



Ontwikkeling van gebiedsdekkende kaartlagen van gemodelleerde bodemeigenschappen op basis van het bodemkoolstofmonitoringnetwerk Cmon

Kaartlagen van gemodelleerde bodemeigenschappen op basis van Cmon

Dit rapport beschrijft de methodologische uitwerking en de resultaten van het opstellen van gebiedsdekkende kaartlagen van gemodelleerde bodemeigenschappen op basis van de gegevens uit het Vlaamse bodemkoolstofmonitoringnetwerk Cmon.

COLOFON

Verantwoordelijke uitgever

Toon Denys, secretaris-generaal

Departement Omgeving

Koning Albert-II laan 15 bus 553, 1210 Brussel (postadres)

vlaanderen.be/omgeving

Een uitgave van het Departement Omgeving, Afdeling Beleid

beleid.omgeving@vlaanderen.be

Auteurs

Davor Josipovic - RoVaD

Katrien Oorts, Joost Salomez, Dries Luts en Sabine Buyle - Team bodem, Departement Omgeving

Publicatiedatum

Mei 2026

Depotnummer

D/2026/3241/187

Wijze van citeren

Departement Omgeving (2026). Ontwikkeling van gebiedsdekkende kaartlagen van gemodelleerde bodemeigenschappen op basis van het bodemkoolstofmonitoringnetwerk Cmon.

INHOUDSTAFEL

1	Inleiding	5
1.1	Symbolen en acronymen	6
2	Methodologie	7
2.1	Algemeen model	8
2.1.1	Verwachtingswaarde	9
2.1.2	Standaardfout	10
2.1.3	Distributie	10
2.1.4	Aleatorische variabiliteit	10
2.2	Samengesteld model	11
2.2.1	Verwachtingswaarde	13
2.2.2	Standaardfout	14
2.2.3	Distributie	15
2.2.4	Aleatorische variabiliteit	16
2.3	Verklarende covariaten	16
2.4	Model selectie	18
2.4.1	Generalisatiefout	18
2.4.2	Covariatenselectie	20
2.4.3	Hyperparameters	21
2.5	Analyse en reproduceerbaarheid	21
2.6	Aanpassingen	23
2.6.1	Vergelijking met Seedling	23
2.6.2	Vergelijking met Wittemans et al. (2023)	25
3	Resultaten.....	26
3.1	pH (KCl)	27
3.1.1	0-10cm	27
3.1.2	10-30cm	28
3.1.3	30-60cm	28
3.1.4	60-100cm	29
3.2	Organische C	29
3.2.1	0-10cm	29
3.2.2	10-30cm	30
3.2.3	30-60cm	30
3.2.4	60-100cm	31
3.3	Organische C voorraden	31
3.3.1	0-30cm	31
3.3.2	0-100cm	32

1 INLEIDING

Gebiedsdekkende bodemdata, zoals het percentage klei, zand en leem, bulkdensiteit, organische koolstof (zowel in percentage als voorraad) en pH (KCl), vormen een essentiële basis voor het begrijpen van de bodem als systeem. Deze parameters beschrijven samen belangrijke fysische en chemische eigenschappen van de bodem en geven inzicht in hoe bodems functioneren en reageren op gebruik en beheer. Doordat deze gegevens ruimtelijk continu beschikbaar zijn, laten ze toe om verschillen binnen Vlaanderen in kaart te brengen en gericht te analyseren.

Daarnaast spelen deze bodemdata een cruciale rol in modellering. Ze vormen essentiële invoer voor modellen die processen zoals waterinfiltratie, afstroming, nutriëntentransport, koolstofopslag en gewasgroei simuleren. Door het gebruik van gebiedsdekkende datasets kunnen dergelijke modellen ruimtelijk expliciet worden toegepast, wat leidt tot meer nauwkeurige voorspellingen en scenarioanalyses. Dit maakt het mogelijk om effecten van klimaatverandering, landgebruiksveranderingen en beheermaatregelen te evalueren en ondersteunen ze zo een meer onderbouwde en toekomstgerichte besluitvorming.

Dit rapport beschrijft de methodologische uitwerking en de resultaten voor het opstellen van gebiedsdekkende *Digital Soil Maps* (DSM) op basis van de data van het Vlaamse koolstofmonitoringnetwerk Cmon. Het vormt een inhoudelijke en technische voortzetting van de DSM-benadering zoals ontwikkeld in Wittemans et al. (2023), aangevuld met inzichten uit het Europese EJP SOIL-programma (Tondini, 2026).

Wittebans et al. (2023) bevat een gedetailleerde analyse van methodologieën voor de kartering van bodemdescriptoren en onderbouwt de keuze voor een Random Forest-gebaseerde DSM-benadering. De implementatie is echter slechts een eerste operationele iteratie. Verschillende methodologische aspecten – waaronder modelvalidatie, onzekerheidskwantificatie, covariatenselectie en reproduceerbaarheid – zijn vatbaar voor verdere verfijning. In dit rapport zetten we dit werk verder.

Tondini (2026) is het resultaat van EJP SOIL werkpakket WP6.2, waarin de DSM-workflow van ISRIC – genaamd Seedling en ontwikkeld als voortzetting van SoilGrids 2.0 – gebruikt werd. Deze workflow integreert onder meer dieptemodellering, systematische covariatenselectie, cross-validatie en expliciete onzekerheidskwantificatie van voorspellingen. Hoewel het een methodologisch meer uitgewerkte en geharmoniseerde DSM-workflow biedt, zijn er voor de Vlaamse context toch aanpassingen aan de generieke workflow nodig.

Dit rapport beschrijft geen finale methodologie, maar een momentopname van de huidige inzichten en methodologische keuzes. Het document dient daarom niet alleen als een onderzoeksrapport, maar ook als een referentiekader voor toekomstige DSM-implementaties binnen de Vlaamse context.

De inhoud van dit rapport is als volgt ingedeeld. Sectie 2 beschrijft de huidige DSM-methodologie. Sectie 3 presenteert de resultaten op basis van deze methodologie. Sectie 4 valideert deze resultaten. Sectie 5 tot slot bespreekt de beperkingen van de huidige aanpak en verkent mogelijke pistes voor toekomstige ontwikkelingen.

1.1 SYMBOLEN EN ACRONYMEN

Hoewel alle symbolen en acroniemen bij hun eerste gebruik zijn gedefinieerd, geven we hier als referentie een beknopt overzicht met een korte uitleg (Tabel 1 & Tabel 2).

Tabel 1: Overzicht van veelgebruikte symbolen

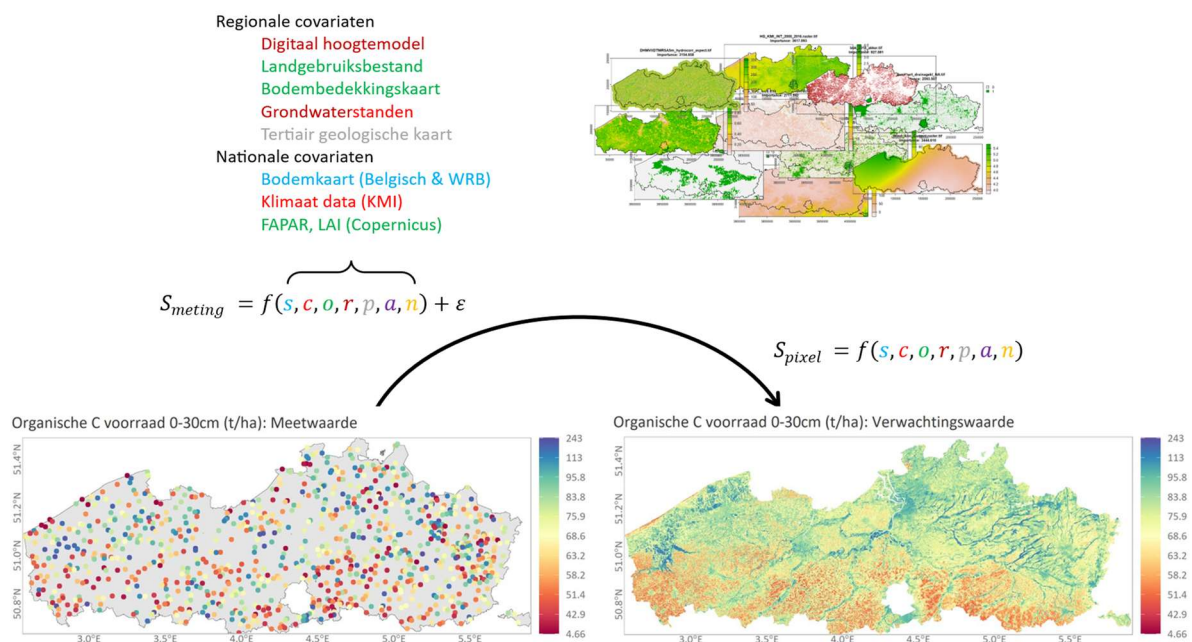
$Y, \mathcal{F}, y$	Stochastische variabele en bijhorende (theoretische) distributie van alle metingen van een bodemdescriptor binnen Vlaanderen, en y een willekeurige trekking uit die distributie.
$Y_l, \mathcal{F}_l, y_l$	Stochastische variabele en bijhorende (theoretische) conditionele distributie van alle herhaalde metingen op een specifieke fysieke locatie l , d.w.z. $Y_l = Y \mid l$ (i.e., Y conditioneel de locatie), en y_l een willekeurige trekking uit die distributie.
$\mu_l, \sigma_l^2, \varepsilon_l$	Eerste moment en tweede centrale moment van Y_l . DSM-model (Vergelijking 1) stelt dat $Y_l = \mu_l + \varepsilon_l$ zodat $\mathbb{E}[Y_l] = \mu_l$ en $\text{Var}[Y_l] = \sigma_l^2 = \text{Var}[\varepsilon_l]$.
n	Het aantal observaties. Dit is 990, het huidig aantal Cmon metingen op een bepaalde diepte, tenzij anders vermeld.
λ_l	Set van SCORPAN covariaten op fysieke locatie l .
L, l	Set van alle fysieke locaties l binnen een bepaald landgebruik. Let wel: binnen het DSM-model (Vergelijking 1) wordt een locatie volledig bepaald door λ_l . Met andere woorden, twee fysiek verschillende locaties kunnen door het model als identiek beschouwd worden indien hun λ_l identiek zijn.
B	Het aantal bomen binnen een Random Forest model.
$f, \hat{f}, f^*$	Geïdealiseerde modelfunctie $f(\cdot; \theta)$ geparametriseerd door θ waarvan f de verkorte notatie is. Dan is $\hat{f} := f(\cdot; \hat{\theta})$ en $f^* := f(\cdot; \theta^*)$ waarbij $\hat{\theta}$ en θ^* de optimale parameters zijn op basis van respectievelijk een steekproef en de populatie.
$\mathcal{L}$	Objectieve functie, soms ook verlies- of kostfunctie genoemd. Is de maatstaf die bepaalt hoe de afwijking tussen modelvoorspellingen en waarnemingen wordt geëvalueerd, en dus welk model als “beste” wordt beschouwd. Zie sectie 2.1.1.
c^*	Het optimaal aantal covariaten in een DSM-model. Zie sectie 2.4.2.
K, R, m_k	Het aantal folds (K) en repetities (R) binnen R-CV. m_k is het aantal observaties binnen een k^{de} fold. Om notatie te verlichten wordt aangenomen dat $\forall k: m_k = m$.
$\mathcal{T}, \mathcal{T}_k, \mathcal{T}_k^c$	Volledige Cmon-dataset met bodemdescriptormetingen en bijhorende covariaten λ_l per locatie l . $\mathcal{T}_k$ is een indexverzameling van $\mathcal{T}$ gebruikt als trainingsdata in de k^{de} fold van CV, en $\mathcal{T}_k^c$ het complement ervan gebruikt als testdata.

Tabel 2: Overzicht van veelgebruikte acroniemen

ALR	Additive Log-Ratio (sectie 2.2)
CV, R-CV	(Repeated) Cross-validation (sectie 2.4)
GRF	Generalized Random Forest (sectie 5)
MC	Monte Carlo
MSE, RMSE	(Root) Mean Squared Error (sectie 2.1.1 & 4.1)
MSPE, RMSPE	(Root) Mean squared prediction error (sectie 2.1.4)
RF	Random Forest (sectie 2.1)
RFE	Recursive feature elimination (sectie 2.4.2)

2 METHODOLOGIE

Om Cmon-plotmetingen te vertalen naar gebiedsdekkende kaarten wordt gebruikgemaakt van een gestandaardiseerde DSM-workflow. Drie stappen in deze workflow zijn essentieel voor modellering. De eerste stap is het verzamelen van verklarende variabelen (covariaten) die de ruimtelijke patronen helpen verklaren. De tweede stap is het modelleren van bodemkenmerken waarbij de Cmon meetwaarden gerelateerd worden aan de geselecteerde covariaten door middel van een functie (i.e., het DSM-model). In de laatste (derde) stap wordt deze functie gebruikt om waarden op rasterbasis te voorspellen. Dit leidt tot continue kaarten van bodemdescriptoren zoals organische C, pH, etc. Dit proces is schematisch voorgesteld in Figuur 1.



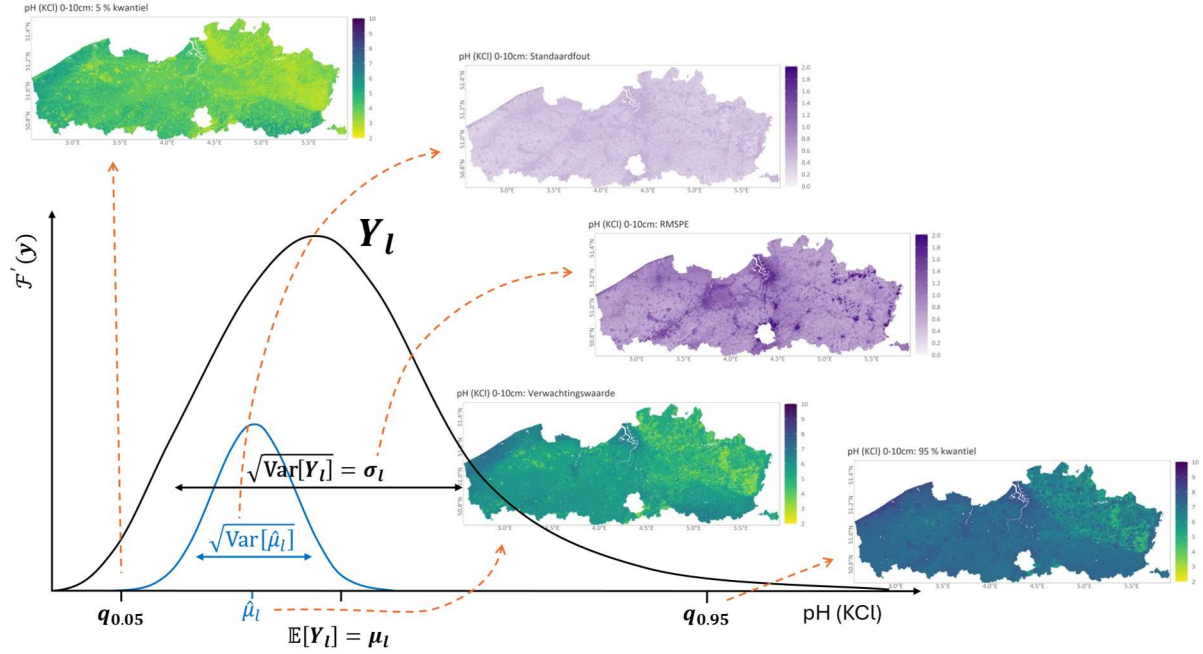
Figuur 1: Schematische voorstelling van de Cmon DSM-workflow. Linksonder: Cmon resultaten van de organische C voorraad 0-30cm (t/ha) die als input dienen (Oorts et al., 2025a). Boven: regionale en nationale covariaten die als verklarende variabelen worden gebruikt. (Bronnen van covariaten zijn ter illustratie en niet volledig.) Deze covariaten worden gecombineerd in een model dat vervolgens voorspellingen maakt van de organische C voorraad op rastercellen van 10×10 m (rechtsonder).

In de volgende secties zijn deze stappen in meer detail uitgewerkt. In sectie 2.1 is het DSM-model formeel gedefinieerd, samen met verwachtingswaarden, standaardfouten en predictieve

2.1 ALGEMEEN MODEL

$$Y_l = \mu_l + \varepsilon_l \quad (1)$$

Voor besluitvorming is het in de eerste plaats belangrijk om de ware waarde, uitgedrukt als de verwachtingswaarde $\mathbb{E}[Y_l] = \mu_l$, zo goed mogelijk te bepalen. Hieruit kan dan bijvoorbeeld de totale koolstofvoorraad binnen een landgebruik L exact bepaald worden als $\sum_{l \in L} \mu_l$, wat ook één van de primaire doelstellingen is van het Cmon meetnet. Echter, in praktijk is de ware μ_l onbekend, maar DSM-modellen leveren wel een schatter voor μ_l , nl. $\hat{\mu}_l$. Vandaar dat het essentieel is om ook de standaardfout van deze schatter $\hat{\mu}_l$ te rapporteren aangezien deze een indicatie geeft in welke buurt van $\hat{\mu}_l$ de ware μ_l zich bevindt. De ware μ_l is vrij abstract en moeilijk verifieerbaar, vandaar dat men in de praktijk ook geïnteresseerd is in de spreiding van meetwaarden op een locatie. Die kan op twee manieren worden beschreven. Eerste mogelijkheid is het schatten van de distributie $\mathcal{F}_l$ van Y_l . Hieruit kunnen vervolgens kwantielen $q_{0.05}$ en $q_{0.95}$ afgeleid worden die dan samen het 90 % predictie interval bepalen voor een toekomstige meetwaarde op die locatie. Een tweede mogelijkheid is het rechtstreeks schatten van de aleatorische variabiliteit $\text{Var}[Y_l] = \sigma_l^2$. Afhankelijk van de toepassing kan één van de twee de voorkeur hebben. Overzicht van deze begrippen is gegeven in Figuur 2.



Figuur 2: Overzicht van de belangrijkste begrippen van de conditionele distributie Y_l op een locatie l . pH (KCl) op 0-10cm diepte is hier als voorbeeld genomen. De kleurgradaties in de kaarten zijn lineair. Voor de verwachtingswaarde en de twee kwantielen is de schaal gelijkgesteld om directe vergelijking te vergemakkelijken. Zo ook voor de standaardfout en RMSPE.

Random Forest (RF) modellen voldoen doorgaans aan al de bovenvermelde vereisten. Voor dit project wordt gebruik gemaakt van *ranger* (Wright, 2017) als de basis RF-implementatie. In de volgende secties wordt elk van deze concepten in meer detail besproken en wordt ook de theorie, op basis waarvan ze berekend zijn, gegeven.

2.1.1 Verwachtingswaarde

Zoals vermeld in de vorige sectie is de primaire doelstelling van het DSM-model het schatten van de locatie-conditionele verwachtingswaarde $\mathbb{E}[Y_l] = \mu_l$ van een bodemdescriptor. Wanneer een (niet-lineaire) regressiefunctie f wordt gefit door minimalisatie van de gekwadrateerde fout, komt de optimale functie f^* op populatieniveau net overeen met deze verwachtingswaarde. Formeel geldt:

$$f^* = \arg \min_f \left(\mathbb{E}_{Y,l} \left[\underbrace{(Y - f(\lambda_l))^2}_{\mathcal{L}} \mid l \right] \right) \rightarrow f^*(\lambda_l) = \mathbb{E}[Y_l] = \mu_l \quad (2)$$

Om deze gelijkheid te garanderen, mag de bodemdescriptor niet getransformeerd worden en moet de objectieve functie $\mathcal{L}$ de gekwadrateerde fout zijn. In de brede literatuur wordt soms ofwel een log-transformatie toegepast op geobserveerde y – omdat wordt aangenomen dat Y_l lognormaal is –, ofwel wordt het absoluut verschil als $\mathcal{L}$ gebruikt. In beide gevallen resulteert deze correctieve ingreep in een f^* die niet langer gelijk is aan de locatie-conditionele verwachtingswaarde, maar aan de locatie-conditionele mediaan, en ondergraaft ze daarmee van meet af aan de doelstelling van het schatten van de verwachtingswaarde.

In de praktijk beschikken we uiteraard niet over de volledige populatie, maar slechts over een steekproef. Daarnaast kan het model doorgaans ook niet alle spatiale variabiliteit verklaren. Het model levert daarom met de bovenvermelde methode (Vergelijking 2) een schatter $\hat{f}$ voor de geïdealiseerde f^* . Vervolgens kunnen we ons afvragen in welke mate deze schatter een consistente en unbiased schatter is van de ware μ_l . Voor Random Forest-modellen zijn deze eigenschappen – consistentie en unbiasedness – niet triviaal en pas relatief recent formeel bewezen voor een

specifieke vorm genaamd *Generalized Random Forests* (GRF). Deze aspecten worden verder besproken in sectie 5.

2.1.2 Standaardfout

Het doel van deze sectie is het kwantificeren van de onzekerheid op de geschatte locatie-conditionele verwachtingswaarde $\hat{\mu}_l$. Volgens de maximum likelihood theorie is deze schatter – onder gebruikelijke aannames – asymptotisch normaal verdeeld (cfr. GRF, sectie 5). In dat geval volstaat het om de variantie $\text{Var}[\hat{\mu}_l]$ te bepalen om de spreiding rond de ware μ_l te karakteriseren. In het kader van RF-modellen wordt deze onzekerheid doorgaans bepaald via een variantieschatter gebaseerd op de *Jackknife-after-Bootstrap* geïntroduceerd door Efron (1992). Specifiek voor de berekeningen in dit project focussen we ons op *Infinitesimal Jackknife* (IJ) voor RF uitgewerkt door Wager et al. (2014).

De toepassing van deze IJ-variantieschatter brengt echter twee praktische problemen mee. Een eerste probleem is dat de IJ-variantieschatter biased is. Hoewel de bias analytisch bepaald en dus gecorrigeerd kan worden, kan deze correctie in sommige gevallen leiden tot negatieve variantieschattingen. Het aantal negatieve variantieschattingen hangt af van het aantal observaties n en van het aantal bomen B , en resulteert bij wijze van illustratie voor dit project in zo'n 40 % negatieve varianties bij $B = 500$, en 17 % bij $B = 2000$.

Een tweede probleem betreft de oplossing die in de *ranger* package wordt toegepast om negatieve varianties te corrigeren. Hierbij wordt een empirische Bayesiaanse correctie (cfr. “Empirical Bayes calibration of noisy variance estimates” in [ranger](#) source code) toegepast op de schatters. Omdat de predicties in kleine blokken worden uitgevoerd, gebeurt deze correctie onafhankelijk per blok. Hierdoor kunnen de variantieschattingen tussen aangrenzende blokken systematisch verschillen wat in ruimtelijke toepassingen kan leiden tot zichtbare strookpatronen.

Om deze artefacten te vermijden, wordt in dit project gebruik gemaakt van de ongecorrigeerde IJ-variantieschatting. De bias wordt hierbij gecontroleerd door het aantal bomen B voldoende groot te kiezen, typisch ongeveer zes maal het aantal trainingspunten n . Deze strategie heeft echter ook een belangrijk nadeel: naarmate de omvang van de Cmon-dataset toeneemt, moet het aantal bomen proportioneel worden verhoogd om dezelfde stabiliteit van de variantieschatting te behouden wat op termijn de computationele kosten significant verhoogt. De ontwikkeling van efficiëntere methoden voor onzekerheidskwantificatie binnen RF-gebaseerde DSM-modellen blijft daarom een belangrijk onderwerp voor verder onderzoek.

2.1.3 Distributie

In sommige toepassingen is het wenselijk om de volledige conditionele distributie van de bodemdescriptor op een locatie te kennen. In dit project wordt de conditionele distributie $\mathcal{F}_l$ van Y_l geschat aan de hand van de theorie uitgewerkt in Meinshausen (2006). Deze methode is reeds geïmplementeerd in de *ranger* package – weliswaar als een Monte Carlo benadering (MC-benadering) van het oorspronkelijke idee – en wordt in dit project zonder verdere aanpassingen gebruikt. Standaard worden de kwantielen $q_{0.05}$, $q_{0.5}$ en $q_{0.95}$ berekend, waarbij $q_{0.05}$ en $q_{0.95}$ samen een 90 % predictie-interval vormen voor een toekomstige meting op locatie l , en $q_{0.5}$ overeenkomt met de locatie-conditionele mediaan.

2.1.4 Aleatorische variabiliteit

Tot slot is voor sommige toepassingen ook de schatting van de aleatorische variabiliteitsfunctie $\text{Var}[Y_i] = \sigma_i^2$ van belang. In principe kan deze rechtstreeks worden afgeleid uit de MC-samples van de geschatte distributie $\hat{F}_i$ uit de vorige sectie. In de praktijk is deze benadering echter zeer gevoelig

voor starten van $\hat{\mathcal{F}}_l$, die mogelijk sterk worden bepaald door keuzes in hyperparameters. Om deze reden wordt in dit project gekozen voor een alternatieve benadering gebaseerd op de volgende identiteit op populatieniveau:

$$\text{Var}[Y_l] = \mathbb{E}[(Y_l - \mu_l)^2] = \mathbb{E}[\varepsilon_l^2] = \text{Var}[\varepsilon_l] = \sigma_l^2 \quad (3)$$

Dit suggereert een tweestapsprocedure voor het modelleren van de variantie. In een eerste stap wordt de verwachtingswaarde μ_i geschat. In een tweede stap worden de gekwadrateerde residuen gemodelleerd als functie van dezelfde covariaten. Dit principe is in de econometrie bekend als tweestapsprocedure voor gegeneraliseerde kleinste kwadraten (Greene, 2003) en is ook onderzocht voor niet-parametrische regressie (Fan & Yao, 1998).

Een belangrijke kanttekening bij deze aanpak is dat in de praktijk de ware verwachtingswaarde μ_l wordt vervangen door de schatter $\hat{\mu}_l$. De residuen worden dus berekend als $\hat{\varepsilon}_l = y_l - \hat{\mu}_l$. Het modeleren van $\mathbb{E}[\hat{\varepsilon}_l^2]$ komt daardoor overeen met het modelleren van de *mean squared prediction error* (MSPE, cfr. Mendez & Lohr, 2011) in plaats van $\mathbb{E}[\varepsilon_l^2]$. Onder de aanname dat ε_l en $\hat{\mu}_l$ onafhankelijk zijn en dat het model unbiased is, kan de volgende relatie tussen de twee worden afgeleid:

$$\begin{aligned}\mathbb{E}[\hat{\varepsilon}_l^2] &= MSPE_l = \mathbb{E}[(Y_l - \hat{\mu}_l)^2] \\ &= \mathbb{E}[\varepsilon_l^2] + (\mathbb{E}[\hat{\mu}_l] - \mu_l)^2 + \mathbb{E}[(\hat{\mu}_l - \mathbb{E}[\hat{\mu}_l])^2] \\ &= \sigma_l^2 + \text{Bias}^2[\hat{\mu}_l] + \text{Var}[\hat{\mu}_l]\end{aligned}\quad (4)$$

Vergelijking 4 toont aan dat het rechtstreeks modelleren van de gekwadeerde residuen $\hat{\epsilon}_l^2$ resulteert in een systematisch opwaarts vertekende schatting van de aleatorische variantie σ_l^2 – wat trouwens ook het geval is indien de variantie zou afgeleid worden uit $\hat{\mathcal{F}}_l$. Bij interpretatie van de MSPE-kaarten moet men daarom bewust zijn van dit effect. Indien een zuivere schatting van de aleatorische variabiliteit σ_l^2 vereist is, moet voor deze bias expliciet worden gecorrigeerd:

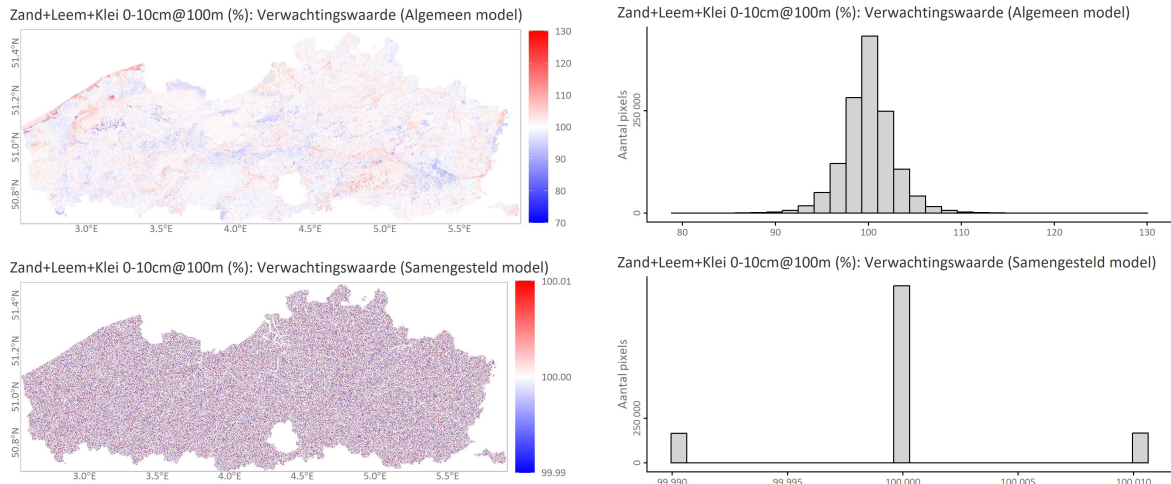
$$\hat{\sigma}_l^2 = MSPE_l - \text{Var}[\hat{\mu}_l] - \text{Bias}^2[\hat{\mu}_l]$$

Kortom, $MSPE_l$ is rechtstreeks gemodelleerd en dus gekend, $\text{Var}[\hat{\mu}_l]$ is berekenbaar (sectie 2.1.2), terwijl $\text{Bias}^2[\hat{\mu}_l]$ in de praktijk niet identificeerbaar is, waardoor een volledige biascorrectie doorgaans niet haalbaar is.

Een belangrijke voorwaarde in Vergelijking 4 is de onafhankelijkheid tussen ε_l en $\hat{\mu}_l$. Deze voorwaarde wordt in de praktijk benaderd door gebruik te maken van out-of-bag (OOB) schatters, waarbij Y_l niet gebruikt wordt bij het bepalen van $\hat{\mu}_l$. Indien deze onafhankelijkheid niet geldt, verschijnt er een bijkomende term in de decompositie, namelijk $2\mathbb{E}[\varepsilon_l(\mu_l - \hat{\mu}_l)]$. Deze term is in het algemeen niet identificeerbaar. Voor het eenvoudigste model zonder covariaten, waarbij $\hat{\mu}_l$ gelijk is aan het steekproefgemiddelde, kan aangetoond worden dat deze term gelijk is aan $-\sigma_l^2/n$ als gevolg van de Bessel-correctie. In dat geval leidt de MSPE tot een systematische onderschatting van de ware variantie σ_l^2 . Dit benadrukt dan ook het belang van out-of-sample voorspellingen voor een consistente schatting van de aleatorische variabiliteit.

2.2 SAMENGESTELD MODEL

Sommige bodemdescriptoren zoals zand, leem en klei (Z_l, L_l, K_l) zijn compositionele bodemdescriptoren. Dit betekent dat hun meetwaarden op een locatie proporties van een geheel zijn en dus voldoen aan de voorwaarde $Z_l + L_l + K_l = 100\%$. Wanneer deze bodemdescriptoren apart gemodelleerd worden, gebruikmakend van het DSM-model uit Vergelijking 1, wordt deze gelijkheid niet noodzakelijk gerespecteerd (cfr. Figuur 3).



Figuur 3: Vergelijking van $\mathbb{E}[Z_l] + \mathbb{E}[L_l] + \mathbb{E}[K_l]$ op basis van algemeen enkelvoudig model (boven) en samengesteld model (onder) bij 100×100m. Het gebruik van algemeen enkelvoudig model induceert ongeveer 3 % fout op totaal terwijl de precisie van het samengesteld model overeenkomt met de ingestelde precisie (± 0.01 %, cfr. Tabel 6)

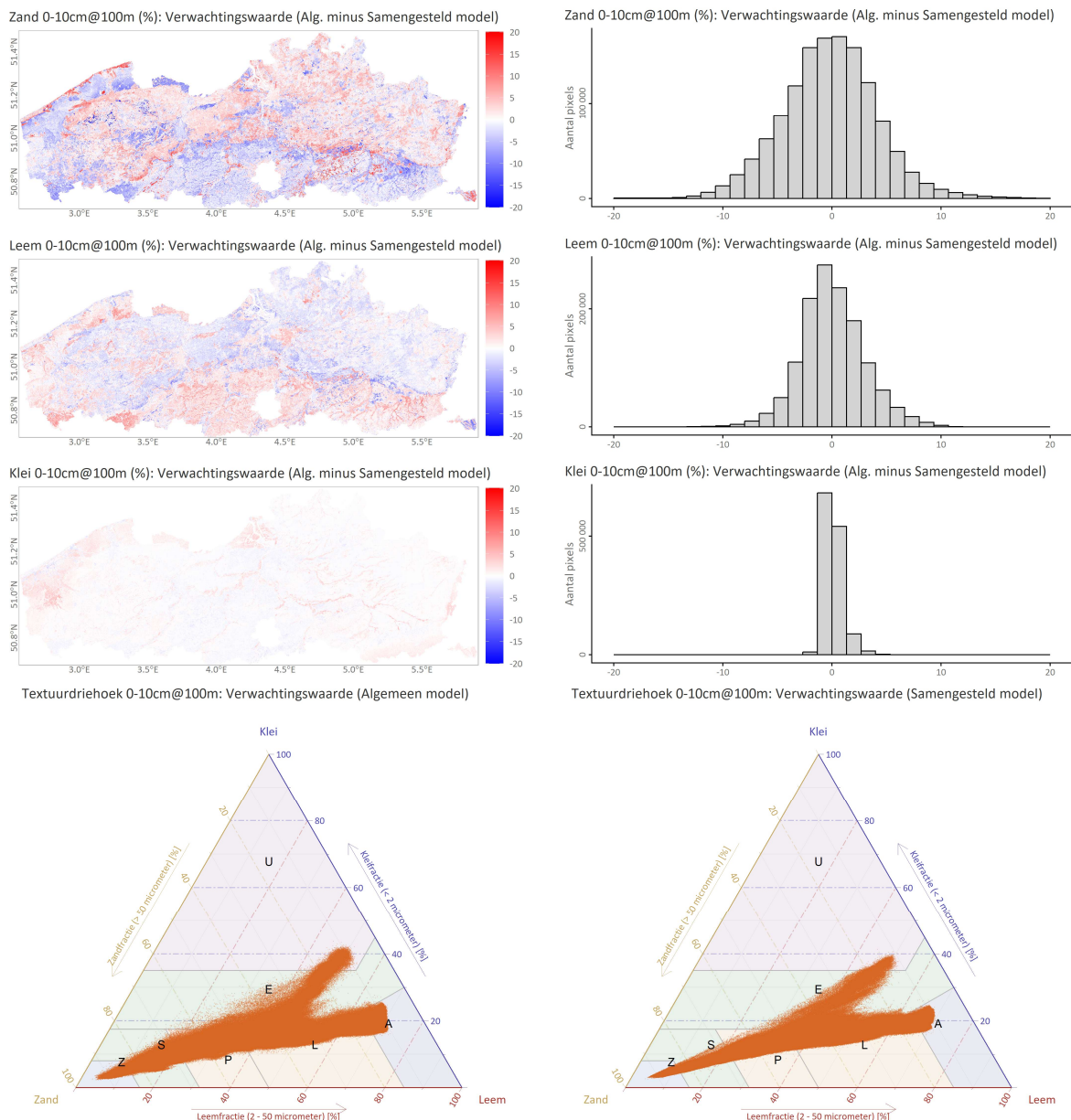
Een gebruikelijke oplossing voor het modeleren van deze descriptoren is om de proporties – dit is, elementen van een positief simplex $\mathbb{S}^2$ – te transformeren naar een onbegrensde $\mathbb{R}^2$ ruimte via een specifieke transformatiefunctie (Aitchison, 1982). De interpretatie is dan dat de drie descriptoren (Z_l, L_l, K_l) realisaties zijn van een onderliggende tweedimensionale latente variabele (A_l, B_l), verkregen via een specifieke transformatiefunctie. Deze aanpak wordt gebruikt in SoilGrids 2.0 (Poggio et al., 2021) en in Seedling (Genova et al., 2024), waarop dit project verder bouwt.

In praktijk vertaalt deze aanpak zich in het afzonderlijk modeleren van A_l en B_l door twee DSM-modellen op basis van Vergelijking 1. Hierdoor kunnen de conditionele marginale distributies $\mathcal{F}_A$ en $\mathcal{F}_B$ geschat worden (cfr. sectie 2.1.3), maar niet hun gezamenlijke afhankelijkheidsstructuur. Impliciet wordt dus aangenomen dat de aleatorische componenten van A_l en B_l onafhankelijk zijn.

Daarnaast is er nog de vraag welke transformatiefunctie het best gebruikt wordt. Er bestaan verschillende bijectieve transformatiefuncties voor het projecteren van een simplex naar een Euclidisch vlak en omgekeerd. Aitchison (1982) bespreekt er drie, waarvan *additive log-ratio* (ALR) reeds wijdverspreid is (cfr. Tabel 3). De transformatiefunctie is bepalend voor de joint distributie van (A_l, B_l). Aangezien onze aanpak de marginale distributies al kan modelleren (cfr. sectie 2.1.3), heeft de keuze van een transformatiefunctie voornamelijk betrekking op de projectie van de afhankelijkheidsstructuur tussen A_l en B_l . We gaan er hier van uit dat deze structuur minder bepalend is dan de marginale structuur, wat het belang van een transformatiefunctie beperkt. Vandaar gemakshalve de keuze voor de ALR-functie. Het expliciet modelleren van de volledige afhankelijkheidsstructuur of het selecteren van een transformatiefunctie die het minst bepalend is door deze structuur vereist verder onderzoek, en zou een mogelijke uitbreiding vormen van het huidige kader.

Tabel 3: *Additive log-ratio* (ALR) transformatiefunctie (links) en de inverse (rechts). De functie g is de verkorte notatie voor de inverse ALR-functie

$A_l = \log \frac{Z_l}{K_l}$ $B_l = \log \frac{L_l}{K_l}$	$Z_l = \frac{e^{A_l}}{1 + e^{A_l} + e^{B_l}} = g_Z(A_l, B_l)$ $L_l = \frac{e^{B_l}}{1 + e^{A_l} + e^{B_l}} = g_L(A_l, B_l)$ $K_l = \frac{1}{1 + e^{A_l} + e^{B_l}} = g_K(A_l, B_l)$
---	---



Figuur 4: Verschil in geschatte verwachtingswaarde tussen algemeen model en samengesteld model voor zand, leem en klei op 0-10cm diepte op gelijke schaal om vergelijking te vergemakkelijken. Voor de textuurdriehoek zijn de schattingen van het algemeen model genormaliseerd met $z + l + k$, zodat de som van hun componenten altijd gelijk is aan 100 %

2.2.2 Standaardfout

In sectie 2.1.2 stelden we dat de schatting van $\text{Var}[\hat{\mu}_l]$ de onzekerheid van $\hat{\mu}_l$ volledig karakteriseert op voorwaarde dat deze (asymptotisch) normaal verdeeld is. In het samengesteld model worden de bodemfracties echter bekomen via een niet-lineaire inverse ALR-transformatie van de onderliggende variabelen A_l en B_l . Daardoor is de schatter $\hat{z}$ doorgaans niet meer asymptotisch normaal verdeeld. De variantie $\text{Var}[\hat{z}]$ blijft wel een betekenisvolle maat voor de spreiding van de schatter, maar kan niet meer rechtstreeks gebruikt worden om symmetrische betrouwbaarheidsintervallen af te leiden.

Ervan uitgaand dat $\hat{a}$ en $\hat{b}$ asymptotisch normaal verdeeld zijn, biedt de delta-methode een computationeel efficiënte manier om de propagatie van onzekerheden door een niet-lineaire transformatie te benaderen. Concreet voor $\hat{z}$ kan de variantie benaderd worden door:

Er zijn in Vergelijking 6 enkele belangrijke zaken om bij stil te staan. De eerste benadering ($\hat{\approx}$) steunt op de huidige beste benadering $\mathcal{Z} \approx g_Z(a, b)$. Het is een benadering daar de strikte gelijkheid volgens Vergelijking 5 niet opgaat. Dit is dan ook een huidige beperking waar verder onderzoek naar nodig is. Bijvoorbeeld, een betere benaderingsfunctie, zoals afgeleid in Lark & Bishop (2007), zou in de plaats van g_Z gebruikt kunnen worden. De tweede benadering ($\hat{\approx}$) is te wijten aan de delta-methode zelf. Simulaties tonen dat de nauwkeurigheid van deze tweede benadering kan afnemen wanneer de voorspelde bodemdescriptoren dicht bij de grenzen van de simplex liggen. Verder worden de partiële afgeleiden geëvalueerd in $\hat{a}$ en $\hat{b}$ en er wordt er impliciet van uitgegaan dat $\text{Cov}(\hat{a}, \hat{b}) = 0$, wat al besproken is in de inleiding voor samengestelde modellen (sectie 2.2).

MC-geïnspireerde methoden, zoals *resampling* of *jackknife-after-bootstrap*, zouden in principe ook gebruikt kunnen worden om de variantie van $\hat{z}$ te schatten. Deze methoden zijn echter aanzienlijk computationeel intensiever. Hoewel deze methodes asymmetrische onzekerheidsintervallen kunnen opleveren, introduceren ze bijkomende MC-variabiliteit en worden ze evengoed beperkt door $z \approx q_z(\alpha, \ell)$ bij gebrek aan een betere benaderingsfunctie.

De MC samples uit de samengestelde distributie (Z_l, L_l, K_l) zijn het resultaat van de inverse ALR-transformatie (Tabel 3) op MC samples uit $\hat{\mathcal{F}}_A$ en $\hat{\mathcal{F}}_B$ (cfr. sectie 2.1.3). Hierop worden vervolgens marginale kwantielen $q_{0.05}$, $q_{0.5}$ en $q_{0.95}$ berekend, waarbij $q_{0.05}$ en $q_{0.95}$ samen een 90 % predictie-interval vormen voor een toekomstige meting op locatie l , en $q_{0.5}$ overeenkomt met de locatie-conditionele mediaan. Let wel dat dit marginale kwantielen zijn, met andere woorden hun som is niet noodzakelijk 100 %.

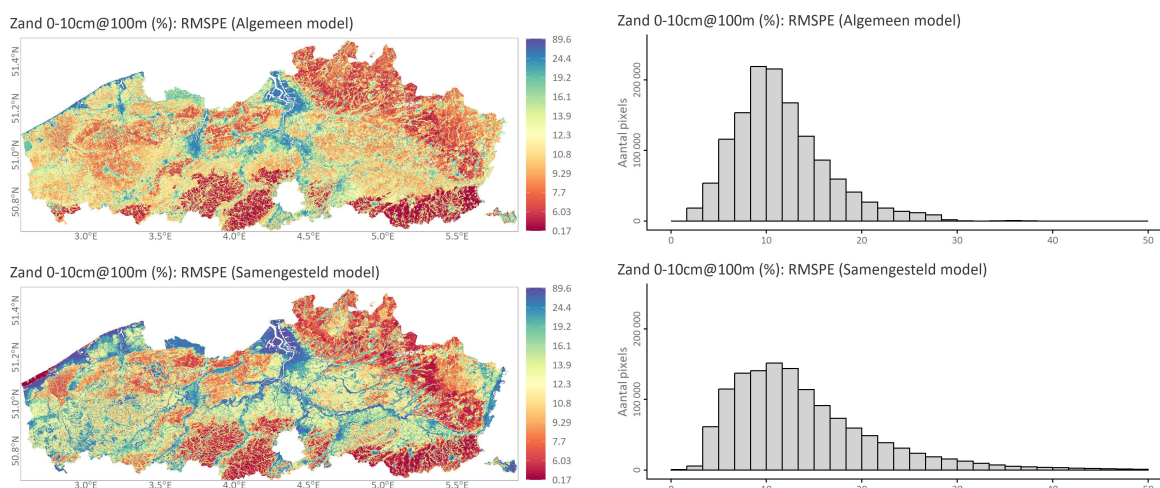
2.2.4 Aleatorische variabiliteit

Voor samengestelde modellen wordt de aleatorische variabiliteit, analoog aan sectie 2.1.4, uitgedrukt via MSPE. Voor een bodemfractie zoals zand geldt:

$$\text{Var}[Z_l] = \text{Var}[g_z(A_l, B_l)] \approx \left(\frac{\partial g_z}{\partial A_l}\right)^2 \text{Var}[A_l] + \left(\frac{\partial g_z}{\partial B_l}\right)^2 \text{Var}[B_l]$$

Deze benadering volgt rechtstreeks uit de delta-methode (**) en is qua vorm gelijkaardig aan Vergelijking 6 in sectie 2.2.2. De partiele afgeleiden zijn geëvalueerd in $\hat{a}$ en $\hat{b}$ en er wordt impliciet van uitgegaan dat $\text{Cov}(A_l, B_l) = 0$, zoals besproken in de inleiding voor samengestelde modellen (sectie 2.2).

Ter illustratie wordt in Figuur 6 de resulterende RMSPE-kaart vergeleken met deze bekomen uit het enkelvoudig model voor zand bij 0-10cm. Net zoals bij de standaardfouten (sectie 2.2.2) vertoont de distributie van de RMSPE in het samengesteld model een grotere spreiding, met meer uitgesproken extreme waarden. De orde van grootte blijft echter vergelijkbaar. Eenzelfde patroon wordt waargenomen voor leem en klei. Hier kan men in het algemeen dus ook besluiten dat in ruil voor het respecteren van compositievoorwaarde (som = 100%) de lokale onzekerheid lichtjes verhoogt.



Figuur 6: RMSPE van zand bij 0-10 cm op basis van 100x100m algemeen model (boven) en samengesteld model (onder). De kaart-kleuren zijn op basis van de kwantielen van de kaart van het samengesteld model om een directe vergelijking te vergemakkelijken

2.3 VERKLARENDE COVARIATEN

Het DSM-model in Vergelijking 1 maakt gebruik van verklarende variabelen, of covariaten, die als input dienden voor het model en de ruimtelijke patronen verklaren. Deze worden doorgaans ingedeeld volgens het SCORPAN-raamwerk (*Soil, Climate, Organisms, Relief, Parent material, Age, Spatial position*). Afhankelijk van de herkomst kunnen covariaten regionaal (Vlaanderen), nationaal (België) of Europees beschikbaar zijn. Deze indeling bepaalt ook de reikwijdte van de resulterende kaarten: zo kunnen regionale covariaten enkel worden toegepast voor kartering van Vlaanderen, terwijl Belgische of Europese covariaten een bredere toepasbaarheid toelaten (e.g., kartering van Wallonië of een Europees land op basis van Cmon data). Regionale covariaten zijn doorgaans representatiever dan Europese covariaten. Tabel 4 geeft een overzicht van de Vlaamse en nationale covariaten die gebruikt werden voor modelering.

Tabel 4: Overzicht van de 544 verklarende variabelen gebruikt door het DSM-model

Naam	Beschrijving	SCORPAN factor	Type	Aantal rasters	Resolutie
bbk_2018	Bodembedekkingskaart 2018	Organism	Categorical	6	1x1m
bodemkaart_drainagekl	Bodemkaart - drainage class	Soil	Categorical	16	1:20.000
bodemkaart_fase	Bodemkaart - fase	Soil	Categorical	8	1:20.000
bodemkaart_gegeneralali	Bodemkaart - Gegeneraliseerde legende	Soil	Categorical	37	1:20.000
bodemkaart_profielont	Bodemkaart - Profielontwikkelingsgroep	Soil	Categorical	16	1:20.000
bodemkaart_streek	Bodemkaart - Bodemstreek	Soil	Categorical	15	1:20.000
bodemkaart_substraa	Bodemkaart - Substraat	Soil	Categorical	38	1:20.000
bodemkaart_textuurk	Bodemkaart - Textuurklasse	Soil	Categorical	40	1:20.000
bodemkaart_variante_2	Bodemkaart - Variante van de profielontwikkelng	Soil	Categorical	19	1:20.000
bodemkaart_variante_v_b	Bodemkaart - Variante van het moedermateriaal	Soil	Categorical	40	1:20.000
DD_Current	Aantal droge dagen	Climate	Continuous		1:20.000
DHMVIIDTMRSA5m_hydrocorr	Digitaal hoogtemodel - hydrocorr	Relief	Continuous		5x5m
DHMVIIDTMRSA5m_hydrocorr_***	Digitaal hoogtemodel derivaten – Aspect, Cgene, Cprof, Cplan, Ctang, Clong, Ccros, Cmini, Cmaxi, Ctota, Croto, Flowdir, Roughness, Slope, Tpi, Tri, Twi swi	Relief	Continuous	17	5x5m
FAPAR_zomer_copernicus	FAPAR zomer Copernicus	Organism	Continuous		
FD_Current	Aantal vorstdagen	Climate	Continuous		
GHG	Gemiddeld Hoogste Grondwaterstand	Relief	Continuous		
GLG	Gemiddeld Laagste Grondwaterstand	Relief	Continuous		
GS_Current	Lengte groeiseizoen	Climate	Continuous		
HG_KMI_GEW_2000_2016	Gewicht hittegolven	Climate	Continuous		
HG_KMI_INT_2000_2016	Intensiteit hittegolven	Climate	Continuous		
HG_KMI_LEN_2000_2016	Lengte hittegolven	Climate	Continuous		
HG_KMI_NR_2000_2016	Aantal hittegolven	Climate	Continuous		
HGD_2000_2016	Aantal hittegolfdagen	Climate	Continuous		
HGD_NR_2000_2016	Hittestress (aantal hittegolfgraadaddagen)	Climate	Continuous		
HGD_overschrijding_2000_2016	Mate van overschrijding hittedrempel	Climate	Continuous		
LAI_zomer_copernicus	Leaf Area Index	Climate	Continuous		
LDS_current	Lengte droogteperiode	Climate	Continuous		
lu_landgebruik_vlaa_2022_v3_akker	Landgebruikskaart Vlaanderen 2022 versie 3	Organism	Categorical	11	10x10m
niveau1_vla_2022_v3_afgedekt	Landgebruiksbestand Vlaanderen 2022 versie 3 niveau 1	Organism	Categorical	21	10x10m
PET_Ann_Current	Totale verdamping jaar	Climate	Continuous		
PET_DJF_Current	Totale verdamping winter	Climate	Continuous		
PET_JJA_Current	Totale verdamping zomer	Climate	Continuous		
Pr_1yr_Current	Extreme neerslag eens per jaar	Climate	Continuous		
Pr_20yr_Current	Extreme neerslag eens in de 20 jaar	Climate	Continuous		
Pr_Ann_Current	Totale hoeveelheid neerslag jaar	Climate	Continuous		
Pr_DJF_Current	Totale hoeveelheid neerslag winter	Climate	Continuous		
Pr_JJA_Current	Totale hoeveelheid neerslag zomer	Climate	Continuous		
r_geologisch_typeov	Typeov (code voor de geologische formatie)	Parent material	Categorical	25	1:50.000
R20mm_current	Aantal dagen met zware neerslag	Climate	Continuous		

$$\text{Err}[n] = \mathbb{E}_{\mathcal{T}} \left[\mathbb{E}_{Y'_l, \lambda'_l} [\mathcal{L}(Y'_l, \hat{\mu}(\lambda'_l)) \mid \mathcal{T}] \right] \stackrel{\mathcal{L}=\hat{\varepsilon}_l^2}{=} \sigma_l^2 + \text{Bias}^2[\hat{\mu}] + \text{Var}[\hat{\mu}] \quad (7)$$

waarbij (Y'_l, λ'_l) een nieuwe observatie met locatiegegevens uit de populatiedistributie van Cmon voorstelt, $\mathcal{T}$ de dataset van grootte n gebruikt voor modelering, en $\mathcal{L}$ de objectieve functie. In sectie 2.1.1 werd geargumenteed in het voordeel van de kleinstekwadratenmethode, wat overeenkomt met de gekwadraterde fout als de objectieve functie $\mathcal{L}$. In dat geval kan de verwachte generalisatiefout worden ontbonden in bias en variantie van $\hat{\mu}$.

In eenvoudige termen stelt Vergelijking 7 dat de generalisatiefout overeenkomt met de verwachte gekwadraterde voorspellingsfout voor een nieuwe observatie, gemiddeld over alle mogelijke trainingssets van grootte n die uit de onderliggende populatie getrokken kunnen worden. Aangezien de volledige Cmon-populatie niet beschikbaar is, wordt deze grootte geschat op basis van de beschikbare steekproef. De theorie (Hastie et al., 2009, p161; Nadeau & Bengio, 2003) leert dat cross-validatie (CV) een unbiased schatter van de verwachte generalisatiefout $\text{Err}[n - m]$ levert:

$$CV_K = \frac{1}{K} \sum_{k=1}^K \frac{1}{m_k} \sum_{i \in \mathcal{T}_k^c} \mathcal{L}(\hat{f}_{\mathcal{T}_k}(\lambda_i), y_i) \stackrel{\mathcal{L}=\hat{\varepsilon}_l^2}{=} \frac{1}{n} \sum_{k=1}^K \sum_{i \in \mathcal{T}_k^c} \hat{\varepsilon}_i^2 = MSE_K \quad (8)$$

waarbij K het aantal folds is, m_k het aantal observaties binnen fold k , $\mathcal{T}_k$ en $\mathcal{T}_k^c$ respectievelijk de training en testindices van de k -fold, en $\hat{f}_{\mathcal{T}_k}$ de gefitte modelfunctie op $\mathcal{T}_k$ indices. Omdat $\mathcal{L}$ de gekwadraterde fout is (cfr. sectie 2.1.1) resulteert dit in de MSE -schatter op basis van K -fold cross-validatie. Om de notatie te vereenvoudigen wordt verder verondersteld dat het totaal aantal observaties n deelbaar is door K , wat maakt dat m gelijk is voor alle folds.

In dit project wordt gebruik gemaakt van *repeated cross-validation* (R-CV), waarbij de willekeurige partitionering van de dataset in K folds R maal wordt herhaald om de variabiliteit ten gevolge van één enkele fold-indeling te reduceren. De uiteindelijke schatter wordt dan gegeven door

$$CV_{K,R} = \frac{1}{R} \sum_{r=1}^R CV_K^{(r)} \quad (9)$$

In de implementatie van dit project wordt gebruik gemaakt van $K = 10$ folds en $R = 10$ repetities.

Omdat MSE zelf een schatter is, is de werkelijke generalisatiefout onbekend en onderhevig aan steekproefvariabiliteit. Om verschillende modellen met elkaar te kunnen vergelijken, is het belangrijk om ook de onzekerheid op deze schatter te kwantificeren. Verder bouwend op Vergelijking 9 kan de variantie van MSE formeel worden geschreven als:

$$\begin{aligned} \text{Var}[MSE] &= \text{Var} \left[\frac{1}{mKR} \sum_{r=1}^R \sum_{k=1}^K \sum_{i \in \mathcal{T}_{k,r}^c} \hat{\varepsilon}_i^2 \right] \\ &\stackrel{\hat{\xi}_{k,r} \text{ i.i.d.}}{=} \frac{1}{K^2 R^2} \left(KR \text{Var}[\hat{\xi}] + \sum_{i \neq j} \text{Cov}[\hat{\xi}_i, \hat{\xi}_j] \right) \\ &\stackrel{\hat{\varepsilon}_i^2 \text{ i.i.d.}}{=} \frac{\text{Var}[\hat{\varepsilon}^2]}{nR} \end{aligned} \quad (10)$$

waarbij $\hat{\xi}_{k,r} = \frac{1}{m_k} \sum_{i \in \mathcal{T}_{k,r}^c} \hat{\varepsilon}_i^2$.

De laatste term in Vergelijking 10 is de meest naïeve benadering die ervan uitgaat dat alle fouttermen $\hat{\varepsilon}_i^2$ onafhankelijk en identiek verdeeld zijn. Deze aanname is echter niet realistisch. Ten eerste werd in sectie 2.1 reeds uitgegaan van heteroskedasticiteit, waardoor de fouttermen niet

identiek verdeeld zijn. Daarnaast zijn in een cross-validatiesetting observaties uit verschillende folds niet onafhankelijk, aangezien de bijhorende modellen getraind worden op overlappende trainingsets. Hierdoor wordt de variantie systematisch onderschat.

Een meer realistische benadering wordt gegeven door de voorlaatste term in Vergelijking 10 waarbij gemiddelde fout $\hat{\xi}$ binnen elke fold (en repetitie) beschouwd wordt als identiek verdeeld, en de covariantie met een eerste-orde benadering bepaald wordt als $\text{Cov}[\hat{\xi}_i, \hat{\xi}_j] = m/n = 1/(K - 1)$. Deze aanpak, oorspronkelijk van Nadeau & Bengio (2003), leidt tot de volgende praktische schatter:

$$\text{Var}[MSE] = \left(\frac{1}{RK} + \frac{1}{K - 1} \right) \text{Var}[\hat{\xi}] \quad (11)$$

Voor de R-CV configuratie in dit project ($K = 10$, $R = 10$) resulteert deze Nadeau & Bengio (2003)-benadering in een variantieschatting die ongeveer 21 % hoger ligt dan de standaardbenadering in *glmnet* waar $\text{Var}[MSE] = \text{Var}[\hat{\xi}]/K$ wordt gebruikt. Het is ook belangrijk om op te merken dat deze schatter een heuristische benadering is, en niet noodzakelijk unbiased. Later werk van Bengio & Grandvalet (2004) toont aan dat er in het algemeen geen unbiased schatter bestaat voor de variantie van CV. Desondanks vormt deze benadering in de praktijk een bruikbare schatting voor het construeren van betrouwbaarheidsintervallen en voor hypothese-tests tussen verschillende modellen. In Tabel 7 rapporteren we daarom naast de standaardfout ook de MSE_0 van het referentiemodel – het model dat altijd het globaal gemiddelde voorspelt.

Tot slot dient benadrukt te worden dat de generalisatiefout $\text{Err}[n]$ in Vergelijking 7 gedefinieerd is voor een model dat wordt getraind op n observaties. Met andere woorden, de verwachte generalisatiefout is afhankelijk van de grootte van de trainingsdata. Bij het vergelijken van modellen gebaseerd op datasets van verschillende grootte (bijvoorbeeld LUCAS versus Cmon) moet daarom met deze nuance altijd rekening worden gehouden.

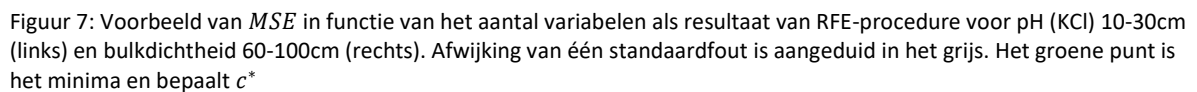
2.4.2 Covariatenselectie

De primaire en meest invloedrijke component van een model zijn de covariaten λ_l in het DSM-model (Vergelijking 1). Voor selectie van beste c^* -aantal covariaten wordt *Recursive Feature Elimination* (RFE) gebruikt. Deze procedure brengt voor klassieke Random Forest modellen enkele praktische uitdagingen met zich mee. Ten eerste zijn Random Forest importance-maten gevoelig voor correlatie tussen covariaten. In aanwezigheid van sterk gecorreleerde variabelen wordt de voorspellende kracht verdeeld over meerdere covariaten, waardoor het belang van de belangrijkste covariaten afneemt naar gelang het aantal covariaten in het model waarmee ze gecorreleerd zijn. Ten tweede wordt bij het construeren van bomen slechts een willekeurige subset van covariaten beschouwd bij elke splitsing, wat bijkomende variabiliteit in de rangschikking introduceert.

Om deze instabiliteit te beperken, wordt de initiële rangschikking van covariaten bepaald met behulp van XGBoost (Chen & Guestrin, 2016). Dit algoritme gebruikt eveneens CART-bomen (Classification And Regression Trees), maar hanteert een boostingstrategie. De belangrijkheid van een covariaat wordt hierbij bepaald door de totale reductie in MSE die aan die covariaat kan worden toegeschreven over alle bomen in het ensemble. In aanwezigheid van sterk gecorreleerde covariaten heeft deze procedure de neiging één dominante te selecteren, terwijl redundante covariaten weinig bijkomende reductie van MSE opleveren en dus lager in de rangorde terechtkomen. Bij een vaste random seed resulteert dit in een stabiele en reproduceerbare rangschikking van de belangrijkste covariaten.

De modelselectie verloopt vervolgens stapsgewijs. Eerst wordt binnen elke R-CV run (10 folds $\times$ 10 herhalingen) een XGBoost-model gefit om een rangschikking van covariaten te bekomen. Vervolgens worden Random Forest modellen gefit met incrementele subsets van covariaten volgens deze

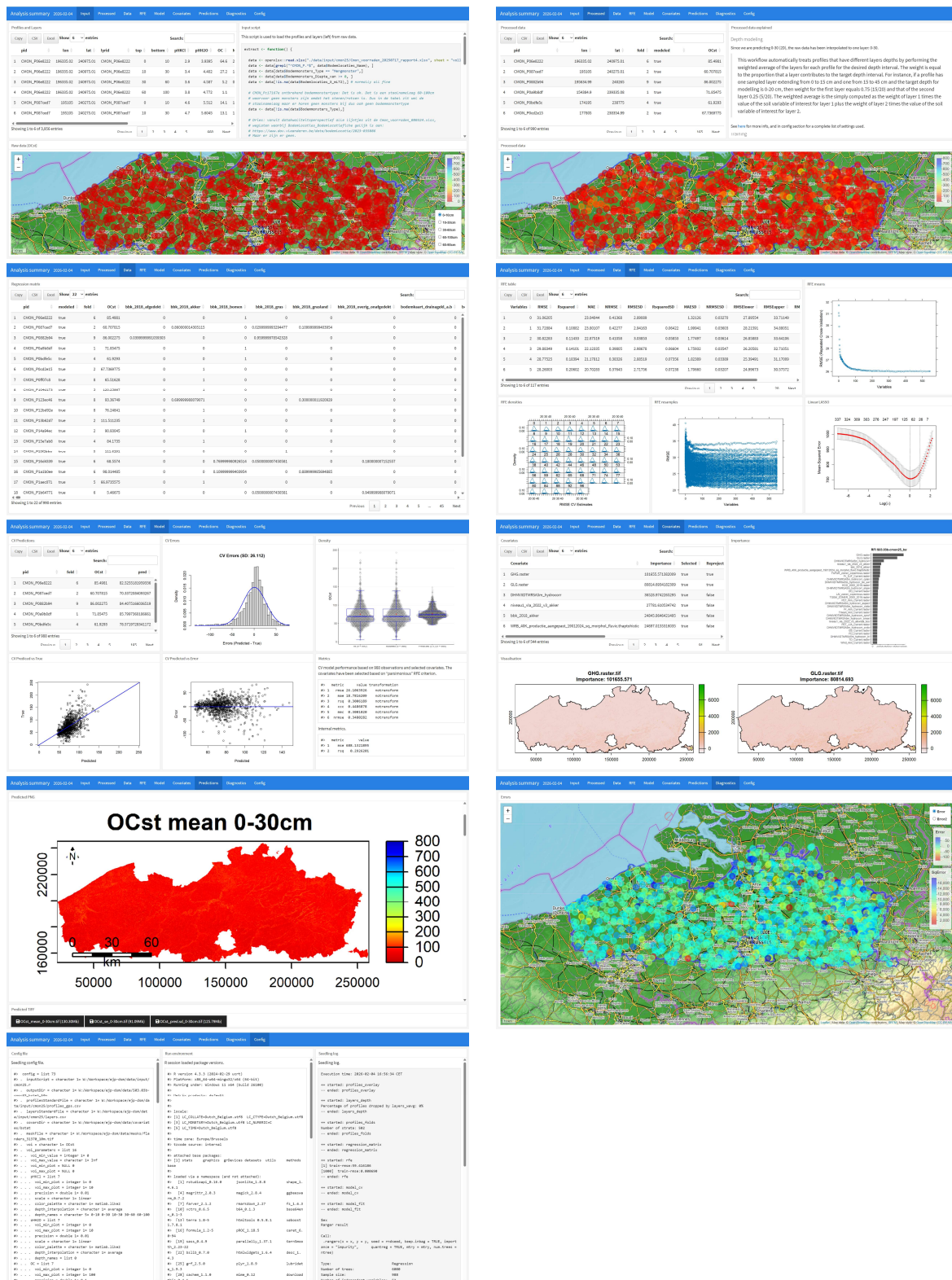
Meestal verloopt deze procedure feilloos zoals in Figuur 7 (links) afgebeeld waar een duidelijke c^* uit af te leiden valt. Echter, in sommige gevallen zoals afgebeeld in Figuur 7 (rechts), blijft de MSE -curve dalen naarmate er meer covariaten worden toegevoegd aan het model, met vaak een kronkelige staart te wijten aan sampling error en hoge variabiliteit in MSE . Dan is c^* bepalen minder evident. Omwille van dit effect, en omdat c^* een structurele overschatting is, bestaat in de literatuur een vuistregel om c te nemen dat één MSE -standaardfout verwijderd is van c^* . (Cfr. Hastie et al., 2009). Dit geeft dan een beter generaliseerbaar en interpreteerbaar model, waarbij het neveneffect op voorspellingen beperkt blijft. In dit project, om discussies rond de juiste vuistregel te omzeilen en omdat de nadruk niet ligt op interpreteerbaarheid van het model maar op de geproduceerde kaarten, is er gekozen om het beste model te nemen. Daarnaast hebben GRF-modellen (cfr. sectie 5) waartoe een overschakeling in de toekomst gepland is, dit probleem niet.



Geen fine-tuning van parameters is uitgevoerd. Parameter $mtry$ is globaal aangepast van $\lfloor \sqrt{c} \rfloor$ naar $\lfloor \min\{\sqrt{c} + 10, c\} \rfloor$ om de generalisatiefout te stabiliseren voor kleine c (het aantal covariaten).

Reproduceerbaarheid en transparantie van de modellering van de finale kaarten zijn van groot belang. Hiertoe is een uitbreiding op de Seedling-workflow ontwikkeld die voor elk scenario automatisch een uitgebreid HTML-rapport genereert. Zo'n rapport fungeert als een gemakkelijk leesbare, zelf-gedocumenteerde samenvatting van de volledige workflow: van initiële configuratie, datavoorbewerking en covariaten-selectie tot modeldiagnostiek en uiteindelijke kaartproductie. De

finale kaarten worden daarbij rechtstreeks in het rapport weergegeven, met de mogelijkheid om de bijhorende GeoTIFF-bestanden te exporteren. Figuur 8 toont een voorbeeld van zo'n HTML-rapport. Zo kan de gebruiker eenvoudig alle relevante informatie terugvinden die aan de basis ligt van de productie van de desbetreffende kaarten.



Figuur 8: Voorbeeld van een HTML-samenvatting van de kartering van koolstofvoorraad op 0-30 cm diepte. De verschillende tabs tonen de belangrijkste stappen in de Seedling workflow (te lezen van links naar rechts en van boven naar onder). *Input*: oorspronkelijke data en de pre-processing script. *Processed*: diepte-modellering, in dit geval van 0-

2.6 AANPASSINGEN

Tabel 5: Overzicht van belangrijkste verschillen tussen de huidige methode, en de methodes voorafgaand aan dit project

Component	Wittemans et al. (2023)	Seedling (v3.13)	Dit project
Model (package)	RF (<i>RandomForest</i>)	RF (<i>ranger</i>)	RF (<i>ranger</i>)
Standaardfout	Standaardafwijking op 2 quasionafhankelijke voorspellingen	-	Jackknife (2.1.2) en delta-methode (2.2.2)
Distributie	-	Kwantielen	Kwantielen
Aleatorische variabiliteit	-	-	MSPE (sectie 2.1.4)
Compositionele data	-	ALR	ALR (sectie 2.2)
Modelselectie	Alle covariaten	RFE (caret)	Gemodificeerde RFE (sectie 2.4)
Samenvattende analyse	-	-	HTML-rapport (sectie 2.5)

De DSM-workflow gebruikt in dit project is grotendeels gebaseerd op de bestaande Seedling-
implementatie. Deze workflow bestaat uit een aantal opeenvolgende stappen, waarbij de eerste en
laatste stap uitbreidingen zijn die specifiek in dit project werden ontwikkeld:

- Deze workflow wordt aangestuurd via configuratiebestanden waarin verschillende scenario's worden gedefinieerd. Naast de methodologische uitbreidingen beschreven in sectie 2 werden ook een aantal praktische aanpassingen doorgevoerd om beter aan te sluiten bij de noden van dit project. De details van alle wijzigingen ten opzichte van Seedling (v3.13) zijn [hier](#) raadpleegbaar. Hieronder worden alleen de belangrijkste wijzigingen samengevat.

- Dieptemodellering is uitgebreid. Naast gewogen gemiddelden (voor concentraties) is ook gewogen sommatie geïmplementeerd (voor contentvariabelen). Deze methode is

- Toch blijven er, ondanks deze aanpassingen, enkele belangrijke beperkingen bestaan:

- Daarnaast lijkt de oorspronkelijke opzet van Seedling gericht op het vergelijken van verschillende modellen. Binnen de context van dit project ligt de nadruk echter evengoed op het zo nauwkeurig mogelijk schatten van verwachtingswaarden, standaardfouten, predictieve distributies en aleatorische variabiliteit voor zowel enkelvoudige als samengestelde modellen. Veel van de aangebrachte aanpassingen kunnen daarom ook begrepen worden als een transformatie van Seedling van een generieke Machine Learning DSM-workflow naar een meer statistisch gefundeerde DSM-workflow.

Het initiële script van Wittemans et al. (2023) volgt in grote lijnen een workflow die vergelijkbaar is met die van Seedling (cfr. sectie 0). Echter, de implementatie van zo'n 20-tal lijnen effectieve code is aanzienlijk compacter en kan herleid worden tot de volgende stappen: preprocessing, modellering en predictie. In dat opzicht kan het script worden gezien als een eerste werkbare implementatie voor het genereren van DSM-kaarten, maar schiet te kort op verschillende vlakken:

- Samengevat kan het script van Wittemans et al. (2023) worden beschouwd als een eerste iteratie die de basis legt voor latere ontwikkelingen. Seedling, en daaropvolgend huidig project, vertegenwoordigt een veel verdere iteratie van deze ontwikkeling. Om deze redenen werd het script van Wittemans et al. (2023) in dit project enkel geanalyseerd ter vergelijking en inspiratie, maar niet gebruikt voor de productie van de finale kaarten.

De geproduceerde kaartlagen zijn in [GeoTIFF-formaat](#) met 4-byte floating-point precisie (FLT4S) en gecomprimeerd met [ZSTD](#) (level 22). De werkelijke precisie is echter manueel beperkt voor elke bodemdescriptor om de kaartgrootte beperkt te houden zonder praktisch bruikbaar verlies van informatie. Deze precisie werd zodanig gekozen dat ze minstens een factor twee kleiner is dan de kleinste standaardfout. Een overzicht van de bodemdescriptoren, de dieptelagen en de gebruikte precisie is gegeven in Tabel 6. Alle kaartlagen zijn gebaseerd op de Cmon 2025-brondata (Oorts et al., 2025a).

Bodemdescriptor	Diepte	Model	Precisie
pH (KCl)	0-10cm, 10-30cm, 30-60cm, 60-100cm	Algemeen model	0.01
Organische C (%)	0-10cm, 10-30cm, 30-60cm, 60-100cm	Algemeen model	0.01
Organische C voorraad (t/ha)	0-30cm, 0-100cm	Algemeen model	0.5
Totale bulkdichtheid (g/cm³)	0-10cm, 10-30cm, 30-60cm, 60-100cm	Algemeen model	0.001
Zand, Leem, Klei (%)	0-10cm, 10-30cm, 30-60cm, 60-100cm	Samengesteld model	0.01

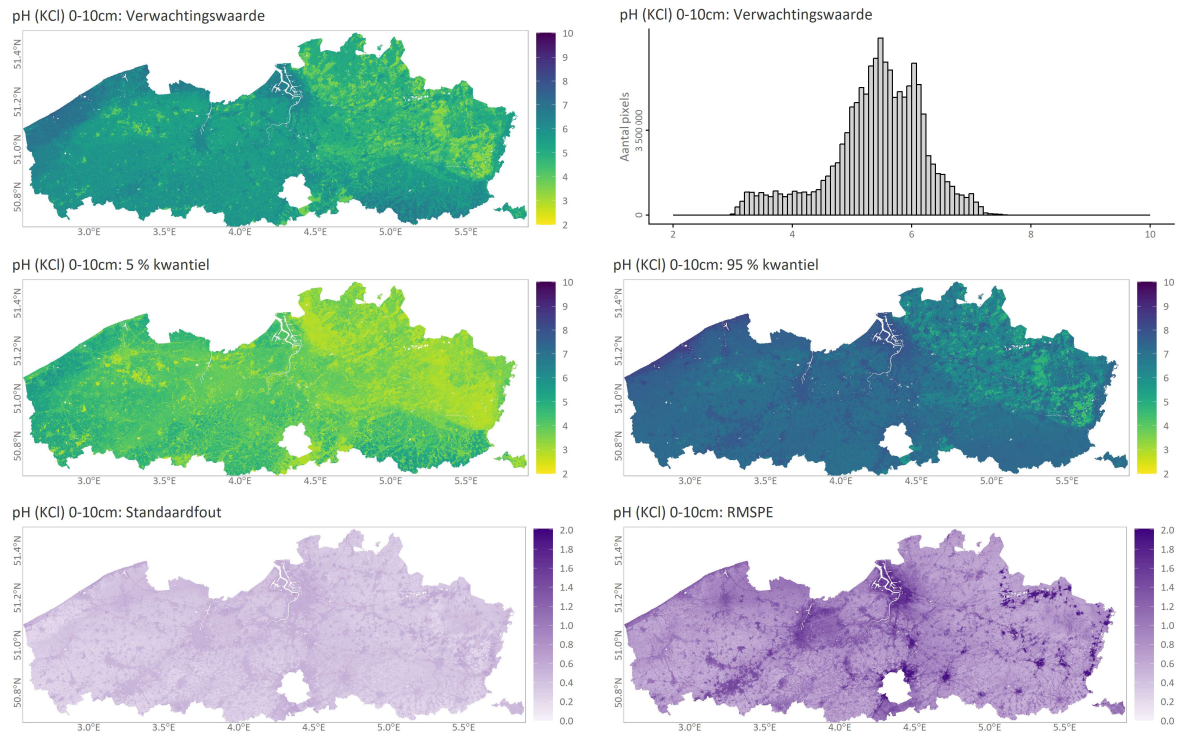
- Verwachtingskaart: schatter van de locatie-conditionele verwachtingswaarde μ_l die geïnterpreteerd kan worden als de geïdealiseerde ware waarde van de bodemdescriptor (cfr. sectie 2.1.1 en Figuur 2). Daarnaast wordt het histogram van de pixelwaarden weergegeven om de globale verdeling over Vlaanderen te visualiseren.
- De 5% en 95% kwantielkaarten: samen geven ze een bereik waarbinnen met 90 % zekerheid een toekomstige meting kan verwacht worden. (cfr. sectie 2.1.3 en Figuur 2).
- Standaardfout-kaart: schatting van de onzekerheid op $\hat{\mu}_l$ (cfr. sectie 2.1.2 en Figuur 2).
- RMSPE-kaart: vierkantswortel van de *mean squared prediction error* (RMSPE), die fungeert als een conservatieve schatter van de aleatorische variabiliteit σ_l (cfr. sectie 2.1.4 en Figuur 2). Deze waarde kan geïnterpreteerd worden als een locatie-specifieke standaarddeviatie van de predictieve distributie en geeft dus een indicatie van de deviatie tussen een toekomstige meting en de voorspelde verwachtingswaarde $\hat{\mu}_l$.

[illegible]

bulkdichtheid), zodat contrast behouden blijft in het grootste deel van het spectrum. Kaarten van de mediaan ($q_{0.5}$) zijn eveneens beschikbaar, maar hier niet weergegeven.

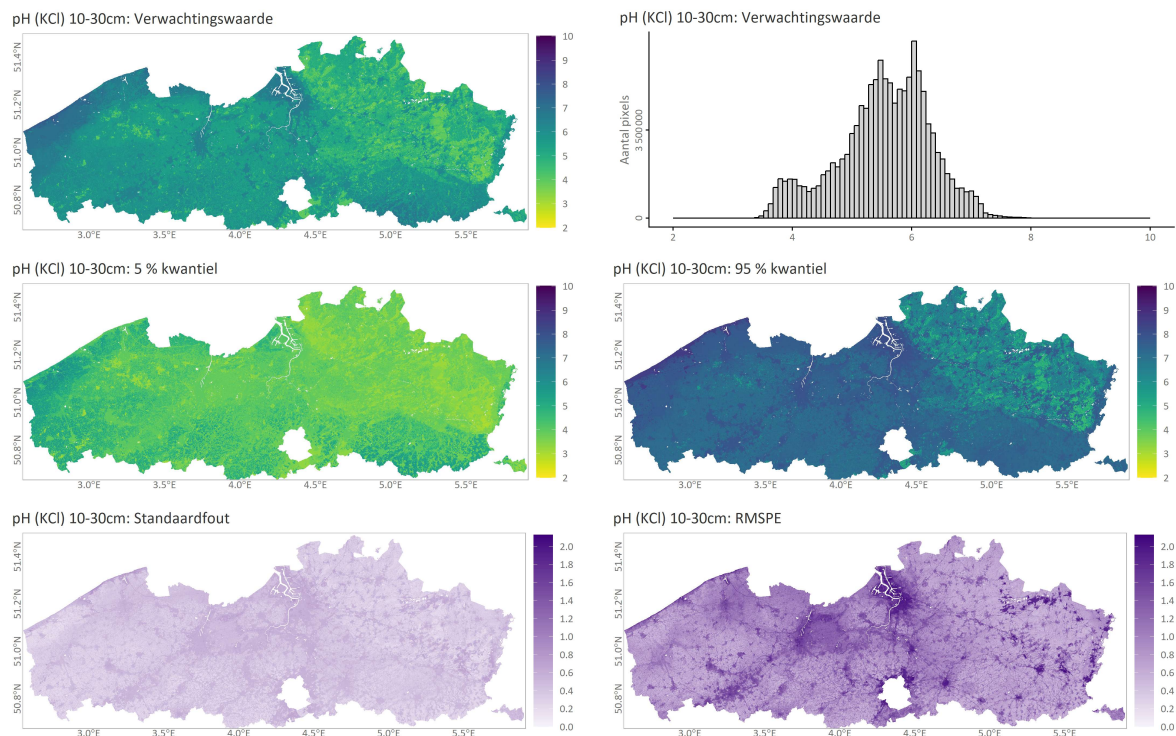
3.1 PH (KCL)

3.1.1 0-10cm



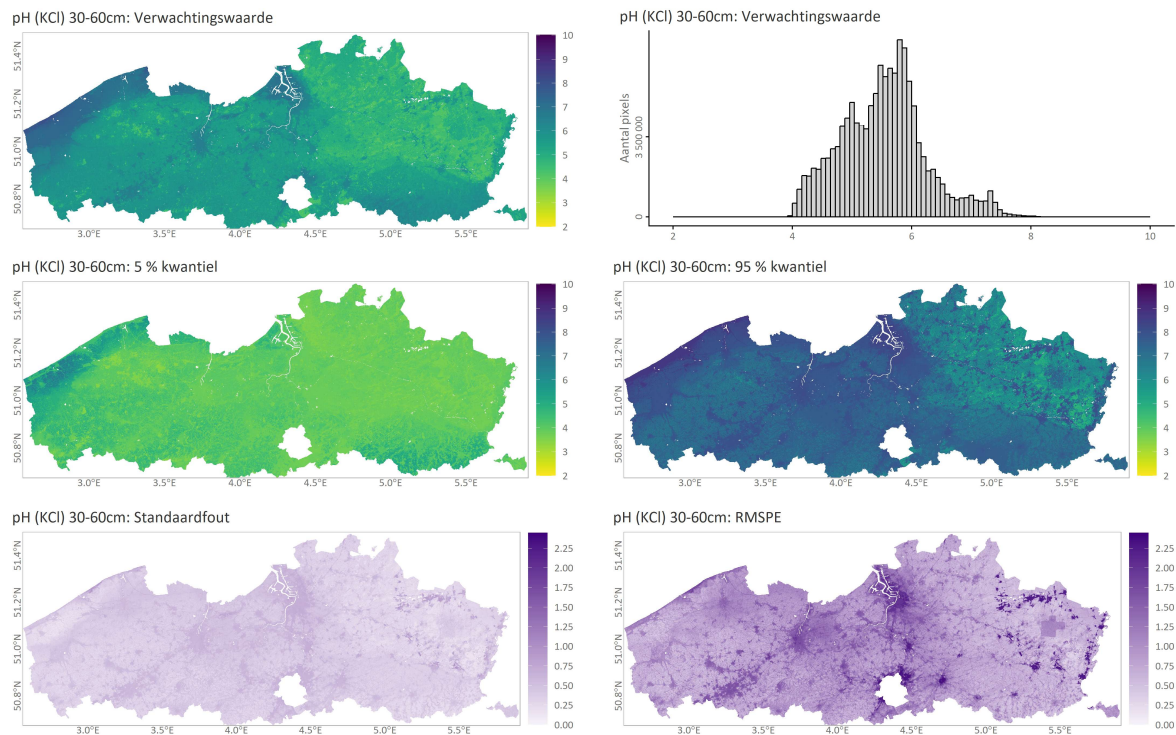
Figuur 9: pH (KCL) 0-10cm

3.1.2 10-30cm



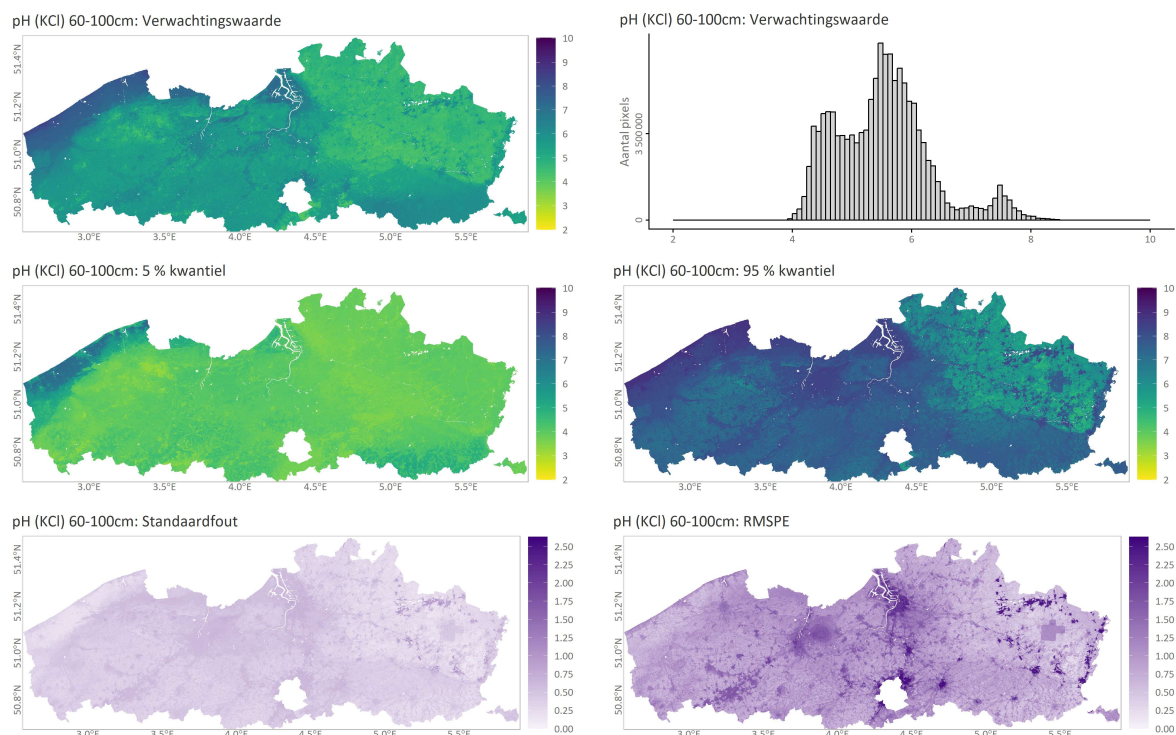
Figuur 10: pH (KCl) 10-30cm

3.1.3 30-60cm



Figuur 11: pH (KCl) 30-60cm

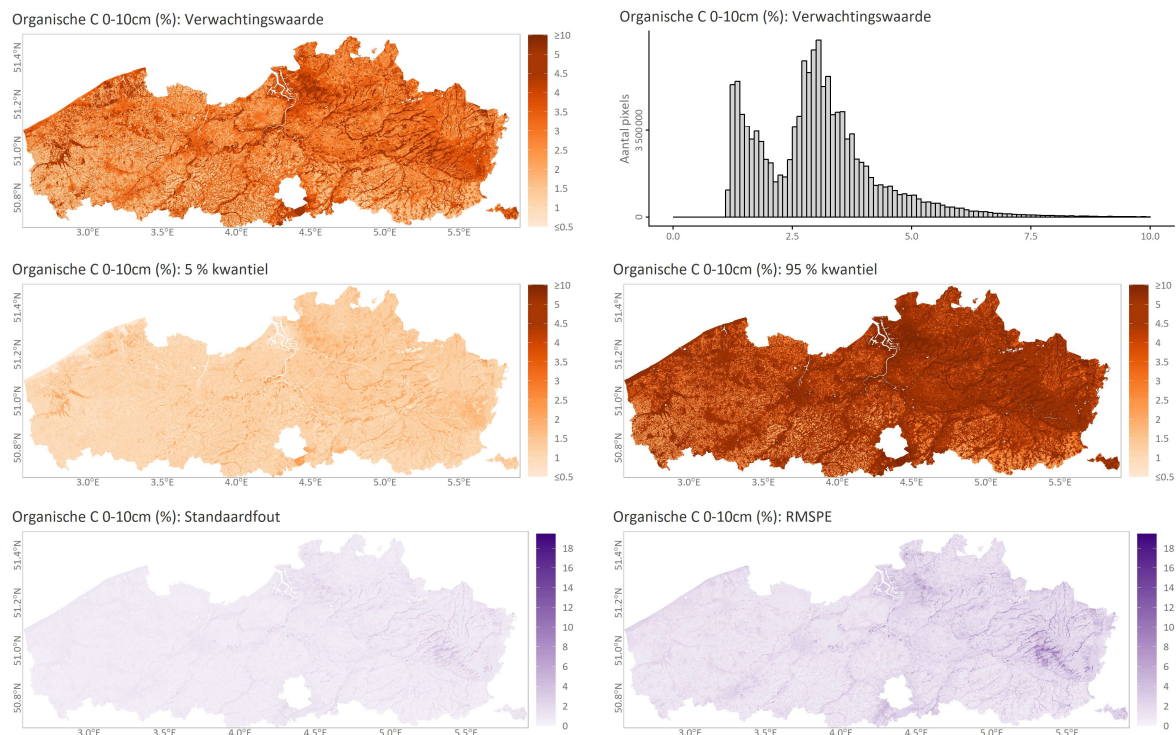
3.1.4 60-100cm



Figuur 12: pH (KCl) 60-100cm

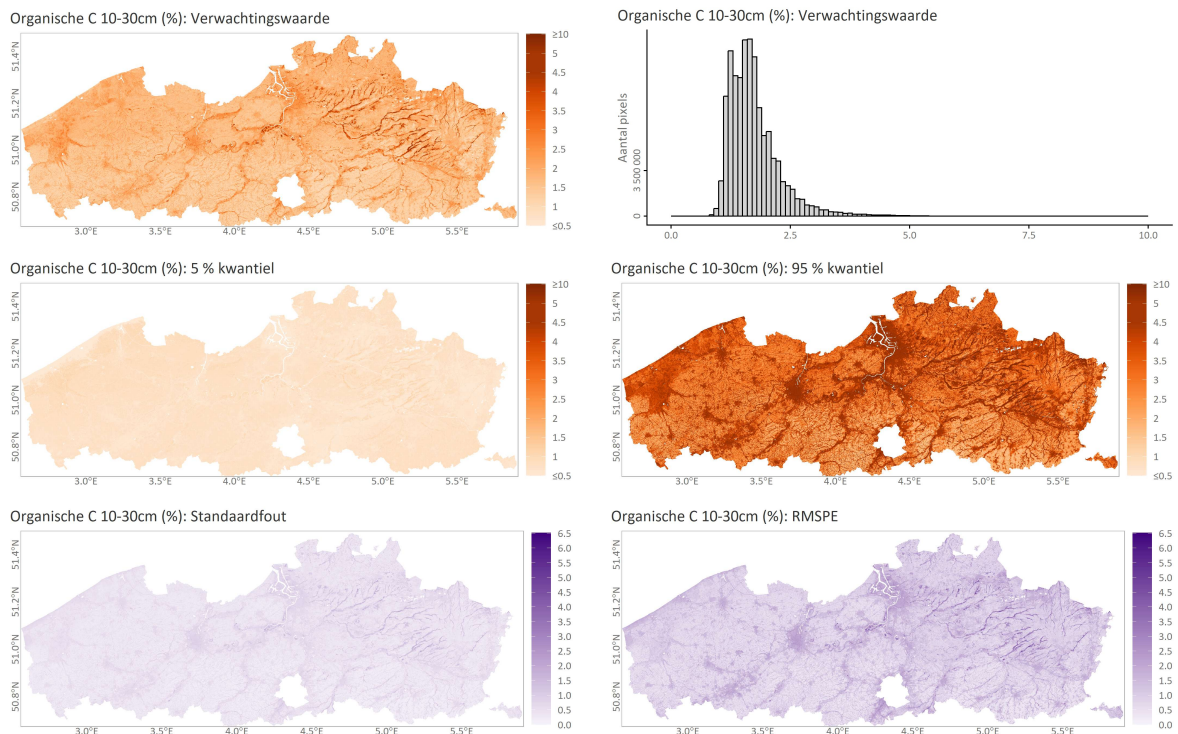
3.2 ORGANISCHE C

3.2.1 0-10cm



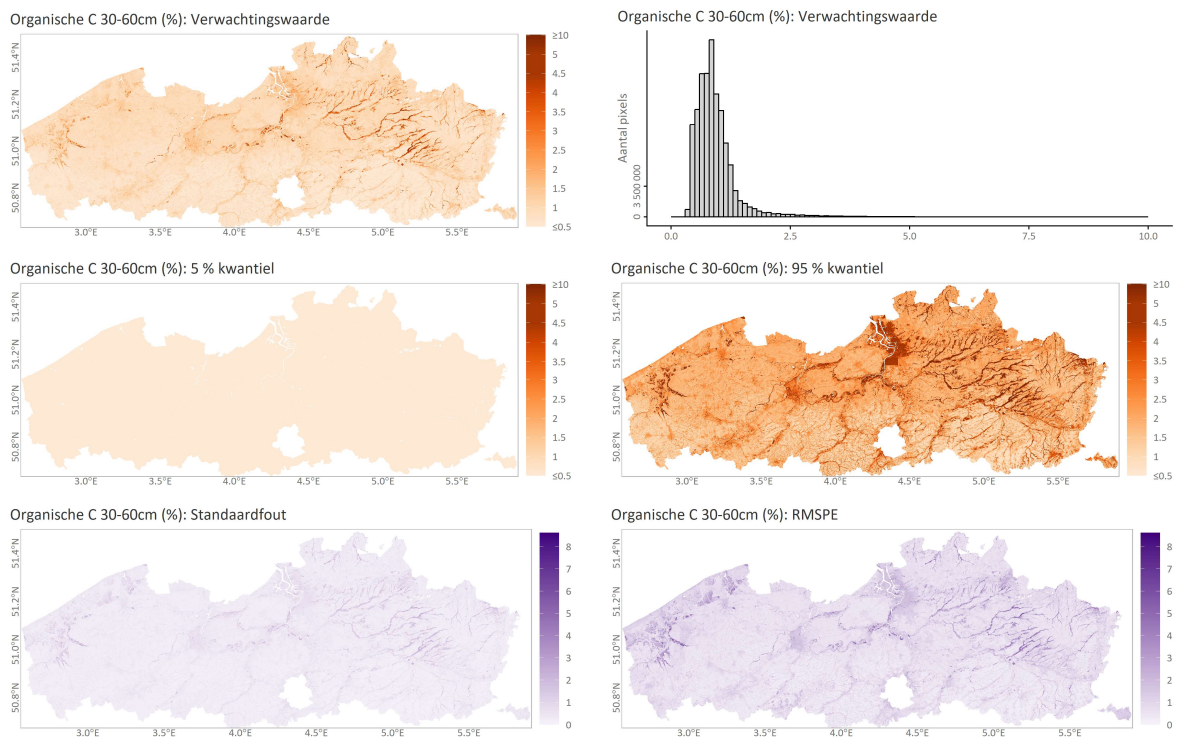
Figuur 13: Organische C 0-10cm (%)

3.2.2 10-30cm



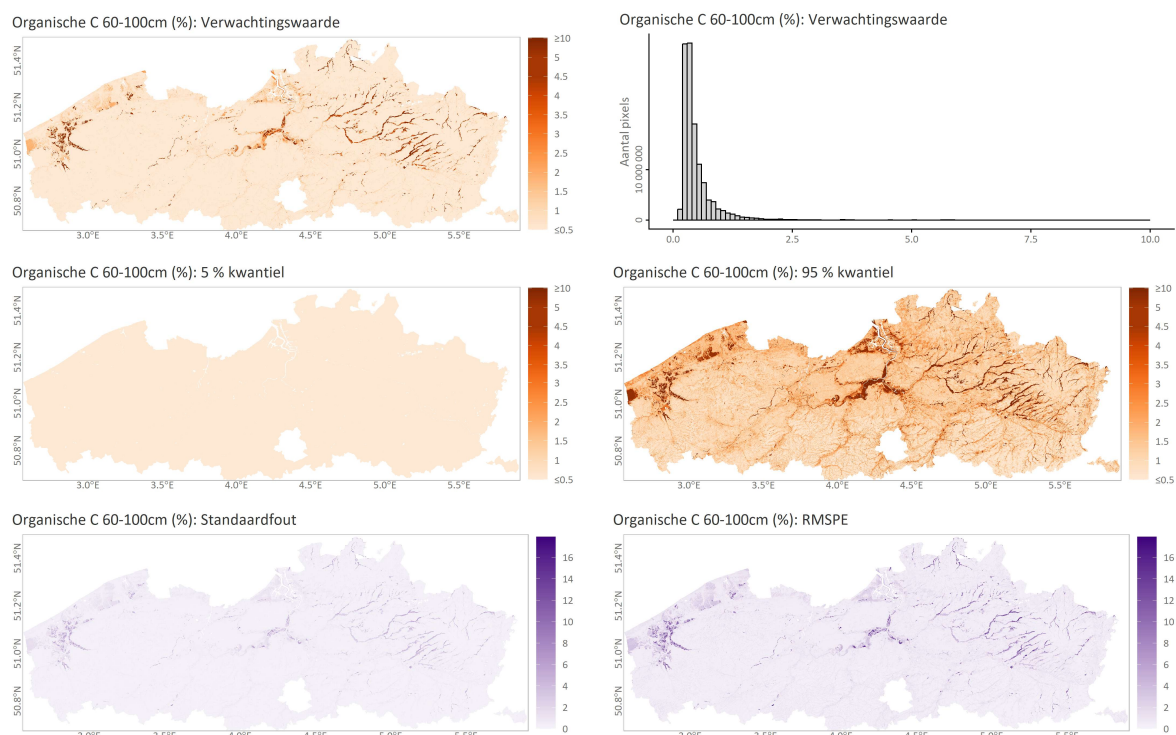
Figuur 14: Organische C 10-30cm (%)

3.2.3 30-60cm



Figuur 15: Organische C 30-60cm (%)

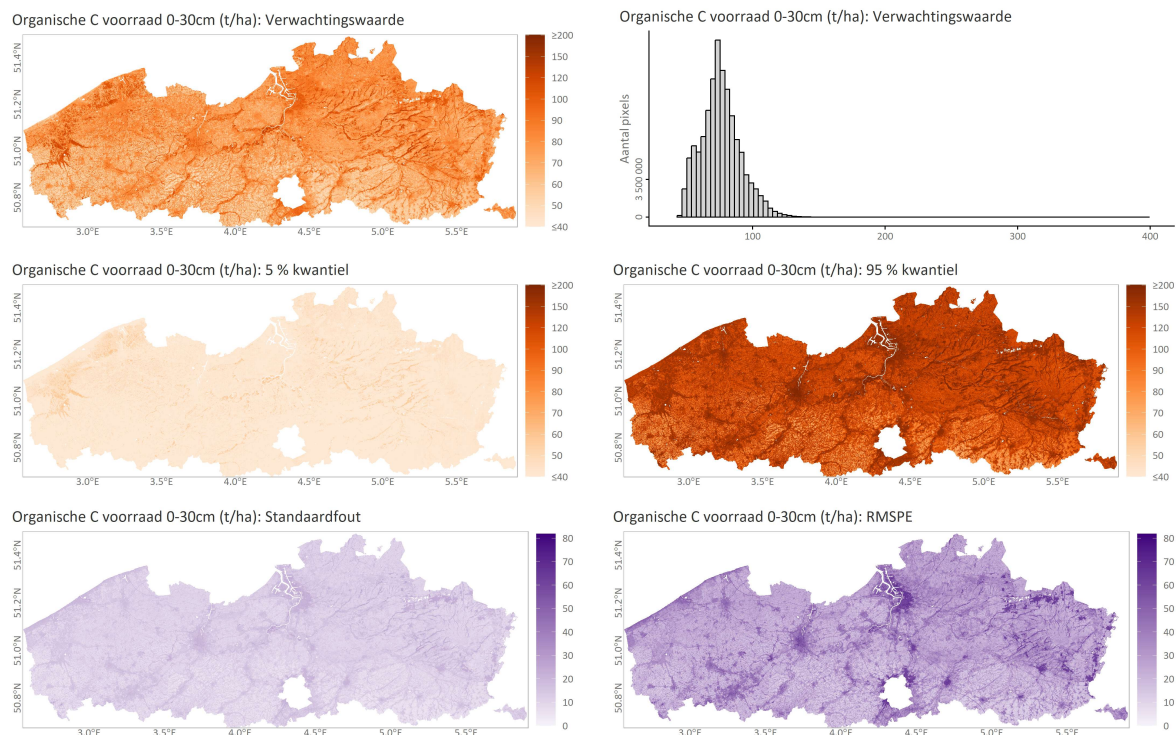
3.2.4 60-100cm



Figuur 16: Organische C 60-100cm (%)

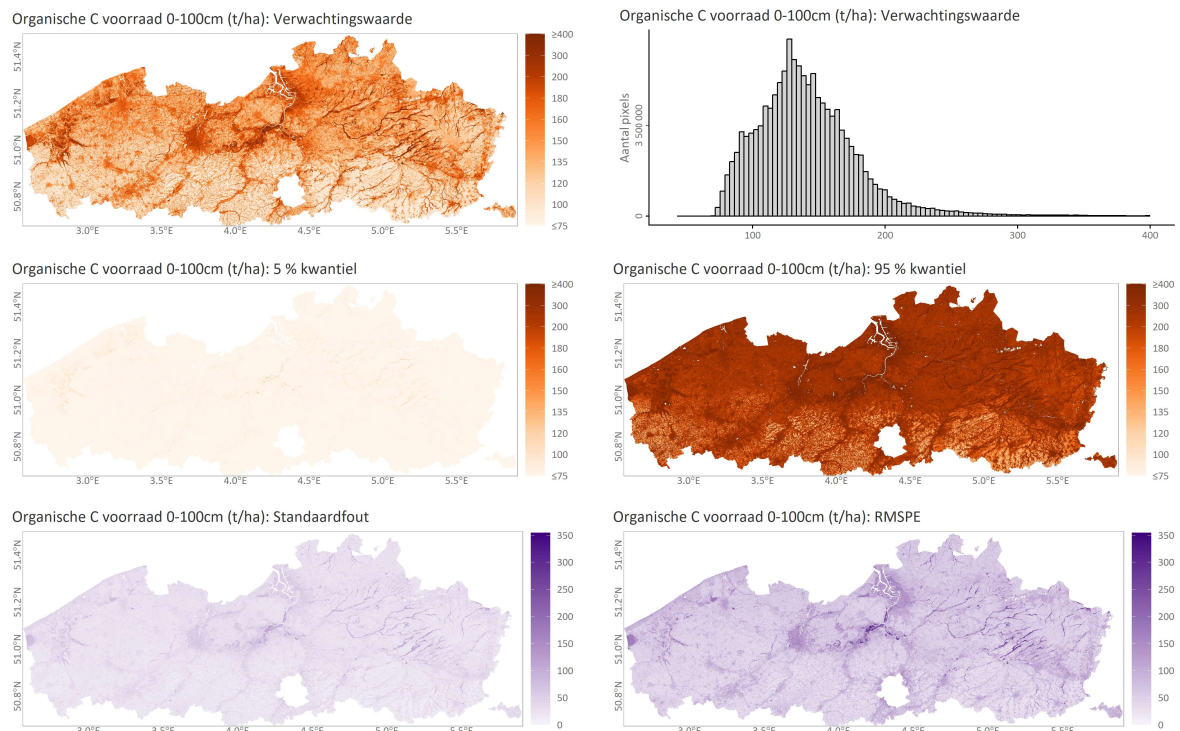
3.3 ORGANISCHE C VOORRADEN

3.3.1 0-30cm



Figuur 17: Organische C voorraden 0-30cm (t/ha)

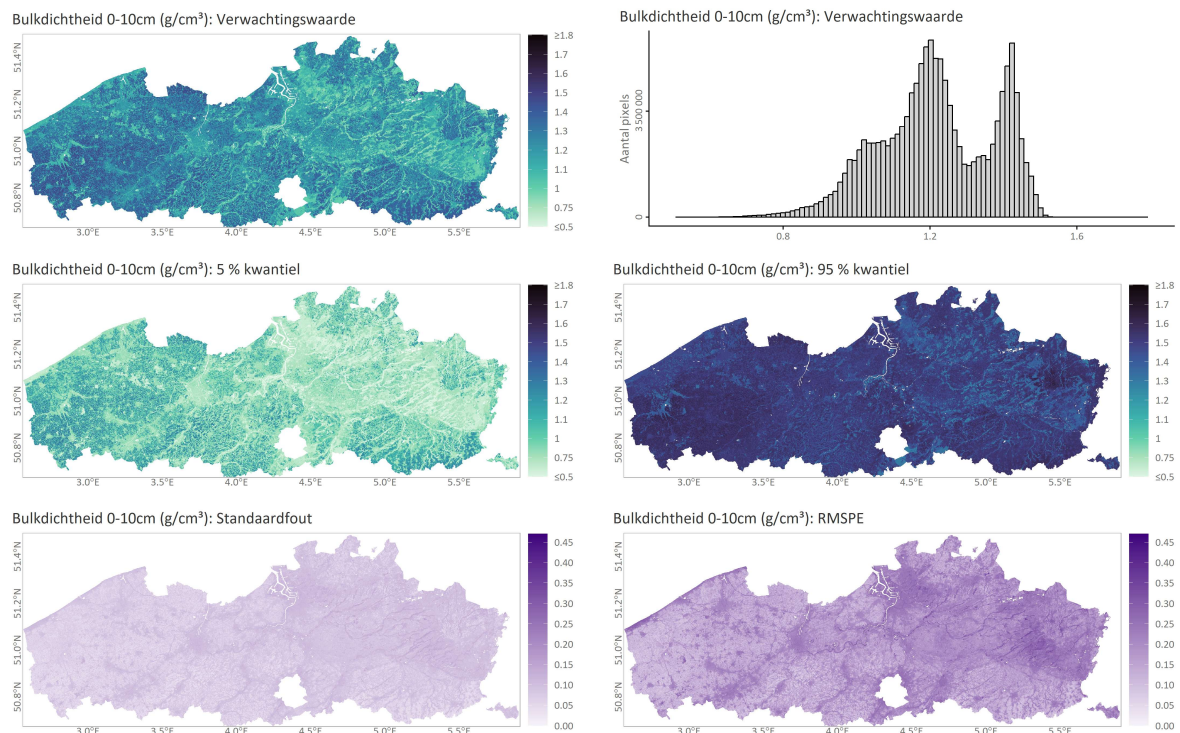
3.3.2 0-100cm



Figuur 18: Organische C voorraden 0-100cm (t/ha)

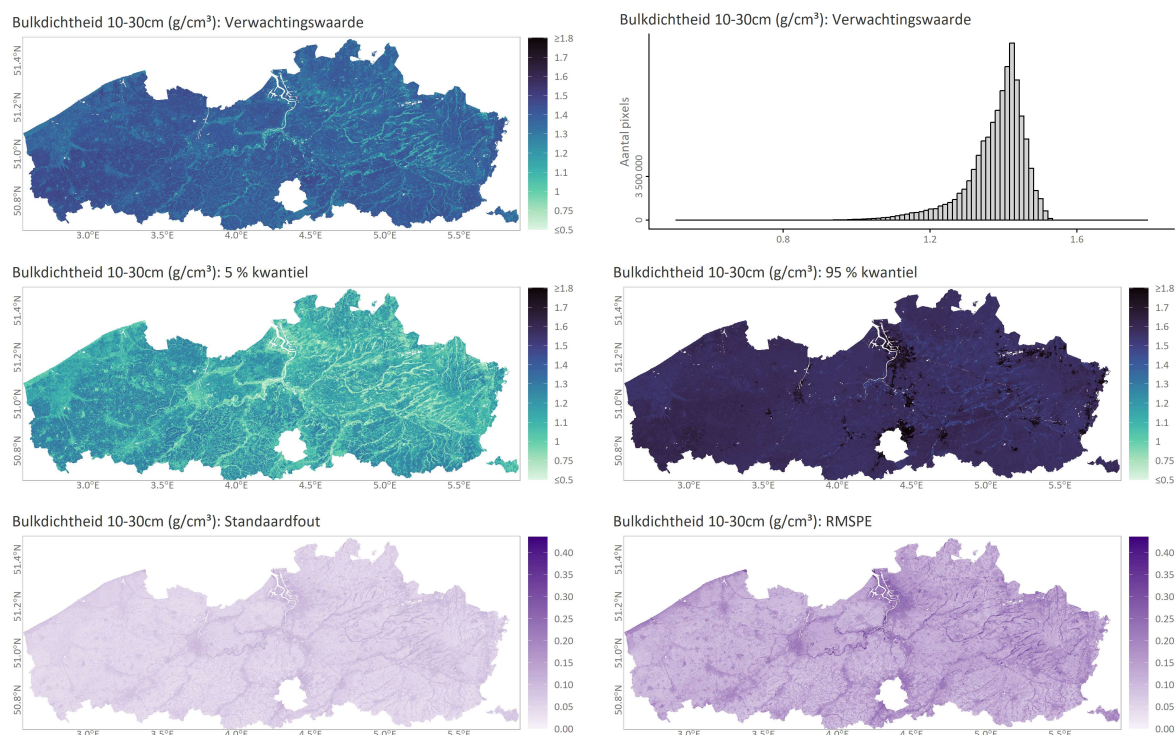
3.4 BULKDICHTHEID

3.4.1 0-10cm



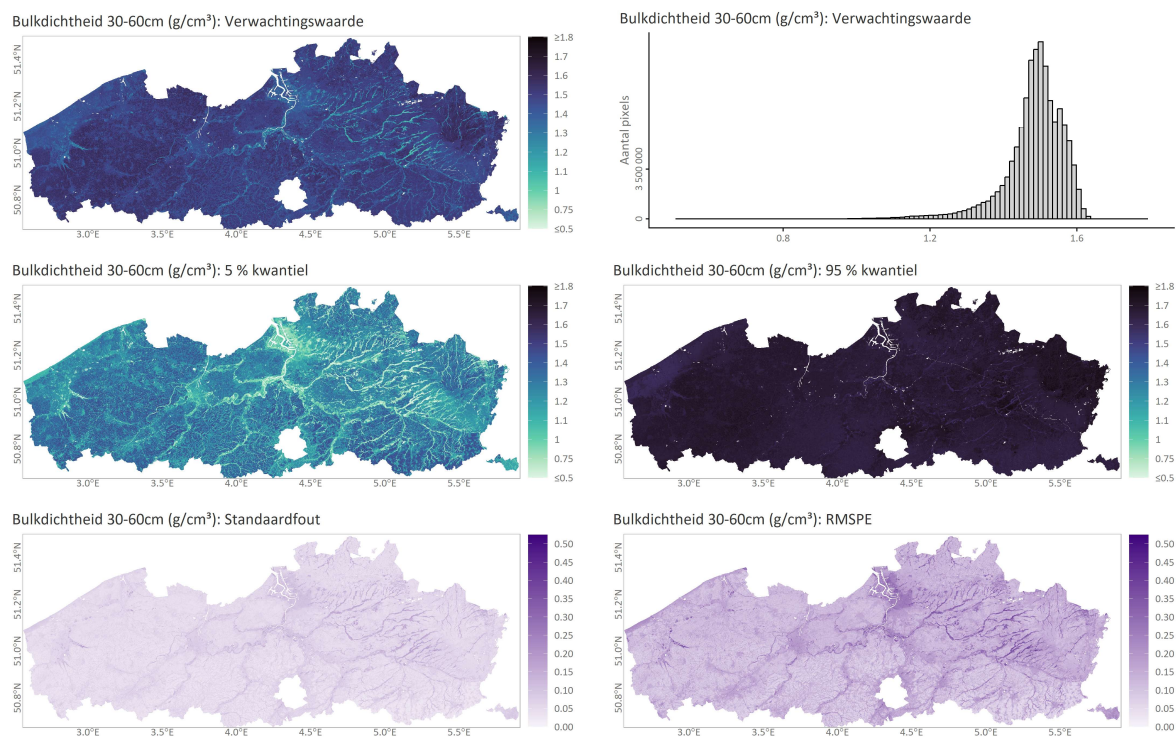
Figuur 19: Bulkdichtheid 0-10cm (g/cm³)

3.4.2 10-30cm



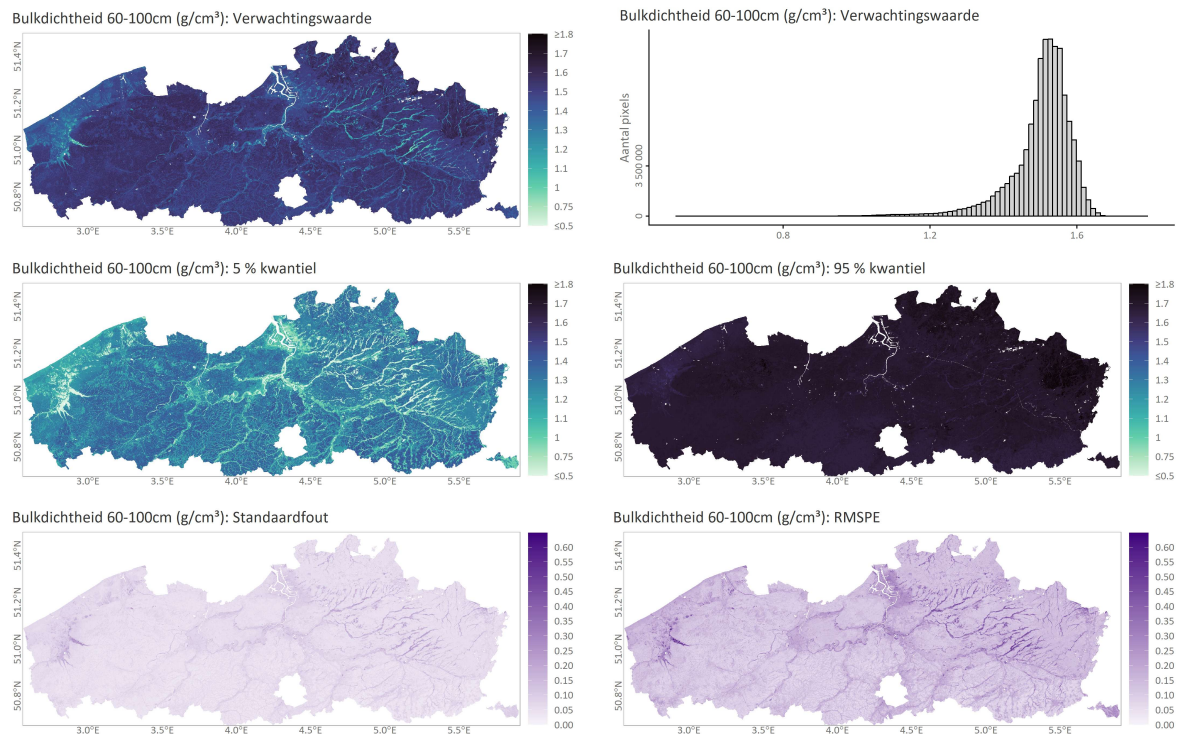
Figuur 20: Bulkdichtheid 10-30cm (g/cm^3)

3.4.3 30-60cm



Figuur 21: Bulkdichtheid 30-60cm (g/cm^3)

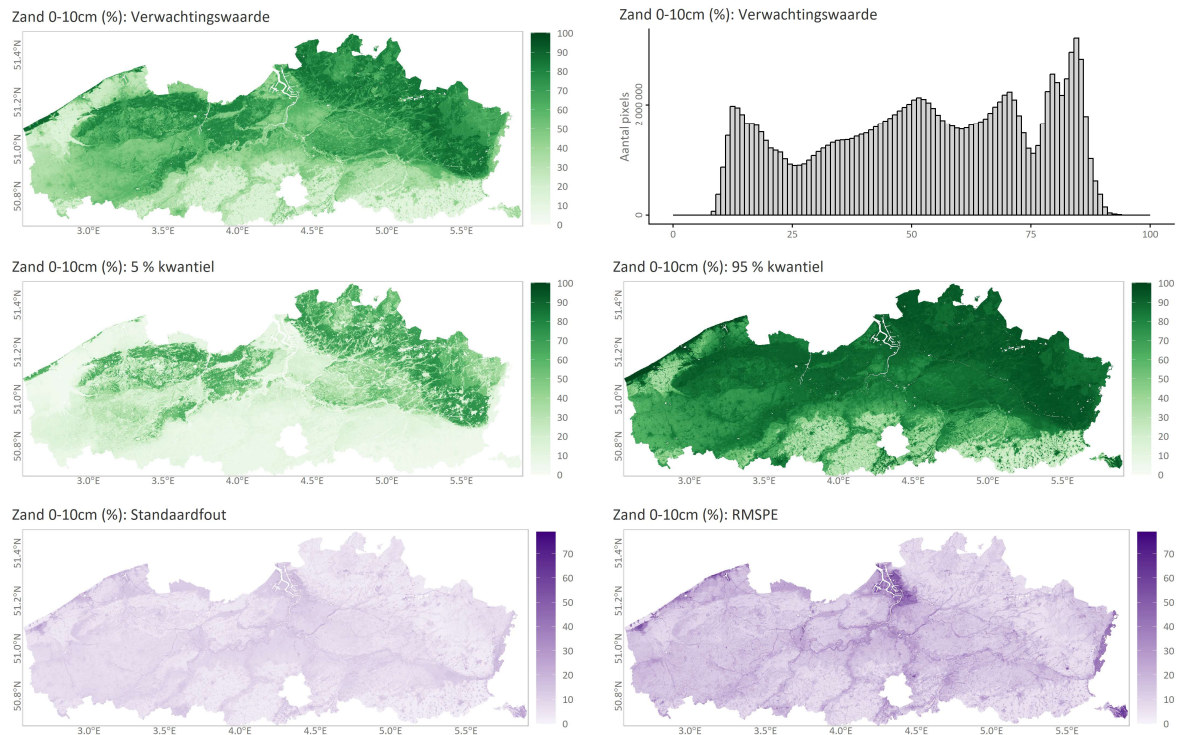
3.4.4 60-100cm



Figuur 22: Bulkdichtheid 60-100cm (g/cm³)

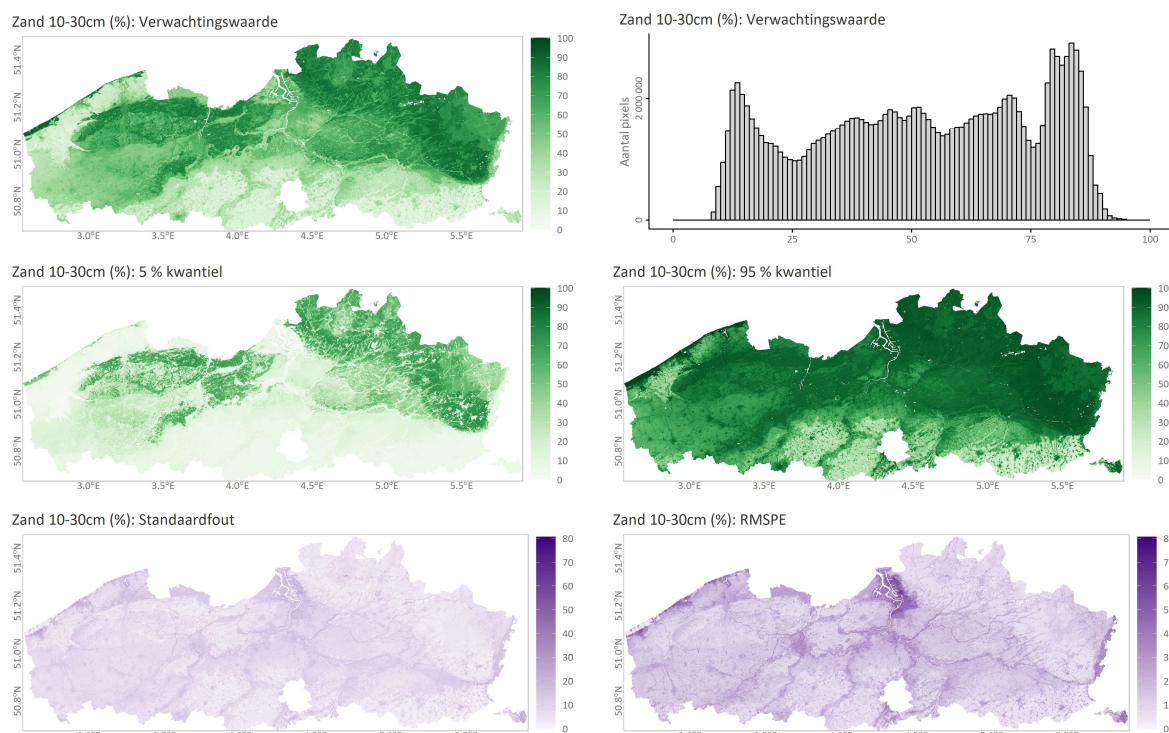
3.5 ZAND

3.5.1 0-10cm



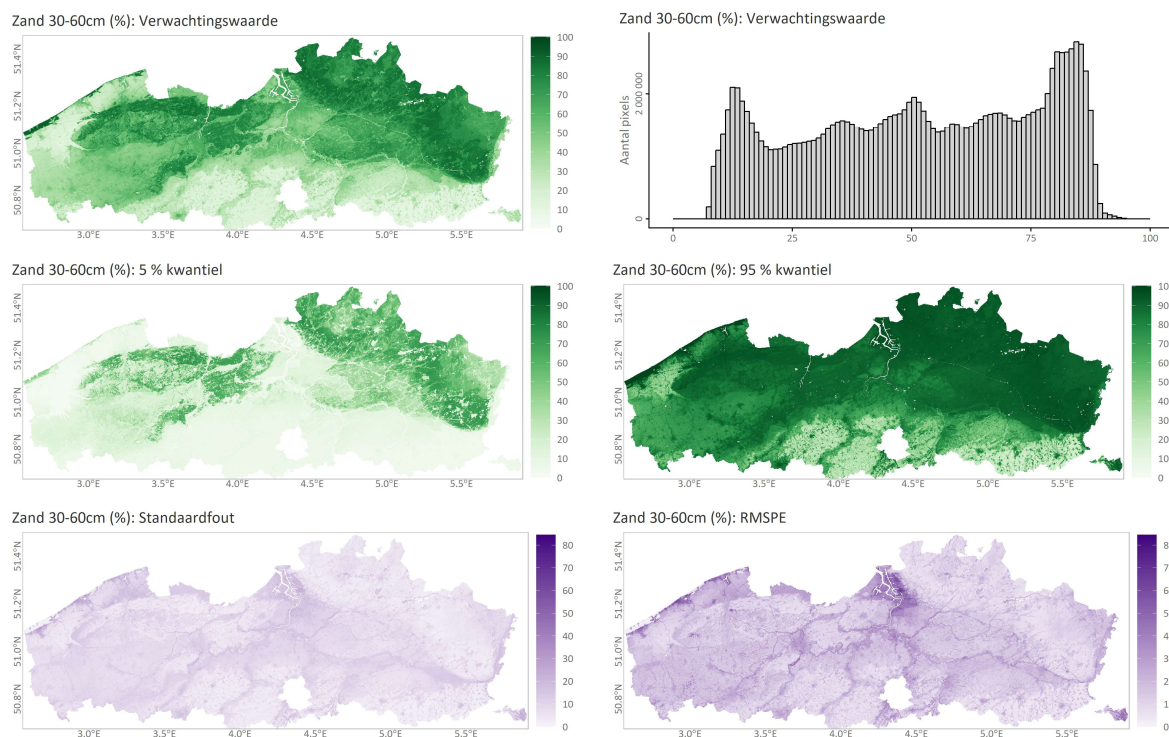
Figuur 23: Zand 0-10cm (%)

3.5.2 10-30cm



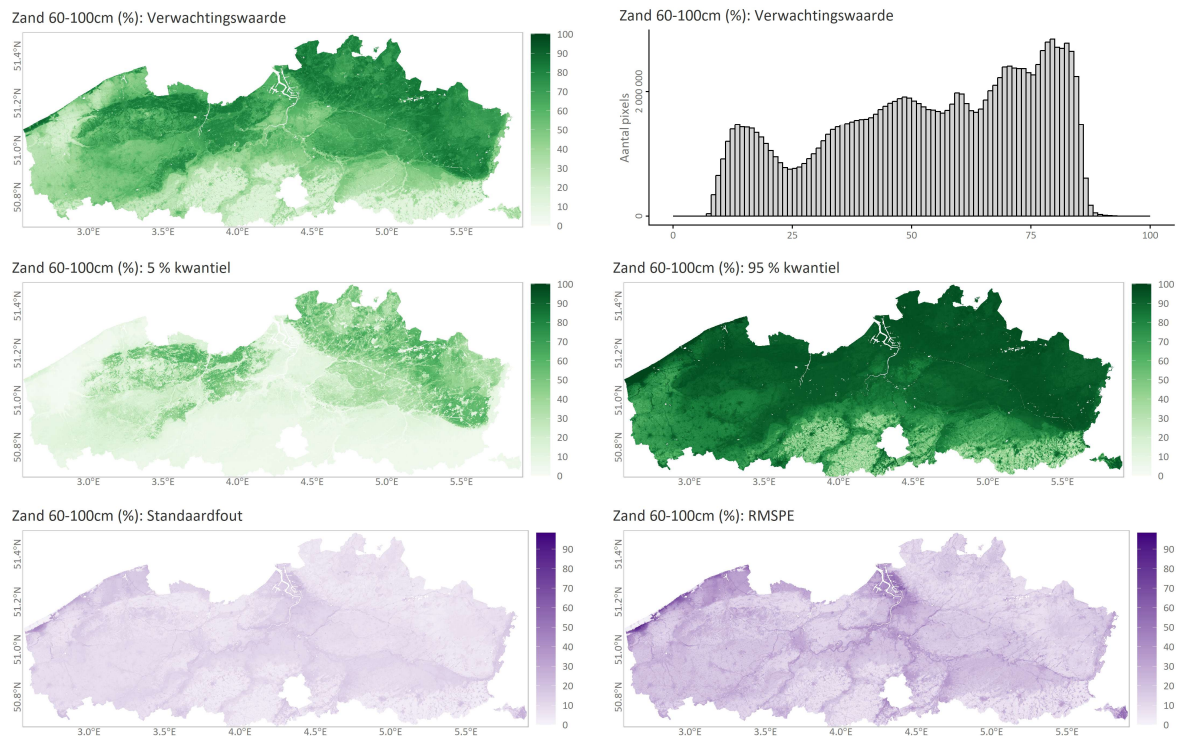
Figuur 24: Zand 10-30cm (g/cm³)

3.5.3 30-60cm



Figuur 25: Zand 30-60cm (g/cm³)

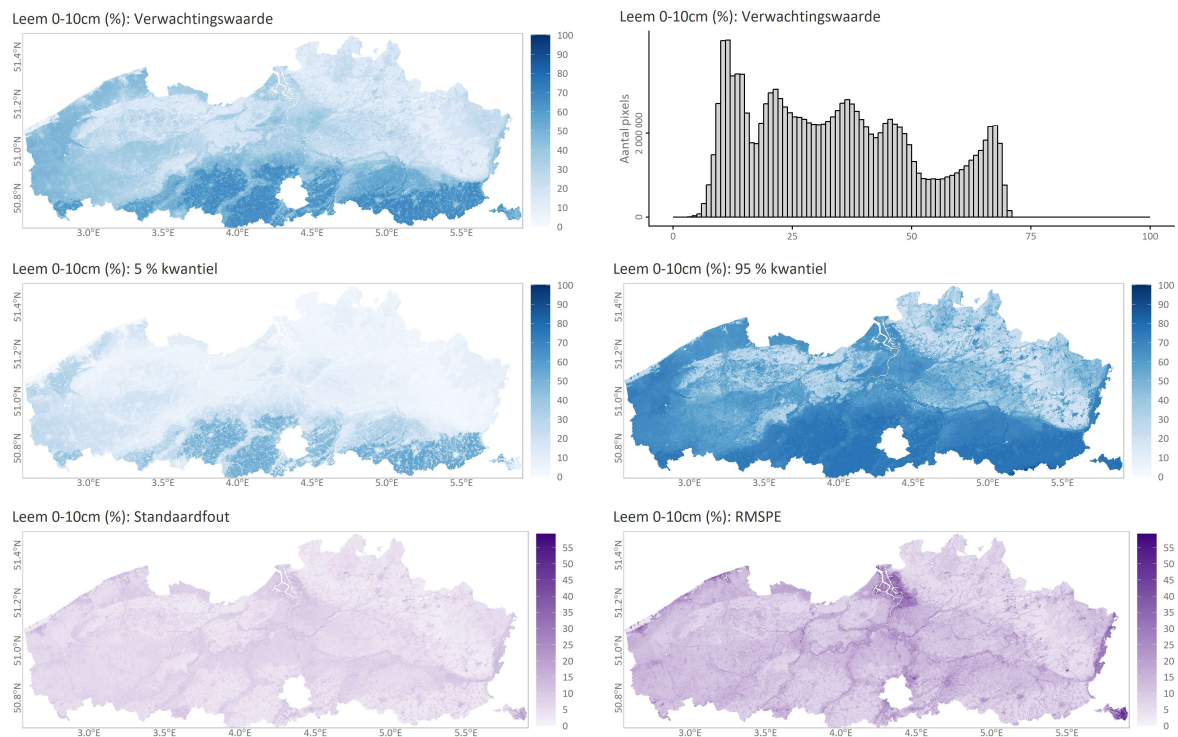
3.5.4 60-100cm



Figuur 26: Zand 60-100cm (g/cm³)

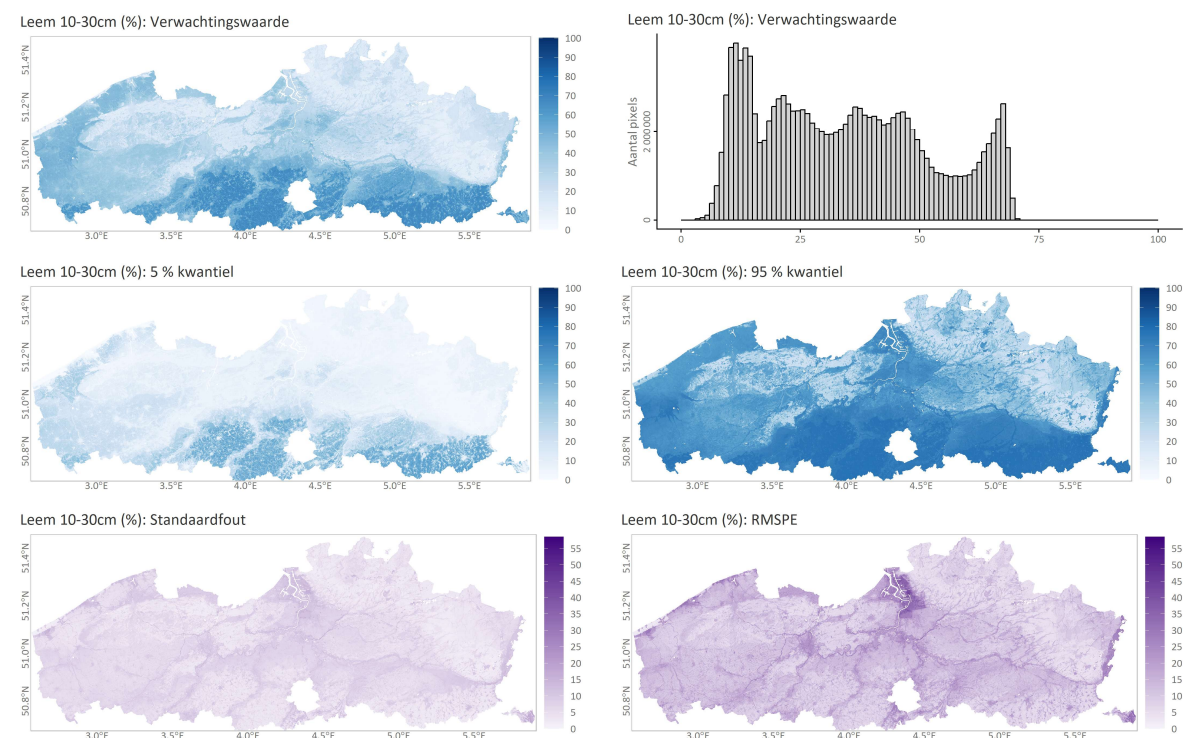
3.6 LEEM

3.6.1 0-10cm



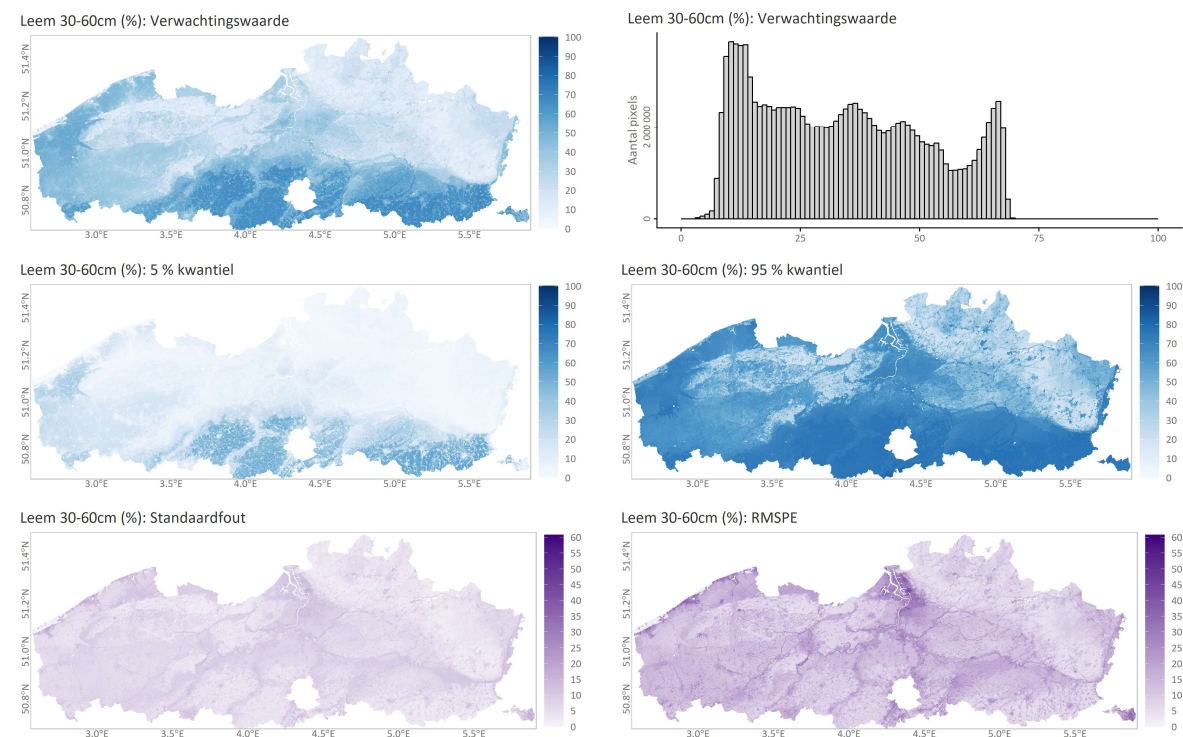
Figuur 27: Leem 0-10cm (%)

3.6.2 10-30cm



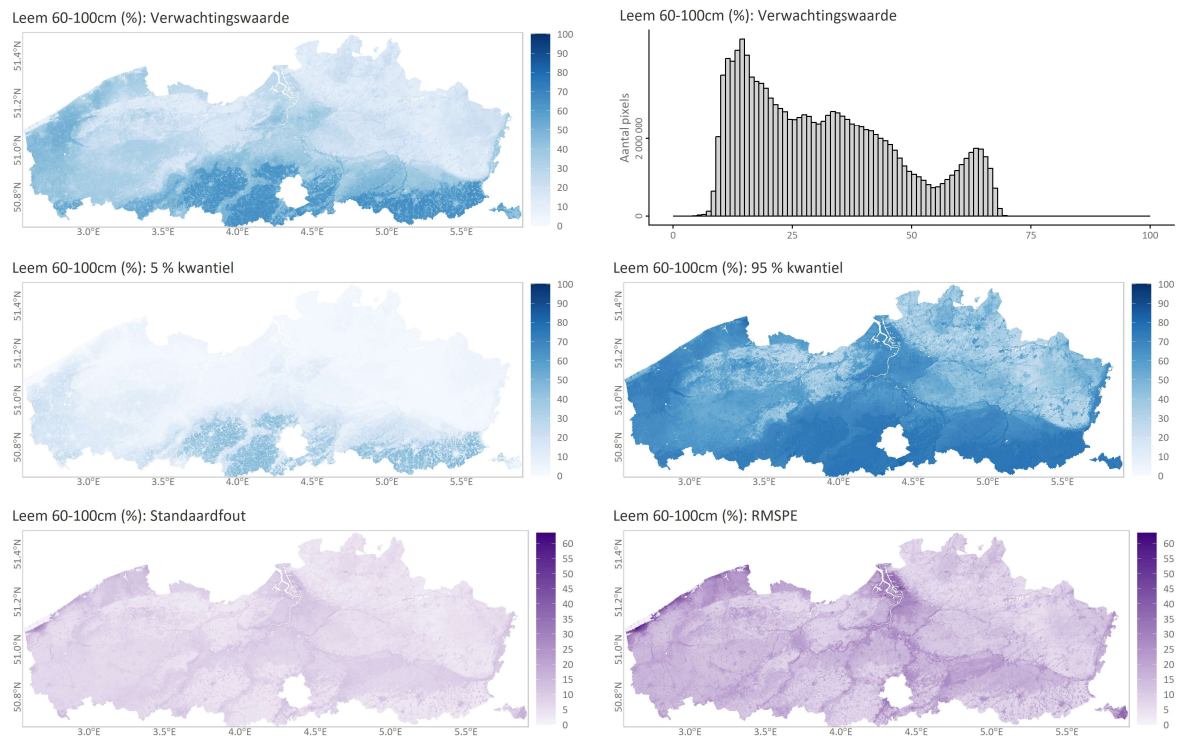
Figuur 28: Leem 10-30cm (g/cm³)

3.6.3 30-60cm



Figuur 29: Leem 30-60cm (g/cm³)

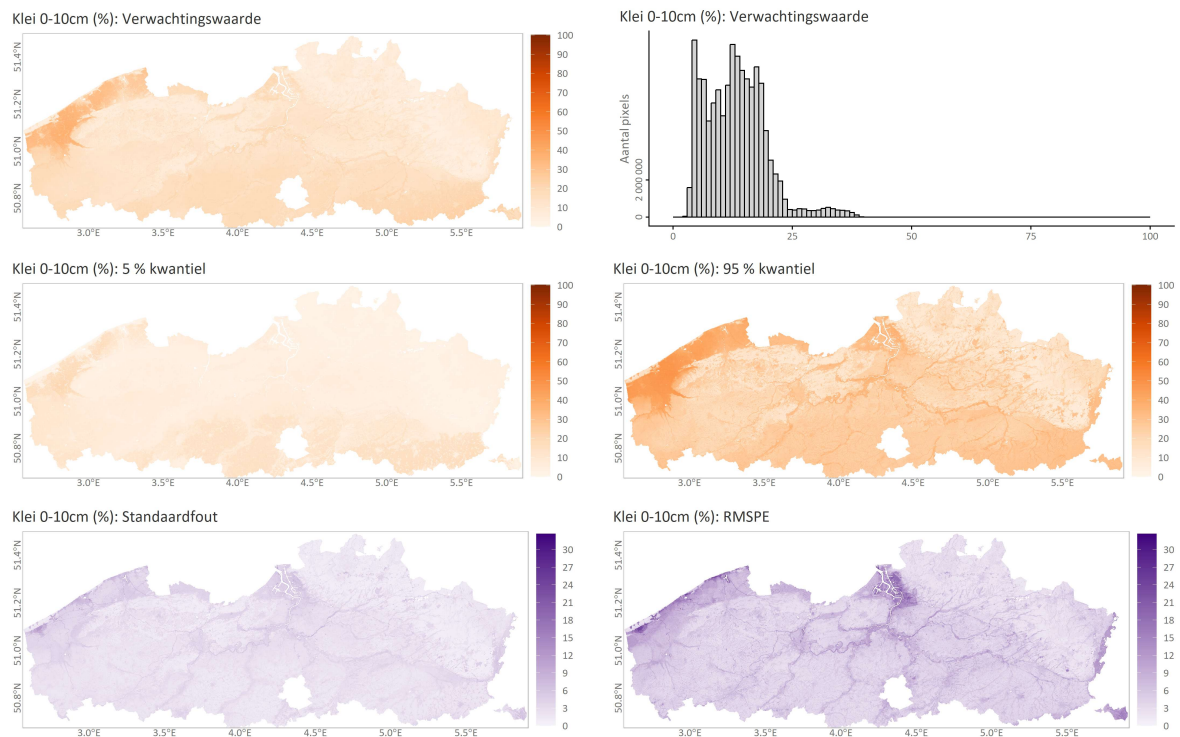
3.6.4 60-100cm



Figuur 30: Leem 60-100cm (g/cm^3)

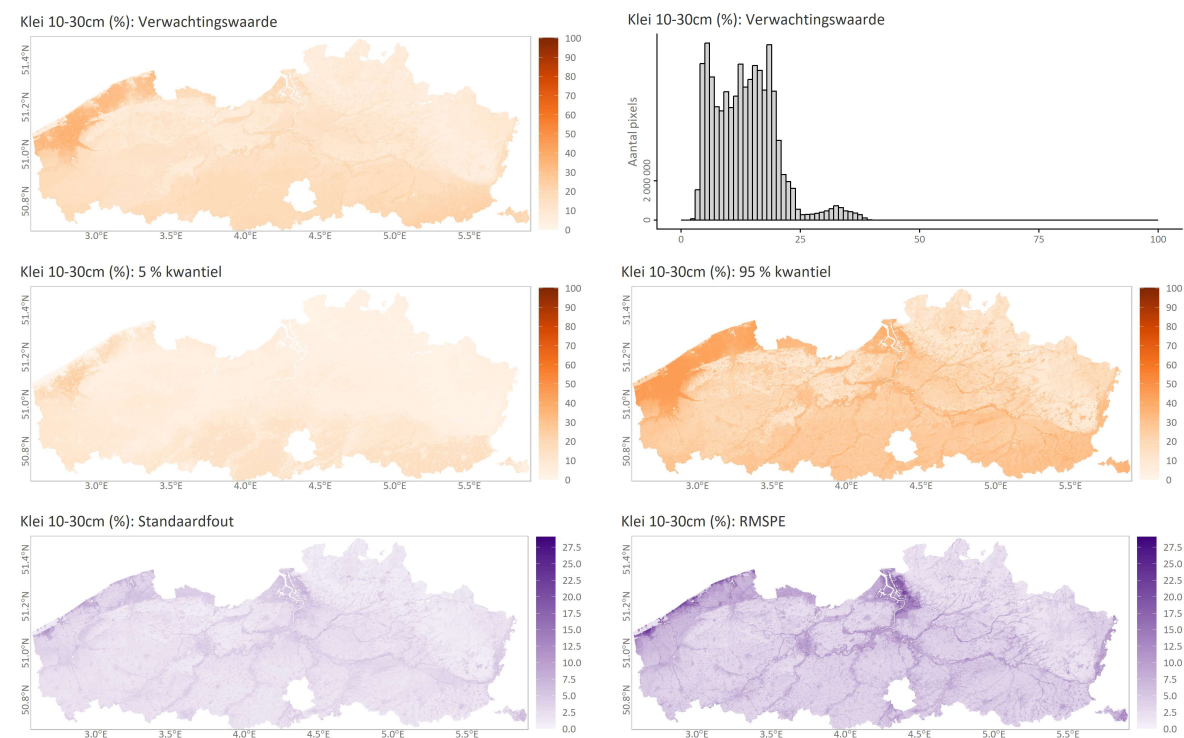
3.7 KLEI

3.7.1 0-10cm



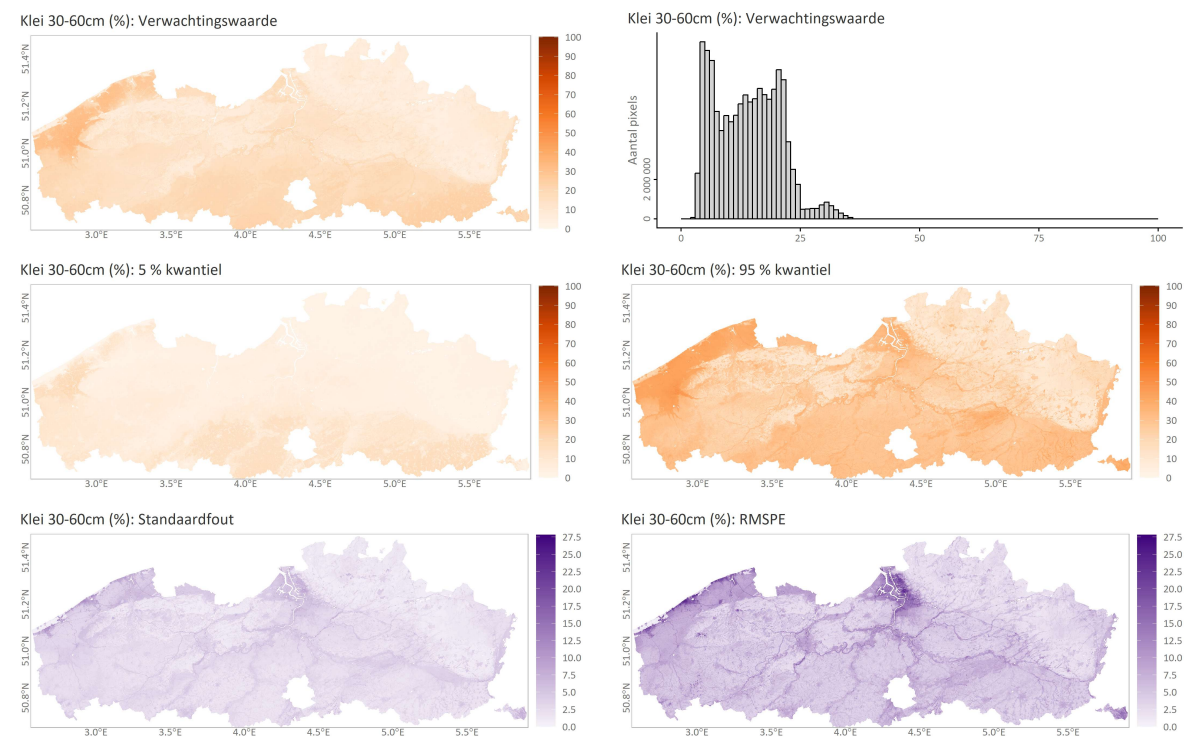
Figuur 31: Klei 0-10cm (%)

3.7.2 10-30cm



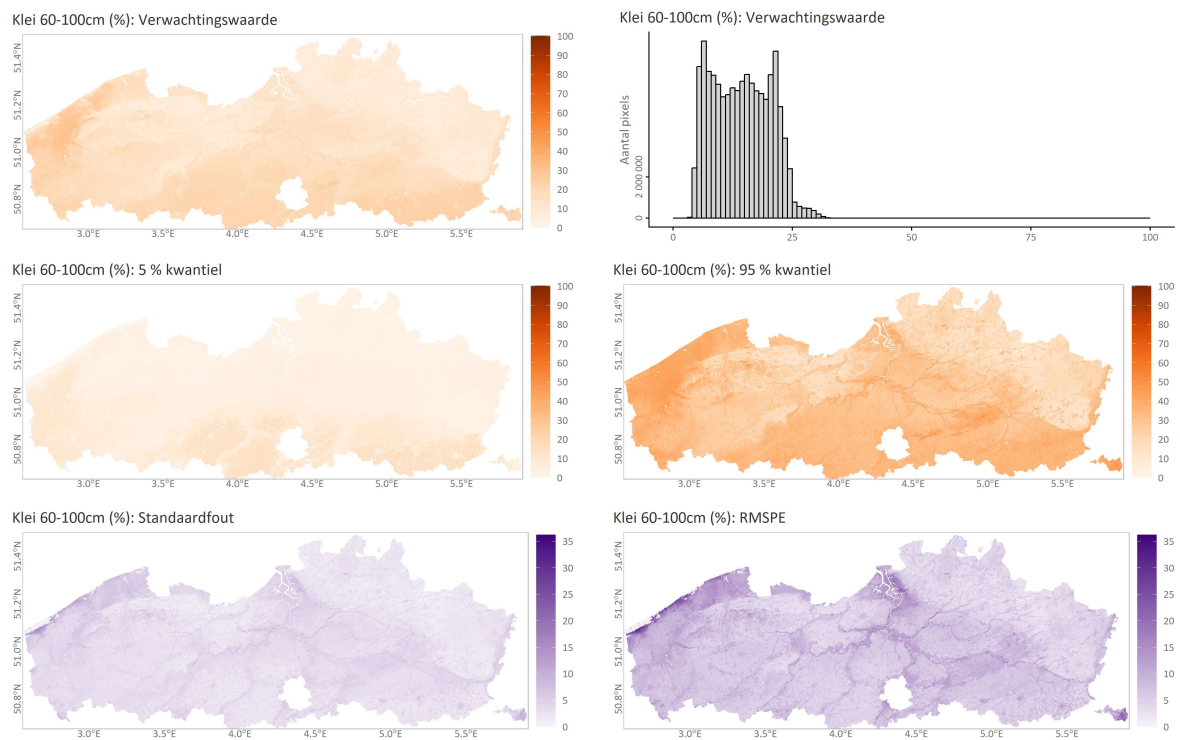
Figuur 32: Klei 10-30cm (g/cm³)

3.7.3 30-60cm



Figuur 33: Klei 30-60cm (g/cm³)

3.7.4 60-100cm



Figuur 34: Klei 60-100cm (g/cm^3)

4 VALIDATIE

In deze sectie worden de geproduceerde kaarten uit sectie 3 gevalideerd aan de hand van enkele technieken. Een eerste manier (sectie 4.1) om kaarten te valideren, is op basis van de verwachte generalisatiefout. Een andere manier, uitgewerkt in sectie 4.2, is om te kijken of de afgeleide schatters voor landgebruiken overeenkomen met schatters op basis van ruwe data. Distributies worden vergeleken in sectie 4.3. Tot slot is er de visuele inspectie in sectie 4.4.

4.1 GENERALISATIEFOUT

In sectie 2.4 is geargumenteerd dat cross-validatie (CV) gebruikt kan worden als schatter voor de verwachte generalisatiefout. Verder is ook gesteld dat omwille van het gebruik van de gekwadrateerde fout als objectieve functie, de CV-schatter overeenkomt met de *mean squared error* (MSE). Op basis van deze grootheid worden twee bijkomende statistieken gerapporteerd die gemakkelijker interpreteerbaar zijn:

$$RMSE = \sqrt{MSE} \text{ en } R^2 = 1 - \frac{MSE}{MSE_0}$$

De $RMSE$ kan worden geïnterpreteerd als de gemiddelde fout wanneer meetwaarden y_i worden vergeleken met de modelvoorspellingen $\hat{\mu}_i$. De R^2 geeft de proportie van de variantie weer die door het model wordt verklaard ten opzichte van de variantie van het referentiemodel – i.e., het model dat voor alle observaties het globale gemiddelde voorspelt. MSE_0 is de schatter van de verwachte generalisatiefout op basis van dit referentiemodel.

Tabel 7 geeft een overzicht van deze metriekeken per scenario. Op basis van de gerapporteerde *p*-waarden kan worden gesteld dat voor pH (KCI) en textuur het DSM-model statistisch significant lagere voorspellingsfouten oplevert dan het referentiemodel. Voor bulkdichtheid neemt de statistische significantie af met toenemende diepte tot slechts beperkt significant op 5 % significantieniveau.

Voor organische C zijn alle p -waarden groter dan 5 % wat impliceert dat er geen statistisch significant bewijs is dat pixelvoorspellingen een lagere fout hebben dan het globaal gemiddelde model. Er zijn verschillende mogelijke redenen hiervoor:

- Het percentage organische C wordt sterk bepaald door het beheer van een bepaalde locatie (bv. input van organisch materiaal in akkers). Dit beheer wordt enkel ruwweg via het landgebruik meegenomen als covariaat in het model. Maar binnen elk landgebruik bepaalt het beheer van elk perceel afzonderlijk sterk het organische C percentage. Vermits hier geen gegevens over zijn, werd dit niet meegenomen in het model wat de slechtere statistische metriecken van organische C percentage en organische C voorraden (deels) zouden kunnen verklaren.
- Het percentage organische C heeft grote uitschieters. Door het weglaten van de organische C percentages > 10% verbeteren de metriecken van het model. De vraag is of een slechter model op basis van alle waarden voor organische C toch niet te verkiezen is boven een model waarbij de hoogste waarden niet worden beschouwd.
- Het DSM-model haalt het landgebruik akker eruit als predictor maar niet als hoofdpredictor. In het jaarrapport van Cmon werden significant lagere organisch C percentage

gerapporteerd voor akker in vergelijking met de andere landgebruiken. Is het beter om voor dit landgebruik de organische C percentages apart te modelleren?

- De GLG en GHG kaarten hebben een grote invloed op de modellering van organische C percentages. Deze kaarten zijn echter niet accuraat voor locaties binnen de eerste meter beneden maaiveld. Dit kan de gemodelleerde waarde beïnvloeden.

Voor organische C voorraden zijn de p -waarden iets beter aangezien ze berekend worden als product van bulkdichtheid en organische C. Hierdoor vertaalt deze onzekerheid zich ook hier in een afnemende modelprestatie met toenemende diepte, maar minder sterk. De p -waarde van 0.0019 voor koolstofvoorraden op 0-30cm diepte wijst op sterk bewijs dat de modelvoorspellingen – op deze diepte althans – wel degelijk een lagere fout vertonen dan het referentiemodel.

Tabel 7: Overzicht van MSE , MSE_0 , $RMSE$ en R^2 voor elke scenario. Standaardfouten van $RMSE$ en R^2 zijn voorlopig niet gegeven. MSE_0 is gebaseerd op het referentiemodel dat voor alle observaties het globale gemiddelde voorspelt. De p -waarden zijn bepaald aan de hand van gecorrigeerde herhaalde t-test (Nadeau & Bengio, 2003)

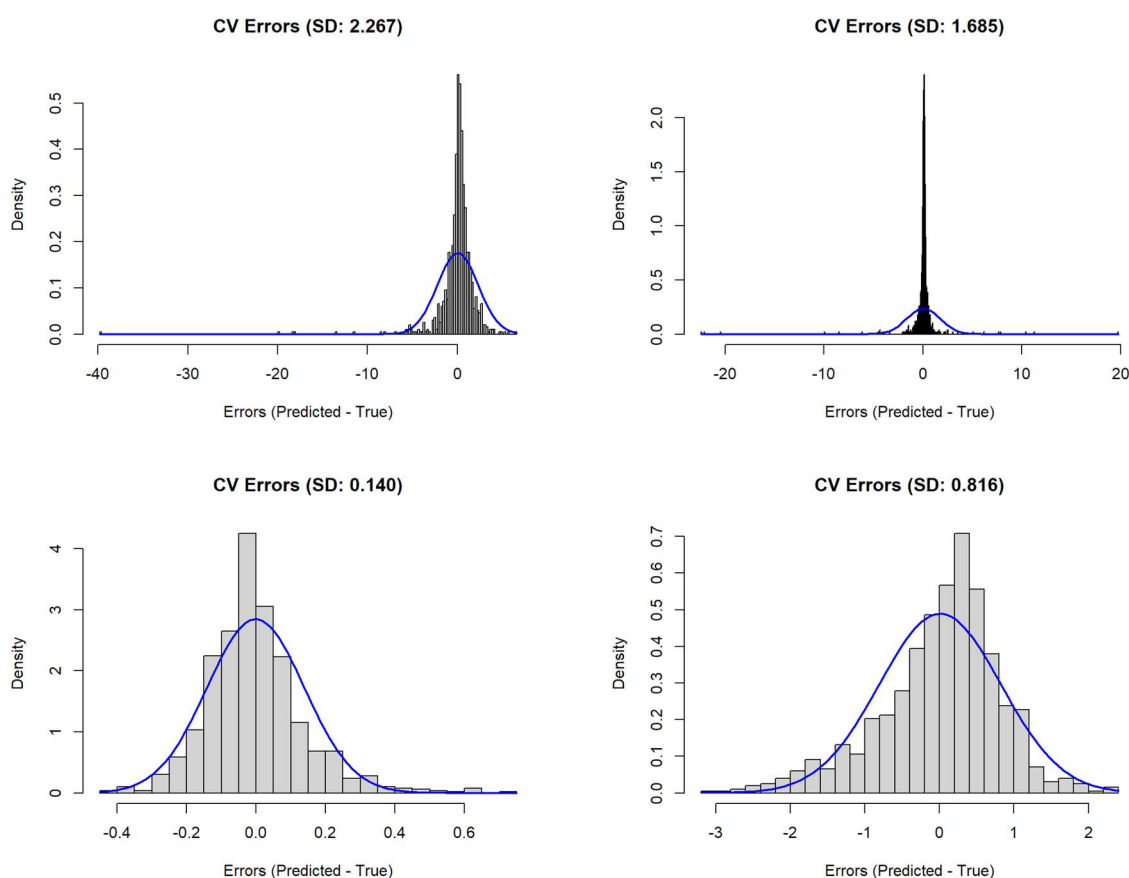
Bodemdescriptor	Diepte	<i>MSE</i>	<i>MSE</i> ₀	Pr(<t)	<i>RMSE</i>	<i>R</i> ²
pH (KCl)	0-10cm	0.666 ±0.035	1.793 ±0.037	<0.0001	0.82	0.63
	10-30cm	0.692 ±0.037	1.626 ±0.039	<0.0001	0.83	0.57
	30-60cm	0.702 ±0.039	1.450 ±0.049	<0.0001	0.84	0.52
	60-100cm	0.725 ±0.040	1.598 ±0.044	<0.0001	0.85	0.55
Totale bulkdichtheid (g/cm³)	0-10cm	0.0332 ±0.0019	0.0704 ±0.0030	<0.0001	0.18	0.53
	10-30cm	0.0198 ±0.0014	0.0314 ±0.0024	0.0003	0.14	0.37
	30-60cm	0.0200 ±0.0017	0.0306 ±0.0029	0.0025	0.14	0.35
	60-100cm	0.0236 ±0.0024	0.0361 ±0.0043	0.0103	0.15	0.35
Organische C (%)	0-10cm	5.1 ±1.8	7.5 ±2.0	0.1977	2.27	0.31
	10-30cm	1.09 ±0.21	1.54 ±0.30	0.1174	1.04	0.29
	30-60cm	1.04 ±0.39	1.41 ±0.51	0.2822	1.02	0.27
	60-100cm	3.1 ±1.1	4.7 ±2.1	0.2509	1.75	0.35
Organische C voorraad (t/ha)	0-30cm	682 ±50	973 ±64	0.0011	26.12	0.30
	0-100cm	4473 ±594	6685 ±1033	0.0399	66.88	0.33
Zand (%)	0-10cm	165 ±20	883 ±26	<0.0001	12.8	0.81
	10-30cm	164 ±19	911 ±22	<0.0001	12.8	0.82
	30-60cm	196 ±14	979 ±34	<0.0001	14.0	0.80
	60-100cm	273 ±19	952 ±43	<0.0001	16.5	0.71
Leem (%)	0-10cm	89.8 ±7.4	513 ±22	<0.0001	9.5	0.83
	10-30cm	89.8 ±6.2	523 ±20	<0.0001	9.5	0.83
	30-60cm	114.3 ±6.4	551 ±24	<0.0001	10.7	0.79
	60-100cm	151.7 ±9.2	536 ±26	<0.0001	12.3	0.72
Klei (%)	0-10cm	28.0 ±4.9	82.2 ±8.6	<0.0001	5.3	0.66
	10-30cm	27.0 ±5.4	85.2 ±8.0	<0.0001	5.2	0.68
	30-60cm	30.3 ±3.6	88.9 ±6.6	<0.0001	5.5	0.66
	60-100cm	44.9 ±4.5	93.9 ±7.5	<0.0001	6.7	0.52

Als MSE_0 referentie wordt hier het globaal gemiddelde gebruikt. In toekomstig werk kan overwogen worden om de modelprestatie ook te vergelijken met een model gebaseerd op de vijf hoofdlandgebruiken waarop Cmon initieel is afgestemd. Een dergelijke vergelijking zou een realistischer beeld van de preformantie geven.

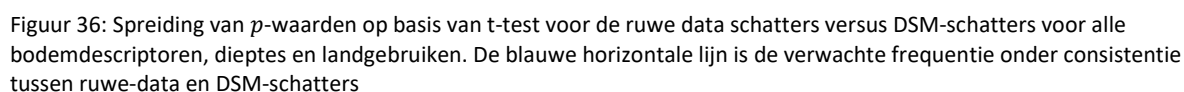
Landgebruik	Cmon ruwe data					Cmon DSM-kaarten			
	Bodem-descriptor	Diepte (cm)	Aantal	Gemiddelde	Standaard-deviatie	Gemiddelde	Standaard-deviatie	Gemiddelde $Pr(> t)$	Standaard-deviatie $Pr(> t)$
Ruimtebeslag	OC	10-30	124	1.58 ±0.09	0.96 ±0.21	1.74	1.09	0.1836	0.6543
Ruimtebeslag	OC	30-60	124	0.89 ±0.07	0.80 ±0.25	1.00	0.90	0.2645	0.7726
Ruimtebeslag	OC	60-100	124	0.48 ±0.05	0.58 ±0.18	0.58	1.25	0.1479	0.0112
Ruimtebeslag	OCst	0-30	124	76.0 ±2.6	29.0 ±3.3	78.5	28.7	0.5048	0.9493
Ruimtebeslag	OCst	0-100	124	140.6 ±6.3	70.0 ±14.0	151.9	71.2	0.2046	0.9494

Tabel 8 omvat in totaal 2x70 tests. Onder de nulhypothese van gelijke schatters wordt verwacht dat louter door toeval ongeveer 5% van de p -waarden onder 0.05 gaat liggen. Voor de gemiddelden met significante afwijking ($p < 0.05$) blijkt dat DSM-schattingen van organische C en pH (KCl) systematisch en tot op één na hoger zijn, terwijl voor bulkdichtheid ze tot op één na lager zijn. Voor de standaarddeviaties met significante afwijking ($p < 0.05$) zijn de DSM-schattingen consistent hoger voor organische C en lager voor BD en pH (KCl). Het feit dat er zowel positieve als negatieve significante afwijkingen bestaan lijkt wel te suggereren dat Vergelijking 12 niet biased is.

Figuur 35 (boven) toont dat scenario's met de grootste discrepanties in standaarddeviatie gepaard gaan met sterk langstaartige predictiefoutverdelingen. Voor de gemiddelden met grootste discrepanties (onder) is dit verband minder uitgesproken: variabelen zoals pH (KCl) vertonen relatief symmetrische foutverdelingen en een hoge R^2 ($\geq 50\%$), maar toch ook vaak een lage p -waarde, wat suggereert dat de geobserveerde verschillen een andere oorsprong hebben.



Er dient benadrukt te worden dat de focus hier niet ligt op individuele verschillen tussen ruwe-data schatters en DSM-gebaseerde schatters, maar op een vergelijking van beide methodologieën als geheel. Zoals reeds aangehaald in de inleiding zouden beide benaderingen, indien de onderliggende aannames geldig zijn, consistente schattingen van dezelfde populatiegrootheden moeten opleveren. Een meer formele evaluatie van deze globale consistentie kan worden uitgevoerd door de verdeling van de resulterende p -waarden te analyseren. Onder de nulhypothese dat beide methodologieën systematisch dezelfde grootheden schatten, worden de p -waarden verwacht uniform verdeeld te zijn. Afwijkingen van deze uniforme verdeling zouden wijzen op systematische discrepanties tussen de benaderingen. Figuur 36 toont de histogrammen van de p -waarden voor zowel de gemiddelden als de standaarddeviaties. Over het algemeen is de verdeling compatibel met uniformiteit, zij het met een lichte maar duidelijke verschuiving naar lagere waarden. Deze bias is meer uitgesproken voor de gemiddelden dan voor de standaarddeviaties, wat enigszins onverwacht is gezien de sterke prestaties van RF-modellen voor verwachtingswaarden. De oorzaak zou verder onderzocht moeten worden. Een mogelijke verklaring is dat Vergelijking 12 impliciet veronderstelt dat de locatie-conditionele DSM-schatters finite-sample unbiased zijn. Indien dit in de praktijk niet geldt, kan dit leiden tot systematische afwijkingen. Daarom wordt in sectie 5 een mogelijke verbetering van deze aanpak voorgesteld door gebruik van *Generalized Random Forests* (GRF). Niettemin zijn de resultaten bemoedigend, aangezien deze DSM-schatters het resultaat zijn van een eerste iteratie waarin nog een paar tekortkomingen aanwezig zijn (cfr. sectie 2).



- Ruwe-data schatters veronderstellen representatieve steekproeven binnen elk landgebruik. Schending van deze aanname (bijvoorbeeld door systematische ontoegankelijkheid van bepaalde locaties) kan leiden tot bias.
- DSM-schatters maken gebruik van covariaten om ruimtelijke patronen te verklaren. Hierdoor kunnen zij in principe corrigeren voor bepaalde vormen van sampling bias maar door de toegenomen complexiteit zijn er ook veel meer zaken die mis kunnen gaan.

mate dit effectief gerealiseerd wordt, en onder welke voorwaarden deze schattingen betrouwbaar zijn, vereist echter verder onderzoek binnen het vervolg van het Cmon-project.

4.3 DISTRIBUTIE

In deze sectie worden de geobserveerde distributies van meetwaarden vergeleken met voorspelde distributies.

4.3.1 Organische C

Voor organische C lijkt de distributie van de voorspelde verwachtingswaarden (Figuur 37, linksboven) visueel niet overeen te komen met de distributie van de oorspronkelijk gemeten waarden (rechtsboven). In het bijzonder vertonen de voorspellingen een bimodale structuur, terwijl de oorspronkelijke data geen duidelijke bimodaliteit vertonen. Dit roept de vraag op of het model wel correct is.

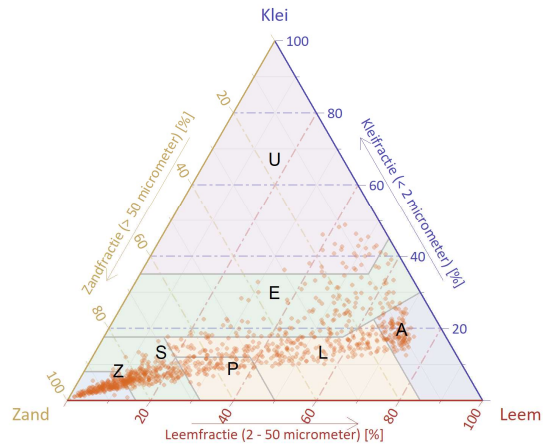
Deze vergelijking is echter misleidend. De gemeten waarden $Y_l = \mu_l + \varepsilon_l$ en de voorspelde verwachtingswaarden $\hat{\mu}_l$ beschrijven fundamenteel verschillende grootheden. De metingen omvatten zowel de ware waarde als alle meet- en ruimtevariabiliteit, terwijl $\hat{\mu}_l$ een schatting is van de ware waarde μ_l . Bijgevolg zal de distributie van $\hat{\mu}_l$ in het algemeen minder spreiding vertonen dan die van Y_l maar is er verder geen directe relatie tussen hun vorm.

Dit laatste kan geïllustreerd worden met een eenvoudig gedachte-experiment. Stel dat de ware waarde ruimtelijk constant is, terwijl metingen onderhevig zijn aan additieve Gaussiaanse ruis. In dat geval zullen de geobserveerde waarden een Gaussiaanse verdeling volgen. De vorm van de verwachtingswaarden zal echter maar één enkele piek vertonen. De distributies van Y_l en μ_l zijn dus qua vorm mogelijk erg verschillend en niet noodzakelijk rechtstreeks vergelijkbaar.

Een meer zinvolle vergelijking wordt bekomen door de meetdistributie $Y_l = \mu_l + \varepsilon_l$ te reconstrueren op basis van de schatters $\hat{\mu}_l + \hat{\varepsilon}_l$ van het DSM-model. Gegeven dat enkel de eerste twee momenten $\mathbb{E}[Y_l] = \mu_l$ en $\text{Var}[Y_l] = \sigma^2$ expliciet gemodelleerd zijn, is een bijkomende aanname nodig over de vorm van Y_l . In deze analyse wordt aangenomen dat Y_l lognormaal verdeeld is. Vervolgens kan de theoretische distributie bepaald worden waaruit gesampled zou worden indien een random sample zou genomen worden uit de huidig bemonsterde Cmon locaties. Deze distributie is weergegeven in Figuur 37 linksonder. De bimodaliteit is verdwenen omdat de ruis tussen de twee modi gesuperponeerd wordt. Zo ontstaat een distributie die veel beter aansluit bij de geobserveerde data. Tot slot geeft Figuur 37 rechtsonder de theoretische distributie weer die zou worden bekomen indien alle mogelijke locaties in Vlaanderen zou worden bemonsterd.

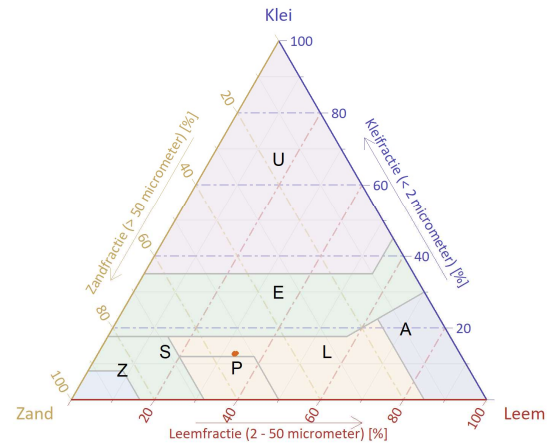
% leem, 35 % klei, 15 % zand anderszids. Tot slot moet, net zoals in voorgaande sectie 4.3.1, opgemerkt worden dat distributies van observaties niet rechtstreeks vergeleken mogen worden met distributies van voorspelde verwachtingswaarden. De overeenkomstige uit het model afgeleide meetwaardedistributie is hier echter niet weergegeven.

Textuurdriehoek 0-10cm: Meeting

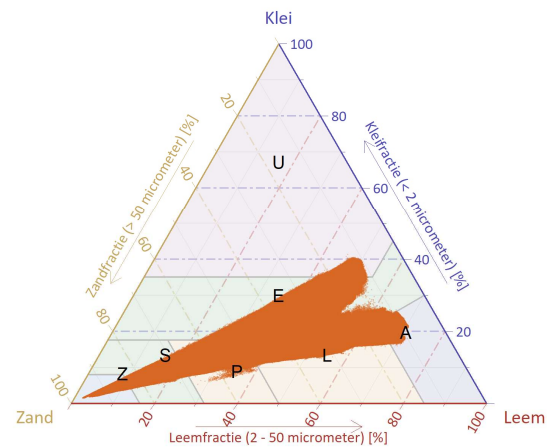
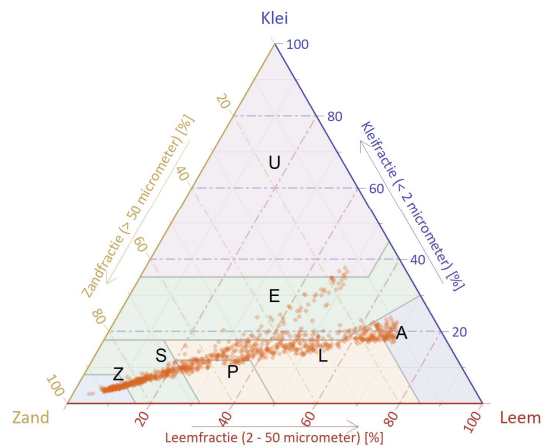


Textuurdriehoek 0-10cm: Verwachtingswaarde (data)

Textuurdriehoek 0-10cm: Globaal gemiddelde

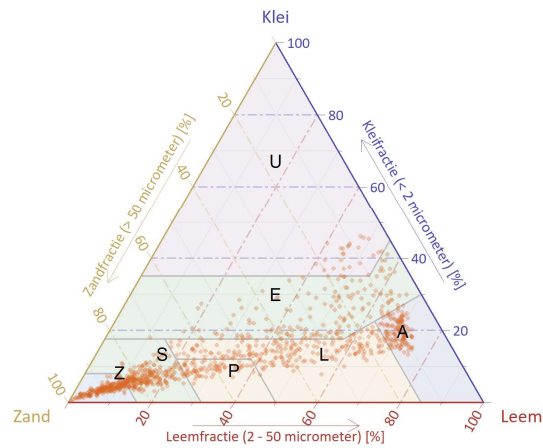


Textuurdriehoek 0-10cm: Verwachtingswaarde (kaart)

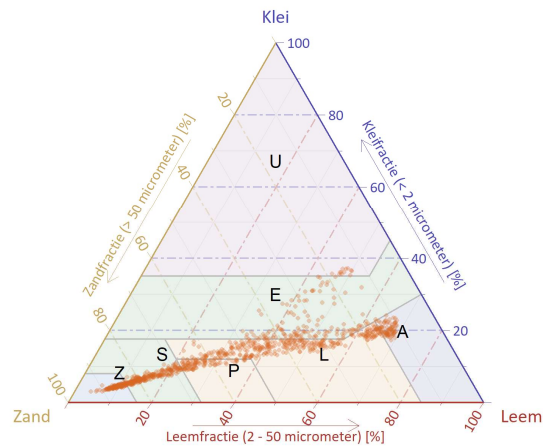


Figuur 38: Originele metingen van textuur bij 0-10cm (linksboven), globaal gemiddelde model (rechtsboven), DSM model voorspellingen op basis van originele data (linksonder) en pixel-waarden van de 10x10m verwachtingskaart (rechtsonder). Alle voorspellingen zijn op basis van 10-fold cross-validatie

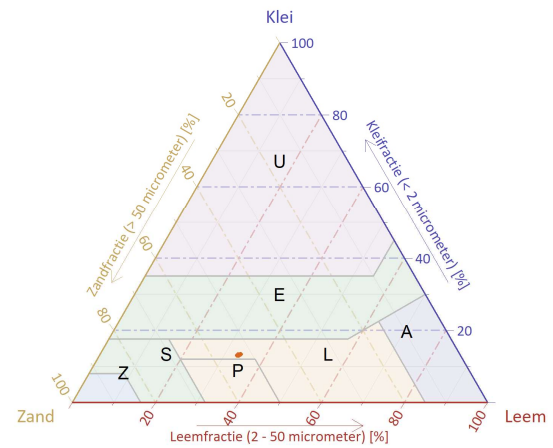
Textuurdriehoek 10-30cm: Meeting



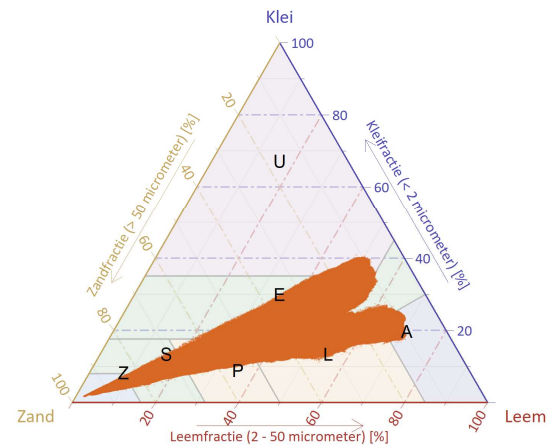
Textuurdriehoek 10-30cm: Verwachtingswaarde (data)



Textuurdriehoek 10-30cm: Globaal gemiddelde



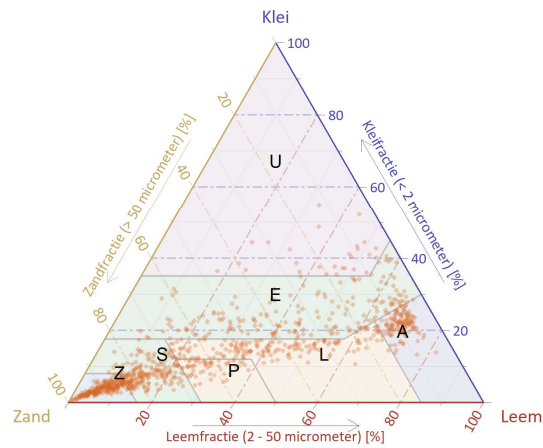
Textuurdriehoek 10-30cm: Verwachtingswaarde (kaart)



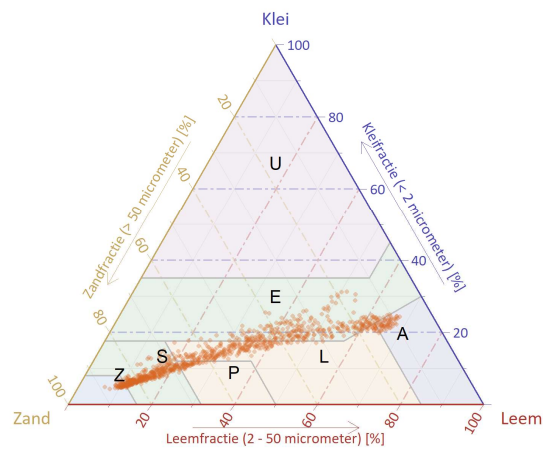
Figuur 39: Originele metingen van textuur bij 10-30cm (linksboven), globaal gemiddelde model (rechtsboven), DSM model voorspellingen op basis van originele data (linksonder) en pixel-waarden van de 10x10m verwachtingskaart (rechtsonder). Alle voorspellingen zijn op basis van 10-fold cross-validatie

////////////////////////////////////

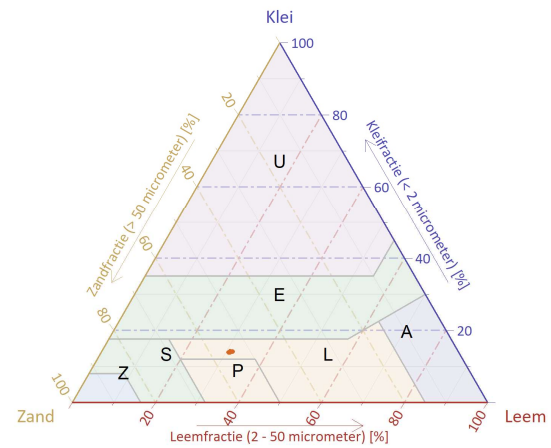
Textuurdriehoek 60-100cm: Meeting



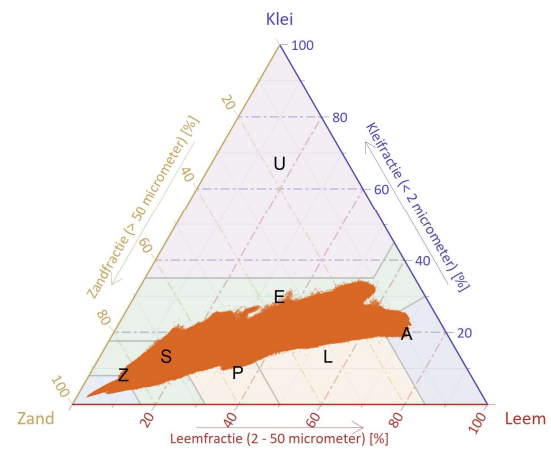
Textuurdriehoek 60-100cm: Verwachtingswaarde (data)



Textuurdriehoek 60-100cm: Globaal gemiddelde



Textuurdriehoek 60-100cm: Verwachtingswaarde (kaart)

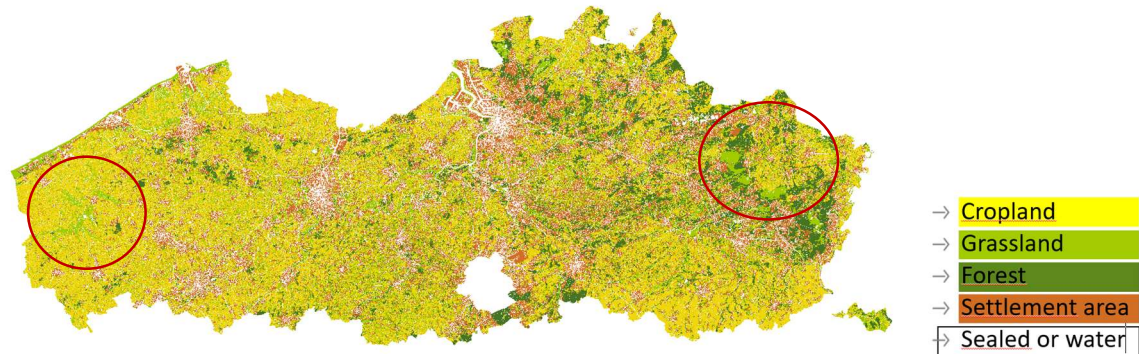


Figuur 41: Originele metingen van textuur bij 60-100cm (linksboven), globaal gemiddelde model (rechtsboven), DSM model voorspellingen op basis van originele data (linksonder) en pixel-waarden van de 10x10m verwachtingskaart (rechtsonder). Alle voorspellingen zijn op basis van 10-fold cross-validatie

4.4 VISUELE VALIDATIE

4.4.1 Patronen in covariaten

4.4.1.1 Landgebruik

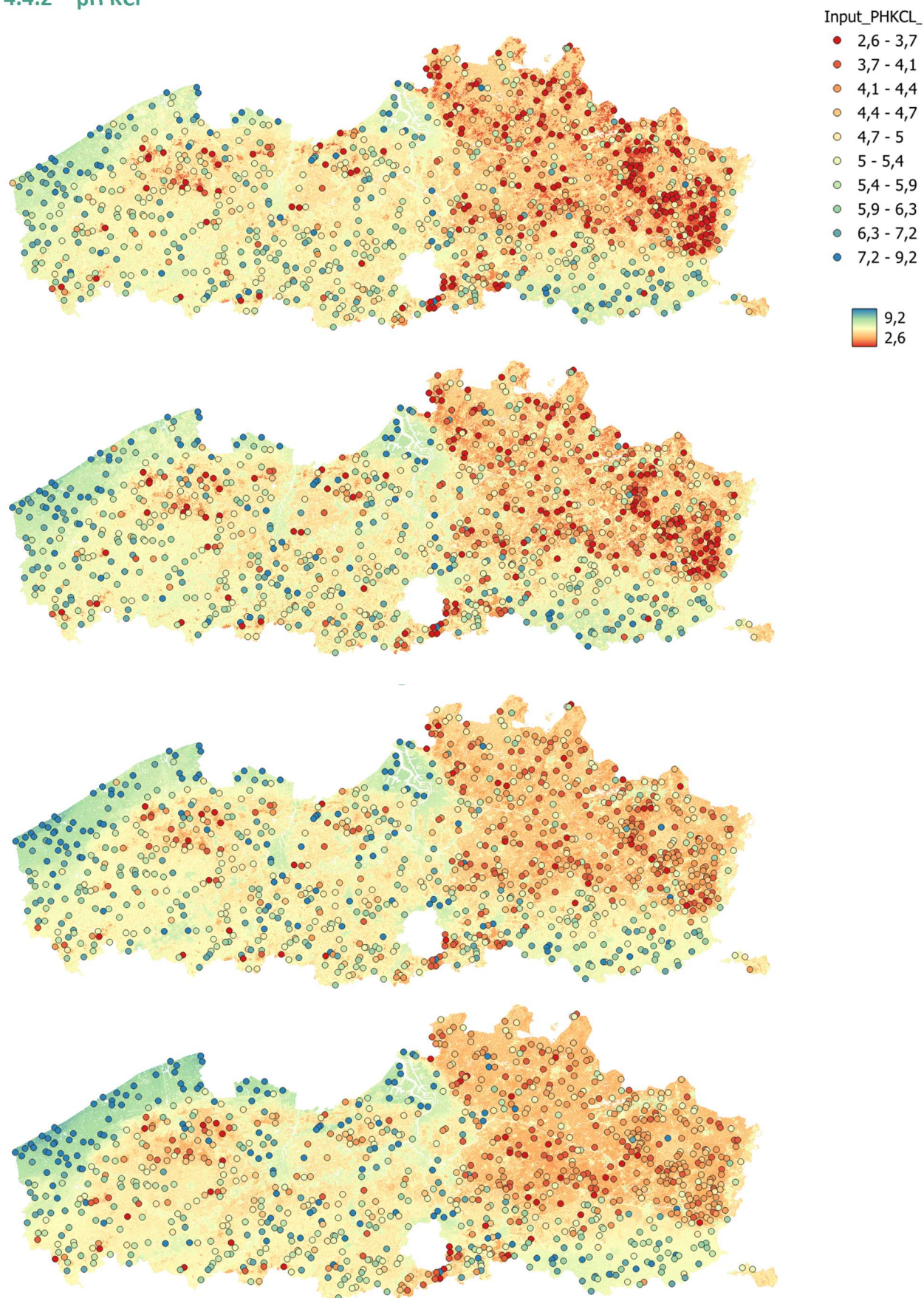


Figuur 42: Landgebruikskartaal afgeleid van de Cmon-landgebruikskartaal (Oorts et al., 2025b)

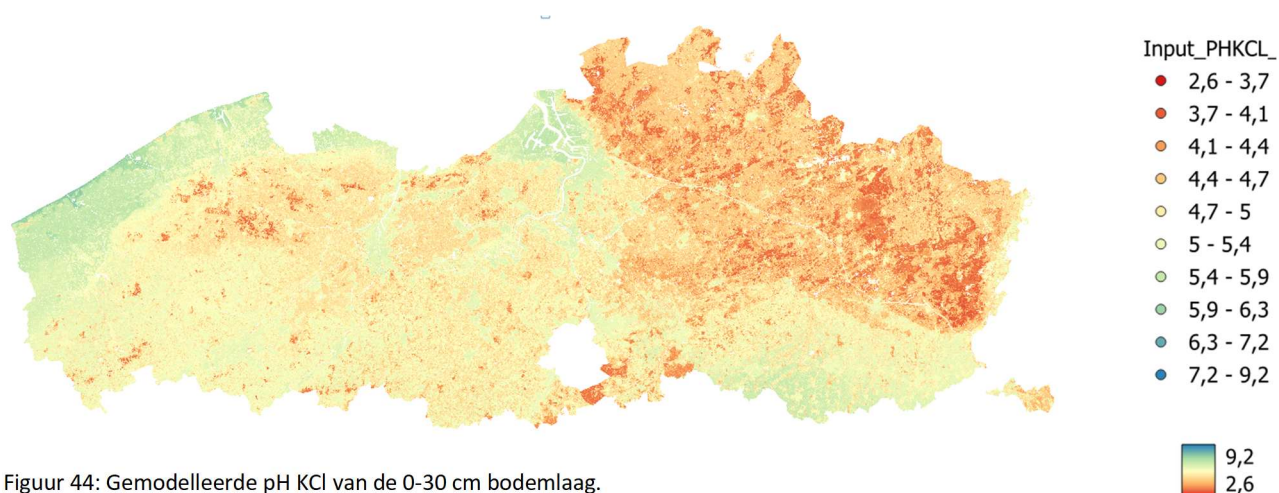
In de kaart van het landgebruik zijn 2 patronen zichtbaar die terugkomen in verschillende kaartlagen:

- Bij de Polders zie je vingers grasland uitsteken.
- In Limburg zie je in de mijnstreek een boog van bos en grasland.

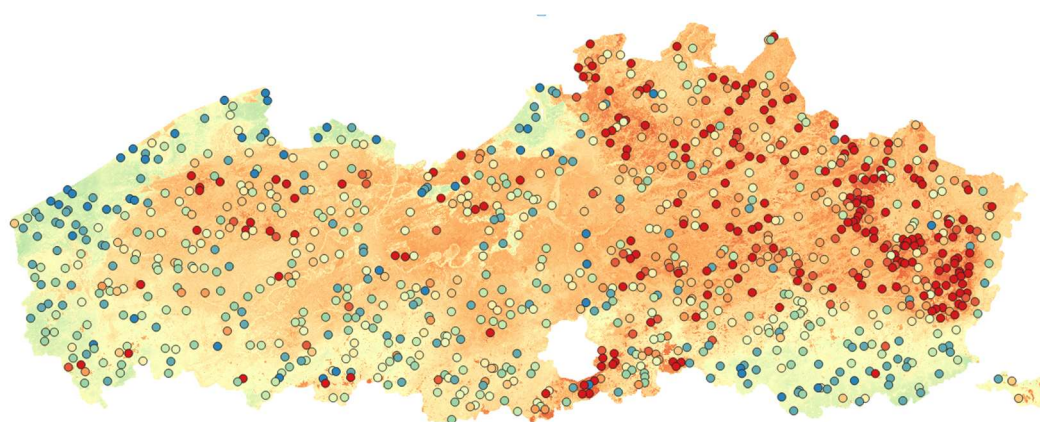
4.4.2 pH KCl



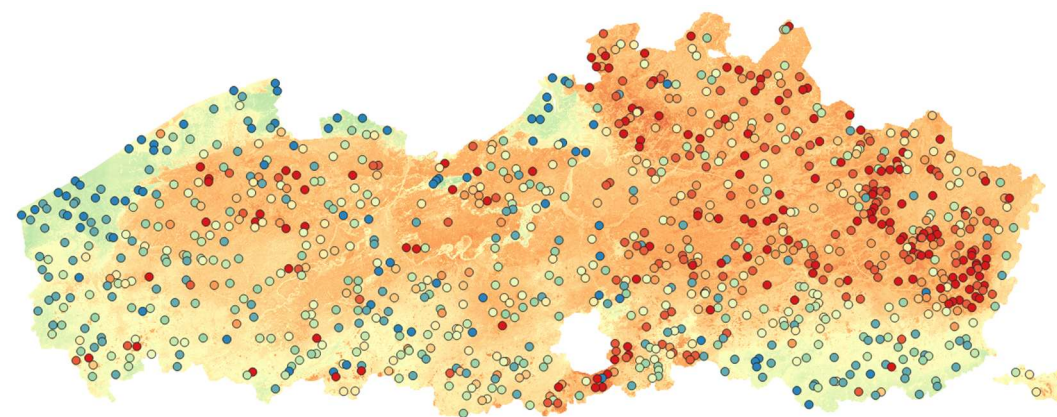
Figuur 43: Gemeten en gemodelleerde pH KCl van de 0-10, 10-30, 30-60 en 60-100 cm bodemlaag.



Figuur 44: Gemodelleerde pH KCl van de 0-30 cm bodemlaag.



Figuur 45: Gemodelleerde pH KCl van de 0-10 cm bodemlaag op basis van Bodemdatabank-STAT vergeleken met de huidige Cmon-metingen van die bodemlaag



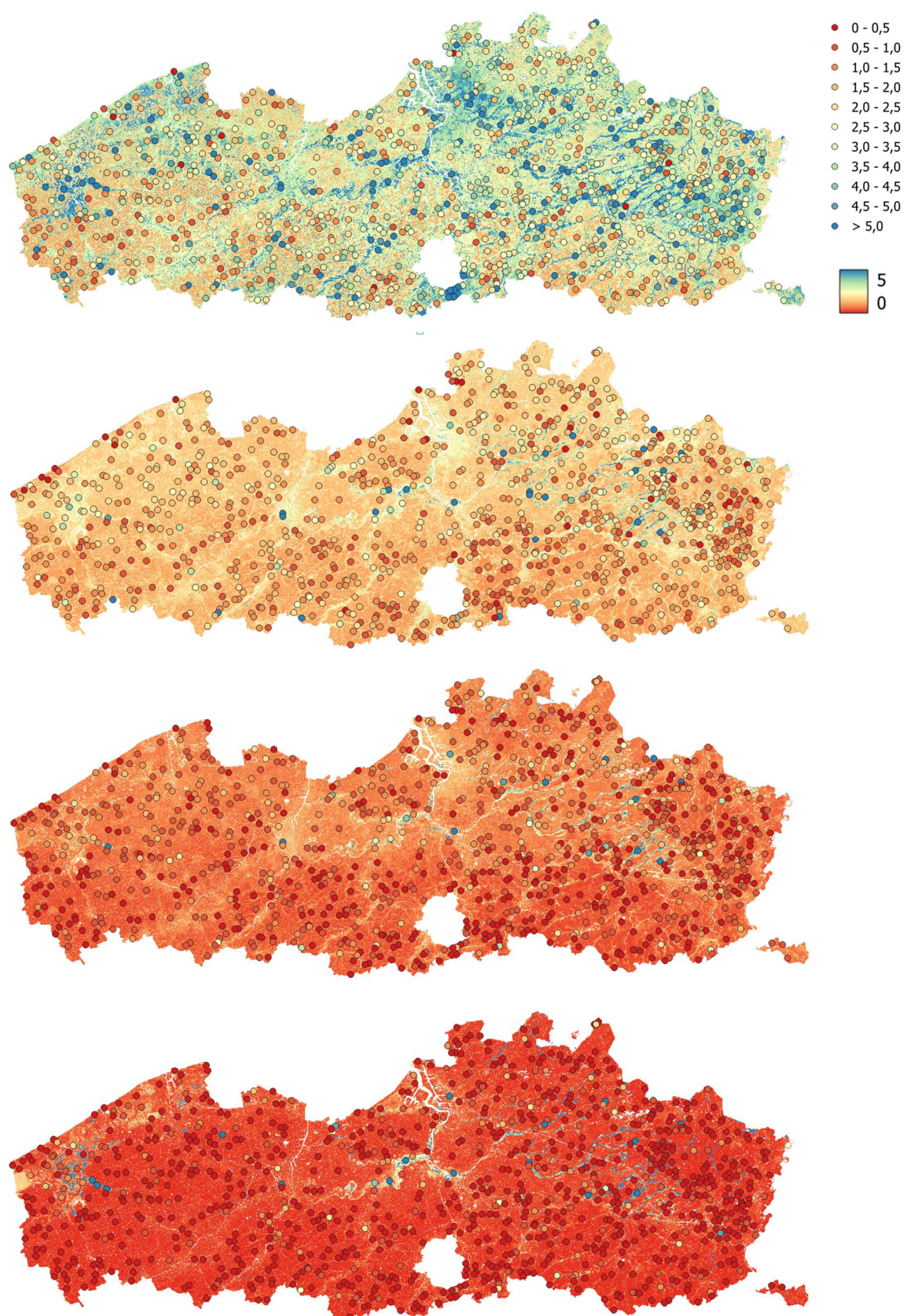
Figuur 46: Gemodelleerde pH KCl van de 10-30 cm bodemlaag op basis van Bodemdatabank-STAT vergeleken met de huidige Cmon-metingen van die bodemlaag

Bij de visuele controle van de pH KCl kaartlagen stelden we volgende zaken vast:

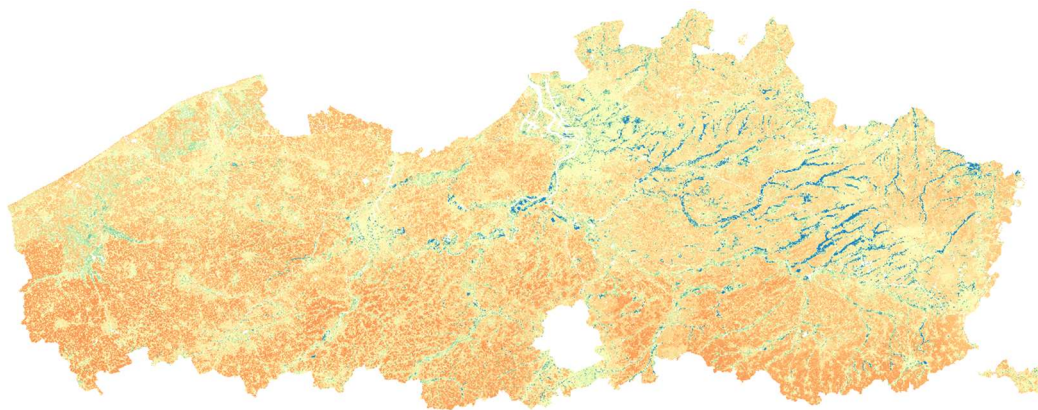
- De polders hebben de hoogste pH KCl.

- De pH KCl in de polders stijgt duidelijk met de diepte. Dit kan wellicht verklaard worden door de opname van basische kationen uit de bovenste lagen en het feit dat wortels verzurend werken.-.
- Algemeen wordt gesteld dat de Vlaamse zandstreek een lagere pH heeft, maar dit komt niet tot uiting uit de metingen en de modellering. Weegt het landgebruik zwaarder door dan de textuur in de Vlaamse zandstreek?
- De Kempen, met grove zandbodems en ook veel bos, hebben in vergelijking met de andere streken bodems met gemiddeld een lagere pH.
- Het landgebruik bos heeft ten opzichte van de andere landgebruiken duidelijk de bodems met de laagste pH. Hierbij vormt de bosgordel in de mijnstreek een zeer uitgesproken patroon.
- De vergelijking van de huidige Cmon-metewaarden met de gemodelleerde gebiedsdekkende kaartlagen op basis van de historische dataset van de Belgische bodemkartering (1947 – 1975) toont duidelijk een algemeen hogere pH KCl op basis van de huidige metingen in vergelijking met de historische metingen. Dit geldt het meest uitgesproken voor de Vlaamse Zandstreek en de Zandleemstreek. Mogelijke verklaringen zouden kunnen zijn dat er momenteel meer aandacht is voor de pH van de bodem in vergelijking met 50 jaar geleden en/of dat de kalibratie van de pH-metingen ten tijde van de bodemkartering minder nauwkeurig was.

4.4.3 Organische C



Figuur 47: Gemeten en gemodelleerde organische C percentages van de 0-10, 10-30, 30-60 en 60-100 cm bodemlaag.

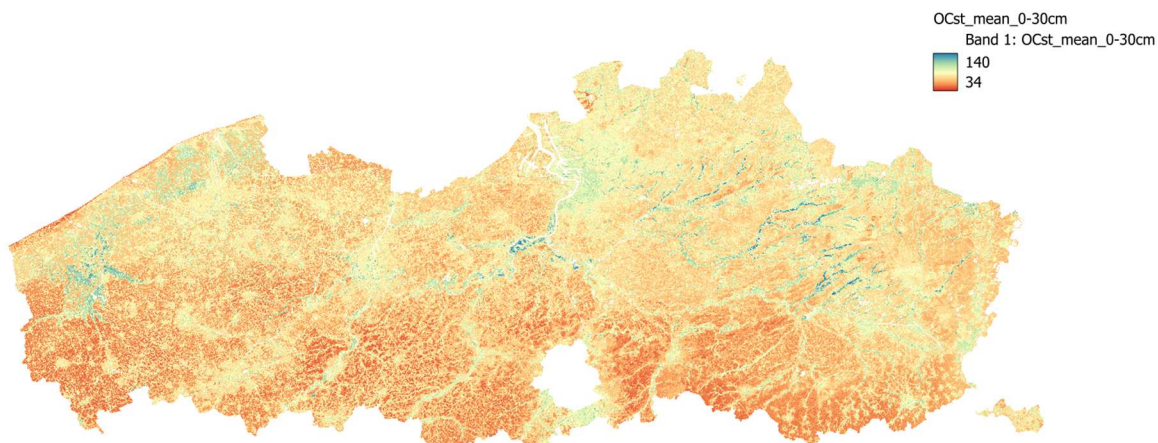


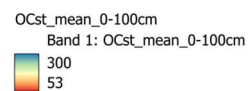
Figuur 48: Gemodelleerde organische C percentages van de 0-30 cm bodemlaag.

We stelden volgende zaken vast bij de visuele controle van de kaarten:

- De valleien bevatten duidelijk hogere percentages organische C.
- Het organische C percentage daalt sterk met de diepte. In de 0-30 cm bodemlaag is er een groot verschil tussen de 0-10 en 10-30 cm bodemlaag.
- Ondanks de slechtere metrieken voor organische C percentage lijken de patronen op de kaartlagen voor organische C percentage wel te kloppen.

4.4.4 Organische C voorraden

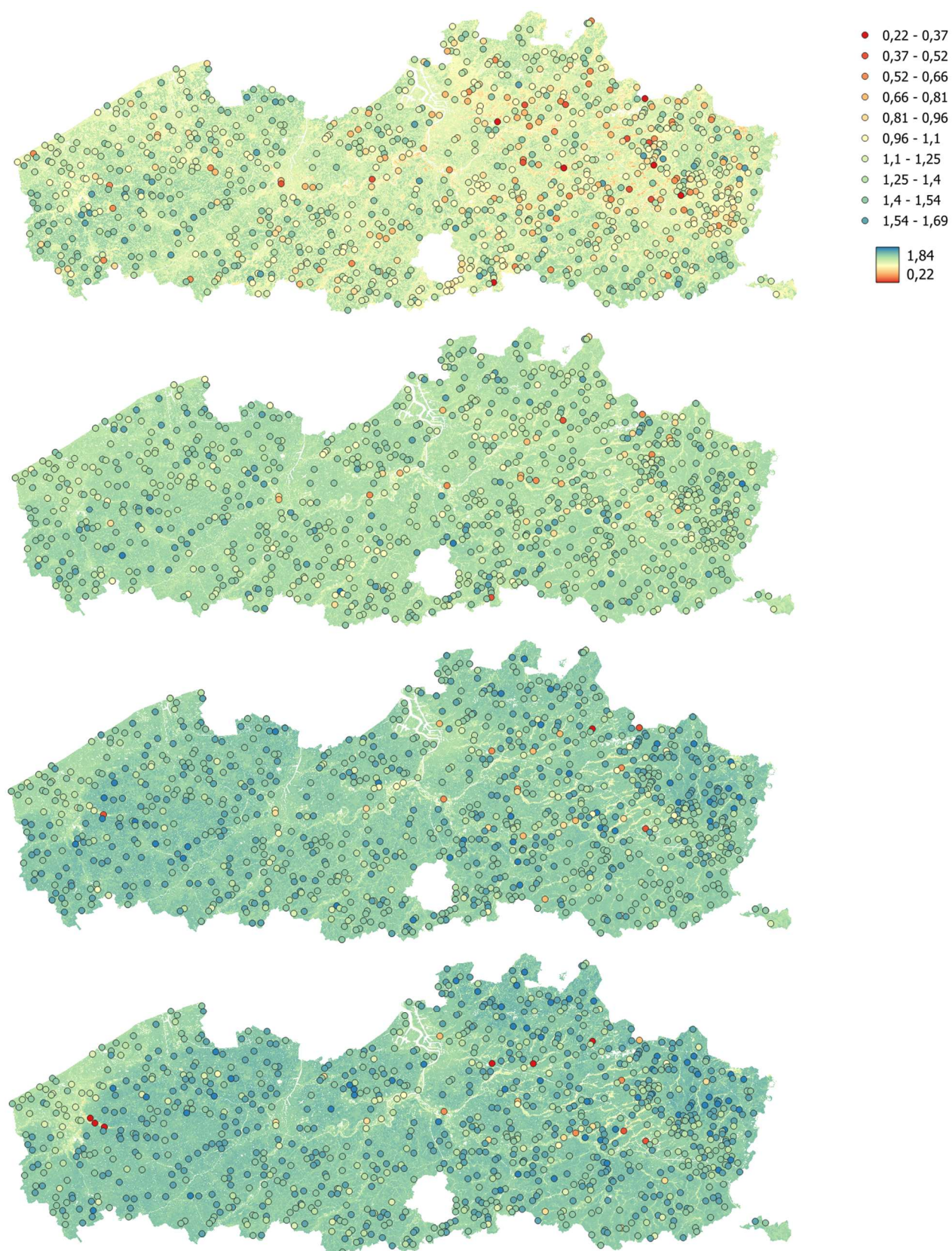




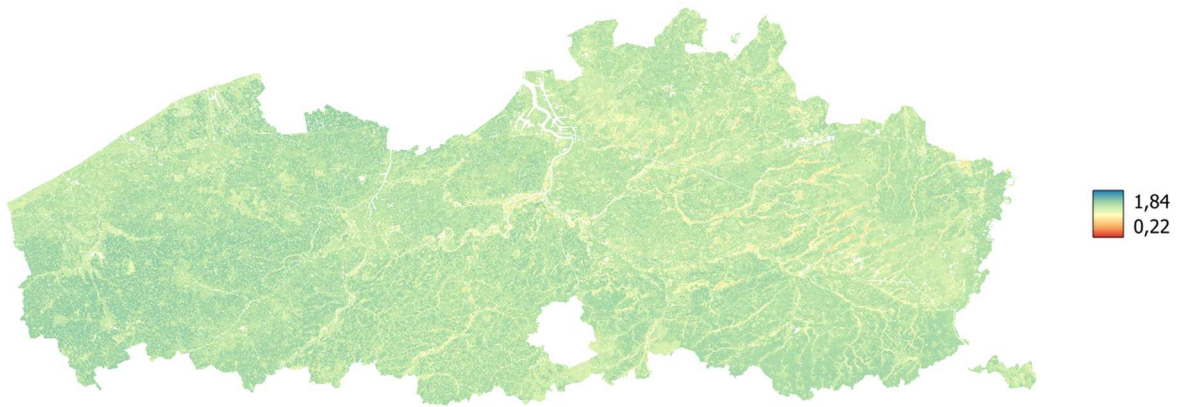
Figuur 49: Gemodelleerde organische C voorraden van de 0-30 en 0-100 cm bodemlaag met (alle waarden boven respectievelijk 140 en 300 t ha⁻¹ hebben dezelfde kleur).

- De ‘poldervingers’ hebben een duidelijk hogere C-voorraad, te wijten aan grasland als landgebruik.
- De hogere waarden in de Moeren (uitgebreikte polders) kunnen verklaard worden door nog een resterende aanwezigheid van veen.
- De valleien bevatten duidelijk hogere voorraden organische C, wellicht ook hier grotendeels te wijten aan het voorkomen van veen.

4.4.5 Bulkdichtheid



Figuur 50: Gemeten en gemodelleerde bulkdensiteit van de 0-10, 10-30, 30-60 en 60-100 cm bodemlaag (g cm^{-3}).

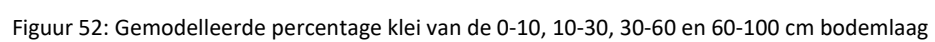


Figuur 51: Gemodelleerde bulkdensiteit van de 0-30 cm bodemlaag (g cm^{-3}).

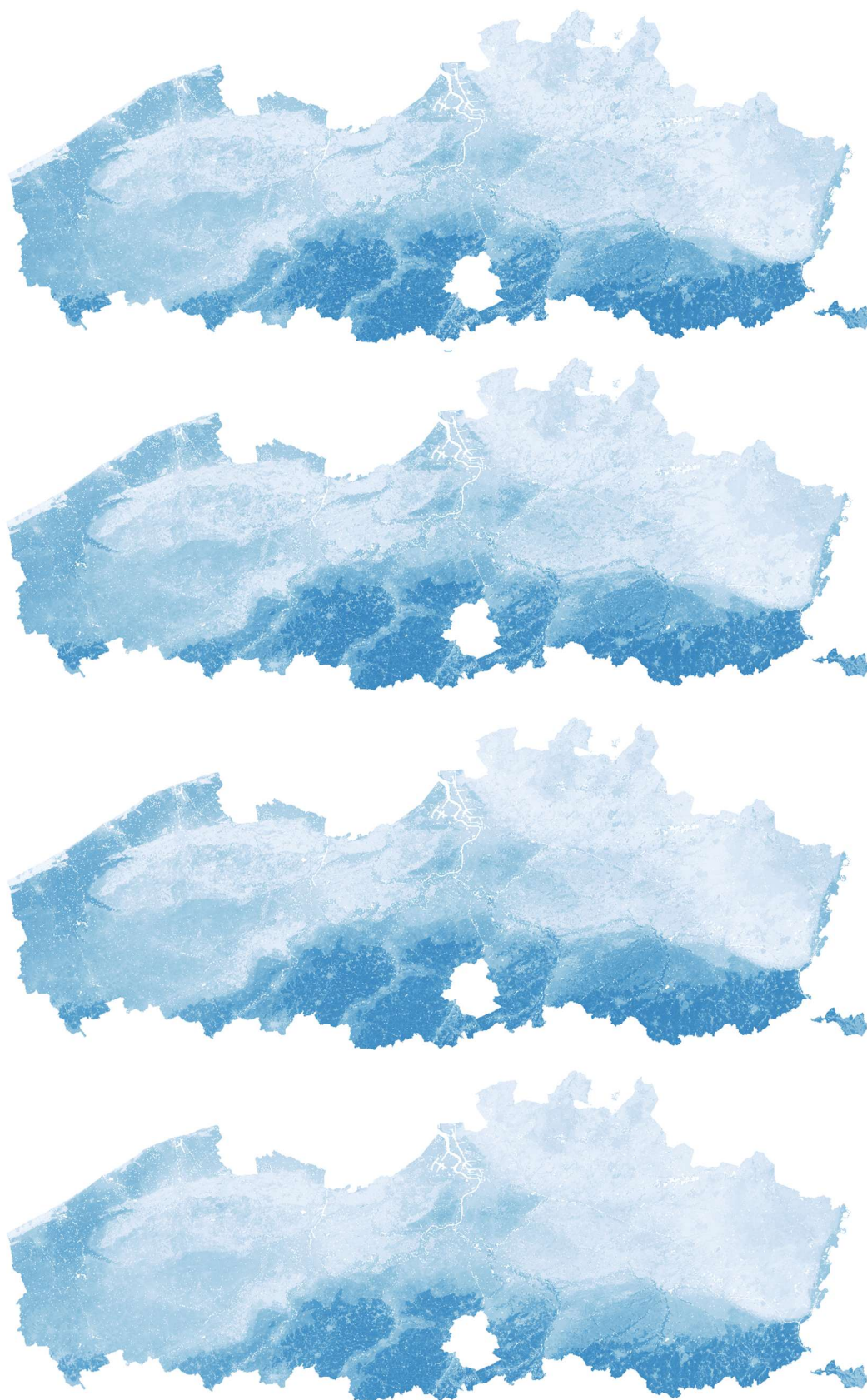
Voor de kaartlagen van de gemodelleerde bulkdensiteit stellen we vast dat:

- De bulkdensiteit daalt met de diepte.
- De bulkdensiteit volgt een omgekeerd patroon in vergelijking met het organische C percentage: hoe hoger het organische C percentage, hoe lager de bulkdensiteit.
- De polders (klei) hebben de laagste bulkdensiteit over alle bodemlagen.
- Een groot deel van West- en Oost-Vlaanderen hebben duidelijk hoger bulkdensiteiten; een gevolg van de meer intensieve landbouwpraktijken (vb. groenteteelt)?

4.4.6.1 % klei

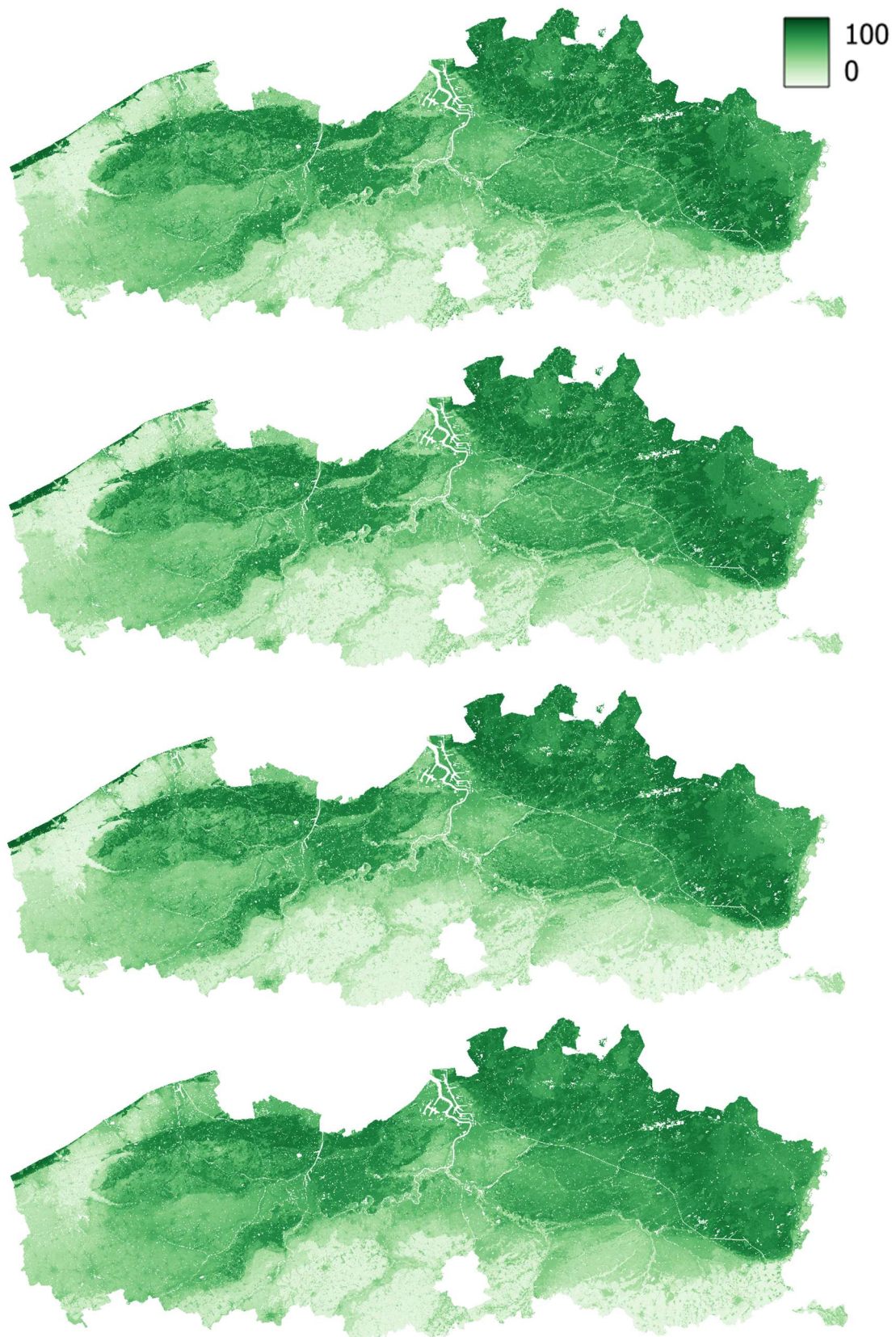


4.4.6.2 % leem



Figuur 53: Gemodelleerde percentage leem van de 0-10, 10-30, 30-60 en 60-100 cm bodemlaag

4.4.6.3 % zand



Figuur 54: Gemodelleerde percentage zand van de 0-10, 10-30, 30-60 en 60-100 cm bodemlaag

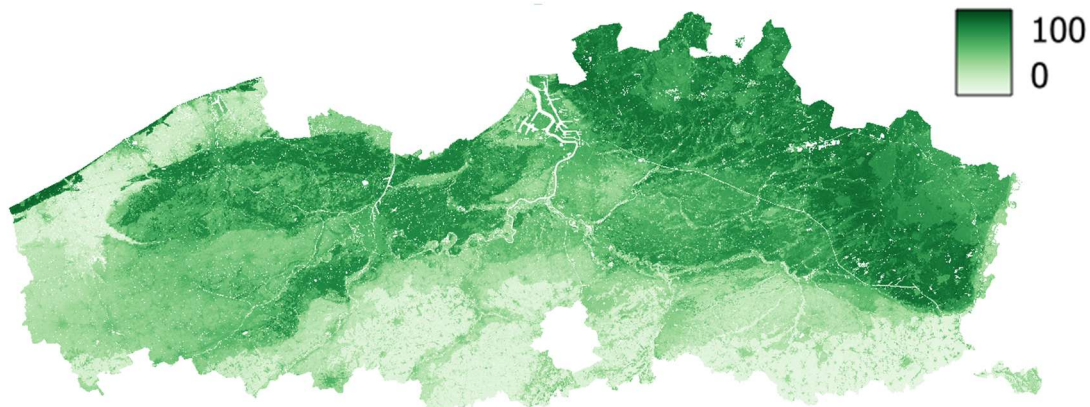
4.4.6.4 % klei, % leem en % zand in de 0-30 cm bodemlaag



Figuur 55: Gemodelleerde percentage klei van de 0-30 cm bodemlaag



Figuur 56: Gemodelleerde percentage leem van de 0-30 cm bodemlaag



Figuur 57: Gemodelleerde percentage zand van de 0-30 cm bodemlaag

4.4.7 Besluit visuele controle

Algemeen kan gesteld worden dat de patronen van de gemodelleerde bodemeigenschappen een goede overeenkomst vertonen met de gemeten puntdata en de verwachte patronen van de bodemeigenschappen in Vlaanderen. Deze gemodelleerde bodemeigenschappen laten evenwel niet toe om uitspraken te doen op beperkte oppervlaktes (vb. perceel, project).

5 VOORUITZICHTEN

Doorheen dit rapport werden verschillende methodologische en praktische beperkingen geïdentificeerd. In deze sectie worden de belangrijkste pistes voor verder onderzoek samengevat.

Een eerste en fundamenteel aandachtspunt betreft de statistische eigenschappen van de gebruikte RF-modellen (cfr. einde sectie 2.1.1). De klassieke RF-modellen gebaseerd op Breiman (2001) zijn empirisch succesvol, maar de theoretische eigenschappen (bijv. consistentie, unbiasedness) zijn nooit bewezen. Voor toepassingen zoals het schatten van koolstofvoorraden per landgebruik is dit problematisch, aangezien dergelijke toepassingen impliciet vereisen dat lokale schatters $\hat{\mu}_l$ en $\hat{\sigma}_l$ betrouwbaar en aggregaerbaar zijn. De resultaten in sectie 4.2 suggereren dat er in de huidige aanpak mogelijks nog systematische bias aanwezig is. Methoden zoals *Generalized Random Forests* (GRF, Tibshirani et al., 2025) bieden hier een theoretisch beter onderbouwd alternatief, met bewezen consistentie en asymptotische unbiasedness en normaliteit (Wager & Athey, 2018), maar vereisen een (substantiële) herwerking van de huidige workflow en zullen ongetwijfeld gepaard gaan met een reductie in R^2 van 0 tot 10 %. En dan nog blijft een belangrijk vereiste – de finite-sample unbiasedness – onbewezen voor non-lineaire modellen zonder extra beperkingen op hun structuur. Deze randvoorwaarde is echter nodig voor het berekenen van unbiased schatters van bodemdescriptoren voor landgebruiken op basis van DSM-kaarten.

Nauw verbonden hiermee is de kwantificatie van onzekerheid. De huidige variantieschatters voor $\hat{\mu}_I$, gebaseerd op jackknife-methoden, zijn zowel computationeel intensief als biased (cfr. sectie 2.1.2). Daarnaast steunen besluitvormende toepassingen (e.g., betrouwbaarheidsintervallen) op asymptotische normaliteitsaannames die niet gegarandeerd zijn (cfr. vorige paragraaf en sectie 2.2.2). Alternatieve methoden voor efficiënte en robuuste onzekerheidskwantificatie vormen dan ook een belangrijke onderzoekslijn.

Een meer algemene vraag is of best gewerkt wordt met één algemeen gefit model waaruit vervolgens alle statistieken gehaald worden, of meerdere gespecialiseerde modellen per statistiek. Momenteel ligt de voorkeur bij het laatste. Bijvoorbeeld voor het modeleren van MSPE (sectie 2.1.4) wordt een tweede model gefit. Het is echter nog niet helemaal duidelijk welke voor- en nadelen dit met zich meebrengt. De systematische overschatting van geaggregeerde variantie bij organische C in sectie 4.2 is mogelijk hieraan gerelateerd en zou verder onderzocht moeten worden.

Voor samengestelde modellen (sectie 2.2) blijven er eveneens belangrijke open vragen. Bijvoorbeeld de juiste impact van de ALR-transformatiefunctie op de schatters van onzekerheid en kwantielen, alsook de beste manier voor het modeleren van hun correlatiestructuur.

Op praktisch vlak zijn er verschillende uitbreidingsmogelijkheden op de workflow. De huidige workflow ondersteunt geen ontbrekende (NA) waarden in covariaten, wat kan leiden tot verkeerde covariatenselectie en mogelijk resulterende bias (cfr. einde sectie 0). Verder is de codebase sterk afgestemd op de huidige toepassing en in de huidige vorm niet geschikt voor breder gebruik. Tot slot zijn ook bepaalde afleidingen zoals Vergelijking 6 en 12 voorlopig alleen analytisch afgeleid, maar niet via simulaties gevalideerd.

Ook op het vlak van covariaten blijft er ruimte voor verbetering. Bepaalde SCORPAN-componenten, zoals *age* en *spatial position*, zijn momenteel niet opgenomen. Bijvoorbeeld tijd en ruimtelijke coördinaten zouden meegenomen kunnen worden. Daarnaast ontbreken ook mogelijk impactvolle variabelen, zoals het landbeheer.

Tot slot zijn er ook toepassingsgerichte onderzoeksvragen. In Oorts et al. (2025a) was de conclusie dan het detecteren van 4 % veranderingen in organische C voorraden met klassieke

Verder onderzoek en implementatie van deze verbeteringen is belangrijk om de DSM-kaarten te laten evolueren van exploratieve producten naar meer robuuste instrumenten voor kwantitatieve besluitvorming.

6 PUBLICATIE VAN DE KAARTLAGEN MET GEMODELLEERDE BODEMEIGENSCHAPPEN

De resulterende datasets van deze onderzoeksopdracht (Tabel 9) zijn als multiband raster gepubliceerd in de Databank Ondergrond Vlaanderen (DOV). Ze zijn beschikbaar in de online [DOV-verkenner](#). In de DOV boomstructuur vind je ze onder Bodem > Gemodelleerde bodemeigenschappen > Recente bodemdata. Meer info vind je op DOV-webpagina [Gebiedsdekkende gemodelleerde bodemeigenschappen - Virtuele bodemanalyse](#).

Tabel 9: Overzicht en locatie van de beschikbare kaartlagen in de DOV-verkenner

Naam DOV-kaartlaag	Bodemlagen	Eenheid	Locatie in de boomstructuur in de online DOV-verkenner
Gemodelleerde fractie klei - Recente bodemdata	0-10 cm 10-30 cm 30-60 cm 60-100 cm	%	Bodem > Gemodelleerde bodemeigenschappen > Recente bodemdata > Textuur
Gemodelleerde fractie leem - Recente bodemdata	0-10 cm 10-30 cm 30-60 cm 60-100 cm	%	Bodem > Gemodelleerde bodemeigenschappen > Recente bodemdata > Textuur
Gemodelleerde fractie zand - Recente bodemdata	0-10 cm 10-30 cm 30-60 cm 60-100 cm	%	Bodem > Gemodelleerde bodemeigenschappen > Recente bodemdata > Textuur
Gemodelleerde pH KCl - Recente bodemdata	0-10 cm 10-30 cm 30-60 cm 60-100 cm	-	Bodem > Gemodelleerde bodemeigenschappen > Recente bodemdata > pH
Gemodelleerde percentage organische koolstof (t per ha) - Recente bodemdata	0-10 cm 10-30 cm 30-60 cm 60-100 cm	%	Bodem > Gemodelleerde bodemeigenschappen > Recente bodemdata > Organische koolstof

- 30-60 cm RMSPE
- 60-100 cm RMSPE

Voor de voorraden organische koolstof in de 0-30 en 0-100 cm bodemlaag zijn telkens het gemiddelde, 5% percentiel en 95% percentiel beschikbaar in de online DOV-verkenner met als extra in de download de RMSPE (fout op de predictie).

De gemodelleerde bodemeigenschappen kunnen nooit een staalname met bijhorende bodemanalyse op een bepaalde locatie vervangen, maar ze zijn wel een praktische tool bij vooronderzoek of als basisdata voor gebiedsdekkende modellering. Deze kaartlagen mogen dus niet op perceelsniveau geïnterpreteerd of gebruikt worden.

