitud vaid juhuslikkusest.

Aasta	Enne		Pärast		Z statistik
	Maatriksi kirjeldus	KHAT	Maatriksi kirjeldus	KHAT	
2018	Eksimismaatriks, eksed sees	0,8040	Eksimismaatriks, eksed eemaldatud	0,8353	2,7606
2019	Eksimismaatriks, eksed sees	0,8464	Eksimismaatriks, eksed eemaldatud	0,8715	2,4890

3.3 Mustkesa klassi puhastamine

Eelnevate mudeli iteratsioonide ja mustkesa klassi umbmäärase definitsiooni põhjal võis eeldada, et selles klassis esineb olulisel hulgal põlde, mis sinna tegelikult ei peaks kuuluma.

Seetõttu vaadati käsitsi läbi kõik *parcelv_narrow* tabelis olevad põllud, mille kultuuriks on märgitud „mustkesa“ ja märgiti ära, milliste põldude aegread (NDVI + coh) ei vasta mustakesa definitsioonile. Mustakesa määratlus on kohati küll segane, aga lähtuti põhimõttest, et enne juuni keskpaika peab olema põllu pind kas ümberpööratud (kündmisjälg) või taimestikku väga vähe (NDVI väga madal), ning nö mustana on põldu hoitud vähemalt kuni augusti lõpuni. PRIA definitsioon: „*Mustkesa on kesa, kus vegetatsiooniperioodil toimub maa ettevalmistamine järgnevate kultuuride külviks ning eelkultuure ei kasvatata. Mustkesa tuleb harida soovitavalt iga 2-3 nädala tagant, et umbrohutaimed ei kasvaks seal kõrgemaks kui 5 cm. Erandjuhul tulenevalt ilmastikust aktsepteeritakse mustkesana maad, millel umbrohtude kõrgus ei ületa 20 cm ja taimed ei ole jõudnud õitsemisfaasi.*“ (Jaanus Ainso, e-kiri, 05.11.2019).

Piiripealseid ja vaieldavaid juhte oli palju, kuid märkisime andmekogus ekseteks suure hulga põldusid:

Aasta	Kontrollitud põlde	Eemaldatud	Kasutatavaid
2018	753	365	388
2019	449	199	250

Pärast mudeli (ainult detailse klassifikatsiooni puhul) uuesti treenimist puhastatud mustakesa põldudega, oli märgata mõningast õigsuse paranemist mustkesa klassis:

Aasta	Mustakesa klassifitseerimisõigsus enne	Mustakesa klassifitseerimisõigsus pärast
2018	0.65	0.74
2019	0.41	0.67

Mõningal määral muutusid ka õigsused teistes klassides. See oli oodatav, kuna mudelit sobitate muutunud andmestikuga.

3.4 Mudeli arhitektuur ja hüperparameetrid

Mudel koosneb sisendkihist, sellele järgnevast tensori mõõtmeid vähendavast kihist (*flatten layer*) ja kahest täissidusast vahekihist (*dense)*, mis on omavahel täielikult ühendatud (i k *fully connected layers*). Mõlema täissidusa kihi järel on on lisatud ka andmeid normeeriv kiht (*batch normalization layer*), mis vähendab mudeli ülesobitamise ohtu. Esimeses täissidusas kihis kasutati aktivatsiooniks mittenegatiivset lineaarfunktsiooni (*Rectified Linear Unit*), teises sigmoid funktsioon ja väljundkihis *softmax* aktivatsioonifunktsiooni, mis annab igale põllule tõenäosused kõigi võimalike nimekirjas olevate kultuurigruppide kohta (28) nii, et tõenäosuste summa on 100%. Mudeli arhitektuuri ja kihtide mõõtmeid vt Joonis 12.

Katsetati ka sisendkihi järele ühemõõtmelise sidumkihi (1D CNN) lisamist, mis andis 2019. aasta puhul küll 0,006 valideerimisõigsuse (*validation accuracy*) kasvu, ent samas kasvas oluliselt ka kadu valideerimisandmestikul (0,3834 → 0,7217). Kombineeritud kahe aasta andmestiku puhul aga ei paranenud ka valideerimisõigsus. Seetõttu otsustati mitte kaasata 1D CNN kihti mudeli arhitektuuri.

Samuti ei parandanud ResNet (residual neural network) arhitektuur märgatavalt mudeli meetrikuid ja tulemused olid sarnased või veidi kehvemad. Samas kasvas treeningule kuluv aeg ligi 30%, sest ResNet närvivõrk on keerulisem. Kahe lineaarse närvivõrguga saavutame antud juhul sama tulemuse kiiremini, mistõttu eelistasime seda. Suurema arvu näidise puhul (andmed rohkematest aastatest) võib ResNet arhitektuur tulevikus siiski paremini sobida.

Layer (type)	Output Shape	Param #
main_input (InputLayer)	[(None, 153, 25)]	0
flatten (Flatten)	(None, 3825)	0
dense1 (Dense)	(None, 3825)	14634450
batch_normalization (BatchNo	(None, 3825)	15300
dense2 (Dense)	(None, 256)	979456
batch_normalization_1 (Batch	(None, 256)	1024
outcome_output (Dense)	(None, 28)	7196

Total params: 15,637,426
Trainable params: 15,629,264
Non-trainable params: 8,162

Joonis 12. Tuvastusmudeli arhitektuur. Antud juhul on näide 2019. aasta detailsema klassifikatsiooniga.

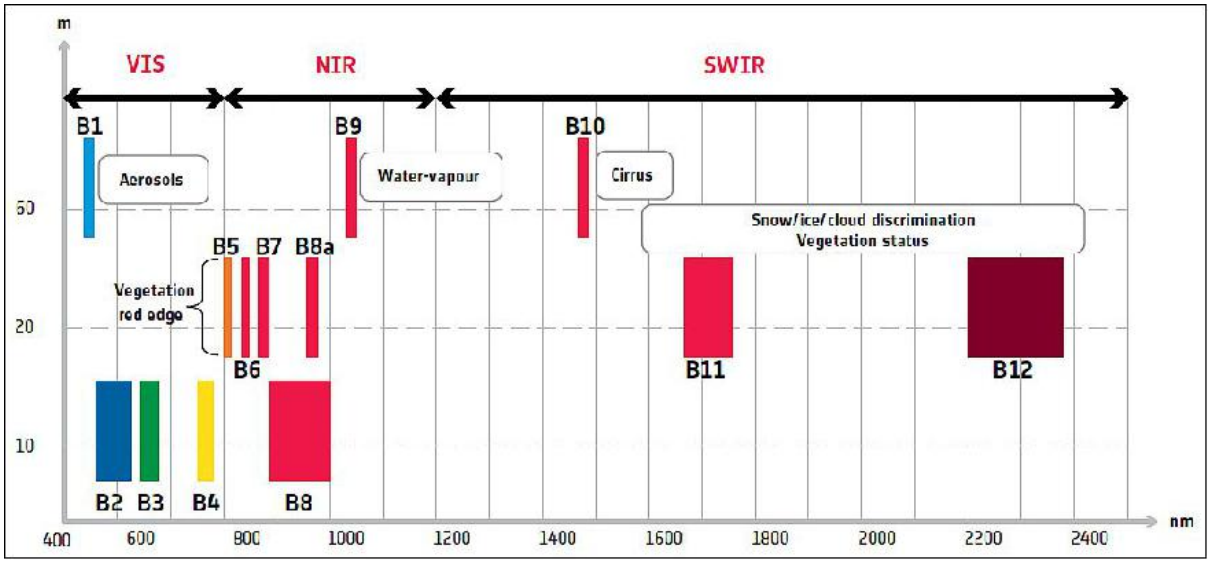
Mudeli hüperparameetrid olid järgmised:

- kao funktsioon (*loss function*) = 'categorical_crossentropy'
- monitooritud meetrik (*monitored metric*) = 'val_custom_f1'
- õpisamm (*learning rate*) = 0.0001
- ploki suurus (*batch size*) = 32
- epohhide arv (*epochs*) = 300
- varase lõpetamise lävi (*early stop patience*) = 10
- õpisammu vähendamise kordaja (*reduce factor*) = 0.2
- õpisammu vähendamise lävi (*reduce patience*) = 10

Erinevate parameetritega mudelite treenimisel jõuti kõige parema tulemuseni järgmise komplektiga (S1 parameetrid sinisega, S2 parameetrid punasega, ilmastiku ja asukoha parameetrid mustaga ning halli värviga on indikaatorveerud): "parcel_id", "s1_date", "cohvv_median", "cohvh_median", "vhvv_median", "s0vv_median", "s0vh_median", "b3_median", "b4_median", "b6_median", "b8a_median", "b11_median", "b12_median", "ndvi_median", "ndwi_median", "ndvire_median", "tc_vegetation_median", "misra_yellow_vegetation_median", "psri_median", "wri_median", "temperature_mean", "precipitation_sum_mean_3h", "precipitation_sum_3h", "precipitation_sum_mean_24h", "precipitation_sum_24h", "x_norm_loc", "y_norm_loc".

Sentinel-2 spektrikanalite valikul on jäetud välja valitute suhtes mõned väga lähedased naaberkanalid, mille andmed tõenäoliselt väga palju uut infot ei lisanuks (vt Joonis 13). Sentinel-2 kanalite põhjal arvutatud indeksitest jäeti välja *tc_wetness* ja *tc_brightness*. Sademete puhul on andmekomplektis ka indikaatorveerud (halli värvi), mis on mudeli sisendandmete eripära.

Väikse tunnustekomplektiga (NDVI vs NDVI+mullatüüp) katsetuste käigus selgus, et teatud klassifikatsiooniga (agrorühmitamise alusel loodud 3 erinevat üldistustaset ehk *soil_type, soil_type2* ja *soil_type3*) mullaandmete lisamisel muutus nii valideerimiskogu täpsus kui ka kärbitud eksimismaatriksi KHAT statistik oluliselt kehvemaks. Samas muldade ordineeritud maatriksi, kus telgeteks litoloogiline-geneetiline ning üldistatud niiskuse telg, põhjal loodud klassifikatsioon (*soil_type4*) näitas sama katse puhul olulist mudeli täpsuse suurenemist (Z-statistik 4,73). Suurema tunnuskomplektiga mudelile mullatüübi (*soil_type4*) lisamine olulist täpsuse suurenemist ei andnud (Z<1,96). Seetõttu pole tõenäoliselt ka operatiivteenuse jaoks tunnuskomplekti valides mullainfo kaasamine oluline.



Joonis 13. Sentinel-2 kanalite jaotus ruumilise lahutuse ja lainepikkuse järgi (allikas: https://directory.eoportal.org/web/eoportal/satellite-missions/c-missions/copernicus-sentinel-2).

3.5 Kasutatud tarkvara ja riistvara

Arvuti riistvara:

Protsessor: Intel® Core™ i7-8700K (6 tuuma, 12 lõime)

Graafikakaart: NVIDIA Corporation GP104 [GeForce GTX 1070]

Muutmälu: 64 GiB

Kasutatud tarkvarapaketid:

Python 3.7.3

Pandas 0.25.3

Keras 2.3.1

Tensorflow-GPU 2.2.0

Matplotlib 3.1.2

4 Mudeli sobitamise tulemused ja arutelu

4.1 Mudeli täpsuse ja sobivuse hindamine

Mudeli täpsuse ja sobivuse hindamiseks kasutatakse klassikaliselt järgmisi meetrikuid:

- **täpsus** (*precision*) = konkreetse klassi õigesti klassifitseeritud näidised / kõik selleks klassiks klassifitseeritud nädised (TP/(TP + FP))*
- **saagis** (*recall*) = konkreetse klassi õigesti klassifitseeritud näidised / kõik selle klassi tegelikud näidised (TP/(TP + FN))
- **F1 skoor** (*F1 Score*) = 2 × (täpsus × saagis)/(täpsus + saagis). Kombineerib täpsuse ja saagise üheks meetrikuks.
- **õigsus** (*accuracy*) = õigesti klassifitseeritud põldude arv / kõigi põldude arv ((TP + TN)/(TP + TN + FP + FN)
- **kadu** (*loss*)

* TP – True Positive; FP – False Positive; TN – True Negative; FN – False Negative

Kadu iseloomustab seda, kui halb mudeli ennustus on üksiku näidise puhul. Kui mudel on perfektne, siis kadu on 0, vastupidisel juhul kadu kasvab. Mudeli treenimise eesmärgiks on leida sellised kaalud, mille puhul keskmine kadu üle kõigi näidiste on madal. Tegemist ei ole protsendilise näitajaga, vaid antud kogu kõigi näidiste puhul tehtud vigade summaga. Funktsioone kao arvutamiseks on mitmeid, antud juhul on mudeli sobitamisel kasutatud *categorical crossentropy* kaofunktsiooni.

Kui püüame eelneva panna põllukultuuride konteksti, võiks näide olla järgmine:

Täpsus näitab, kui palju oli tegelikult kartulipõlde kõigi mudeli poolt kartulipõlluks klassifitseeritud põldude seas. Juhtub, et mudel klassifitseerib ka näiteks mõned peedi või porgandipõllud kartuliks (valepositiivsed).

Saagis näitab, kui suure osa tegelikest kartulipõldudest mudel üles leidis. Juhtub, et vahel mudel määrab kartulipõllu hoopis peedipõlluks (valenegatiivsed).

Meetrikute arvutuskäigud on illustreeritud Tabel 9, mis kujutab hüpoteetilist eksimismaatriksit nelja kultuuriga klassifikatsioonile, kus andmestikus oli igat kultuuri 10 põldu.

Tabel 9. Meetrikute arvutamise põhimõtted eksimismaatriksil. Maatriksi ridadel on kultuuride tegelikud sildid ja tulpades ennustatud sildid.

Ennustatud *(predicted)*

Tegelik *(true)*		Kultuur 1	Kultuur 2	Kultuur 3	Kultuur 4	**Täpsus** (precision)		Saagis (recall)		**Õigsus** (accuracy)	F1 skoor (F1 score)
	Kultuur 1	9	1	0	0	9/(9+0+0+1)=	0,90	9/(9+1+0+0)=	0,90	0,95	0,90
	Kultuur2	0	7	3	0	7/(1+7+2+1)=	0,64	7/(0+7+3+0)=	0,70	0,83	0,67
	Kultuur 3	0	2	7	1	7/(0+3+7+3)=	0,54	7/(0+2+7+1)=	0,70	0,78	0,61
	Kultuur 4	1	1	3	5	5/(0+0+1+5)=	0,83	5/(1+1+3+5)=	0,50	0,85	0,63

Millist meetrikut jälgida mudeli soorituse hindamisel? Sellele küsimusele pole ühest vastust ning meetriku valik sõltub suuresti klassifitseerimisülesande ja andmestiku iseloomust. Tasakaalust väljas klassidega andmestiku puhul (nagu meie põllukultuuride andmestik on), ei ole üldõigsus kindlasti kõige parem näitaja, mida kasutada parima mudeli valimisel. Mingitel juhtudel, kui on väga arvukad ja domineerivad klassid (nagu meie puhul näiteks heintaimed), võib pelgalt õigsuse meetriku põhjal tehtud mudeli valikud olla isegi eksitavad.

Siin projektis otsustasime kasutada F1 skoori parima mudeli valimisel. See meetrik kombineerib täpsuse ja saagise. Mudeli treenimise etapis arvutatakse pärast iga epohhi keskmistatud ja kaalutud F1 skoor üle kõigi kultuurigruppide ja parim mudel valitakse just selle meetriku põhjal. Kaalumine tehakse valideerimiskogu põldude arvuga igas klassis. Sel viisil ei diskrimineeri ega unusta mudel ka vähemarvukaid klasse ja nende omadusi.

Eksimismaatriksid (*confusion matrices*) on oluliseks täienduseks hindamaks mudeli sobivust eri kultuurigruppide lõikes. Samuti on eksimismaatriksi tabelitest võimalik saada infot selle kohta, milliseid klasse omavahel segamini aetakse. Väärtused eksimismaatriksi peadiagonaalil näitavad iga klassi klassifitseerimise saagist, ehk antud klassis õigesti klassifitseeritud põldude arvu jagatist kõigi selle klassi põldude arvuga testkogus.

Näide: klassis „Kartul“ on testkogus 39 põldu (ehk neil on küljes tõene silt „Kartul“). Mudel, mis tõesest sildist ei tea midagi, klassifitseerib need põllud järgnevalt: ühe määrab klassi

„Liblikõieliste segud (alla 80%)“ ja 38 tükki klassi „Kartul“. Normeeritud eksimismaatriksil kajastub see tulemus selliselt, et kartuli tegelike siltide real on väärtused 0,03 (1/39=0,03) ja 0,97 (38/39=0,97). Esimene neist siis ennustatud siltide liblikõieliste klassi tulbas ja teine kartuli tulbas, ehk peadiagonaalil. Meie eksimismaatriksite puhul peadiagonaalil esitatud hinnanguid tuntakse ka kui antud klassi tegijatäpsus (Producer’s accuracy).

Vähendamaks eksimismaatriksite visuaalse võrdlemise subjektiivsust kasutasime mudeli arenduse varajases faasis ka Kappa statistilise analüüsi meetodit, mis võimaldab hinnata kahe eksimismaatriksi erinevust ja selle erinevuse statistilist usaldusväärsust (Bishop jt 1975 viidatud Congalton ja Green, 2008 järgi). Sel viisil saab erineva tunnuskomplektiga treenitud mudelite eksimismaatrikseid võrrelda ja hinnata, kas mingi tunnuse lisamine muutis tulemust paremaks või halvemaks.

Eksimismaatriksite võrdlemiseks ja vajalike statistikute arvutamiseks loodi spetsiaalne Python skript, milles rakendatud matemaatilised ja statistilised arvutused lähtusid Congaltoni ja Greeni (2008) raamatus kirjeldatud loogikast ja valemitest.

Järgnevalt on loetletud erinevad sammud, mida skript teeb:

1. mudeli algupärasest eksimismaatriksist eraldatakse 8 vähe levinud kultuurigruppi, mida soovisime maatriksite võrdlemisel mitte arvestada;
2. saadud alam-maatriksile arvutatakse kaks statistilist näitajat:
    - K – KHAT, näitab kui palju erineb maatriks juhuslikkusest (0 väärtus)
    - var(K) – KHATi varieeruvus, vajalik Z väärtuse arvutamiseks;
3. kahele erinevale eksimismaatriksile tehakse võrdlustest, kasutades eelmises punktis arvutatud statistikuid ja leitakse Z väärtus, mis näitab, kas nende maatriksite erinevus on statistiliselt oluline. Kui Z väärtus on suurem kui 1,96, võib järeldada ($p < 0,05$), et võrreldud eksimismaatriksid on tõepoolest oluliselt erinevad ja erinevus pole tingitud vaid juhuslikkusest;
4. väljastatakse uus eksimismaatriks koos arvutatud statistiliste näitajatega.

Lõplike tulemuste hindamisel jäime F1 skoori, saagise ja täpsuse juurde, ning eelpool kirjeldatud Kappa statistilise analüüsi meetodit enam ei kasutanud.

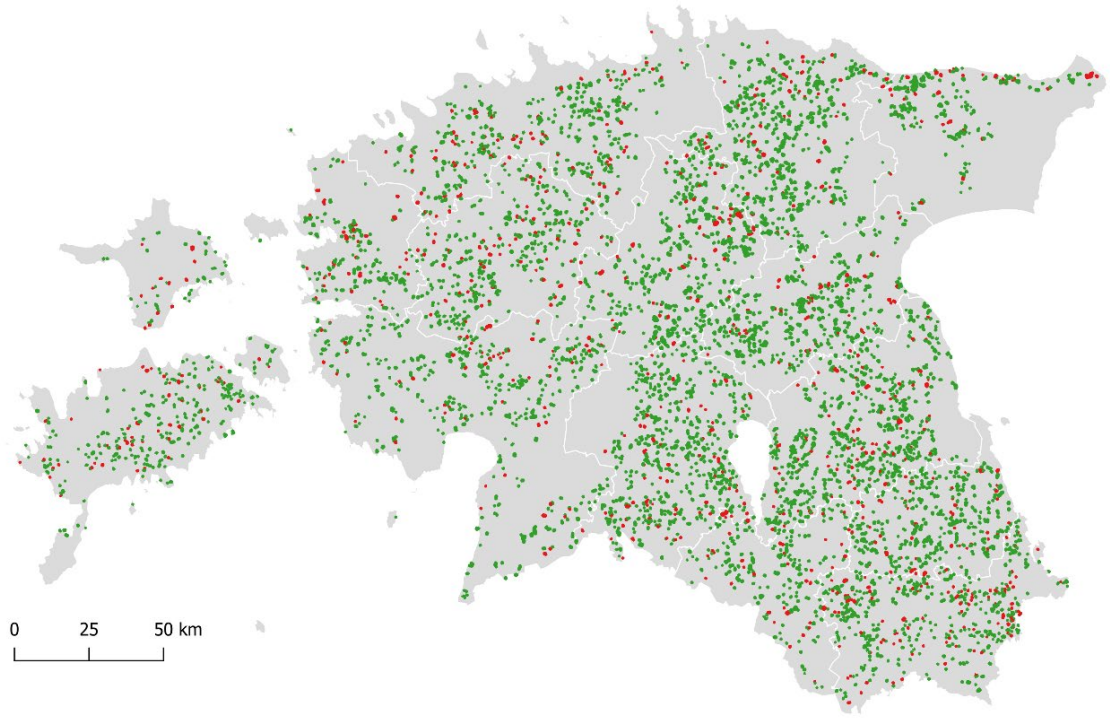
4.2 Tulemused kahe aasta andmestikul

Mudeli arendamise varasemates iteratsioonides sobitasime mudeleid eraldi 2018. ja 2019. aasta andmetel. Seejärel aga liitsime andmestikud ja lõplikus andmekogus on põllud nii 2018. kui 2019. aastast. Mudelid, mille meetrikuid ja täpsushinnanguid järgnevalt kirjeldame, on sobitatud just selle koondandmestiku peal. Usume, et kaasates andmestikku näidiseid erinevatest aastatest (erinevad ilmastikuolud), suudame luua universaalsema ja päris maailmas paremini kasutatava mudeli, isegi kui üldistatud täpsushinnangud mõnevõrra halvenevad. Varasemate tulemustega eraldi aastate andmestike peal on võimalik tutvuda tulemites D4.5 ja D4.6.

Eelpool kirjeldatud andmestiku (vt ptk 2.4) ja mudeli arhitektuuriga (vt ptk 3.4) jõuti järgmiste tulemusteni (meie poolt parimaks hinnatud mudeli puhul):

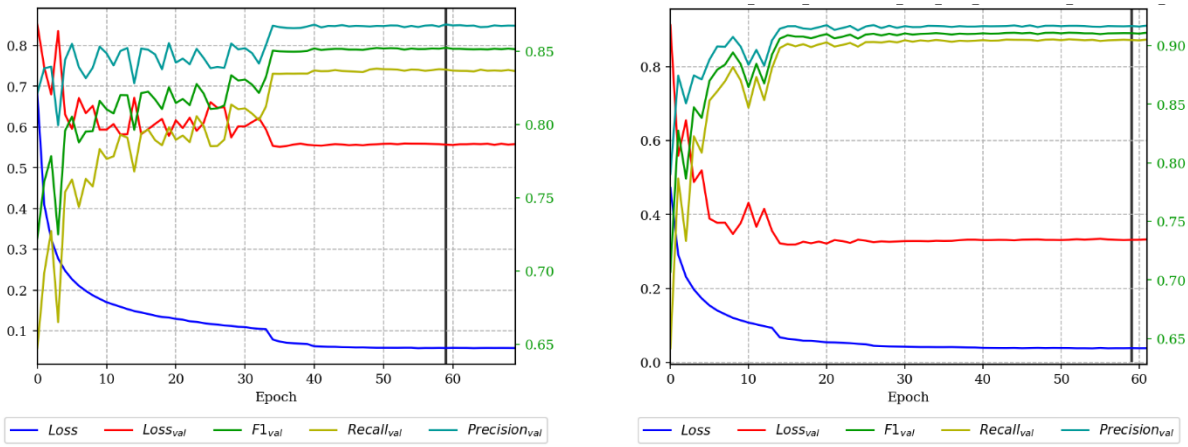
	2018+2019	
	Detailne	Üldine
Treeningkogu F1 skoor	0,9807	0,9883
Treeningkogu kadu	0,0584	0,0380
Valideerimiskogu F1 skoor	0,8524	0,9109
Valideerimiskogu kadu	0,5567	0,3331

Joonis 14 näitab detailse klassifikatsiooni testkogu põldude jaotust ruumiliselt. Kahe aasta andmestikust valiti juhuslikult testkogusse 5773 põldu, millest mudel klassifitseeris õigesti 4880 ja valesti 893 põldu. Kuna juhuvalim oli piisavalt suur, siis katavad sellesse kuuluvad põllud Eesti ala ühtlaselt.

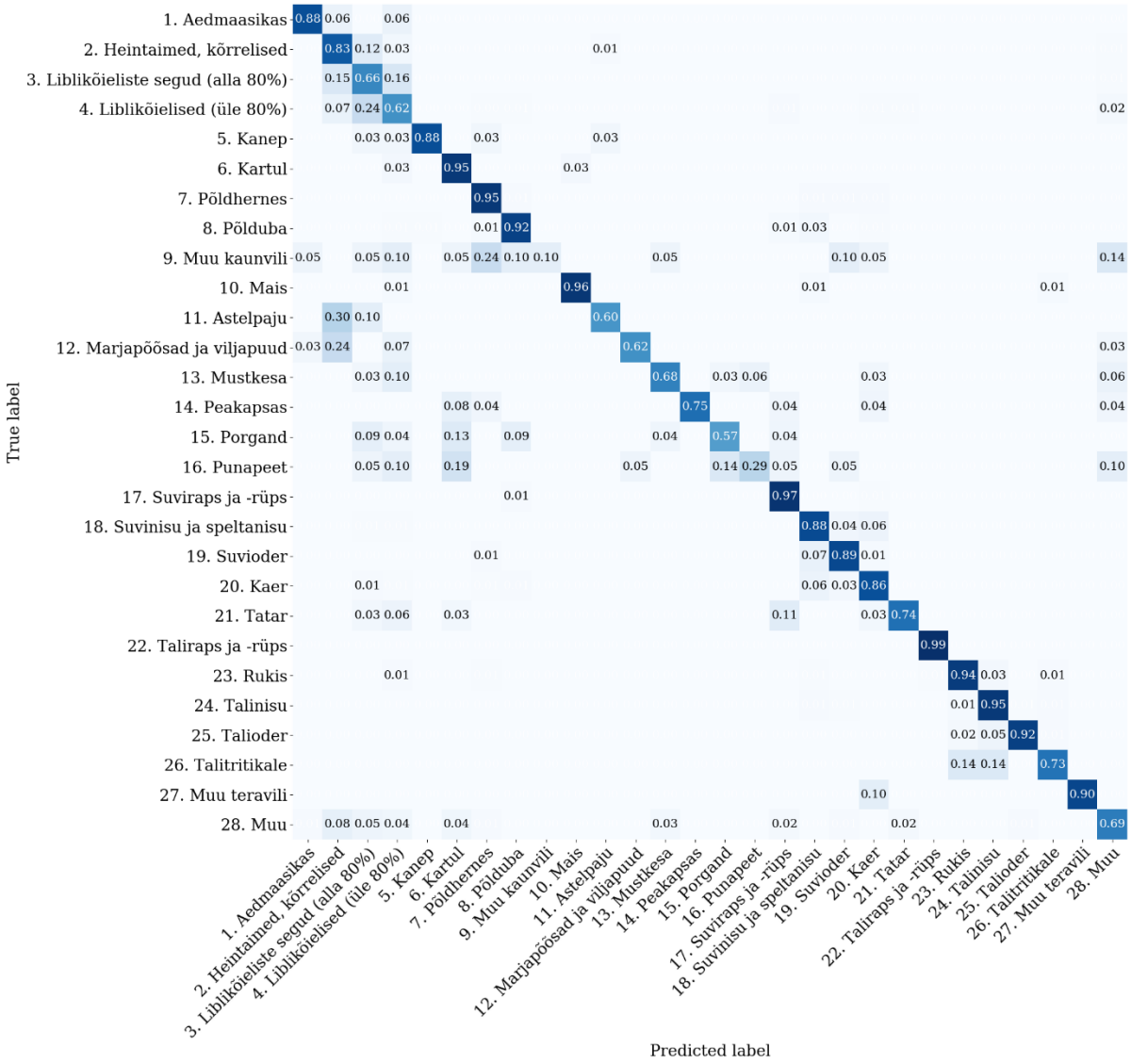


Joonis 14. Juhuslikult testkogusse valitud põldude (5773) jaotus Eesti alal. Rohelised põllud klassifitseeriti õigesti (4880) ja punased valesti (893). Testkogusse kaasati põlde nii 2018 kui 2019 aastast, osaliselt võivad nad ruumiliselt kattuda.

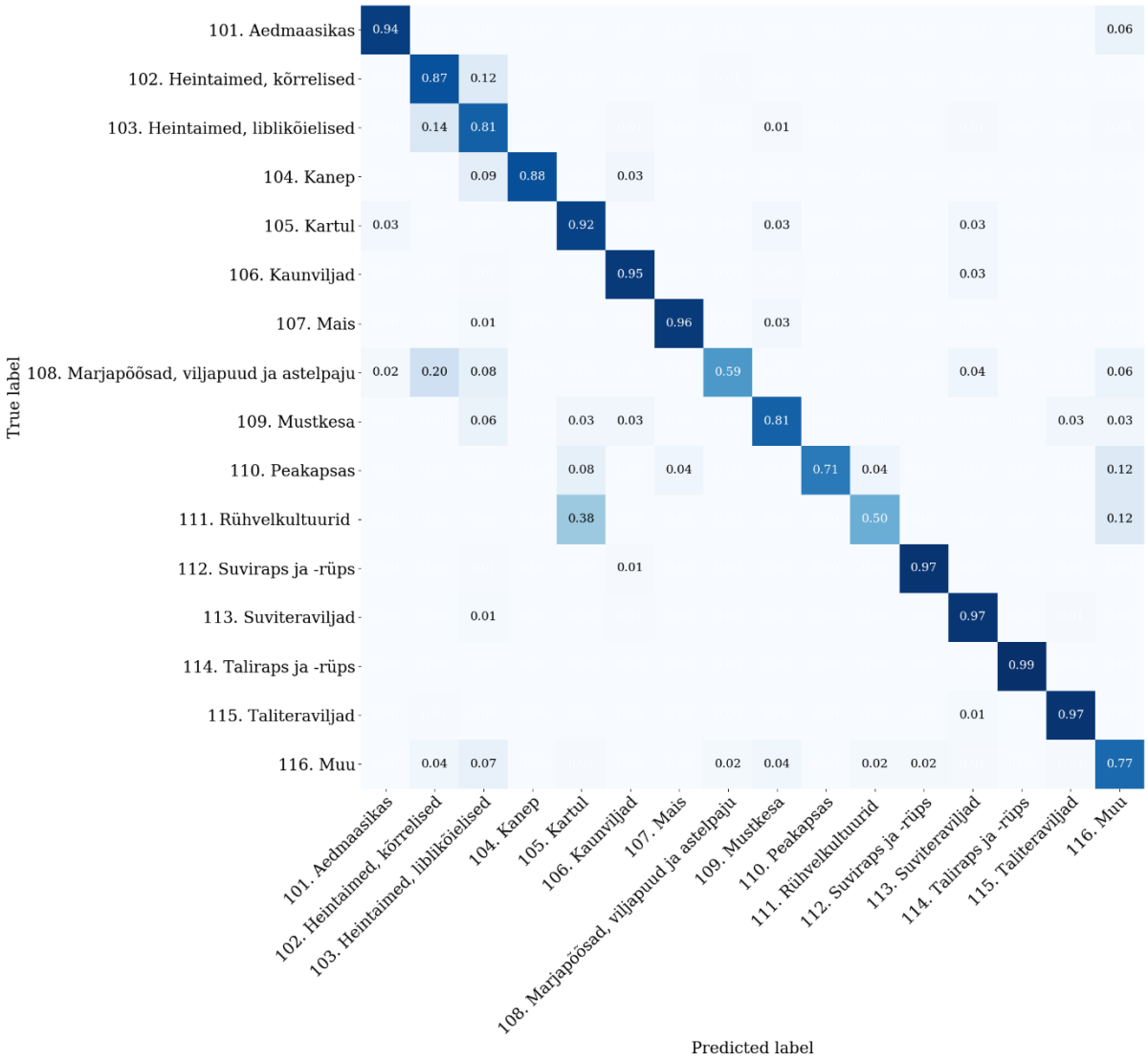
Meetrikute täpsemad selgitused on peatükis 4.1. Treenimiste ajalood on kujutatud Joonis 15, eksimismaatriksid Joonis 16 ja Joonis 17. Põllukultuurigruppide tuvastamise võimekust täpsuse klasside kaupa kujutavad Tabel 10 ja erinevad meetrikud mudeli hinnangutele testkogul Tabel 11 ja Tabel 12.



Joonis 15. Mudeli treenimise ajalugu. Must joon näitab epohhi, mille mudel hinnati parimaks. Vasakul vertikaalsel teljel on kadude (Loss) skaala ja paremal valideerimisandmestiku F1 skoori (F1), saagise (Recall) ja täpsuse (Precision) skaala. Vasakul on detailse klassifikatsiooni mudel ja paremal üldise klassifikatsiooni mudel.



Joonis 16. Kahe aasta andmete peal treenitud detailse klassifikatsiooni mudeli normeeritud eksimismaatriks testandmestikul.



Joonis 17. Kahe aasta andmete peal treenitud üldise klassifikatsiooni mudeli normeeritud eksimismaatriks testandmestikul.

Tabel 10. Põllukultuurigruppide tuvastamise saagis detailse ja üldise klassifikatsiooni puhul.

Saagis (recall)	Kultuurigrupid (detailne)	Kultuurigrupid (üldine)
> 90 %	Kartul; Põldhernes; Põlduba; Mais; Suviraps ja -rüps; Taliraps ja -rüps; Rukis; Talinisu; Talioder; Muu teravili.	Aedmaasikas; Kartul; Kaunviljad; Mais; Suviraps ja -rüps; Suviteraviljad; Taliraps ja -rüps; Taliteraviljad;
80–90 %	Aedmaasikas; Heintaimed, kõrrelised; Kanep; Suvinisu ja speltanisu; Suvioder; Kaer;	Heintaimed, kõrrelised; Heintaimed, liblikõielised; Kanep; Mustkesa;
70–80 %	Peakapsas; Tatar; Talitritikale;	Peakapsas; Muu;
60–70 %	Liblikõieliste segud (alla 80%); Liblikõielised (üle 80%); Astelpaju; Marjapõõsad ja viljapuud; Mustkesa; Muu.	-
50–60 %	Porgand.	Marjapõõsad, viljapuud ja astelpaju; Rühvelkultuurid;
< 50 %	Muu kaunvili; Punapeet.	-

Tabel 11. Testkogu hinnangute meetrikud klasside kaupa (detailsem klassifikatsioon). Saagis on näitaja, mida on kujutatud ka eksimismaatriksi peadiagonaalil.

Kultuur		Meetrikud		
Kood	Nimi	**Täpsus** (precision)	Saagis (recall)	F1 skoor (F1 score)
1	Aedmaasikas	0,7	0,88	0,78
2	Heintaimed, kõrrelised	0,78	0,83	0,81
3	Liblikõieliste segud (alla 80%)	0,67	0,66	0,66
4	Liblikõielised (üle 80%)	0,63	0,62	0,63
5	Kanep	0,88	0,88	0,88
6	Kartul	0,7	0,95	0,8
7	Põldhernes	0,92	0,95	0,94
8	Põlduba	0,84	0,92	0,88
9	Muu kaunvili	0,4	0,1	0,15
10	Mais	0,97	0,96	0,96
11	Astelpaju	0,55	0,6	0,57
12	Marjapõõsad ja viljapuud	0,86	0,62	0,72
13	Mustkesa	0,68	0,68	0,68
14	Peakapsas	1	0,75	0,86
15	Porgand	0,76	0,57	0,65
16	Punapeet	0,67	0,29	0,4
17	Suviraps ja -rüps	0,93	0,97	0,95
18	Suvinisu ja speltanisu	0,85	0,88	0,87
19	Suvioder	0,91	0,89	0,9
20	Kaer	0,88	0,86	0,87
21	Tatar	0,74	0,74	0,74
22	Taliraps ja -rüps	0,99	0,99	0,99
23	Rukis	0,92	0,94	0,93
24	Talinisu	0,97	0,95	0,96
25	Talioder	0,96	0,92	0,94
26	Talitritikale	0,75	0,73	0,74
27	Muu teravili	0,69	0,9	0,78
28	Muu	0,74	0,69	0,72

Tabel 12. Testkogu hinnangute meetrikud klasside kaupa (üldine klassifikatsioon). Saagis on näitaja, mida on kujutatud ka eksimismaatriksi peadiagonaalil.

Kultuur		Meetrikud		
Kood	Nimi	**Täpsus** (precision)	Saagis (recall)	F1 skoor (F1 score)
101	Aedmaasikas	0,79	0,94	0,86
102	Heintaimed, kõrrelised	0,83	0,87	0,85
103	Heintaimed, liblikõielised	0,82	0,81	0,82
104	Kanep	0,93	0,88	0,9
105	Kartul	0,75	0,92	0,83
106	Kaunviljad	0,97	0,95	0,96
107	Mais	0,99	0,96	0,97
108	Marjapõõsad, viljapuud ja astelpaju	0,81	0,59	0,68
109	Mustkesa	0,62	0,81	0,7
110	Peakapsas	1	0,71	0,83
111	Rühvelkultuurid	0,5	0,50	0,5
112	Suviraps ja -rüps	0,99	0,97	0,98
113	Suviteraviljad	0,95	0,97	0,96
114	Taliraps ja -rüps	1	0,99	1
115	Taliteraviljad	0,98	0,97	0,98
116	Muu	0,84	0,77	0,81

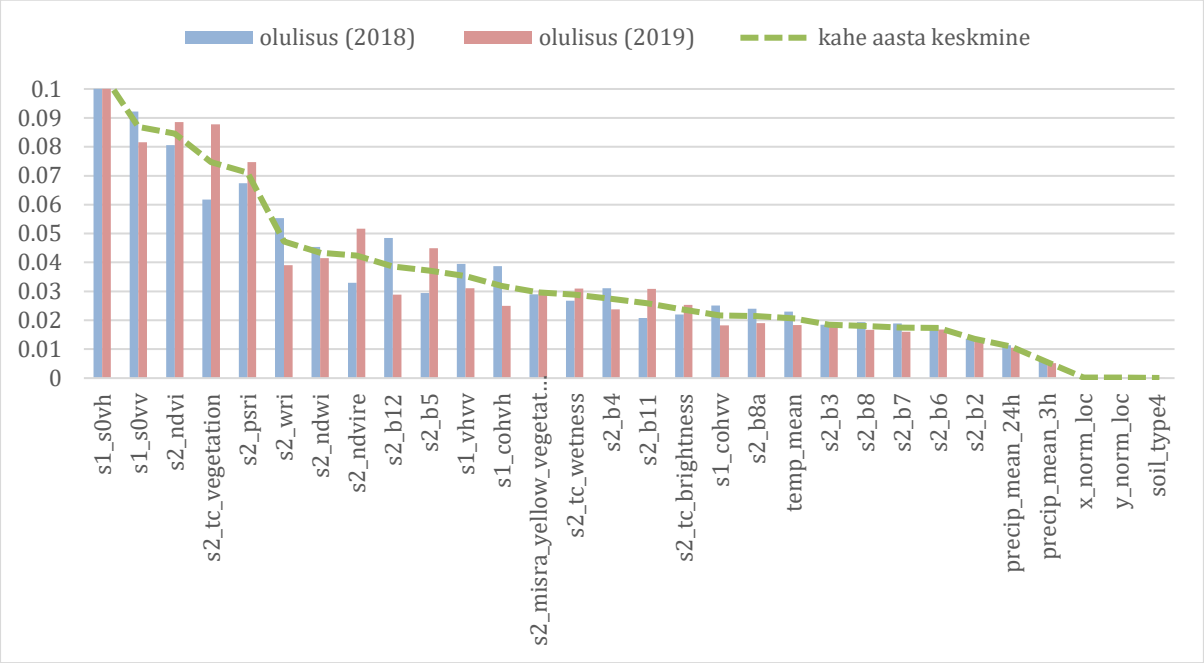
4.3 Tunnuste olulisuse hindamine

Iga konkreetse tunnuse panust ehk olulisust mudeli lõpptulemusse on närvivõrkudes keerulisem hinnata kui näiteks klassikaliste klassifitseerimisalgoritmide puhul. Viimaste puhul on erinevates andmetöötlusepakettides olemas eeldefineeritud funktsoonid tunnuste olulisuse hindamiseks.

Antud juhul kasutati *scikit-learn* (https://scikit-learn.org/stable/) teeke ja ehitati sama sisendandmestiku peale uus otsustusmetsa masinõppemudel. Lisaks tunnuste olulisuse hindamisele pakub alternatiivne mudel head võrdlusmaterjali närvivõrkudega saavutatud tulemuste täpsuse hindamiseks.

Klassifitseerimisel kasutati esialgu kõiki olemasolevaid tunnuseid (vt pt 2.1). Otsustusmetsa klassifikaatorilt on ühe omadusena võimalik pärida tunnuste olulisuse hinnanguid

(*feature_importances_*). Tuntud ka kui Gini-olulisus või ebapuhtuse-põhine tunnuste olulisus (*impurity based*), näitab see hinnang, kui tihti mingit tunnust kasutati puutippude lahknemistel[1]. Mõningase järeltöötluse abil on võimalik hinnata iga tunnuse üle kõigi päevade summeeritud olulisust (Joonis 18).



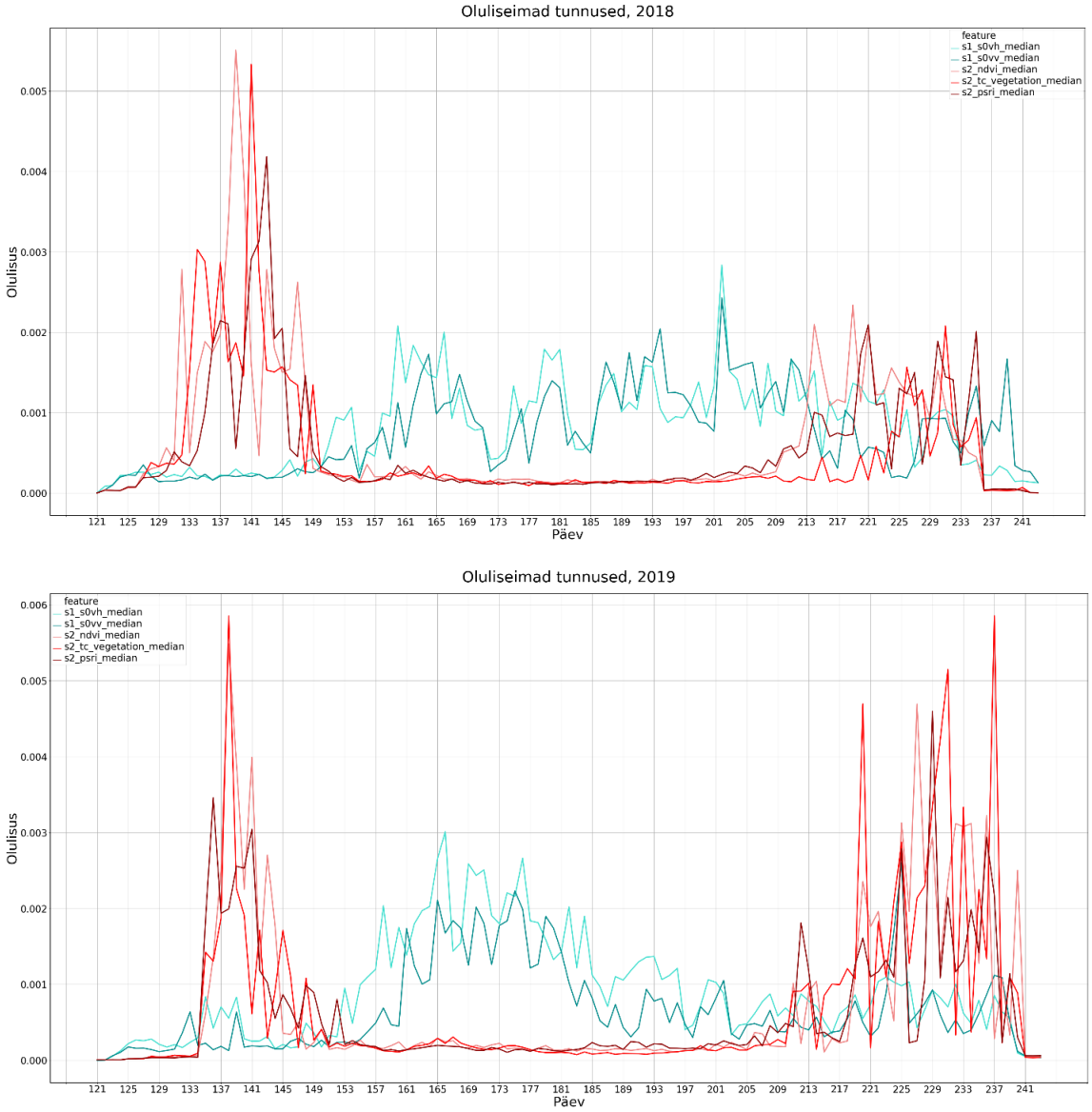
Joonis 18. Tunnuste Gini-olulisus 2018. ja 2019. aasta treeningandmete sobitamisel. Järjestatud kahe aasta keskmise väärtuse järgi, mis on ka kujutatud halli joonena graafikul. Mida suurem väärtus, seda olulisem tunnus.

Lisaks iga tunnuse erinevale olulisusele, võivad tunnuste olulisused ka ajas muutuda. Kui vaadata 5 oluliseima tunnuse olulisust ajas (Joonis 19), siis on märgata, et optilise satelliidi Sentinel-2 tunnused (graafikul punaka värviga) on olulisemad hooaja alguses ja lõpus, samas kui radarsatelliidi Sentinel-1 tunnused (graafikul rohekad) omavad suuremat mõju hooaja keskel. Sellisel mustril võib olla mitu põhjust:

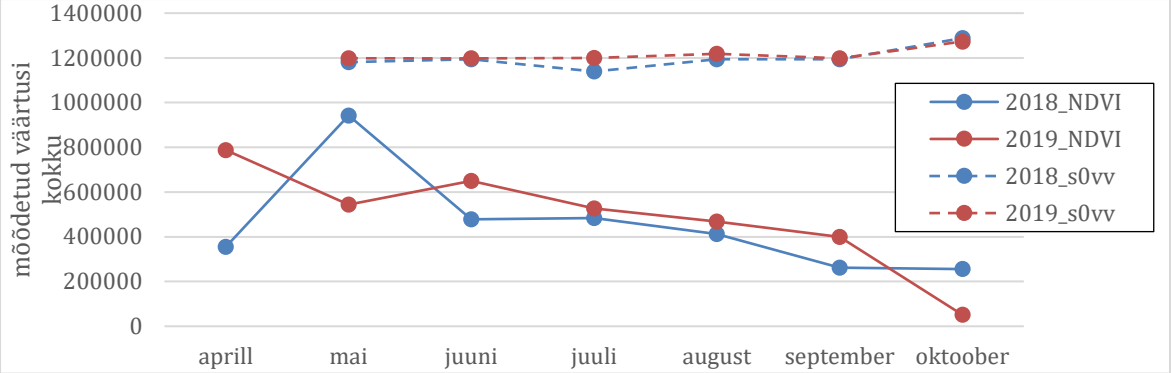
- S2 tunnused on olulisemad mais ja augustis, sest mais on vastavalt erinevate põllukultuuride erinevale arengule nende "värv" ajas erinev ning ka augustis, saagi valmimise faasis on nende "värv" erinev. Samas kui juunis-juulis on peaaegu kõik kultuurid ühtlaselt rohekates toonides ning suurt eristust S2 värvi järgi teha ei saa.
- S1 on tundlik taimede struktuurile ja veesisaldusele ning ka juunis-juulis on erinevate kultuuride vahel võimalik leida erinevusi. Juunis-juulis toimuvad paljudel kultuuridel ka strukturaalsed muutused, näiteks viljapeade moodustumine, viljade valmimine jne.

[1] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html?highlight=gini%20impurity#sklearn.ensemble.RandomForestClassifier.feature_importances_

- Eestis on üldiselt juunis ja juulis palju pilvi, mis häirivad optilise satelliidi ülesvõtteid. Pilvedest rikutud tühimikud täidab edukalt S1, mis panustab seetõttu ka enam kultuuride eristamisse. Võiks arvata, et ka selle tõttu on kevadel optilised tunnused olulisemad, sest on lihtsalt rohkem mõõtmisi. Joonis 20 aga seda hüpoteesi päris ei kinnita. 2018 aasta puhul on tõesti mais oluliselt rohkem mõõtmisi kui suvekuudel või sügisel, aga 2019 mai on jällegi mõõtmiste hulgalt pigem tagasihoidlik. Ilmselt pilvisusest tunnuste hooajaline muster (eelkõige kesksuvine S2 väheolulisus) siiski väga mõjutatud pole.



Joonis 19. Viie oluliseima tunnuse olulisus päevade lõikes. Punakate toonidega on kujutatud Sentinel-2 tunnused ja sinakas-rohekate toonidega Sentinel-1 tunnused.



Joonis 20. Kõik andmebaasis olevad S2 NDVI ja S1 s0vv mõõtmised kuude kaupa aastatel 2018 ja 2019.

4.4 Otsustuspuudemetsa klassifitseerimistulemused

Tunnuste olulisuse hindamise kõrvaltulemusena (või pigem eeldusena) tekkisid juhusliku otsustusmetsa klassifikaatori (*Random Forest*) tulemused, mis on väärtuslik võrdlusmaterjal närvivõrkude mudeli tulemustele.

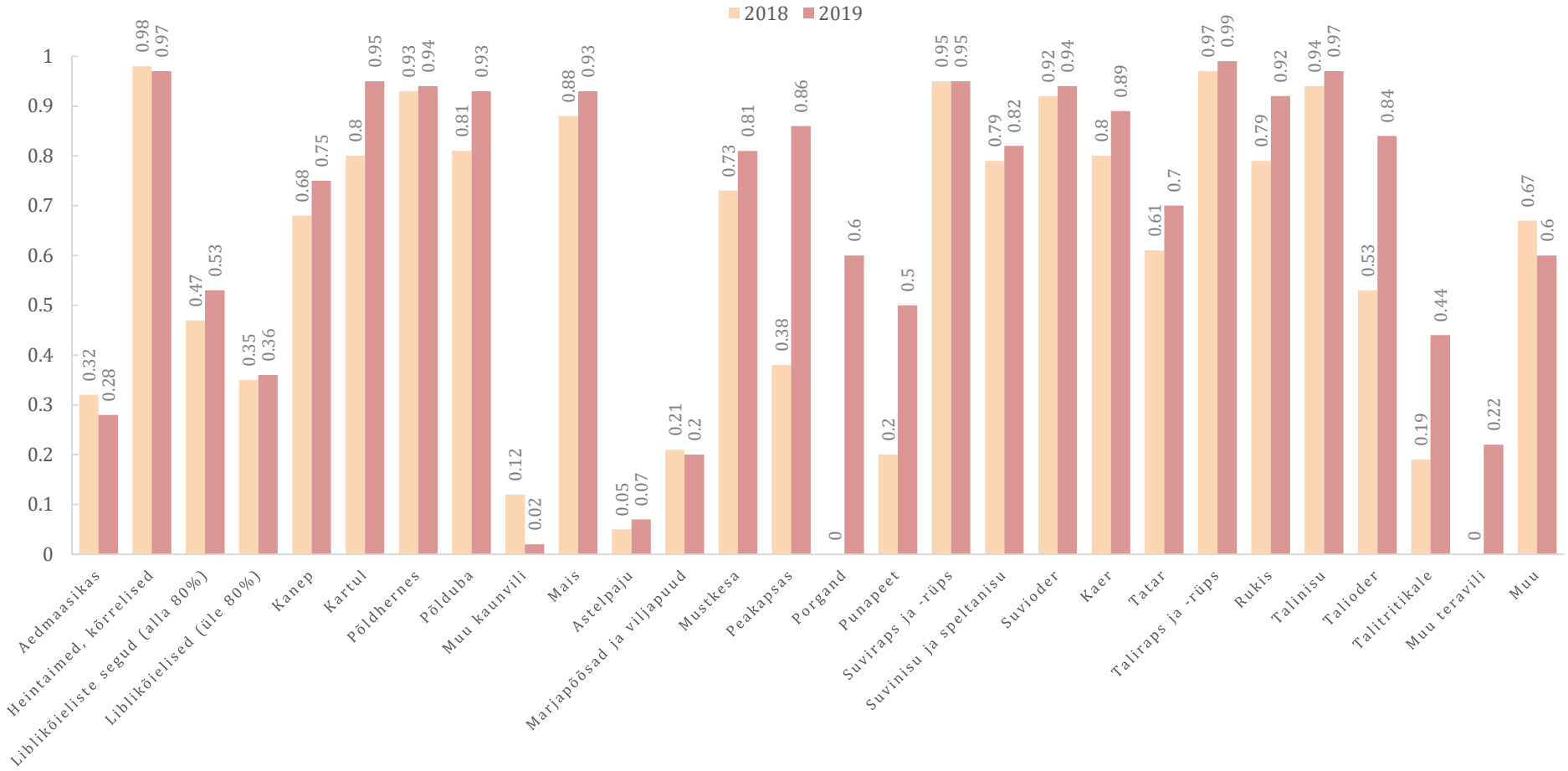
Kuigi sisendandmestik on mõlemal mudelil sama, tuleb rõhutada et eeltöötluse sammud olid oluliselt erinevad ja sõltusid antud mudeli arhitektuuri iseärasustest ja andmeformaadi eeldustest. Samuti polnud treening- ja testandmestiku sisud identsed ning otsustuspuudemetsa puhul ei viinud me treeningkogu klasside näidised tasakaalu. Seetõttu peaks seda võrdlust võtma kui närvivõrkude-põhise lähenemise kaudset valideerimist alternatiivse klassifitseerimisalgoritmiga.

Kokkuvõttes olid tulemused üsna sarnased. Klasside ülene kaalutud keskmine F1 skoor oli 2018. aasta testandmestiku puhul **0,83** ja 2019. aastal **0,86**, mis on suhteliselt kõrge ja võrreldav närvivõrkude kahe aasta mudeli valideerimisandmestiku F1 skooriga (**0,85**). Otsustusmetsa mudeli klassifitseerimistulemused on toodud allpool (Tabel 13, Joonis 21 ja Joonis 22).

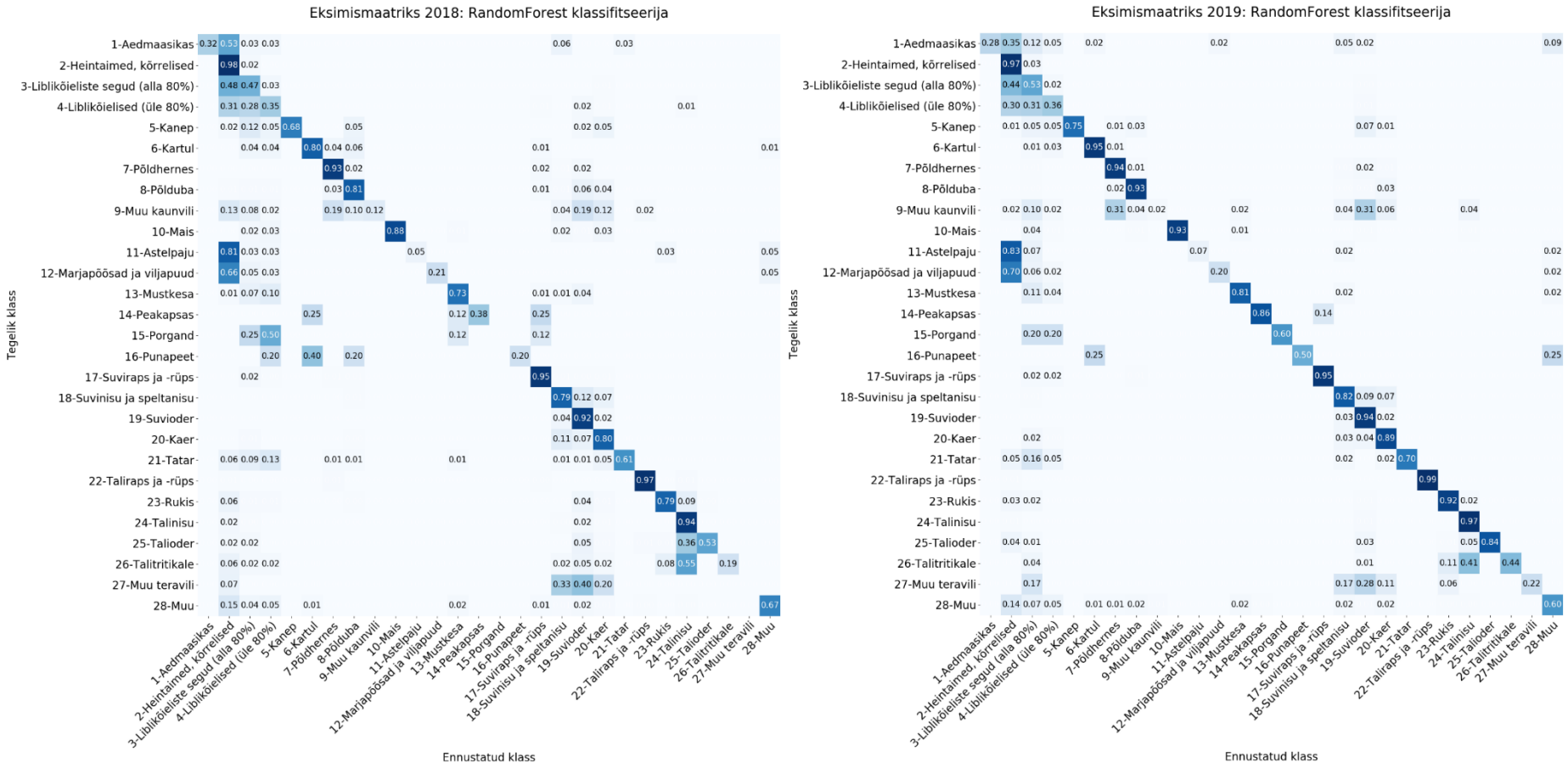
Tabel 13. 2018. ja 2019. aasta testandmestiku klassifitseerimistulemused otsustusmetsa klassifikaatoriga.

Klass	Kultuur	2018		2019	
		saagis	põlde	saagis	põlde
1	Aedmaasikas	0.32	34	0.28	43
2	Heintaimed, kõrrelised	0.98	9619	0.97	9472
3	Liblikõieliste segud (alla 80%)	0.47	2291	0.53	2429
4	Liblikõielised (üle 80%)	0.35	894	0.36	753
5	Kanep	0.68	57	0.75	73
6	Kartul	0.8	82	0.95	79

7	Põldhernes	0.93	554	0.94	626
8	Põlduba	0.81	335	0.93	201
9	Muu kaunvili	0.12	52	0.02	48
10	Mais	0.88	124	0.93	167
11	Astelpaju	0.05	37	0.07	46
12	Marjapõõsad ja viljapuud	0.21	62	0.2	54
13	Mustkesa	0.73	67	0.81	54
14	Peakapsas	0.38	8	0.86	7
15	Porgand	0	8	0.6	5
16	Punapeet	0.2	5	0.5	4
17	Suviraps ja -rüps	0.95	704	0.95	399
18	Suvinisu ja speltanisu	0.79	1537	0.82	1131
19	Suvioder	0.92	2441	0.94	2031
20	Kaer	0.8	1071	0.89	991
21	Tatar	0.61	82	0.7	44
22	Taliraps ja -rüps	0.97	407	0.99	738
23	Rukis	0.79	247	0.92	533
24	Talinisu	0.94	1200	0.97	1902
25	Talioder	0.53	130	0.84	276
26	Talitritikale	0.19	62	0.44	85
27	Muu teravili	0	15	0.22	18
28	Muu	0.67	212	0.6	227
Kaalutud keskmine F1 skoor		0.83	22337	0.86	22436



Joonis 21. Otsustuspuude metsa klassifitseerimise saagised kahe aasta testandmestikul.



Joonis 22. Otsustusmetsa 2018. (vasakul) ja 2019. aasta (paremal) testandmestiku klassifitseerimise eksimismaatriksid.

4.5 Tulemused erinevatel ajahetkedel hooaja jooksul

Võimaliku operatiivsüsteemi jaoks on oluline kultuure tuvastada hooaja võimalikult varajases faasis, et kahtlaste põldude puhul oleks võimalik teostada kohapealset kontrolli. Katsetasime kahe aasta mudelit (detailse ja üldise klassifikatsiooniga) sobitada kahe varasema kuupäeva seisuga – 24. juuni ja 1. august. Tulemused on toodud Tabel 14.

Tabel 14. Kahe aasta mudeli detailse ja üldise klassifikatsiooni mudeli meetrikud erinevatel ajahetkedel

	24. juuni		1. august		1. september	
	Detailne	Üldine	Detailne	Üldine	Detailne	Üldine
Treeningkogu F1 skoor	0,9306	0,9730	0,9659	0,9825	0,9807	0,9883
Treeningkogu kadu	0,2024	0,0844	0,1057	0,0564	0,0584	0,0380
Valideerimiskogu F1 skoor	0,7442	0,8653	0,8394	0,8993	0,8524	0,9109
Valideerimiskogu kadu	0,7967	0,4780	0,5474	0,3520	0,5567	0,3331

Tuleb silmas pidada, et erinevate lõppkuupäevadega treenides kasutasime sama arhitektuuriga mudelit, mis ei pruugi lühendatud aegridade puhul olla alati optimaalne.

Nagu eelnevast tabelist näha, siis hooaja edenedes klassifikatsiooni F1 skoor paraneb, mis oli ka oodatav. Juuli andmete kaasamine parandas tulemust oluliselt, samas kui 1. augusti ja 1. septembri erinevus oli väiksem.

Üle klasside kaalutud meetrikute väärtused ei anna terviklikku pilti kultuurigruppide kaupa. Erinevate kuupäevade eksimismaatriksitest on näha, et üle 90% saagis saavutati 24. juuniks järgmistes detailse klassifikatsiooni klassides: aedmaasikas, põldhernes; taliraps ja -rüps, rukis, talinisu. 1. augustiks lisandusid eelnevatele üle 90% õigsusega veel kartul, põlduba, mais, suviraps- ja rüps, talioder. Aedmaasika saagis aga langes taas alla 90%. 1. septembriks saavutasid üle 90% õigsuse kõigile eelnevatele lisaks veel vaid muu teravili.

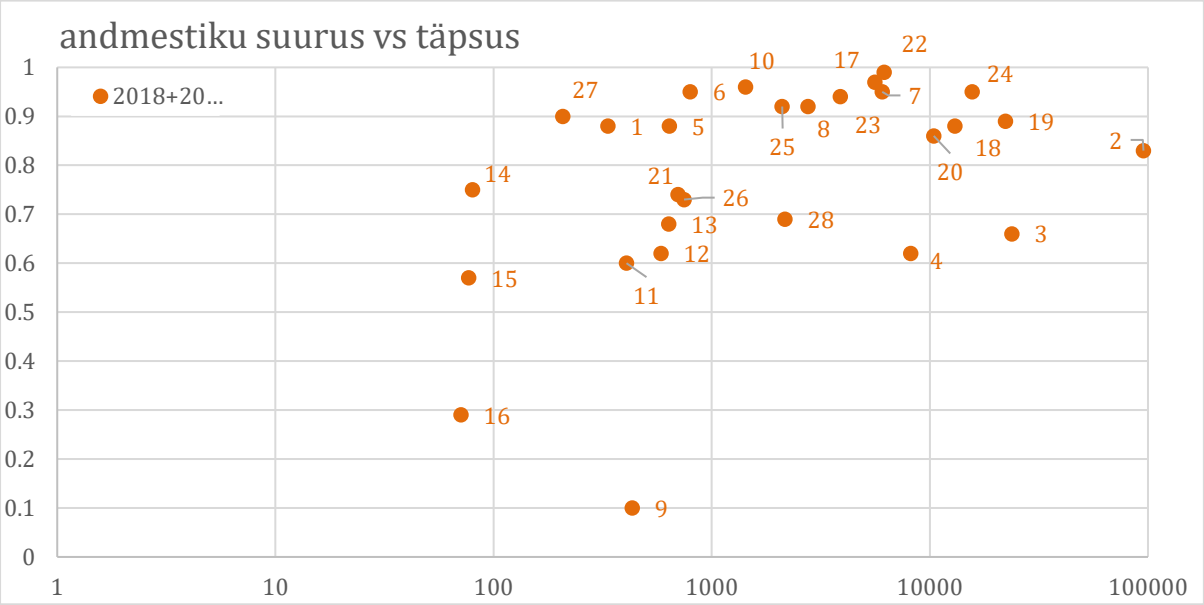
Eksimismaatriksid (ka üldise klassifikatsiooni omad) on leitavad aruandele kaasatud jooniste hulgast.

4.6 Tulemuste tõlgendamine ja arutelu

Kuna sügavõppe mudeli täpsus sõltub otseselt etteantud sisendandmete kvaliteedist ja hulgast, siis teatud käitumist on võimalik seletada sisendandmete analüüsiga. Kui lugeda kokku, mitu mudeli treenimisel kasutatavat põldu iga põllukultuurigrupi kohta andmekogus on (Tabel 4), siis kõige vähemarvukad kultuurid (punapeet, porgand, peakapsas jt) on just need, mida mudel suhteliselt kehvemini tuvastada suudab. Väike hulk treening- ja valideerimisandmeid mõjutab oluliselt mudeli sooritust. Madala täpsuse põhjuseks võib olla ka nende kultuurigruppide aegridade suhteline sarnasus teiste (suuremate) kultuurigruppide aegridadega.

Joonis 23 on kujutatud klasside suuruste ja klasside saagise suhet. Üldiselt on näha seost, et mida suurem hulk näidiseid, seda suurem saagis. Umbes 1000 näidise juures võib märgata küllastumist, ehk ideaaljuhul võiks iga klassi kohta olla õpetusandmekogus u 1000 esinduslikku ja kvaliteetset põldu.

Samuti tulevad välja klassid 2, 3 ja 4 (liblikõielised), mille puhul on küll treeningkogu suur, ent saagis jääb madalaks (eriti klasside 3 ja 4 puhul). Ilmselt on nende klassid täpne eristamine olemasoleva andmestikuga väga raske, kui mitte võimatu ülesanne. Liblikõieliste klassi eristusvõimalusi analüüsime allpool.



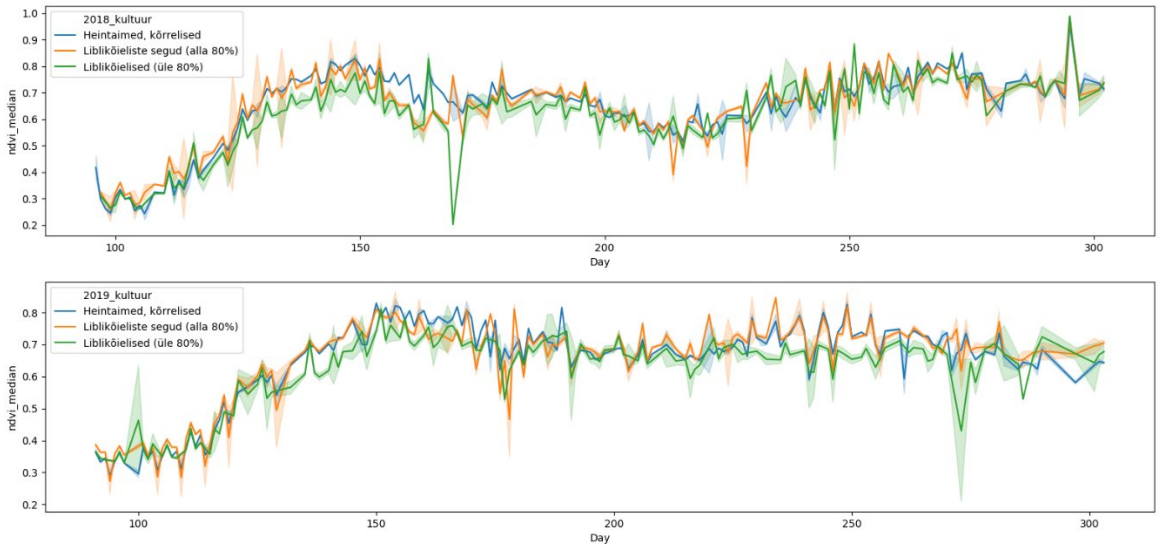
Joonis 23. Erinevate kultuuriklasside (detailne klassifikatsioon) andmekogu suurused (x teljel) klassifitseerimise saagised (y-teljel).

Kuna osade detailse klassifikatsiooni (*narrow*) kultuurigruppide sobiva geomeetriga põldude koguarv ja osakaal kogu taotletud pinnast moodustas vähem kui 1% ning õpetusandmete hulk nendes gruppides oli väike, ei ole mudeli ennustusvõime nendes klassides nii oluline kui arvukamalt Eestis kasvatatavate kultuuride puhul. Meie hinnangul on sellisteks „vähetähtsateks“ kultuurideks aedmaasikas (1), muu kaunvili (9), astelpaju (11), marjapõõsad ja viljapuud (12), peakapsas (14), porgand (15), punapeet (16), muu teravili (27). Nende kultuuride puhul jäävad ka lõpliku mudeli saagised 0,1 – 0,75 vahele (välja arvatud aedmaasikas (0,88) ja muu teravili (0,9)), mis ilmselt pole operatiivteenuse jaoks piisav tulemus (Tabel 15). Samas on ilmselt võimalik ka nende klasside tulemusi parandada, kui kaasata rohkem kvaliteetseid näidised treeningkogusse ja mitte kasutada meie enda loodud sünteetilisi aegridu.

Tabel 15. Kasvupinnalt väikeste kultuuride tuvastustäpsused

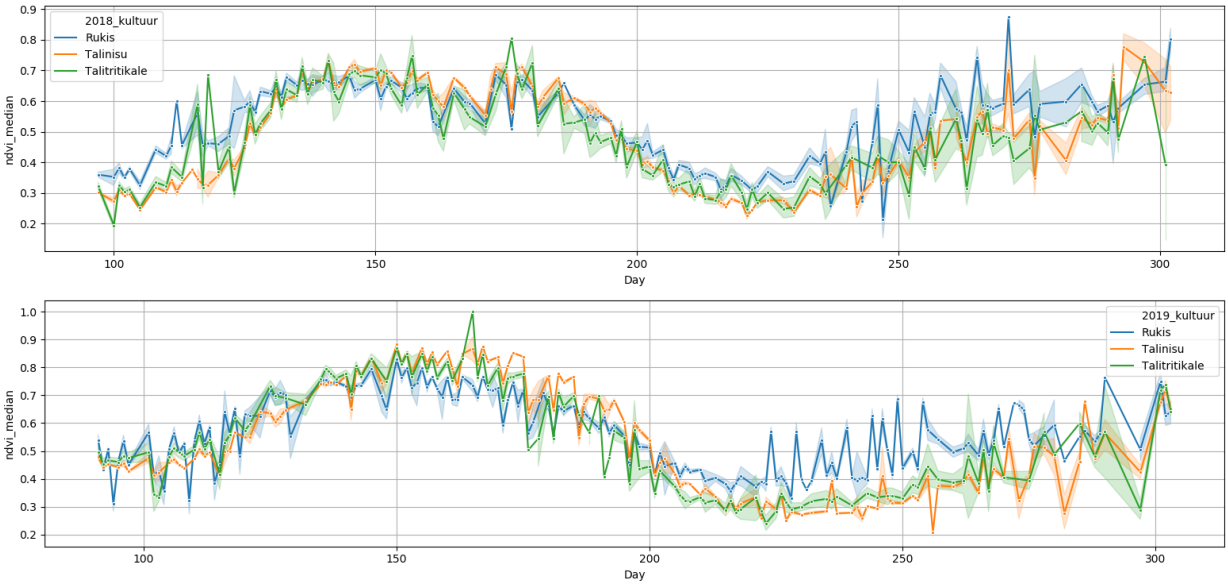
Kultuur		Meetrikud		
Kood	Nimi	**Täpsus** (precision)	Saagis (recall)	F1 skoor (F1 score)
1	Aedmaasikas	0,7	0,88	0,78
9	Muu kaunvili	0,4	0,1	0,15
11	Astelpaju	0,55	0,6	0,57
12	Marjapõõsad ja viljapuud	0,86	0,62	0,72
14	Peakapsas	1	0,75	0,86
15	Porgand	0,76	0,57	0,65
16	Punapeet	0,67	0,29	0,4
27	Muu teravili	0,69	0,9	0,78

Samas on tuvastustäpsus suhteliselt madal ka mõnes suures kultuurigrupis, näiteks gruppides 3 (liblikõieliste segud alla 80%) ja 4 (liblikõielised üle 80%), mida mudel ajab segamini just omavahel ja grupiga 2 (heintaimed, kõrrelised). Kui visualiseerida nende gruppide kõigi põldude satelliitmõõtmiste põhiseid tunnuseid, siis paistabki nende üldine kasvukäik väga sarnane, mistõttu on ka mudeli eksimused arusaadavad. Joonis 24 on toodud näitena NDVI väärtuste joondiagrammid. Aruandega kaasas olevas jooniste kogus on ka teiste tunnuste ajalised käigud.

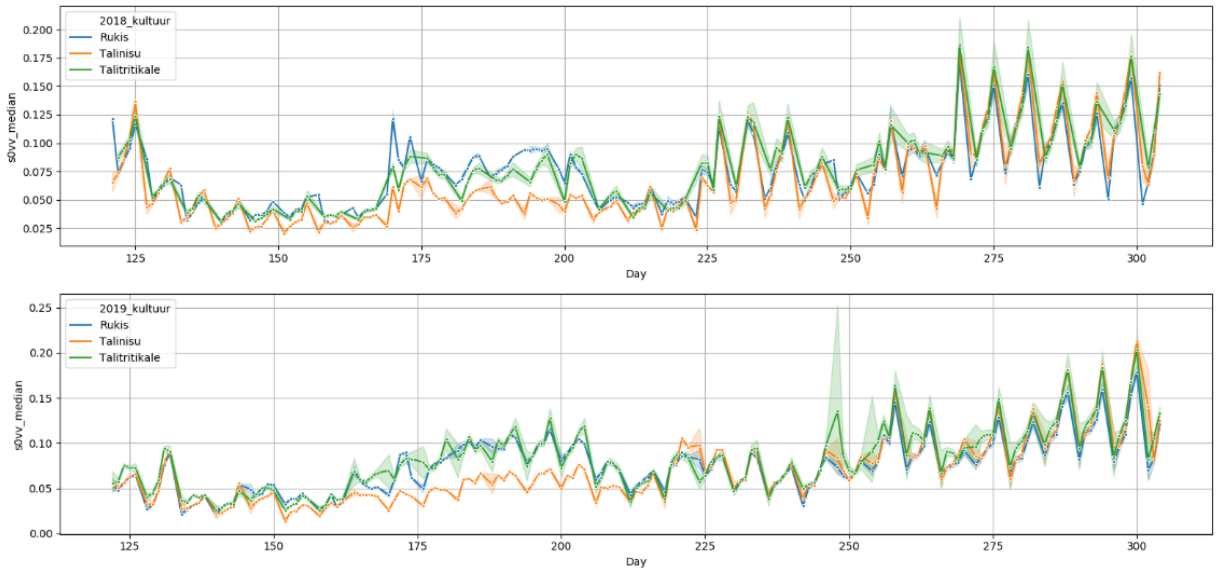


Joonis 24. Kõigi NDVI väärtuste joondiagrammid (üleval 2018 ja all 2019) kolme kultuurigruppi (2, 3, 4) kuuluvate kõigi põldude puhul.

Talitritikalet (26) näib olevat raske eristada talinisust ja rukkist, mille ristand ta on. Optiliselt satelliidilt saadud tunnuste puhul ongi need klassid üpris sarnased (vt NDVI käike Joonis 25), kus küll rukis eristub neist kõige rohkem. Radaritunnuste järgi (vt Joonis 26) eristub nendest kolmest jällegi kõige paremini talinisu. Talitritikale, mille tunnuste aegread kattuvad kas ühe või teisega või jäävad nende vahele, ongi seetõttu ilmselt raskemini eristatav.



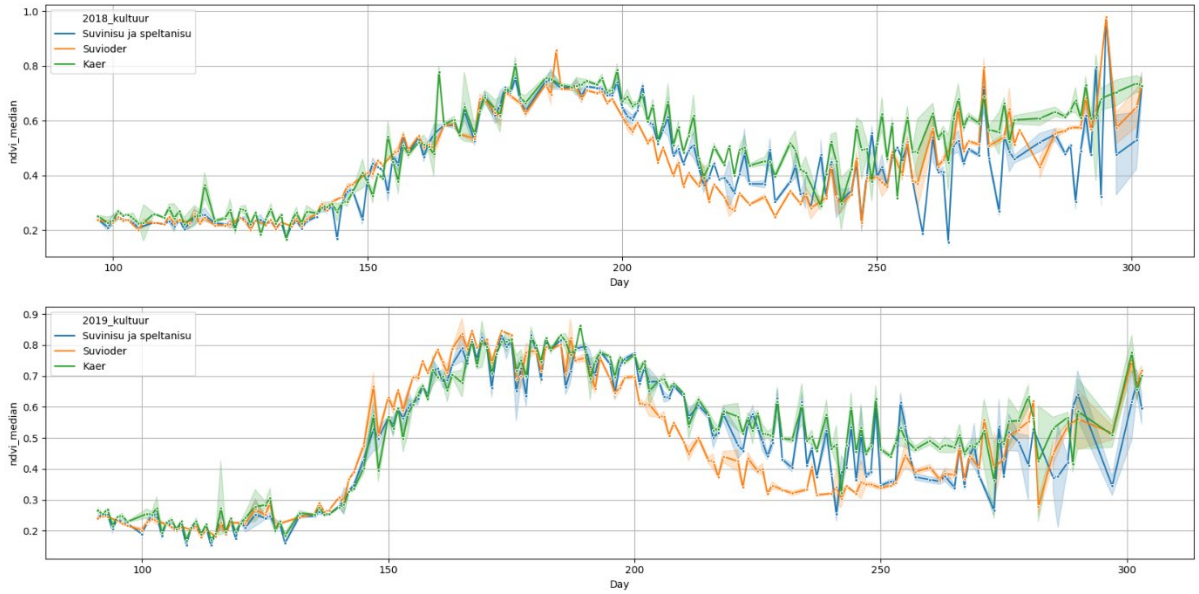
Joonis 25. Kõigi NDVI väärtuste joondiagrammid rukki, talinisu ja talitritikale klassidesse kuuluvate kõigi põldude puhul.



Joonis 26. Kõigi s0vv väärtuste joondiagrammid rukki, talinisu ja talitritikale klassidesse kuuluvate kõigi põldude puhul.

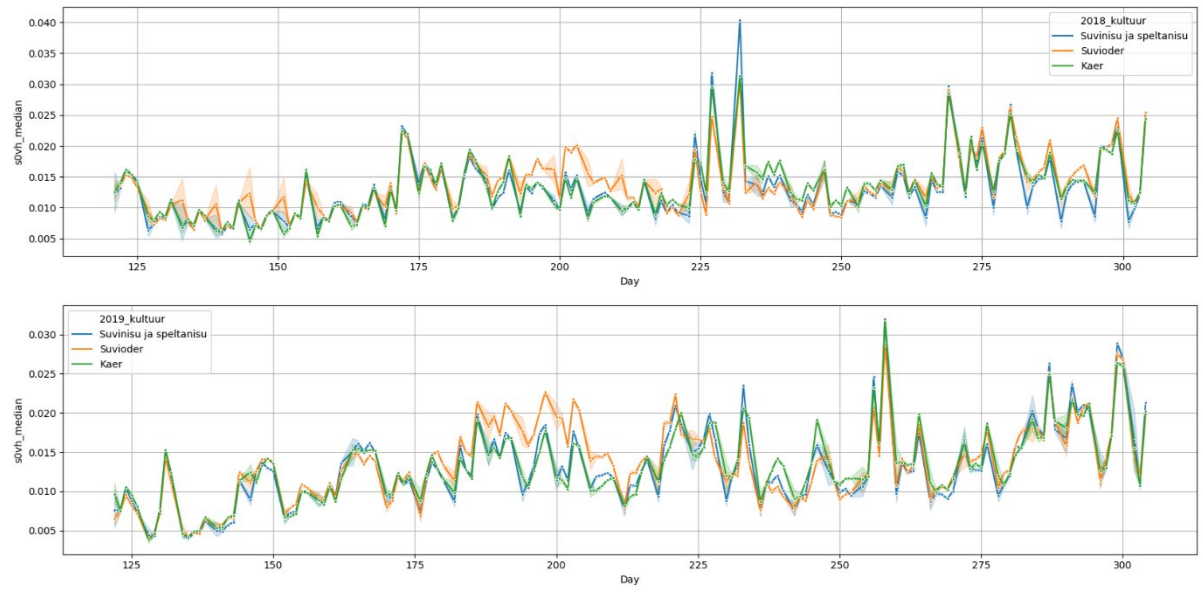
Teraviljadest on tuvastusõigsus madalam tatral (klass 21) (<80%). Tatrapõldudele ennustab mudel liblikõieliste klasse (3,4) ning veidi ka suvirapsi ja -rüpsi (22) ning muu klassi (28).

Selgesti tuleb välja ka mõningane mudeli hinnangute segadus suvinisu ja speltanisu (18), suviodra (19) ja kaera (20) vahel, mille saagised jäävad napilt alla 0,9. Peamiselt ajab mudel segamini neid just omavahel. Joonis 27 kujutab kõigi nendesse klassidesse kuuluvate põldude NDVI väärtusi ajas, kust on näha, et suviodra NDVI väärtused on vähemalt hooaja lõpus erinevad kaerast ja suvinisust. Kaer ja suvinisu on väga sarnase NDVI käiguga.



Joonis 27. Kõigi NDVI väärtuste joondiagrammid suvinisu ja speltanisu, suviodra ja kaera klassidesse kuuluvate kõigi põldude puhul.

Ka radarsatelliidi parameetri s0vh aegrea põhjal eristub suvioder kõige selgemini nendest kolmest kultuurist (Joonis 28). Põhjust, miks mudel neid kolme kultuuri väga täpselt eristada ei suuda, on keeruline leida. On selge, et nende kasvukäigud ei ole väga erinevad, ent iseärasusi leidub (eriti suviodra puhul). Üheks põhjuseks võivad olla ka valesti sildistatud näidised, sest eksete eemaldamine polnud tõenäoliselt täiuslik ja õpetusandmekogusse jäi siiski teatav hulk valede siltidega põlde.

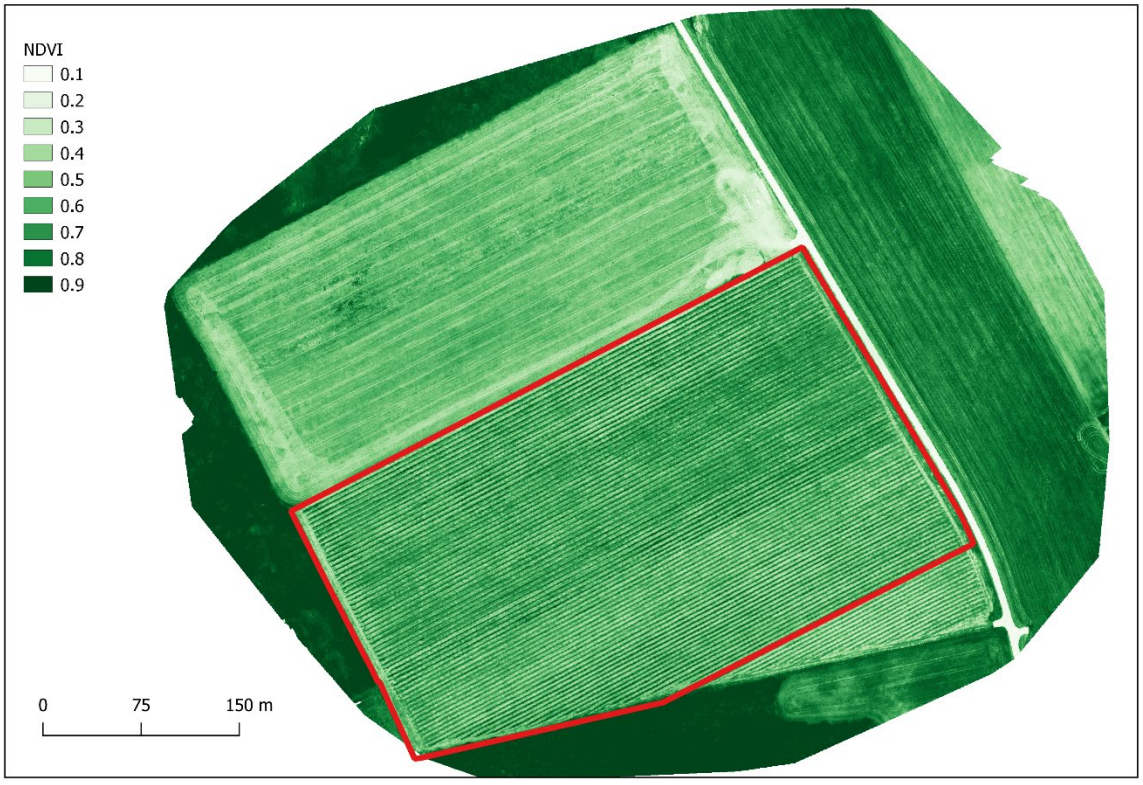


Joonis 28. s0vh väärtuste joondiagrammid suvinisu ja speltanisu, suviodra ja kaera klassidesse kuuluvate kõigi põldude puhul.

Problemaatiline klass on ka „muu“, mille puhul detailse klassifikatsiooni mudel paigutab ekslikult põlde eelkõige heintaimede klassidesse (1, 2, 3) aga vahel ka mujale. Sarnast käitumist võib täheldada ka jämedama klassifikatsiooni „muu“ klassi puhul.

Hoolimata mustkesa klassi erikohtlemisest eeltöötluse etapis (käsitsi eksete eemaldamine), ei suutnud me selle tuvastustäpsusi väga kõrgeks tõsta. Saagis 0,68 detailsema ja 0,81 üldise klassifikatsiooni puhul pole kehvad tulemused, ent endiselt aetakse seda segamini liblikõieliste ja erinevate juurviljadega. Vähem teraviljadega. Tegemist on keerulise klassiga, mille definitsioon on ähmane ja põldude väljanägemise varieerumine ilmselt suur.

Jämedama klassifikatsiooni grupp „Marjapõõsad, viljapuud ja astelpaju“ on kehva tuvastustäpsusega ja seda aetakse peaaegu eranditult segamini kõrreliste ja liblikõieliste heintaimedega. Senise mudeli arendustöö käigus muutub üha ebatõenäolisemaks, et põõsaid ja viljapuid on antud tunnuskomplektiga võimalik piisava täpsusega tuvastada. Vigade põhjuseks on ilmselt asjaolu, et ülevalt satelliidi vaatepunktist ongi marjapõõsaste ja viljapuude aiad suuresti rohttaimedega kaetud ning saadud tunnuskomplektid heintaimede tunnuskomplektidest palju ei erine. Joonis 29 ja Joonis 30 illustreerivad seda, kus mustsõstrapõllu viiruline struktuur hästi välja tuleb. Arvestades, et satelliitfoto lahutus on jämedam kui droonikaameral, sulavad selle pikslites põõsaread ja reavahed kindlasti kokku.



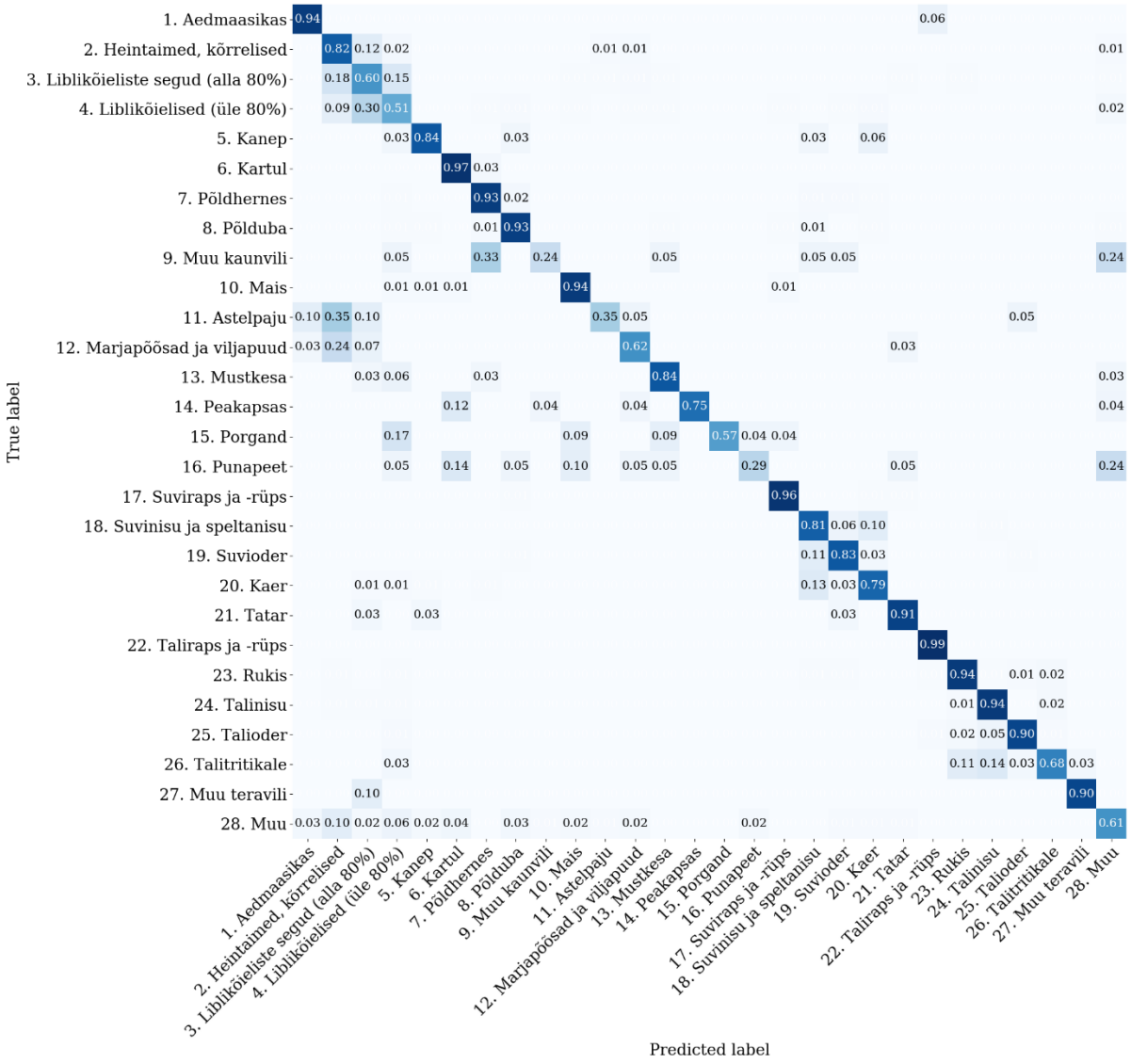
Joonis 29. Droonifoto põhjal loodud NDVI pilt mustsõstrapõllust Verevi küla lähistel (ümbritsetud punase joonega). Mustsõstrapõllu kõrval (põhjapool) on niidetud rohumaa ja nurgas on näha ka sõnnikuhunnikuid. Pildistamise kuupäev: 28.07.2020.

Joonis 30. Mustsõstrapõllu foto, kus on näha põõsad ja niidetud reavahed. Pildistatud kagu suunas. Pildistamise kuupäev: 28.07.2020.

Treenisime närvivõrkude-põhist detailse klassifikatsiooni kahe aasta mudelit ka ainult 5 kõige olulisema tunnusega: 's0vv_median', 's0vh_median', 'ndvi_median', 'tc_vegetation_median' ja 'psri_median'. Saadud mudeli valideerimisandmestiku F1 skoor oli **0,8126** ja kadu **0,6389**. Suurema hulga tunnustega olid need meetrikud vastavalt 0,8524 ja 0,5567.

Viie tunnusega mudeli kärbitud eksimismaatriks testandmestikul on Joonis 31.

Kuigi üldine klassifitseerimisõigsus on veidi väiksem, saab öelda, et väga limiteeritud tunnuste hulgaga on võimalik saavutada üle 0,9 saagis järgmiste kultuuride puhul: aedmaasikas, kartul, põldhernes, põlduba, mais, suviraps ja -rüps, tatar, taliraps ja -rüps, rukis ja talinisu, talioder ja muu teravili. Ehk võrreldes suurema tunnuste arvuga mudeliga kõige täpsemini (saagis > 0,9) ennustatud klasside hulk isegi kasvas aedmaasika ja tatra arvelt.



Joonis 31. Viie oluliseima tunnusega sobitatud kahe aasta mudeli klassifitseerimise saagised testkogul.

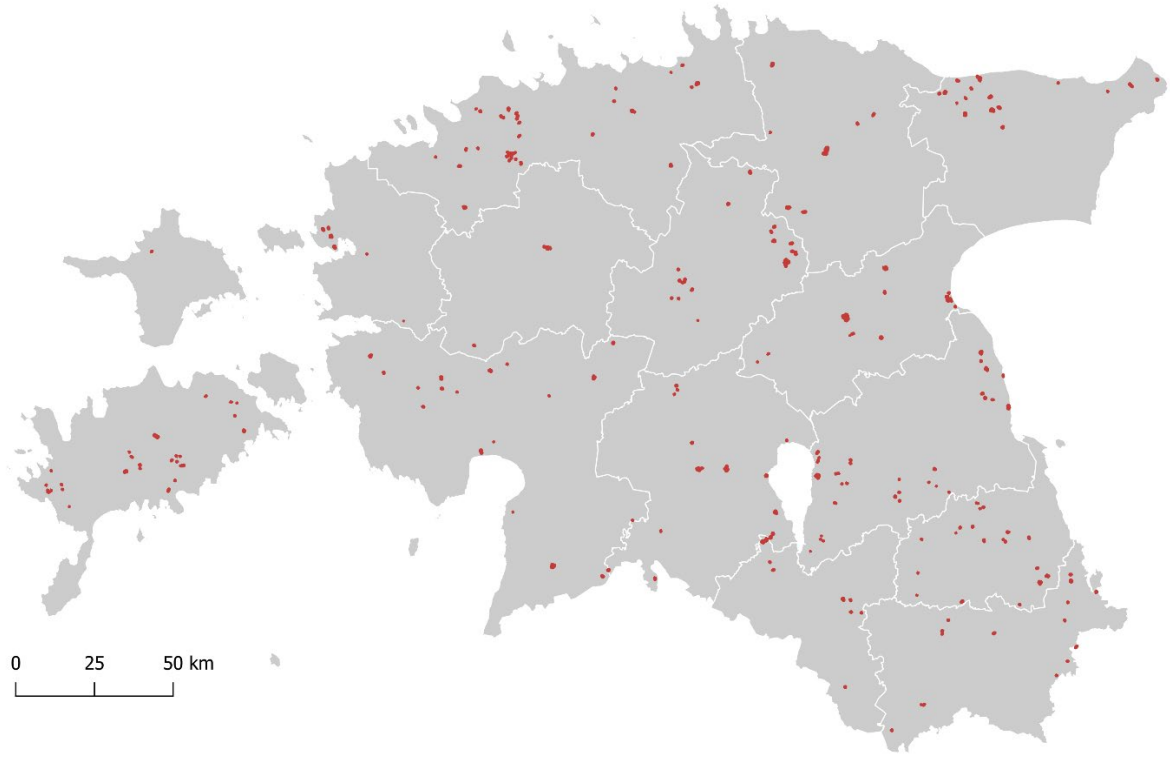
4.7 Mudeli sooritus kontrollitud testkogul

Kui algselt valiti testkogusse põlde juhuslikult, siis mudeli viimases arendustsüklis kohaldasime kogu andmete eeltöötlust viisil, et saaksime testkogusse määrata konkreetseid põlde, mille puhul meil on olemas maapealsed kontrollandmed ja mille sildi õigsuses võib olla kindel. Ilmselt on ka juhuvalimi puhul suur osa siltidest tõepärased, ent PRIA andmetel on talunike taotlusel esitatud andmetes nii tahtmatuid kui tahtlikke eksimusi.

Kuna maapealseid kontrollitud andmeid ei ole iga kultuurigrupi jaoks piisavalt, et kogu testkogu nendega ära täita, siis lisaks paigutati sinna põlde ka juhuvalimist.

Usaldusväärseid põlde kogusime PRIA 2019. aasta välikontrolli sattunud põldude hulgast, Eesti Maaülikooli ja Tartu Ülikooli uurimisgruppide droonilendudest (millest annab täpsema ülevaate RITA tulem D4.9) ja vajadusel ka 2019. a Maa-ameti ortofotodelt, mille põhjal määrati peamiselt vilja- ja marjapuude istandikke.

Valisime igasse kultuurigruppi vähemalt 10 põldu võimalikult hajutatult üle Eesti. Kokku sai kontrollitud testkogusse 298 põldu, nende ruumilist paigutust iseloomustab Joonis 32.



Joonis 32. Kontrollitud testkogu põldude jaotus Eesti alal.

Kontrollitud testkogu tulemused on esitatud Tabel 16. Nagu näha, siis klassides, kus juba suure testkogu peal olid mudeli hinnangud täpsed, on need kontrollitud testkogul isegi veel täpsemad. Näiteks ennustati õigesti kõik põllud klassides 2, 3, 6, 7, 17, 19, 22, 24, 25. Samas mõnes klassis oli mudeli sooritus eriti kehv (sarnaselt suurele testkogule) – näiteks „Muu kaunvili“ klassis ei määratud ühtegi põldu õigesse klassi ning punapeedil sai õige sildi külge

vaid üks põld. Üldiselt on kontrollitud testkogu tulemused heas kooskõlas suure juhuslikult valitud testkogu tulemustega (Tabel 11).

Tabel 16. Mudeli hinnangud kontrollitud testkogul. Veerg „Ennustatud klassid“ kirjeldab valede hinnangute klasse. Näide: 10-st aedmaasika põllust 9 klassifitseeriti õigesti ja 1 valesti. Valesti määratud põldu pidas mudel liblikõieliseks rohumaaks liblikõieliste osakaaluga üle 80%.

Kultuur						
Kood	Nimi	Põlde	Õige hinnang	Vale hinnang	Ennustatud klassid	Saagis
1	Aedmaasikas	10	9	1	[4]	0,90
2	Heintaimed, kõrrelised	10	10	0		1,00
3	Liblikõieliste segud (alla 80%)	10	10	0		1,00
4	Liblikõielised (üle 80%)	12	10	2	[3, 3]	0,83
5	Kanep	10	9	1	[11]	0,90
6	Kartul	10	10	0		1,00
7	Põldhernes	12	12	0		1,00
8	Põlduba	10	9	1	[5]	0,90
9	Muu kaunvili	10	0	10	[7, 7, 7, 7, 20, 28, 28, 28, 1, 7]	0,00
10	Mais	11	10	1	[18]	0,91
11	Astelpaju	10	5	5	[2, 3, 2, 2, 2]	0,50
12	Marjapõõsad ja viljapuud	12	9	3	[2, 2, 2]	0,75
13	Mustkesa	10	4	6	[20, 4, 28, 4, 16, 4]	0,40
14	Peakapsas	10	6	4	[20, 7, 28, 6]	0,60
15	Porgand	10	6	4	[6, 3, 3, 8]	0,60
16	Punapeet	10	1	9	[6, 15, 15, 19, 4, 28, 28, 17, 12]	0,10
17	Suviraps ja -rüps	14	14	0		1,00
18	Suvinisu ja speltanisu	11	9	2	[19, 5]	0,82
19	Suvioder	14	14	0		1,00
20	Kaer	10	8	2	[18, 8]	0,80
21	Tatar	10	9	1	[4]	0,90
22	Taliraps ja -rüps	12	12	0		1,00
23	Rukis	10	9	1	[24]	0,90

24	Talinisu	10	10	0		1,00
25	Talioder	10	10	0		1,00
26	Talitritikale	10	9	1	[23]	0,90
27	Muu teravili	10	9	1	[20]	0,90
28	Muu	10	6	4	[4, 25, 6, 17]	0,60
Kokku		298	239	59	Keskmine kaalumata saagis:	0,80

4.8 Soovitused ja mõttekohad põllukultuuride tuvastamise operatiivsüsteemi loomiseks

- Temperatuuri, sademete, mullastiku ja põllu asukoha andmed, sellisel kujul nagu neid käesolevas töös kasutati, tundavad omavat väga väikest mõju klassifitseerimisõigsusele, mistõttu nende kasutamine operatiivsetes teenustes ei pruugi olla otstarbekas. Eriti arvestades nende eeltöötluse kohmakust ja põhimõttelist erinevust satelliidiandmete töötlusest, mille jaoks on juba olemas toimiv automaatne töötlusahel. Sademe- ja temperatuuriandmestiku olulisus ja kasutatavus võib muutuda olulisemaks, kui kasutada neid kumulatiivselt, ehk summeerida hooaja algusest alates kuni satelliitpildi ülesvõtteni. Sellisel juhul sisaldaksid need rohkem infot antud hooaja klimaatiliste tingimuste kohta. Seda antud arendustöö käigus ei katsetatud.
- Kui arvutusvõimsus ja/või ressursikasutus seab piire kasutatavate Sentinel-1 ja -2 tunnuste arvule, on tunnuskomplekti võimalik oluliselt kärpida, säilitades seejuures mudeli hea sooritusvõime. Kindlasti peaks olema kaasatud tunnuseid nii optilise kui radarsatelliidi tunnuste hulgast, sest need täiendavad teineteist ja annavad koos paremaid tulemusi kui eraldi. Seda kinnitavad nii varasemad uurimistööd kui ka meie olulisemate tunnuste analüüs. Meie andmetel on kõige olulisemad tunnused 's0vv_median', 's0vh_median', 'ndvi_median', 'tc_vegetation_median' ja 'psri_median'.
- Suur osa taliteraviljadest (rukis, talinisu ja talioder) on ilmselt ka operatiivsüsteemis kõrge täpsusega eristatavad – nii eraldi kui ühise taliteraviljade klassina, juba hooaja keskel. Erandiks on talitritikale, mida mudel ennustab ka rukkiks ja talinisuks ning mille täpne klassifitseerimine võib olla keerukam.
- Suviteraviljade (suvinisu ja speltanisu, suvioder, kaer, tatar) omavahelise eristamise täpsus hooaja keskel (24. juuni) on selgelt kehvem kui augusti alguses ja enne augustit nende eristamine mõttekas pole. Ka hooaja lõpus ei küündi suviteraviljade saagised üle 0,9 (jäädes vahemikku 0,74-0,89). Operatiivsüsteemis võiks kaaluda nende eristamist ühise „suviteraviljade“ grupina, mille tuvastustäpsus on kõrge (saagis 0,97).

- Rapsi- ja rüpsi (nii tali- kui suvevormi) ning maisi tuvastamine tundub mudelile lihtsa ülesandena. Kui talirapsi võiks operatiivsüsteemis tuvastada juba juuni lõpus (pärast õitsemist), siis suvirapsi ja maisi tuvastamisega peaks ootama vähemalt augustini.
- Marjapuude ja viljapuude (sh astelpaju) tuvastamine satelliitseirega näib väga keeruline ning kaaluda tuleks teisi andmeallikaid, mis võivad töötada (nt aerolaserskaneerimise andmed, kust piisavalt tiheda punktipilve korral võiksid iseloomulikud võrade kujud välja joonistuda). Nende kultuuride istandike eripära tõttu, kus puude või põõsaste read vahelduvad rohuribadega, aetakse neid peamiselt just heintaimedega segamini.
- Köögiviljadest võiks operatiivsüsteemi kaasata kartuli, või luua uue suurema klassi, kuhu peale kartuli kuuluvad ka peakapsas, porgand ja punapeet. Kapsa, porgandi ja punapeedi eraldi tuvastamine oli hetkel ebatäpne ja neid aeti segamini erinevate klassidega, ent kõige rohkem just kartuliga.
- Põldoa ja põldherne võiks kaasata operatiivsüsteemi, nende tuvastamisel on mudel täpne.
- Hetkel on mudeli ennustuse etapis määratud põllule see kultuur, mille tõenäosus on mudeli arvates suurim. Näiteks, olgu meil 3 kultuuriga klassifikatsioon ja mudeli poolt tehtud hinnang konkreetse põllu klassikuuluvuse kohta oleks järgmine: kultuur 1 – 33%, kultuur 2 – 34%, kultuur 3 – 33%. Sellisel juhul omistatakse põllule kultuur 2 klass, kuigi on selge, et mudel on tegelikult üpris segaduses. Operatiivsüsteemis peaks kindlasti kehtestama lävendid, kui suur peab klassikuuluvuse tõenäosus olema, et see tõepoolest põllule määrataks. Kui mudeli tulemus jääb alla lävendi, liigub põld nö „kollasesse klassi“, mille kohta me ühest hinnangut anda ei oska. Sõltuvalt kasutusjuhust saab anda selliste põldude korral mitme klassikuuluvuse tõenäosused, mis on kõige tõenäolisemad kultuurid, mis antud põllul kasvavad, ning edasised otsused ja menetlemine jäävad PRIA spetsialistide teha.

5 Kokkuvõte

RITA1/02-52 „Kaugseire andmete kasutuselevõtt avalike teenuste väljatöötamisel ja arendamisel“ põllumaade alamprojekti raames töötati välja Eesti oludesse sobiv põllukultuuride tuvastamise metoodika. Peamiseks andmeallikaks olid Sentinel-1 ja -2 aegread, rakendati närvivõrkude masinõppe mudeleid. Töötati läbi kogu Eestit kattev 2018. ja 2019. aasta andmestik ning uuriti saadud täpsusi ja eksimise põhjuseid.

Valideerimiskogul arvutatud F1-skoordid olid 0,85 (kitsam 28 kultuurigrupiga jaotus) ja 0,91 (laiem 16 kultuurigrupiga jaotus). Enamlevinud kultuurigruppide puhul olid testkogu saagised (*recall*) üle 0,9 ning metoodika võib lugeda operatiivkasutusse võtmiseks küpseks. Väga soovitav on realiseerida see koos „kollase klassiga“, et väiksema usaldusväärsusega tulemused oleksid selgelt eristatud, otseseid vigu oleks vähe ning süsteemi lõppkasutajad saaksid mudeli väljundit usaldada.

Võrreldes mudeli varasema versiooniga (tulem D4.6) on käesolevas lõpparuandes kirjeldatud mudel seadistatud suuri ja väikseid kultuurigruppe võrdsemalt käsitlema. Kitsama klassifikatsiooni (28 kultuurigruppi) keskmine saagis tõusis tänu klasside tasakaalu arvestamisele 0.73-lt 0.78-le, kuid samas suurenes üle-eestiliselt potentsiaalselt valesti klassifitseeritavate põldude hulk 13 204-lt 18 582-le, sest osade suuremate klasside (nt. kõrrelised heintaimed ja suvioder) täpsus käesolevas mudeli versioonis langes. Operatiivkasutuse jaoks tuleb vastavalt PRIA vajadustele leida mudeli seadistamisel mõistlik tasakaal. Kas olulisem on kõigi klasside (ka vähelevinumate) võimalikult võrdne käsitlemine ja kõigi klasside puhul keskmise ja kõrgema täpsuse saavutamine või suurem täpsus just levinumate kultuurigruppide puhul?

Avalik prototüüptarkvara testandmestikul hinnangute andmiseks on alla laetav aadressilt: https://bitbucket.org/kappazeta/rita-evaluator/src/master/

Aruandes kirjeldatud eksimismaatriksid ja muud olulised joonised antakse üle ka eraldi failidena. Samuti tehakse kättesaadavaks PostgreSQL andmebaasid kogu andmestikuga nii 2018 kui 2019 aasta kohta.

6 Viited

D4_4 Põllukultuuride tuvastusmudeli kirjeldus koos täpsushinnangutega, 1. iteratsioon – sügis 2019

D4_5 Põllukultuuride tuvastusmudeli kirjeldus koos täpsushinnangutega, 2. iteratsioon – kevad 2020

D4_5_Lisa1_Pollukultuuride_tunnusvektorite_aegridade_rühmitamine_klasterdamise_abil_ja_eksete_eemaldamise_metoodika.docx

D4_6 Põllukultuuride tuvastusmudeli kirjeldus koos täpsushinnangutega, 3. iteratsioon – suvi 2020

Congalton, R. G., & Green, K. (2008). Assessing the Accuracy of Remotely Sensed Data: Principles and Practices, Second Edition (2 edition). CRC Press.