

Optimalisatie van natuurbeheer met behulp van machine learning

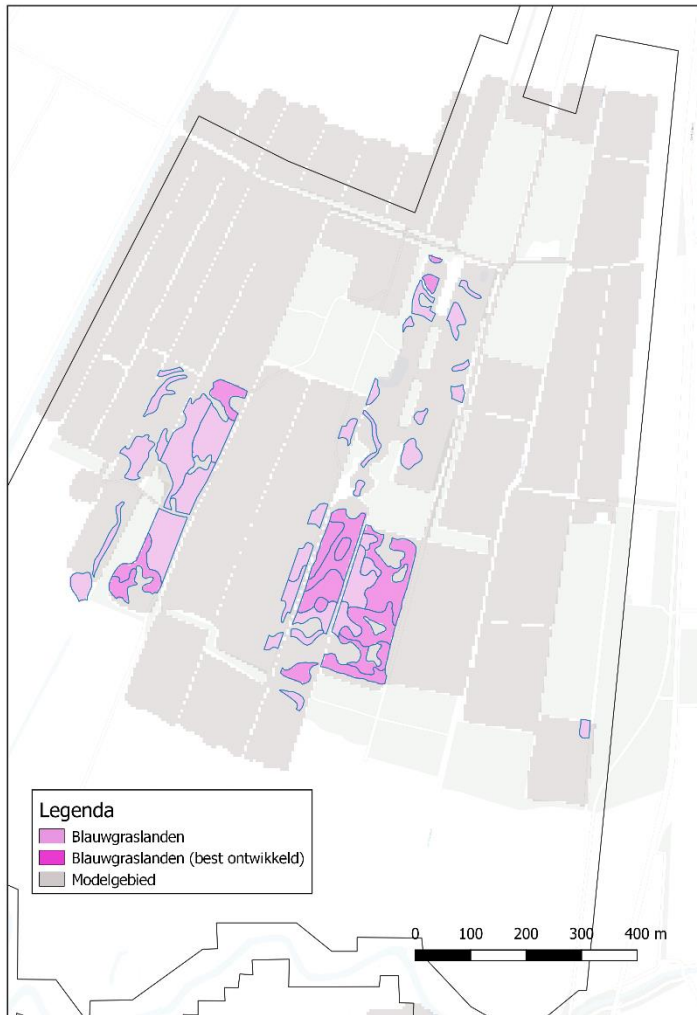
Kees van Immerzeel (Sweco), Willem Molenaar (Provincie Fryslân)

In dit artikel wordt verkend of de techniek van ‘machine learning’ bruikbaar is om het beheer in een natuurgebied te optimaliseren. Het gaat om het schraallandgebied Rome bij Drachten. Het geconstrueerde model is gevoed met abiotische gegevens die samenhangen met het voorkomen van blauwgrasland. Het model blijkt goed in staat te zijn het huidige voorkomen van blauwgrasland te verklaren. In een geselecteerd perceel blijkt een verhoging van de grondwaterstand bij te dragen aan de ontwikkeling van blauwgrasland. De techniek kan helpen om ook in andere natuurgebieden de locatie specifieke optimale groeiomstandigheden te vinden voor natuurlijke vegetaties.

In natuurgebieden wordt door het nemen van beheermaatregelen getracht de ontwikkeling van specifieke vegetaties (natuurdoeltypen) te stimuleren. Welke maatregelen het meest effectief zijn is daarbij op voorhand niet altijd duidelijk. Er zijn diverse factoren (ecologische randvoorwaarden) die samenhangen met het voorkomen van vegetaties. In grote lijnen zijn die factoren vaak wel bekend. Denk bijvoorbeeld aan de grondwaterstanden, bodemtype, voedingstoestand en zuurgraad (pH). Voor de meeste vegetatietypen zijn deze factoren tot op zekere hoogte ook gekwantificeerd met een bepaald bereik. Bijvoorbeeld: de gemiddelde voorjaarsgrondwaterstand (GVG) ligt tussen 10-25 centimeter onder het maaiveld (cm-mv).

In de praktijk blijkt dat er in elk natuurgebied dusdanig specifieke condities aanwezig zijn dat niet duidelijk is welke factor, of combinatie van factoren, het meest bepalend is voor de aanwezige vegetatie. Daardoor is het niet altijd duidelijk welke beheermaatregelen genomen moeten worden om de situatie te optimaliseren.

Dit is ook het geval in het natuurterrein Rome, een schraallandcomplex nabij Drachten en onderdeel van het Natura 2000-gebied Van Oordt's Mersken. In dit gebied is op 7,5 procent van het oppervlak het habitatype Blauwgraslanden aanwezig (zie afbeelding 1). Optimalisatie van het beheer is hier gewenst, aangezien dit habitatype onder druk staat: het oppervlak en de kwaliteit nemen af, terwijl er vanuit Natura 2000 een uitbreidingsdoel geldt.



Afbeelding 1. Vóórkomen van de blauwgraslanden in Rome (2016)

In dit artikel wordt verkend of het mogelijk is om met ‘machine learning’ de belangrijkste factoren te identificeren die samenhangen met het vóórkomen van het blauwgrasland om zo te komen tot een optimaal beheer. Het model dat daarbij door ‘machine learning’ ontstaat is vervolgens gebruikt om het effect van beheermaatregelen op het vóórkomen van blauwgrasland in een geselecteerd perceel te voorspellen.

Werkwijze

In het modelgebied (afbeelding 1) zijn met een resolutie van 5x5 meter de volgende gegevens (vlakdekkend) beschikbaar die mogelijk samenhangen met het voorkomen van blauwgrasland:

- Grondwaterstanden: GHG, GVG, GLG en de GT (respectievelijk de gemiddelde hoogste/voorjaar/laagste grondwaterstand en de hiervan afgeleide grondwatertrap; gezamenlijk worden deze afgekort als GxG’s).
- Bodemtype;
- Keileemdiepte.

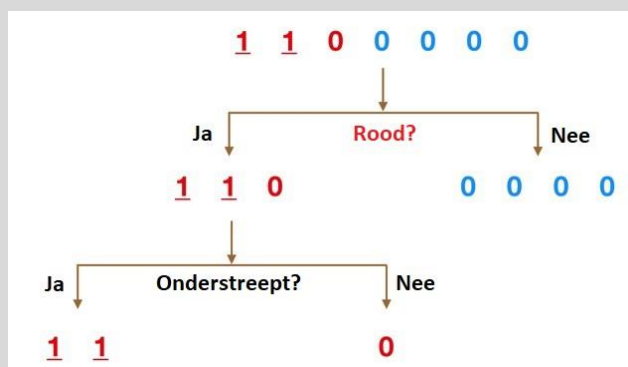
Het is bekend dat er meer parameters zijn die het vóórkomen van blauwgrasland bepalen. Maar alleen van de bovenstaande gegevens is vlakdekkende informatie voorhanden waarmee het model gevoed kan worden.

De helft van deze gegevens is gebruikt om het model te construeren; de andere helft van de waarnemingen (het testgedeelte van de gegevens) is gebruikt om het model te valideren.

Het model is geconstrueerd met de 'Random Forest'-techniek (RF). Dit is een techniek die kan worden ingezet bij classificatieproblemen. Bij zulke problemen wordt getracht om, op basis van meerdere kenmerken, te voorspellen tot welke klasse een waarneming behoort. In dit onderzoek gaat het om het voorspellen om van de klasse 0 (=geen blauwgrasland) of 1 (wel blauwgrasland). Er zijn veel publicaties over deze techniek beschikbaar [1]. Op deze plaats wordt volstaan met een korte toelichting op de RF-techniek (kader).

Beslisbomen

Een beslisboom vormt het hart van de RF-techniek. Hieronder is een voorbeeld opgenomen van een beslisboom.



Afbeelding 2. Beslisboom

Stel: de dataset bestaat uit nullen en enen, waarbij sommige cijfers rood zijn en andere blauw. Daarnaast zijn sommige cijfers onderstreept. Bij het construeren van de beslisboom is bij iedere vertakking de vraag: welk kenmerk verdeelt de waarnemingen zo, dat de resulterende subgroepen zo verschillend mogelijk van elkaar zijn? In het voorbeeld wordt als eerste geselecteerd op kleur, want daardoor zijn de resulterende subgroepen zo verschillend mogelijk. Al dan niet onderstreept zijn is het tweede kenmerk waarop wordt geselecteerd, etc.

In de RF-techniek wordt niet één maar een groot aantal beslisbomen samengesteld om te komen tot een classificatie. Bij de constructie van iedere boom wordt slechts een deel van de beschikbare parameters gebruikt. Welke parameters dat zijn wordt door het toeval bepaald. Iedere boom in 'het bos van bomen' resulteert in een 'eigen' voorspelling (klasse). Door de voorspellingen van alle bomen samen in ogenschouw te nemen ontstaat voor iedere klasse een verwachtingswaarde (kans). De beslisbomen die op deze manier worden geconstrueerd geven voorspellingen die relatief onafhankelijk van elkaar zijn. Door de voorspellingen van alle beslisbomen in zijn geheel te beschouwen ontstaat een robuuste voorspelling.

De constructie van het RF-model komt neer op het construeren van de beslisbomen waarmee voorspellingen kunnen worden gedaan. Dit wordt het ‘trainen’ van het model genoemd. Daarna is het model gevalideerd met het testgedeelte van de gegevens.

Na de validatie is dit model gebruikt om voor een geselecteerd perceel in het modelgebied het effect te onderzoeken van wijzigingen in de GxG's op het voorspelde vóórkomen van blauwgrasland. Zo kan duidelijk worden of het aanpassen van het waterhuishoudkundige beheer bij kan dragen aan het uitbreiden van blauwgrasland.

Resultaten modelvalidatie

Na de constructie van het model (het ‘trainen’) is de juiste werking ervan gecontroleerd door met het testgedeelte van de gegevens voorspellingen te doen over het voorkomen van blauwgrasland. Tabel 1 geeft het resultaat van de modelvalidatie weer.

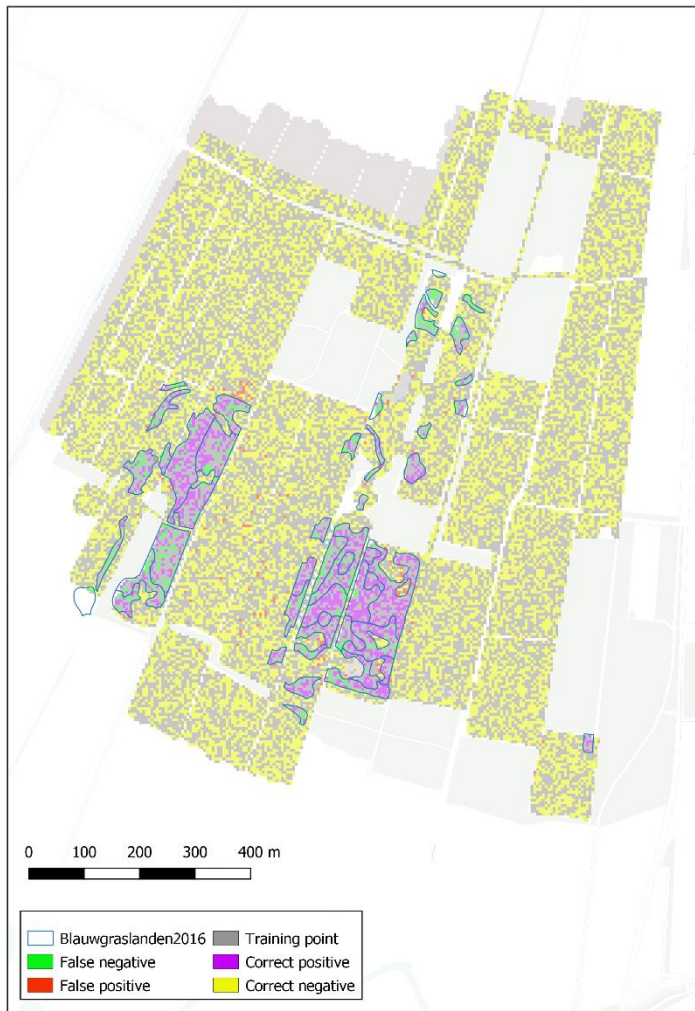
Tabel 1. Resultaat van de modelvalidatie. 1=blauwgrasland; 0=geen blauwgrasland

Voorspelling	Referentie	
	0	1
0	92%	2%
1	1%	5%
Totaal	93%	7%

Op 7 procent van de 24.768 testpunten komt blauwgrasland voor (gemeten). Op 5 procent van de testpunten voorspelt het model blauwgrasland waar het er inderdaad is; in 1 procent van de gevallen voorspelt het model blauwgrasland waar het er in werkelijkheid niet is. In 93 procent van de testpunten waar geen blauwgrasland voorkomt, voorspelt het model dit ook.

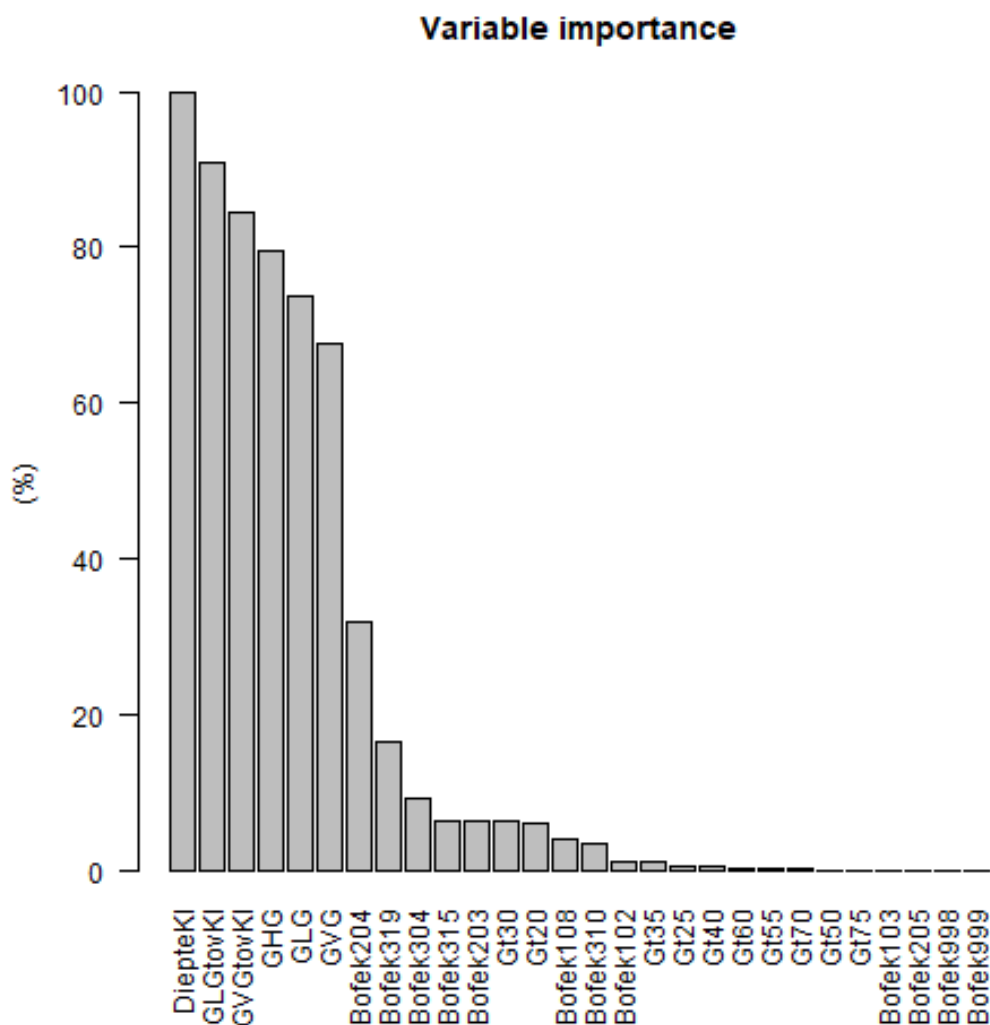
Een toets voor de kwaliteit van het model is de ‘kappa statistic’. De kappawaarde van het model is 0.77, wat geldt al een goede fit [2].

Afbeelding 3 geeft het kaartbeeld van de voorspellingen die zijn gedaan met de testdataset. Juiste voorspellingen zijn vooral gedaan waar de keileem relatief ondiep aanwezig is. Dat is het geval bij de zuidoostelijke percelen met blauwgrasland.



Afbeelding 3. Voorspelling met de testdataset. 'False negative'= geen blauwgrasland voorspeld waar het er wel is; 'False positive'= blauwgrasland voorspeld waar het er niet is; 'Correct positive'= blauwgrasland voorspeld waar het er inderdaad is; 'Correct negative'= geen blauwgrasland voorspeld waar het er inderdaad niet is

Afbeelding 4 toont de 'variable importance' van de verschillende parameters in het model. Deze waarde 'variable importance' geeft het relatieve belang aan van een parameter bij het structureren van de gegevens [3]. Concreet in dit geval: wat is het relatieve belang van een modelparameter bij de voorspelling van het vóórkomen van blauwgrasland?

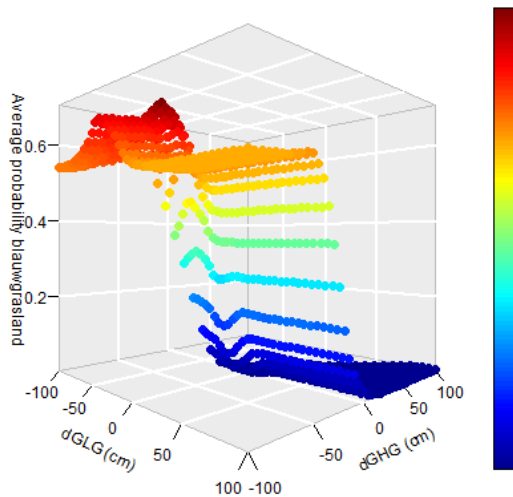


Afbeelding 4. Het relatieve belang van verschillende parameters in het model. GVG/GLG/GHG= resp. gemiddelde voorjaars- laagste en hoogste grondwaterstand (cm-mv); Gt=Grondwatertrap; GVGtovKI / GLGtovKI = resp. gemiddelde voorjaars- en laagste grondwaterstand t.o.v. het keileem niveau (cm); Bofek=code bodemfysische eenheid; DiepteKI=diepte keileem (cm-mv)

Uit afbeelding 3 blijkt dat de parameters die samenhangen met de diepte van de keileem (DiepteKI, GLGtovKI, GVGtovKI) relatief belangrijk zijn in het model om een juiste voorspelling te kunnen doen.

Resultaten beheeroptimalisatie

Bij het beheer van blauwgraslanden vormt het waterbeheer een belangrijk stuurmechanisme waarmee de grondwaterstanden kunnen worden beïnvloed. Daarom is er met het model nagegaan wat het effect is van het aanpassen van de grondwaterstanden. Hiertoe is er een perceel geselecteerd waarvan het maaiveld iets hoger ligt dan de percelen waar nu blauwgrasland voorkomt. Voor dit perceel zijn alternatieve datasets gemaakt met verschillende GLG/GHG-combinaties. Om de optimale GxG-combinatie te vinden is bij iedere combinatie de gemiddelde kans op blauwgrasland berekend. Afbeelding 5 toont de relatie tussen de wijzigingen in de GHG/GLG en de berekende gemiddelde kans op blauwgrasland in het perceel.

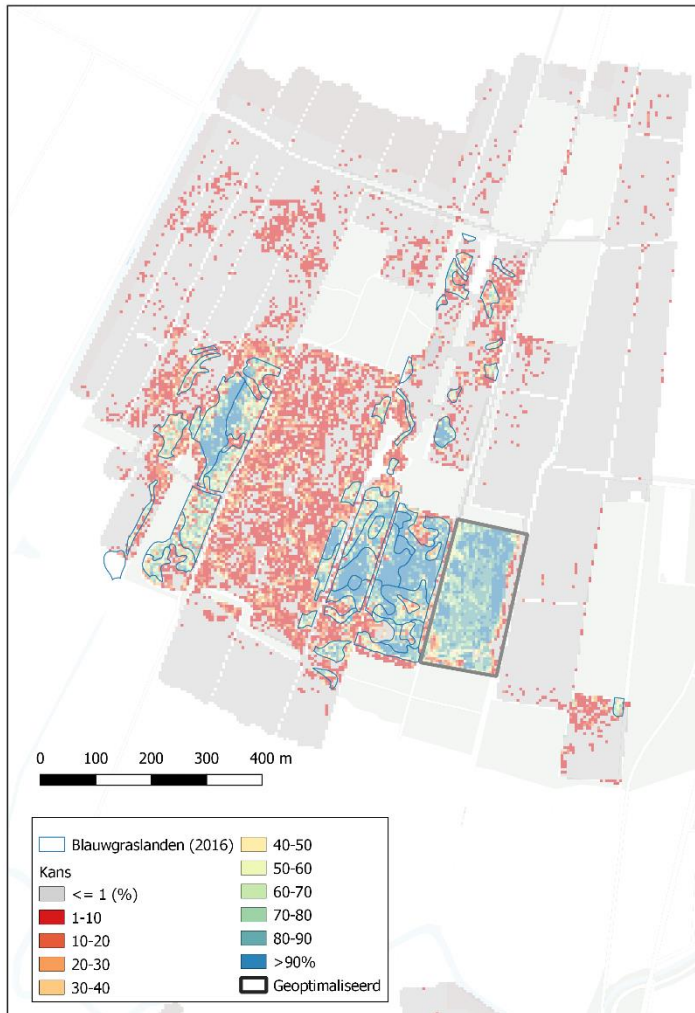


Afbeelding 5. Berekende gemiddelde kans op blauwgrasland in een perceel bij een wijziging van de GxG's. Een negatieve verandering van de GxG betekent een verhoging van de grondwaterstand (natter)

Uit afbeelding 5 blijkt dat een 'drogere' GHG in combinatie met een 'drogere' GLG niet leidt tot een verhoging van de kans op blauwgrasland in dit perceel. Bij een 'nattere' GHG en een 'drogere' GLG is dit wel het geval. De grootste kans op blauwgrasland wordt voorspeld bij een combinatie van een 'nattere' GHG en een 'nattere' GLG': respectievelijk ca. -2 cm-mv en ca. 53 cm-mv. De voorspelde kans op blauwgrasland is dan 70%, wat betekent dat het model blauwgrasland voorspelt (want de kans erop is groter dan 50%). Afbeelding 5 toont de ruimtelijke verdeling van de berekende kans op blauwgrasland in het geoptimaliseerde perceel.

Uit de resultaten van de testberekening (afbeelding 3) blijkt dat bij ongewijzigde GxG's in het geselecteerde perceel geen blauwgrasland wordt voorspeld. Dit resultaat is begrijpelijk: door de iets hogere ligging van het perceel zonder blauwgrasland zijn de GxG's er 'droger' dan in het naastgelegen perceel waar in de huidige situatie wel blauwgrasland aanwezig is.

Met het model kan dus voor een perceel worden bepaald welke verandering in grondwaterstand nodig is om de situatie voor blauwgrasland te optimaliseren. Omdat het model is getraind met informatie van dit gebied, is het gevonden optimum gebaseerd op de specifieke omstandigheden in dat gebied.



Afbeelding 6. Berekende kans op blauwgrasland. In het geselecteerde perceel zijn de GxG's geoptimaliseerd

Conclusie

Machine learning is in het schraallandgebied Rome bruikbaar gebleken om in een perceel de optimale grondwaterstanden (GxG's) te vinden voor de ontwikkeling van blauwgrasland. Dat biedt perspectief om de techniek ook in andere natuurgebieden in te zetten voor een locatiespecifieke optimalisatie van het beheer.

Evaluatie

Een voorwaarde om succesvol een vegetatievoorspellingsmodel te maken met machine learning is de beschikbaarheid van voldoende gegevens. Zonder informatie over de keileemdiepte zou bijvoorbeeld de voorspelling in het natuurgebied Rome beduidend slechter zijn geweest.

De verwachting is dat het model dat is getraind voor het natuurgebied Rome alleen voorspellende waarde heeft in andere gebieden als deze gebieden een vergelijkbare bodemopbouw, ligging en waterhuishouding hebben.

Het model geeft geen uitsluitel over de processen die mogelijk samenhangen met het vóórkomen van blauwgrasland. Het feit dat de keileemdiepte in dit gebied een belangrijke factor is om te komen tot

een goed verklarend model kan wel aanleiding zijn om te kijken welk proces samenhangt met de keileemdpte.

In dit geval heeft bijvoorbeeld het ondiep vóórkomen van keileem vermoedelijk te maken met een positief effect op de vochtvoorziening, zuurgraad en basenverzadiging door capillaire opstijging van baserijk grondwater in het zomerseizoen. Nader onderzoek zou kunnen uitwijzen of dat inderdaad het geval is.

De resultaten van dit onderzoek kunnen worden gereproduceerd met de R-scripts op GitHub (<https://github.com/KVIsweco/AnalyseBlauwgraslandRome>).

Referenties

1. https://en.wikipedia.org/wiki/Random_forest. Geraadpleegd op 24 april 2020
2. Landis, J.R.; Koch, G.G. (1977). *The measurement of observer agreement for categorical data*. *Biometrics*. 33 (1): 159–174.
3. <http://hdl.handle.net/2268/170309>, geraadpleegd op 24 april 2020 .