Vereniging van Rijnwaterbedrijven

Imputeren van ontbrekende waarden in RIWA-base

Procedure voor imputeren en beoordelen van waarden in RIWA-base

Aart Smits (eauQstat)
Eit van der Meulen (AMO)
Gerrit van de Haar (RIWA)
Paul Baggelaar (Icastat, eindredactie)

Vereniging van Rijnwaterbedrijven - RIWA • Amsterdam, januari 2014

www.riwa.org

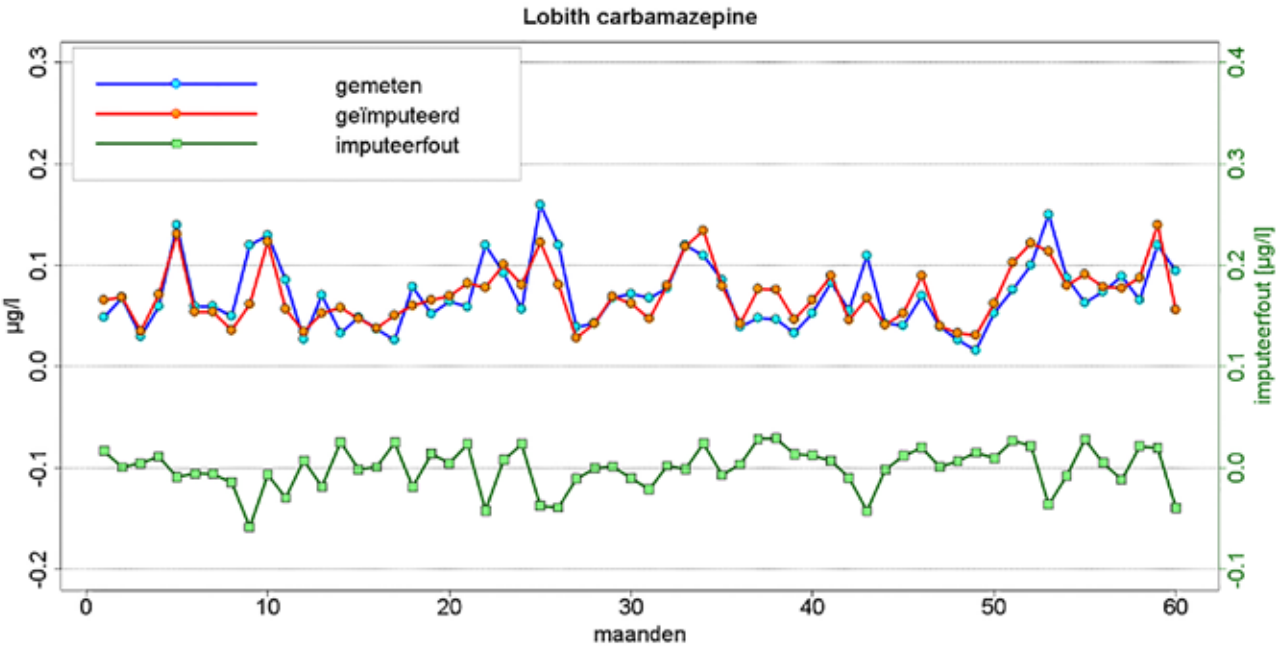
Samenvatting

RIWA, de Vereniging van Rivierwaterleidingbedrijven in Nederland en België, streeft ernaar om haar doelstellingen te bereiken met degelijk onderbouwde informatie en steekt dan ook veel energie in het monitoren van de oppervlaktewaterkwaliteit. Maar het komt regelmatig voor dat er gaten vallen in meetreeksen, wat het beeld vertroebelt. RIWA heeft daarom behoefte aan een geschikte methode om gaten in tijdreeksen op te vullen, door het imputeren (schatten) van ontbrekende waarden.

Dit rapport belicht een studie die is uitgevoerd om tot de meest geschikte imputeermethode te komen. Daarvoor zijn eerst verschillende mogelijk geschikte imputeermethoden geselecteerd. Vervolgens zijn met Monte Carlosimulaties de imputeerprestaties van drie methoden vergeleken, namelijk Random Forest, meervoudige lineaire regressie en het neurale netwerk, die alle drie gebruik maken van relaties met andere tijdreeksen. Random Forest is op te vatten als een soort niet-lineaire regressiemethode en is net als het neurale netwerk afkomstig uit het onderzoeksveld Machinaal Leren.

De imputeerprestaties zijn vergeleken bij toepassing op kunstmatige tijdreeksen en bij toepassing op enkele honderden praktijkreeksen uit de database van RIWA (RIWA-base). Daarbij bleek dat Random Forest over het algemeen beter presteerde dan meervoudige lineaire regressie en ook robuuster was. Het neurale netwerk scoorde doorgaans het slechtst (zij het met de kanttekening dat een pragmatische keuze is gedaan voor een breed toepasbare structuur van het neuraal netwerk, zodat mogelijk niet al zijn capaciteiten zijn benut). Random Forest bleek overigens niet alleen geschikt om gaten in tijdreeksen op te vullen, maar bleek ook perspectieven te bieden om de waarden van de reeks te controleren.
Bij toepassing op praktijkreeksen bleken de imputeerprestaties verschillend per parametergroep. Zo bleken de klassieke parameters over het algemeen betrouwbaarder te imputeren dan biologische en hydrobiologische parameters en een aantal organische microparameters. Dit komt vermoedelijk doordat er bij de klassieke parameters minder uitschieters voorkomen. Het kan ook komen doordat er bij die parameters over het algemeen minder meetfouten optreden.

Onderstaand een voorbeeld van de imputeerprestaties van Random Forest. Van een vijfjaarsreeks van maandwaarden van carbamazepine te Lobith is stapsgewijs elke waarde verwijderd en vervolgens geïmputeerd. De blauwe lijn toont de oorspronkelijke reeks en de rode lijn de geïmputeerde reeks. De groene lijn toont de verschilreeks (deze is geschaald op de rechteras).



Deze reeks blijkt goed te imputeren. De imputeerfouten blijven beperkt tot minder dan 0,05 µg/l en de imputanten benaderen de metingen goed.

Dit rapport presenteert tenslotte een procedure voor toepassing van Random Forest op RIWA-base. De procedure is ontwikkeld om ontbrekende waarden te imputeren van vijfjarige meetreeksen op maandbasis en kan ook dienen om de maandwaarden van die reeksen te beoordelen.

Inhoudsopgave

1

Inleiding

Dit rapport beschrijft de ontwikkeling van een procedure voor het opvullen van gaten in meetreeksen met geschatte waarden - aangeduid als imputeren - en voor het beoordelen van meetwaarden. De ontwikkelde procedure is bedoeld voor toepassing op RIWA-base, de database van RIWA.

1.1 Probleemstelling

RIWA, de Vereniging van Rivierwaterleidingbedrijven in Nederland en België, streeft ernaar om haar doelstellingen te bereiken met degelijk onderbouwde informatie. Eén van de belangrijkste bronnen voor deze informatie is het RIWA-meetnet, dat gegevens verzamelt van de oppervlaktewaterkwaliteit van verschillende locaties in de stroomgebieden van Rijn en Maas. De meest relevante waterkwaliteitsparameters worden minstens vierwekelijks gemeten (basismeetprogramma), om een goed beeld te kunnen krijgen van het verloop gedurende het jaar. Verder stelt dit ook in staat om verantwoorde uitspraken te doen over structurele veranderingen over langere perioden.

De meetgegevens van het meetnet worden opgeslagen in een database (RIWA-base), zodat regelmatig bruikbare statistische informatie kan worden afgeleid en gepresenteerd over toestand en trends van de waterkwaliteit. Zo publiceert de RIWA al vanaf eind jaren zestig de maandgemiddelden van alle gemeten oppervlaktewaterkwaliteitsparameters in jaarrapporten. Aanvullend worden inmiddels per parameter ook kengetallen over het betreffende meetjaar vermeld, zoals aantal meetwaarden, minimum, maximum en enkele percentielen. En sinds enkele jaren wordt voor elke parameter het resultaat van statistische trendanalyse van de deelreeks van de laatste vijf jaar weergegeven in een pictogram (het RIWA-pict), waarbij de kleur aangeeft hoe de toestand in het recentste meetjaar zich verhoudt tot de IAWR-streefwaarde uit het Donau-, Maas- en Rijnmemorandum van 2008. Maar als een meetreeks niet minstens over 10 maanden van een beschouwd jaar meetwaarden bevat, dan wordt zijn informatie inhoud te gering geacht om in aanmerking te komen voor de standaardbeoordeling door de RIWA.

Het komt helaas regelmatig voor dat meetreeksen zijn onderbroken, waardoor het moeilijker is om tot statistisch onderbouwde uitspraken te komen over normoverschrijdingen en trends. Dit kan meerdere oorzaken hebben, zoals veranderingen in analysemethode en/of uitvoerend laboratorium, misvattingen over het analysepakket of (tijdelijke) bezuinigingen. Hierdoor zijn betrouwbare, op meetgegevens van de betreffende parameters gefundeerde uitspraken, niet meer mogelijk. Om na te gaan of het imputeren[1] van onvolledige meetreeksen een oplossing voor dat probleem kan bieden, is RIWA een studie naar de mogelijkheden van het imputeren gestart. Als aanvullende toepassingen worden gezien het repareren van vertekende meetreeksen - zoals ontstaan door het toepassen van een verkeerde bemonsterings- of analysemethode[2] - en het identificeren van individuele verdachte meetwaarden. We kunnen namelijk stellen dat een meetwaarde verdacht is, als deze sterk afwijkt van de voor dat meettijdstip geïmputeerde waarde. Dit geeft overigens nog geen rechtvaardiging om de meetwaarde te verwijderen, want daarvoor moeten we zeker weten dat het een fout betreft.

---

[1] *Imputeren is het vervangen van een ontbrekende waarde door een schatting. Die schatting wordt aangeduid als imputant.*

[2] *Voorbeelden van vertekende meetreeksen zijn die van AOX, waaraan RIWA en IAWR in 2003 een uitgebreide enquête hebben besteed en DOC in 2005.*

Voor wat betreft de meetperiode 2008 t/m 2012 bevat RIWA-base 2.941 meetreeksen uit het Nederlandse deel van het Rijnstroomgebied, waarvan 1.353 voldoen aan de eerder genoemde eis van minstens 10 maanden per jaar met een meetwaarde. De 1.588 meetreeksen die niet voldoen aan de eis betreffen parameters die geschrapt zijn uit het meetprogramma, of recentelijk zijn toegevoegd maar nog geen vijf jaar volledig zijn gemeten. Een relevant deel van deze 1.588 reeksen - namelijk 525 reeksen - voldoet echter bijna aan de eis, aangezien deze over de betreffende vijfjaarsperiode in hooguit 12 maanden geen metingen bevatten. Deze reeksen komen daarom zeker in aanmerking voor het imputeren, zodat hun informatie niet onbenut hoeft te blijven. Voor wat betreft de in RIWA-base aanwezige meetreeksen uit het stroomgebied van de Maas komt eveneens een groot aantal in aanmerking voor het imputeren.

1.2 Over dit rapport

Na deze inleiding vat hoofdstuk 2 twee voorstudies samen van het imputeren van RIWA-reeksen, zodat we de draad op kunnen pakken. In hoofdstuk 3 is vervolgens beschreven hoe Random Forest uit de bus kwam als meest geschikte imputeermethode voor de procedure. In hoofdstuk 4 worden de keuzen voor de instellingen van Random Forest onderbouwd. Hoofdstuk 5 presenteert het voorstel van de procedure voor het imputeren van ontbrekende waarden en voor het beoordelen van meetwaarden, bedoeld voor toepassing op RIWA-base. En tenslotte geeft hoofdstuk 6 een aantal illustraties van het imputeren van ontbrekende waarden van meetreeksen uit RIWA-base. Het hoofddeel van dit rapport sluit af met de alfabetisch gerangschikte lijst van de literatuurverwijzingen.

Dit rapport heeft 3 bijlagen. Bijlage 1 vermeldt de resultaten van de multicriteria-analyse voor de voorselectie van de imputeermethode. Bijlage 2 licht de drie in deze studie vergeleken imputeermethoden toe, namelijk lineaire regressie, Random Forest en het neurale netwerk. Bijlage 3 geeft formele beschrijvingen van de voor de beoordeling gebruikte kunstmatige tijdreeksprocessen.

1.3 Tip voor een verkorte route door dit rapport

Aangezien dit rapport grotendeels handelt over wiskundige en statistische methoden, is het onvermijdelijk dat het veel technische gedeelten bevat die niet voor iedereen interessant zijn. Maar diegenen die in ieder geval de belangrijkste boodschappen willen meenemen kunnen volstaan met de volgende rapportgedeelten:

- de samenvatting;
- de afsluitende paragrafen van de hoofdstukken 2, 3 en 4, waar de kernpunten van het hoofdstuk zijn samengevat;
- § 3.2.2: Het principe van de werking van Random Forest;
- hoofdstuk 5: Procedure imputeren en beoordelen RIWA-base;
- hoofdstuk 6: Imputeren met Random Forest: enkele voorbeelden.

2

Voorstudies imputeren en beoordelen RIWA-reeksen

In het ontwikkelingstraject van de procedure voor het imputeren en beoordelen van RIWA-reeksen zijn er reeds twee voorstudies verricht. Om de draad op te kunnen pakken worden deze hieronder kort samengevat.

2.1 Voorstudie 2010

Uit een eerste verkennende studie van het imputeren van RIWA-reeksen is gebleken dat dit betrouwbaardere resultaten oplevert als gebruik wordt gemaakt van statistische relaties met andere parameters van dezelfde meetlocatie en/of van dezelfde parameter op naastliggende meetlocaties [Smits en Baggelaar, 2010]. Daarbij zijn de imputeerprestaties vergeleken van transfermodellering (volgens de Box-Jenkins-methode) en een neuraal netwerk[3], bij toepassing op een meetreeks van amidotrizoïnezuur (een röntgencontrastmiddel) in de Rijn bij Lobith. Het neuraal netwerk bleek betrouwbaardere imputanten voor die reeks op te leveren.

Als praktische toepassing zijn in die studie vervolgens ontbrekende waarden van een tiental röntgencontrastmiddelen geïmputeerd met behulp van een neuraal netwerk. Hierdoor konden de onderbroken reeksen van deze stoffen gemeten bij Lobith, Nieuwegein en Andijk zodanig worden aangevuld, dat ze toch in aanmerking kwamen voor de standaardverwerking voor het jaarrapport 2010 van de Rijn. Gezien de hoge kosten van de chemische analyse van röntgencontrastmiddelen kon zo kapitaal- en informatieverlies worden voorkomen.

2.2 Voorstudie 2012

Aangezien uit de eerste verkennende studie bleek dat de statistische relaties met andere meetreeksen van groot belang zijn bij het imputeren, is in een aansluitende studie getracht de relaties tussen meetreeksen van RIWA-base in kaart te brengen [Baggelaar et al., 2012]. Gezien het grote aantal meetreeksen in RIWA-base – ruim 11.500 – was het echter ondoenlijk om álle mogelijke onderlinge relaties van meetreeksen in beeld te brengen en is het onderwerp pragmatisch afgebakend.

Eerst is een clusteranalyse uitgevoerd van de parameters gemeten in de Rijn bij Lobith, gericht op het opsplitsen van de verzameling parameters in min of meer homogene groepen parameters. Elke reeks is daartoe omgezet naar een reeks maandwaarden. Maar de clusteranalyse bleek slechts beperkt mogelijk met de gegevens van RIWA-base, doordat veel meetreeksen gaten vertonen. Bij het ontbreken van een waarde van één van de betrokken parameters doet de betreffende maand namelijk niet meer mee voor de clusteranalyse, ongeacht of er voor de andere parameters wél waarden beschikbaar zijn. Uiteindelijk resteerde er slechts een beperkte gegevensset voor de clusteranalyse (58 parameters), die leidde tot 35 redelijk verklaarbare clusters.

Vervolgens is van een zevental bij Lobith gemeten parameters nagegaan welke relaties deze vertonen met andere parameters gemeten bij Lobith, zowel over een lange periode (1976 t/m 2011), als over een korte recente periode (2006 t/m 2011), eveneens op maandbasis.

---

[3] *Het wordt ook wel aangeduid als kunstmatig neuraal netwerk, om het duidelijker te onderscheiden van het neuraal netwerk in het menselijk lichaam. Maar we gaan er van uit dat die begripsverwarring hier niet zal spelen en gebruiken daarom de verkorte aanduiding.*

De statistische relatie tussen twee meetreeksen is daarbij gekwantificeerd met de Spearman-rangcorrelatiecoëfficiënt. Dit is het robuuste equivalent van de 'gewone' correlatiecoëfficiënt en is gebaseerd op de rangnummers van de waarden, in plaats van op de waarden zelf, zodat deze niet of nauwelijks kan worden vertroebeld door uitschieters. Het is ook de betere correlatiemaat als de waarden van één of beide betrokken meetreeksen niet afkomstig zijn uit een normale kansverdeling. Dit laatste is meestal het geval bij parameters van de oppervlaktewaterkwaliteit.

Slechts één van de zeven parameters – beryllium – bleek over geen van beide perioden gerelateerd aan een andere parameter. AMPA bleek alleen over de lange periode gerelateerd aan minstens één parameter (in feite aan vier parameters, namelijk carbofuran, aldicarb, oxamyl en aldicarbsulfoxide, zij het in al deze gevallen sterk negatieve relaties). De vijf andere beschouwde parameters (ammonium, cobalt, carbamazepine, MTBE en jomeprol) bleken over beide perioden sterk of zeer sterk gerelateerd aan minstens één parameter. Zeer sterke relaties waren er slechts voor ammonium over de lange periode (met nitraat) en voor cobalt over de recente periode (met nikkel en vanadium).

Voor het imputeren aan de hand van een gerelateerde meetreeks verdient het de voorkeur gebruik te maken van de statistische relatie zoals vastgesteld over een korte recente periode, want over een lange periode kan er sprake zijn van veranderingen in de relatie, door geleidelijke trends. Er blijkt voor vijf van de zeven beschouwde parameters sprake van een sterke relatie met één of meer andere parameters over de korte recente periode.

Tenslotte is voor elk van deze zeven geselecteerde parameters nagegaan welke kruisrelaties ze vertonen met dezelfde parameter gemeten bij Nieuwegein, voor verschillende tijdsverschuivingen (maanden). Voor zes van deze parameters bleek er sprake van een duidelijke correlatie tussen de concentratie bij Nieuwegein en die bij Lobith van dezelfde maand, dus zonder tijdsverschuiving. De relaties bleken met tijdsverschuivingen minder sterk. Voor cobalt bleek er daarentegen zelfs zonder tijdsverschuiving nauwelijks sprake van correlatie.

2.3 Kernpunten voorstudies

- Het imputeren levert betrouwbaardere resultaten op als gebruik wordt gemaakt van statistische relaties met andere parameters van dezelfde meetlocatie en/of van dezelfde parameter op naastliggende meetlocaties.
- Bij een eerste verkennende exercitie bleek het imputeren met een neuraal netwerk van een meetreeks van amidotrizoïnezuur in de Rijn bij Lobith bruikbare imputanten op te leveren.
- Uit een steekproef bleek dat RIWA-reeksen onderlinge statistische relaties vertonen die benut kunnen worden voor het imputeren. Het betreft zowel relaties met andere parameters van dezelfde meetlocatie, als relaties met parameters op naastliggende meetlocaties.
- Het is aan te bevelen om de statistische relaties tussen meetreeksen te beoordelen en te vergelijken op basis van de Spearman-rangcorrelatiecoëfficiënt. Deze wordt namelijk niet of nauwelijks vertroebeld door uitschieters en is ook een betere correlatiemaat als de waarden van één of beide betrokken reeksen niet afkomstig zijn uit een normale kansverdeling, wat meestal het geval is bij parameters van de oppervlaktewaterkwaliteit.
- Voor het imputeren aan de hand van een gerelateerde meetreeks verdient het de voorkeur gebruik te maken van de statistische relatie zoals vastgesteld over een korte recente periode, want over een lange periode kan er sprake zijn van geleidelijke veranderingen in de relatie.

3

Selectie meest geschikte imputeermethode

Zoals beschreven in het voorgaande hoofdstuk bleek het neurale netwerk bij eerste verkenningen tot bruikbare imputanten te kunnen leiden, maar er zijn meer methoden beschikbaar om te imputeren. Dit hoofdstuk beschrijft hoe stapsgewijs de meest geschikte imputeermethode voor de te ontwikkelen procedure is geselecteerd.

3.1 Beoordeling mogelijkheden van verschillende imputeermethoden

Er zijn de laatste 15 à 20 jaar meerdere imputeermethoden ontwikkeld. In veel gevallen zijn deze toegespitst om problemen op te lossen die werden veroorzaakt door de non-respons bij enquêtes (vaak met niet-numerieke gegevens), maar veel daarvan kunnen ook worden toegepast op problemen met numerieke gegevens.

We kunnen op hoofdlijnen de volgende twee groepen imputeermethoden onderscheiden:

1. Methoden die enkelvoudig imputeren: elke ontbrekende waarde wordt vervangen door één imputant.
2. Methoden die meervoudig imputeren: elke ontbrekende waarde wordt vervangen door meerdere imputanten (veelal zijn dit er vijf). Voorbeelden zijn Amelia[4] [Honaker et al., 2012] en MICE [Van Buren and Groothuis-Oudshoorn, 2011].

Meervoudig imputeren heeft voor- en nadelen

Het voordeel van meervoudig imputeren is dat uit de spreiding van de imputanten voor een bepaalde waarde kan worden afgeleid hoe betrouwbaar het imputeren is. Een nadeel is echter dat er dan speciale statistische verwerkingsmethoden nodig zijn, aangezien de imputanten op een bepaalde wijze moeten worden samengevoegd. Dergelijke speciale verwerkingsmethoden zijn al wel ontwikkeld voor het schatten van kengetallen, maar nog niet voor het schatten van trends, terwijl juist de trendanalyse voor de RIWA een belangrijke toepassing is. Een ander nadeel van het meervoudig imputeren is dat daarvoor een complexe en foutgevoelige database moet worden opgezet, met meerdere versies van elke meetreeks. Gezien deze nadelen zijn de meervoudige imputeermethoden niet bij onze beoordeling betrokken.

Beschouwde enkelvoudige imputeermethoden

De geschiktheid van een imputeermethode zal niet alleen afhangen van de onderliggende wiskundige algoritmes, maar ook van de mogelijkheden die zijn softwarematige implementatie bieden. Vandaar dat beide elementen zijn meegenomen bij de beoordeling.

Er zijn acht enkelvoudige imputeermethoden beoordeeld met behulp van een multicriteria-analyse (MCA). Tabel 3.1 vermeldt welke methoden dit betreft en ook tot welke groep ze behoren. In bijlage 1 worden deze methoden nader toegelicht.

---

[4] *De naam verwijst toepasselijk naar de wellicht bekendste 'missing person' uit de moderne geschiedenis, namelijk Amelia Earhart, een Amerikaanse pilote. Zij werd in 1932 beroemd, toen zij als eerste vrouw in een solovlucht de Atlantische Oceaan overstak. Maar in juli 1937 verdween ze (met haar navigator) spoorloos in de Grote Oceaan, tijdens wat de langste vlucht tot dan toe moest worden. Er zijn sindsdien vele expedities uitgezet om ze te zoeken, maar steeds zonder succes.*

Tabel 3.1: De acht imputeermethoden die met behulp van multicriteria-analyse zijn beoordeeld op geschiktheid voor toepassing in de te ontwikkelen procedure. Zie bijlage 1 voor een toelichting op de methoden.

Aanduiding	Module/Taal	Groep	Werkveld
NN-MBP	MBP/MBP	Neuraal Netwerk	Machinaal Leren
NN-R	NN/R		
SOM+EOF	ICS/Matlab		
RF-missForest	missForest/R	Niet-lineaire regressie	
BJ+interv	TSA/R	Tijdreeksmodellering	Statistiek
mtsdi-R	mtsdi/R		
Imputation-R	imputation/R	Lineaire regressie	
IRMI-R	VIM/R		

Multicriteria-analyse

Bij de multicriteria-analyse zijn door drie deskundigen[5] per imputeermethode scores van 0 t/m 10 toegekend aan twaalf aspecten, die voor toepassing door de RIWA van belang zijn te achten. Evenzo is aan elk aspect een gewicht van 0 t/m 10 toegekend, dat weergeeft welk belang dat aspect heeft voor de RIWA. De twaalf aspecten zijn vermeld in tabel 3.2, evenals het gemiddelde gewicht en het daaruit berekende relatieve gewicht (deze laatste sommeren tot 1). Per aspect is het gemiddelde van de drie scores vermenigvuldigd met het relatieve gewicht en vervolgens zijn die twaalf producten gesommeerd tot de totaalscore. Door te vermenigvuldigen met het relatieve gewicht zijn ook de totaalscores weer in de schaal 0 t/m 10, wat ze inzichtelijker maakt.

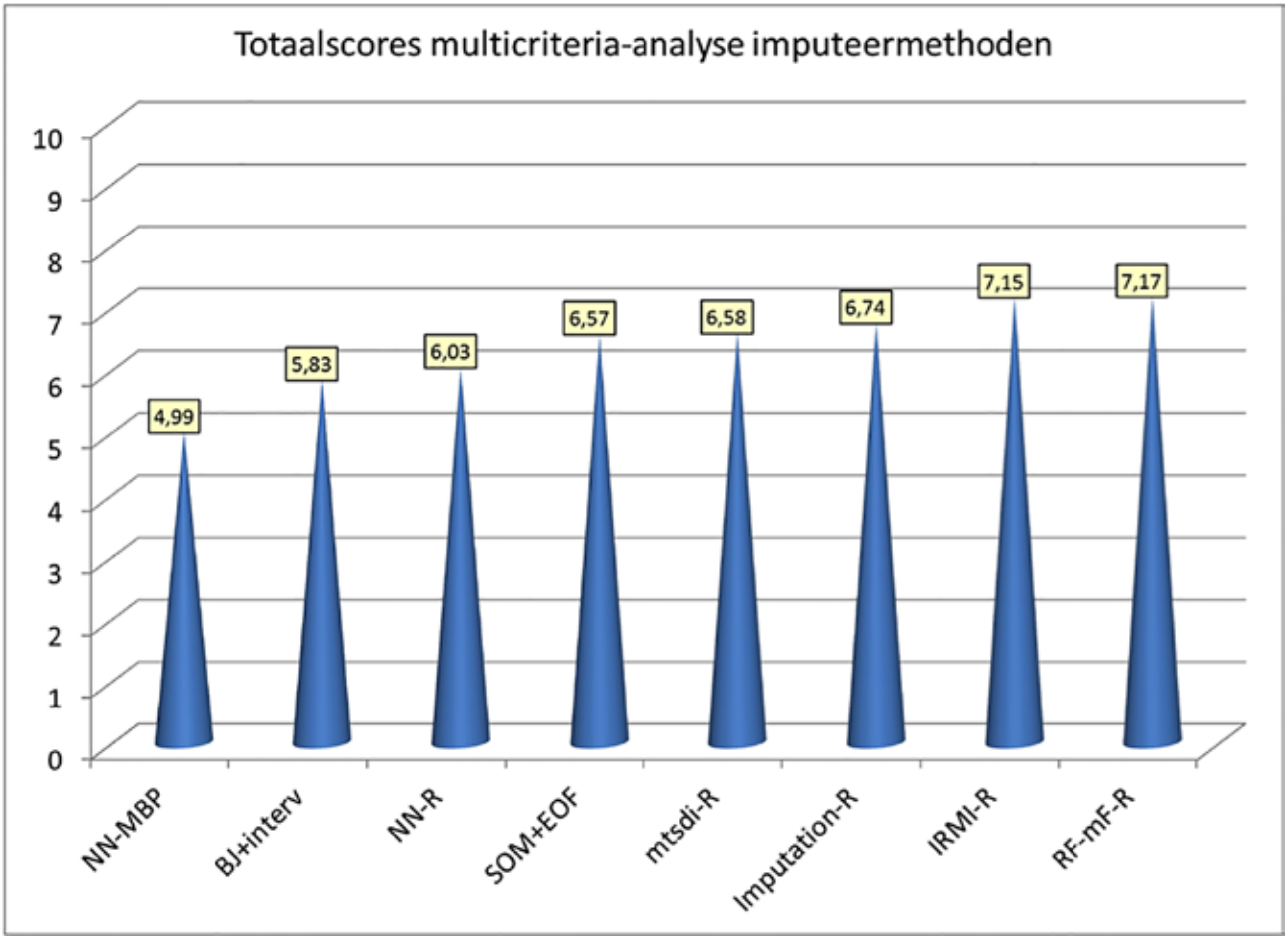
Tabel 3.2: De aspecten waaraan per imputeermethode scores zijn toegekend door drie deskundigen, alsmede het per aspect toegekende gemiddelde gewicht en het daaruit resulterende relatieve gewicht.

Aspect	gemiddeld gewicht	relatief gewicht
Betrouwbaarheid resultaten te toetsen	8,50	0,10
Te automatiseren	9,75	0,11
Kosten software	6,00	0,07
Uiteindelijke hoeveelheid werk	9,25	0,11
Te doorgronden	8,75	0,10
Maakt gebruik van zoveel mogelijk informatie	7,50	0,09
Vergt nog maar weinig uitzoekwerk	5.75	0,07
Geschikt gebleken voor hetzelfde soort gegevens als in RIWA-base	7,00	0,08
Betrouwbare bron	7,25	0,08
Mogelijkheid om te loggen	5,00	0,06
Imputanten kenbaar te maken in RIWA-base	7,00	0,05
Geeft kwantitatieve maat voor succes imputeerproces	8,00	0,09

[5] *Het betreft een data-analist/statistisch programmeur, een RIWA-base-deskundige en een milieustatisticus.*

De tabel met de scores per combinatie van imputeermethode en aspect en met de daaruit berekende totaalscores is vermeld in bijlage 1. De totaalscores zijn per imputeermethode ook weergegeven in het kegeldiagram van figuur 3.1.

Figuur 3.1: Kegeldiagram van de totaalscores van de multicriteria-analyse van de imputeermethoden.
Zie bijlage 1 voor een toelichting op de methoden.



Bij de multicriteria-analyse kregen de imputeermethoden IRMI-R en RF-missForest-R de hoogste totaalscores, waarbij het verschil tussen beide minimaal was. IRMI-R is een toepassing van meervoudige lineaire regressie, terwijl RF-missForest-R een toepassing is van Random Forest, op te vatten als een niet-lineaire regressiemethode (zie § 3.2.2 voor het principe van de werking van Random Forest).

3.2 Vergelijking prestaties imputeermethoden

3.2.1 Beschouwde methoden en meetreeksen

Met Monte Carlosimulaties zijn de imputeerprestaties beoordeeld van de volgende methoden:

1. Random Forest (RF);
2. Lineaire regressie (LR);
3. Neuraal netwerk (NN).

De principes van deze drie methoden zijn toegelicht in bijlage 2. Aangezien Random Forest de onbekendste van de drie, is wordt het principe van zijn werking ook wat minder formeel toegelicht in de volgende paragraaf (§ 3.2.2).

Zoals beschreven in de voorgaande paragraaf kregen de imputeermethoden IRMI en Random Forest bij de multicriteria-analyse de hoogste scores. Die scores waren nog slechts gebaseerd op kwalitatieve aspecten. Maar bij een eerste praktijkbeoordeling van deze methoden aan de hand van reeksen uit RIWA-base bleek de softwarematige implementatie van IRMI numeriek zeer instabiel. Daarom is besloten het te vervangen met de Matlabmodule Regress, die hetzelfde modelleerprincipe als IRMI toepast, namelijk meervoudige lineaire regressie. Verschillen zijn echter dat Regress de gebruikelijkere kleinste-kwadraten-schatter hanteert en robuust is bij sterke onderlinge correlaties van de predictoren (multicollineariteit).
Verder is het neurale netwerk toegevoegd als een soort maatstafmethode, aangezien daar reeds bij de eerste imputeerexercities positieve ervaringen mee zijn opgedaan [Smits en Baggelaar, 2010].

De kwantitatieve prestaties van deze drie imputeermethoden zijn vergeleken door ze toe te passen op twee datasets, elk met een groot aantal tijdreeksen van 5 jaar lengte, met 12 maandwaarden per jaar, om aan te sluiten op de intentie van de RIWA om reeksen van die lengte te imputeren.[6]
De twee beschouwde datasets betreffen:

1. een dataset met kunstmatige meetreeksen, met verschillende gradaties van eigenschappen zoals scheefheid, seizoenseffecten, percentage gecensureerde waarden en mate van correlatie tussen de predictoren;
2. een dataset met díe meetreeksen uit RIWA-base van de meetlocaties Lobith, Nieuwegein, Nieuwersluis en Andijk, die van 2007 t/m 2011 60 maandwaarden bevatten (dus zonder ontbrekende maandwaarden).

De prestaties van de imputeermethoden bij toepassing op de tweede dataset zijn uiteraard het relevantst, aangezien die reeksen afkomstig zijn uit het bedoelde toepassingsgebied. Het meenemen van de dataset met kunstmatige reeksen is vooral bedoeld om een beeld te krijgen van de invloeden van reekseigenschappen op de imputeerprestaties. Bij kunstmatige reeksen kunnen er immers meer en specifiekere variaties worden aangebracht in reekseigenschappen.

We vergelijken de simuleerprestaties van de drie methoden voor verschillende situaties, waarbij steeds één of meer missende waarden van een reeks zijn geïmputeerd met behulp van relaties met andere reeksen (de predictoren). Maar eerst geven we in de volgende paragraaf een toelichting op het principe van de werking van Random Forest, aangezien dat voor velen de onbekendste van de drie beschouwde methoden zal zijn.

[6] *Voor de jaarrapportages worden immers statistische verwerkingen uitgevoerd op maandbasis en op jaarbasis.*

3.2.2 Het principe van de werking van Random Forest

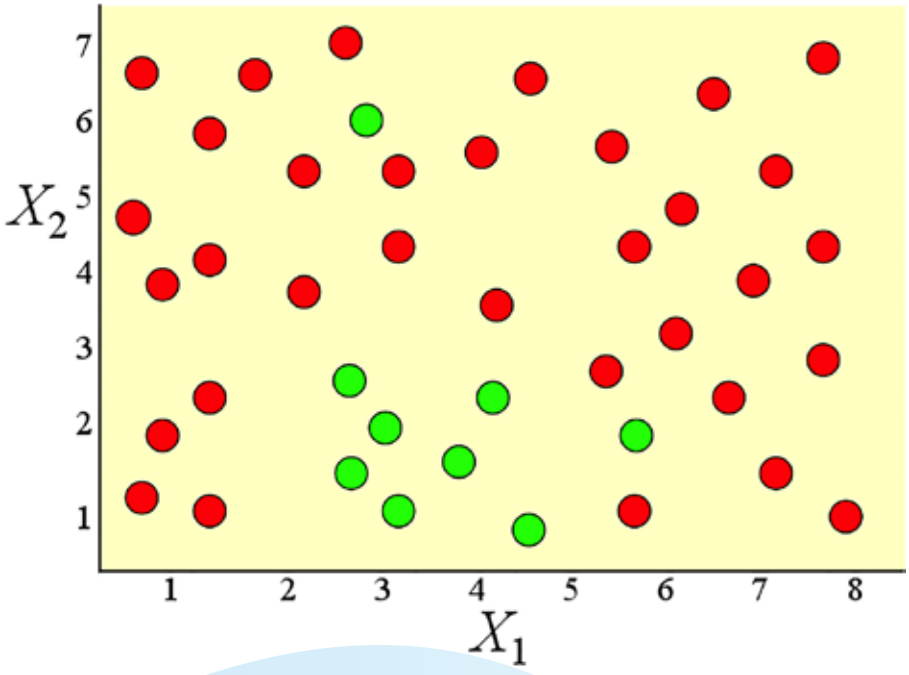
Random Forest [Breiman, 1993, 2001] is net als het neurale netwerk een methode uit het onderzoeksveld Machinaal Leren. Het machinaal leren houdt in dat de methode op basis van een trainingsset met gegevens 'leert' om bij een bepaalde invoer de daarbij behorende uitvoer te schatten.

Zoals de bloemrijke naam al doet vermoeden bestaat Random Forest uit een verzameling beslisbomen (decision trees), waarbij een beslisboom een schema is dat de verschillende mogelijke trajecten naar een bepaald doel (de uitvoer) weergeeft. Dit schema bestaat uit een opeenvolging van binaire beslisregels.

We zullen hieronder de werking van een beslisboom toelichten voor een eenvoudige dataset met twee predictoren (de verklarende variabelen X_1 en X_2) en een uitvoerreeks (Y) die slechts twee mogelijke waarden heeft, zoals A of B (maar vele andere voorbeelden zijn mogelijk, zoals 0 of 1, goed of slecht, winst of verlies, etc.).
De onderstaande figuur toont links zeven willekeurige records van de dataset en rechts een tweedimensionale weergave van de gehele dataset, waarbij voor de overzichtelijkheid een waarde A is weergegeven met een rode cirkel en een waarde B met een groene cirkel.

Figuur 3.2: Dataset met twee predictoren (verklarende variabelen) en één uitvoerreeks, om de werking van een beslisboom van Random Forest toe te lichten. Links zeven willekeurige records van de dataset en rechts een tweedimensionale weergave van de gehele dataset.

X_1	X_2	Y
6,1	3,1	A
3,2	5,4	A
2,9	1,9	B
1,6	6,6	A
5,5	1,8	B
2,5	2,6	B
7,9	1,0	A
...	...	...

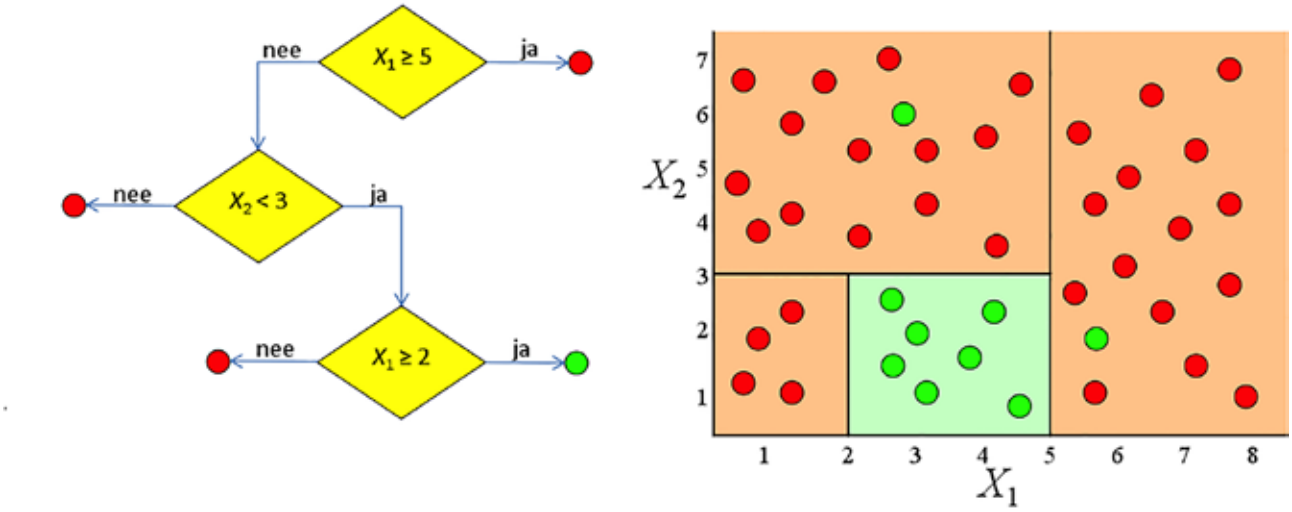


De beslisboom wordt als volgt serieel opgebouwd uit een aantal knooppunten.

1. Bij het eerste knooppunt wordt bepaald bij welke waarde van één van de predictoren (X_1 of X_2) de verzameling *Y*-waarden in de twee meest homogene delen wordt opgesplitst. Dit blijkt de waarde 5 van X_1 te zijn. Bij $X_1 \geq 5$ is *Y* bijna altijd rood (A).
2. Bij het tweede knooppunt wordt bepaald bij welke waarde van één van de predictoren de deelverzameling *Y*-waarden met $X_1 < 5$ in de twee meest homogene delen wordt opgesplitst. Dit blijkt de waarde 3 van X_2 te zijn. Bij $X_1 < 5$ en $X_2 \geq 3$ is *Y* bijna altijd rood (A).
3. Bij het derde knooppunt wordt bepaald bij welke waarde van één van de predictoren de deelverzameling *Y*-waarden met $X_1 < 5$ en $X_2 < 3$ in de twee meest homogene delen wordt opgesplitst. Dit blijkt de waarde 2 van X_1 te zijn. Bij $2 \leq X_1 < 5$ en $X_2 < 3$ is *Y* altijd groen (B), terwijl deze bij $X_1 < 2$ altijd rood (A) is.

De resulterende beslisboom is in onderstaande figuur links weergegeven. Rechts zien we hoe de dataset door de splitsingswaarden van de drie knooppunten in zo homogeen mogelijke deelverzamelingen is opgesplitst.

Figuur 3.3: Links: de beslisboom die is afgeleid voor de voorbeelddataset, met drie knooppunten. Rechts: Opsplitsing van de dataset door de splitsingswaarden van de drie knooppunten.



Het principe van de beslisboom is dus dat de verzameling *Y*-waarden door de optimale splitsingswaarden van de predictoren wordt opgesplitst in zo homogeen mogelijke deelverzamelingen. In het bovenstaande voorbeeld is dit uitgewerkt voor een uitvoerreeks *Y* met slechts twee nominale uitkomsten (A en B), maar het principe werkt net zo goed voor een uitvoerreeks *Y* met numerieke uitkomsten, zoals de RIWA-reeksen van de oppervlaktewaterkwaliteit.

Beslisbomen zoals boven bestonden al lange tijd, maar werden opgevat als zwakke voorspellers, aangezien ze doorgaans geen zeer betrouwbare schattingen opleverden (het bovenstaande voorbeeld moet dan ook worden gezien als een bijna ideale situatie, waar wél zeer betrouwbare schattingen resulteren). De vondst van Breiman was dat door vele beslisbomen te combineren en ook enkele stochastische elementen toe te voegen een veel sterkere voorspeller ontstond, die hij toepasselijk aanduidde als Random Forest.

Meer technische details

De door Breiman toegevoegde stochastische elementen zijn de volgende:

1. Elke beslisboom wordt afgesteld met een trainingsset die even groot is als de oorspronkelijke dataset, maar doordat elk record wordt getrokken mét teruglegging, volgt uit kansberekening dat de trainingsset gemiddeld circa 63% van de records van de oorspronkelijke dataset bevat, waarbij meerdere records tweemaal of vaker voorkomen. De overige 37% records, aangeduid als de OOB-set (Out-of-Box-set), wordt gebruikt om een empirische en zuivere indruk te krijgen van de betrouwbaarheid van de voorspelling. De OOB-set is immers niet gebruikt voor het trainen van de beslisboom. Daartoe worden voor elke boom de *Y*-waarden van zijn OOB-set voorspeld, met behulp van de bijbehorende waarden van de predictoren. Vervolgens worden alle voorspellingen van een bepaalde *Y*-waarde gemiddeld over de betrokken beslisbomen en vergeleken met de werkelijke waarde *Y*. Dit schept een groot voordeel ten opzichte van andere imputeermethoden, aangezien daarmee bij elke imputatie ook een maat kan worden geleverd voor zijn nauwkeurigheid.
2. Per knooppunt wordt aselect een aantal predictoren uit de predictorenset getrokken, waarmee vervolgens de optimale splitsingswaarde van dat knooppunt wordt bepaald. Doorgaans wordt het aantal predictoren per knooppunt gesteld op $p/3$, afgerond naar boven, waarbij p het totaal aantal predictoren is.

Bij Random Forest stopt het vertakken van een deel van de beslisboom als daardoor een eindpunt (blad) minder *Y*-waarden zou gaan bevatten dan het daarvoor ingestelde minimum. Dat minimum is doorgaans ingesteld op 3, 4 of 5 waarden. Als een blad meer dan één *Y*-waarde bevat, geeft dat blad hun gemiddelde als schatting. Daarmee kan een geschatte waarde niet kleiner zijn dan het minimum van *Y* en niet groter dan het maximum van *Y*. Bij imputeren kan dit er toe leiden dat de variatie in een reeks enigszins wordt afgevlakt.

3.2.3 Vergelijking imputeerprestaties bij toepassing op kunstmatige reeksen

De imputeerprestaties van de drie methoden zijn eerst vergeleken bij toepassing op de dataset met kunstmatige reeksen. Daarbij is de volgende simulatieopzet toegepast:

1. Genereer 10 reeksen *X* (de predictoren), elk bestaande uit 60 onafhankelijke trekkingen uit een normale kansverdeling en combineer deze zodanig dat de resulterende reeks *Y* voldoet aan bepaalde proceskarakteristieken (zie ook tabel 3.3).

2. Verwijder 15 waarden van *Y*.

3. Imputeer de missende waarden van *Y* vervolgens met behulp van de predictoren en doe dit afzonderlijk met Random Forest, lineaire regressie en het neurale netwerk.

4. Bepaal voor elk van de drie imputeermethoden afzonderlijk van elke geïmputeerde waarde de imputeerfout als:
   imputeerfout = geïmputeerde waarde – oorspronkelijke waarde [4]
   en de relatieve imputeerfout als:
   %imputeerfout = 100%*imputeerfout / oorspronkelijke waarde [5]

5. Bepaal voor elk van de drie imputeermethoden afzonderlijk van elk jaar waarvan één of meer waarden zijn geïmputeerd de schattingsfout van elk van de jaarkengetallen die relevant zijn voor de RIWA-rapportages, namelijk minimum, 10-percentiel (P10), mediaan, gemiddelde, 90-percentiel (P90), maximum en standaardafwijking als:
   schattingsfout = geschat jaarkengetal – oorspronkelijk jaarkengetal [6]
   en de relatieve schattingsfout als:
   %schattingsfout = 100%*schattingsfout / oorspronkelijk jaarkengetal [8]

6. Herhaal voor elk proces de voorgaande stappen 1 t/m 5 200 maal. Aangezien elk proces een stochastisch element heeft, resulteren er 200 verschillende reeksen per proces, zij het wel alle met dezelfde proceskarakteristieken.

7. Bepaal tenslotte voor elk van de drie imputeermethoden afzonderlijk de statistische kenmerken van de verdeling imputeerfouten en van de verdeling schattingsfouten van het jaarkengetal en leid daaruit prestatiescores af voor de drie methoden (zie ook verder).

Uit de 10 normaal verdeelde predictorenreeksen zijn 11 verschillende soorten processen gesimuleerd, met de in tabel 3.3 vermelde kenmerken. Bijlage 3 bevat de formele beschrijvingen van deze processen.

Tabel 3.3: Kenmerken van de elf kunstmatige processen die zijn gesimuleerd om de prestaties van de drie imputeermethoden te kunnen vergelijken.

Nummer	Omschrijving proces	Predictoren X(10)	Ruis
1	Lineair	Normaal verdeeld	Normaal verdeeld
2	Niet-lineair (1/X)	Normaal verdeeld	Normaal verdeeld
3	Niet-lineair (X^3)	Normaal verdeeld	Normaal verdeeld
4	Niet-lineair ($\sqrt{X}$)	Normaal verdeeld	Normaal verdeeld
5	Lineair	Normaal verdeeld	Normaal verdeeld, met autocorrelatie
6	Lineair	Normaal verdeeld	Lognormaal verdeeld, met autocorrelatie
7	Lineair	Normaal verdeeld	Uniform verdeeld, met autocorrelatie
8	Niet-lineair (e^X)	Normaal verdeeld	Lognormaal verdeeld, met autocorrelatie
9	Lineair met seizoenseffecten	Normaal verdeeld, met seizoenseffecten	Normaal verdeeld
10			
11			

Elk van deze elf processen is gesimuleerd met verschillende instellingen voor:
- de collineariteit (dit is de mate waarin de predictoren onderling gecorreleerd zijn), met als instellingen onderlinge correlatiecoëfficiënten van 0 (geen collineariteit) en 0,5 (matige collineariteit);
- het percentage gecensureerde waarden van de *Y*-reeks en de predictoren, met de volgende combinaties als instellingen [0% 0%], [25% 0%], [25% 25%], [50% 0%], [50% 50%], [75% 0%] en [75% 75%]);
- het percentage extreme waarden van de *Y*-reeks, met als instellingen 0 en 10%.

Verder zijn de drie imputeermethoden per proces nog vergeleken bij:
- het imputeren van verschillende aantallen ontbrekende waarden, namelijk achtereenvolgens 1, 6, 11, .. en 41 ontbrekende waarden van de 60 waarden;
- het imputeren met verschillende aantallen predictoren, namelijk 10 en 5.

In totaal zijn er voor elk proces dus vele verschillende situaties vergeleken. Om daarbij het stochastische element te verdisconteren is elke situatie 200 maal gesimuleerd, met steeds aselect genomen beginwaarden.

Voor het vergelijken van de prestaties van de drie methoden is gebruik gemaakt van prestatiescores. Deze scores zijn afzonderlijk toegekend op basis van:

1. kenmerken van de kansverdeling van de imputeerfout;
2. kenmerken van de kansverdeling van de schattingsfout van het jaarkengetal. Dit afzonderlijk voor de jaarkengetallen minimum, P10, mediaan, gemiddelde, P90, maximum en standaardafwijking.

De beschouwde kenmerken van de kansverdeling van de imputeerfout of van de kansverdeling van de schattingsfout van het jaarkengetal zijn:

- **mediaan**;
- **gemiddelde**;
- **RMSE** (root mean squared error) – Dit is een maat voor de spreiding van de imputeerfout (of schattingsfout), die wordt berekend als de wortel van het gemiddelde van de gekwadrateerde imputeerfout (of schattingsfout);
- **onnauwkeurigheid** - De onnauwkeurigheid van het imputeren is gedefinieerd als de absolute waarde van de relatieve imputeerfout (relatief ten opzichte van de werkelijke waarde), die hooguit in 5% van de gevallen wordt overschreden. Het is dus een soort bijna-bovengrens van de absolute waarde van de relatieve imputeerfout. Deze wordt berekend als het maximum van de volgende twee waarden: 1. de absolute waarde van het 2,5-percentiel (P2,5) van de relatieve imputeerfout en 2. de absolute waarde van het 97,5-percentiel (P97,5) van de relatieve imputeerfout. En de onnauwkeurigheid van het schatten van een bepaald jaarkengetal is gedefinieerd als de absolute waarde van de relatieve schattingsfout (relatief ten opzichte van het werkelijke jaarkengetal), die hooguit in 5% van de gevallen wordt overschreden. Deze bijna-bovengrens van de absolute waarde van de relatieve schattingsfout wordt berekend als het maximum van de volgende twee waarden: 1. de absolute waarde van het 2,5-percentiel (P2,5) van de relatieve schattingsfout en 2. de absolute waarde van het 97,5-percentiel (P97,5) van de relatieve schattingsfout.

De scores zijn eerst bepaald voor elke afzonderlijke combinatie van proces, situatie en simulatie, waarbij de best presterende methode een score 2 kreeg, de één na beste een score 1 en de slechtste een score 0, terwijl bij gelijk presteren de score is gedeeld. Vervolgens is per proces de procentuele score van een imputeermethode vastgesteld als de totale score voor die methode, gedeeld door de totale score voor de drie methoden.

Resultaten van de vergelijking van de drie imputeermethoden

Tabel 3.4 toont voor elk van de elf processen de procentuele scores van de drie imputeermethoden, op basis van kenmerken van de kansverdeling van de imputeerfout, zoals bepaald over alle beschouwde situaties met verschillende instellingen voor collineariteit, percentage gecensureerde waarden van de Y-reeks en de predictoren en het percentage extreme waarden van de Y-reeks en voor verschillende aantallen ontbrekende waarden en voor verschillende aantallen predictoren. De beste procentuele score per drietal imputeermethoden is steeds groen gekleurd, terwijl de slechtste oranje is gekleurd.

Tabel 3.4: Procentuele scores van de drie imputeermethoden op basis van kenmerken van de kansverdeling van de imputeerfout, bepaald over alle beschouwde situaties per proces, met 200 simulaties per situatie.

	RF	LR	NN	RF	LR	NN	RF	LR	NN	RF	LR	NN
Proces	Mediaan			Gemiddelde			RMSE			Onnauwkeurigheid		
1	16	30	54	43	37	20	54	45	1	55	45	0
2	9	33	58	43	42	15	65	34	0	61	38	0
3	14	30	55	42	38	20	48	47	5	47	48	5
4	12	32	56	41	39	19	65	35	0	63	37	0
5	15	33	53	41	36	23	56	43	1	57	43	0
6	12	29	59	43	37	20	57	42	1	57	43	0
7	17	29	55	39	37	24	51	46	3	50	47	2
8	13	32	55	48	31	21	57	43	1	55	44	1
9	10	34	56	43	42	16	66	34	0	66	34	1
10	11	33	56	42	44	14	65	35	0	64	35	0
11	12	31	57	42	43	14	65	35	0	65	35	0

Voor het gemiddelde, de RMSE en de onnauwkeurigheid van de imputeerfout zijn de scores van Random Forest meestal beter dan van lineaire regressie. Alleen de scores voor de mediaan zijn van lineaire regressie beter dan van Random Forest. Het neurale netwerk scoort als beste voor de mediaan, maar de scores van deze methode voor het gemiddelde, de RMSE en de onnauwkeurigheid zijn het slechtst van de drie methoden.

Tabel 3.5 toont voor elk van de elf processen de procentuele scores van de drie imputeermethoden, op basis van kenmerken van de kansverdeling van de schattingsfout van het jaarkengetal, over alle beschouwde situaties.

Tabel 3.5: Procentuele scores van de drie imputeermethoden op basis van kenmerken van de kansverdeling van de schattingsfout van het jaarkengetal, bepaald over alle beschouwde situaties per proces, met 200 simulaties per situatie.

	RF	LR	NN	RF	LR	NN	RF	LR	NN	RF	LR	NN
Proces	Mediaan			Gemiddelde			RMSE			Onnauwkeurigheid		
1	32	41	28	37	46	17	48	47	4	51	46	3
2	31	44	25	34	50	15	50	48	1	50	46	4
3	30	38	32	34	43	23	42	47	10	45	46	9
4	32	43	25	36	50	14	52	47	2	52	46	2
5	31	40	29	37	45	18	49	47	4	52	44	4
6	30	40	30	36	45	19	50	46	4	53	43	4
7	31	39	30	36	44	20	46	47	7	48	46	6
8	31	41	28	36	44	21	47	48	4	50	46	4
9	30	42	28	35	47	18	50	46	4	51	45	4
10	31	42	27	36	47	17	49	47	4	50	45	5
11	31	41	27	36	47	17	49	47	4	50	46	5

De scores voor de mediaan en het gemiddelde van de schattingsfout van de jaarkengetallen zijn het best voor lineaire regressie. Maar de scores voor de RMSE en de onnauwkeurigheid van de schattingsfout van de jaarkengetallen zijn doorgaans het best voor Random Forest. Het toegepaste neurale netwerk heeft vrijwel steeds de slechtste scores.

Conclusies over de imputeerprestaties bij kunstmatige reeksen

Uit de boven beschreven resultaten kunnen de volgende conclusies worden getrokken.

1. Random Forest presteert doorgaans beter dan lineaire regressie als we afgaan op het gemiddelde, de RMSE en de onnauwkeurigheid van de imputeerfout en ook als we afgaan op de RMSE en de onnauwkeurigheid van de schattingsfout van de jaarkengetallen.

2. Lineaire regressie presteert doorgaans beter dan Random Forest als we afgaan op mediaan van de imputeerfout en op de mediaan en het gemiddelde van de schattingsfout van de jaarkengetallen.

3. Het neurale netwerk scoort vaak het beste als we afgaan op de mediaan van de imputeerfout. Maar voor de andere prestatiematen scoort het neurale netwerk bijna altijd als slechtste van de drie imputeermethoden.

3.2.4 Vergelijking imputeerprestaties bij toepassing op reeksen RIWA-base

De imputeerprestaties van de drie methoden zijn ook vergeleken bij toepassing op díe reeksen uit RIWA-base van de meetlocaties Lobith, Nieuwegein, Nieuwersluis en Andijk, die van 2007 t/m 2011 60 maandwaarden bevatten (dus zonder ontbrekende maandwaarden). Dit betreft 781 reeksen, waarvan er echter 321 alleen maar gecensureerde waarden bevatten. Van deze 321 volledig gecensureerde reeksen zijn er 141 constant, doordat alle 60 waarden van de reeks gecensureerd zijn ten opzichte van dezelfde rapportagegrens. De overige 180 volledig gecensureerde reeksen zijn niet constant, aangezien bij elk daarvan de waarden zijn gecensureerd ten opzichte van twee of meer rapportagegrenzen. De 321 volledig gecensureerde reeksen zijn niet betrokken bij de vergelijking, doordat het bij een dergelijke reeks voor de hand ligt een ontbrekende waarde te imputeren met een naïeve methode, zoals het vervangen van de ontbrekende waarde met het gemiddelde van de naastliggende waarden. Er resteren dan voor deze exercitie 460 reeksen die niet volledig gecensureerd zijn.

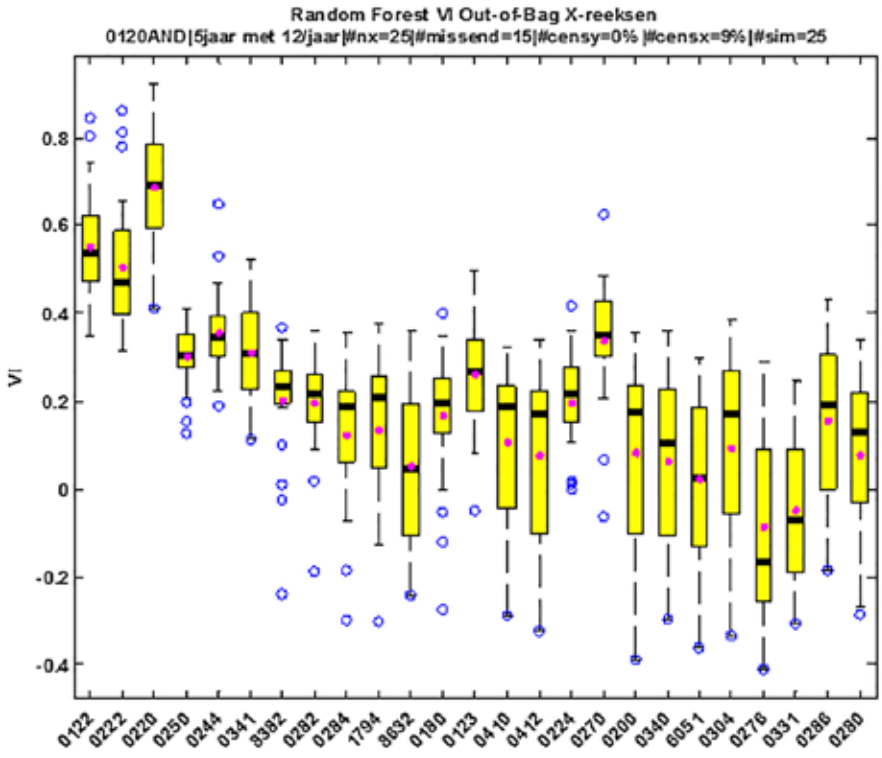
Bij de hiervoor beschreven exercitie met de kunstmatige reeksen waren de predictoren reeds bekend (10 normaal verdeelde reeksen), maar bij de praktijkreeksen uit RIWA-base is dat niet het geval. Daarom is bij de vergelijking ook de invloed van het gekozen aantal predictoren op de imputeerprestaties beschouwd.

In eerste instantie zijn voor elke reeks 25 predictoren van dezelfde meetlocatie geselecteerd op basis van de grootte van de absolute waarde van de Spearman-rangcorrelatiecoëfficiënt. Vervolgens is nagegaan of de imputeerprestaties konden worden verbeterd door nader te selecteren op de predictoren. Deze selectie is uitgevoerd met behulp van de VI (Variable Importance) van elke predictor, zoals bepaald met Random Forest. De VI geeft aan hoe belangrijk de predictor is voor het voorspellen met Random Forest.
Om de VI van een predictor vast te stellen wordt met Random Forest eerst per boom vastgesteld wat de verandering is van de gemiddelde gekwadrateerde imputeerfout (GKI) van de OOB-selectie door permutatie van die predictor. De permutatie houdt in dat de volgorde van de betreffende predictorreeks aselect wordt gewijzigd, zodat een eventuele relatie met de Y-reeks verdwijnt. De VI is dan het gemiddelde van deze verandering over alle bomen, gedeeld door de standaardafwijking van die verandering. Naarmate de predictor relevanter is voor de voorspelling zal de VI groter zijn.

Ter illustratie toont figuur 3.4 een voorbeeld van een VI-plot. Het betreft een exercitie waarbij 15 waarden van een vijfjaarsreeks maandwaarden van de watertemperatuur bij Andijk zijn geïmputeerd met Random Forest. Daarvoor zijn 25 predictoren geselecteerd op basis van de absolute waarde van de Spearman-rangcorrelatiecoëfficiënt met

de watertemperatuur bij Andijk. Er zijn 25 simulaties uitgevoerd, waarbij steeds aselect 15 van de 60 waarden zijn verwijderd en de VI's van de predictoren zijn bepaald. Op die manier konden er voor elke predictor 25 VI's worden geraamd. Deze zijn per predictor samengevat met een boxplot.
Om VI's te bepalen hoeven er overigens geen waarden te worden verwijderd, aangezien de VI al wordt vastgesteld bij het opstellen van het Random Forest. Maar door dit herhaald te doen met steeds andere verwijderde waarden, ontstaat er een realistischere stochastiek in de verdeling van de VI.

Figuur 3.4: VI-plot met boxplots van de VI van 25 predictoren voor het imputeren van de watertemperatuur (120) bij Andijk. De invoervariabelen zijn langs de horizontale as gesorteerd op basis van de absolute waarde van de Spearman-rangcorrelatiecoëfficiënt met de watertemperatuur bij Andijk.



Toelichting op de boxplot
De boxplot is een handzame manier om de kansverdeling van een verzameling waarden eendimensionaal weer te geven. Het middendeel – de box – loopt van het 25-percentiel tot het 75-percentiel. De mediaan (het 50-percentiel) is aangegeven als een horizontale streep in de box. En in bovenstaande boxplots is tevens het gemiddelde aangegeven, als een rode punt. De verticale lijnen strekken zich uit tot de extremen (minimum en maximum), tenzij zo'n extreem verder van de box ligt dan 1,5-maal de boxlengte, in welk geval de lijn stopt bij 1,5 maal de boxlengte. Extremere waarden zijn dan apart gemarkeerd (hier als cirkels).

Verklaring parametercodes
0122: zuurstof ; 0222: waterstofcarbonaat; 0220: koolstofdioxide; 0250: totale hardheid; 0244: calcium; 0341: nikkel, na filtr. over 0,45 µm; 8382: isoproturon; 0282: nitraat als N; 0284: ortho fosfaat als P; 1794: ethyleendiaminetetraethaanzuur (EDTA); 8632: aminomethylfosfonzuur (AMPA); 0180: zuurgraad; 0123: zuurstofverzadiging; 0410: UV-extinctie, 254 nm; 0412: kleurintensiteit, Pt/Co-schaal als Pt; 0224: carbonaat; 0270: ammonium als N; 0200: EGV (elek. geleid.verm., 20 °C); 0340: nikkel; 6051: amidotrizoïnezuur; 0304: mangaan; 0276: organisch gebonden stikstof als N; 0331: koper, na filtr. over 0,45 µm; 0286: totaal fosfaat als P; 0280 nitriet als N

Uit deze figuur blijkt dat voor een aantal predictoren in verschillende simulaties ook negatieve VI's resulteerden. Dit betekent dat de betreffende predictor mogelijk de nauwkeurigheid van het imputeren verlaagt en daarmee schadelijk kan zijn voor het imputeren.
Uit figuur 3.4 blijkt ook dat een grotere rangcorrelatiecoëfficiënt niet altijd een grotere VI hoeft op te leveren. Zo heeft de parameter 220 (koolstofdioxide) een grotere gemiddelde VI, maar een kleinere rangcorrelatiecoëfficiënt dan de parameters 122 (zuurstof) en 222 (waterstofcarbonaat). En de parameter 270 (ammonium) scoort ook relatief hoog bij de VI.

Om het nut van predictorenselectie op basis van de VI's te kunnen beoordelen, hebben we de imputeerprestaties van Random Forest bij toepassing op de 460 RIWA-reeksen vergeleken voor verschillende aantallen predictoren, namelijk voor 25 predictoren en voor vier subsets daarvan. De subsets zijn bepaald door eerst met Random Forest voor elke predictor 25 VI's te bepalen op de wijze zoals boven beschreven. Vervolgens zijn per subset alleen díe predictoren geselecteerd die voldoen aan het bijbehorende criterium:

- subset 1: het 50-percentiel van de VI's is positief;
- subset 2: het 25-percentiel van de VI's is positief;
- subset 3: het 10-percentiel van de VI's is positief.
- subset 4: het 5-percentiel van de VI's is positief.

Het minimum aantal predictoren per subset is bij deze exercities op 5 gezet.

Random Forest is per reeks voor elk van de predictorensets toegepast met 15 van de 60 waarden van de reeks verwijderd. Om daarbij het stochastische element zo goed mogelijk te verdisconteren is elke situatie 25 maal gesimuleerd, waarbij aselect steeds 15 waarden zijn verwijderd.
De afzonderlijke prestaties van de subsets 1 t/m 4 zijn afgezet tegen die van de uitgebreidere set met 25 predictoren. Net zoals bij het vergelijken van de imputeerprestaties bij toepassing op kunstmatige reeksen (zie § 3.2.3) is daarvoor afgegaan op prestatiescores die afzonderlijk zijn toegekend op basis van: 1) kenmerken van de kansverdeling van de imputeerfout en 2) kenmerken van de kansverdeling van de schattingsfout van het jaarkengetal. De scores zijn eerst bepaald voor elke combinatie van reeks en predictorenset, waarbij de best presterende predictorenset een score 1 kreeg en de andere een score 0, terwijl bij gelijk presteren de score is gedeeld. Vervolgens is per predictorenset de procentuele score vastgesteld als de totale score over alle reeksen, gedeeld door de totale score voor de twee predictorensets.

Tabel 3.6 vermeldt de procentuele scores per combinatie van predictorensets, op basis van kenmerken van de kansverdeling van de imputeerfout. Voor elk kenmerk is de beste score groen gekleurd.

Tabel 3.6: Procentuele scores van Random Forest met verschillende predictorensets (RF1 en RF2) bij toepassing op de 460 RIWA-reeksen, op basis van kenmerken van de kansverdeling van de imputeerfout. RF1 omvat steeds 25 predictoren en RF2 een subset daarvan op basis van het positief zijn van respectievelijk de P50, P25, P10 en P5 van de verdeling van de VI.

	RF1	RF2	RF1	RF2	RF1	RF2	RF1	RF2	RF1	RF2
Predictorselectie	Gemidd. #predictoren		Mediaan		Gemiddelde		RMSE		Onnauwkeurigheid	
25 predictoren - P50 VI	25	13,5	44	56	50	50	47	53	46	54
25 predictoren - P25 VI	25	8,5	41	59	51	49	42	58	40	60
25 predictoren - P10 VI	25	6,0	41	59	52	48	38	62	37	63
25 predictoren - P5 VI	25	5,5	37	63	48	52	39	61	38	62

Door het toepassen van de verschillende selectiecriteria neemt het gemiddelde aantal predictoren af van 25 tot uiteindelijk 5,5 (zie de derde kolom van tabel 3.6). Voor drie van de vier kenmerken van de kansverdeling van de imputeerfout is de procentuele score duidelijk hoger bij een subset van de 25 predictoren. Voor het vierde kenmerk (gemiddelde) zijn er weinig verschillen.
De hoogste scores zijn doorgaans voor de subsets gebaseerd op het 10-percentiel en het 5-percentiel van de VI's.

Tabel 3.7 vermeldt de procentuele scores per combinatie van predictorensets, op basis van kenmerken van de kansverdeling van de schattingsfout van het jaarkengetal.

Tabel 3.7: Procentuele scores van Random Forest met verschillende predictorensets (RF1 en RF2) bij toepassing op de 460 RIWA-reeksen, op basis van kenmerken van de kansverdeling van de schattingsfout van het jaarkengetal. RF1 omvat 25 predictoren en RF2 een subset daarvan op basis van het positief zijn van respectievelijk de P50, P25, P10 en P5 van de verdeling van de VI.

	RF1	RF2	RF1	RF2	RF1	RF2	RF1	RF2	RF1	RF2
Predictorselectie	Gemidd. #predictoren		Mediaan		Gemiddelde		RMSE		Onnauwkeurigheid	
25 predictoren - P50 VI	25	13,5	49	51	50	50	48	52	50	50
25 predictoren - P25 VI	25	8,5	48	52	46	54	46	54	48	52
25 predictoren - P10 VI	25	6,0	47	53	43	57	44	56	44	56
25 predictoren - P5 VI	25	5,5	47	53	43	57	45	55	46	54

Uit deze tabel blijkt dat elk van de vier subsets van de 25 predictoren tot betere imputeerprestaties leidt dan de set van 25 predictoren. Voor elk van de vier kenmerken van de kansverdeling van de schattingsfout van het jaarkengetal is de procentuele score namelijk hoger bij een subset van de 25 predictoren. De hoogste scores zijn daarbij voor de subsets gebaseerd op het 10-percentiel en het 5-percentiel van de VI's.

De conclusie van deze exercitie luidt dat het voor toepassing van Random Forest op RIWA-reeksen loont om een subset van predictoren te selecteren. Daarvoor komen het meest in aanmerking subsets gebaseerd op het positief zijn van het 10-percentiel of het 5-percentiel van de VI's.
Vervolgens zijn de prestaties van de drie imputeermethoden bij toepassing op RIWA-reeksen vergeleken voor 25 predictoren en voor de vier subsets daarvan, die zijn bepaald zoals boven beschreven.

Resultaten van de vergelijking van de drie imputeermethoden bij toepassing op RIWA-reeksen

Tabel 3.8 toont voor elk van de predictorensets de procentuele scores van de drie imputeermethoden, op basis van kenmerken van de kansverdeling van de imputeerfout bij toepassing op de 460 RIWA-reeksen. De beste procentuele score per drietal imputeermethoden is weer groen gekleurd, terwijl de slechtste oranje is gekleurd.

Tabel 3.8: Procentuele scores van de drie imputeermethoden op basis van kenmerken van de kansverdeling van de imputeerfout, bij 25 predictoren en bij vier selecties daarvan op basis van het positief zijn van respectievelijk de P50, P25, P10 en P5 van de verdeling van de VI.

	Gemidd.	RF	LR	NN	RF	LR	NN	RF	LR	NN	RF	LR	NN
Predictorselectie	#predictor	Mediaan			Gemiddelde			RMSE			Onnauwkeurigheid		
25 predictoren	25	22	41	37	48	24	28	58	25	17	50	32	18
P50 VI	13,5	24	38	38	45	29	25	54	34	12	46	39	16
P25 VI	8,5	26	37	37	42	33	25	51	39	10	43	43	14
P10 VI	6,0	26	37	36	43	34	23	50	42	9	43	46	11
P5 VI	5,5	26	36	38	45	34	21	50	42	8	44	44	12

Uit deze tabel blijkt dat Random Forest het beste scoort voor het gemiddelde, de RMSE en de onnauwkeurigheid van de imputeerfout, al zijn er voor het laatste kenmerk bij kleinere predictorensets nauwelijks verschillen met de scores voor lineaire regressie. Het neurale netwerk scoort over het algemeen het slechtst.

Tabel 3.9 toont voor elk van de predictorensets de procentuele scores van de drie imputeermethoden, op basis van kenmerken van de kansverdeling van de schattingsfout van het jaarkengetal bij toepassing op de 460 RIWA-reeksen.

Tabel 3.9: Procentuele scores van de drie imputeermethoden op basis van kenmerken van de kansverdeling van de schattingsfout van het jaarkengetal, bij 25 predictoren en bij vier selecties daarvan op basis van het positief zijn van respectievelijk de P50, P25, P10 en P5 van de verdeling van de VI.

Predictorselectie	Gemidd. #predictor	RF	LR	NN	RF	LR	NN	RF	LR	NN	RF	LR	NN
		Mediaan			Gemiddelde			RMSE			Onnauwkeurigheid		
25 predictoren	25	34	36	30	39	31	30	53	28	19	50	30	20
P50 VI	13,5	34	36	30	36	37	27	49	35	16	47	35	18
P25 VI	8,5	34	37	29	36	41	24	47	40	13	45	41	15
P10 VI	6,0	35	38	27	37	42	21	46	43	11	44	42	13
P5 VI	5,5	35	38	27	38	42	20	46	43	11	45	43	12

Uit deze tabel blijkt dat Random Forest het beste scoort voor de RMSE en de onnauwkeurigheid van de schattingsfout van het jaarkengetal, terwijl lineaire regressie doorgaans beter scoort voor mediaan en gemiddelde van die schattingsfout, al zijn de verschillen met Random Forest daar doorgaans gering. Het neurale netwerk scoort weer het slechtst.

Conclusies over de imputeerprestaties bij toepassing op RIWA-reeksen

Bij toepassing op de 460 praktijkreeksen uit RIWA-base blijkt Random Forest over het algemeen betere imputeerprestaties te hebben dan lineaire regressie. Het neurale netwerk presteert het minst van de drie beschouwde imputeermethoden.

3.3 Kernpunten selectie meest geschikte imputeermethode

- Bij het speuren naar de meest geschikte imputeermethode is op basis van literatuuronderzoek een lijst met kandidaten opgesteld, waarna met een eenvoudige vorm van multicriteria-analyse enkele methoden zijn geselecteerd die de meeste perspectieven boden. Dit waren Random Forest en meervoudige lineaire regressie. Voor de praktische methodenvergelijking is ook het neuraal netwerk meegenomen, als een soort maatstaf, aangezien daar al enige ervaring mee was opgedaan in een voorgaande studie. Met elk van deze methoden kunnen ontbrekende waarden in een reeks worden geschat door gebruik te maken van de relaties van deze reeks met andere reeksen. Daarbij kunnen Random Forest en het neuraal netwerk ook gebruik maken van niet-lineaire relaties, dit in tegenstelling tot lineaire regressie. Verder kunnen de twee eerstgenoemde methoden ook overweg met predictoren die onderling sterk gecorreleerd zijn, terwijl lineaire regressie dan schattingsproblemen heeft.
- Aansluitend zijn de imputeerprestaties van de drie methoden vergeleken, zowel bij toepassing op kunstmatig gesimuleerde reeksen als bij toepassing op een doorsnee van meetreeksen uit RIWA-base.
- Over het algemeen bleek uit de vergelijking dat Random Forest betere imputeerprestaties vertoonde dan lineaire regressie. Het neurale netwerk presteerde doorgaans het slechtst van de drie methoden. Maar hier past de kanttekening dat een pragmatische keuze is gedaan voor een breed toepasbare structuur van het neuraal netwerk, zodat mogelijk niet al zijn capaciteiten zijn benut. Een optimale inzet zou een onpraktische inspanning vergen.
- Een bijkomend voordeel van Random Forest is dat deze methode numeriek stabieler is dan meervoudige lineaire regressie, wat een voordeel is bij de beoogde automatische toepassing op het grote aantal meetreeksen van RIWA-base

4

Meest geschikte instellingen Random Forest

Bij het toepassen van Random Forest zijn er meerdere instelmogelijkheden. In dit hoofdstuk gaan we na wat de meest geschikte keuzen zijn voor de belangrijkste instelmogelijkheden, te weten het aantal predictoren, het aantal bomen, het aantal predictoren per knooppunt en het minimum aantal waarden per blad. Verder gaan we na wat een geschikte keuze is voor het maximum aantal imputanten van een vijfjaarsreeks met maandwaarden.

4.1 Aantal predictoren

In § 3.2.3 is al aangetoond dat het bij toepassing van Random Forest op praktijkreeksen van RIWA-base loont om de predictoren te selecteren op basis van de VI (Variable Importance). Wat daarbij de beste selectiemethode is, verdient echter nog nader onderzoek. In deze studie is alleen nog een selectiemethode verkend die bestaat uit de volgende stappen:

1. Selecteer als startset de 25 reeksen die het sterkst correleren met de te imputeren reeks. Dit zijn de reeksen met de grootste absolute waarde van de Spearman-rangcorrelatiecoëfficiënt.
2. Verwijder willekeurig waarden van de te imputeren reeks en bepaal voor elk van de 25 predictoren de VI. Herhaal dit 25 maal met steeds andere verwijderde waarden, zodat er van elke predictor een verzameling VI's resulteert.
3. Selecteer vervolgens díe predictoren waarvan het keuzepercentiel van de verzameling VI's groter is dan nul. Maar selecteer altijd minstens twee predictoren.

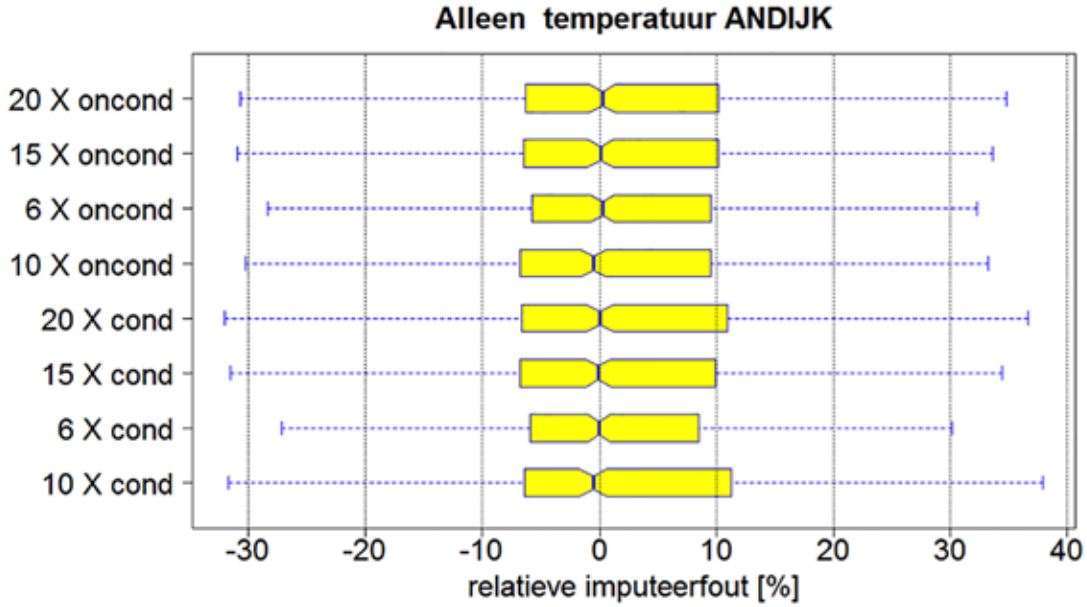
Het keuzepercentiel van stap 3 is achtereenvolgens gesteld op het 50-, 25-, 10 en 5-percentiel. De beste imputeerprestaties bleken daarbij voor de predictorensets gebaseerd op het 10-percentiel en het 5-percentiel van de VI's, met weinig verschil in prestaties tussen deze twee keuzepercentielen (zie § 3.2.4).
Maar er zijn ook selectiemethoden denkbaar, die uitgaan van een eenmalig bepaalde VI per predictor, waarbij vervolgens alle predictoren worden geselecteerd met een VI groter dan een bepaald percentage (zoals 5% of 10%) van de maximale VI. Een objectieve vergelijking van de verschillende selectiemethoden vergt echter een omvangrijke simulatiestudie.

4.1.1 Selectie predictoren op basis van conditionele VI?

Als predictoren onderling sterk zijn gecorreleerd (multicollineariteit) is er een risico dat hun VI's worden overschat. Dit kan worden voorkomen door de conditionele VI's te bepalen [Strobl et al., 2009]. Om na te gaan of predictorenselectie op basis van de conditionele VI tot betere imputeerprestaties leidt dan selectie op basis van de 'gewone' (onconditionele) VI zijn enkele simulatie-exercities uitgevoerd.

Bij de eerste exercitie is voor de recentste vijfjaarsreeks van de watertemperatuur van het IJsselmeer bij Andijk (met 60 maandwaarden) nagegaan in welke mate de imputeerprestaties worden beïnvloed door de predictorenset te selecteren op basis van de conditionele VI in plaats van de onconditionele VI. Door het sterke seizoenseffect is de watertempratuur namelijk sterk gecorreleerd aan een groot aantal predictoren, die ook onderling sterk gecorreleerd zijn.
Bij de exercitie zijn 49 maal, met steeds een maand verschuiving, 12 opeenvolgende waarden van de watertemperatuur verwijderd en vervolgens geïmputeerd met Random Forest, zowel met predictorenselectie op basis van de onconditionele, als de conditionele VI. Er is gestart met achtereenvolgens 20, 15, 10 en 6 predictoren. Voor het bepalen van de VI's zijn 200 bomen gebruikt, terwijl er voor het imputeren 30 bomen zijn gebruikt. In de literatuur wordt namelijk aanbevolen VI's te bepalen met veel bomen. Het aantal predictoren per knooppunt is gesteld op p/3, afgerond naar boven, waarbij p het aantal geselecteerde predictoren is. Tenslotte is het minimum aantal waarden per blad gesteld op 3. De resultaten zijn vermeld in onderstaande figuur 4.1

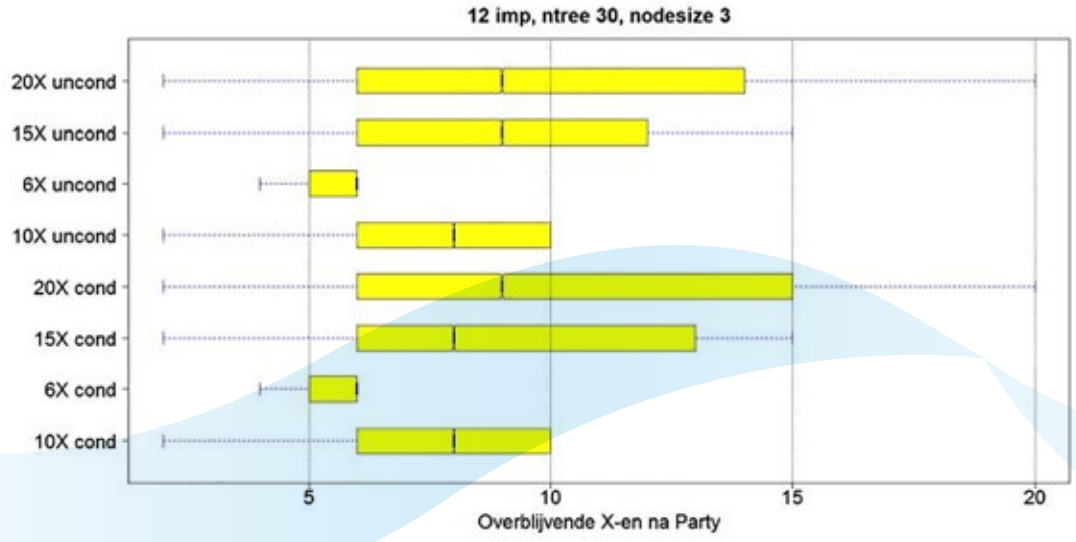
Figuur 4.1: Boxplot van de relatieve imputeerfout bij het imputeren van 12 van de 60 waarden van de recente vijfjaarsreeks van de watertemperatuur bij Andijk, voor verschillende startaantallen predictoren en voor predictorenselectie op basis van de onconditionele VI en de conditionele VI.De uitschieters zijn niet getoond.



Uit bovenstaande figuur blijkt dat de imputeerprestaties bij toepassing op de watertemperatuur bij Andijk nauwelijks veranderen door de predictoren te selecteren op basis van de onconditionele VI in plaats van op de conditionele VI.

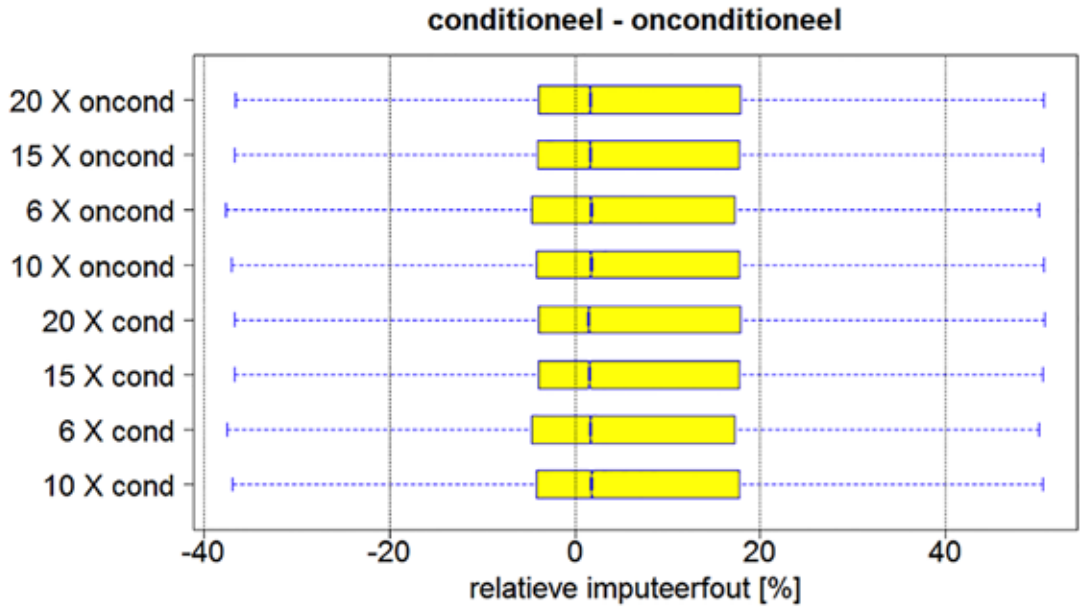
Figuur 4.2 toont de boxplot van het aantal resterende predictoren na selectie op basis van de VI (onconditioneel of conditioneel). Het verschil in aantal resterende predictoren van de twee benaderingen blijkt beperkt. In twee van de vier situaties levert selectie op basis van de onconditionele VI één predictor minder op.

Figuur 4.2: Boxplot van het aantal resterende predictoren na selectie op basis van de VI (onconditioneel of conditioneel).



De hiervoor beschreven exercitie is ook toegepast op de praktijkreeksen uit RIWA-base, exclusief de constante reeksen en de reeksen die alleen gecensureerde waarden bevatten. De resultaten over alle reeksen zijn samengevat in figuur 4.3, in de vorm van boxplots van de relatieve imputeerfout.

Figuur 4.3: Boxplot van de relatieve imputeerfout, na selectie op basis van de VI (onconditioneel of conditioneel), zoals bepaald over de praktijkset van vijfjarige RIWA-reeksen. De verticale as vermeldt ook het aantal predictoren van de startset. Uitschieters zijn niet getoond (de grootste zijn afkomstig van enkele (hydro)biologische parameters).



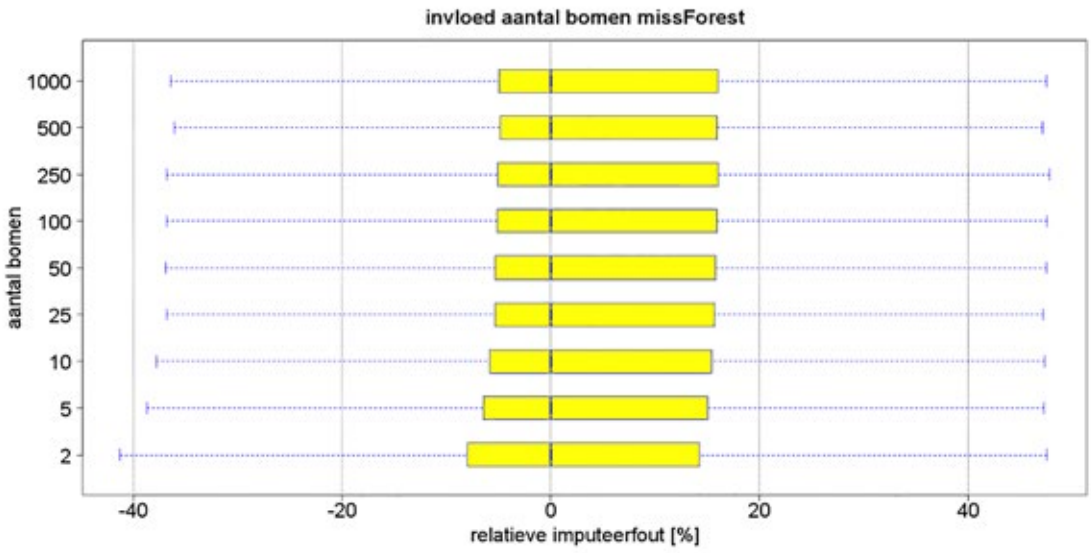
Uit bovenstaande figuur blijkt dat de imputeerprestaties bij toepassing op de praktijkreeksen van RIWA-base niet noemenswaardig veranderen door de predictoren te selecteren op basis van de conditionele VI in plaats van op de conditionele VI.

4.2 Aantal bomen

Experimenteel is de keuze van het aantal bomen (zo ook het aantal predictoren) moeilijk vast te stellen, doordat dit afhangt van het soort proces dat de te imputeren reeks genereert, de sterkte van de correlatie met de predictoren, hun onderlinge correlaties, het aantal missende waarden en het aantal gecensureerde waarden. Het aantal beslisbomen dient in ieder geval meer te zijn dan het aantal predictoren, zodat deze met grote zekerheid allemaal gebruikt worden in het trainingsalgoritme.

Om na te gaan wat de invloed is van het aantal bomen op de imputeerprestaties is een simulatie-exercitie uitgevoerd met de set praktijkreeksen van RIWA-base. Daarbij is van elke reeks met steeds een maand verschuiving een waarde verwijderd en vervolgens geïmputeerd met Random Forest, waarna over alle de gehele set de boxplot van de relatieve imputeerfout is bepaald. Dit is herhaald met verschillende aantallen bomen. De resultaten zijn weergegeven in figuur 4.4.

Figuur 4.4: Boxplot van de relatieve imputeerfout bij toepassing van Random Forest op de set vijfjarige praktijkreeksen uit RIWA-base, voor verschillende aantallen bomen. Uitschieters zijn niet getoond (de grootste zijn afkomstig van enkele (hydro)biologische parameters).



Uit bovenstaande figuur blijkt dat er bij meer dan 25 à 50 bomen geen duidelijke verbeteringen meer zijn in de imputeerprestaties. Maar aangezien de rekentijd wel nadelig wordt beïnvloed door het aantal bomen, stellen we voor om bij het imputeren van RIWA-base met Random Forest 30 bomen te hanteren.

4.3 Aantal predictoren per knooppunt

Bij elk knooppunt van een boom van Random Forest wordt aselect een aantal predictoren geselecteerd uit alle beschikbare predictoren, waaruit vervolgens de optimale splitsingspredictor en bijbehorende splitsingswaarde voor dat knooppunt wordt bepaald, zijnde de predictor met de splitsingswaarde die de grootste interne homogeniteit levert van de twee opgesplitste delen van de te imputeren reeks. In de literatuur wordt overwegend aanbevolen om het aantal predictoren per knooppunt te stellen op $p/3$, afgerond naar boven. Hierbij is p het totaal aantal predictoren.

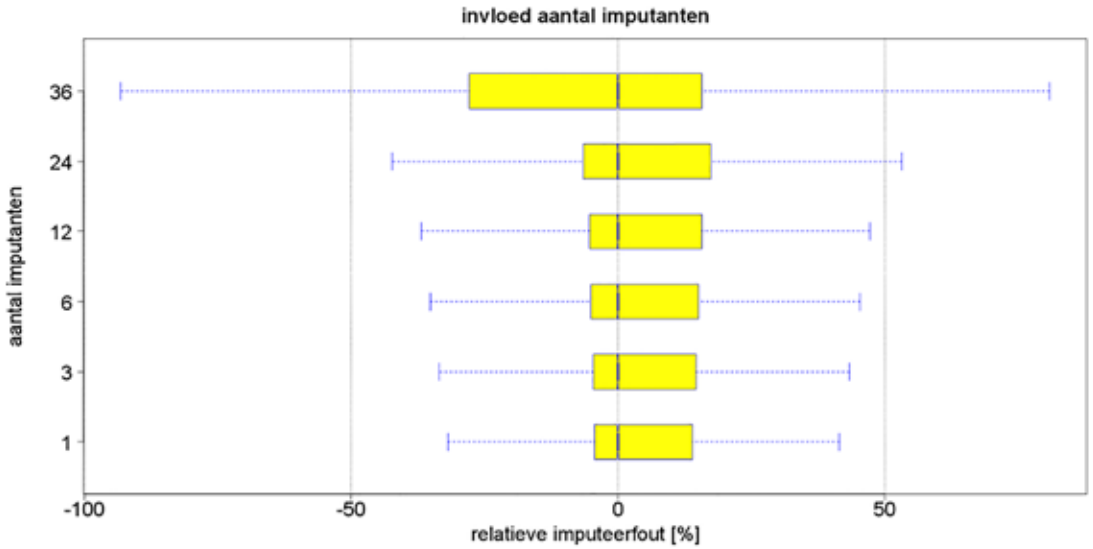
4.4 Minimum aantal waarden per blad

Bij Random Forest stopt het vertakken van een deel van de beslisboom als daardoor een eindpunt (blad) minder *Y*-waarden zou gaan bevatten dan het daarvoor ingestelde minimum. Dat minimum ligt doorgaans op 3, 4 of 5 waarden. Als een blad meer dan één *Y*-waarde bevat, geeft dat blad hun gemiddelde als schatting. We stellen voor om als minimum 3 waarden te hanteren, aangezien uit verkennende exercities met praktijkreeksen bleek dat de imputeerprestaties nauwelijks verschilden met 2, 3, 4 of 5 als grenswaarde.

4.5 Maximum aantal imputanten per reeks

Om na te gaan wat de invloed is van het aantal imputanten per reeks op de imputeerprestaties is een simulatie-exercitie uitgevoerd met de set praktijkreeksen van RIWA-base. Daarbij is van elke reeks met steeds een maand verschuiving een waarde verwijderd en vervolgens geïmputeerd met Random Forest, voor verschillende aantallen imputanten. Per aantal imputanten is vervolgens over de gehele set de verdeling van de relatieve imputeerfout bepaald. De resultaten zijn weergegeven in de boxplots van figuur 4.5.

Figuur 4.5: Boxplot van de relatieve imputeerfout bij toepassing van Random Forest op de set vijfjarige praktijkreeksen uit RIWA-base, voor verschillende aantallen imputanten per reeks van 60 waarden. Uitschieters zijn niet getoond (de grootste zijn afkomstig van enkele (hydro)biologische parameters).



Uit bovenstaande figuur blijkt dat pas bij meer dan 24 imputanten de imputeerprestaties duidelijk afnemen. De calibratieset (minder dan 36 maandwaarden) wordt dan blijkbaar merkbaar te klein.

4.6 Kernpunten meest geschikte instellingen Random Forest

- Start met de 25 predictoren die het best correleren met de te imputeren reeks (te beoordelen met de absolute waarde van de Spearman-rangcorrelatiecoëfficiënt) en selecteer vervolgens een subset daarvan op basis van een grenswaarde voor de VI.
- Pas Random Forest toe met 30 bomen.
- Gebruik $p/3$ predictoren per knooppunt, waarbij p het uiteindelijke aantal geselecteerde predictoren is.
- Stel random Forest in op minimaal 3 waarden per blad.
- Beperk het imputeren tot vijfjaarsreeksen op maandbasis, waarvan minstens één en hoogstens 12 maandwaarden ontbreken.

5

Procedure imputeren en beoordelen RIWA-base

Deze studie is gericht op het ontwikkelen van een procedure voor de volgende werkzaamheden, uit te voeren op de reeksen in RIWA-base:
1. imputeren ontbrekende maandwaarden;
2. beoordelen maandwaarden.

Dit hoofdstuk presenteert een voorstel voor de hoofdlijnen van deze procedure, dat is gebaseerd op de in de voorgaande hoofdstukken beschreven bevindingen van deze studie. In het onderstaande zijn de hoofdlijnen uitgesplitst naar de twee doeleinden van de procedure.

5.1 Voorstel hoofdlijnen imputeren ontbrekende maandwaarden

De voorgestelde stappen om ontbrekende maandwaarden te imputeren zijn de volgende:
1. Selecteer over de te beschouwen vijfjaarsperiode alle reeksen op maandbasis, waarvan minstens één en hoogstens twaalf waarden ontbreken.

De volgende stappen moeten dan per reeks worden uitgevoerd.
2. Selecteer voor de reeks 25 predictoren, elk met 60 waarden (dus zonder ontbrekende waarden), op basis van de grootte van de absolute waarde van de Spearman-rangcorrelatiecoëfficiënt.
3. Bepaal met Random Forest voor elk van de 25 predictoren de VI, met 200 bomen, 9 predictoren per knooppunt en met minimaal 3 waarden per blad aan het eind van elke laatste vertakking. Afhankelijk van de selectiemethode van de uiteindelijke predictorenset worden er per predictor één of meerdere VI's bepaald. Als er meerdere VI's worden bepaald, gebeurt dat door herhaald een willekeurige waarde uit de reeks weg te laten en daarbij de VI's van de predictoren te bepalen.
4. Selecteer vervolgens een subset van de predictoren. Vooralsnog wordt daarbij aanbevolen de predictoren te selecteren waarvan het 5-percentiel van de verzameling VI's groter is dan nul. Andere selectiemethoden zijn ook denkbaar. Maar er dienen minstens twee 2 predictoren te worden geselecteerd.
5. Imputeer de ontbrekende waarden van de reeks met Random Forest, met 30 bomen, $p/3$ predictoren per knooppunt, naar boven afgerond (p is het aantal uiteindelijk geselecteerde predictoren voor de betreffende reeks) en met minimaal 3 waarden per blad.
6. Sla de imputanten op in RIWA-base, met de bijbehorende OOB-fout en tevens met een label dat het een imputant betreft. De OOB-fout wordt hierbij uitgedrukt als standaardafwijking van de imputeerfout (in de meetschaal).

Aandachtspunten

Bij het implementeren van deze procedure zullen ook nog praktijkaspecten aan de orde komen, zoals de te gebruiken software, de wijze van koppeling met RIWA-base, de wijze van opslag van de geïmputeerde waarde (en bijbehorende maat voor de betrouwbaarheid daarvan) en de wijze van presentatie van resultaten van statistische analyses van meetreeksen waarvan één of meer waarden zijn geïmputeerd.

5.2 Voorstel hoofdlijnen beoordelen maandwaarden

De voorgestelde stappen om maandwaarden te beoordelen zijn de volgende:

1. Selecteer over de te beschouwen vijfjaarsperiode alle reeksen op maandbasis, waarvan geen of hooguit twaalf waarden ontbreken.

De volgende stappen moeten dan per reeks worden uitgevoerd.

2. Selecteer voor de reeks 25 predictoren, elk met 60 waarden (dus zonder ontbrekende waarden), op basis van de grootte van de absolute waarde van de Spearman-rangcorrelatiecoëfficiënt.
3. Bepaal met Random Forest voor elk van de 25 predictoren de VI, met 200 bomen, 9 predictoren per knooppunt en met minimaal 3 waarden per blad aan het eind van elke laatste vertakking. Afhankelijk van de selectiemethode van de uiteindelijke predictorenset worden er per predictor één of meerdere VI's bepaald. Als er meerdere VI's worden bepaald, gebeurt dat door herhaald een willekeurige waarde uit de reeks weg te laten en deze vervolgens te imputeren.
4. Selecteer vervolgens een subset van de predictoren. Vooralsnog wordt daarbij aanbevolen de predictoren te selecteren waarvan het 5-percentiel van de verzameling VI's groter is dan nul. Andere selectiemethoden zijn ook denkbaar. Maar er dienen minstens twee 2 predictoren te worden geselecteerd.

Doe de volgende stappen voor elke waarde van de reeks.

5. Verwijder de waarde en imputeer deze vervolgens met Random Forest, met 30 bomen, $p/3$ predictoren per knooppunt, naar boven afgerond (p is het aantal uiteindelijk geselecteerde predictoren voor de betreffende reeks) en met minimaal 3 waarden per blad.
6. Beoordeel uit het verschil tussen de geïmputeerde waarde en de werkelijke waarde of de werkelijke waarde al of niet als verdacht kan worden aangemerkt. Betrek hierbij ook de bijbehorende OOB-fout van het imputeren.

Aandachtspunten

Ook bij het implementeren van deze procedure zullen nog praktijkaspecten aan de orde komen. Zo is te verwachten dat deze procedure bewerkelijk is, omdat hiervoor veel reeksen in aanmerking zullen komen, waarvan achtereenvolgens alle waarden moeten worden verwijderd en vervolgens geïmputeerd. Mogelijk werkt dat niet meer met MS Access en moet daarvoor een ander platform worden gebruikt, zoals Mysql.
Verder zullen er nog criteria voor de beoordeling van het verschil tussen de geïmputeerde waarde en de werkelijke waarde moeten worden opgesteld.

Het is aan te bevelen deze beoordelingsprocedure pas te implementeren, als de imputeerprocedure zich in de praktijk voldoende heeft bewezen.

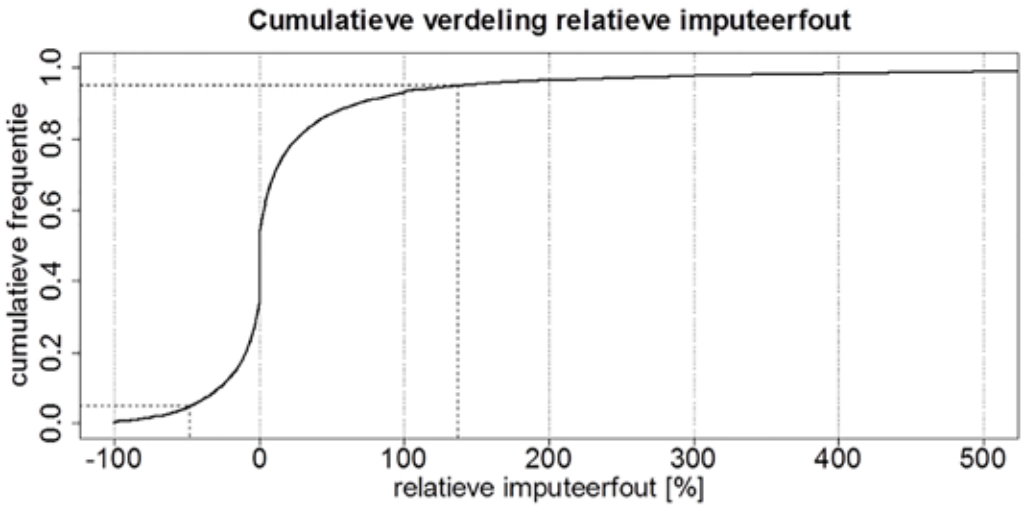
6 Imputeren met Random Forest: enkele voorbeelden

Dit hoofdstuk toont een aantal praktijk voorbeelden van het imputeren van meetreeksen uit RIWA-base met Random Forest.

6.1 Toepassing op praktijkreeksen uit RIWA-base

Om een eerste empirische indruk te krijgen van de nauwkeurigheid die haalbaar is in de praktijk, is Random Forest toegepast op de set praktijkreeksen uit RIWA-base, waarbij per reeks 48 maal (met steeds een maand verschuiving) een voortschrijdende deelreeks van 12 aaneengesloten waarden is verwijderd en vervolgens geïmputeerd. Daarbij is Random Forest ingesteld met 25 predictoren, 30 bomen, 8 predictoren per knooppunt en minimaal 3 waarden per blad. De relatieve imputeerfout bleek daarbij te liggen tussen -100%[7] en 133.410%, met P5=-48,5%, P10=-27,7%, P25=-5,3%, P50=0%, P75=15,7%, P90=68,0% en P95=136,5%. De cumulatieve verdeling is weergegeven in figuur 6.1.

Figuur 6.1: Cumulatieve verdeling van de relatieve imputeerfout bij toepassing op de set praktijkreeksen uit RIWA-base. Daarbij is Random Forest ingesteld met 25 predictoren, 30 bomen, 8 predictoren per knooppunt en minimaal 3 waarden per blad. De grootste uitschieters zijn afkomstig van enkele (hydro)biologische parameters.



[7] *Het minimum van de relatieve imputeerfout bedraagt -100%, aangezien negatieve meetwaarden vrijwel niet voorkomen. Random Forest levert immers alleen imputanten die binnen het bereik van de meetwaarden van de betreffende parameter liggen, zodat de relatieve imputeerfout, gedefinieerd als 100%·(imputant – waarde)/waarde niet lager kan zijn dan -100%.*

6.2 Imputeerprestaties uitgesplitst naar de parametergroepen

Als de relatieve imputeerfouten worden uitgesplitst naar de parametergroepen, blijkt dat er aanzienlijke verschillen kunnen optreden (zie figuur 6.2). Blijkbaar zijn bepaalde parametergroepen veel beter te imputeren dan andere.

Figuur 6.2: Boxplot van de relatieve imputeerfout per parametergroep, bij toepassing op de set vijfjarige praktijkreeksen van maandwaarden uit RIWA-base. Daarbij is Random Forest ingesteld met 25 predictoren, 30 bomen, 8 predictoren per knooppunt en minimaal 3 waarden per blad. De grootste uitschieters zijn niet getoond (die zijn afkomstig van enkele (hydro)biologische parameters).

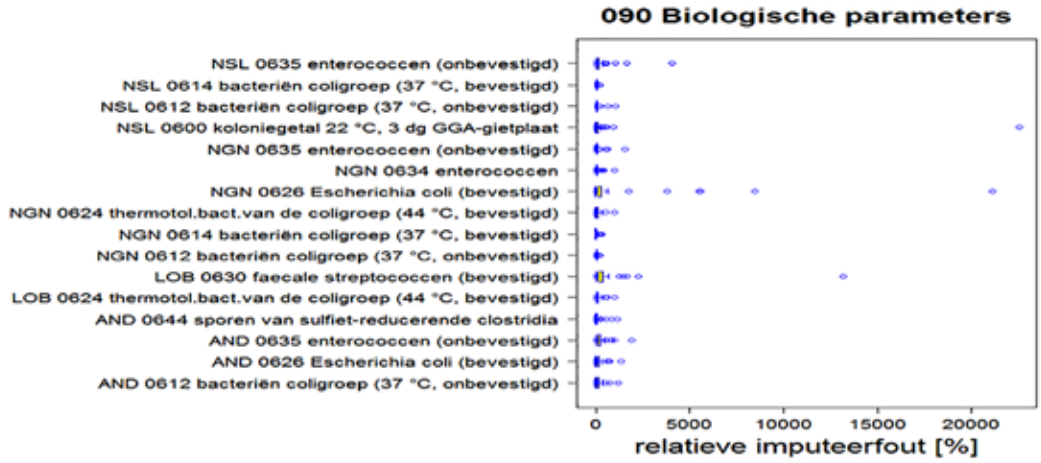
Uit bovenstaande figuur blijkt bijvoorbeeld dat de klassieke parameters (weergegeven in het onderste deel) over het algemeen betrouwbaarder zijn te imputeren dan biologische en hydrobiologische parameters en een aantal organische microparameters. Dit wordt vermoedelijk veroorzaakt doordat er bij de klassieke parameters minder uitschieters voorkomen. Het kan ook komen doordat er bij de klassieke parameters doorgaans minder meetfouten optreden.

In de nu volgende subparagrafen 6.2.1 t/m 6.2.3 wordt specifiek ingezoomd op enkele parametergroepen waar bij het imputeren relatief grote imputeerfouten kunnen optreden.

6.2.1 Imputeerprestaties voor biologische parameters

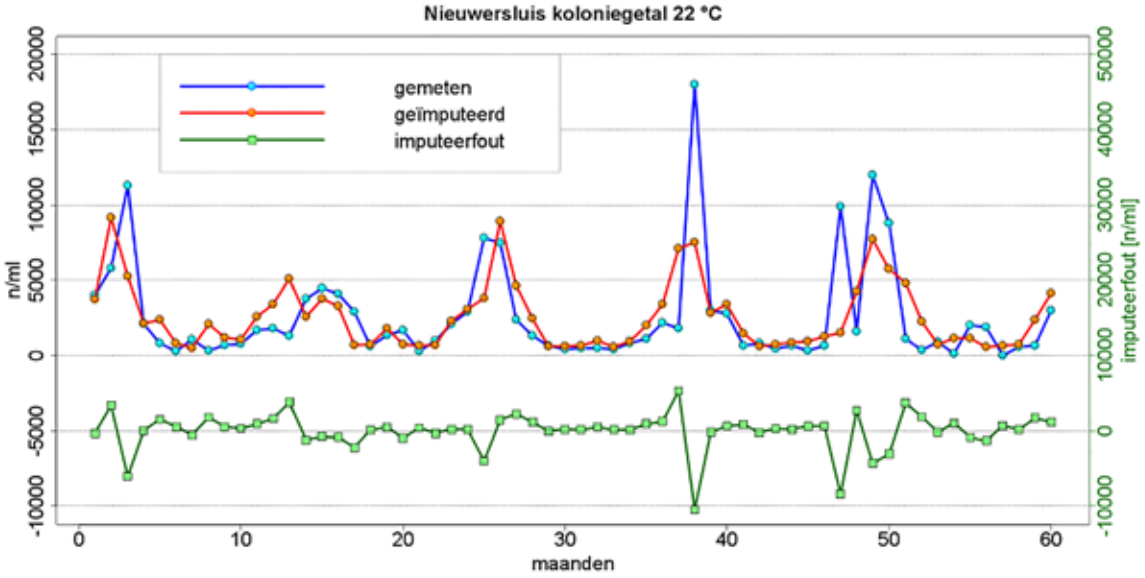
In figuur 6.3 is de boxplot van de relatieve imputeerfout weergegeven per biologische parameter.

Figuur 6.3: Boxplot van de relatieve imputeerfout per biologische parameter, bij toepassing op de set praktijkreeksen van maandwaarden uit RIWA-base.



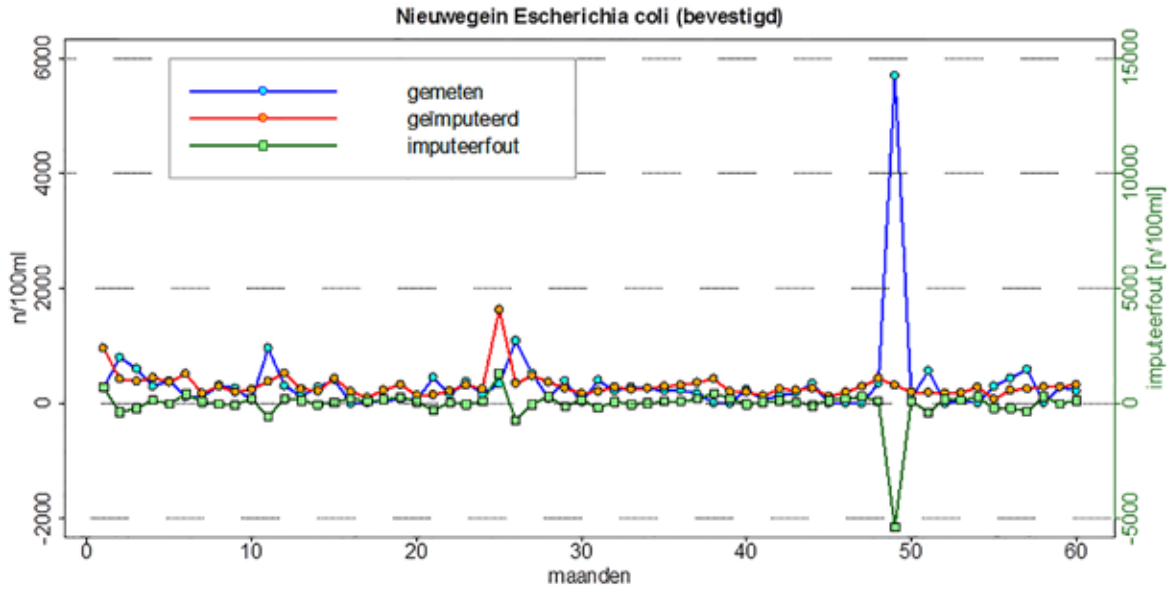
Hieronder zijn de imputeerprestaties voor de twee volgens figuur 6.3 minst betrouwbaar te imputeren biologische parameterreeksen afzonderlijk weergegeven.

Figuur 6.4: Vergelijking van gemeten en geïmputeerde vijfjaarsreeks van maandwaarden van het koloniegetal 22 °C te Nieuwersluis. Elke waarde is verwijderd en vervolgens geïmputeerd.



De geïmputeerde reeks volgt wel min of meer het patroon van de meetreeks, maar de uitschieters zijn niet goed geïmputeerd.

Figuur 6.5: Vergelijking van gemeten en geïmputeerde vijfjaarsreeks van maandwaarden van Escherichia coli (bevestigd) te Nieuwegein. Elke waarde is verwijderd en vervolgens geïmputeerd.

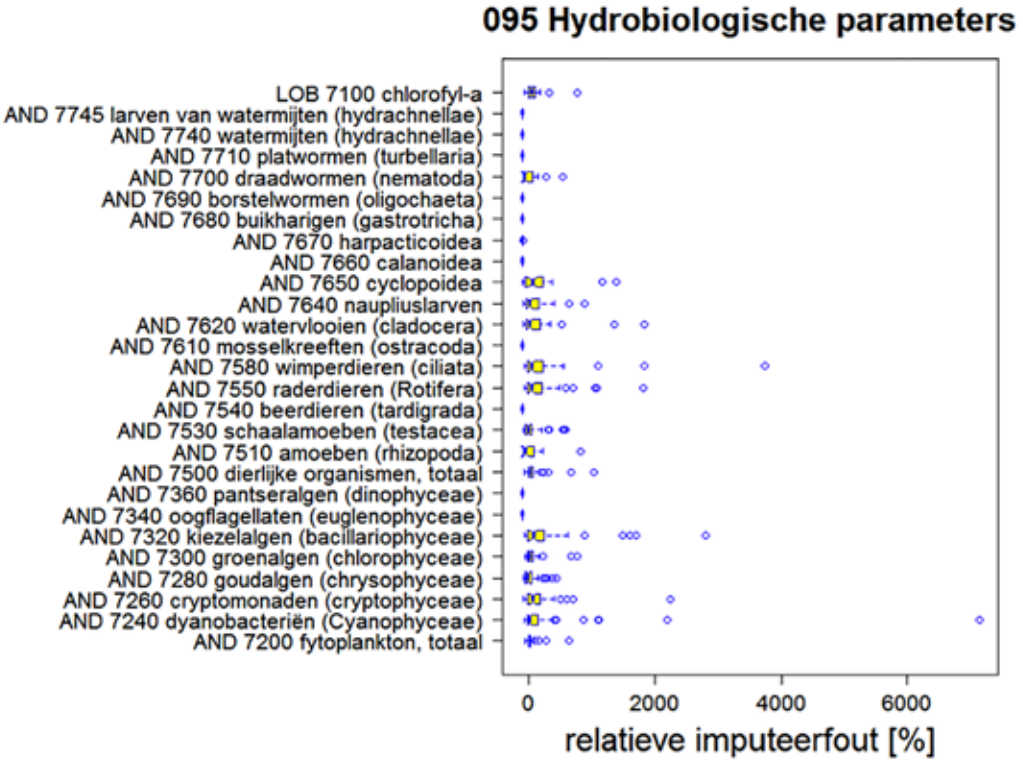


Ook voor deze reeks is de uitschieter niet goed geïmputeerd.

6.2.2 Imputeerprestaties voor hydrobiologische parameters

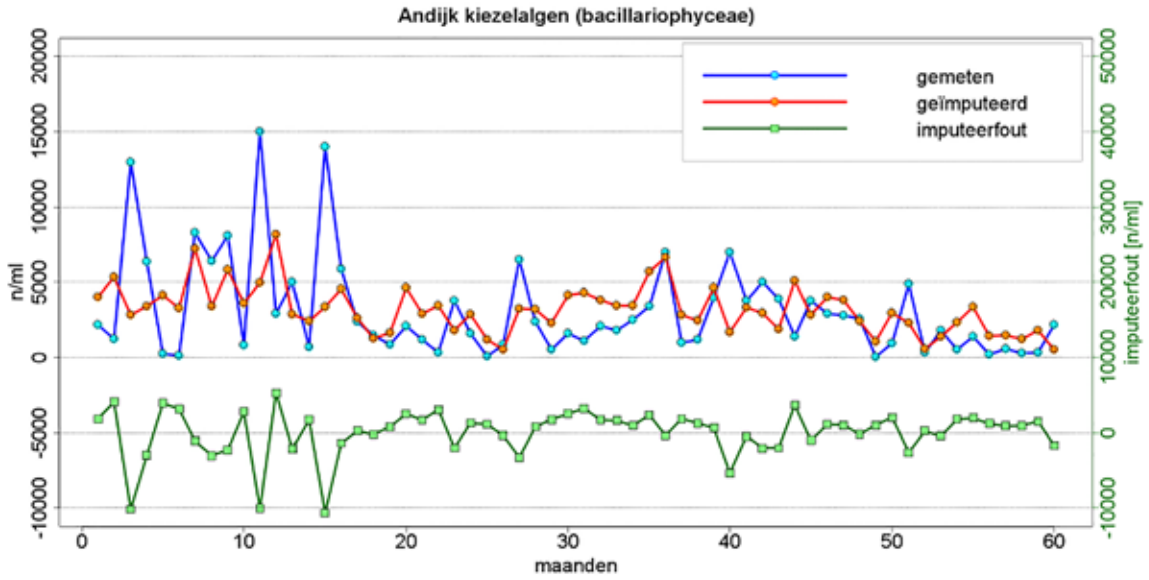
In figuur 6.6 is de boxplot van de relatieve imputeerfout weergegeven per hydrobiologische parameter.

Figuur 6.6: Boxplot van de relatieve imputeerfout per hydrobiologische parameter, bij toepassing op de set praktijkreeksen van maandwaarden uit RIWA-base.



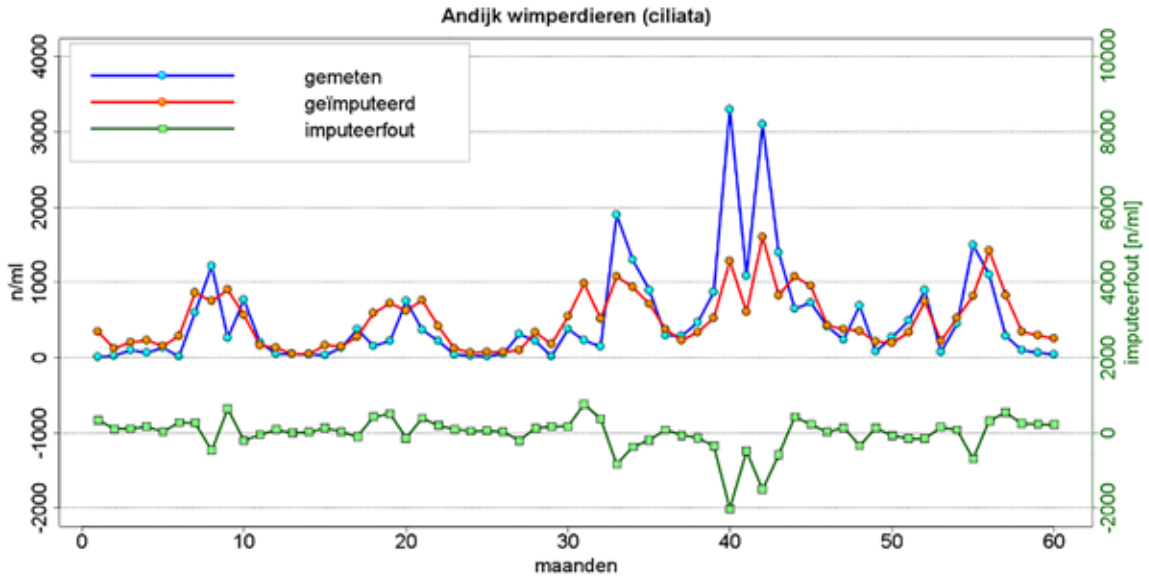
Hieronder zijn de imputeerprestaties voor twee hydrobiologische parameterreeksen afzonderlijk weergegeven.

Figuur 6.7: Vergelijking van gemeten en geïmputeerde vijfjaarsreeks van maandwaarden van kiezelalgen te Andijk. Elke waarde is verwijderd en vervolgens geïmputeerd.



Het niveau wordt doorgaans redelijk geïmputeerd, maar de drie uitschieters worden niet goed geïmputeerd. Blijkbaar bevat geen van de predictoren daar informatie over.

Figuur 6.8: Vergelijking van gemeten en geïmputeerde vijfjaarsreeks van maandwaarden van wimperdieren te Andijk. Elke waarde is verwijderd en vervolgens geïmputeerd.

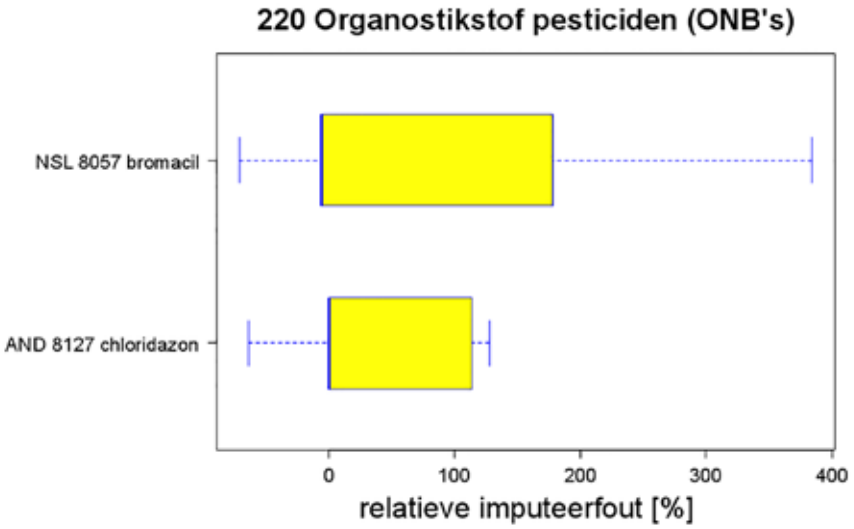


Het patroon van deze reeks wordt door het imputeren goed gereproduceerd, alleen is het ook hier niet mogelijk om de uitschieters goed te imputeren.

6.2.3 Imputeerprestaties voor organostikstof-pesticiden

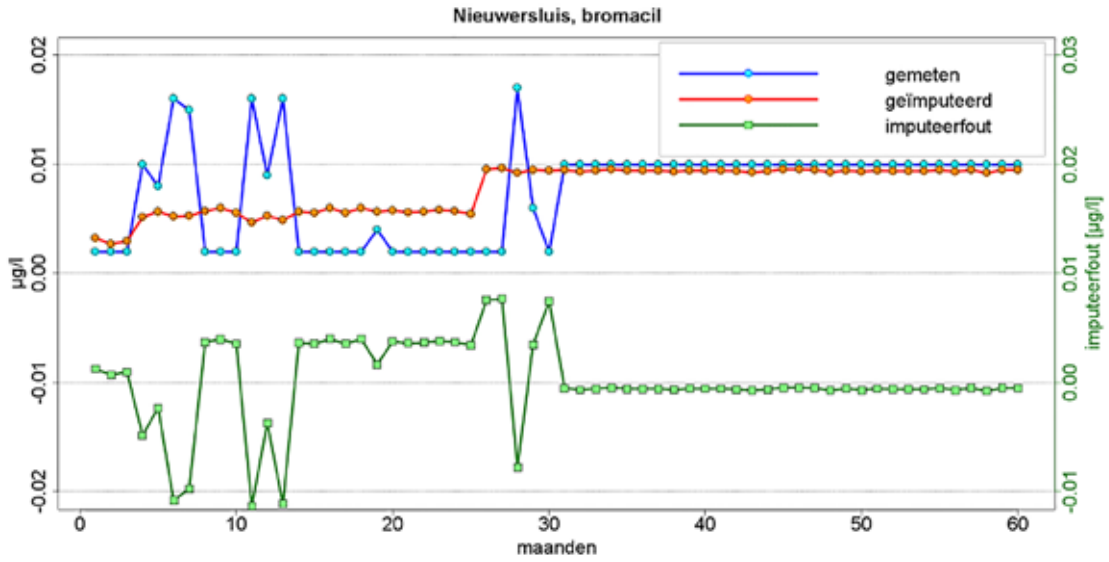
In figuur 6.9 is de boxplot van de relatieve imputeerfout weergegeven per organostikstof-pesticide, met daaronder ook een uitwerking van de bromacilreeks van Nieuwersluis (figuur 6.9).

Figuur 6.9: Boxplot van de relatieve imputeerfout per organostikstof-pesticide, bij toepassing op de set praktijkreeksen van maandwaarden uit RIWA-base.



Gezien de doorgaans zeer lage niveaus van deze parameters zijn de relatieve imputeerfouten nog niet extreem groot.

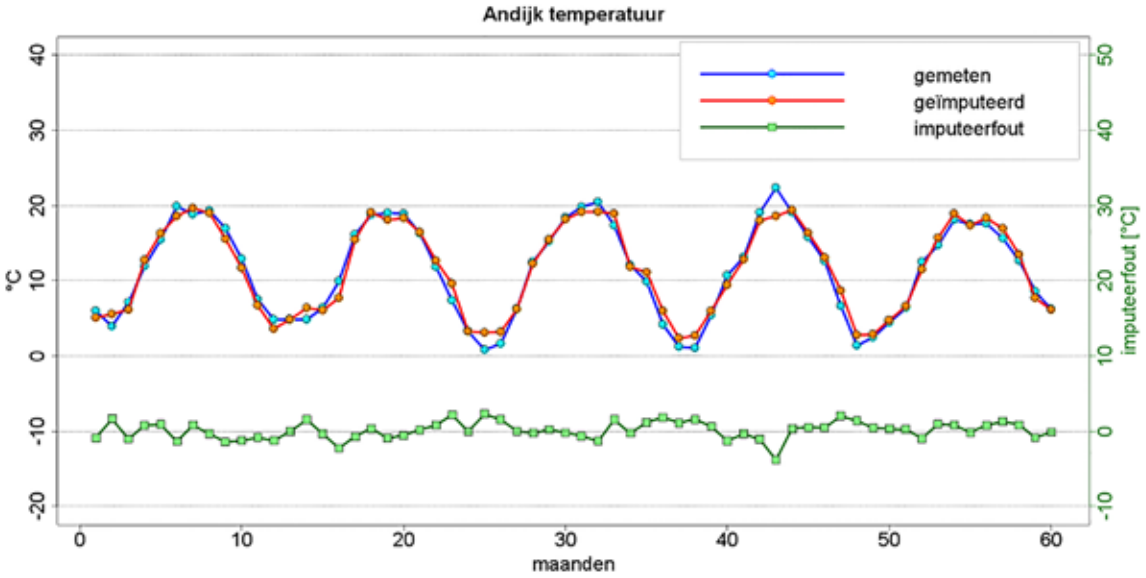
Figuur 6.10: Vergelijking van gemeten en geïmputeerde vijfjaarsreeks van maandwaarden van bromacil te Nieuwersluis. Elke waarde is verwijderd en vervolgens geïmputeerd.



Uit figuur 6.10 blijkt dat de bromacilconcentratie de eerste 2,5 jaar nog zowel lager als hoger dan 0,01 µg/l lag, maar vanaf medio 2009 is constant <0,02 µg/l opgegeven. De analysemethode is toen gewijzigd. Daarvoor lag de rapportagegrens lager, namelijk op 0,005 µg/l. De imputaties volgen wel enigszins het algehele niveau, maar kunnen de pieken en dalen niet reproduceren, al zijn de imputeerfouten in de meetschaal slechts zeer gering.

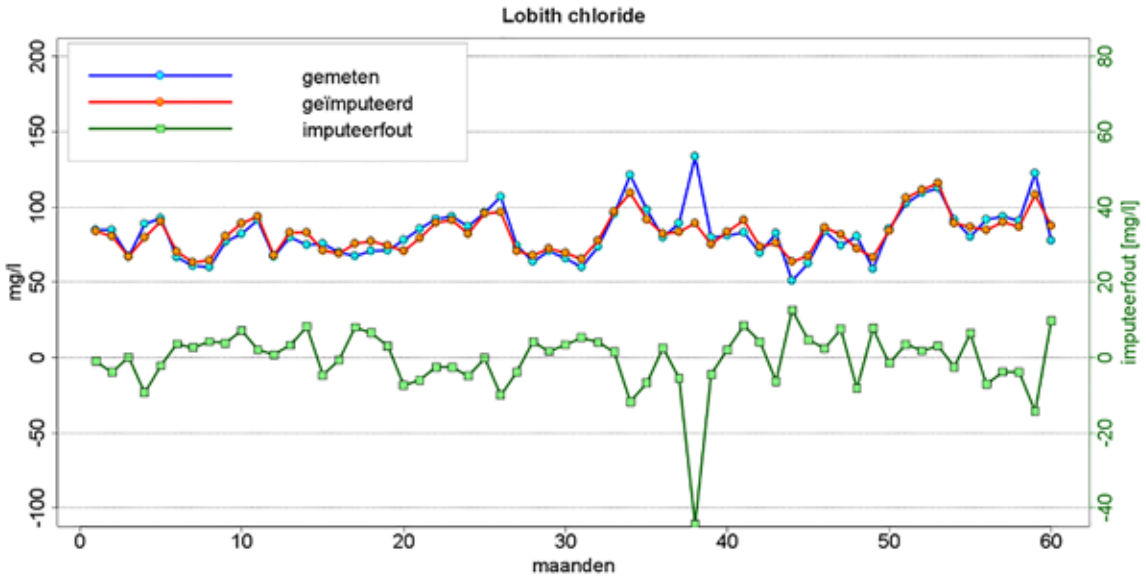
6.2.4 Imputeerprestaties voor enkele willekeurige parameters

Figuur 6.11: Vergelijking van gemeten en geïmputeerde vijfjaarsreeks van maandwaarden van watertemperatuur te Andijk. Elke waarde is verwijderd en vervolgens geïmputeerd.



De imputanten van de watertemperatuur te Andijk zijn zeer betrouwbaar. Dit zal komen doordat deze parameter sterk is gecorreleerd aan een groot aantal andere parameters en daardoor goed is te beschrijven uit predictoren.

Figuur 6.12: Vergelijking van gemeten en geïmputeerde vijfjaarsreeks van maandwaarden van chloride te Lobith. Elke waarde is verwijderd en vervolgens geïmputeerd.

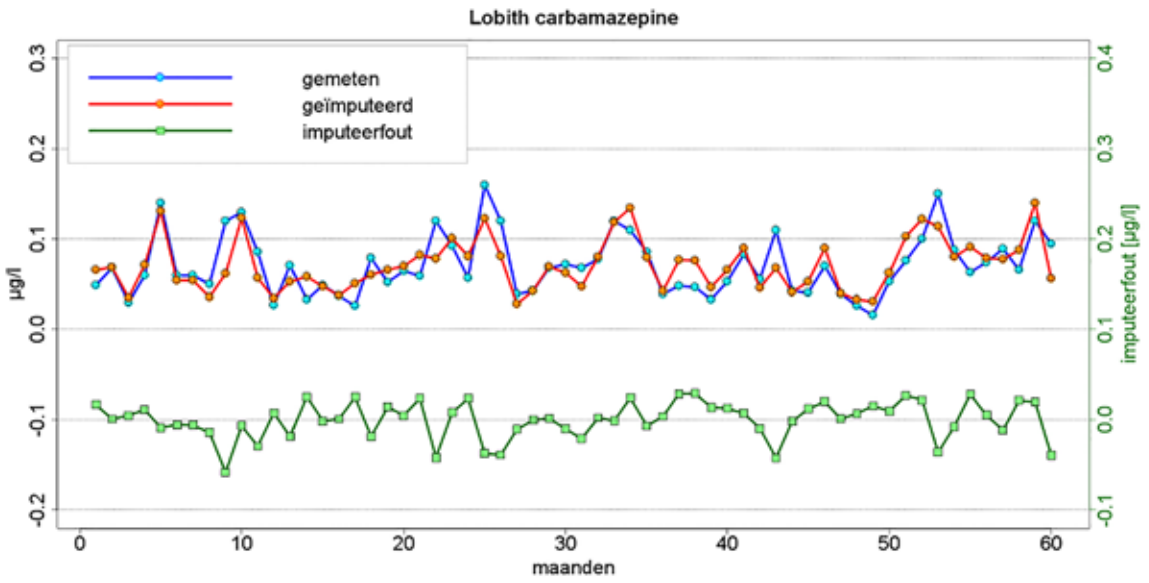


De chlorideconcentratie blijkt over het algemeen zeer goed te imputeren, echter met uitzondering van de ene uitschieter (februari 2010). Maar als we de individuele meetwaarden van chloride in die periode vergelijken met die van EGV (zie tabel 6.1), dan blijkt dat er vermoedelijk sprake is van een fout van het laboratorium bij de bepaling van chloride. De afwijkende chloridewaarde van 24 februari 2010 gaat namelijk niet gepaard met een afwijkende EGV-waarde. Dit is een treffende illustratie van de mogelijkheid om via het imputeren ook waarden te beoordelen.

Tabel 6.1: Analyseresultaten van EGV (elektrisch geleidingsvermogen) en chlorideconcentratie van de Rijn te Lobith, van 27 januari t/m 24 maart 2010.

Monsterdatum	EGV (ms/m)	Chloride (mg/l)
27-01-2010	64	97,5
10-02-2010	62	94,5
24-02-2010	74	172,3
10-03-2010	53	80,7
24-03-2010	58	78,9

Figuur 6.13: Vergelijking van gemeten en geïmputeerde vijfjaarsreeks van maandwaarden van carbamazepine te Lobith. Elke waarde is verwijderd en vervolgens geïmputeerd.



Deze reeks blijkt goed te imputeren. De imputeerfouten blijven beperkt tot minder dan 0,05 µg/l en de imputanten benaderen de metingen goed.

Aangehaalde en geraadpleegde literatuur

Baggelaar, P.K., Van der Meulen, E.C.J. en Smits, A.H. (2012): Verkenning van relaties tussen meetreeksen van het RIWA-meetnet. Icastat, juni 2012, 29 blz.

Box, G.E.P. and Jenkins, G.M. (1976): Time Series Analysis, Forecasting and Control. Holden Day, San Francisco.

Breiman, L., et al. (1993): Classification and Regression Trees. Chapman & Hall, Boca Raton, 1993.

Breiman, L. (2001): Random Forests. Statistics Department, University of California, Berkeley, January 2001.

Grömping, U. (2009): Variable Importance Assessment in Regression: Linear Regression versus Random Forest. American Statistical Association, Vol.63, No.4, The American Statistician, 2009.

Honaker, J., King, G. and Blackwell, M. (2012): Amelia II: A program for missing data. Version 1.6.2, May 2, 2012, 57 blz.

Latinne, P. (2001): Limiting the Number of Trees in Random Forests. Presented to the 2nd International Workshop on Multiple Classifier Systems - MCS'2001, Cambridge, UK.

Mitchell, W.M. (2011): Bias of the Random Forest Out-of-Bag (OOB) Error for certain input parameters. Metabolon Inc., Research Triangle Park, New Caledonia, USA, 2011.

Montillo, A. (2009): Random Forest; guest lecture: statistical foundation of Data Analysis. University of Pennsylvania, Computer Science, February 2009.

Smits, A. en Baggelaar, P.K. (2010): Estimating missing values in time series. RIWA, September 2010, 20 pp.

Sorjamaa, A., Merlin, P., Maillet, B. and Lendasse, A. (2007): SOM+EOF for Finding Missing Values. European Symposium on Artificial Neural Networks proceedings, 25-27 April 2007, pp. 115 – 120.

Strobl, C., Hothorn, T. and Zeileis, A. (2009): Party on! A New, Conditional Variable Importance Measure for Random Forests Available in the party Package. Technical Report Number 050, 2009, Institut für Statistik, Ludwig-Maximilians-Universität München, 4 blz.

Sutton, C. D. (2005): Classification and Regression Trees. Handbook of Statistiscs, Vol. 24. Elsevier, 2005.

Svetnik, V. (2003): Regression Tool Compound Classification and QSAR Modeling. J. Chem. Inf. Comput. Sci. 2003, 43, 1947-1958, 2003.

Van Buuren, S. and Groothuis-Oudshoorn, C. G. M. (2011): mice: Multivariate Imputation by Chained Equations. Journal of Statistical Software, 45(3), 1–67.

Bijlage 1: Resultaten multicriteria-analyse voor selectie imputeermethode

Bij de multicriteria-analyse zijn de onderstaande acht methoden beoordeeld.

1. NN-MBP: (kunstmatig) neuraal netwerk. MBP (een afkorting van Multiple Back-Propagation) is een open source softwareapplicatie. Neurale netwerken behoren tot het werkveld Machinaal Leren[8], waarbinnen systemen worden ontwikkeld die automatisch kunnen leren uit gegevens. Ze worden vooral toegepast voor het modelleren van complexe, niet-lineaire systemen. De relaties tussen de invoer en uitvoer van het systeem worden als het ware 'geleerd' door het automatisch iteratief aanpassen van interne instellingen van het netwerk, zodanig dat één of andere functie (doorgaans is dit de kwadraatsom) van de afwijkingen van modelwaarden en uitvoerwaarden van de trainingsset geminimaliseerd wordt. Het principe en de structuur van een NN heeft veel analogieën met het neurale netwerk van de mens. Als de relaties tussen invoer en uitvoer zijn aangeleerd, dan kunnen eenvoudig ontbrekende uitvoerwaarden worden geschat, door de bijbehorende invoerwaarden aan te bieden aan het NN.

2. NN-R: (kunstmatig) neuraal netwerkmodule in de programmeertaal R (eveneens open source).

3. SOM+EOF: SOM staat voor Self Organising Maps, een niet-lineaire projectiemethode en een soort NN, terwijl EOF staat voor Empirical Orthogonal Functions, een lineaire projectiemethode [Sorjamaa et al., 2007].

4. Random Forest - missForest-R: Random Forest behoort net als NN tot het werkveld Machinaal Leren. Het is op te vatten als een niet-lineaire regressiemethode en wordt ook wel aangeduid als een sequentiële partitie-methode, die gebruik maakt van een groot aantal beslisbomen. Een toelichting op het principe van de werking van Random Forest vindt u in § 3.2.2 en een formelere beschrijving vindt u in bijlage 2. Hier is de Random Forest-applicatie missForest beschouwd, een module in de programmeertaal R.

5. BJ+interv: Box-Jenkins-modellering met interventies. De Box-Jenkins-tijdreeksmodellering stamt uit de jaren zeventig [Box and Jenkins, 1976] en richt zich op het beschrijven van een tijdreeks (de uitvoerreeks) als lineaire functie van zijn eigen verleden en/of van andere, aan de uitvoerreeks gerelateerde tijdreeksen (de invoerreeksen). Uit de geschatte modelparameters kan de grootte en de statistische significantie van de relatie tussen de uitvoerreeks en een invoerreeks worden bepaald. Als de uitvoerreeks ontbrekende waarden bevat, dan kunnen deze worden geschat door voor elke ontbrekende waarde ook een zogenaamde interventiereeks in het model op te nemen. Deze interventiereeks is even lang als de uitvoerreeks en heeft de waarde nul voor de meettijdstippen waar de uitvoer-reeks een waarde heeft en de waarde één voor het meettijdstip waar de betreffende uitvoerwaarde ontbreekt. De geschatte modelparameters van deze interventiereeksen vormen dan de imputanten van de uitvoerreeks.

6. mtsdi-R: de module mtsdi (in de programmeertaal R) maakt voor het imputeren zowel gebruik van correlaties in de tijd als in de ruimte.

7. Imputation-R: Dit pakket kan naar keuze verschillende imputeermethoden toepassen op ontbrekende waarden, zoals vervangen door het gemiddelde, k-nearest neighbor, singular value decomposition (SVD), Singular Value Thresholding en lineaire imputatie. Bij de multicriteria-analyse van § 3.1 is van deze mogelijkheden alleen SVD beoordeeld voor toepassing op RIWA-base.

8. IRMI-R: IRMI staat voor Iterative Robust Model-based Imputation en is gebaseerd op meervoudige lineaire regressie. Het is geïmplementeerd in het imputatiepakket VIM, in de programmeertaal R. Het maakt gebruik van een EM-algoritme (expectation-maximization) om de regressieparameters te schatten, waarna de imputanten met het model kunnen worden geschat.

---

[8] *In de Engelstalige literatuur aangeduid als 'Machine Learning'.*

Onderstaande tabel vermeldt de scores en totaalscores die de acht methoden kregen bij de multicriteria-analyse.

	Aspect	Betrouwbaarheid resultaten te toetsen	Te automatiseren	Kosten software	Uiteindelijke hoeveelheid werk	Te doorgronden	Maakt gebruik van zoveel mogelijk informatie	Vergt nog maar weinig uitzoekwerk	Geschikt gebleken voor hetzelfde soort gegevens als in RIWA-base	Betrouwbare bron	Mogelijkheid om te loggen	Imputanten kenbaar te maken in RIWA-base	Geeft kwantitatieve maat voor succes imputeerproces	Totaalscore
	gewicht	8,50	9,75	6,00	9,25	8,75	7,50	5,75	7,00	7,25	5,00	4,50	8,00	
	rel.gew.	0,10	0,11	0,07	0,11	0,10	0,09	0,07	0,08	0,08	0,06	0,05	0,09	
NN-MBP	score	7	0	10	1	10	5	6	10	8	2	2	0	
	rg x s	0,68	0,00	0,69	0,11	1,00	0,43	0,40	0,80	0,66	0,11	0,10	0,00	4,99
BJ+interv	score	7	5	10	3	10	5	6	5	8	7	2	2	
	rg x s	0,68	0,56	0,69	0,32	1,00	0,43	0,40	0,40	0,66	0,40	0,10	0,18	5,83
NN-R	score	7	7	10	5	10	5	6	8	8	2	2	0	
	rg x s	0,68	0,78	0,69	0,53	1,00	0,43	0,40	0,64	0,66	0,11	0,10	0,00	6,03
SOM+EOF	score	8	6	10	7	4	8	4	9	8	2	2	8	
	rg x s	0,78	0,67	0,69	0,74	0,40	0,69	0,26	0,72	0,66	0,11	0,10	0,73	6,57
mtsdi-R	score	7	5	10	5	10	7	4	8	8	2	2	8	
	rg x s	0,68	0,56	0,69	0,53	1,00	0,60	0,26	0,64	0,66	0,11	0,10	0,73	6,58
Imputation-R	score	8	8	10	5	8	8	4	6	8	2	2	8	
	rg x s	0,78	0,89	0,69	0,53	0,80	0,69	0,26	0,48	0,66	0,11	0,10	0,73	6,74
IRMI-R	score	8	8	10	7	8	8	6	8	8	2	2	7	
	rg x s	0,78	0,89	0,69	0,74	0,80	0,69	0,40	0,64	0,66	0,11	0,10	0,64	7,15
RF-mF-R	score	8	8	10	7	8	8	6	7	8	2	2	8	
	rg x s	0,78	0,89	0,69	0,74	0,80	0,69	0,40	0,56	0,66	0,11	0,10	0,73	7,17

Bijlage 2: Toelichting op lineaire regressie, Random Forest en het neurale netwerk

Principe van lineaire regressie

Bij lineaire regressie wordt gebruik gemaakt van een model dat de van belang zijnde variabele (Y) beschrijft als lineaire functie van één of meer daaraan gerelateerde verklarende variabelen (X_1, X_2, ..., X_p), hierna aangeduid als de predictoren, volgens:

$$y_i = b_0 + b_1 \cdot x_{1i} + b_2 \cdot x_{2i} + ... + b_p \cdot x_{pi} + r_i \qquad [1]$$

waarin y_i de i-de waarde van Y (i=1, 2, ..., N), b_0 de interceptparameter, b_1, b_2, ..., b_p de modelparameters voor respectievelijk X_1, X_2, ..., X_p en x_{1i}, x_{2i}, ..., x_{pi} de i-de waarde van respectievelijk X_1, X_2, ..., X_p en tenslotte r_i het i-de modelresidu.

De parameters van dit model worden (doorgaans) geschat met de Kleinste-Kwadraten-methode, door de volgende kwadraatsom van de modelresiduën (KSr) te minimaliseren:

$$KSr = \sum_{i=1}^{N} \left(y_i - \left(b_0 + b_1 \cdot x_{1i} + b_2 \cdot x_{2i} + ... + b_p \cdot x_{pi} \right) \right)^2 = \sum_{i=1}^{N} r_i^2 \qquad [2]$$

Als de modelresiduën afkomstig zijn uit dezelfde normale kansverdeling en ook onderling onafhankelijk zijn, kunnen er op basis van het model kwantitatieve betrouwbaarheidsuitspraken worden gedaan over de relaties van Y met elk van de predictoren, evenals over voorspellingen van Y.

Uitgaande van de met het model beschreven statistische relaties tussen Y en de predictoren kunnen ook ontbrekende waarden van Y worden geïmputeerd. Het is daarbij overigens niet noodzakelijk dat de modelresiduën voldoen aan bovenstaande voorwaarden.

Principe van Random Forest

Random Forest [Breiman, 1993, 2001] is momenteel een van de populairste methoden uit het onderzoeksveld Machinaal Leren en wordt ook wel aangeduid als een niet-lineaire regressiemethode, of als een sequentiële partitiemethode. Zoals de bloemrijke naam al doet vermoeden bestaat Random Forest uit een verzameling beslisbomen ('decision trees'), waarbij een beslisboom een schema is dat de verschillende mogelijke trajecten naar een bepaald doel weergeeft. Bij Random Forest wordt een aantal (n) beslisbomen gecreëerd met 'bagging' (een afkorting van bootstrap-aggregatie, zijnde een bepaalde manier om aselect elementen uit een verzameling te trekken).

Uitgangspunt vormt de dataset T, bestaande uit N waarden van de van belang zijnde variabele Y en p predictoren X (X_1, X_2, ..., X_p), ook elk met N waarden. We kunnen dit formeel uitschrijven als: $T = \{Y, X\} = \{(y_1; x_{11}, x_{21}, ..., x_{p1}); ...; (y_N; x_{1N}, ..., x_{pN})\}$.

Eerst worden uit de dataset T met bootstrapping n datasets samengesteld (T_1, T_2, ..., T_n). De bootstrapping houdt in dat telkens N willekeurige combinaties ($y_i; x_{1i}, ..., x_{pi}$) via verloting worden geselecteerd uit T, met teruglegging. Deze datasets zijn dus even groot als de volledige dataset T, maar door de trekking met teruglegging komen meerdere combinaties ($y_i; x_{1i}, ..., x_{pi}$) niet voor, terwijl andere combinaties juist herhaald voorkomen. Uit kansrekening volgt dat gemiddeld circa 63% van de combinaties in een bootstrapping-dataset belandt, de overige circa 37% betreft dan één of meer herhalingen van die combinaties. De bootstrapping-dataset wordt de trainingsset of ook wel 'In-Bag'-selectie genoemd, terwijl de resterende niet geselecteerde data wordt aangeduid als de 'Out-of-Bag'-selectie, of kortweg de OOB-selectie.

Voor elke afzonderlijke trainingsset (T_i) wordt iteratief een aparte beslisboom afgeleid. Bij elke vertakking (knooppunt of 'node') van een beslisboom worden uit de p predictoren aselect m predictoren geselecteerd en van elk van deze wordt vervolgens de optimale splitsingswaarde voor de vertakking bepaald, zijnde de waarde van de betreffende verklarende variabele X die Y uitsplitst in twee zo homogeen mogelijke delen. De variabele X die dit het best bewerkstelligt wordt dan de splitsingsvariabele voor die bepaalde vertakking.

Het vinden van de optimale splitsingswaarde van een predictor X geschiedt door het minimaliseren van een doelfunctie. We gaan hier uit van de situatie dat Y bestaat uit numerieke waarden. Random Forest geeft dan een schatting van y_i bij een bepaalde set waarden (x_{1i}, ..., x_{pi}) van de predictoren X, als het gemiddelde van de n schattingen van de n beslisbomen. Als Y daarentegen een nominale variabele is (zoals kleur, politieke partij, geslacht, etc.), dan schat Random Forest y_i als de modus van de n schattingen van de n beslisbomen (dit is de waarde die het vaakst voorkomt). Het schatten van een nominale variabele is van belang bij classificatie-vraagstukken.

In het onderstaande lichten we het vertakken toe en geven vervolgens een voorbeeld van een beslisboom.

Stel dat bij een bepaalde vertakking van een beslisboom aselect m predictoren zijn geselecteerd uit de p beschikbare predictoren. Vind nu díe splitsingswaarde s_j van één van deze variabelen X_j (j=1,...m), die de variabele Y in de twee homogeenste delen splitst. Dit komt neer op het minimaliseren van de volgende doelfunctie L_j van kwadratische afwijkingen van Y-waarden ten opzichte van het gemiddelde van de Y-waarden in dat betreffende deel:

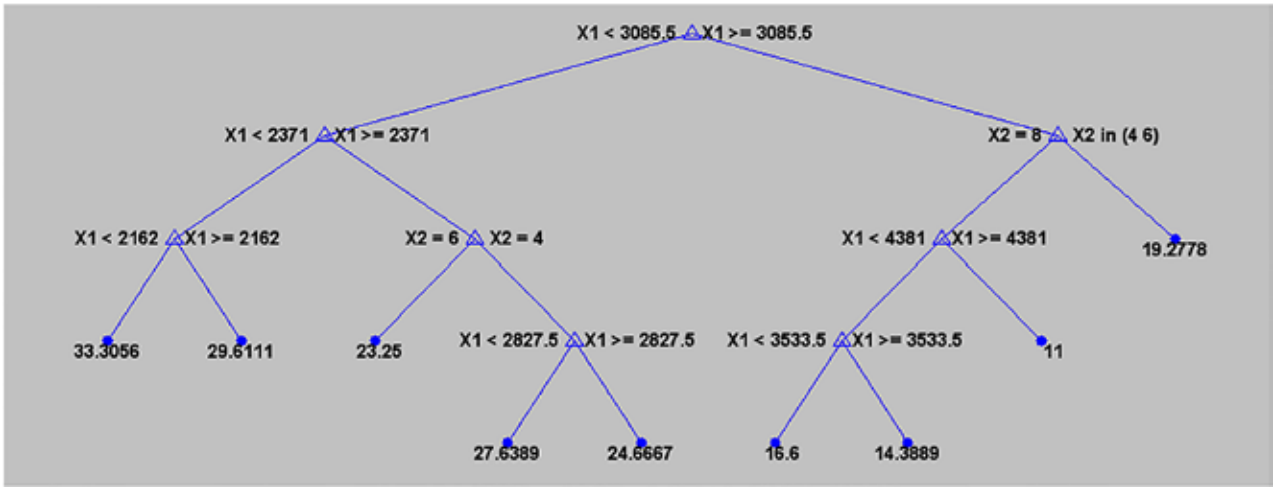
$$L_j = \sum_{x_{ji} \le s_j} \left(y_i - \left(\overline{y} \mid x_{ji} \le s_j \right) \right)^2 + \sum_{x_{ji} > s_j} \left(y_i - \left(\overline{y} \mid x_{ji} > s_j \right) \right)^2 \qquad [3]$$

met $(\overline{y} \mid x_{ji} \le s_j)$ het gemiddelde van het deel van de variabele Y dat correspondeert met de waarden van X_j die kleiner of gelijk zijn aan s_j en met $(\overline{y} \mid x_{ji} > s_j)$ het gemiddelde van het deel van Y dat correspondeert met de waarden van X_j die groter zijn dan s_j.

De minimalisatie van de doelfunctie levert bij elke vertakking in een beslisboom de optimale splitsingsvariabele en de bijbehorende optimale splitsingswaarde op. Dit bewerkstelligt dat de variabele Y bij die vertakking in twee zo homogeen mogelijke delen wordt uitgesplitst.
Het vertakken van een deel van de beslisboom stopt als daardoor een eindpunt (blad) minder Y-waarden zou gaan bevatten dan het daarvoor ingestelde minimum. Bij Random Forest staat dat minimum doorgaans op 3, 4 of 5 waarden. Als een blad meer dan één Y-waarde bevat, geeft dat blad hun gemiddelde als schatting.

Figuur b2.1 toont een voorbeeld van een eenvoudige beslisboom met twee predictoren (p=2), waarbij elke vertakking is gebaseerd op één predictor (m=1).

Figuur b2.1: Voorbeeld van een beslisboom (m=1, p=2). Bij elk knooppunt is de splitsingswaarde gegeven van de predictor X_1 (hier een continue variabele) of X_2 (hier een discrete variabele, met waarden 4, 6 en 8). Als de boom niet verder vertakt kan worden, resulteert de schatting van de variabele Y, zoals bepaald door deze beslisboom.



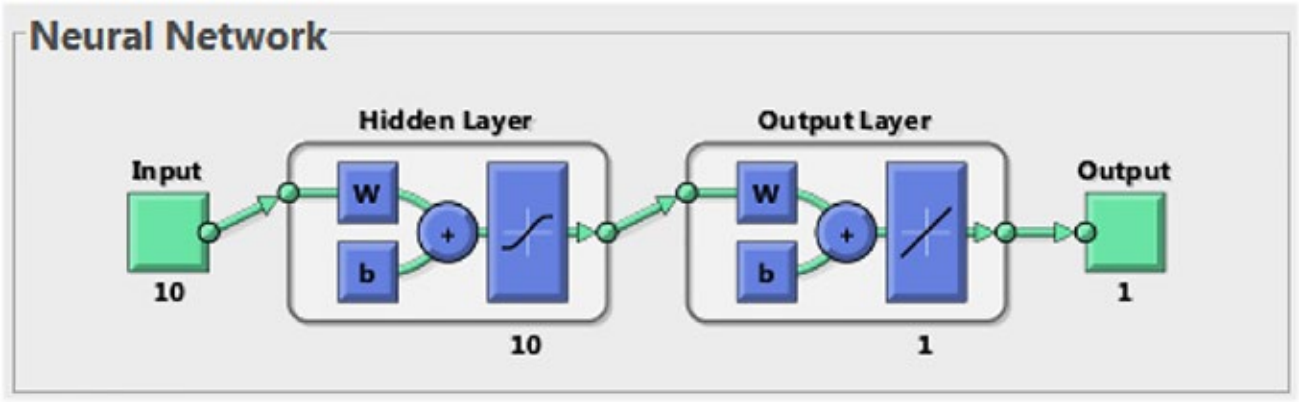
De uiteindelijk door Random Forest bepaalde geschatte waarde van y_i bij de bijbehorende combinatie (x_{1i}, x_{2i}, ..., x_{pi}) is het rekenkundig gemiddelde van de schattingen van de n beslisbomen, waarbij elke afzonderlijke schatting een waarde is die afkomstig is uit de dataset van Y. Daarmee kan een geschatte waarde niet kleiner zijn dan het minimum van Y en niet groter dan het maximum van Y. Het kan er zelfs toe leiden dat de variatie in een reeks enigszins wordt afgevlakt.

De niet gebruikte combinaties (y_i; x_{1i}, ..., x_{pi}) worden de 'Out-Of-Bag'-selectie (OOB-selectie) genoemd. Deze selectie wordt gebruikt om de kansverdeling van de voorspelfout empirisch af te leiden. Daartoe worden voor elke boom de y-waarden van zijn OOB-selectie voorspeld, met behulp van de bijbehorende (x_{1i}, ..., x_{pi})-waarden. Vervolgens worden alle voorspellingen van een bepaalde y_i-waarde gemiddeld over de betrokken beslisbomen en vergeleken met de werkelijke waarde y_i. Dit schept een groot voordeel ten opzichte van andere methoden, aangezien daarmee bij elke imputatie ook een maat kan worden geleverd voor zijn nauwkeurigheid.

Toelichting op neuraal netwerk

In de begintijd van het Machinaal Leren – en ruim vóór de introductie van Random Forest - werden vooral neurale netwerken gebruikt voor het beschrijven van complexe, niet-lineaire processen. In deze studie gebruiken we bij het empirisch vergelijken van enkele imputeermethoden (zie § 3.2) een neuraal netwerk met een aantal predictoren en twee lagen, namelijk een laag met 10 sigmoïde neuronen, die een niet-lineaire relatie kunnen beschrijven en een uitvoerlaag met één lineair neuron, zoals geschematiseerd in onderstaande figuur b2.2.

Figuur b2.2: Voorbeeld van een in deze studie toegepast neuraal netwerk, hier met 10 invoerreeksen en één uitvoerreeks.



Volgens de Matlab-specificaties heeft dit algemene 'feedforward'-neurale netwerk goede eigenschappen voor het beschrijven en voorspellen van data en daarmee voor het imputeren van missende waarden.

Een nadeel van een neuraal netwerk is dat de keuze voor de dimensionering niet vastligt, er zijn namelijk vele mogelijkheden voor de keuze van overdrachtfuncties, het aantal lagen en neuronen. Hier hebben we pragmatisch gekozen voor 10 neuronen in de eerste laag, gezien de door ons gebruikte datasets (maximaal 25 predictoren van 5 jaar met 12 waarden per jaar). Het bleek onpraktisch (te arbeidsintensief) om per situatie een daarop optimaal afgestemd neuraal netwerk te ontwerpen en toe te passen. De in deze studie beschreven prestaties van het neurale netwerk kunnen de daadwerkelijk haalbare prestaties daarom onderschatten. Bij een onjuiste keuze van het aantal neuronen - te weinig dan wel te veel - is er gauw sprake van het overschatten van hoge waarden en het onderschatten van lage waarden. Bovendien is bij te weinig neuronen de convergentie naar optimale parameterwaarden van het neurale netwerk moeizaam. In tegenstelling tot lineaire modellen zijn neurale netwerken daarentegen aanzienlijk robuuster bij een eventuele overparametrisering.

Bijlage 3: Formele beschrijving kunstmatige reeksen

De imputeerprestaties van lineaire regressie, Random Forest en het neurale netwerk zijn beoordeeld bij toepassing op elf kunstmatige processen (zie daarvoor § 3.2.3). Deze processen zijn hieronder gekenschetst.

1) Lineair proces, predictoren normaal verdeeld, normaal verdeelde ruis

$$Y_i = \sum_{k=1}^{10} X_{ki} + c + 2 \cdot a_i \quad \text{voor } i = 1, 2, \ldots, 60$$

Elke X_{ki} is hierbij een aselecte trekking uit een normale kansverdeling met gemiddelde 10 en standaardafwijking 1 en elke a_i is hierbij een aselecte trekking uit een standaardnormale kansverdeling met gemiddelde 0 en standaardafwijking 1. De c is een constante om het gemiddelde van elke reeks op 100 te krijgen. Het betreft een lineair proces.

2) Niet-lineair proces ($1/X$), predictoren normaal verdeeld, normaal verdeelde ruis

$$Y_i = \sum_{k=1}^{10} \frac{1}{X_{ki}} + c + 2a_i \quad \text{voor } i = 1, 2, \ldots, 60$$

3) Niet-lineair proces (X^3), predictoren normaal verdeeld, normaal verdeelde ruis

$$Y_i = \sum_{k=1}^{10} (X_{ki})^3 + c + 2a_i \quad \text{voor } i = 1, 2, \ldots, 60$$

4) Niet-lineair proces ($\sqrt{X}$), predictoren normaal verdeeld, normaal verdeelde ruis

$$Y_i = \sum_{k=1}^{10} \sqrt{X_{ki}} + c + 2a_i \quad \text{voor } i = 1, 2, \ldots, 60$$

5) Lineair proces, predictoren normaal verdeeld, normaal verdeelde ruis met autocorrelatie

$$Y_i = \sum_{k=1}^{10} X_{ki} + c + 2N_i \quad \text{voor } i = 1, 2, \ldots, 60$$
$$N_i = 0{,}5N_{i-1} + a_i$$

6) Lineair proces, predictoren normaal verdeeld, lognormaal verdeelde ruis met autocorrelatie

$$Y_i = \sum_{k=1}^{10} X_{ki} + c + 2\frac{e^{N_i}}{5{,}8747} \quad \text{voor } i = 1, 2, \ldots, 60$$
$$N_i = 0{,}5N_{i-1} + a_i$$

7) Lineair proces, predictoren normaal verdeeld, uniform verdeelde ruis, met autocorrelatie

$$Y_i = \sum_{k=1}^{10} X_{ki} + c + 2N_i \quad \text{voor } i = 1, 2, \ldots, 60$$
$$N_i = 0{,}5N_{i-1} + u_i$$

De u_i is hierbij echter geen aselecte trekking uit een normale kansverdeling, maar uit een uniforme kansverdeling met het bereik [0,1].

8) Niet-lineair proces (e^X), predictoren normaal verdeeld, lognormaal verdeelde ruis met autocorrelatie

$$Y_i = \sum_{k=1}^{10} e^{X_{ki}} + c + 2\frac{e^{N_i+1}}{5{,}8747} \quad \text{voor } i = 1, 2, \ldots, 60$$

$$N_i = 0{,}5N_{i-1} + a_i$$

Elke X_{ki} is hierbij een aselecte trekking uit een normale kansverdeling met gemiddelde 2 en standaardafwijking 1.

9, 10 en 11) Lineaire processen met seizoenseffecten, predictoren normaal verdeeld met verschillende seizoenseffecten, normaal verdeelde ruis.
Dit zijn processen die ontstaan door Fourierreeksen te sommeren.

$$Y_i = c - 0{,}6235\cos(i\pi/6) - 1{,}3501\sin(i\pi/6) - 1{,}1622\cos(i\pi/3) - 0{,}9443\cdot\sin(i\pi/3) + 0{,}8a_i$$

met

$$X_1 = \sin(iw_1),\ X_2 = \sin(2iw_2),\ X_3 = \cos(iw_3),\ X_4 = \cos(2iw_4),\ X_5 = \sin(3iw_5),$$
$$X_6 = \sin(4iw_6),\ X_7 = \cos(3iw_7),\ X_8 = \cos(4iw_8),\ X_9 = \sin(iw_9) + \sin(2iw_9) \text{ en}$$
$$X_{10} = \cos(iw_{10}) + \cos(2iw_{10})$$

Elke w_k is hierbij een trekking uit een normale kansverdeling met gemiddelde $\pi/6$ en standaardafwijking 1. Bij proces 9 is er sprake van periodiciteit in de *Y*- en *X*-reeksen, maar doordat de frequenties duidelijk kunnen verschillen is een *Y*-reeks niet direct gerelateerd met de *X*-reeksen, zoals in de processen 1 t/m 8 wel duidelijk het geval is. Proces 10 is identiek aan proces 9, maar dan met een kleinere standaardafwijking (0,1). Proces 11 is identiek aan proces 9 en 10 maar dan met een nog kleinere standaardafwijking (0,01), zodat er een sterker verband is tussen de *Y*- en de *X*-reeksen.

Colofon

Auteurs Aart Smits (eauQstat)
Eit van der Meulen (AMO)
Gerrit van de Haar (RIWA)
Paul Baggelaar (Icastat, eindredactie)

Fotografie Laurens Hitman; www.hitmanphotography.com

Uitgever RIWA, Vereniging van Rivierwaterbedrijven

Vormgeving Meyson Company, Zaandam

ISBN/EAN 978-90-6683-156-8

Dit rapport is gratis te downloaden via de website www.riwa.org. Desgevraagd kan het worden geprint na bestelling via riwa@riwa.org (printing on demand). Voor de kosten hiervan wordt verwezen naar onze website.

RIWA

Vereniging van
Rijnwaterbedrijven

RIWA-Rijn
Groenendael 6
3439 LV Nieuwegein
T +31306009030
F +31306009039
E riwa@riwa.org
W www.riwa.org