

# Regressie- en classificatietechnieken in hydro-ecologische modellen: toepassing voor vallei-ecosystemen in Vlaanderen

<sup>1</sup> Universiteit Gent  
<sup>2</sup> Instituut voor Natuur- en  
Bosonderzoek  
<sup>3</sup> Universiteit Antwerpen

Hydro-ecologische distributiemodellen voorspellen het voorkomen van soorten en vegetatietypes uitgaande van de hydrologische en hydrogeochemische standplaatscondities. Dergelijke modellen kennen toepassingsmogelijkheden in het natuurbeleid- en natuurbeheer, doordat ze in staat zijn veranderingen in soorten- en vegetatiedistributies als gevolg van veranderende standplaatscondities te modelleren. In deze studie worden twee hydro-ecologische distributiemodellen toegelicht: (i) het logistische regressiemodel en (ii) het random forest model. Op basis van een toepassing van deze modellen in Vlaamse vallei-ecosystemen met voornamelijk grondwatergebonden vegetaties kan besloten worden dat beide modellen in staat zijn om een redelijk accurate, gebiedsdekkende verspreidingkaart te genereren van de voorkomende vegetatietypes.

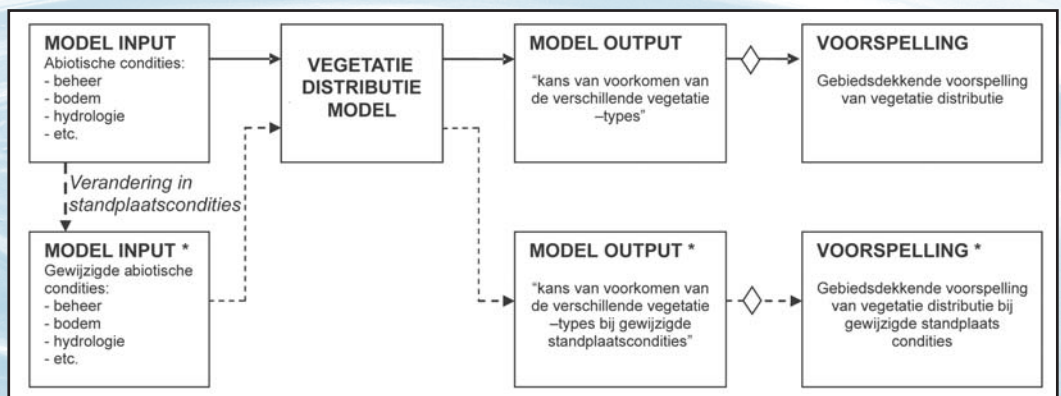
## Inleiding

Sociale, industriële, technologische en landbouwkundige ontwikkelingen bepalen onze omgang met het landschap in grote mate. Ze hebben een degradatie van het milieu en de aanwezige natuur tot gevolg gehad. Deze evolutie gaat nog steeds door: processen als verzuring, vermeting, versnippering en verdroging hebben een negatieve invloed op natuurwaarden. De laatste jaren is de aandacht voor deze processen toegenomen, zowel op het beleidsniveau als op het vlak van dagelijks natuurbeheer. Door gerichte maatregelen tracht men de achteruitgang van het natuurlijke milieu en het verdwijnen van organismen en levensgemeenschappen tegen te gaan en de natuurwaarde opnieuw te verhogen. Zowel bij degradatie als bij restauratie van natuur is een inschatting van de effecten van de veranderingen in de milieuomstandigheden op het ecosysteem een noodzaak. Het laat toe de gevolgen van beleid- en beheersbeslissingen te voorzien. Deze studie focust op een deel van het ecosysteem, nl. de vegetatie. Vegetatieontwikkeling wordt door verschillende factoren bepaald: klimaat, geologie, hydrologie, beheer, verspreiding- en reproductiestrategie van planten. Kennis van de relatie tussen vegetatiestructuur en vegetatiesamenstelling enerzijds en standplaats-

condities anderzijds, is noodzakelijk om voorspellingen te kunnen doen over mogelijke vegetatieontwikkelingen na herstelmaatregelen. De integratie van deze kennis in een hydro-ecologisch model levert een instrument dat efficiënt kan worden ingezet en een waaier van toepassingsmogelijkheden heeft in het natuurbeleid en het natuurbeheer.

Meestal zijn deze hydro-ecologische modellen empirisch van aard, waarbij veldobservaties gerelateerd worden aan standplaatscondities op basis van statistische of theoretische technieken. In deze studie worden twee van deze technieken besproken: (i) de logistische regressie techniek en (ii) de vrij recent ontwikkelde random forest techniek. Vegetatie distributiemodellen die gebruik maken van deze technieken zijn gelijkaardig in modelopbouw. Op basis van de modelinput, welke gebiedsdekkende gegevens bevat over de standplaatscondities wordt door het model een kans op voorkomen van verschillende vegetatietypes berekend. Aan de hand van een eenvoudige beslissingsregel komt men tot een gebiedsdekkende voorspelling van de vegetatiedistributie. Wanneer een verandering in standplaatscondities optreedt, bijvoorbeeld door verdroging, wordt hetzelfde model gebruikt om een gebiedsdekkende vegetatiedistributie te voorspellen bij gewijzigde standplaatscondities (Figuur 1).

Figuur 1: Conceptuele voorstelling van het hydro-ecologisch distributiemodel. Op basis van gebiedsdekkende gegevens aangaande standplaatscondities, worden de kansen van voorkomen van verschillende vegetatietypes berekend. Deze worden tegen elkaar afgewogen d.m.v. de beslissingsregel "het vegetatietype met de hoogste kans van voorkomen is het voorspelde vegetatietype", geformuleerd in  $\diamond$ . Na wijziging van de standplaatscondities, wordt op basis van de nieuwe model input \* de kans van voorkomen van de verschillende vegetatietypes berekend (model output \*), welke na toepassing van beslissingsregel  $\diamond$  leiden tot nieuwe voorspellingen \*.



## Twee hydro-ecologische distributiemodellen

In deze studie worden twee verschillende technieken behandeld die allebei het voorkomen van vegetatietypes in relatie tot standplaatscondities voorspellen, en dus allebei kunnen geïntegreerd worden in een hydro-ecologisch distributie model.

### Logistische regressie

De logistische regressie techniek (Hosmer and Lemeshow, 2000) beschrijft de relatie tussen onafhankelijke variabelen (in deze studie de standplaatscondities) en een binaire respons variabele (in deze studie de aan- of afwezigheid van een vegetatietype) aan de hand van een mathematische functie. Op elke locatie kan een verzameling van  $i$  onafhankelijke variabelen worden gemeten die kan worden voorgesteld als de vector  $\mathbf{x}=(x_1, x_2, \dots, x_i)$ . De conditionele kans dat een vegetatietype voorkomt bij de gegeven standplaatscondities  $\mathbf{x}$  wordt voorgesteld als  $P(\mathbf{x})$  en gemodelleerd als:

$$P(\mathbf{x}) = \exp(g(\mathbf{x})) / (1 + \exp(g(\mathbf{x}))) \quad (1)$$

met  $P(\mathbf{x})$  de kans op aanwezigheid, gegeven de standplaatscondities  $\mathbf{x}$ . De functie  $g(\mathbf{x})$  wordt berekend als een lineaire combinatie van onafhankelijke variabelen, waarbij  $\beta_i$  de regressieparameters zijn:

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_i x_i \quad (2)$$

Een schatter voor  $g(\mathbf{x})$  moet berekend worden voor elk vegetatietype afzonderlijk, waarna de kans op voorkomen van dat vegetatietype berekend wordt aan de hand van vgl. (1). Vervolgens worden de berekende kansen van voorkomen voor de verschillende vegetatietypes tegen elkaar afgewogen, en is het vegetatietype met de hoogste kans van voorkomen het voorspelde vegetatietype.

## Random forests

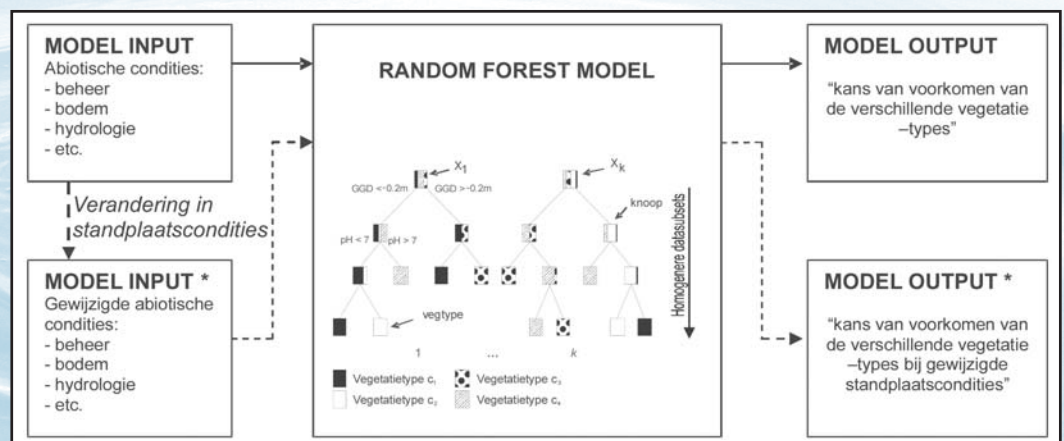
De random forest techniek (Breiman, 2001) is een classificatietechniek waarbij een groot aantal classificatiebomen worden gegenereerd, welke dan samen gevoegd worden om de finale classificatie te berekenen. Elke classificatieboom wordt gegenereerd op basis van een deel ( $X_i$ ) van de originele dataset ( $X$ ). De beslissingsregels die de knopen splitsen maken gebruik van een subset van  $m$  onafhankelijke variabelen (hier dus standplaatscondities), waaruit de beste wordt geselecteerd. Onderstaande pseudocode beschrijft het algoritme om een random forest te genereren bestaande uit  $k$  classificatiebomen (zie ook Figuur 2):

- (i) voor  $i = 1$  tot  $k$  doe:
  - a. neem een deel  $X_i$  bestaande uit  $2/3$  van de elementen van de originele dataset  $X$  (de referentiegegevens);
  - b. gebruik  $X_i$  om een classificatieboom te genereren, waarbij de knopen gesplitst worden op basis van de beste splitsingsvariabele, gekozen uit  $m$  willekeurig gekozen onafhankelijke variabelen;
- (ii) In toepassing voor nieuwe data: bereken de kans van voorkomen door de nieuwe data te laten classificeren door alle  $k$  classificatiebomen uit het random forest. De kans van voorkomen van vegetatietype  $c_i$  wordt gegeven door  $P(c_i) = N_{c_i} / N_{\text{tot}}$  waarbij  $P(c_i)$  de kans van voorkomen van vegetatietype  $c_i$  is,  $N_{c_i}$  het aantal classificatiebomen dat het vegetatietype  $c_i$  classificeert, en  $N_{\text{tot}} (=k)$  het totale aantal classificatiebomen in het random forest.
- (iii) De beslissingsregel "het vegetatietype met de hoogste kans van voorkomen is het voorspelde vegetatietype" wordt gehanteerd om tot de finale vegetatiedistributievoorspelling te komen.

### Toepassing in Vlaanderen

Vallei-ecosystemen vervullen een aantal belangrijke socio-economische en ecologische functies, waarvan het aanvullen van het grondwater, de reductie van sediment transport, overstromings-

Figuur 2: Conceptuele voorstelling van het random forest model, bestaande uit  $k$  classificatiebomen. Elke boom wordt gebouwd op basis van een andere random subdataset  $X_i$ . Deze dataset wordt op de knopen gesplitst in steeds homogener datasets op basis van de standplaatscondities. In de eindknopen komen alle elementen samen met één bepaald vegetatietype.



controle en de bewaring van habitat en biodiversiteit slechts enkele voorbeelden zijn. Niettegenstaande werden in West Europa doorheen de jaren de meeste vallei-ecosystemen drooggelegd en onder cultuur gebracht. Het geïntegreerde waterbeleid van de laatste decennia benadrukt het belang van grondwater gebonden vallei-ecosystemen, waardoor deze prioritair geworden zijn in natuur herstel- en bescherming.

### Studiegebieden en dataset

Het Instituut voor Natuur- en Bosonderzoek heeft met verschillende onderzoek- en monitoringsprojecten gedurende de periode 1990 - 1997 gebiedsdekkende informatie verzameld over een aantal vallei-ecosystemen, waaronder de Vallei van de Zwarte Beek, Vorsdonkbos, Doode Bemde en Snoekengracht. De gebieden werden opgesplitst door een regelmatig raster met rastercellen van 20 m x 20 m (10 m x 10 m in Snoekengracht). Voor elk van die rastercellen werden 14 standplaatsvariabelen opgemeten (zij het door gebiedsdekkende monitoring, zij het door interpolatie van puntmetingen). Abiotische variabelen die hierbij opgemeten of berekend werden zijn: bodemtype, beheer, grondwaterstand en grondwaterkwaliteit (pH, K<sup>+</sup>, Fe<sub>(tot)</sub>, Mg<sup>2+</sup>, Ca<sup>2+</sup>,

SO<sub>4</sub><sup>2+</sup>, Cl<sup>-</sup>, NO<sub>3</sub><sup>-</sup>-N, NH<sub>4</sub><sup>+</sup>-N, H<sub>2</sub>PO<sub>4</sub><sup>-</sup> en de ionenratio IR=100[1/2Ca<sup>2+</sup>]/[1/2Ca<sup>2+</sup> + Cl]). Dezelfde rasters werden gebruikt om de vegetatie van de gebieden te inventariseren. Het voorkomen en de bedekkingsgraad van 85, voornamelijk grondwater gebonden plantensoorten werd geschat gebruik makende van de decimale schaal van Londo. Vervolgens werden deze gegevens geclusterd, gebruik makende van TWINSPAN, tot 11 duidelijk gedefinieerde vegetatietypes (Tabel 1).

Er werd een hydro-ecologische (L) dataset samengesteld waarbij voor elk van de in totaal 1705 rastercellen de standplaatscondities beschreven worden d.m.v. een vector  $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{i14})$  waarin de gemeten waarden voor de 14 onafhankelijke variabelen vervat zitten, en waaraan het overeenkomstige geobserveerde vegetatietype gekoppeld werd (Huybrechts et al., 2002).

### Resultaten en evaluatie

Wegens een gebrek aan een onafhankelijke dataset werd de dataset L willekeurig gesplitst in twee delen voor tweevoudige kruisvalidatie. Zodoende werd elk element van L één maal gebruikt tijdens de model constructie, en één maal tijdens de model evaluatie. Resultaten zijn weergegeven in

Tabel 1: Samenvatting van de vegetatietypes: naam, korte beschrijving en oppervlakte

Naam	Beschrijving	Oppervlakte [ha] (aantal rastercellen)			
		ZB 6.80 (170)	VB 12.80 (320)	DB 20.76 (519)	SG 6.69 (696)
<i>Alno-Padion</i> (Elzen- vogelkersbos)	Bossen van de drogere standplaats met <i>Quercus robur</i> L., <i>Fraxinus excelsior</i> L., <i>Carpinus betulus</i> L. en <i>Alnus glutinosa</i> (L.) Gaertn.				1.47 (147)
<i>Arrhenatherion elatioris</i> (Glanshaver grasland)	Grasland met <i>Arrhenatherum elatius</i> (L.) J.&C. Presl., <i>Anthriscus sylvestris</i> (L.) Hoffm. and <i>Leucanthemum vulgare</i> Lamk.			2.80 (70)	0.91 (91)
<i>Calthion palustris</i> (Dottergrasland)	Soortenrijk mesotroof grasland gedomineerd door <i>Caltha palustris</i> L. en vele <i>Carex</i> -soorten			4.24 (106)	0.95 (95)
<i>Carici elongatae</i> – <i>Alnetum glutinosae</i> (Mesotroof elzenbroek)	Mesotroof elsenbroek met <i>Alnus glutinosa</i> (L.) Gaertn. en <i>Carex acutiformis</i> Ehrh., <i>Lycopus europaeus</i> L. en <i>Solanum dulcamara</i> L. in de kruidlaag		3.16 (79)	1.20 (30)	1.41 (141)
<i>Caricion curto-nigrae</i> (Kleine zeggevegetatie)	Lage vegetatie met kleine <i>Cyperaceën</i> , zoals <i>Carex panicea</i> L. en <i>Carex rostrata</i> Stokes	6.80 (170)	1.12 (28)		
<i>Cirsio-Molinietum</i> (Blauwgrasland)	Vergelijkbaar met <i>Caricion curto-nigrae</i> , maar meer <i>Poaceae</i> en een hogere productiviteit		1.12 (28)		
<i>Filipendulion</i> (Moerassperea ruigte)	Ruigten van beekdalen met <i>Filipendula ulmaria</i> (L.) Maxim., <i>Valeriana officinalis</i> L. en <i>Alopecurus pratensis</i> L.		4.76 (119)	4.16 (104)	1.12 (112)
<i>Magnocaricion</i> (Grote Zeggevegetatie)	Zegge moeras met verschillende hoog opgaande <i>Carex</i> soorten			2.52 (63)	
<i>Magnocaricion met Phragmites</i> (Rietruigte)	<i>Magnocaricion</i> vegetatie met <i>Phragmites australis</i> (Cav.) Steud.			3.72 (93)	0.83 (83)
<i>Phragmitetalia</i> (Rietland)	Hoogproductief rietland, gedomineerd door <i>Phragmites australis</i> (Cav.) Steud.			2.12 (53)	0.27 (27)
<i>Sphagno-Alnetum</i> (Elzen-berkenbroek)	Oligotroof broek met <i>Betula pubescens</i> Ehrh. en <i>Alnus glutinosa</i> (L.) Gaertn., met een dense moslaag van <i>Sphagnum palustre</i> L. en <i>Sphagnum fimbriatum</i> Wilson.		2.64 (66)		

ZB, Zwarte Beek; VB, Vorsdonkbos; DB, Doode Bemde; SG, Snoekengracht.

Figuur 3: Voorspellingen voor de vier studiegebieden op basis van het logische regressiemodel (a, LR model) en het random forest model (b, RF model). De voorspelde vegetatie distributie (○) is afgebeeld bovenop de geobserveerde vegetatie distributie (□).



Figuur 3. Van de 1705 rastercellen werd er voor 1128 (69.3%) een juiste, en voor 524 (30.7%) een foute vegetatievoorspelling gemaakt door het logistische regressie model. Inspectie van Figuur 3 leert ons dat (i) correcte voorspellingen gemaakt werden voor gebieden met geringe diversiteit in vegetatie (Zwarte Beek); (ii) voor meer diverse

gebieden regelmatig foute voorspellingen gemaakt werden en dat (iii) binnen deze gebieden de voorspellingen beter waren voor relatief grote, homogene vegetatieclusters (bv. noordelijk deel van Vorsdonkbos). Het random forest model maakte 1307 (76.7%) juiste en 398 (23.3%) foute voorspellingen. Correcte voorspellingen situeren

zich in het centrum van de vegetatieclusters. Voor rastercellen op de grens tussen twee vegetatietypes en voor geïsoleerde rastercellen zijn de voorspellingen minder accuraat.

We kunnen concluderen dat beide hydro-ecologische modellen in staat zijn om op basis van abiotische standplaatscondities redelijk accuraat de zones aan te geven waar de potenties voor de ontwikkeling van bepaalde vegetatietypes hoog zijn. De afbakening van deze zones op niveau van rastercellen (bv. 20 m x 20 m) is niet perfect. Op de grens van twee vegetatietypes komen overgangszones in de berekeningen naar voor, wat trouwens overeen komt met de situatie op het terrein. Het random forest model presteert iets beter dan de logistische regressie. Beide distributiemodellen kunnen ingezet worden voor beleid- en beheersondersteuning.

### Distributiemodellen in een ruimere context

Ondanks de goede modelprestaties en de ruime toepassingsmogelijkheden, dienen de voorgestelde hydro-ecologische modellen ook in een ruimere ecologische context te worden geplaatst. Beide modellen relateren het voorkomen van vegetatietypes aan de heersende standplaatscondities. Welke standplaatsvariabelen daarbij opgenomen worden is vaak een praktische overweging, maar het is duidelijk dat niet alle standplaatsvariabelen even determinerend zijn voor de vegetatiedistributie, ze zijn niet allemaal causaal verantwoordelijk voor de vegetatierespons. Als gevolg zijn deze empirische modellen moeilijk extrapoleerbaar (in tijd en ruimte), en in principe enkel toepasbaar in de gebieden waarvoor ze ontwikkeld werden. Het gebruik van causale standplaatsvariabelen met een directe impact op de vegetatie zou een eerste stap naar een meer mechanistische, ruimer toepasbare modelbenadering. Maar zelfs al zouden enkel causale variabelen opgenomen worden in deze modellen, dan nog zouden de voorspellingen niet volledig overeenkomen met de observaties daar ecologische processen als competitie, predatie en verbreiding eveneens een invloed hebben op vegetatiedistributies, maar moeilijk te vertalen zijn in kwantitatieve variabelen. Daarom drukken modelresultaten van voorgestelde hydro-ecolo-

gische modellen eerder een habitatgeschiktheid voor de verschillende vegetatietypes uit.

### Dankwoord

Wij wensen het Bijzonder Onderzoeksfonds (BOF, project nummer 011/015/04) van de Universiteit Gent, en het Fonds voor Wetenschappelijk Onderzoek (krediet aan navorsers 1.5.108.03) te bedanken.

### Bibliografie

BREIMAN, L. (2001), Random forests. *Mach. Learn.* 45, 5-32.

HOSMER, D.W. and LEMESHOW S. (2000), *Applied Logistic Regression*. Second edition. Wiley, Chichester.

HUYBRECHTS, W., DE BECKER, P., DE BIE, E., WASSEN E. and BIO, A. (2002), Ontwikkeling van een Hydro-ecologisch model voor valleiecosystemen in Vlaanderen, ITORS-VL. VLINA 00/16. Instituut voor Natuur- en Bosonderzoek, Brussel, België.

PETERS, J., DE BAETS, B., VERHOEST N.E.C., SAMSON R., DEGROEVE, S., DE BECKER, P. and HUYBRECHTS, W. (2007), Random forests as a tool for ecohydrological distribution modelling. *Ecological Modelling*. (In press).

*Ir. Jan Peters*  
*Laboratorium voor Hydrologie en Waterbeheer*  
*Universiteit Gent*  
*Coupure Links 653*  
*9000 Gent*  
*e-mail: jan.peters@ugent.be*  
*Tel. 09 264.61.40*  
*Fax. 09 264.62.36*

*N.E.C. Verhoest<sup>1</sup>, B. De Baets<sup>1</sup>,*  
*P. De Becker<sup>2</sup>, W. Huybrechts<sup>2</sup> en*  
*R. Samson<sup>3</sup>*

<sup>1</sup> *Universiteit Gent*

<sup>2</sup> *Instituut voor Natuur-en Bosonderzoek*

<sup>3</sup> *Universiteit Antwerpen*