KM 44 (1993)
pg. 21-34



# Niet-parametrische regressie: neuraal of statistisch?

Dee Denteneer*
Philips Natuurkundig Laboratorium
Eindhoven

29 Maart 1993

**Abstract**

Neurale netwerken vormen een belangrijke bijdrage tot de ontwikkeling van de regressiemethoden. Minstens even belangrijk, maar helaas veel minder bekend, zijn de niet-parametrische, adaptieve regressiemethoden die binnen de statistiek ontwikkeld worden onder mysterieuze acroniemen als CART en MARS. Dit artikel vraagt aandacht voor niet-parametrische regressiemethoden van zowel neurale als statistische origine. Het zijn flexibele methoden met vele toepassingen binnen de statistiek. Daarnaast presenteert het artikel enig vergelijkend warenonderzoek tussen enerzijds neurale netwerken en anderzijds de statistische methoden. Deze vergelijking leidt tot enige relativering van de soms wat overspannen verwachtingen ten aanzien van de toepassing van de besproken technieken.

*Gebaseerd op een voordracht op 14 december 1992 in het colloquium "Stochastiek en neurale netwerken" aan de UvA. Adres: Philips Natuurkundig Laboratorium; Prof. Holstlaan 4; 5656 AA Eindhoven. Tel: 040-744767. E-mail: dentenee@prl.philips.nl.

# 1 Inleiding

Toepassingen van neurale netwerken lijken alomvertegenwoordigd, of het nu binnen de industriële, medische of financiële wereld is. Deze succesvolle toepassingen gaan enerzijds gepaard met een ongebreideld optimisme over verdere toepassingen binnen welk gebied dan ook en met wat voor gegevens dan ook. Anderzijds gaan deze technieken gepaard met een dédain ten aanzien van de statistiek, waarvan bijvoorbeeld het volgende citaat uit Hecht-Nielsen (1990, geciteerd uit Ripley, 1993) getuigt.

> In essence, in terms of everyday practice, there has been only modest progress in regression analysis since the days of Gauss.

De statistiek heeft weinig vooruitgang geboekt sinds de dagen van Gauss (1777-1855) en heeft niets meer te bieden dan een lineair model voor het het schatten van een regressiefunctie. De neurale methoden daarentegen, zo wordt gesuggereerd, benaderen willekeurige regressiefuncties en zijn door de willekeurige leek te gebruiken.

Dit artikel betreft deze opvattingen. Wij willen de aandacht vestigen op niet-parametrische regressiemethoden: zowel op neurale netwerken als op methoden die zijn ontwikkeld binnen de statistiek. Door de neurale methoden naast de statistische methoden te presenteren hopen wij tenminste één vooroordeel over de statistiek, zoals bijvoorbeeld verwoord door Hecht-Nielsen, weg te nemen.

Verder wijzen wij op de onvolkomenheden, zowel praktisch als theoretisch, van de niet-parametrische methoden, wederom, of het nu neurale of statistische methoden betreft. Het blijkt dat de huidige ontwikkelingen binnen de niet-parametrische regressie niet hebben geleid tot een 'black-box', dat wil zeggen geheel automatisch, modelleringshulpmiddel. Er zijn zelfs fundamentele obstakels om zover te komen. Deze observatie houdt in dat de geschetste ontwikkelingen de statisticus niet overbodig maken, maar juist meer bestaansrecht geven.

In dit artikel worden eerst enkele niet-parametrische, neurale en statistische methoden voorgesteld. Vervolgens wordt het 'onzuiverheid-variantie' dilemma besproken, een fundamenteel probleem voor de toepassingen van alle niet-parametrische regressiemethoden. Hierbij dient Geman, Bienenstock en Doursat (1992) als leidraad. De methoden worden vergeleken op basis van een aantal concrete problemen; hierbij putten wij uit Ripley (1993) en eigen ervaring bij lampontwerp. Een aantal conclusies besluit het verhaal.

# 2 Niet-parametrische regressie

Regressiemethoden houden zich bezig met het modelleren van een afhankelijke variabele, $y$, als functie van een aantal onafhankelijke variabelen, de predictoren, $x_1, \ldots, x_p$,

$$y = f(x_1, \ldots, x_p). \tag{1}$$

Is de variabele $y$ discreet dan spreekt men ook wel van klassificatie in plaats van regressie. Bij het bepalen van de functie $f$ in (1) spelen altijd waarnemingen een rol. Men verzamelt waarnemingen $(y_i, x_i) = (y_i, x_{i1}, \ldots, x_{ip})$ en gebruikt deze om de functie $f$ te preciseren.

Naast deze waarnemingen spelen vaak aannames een rol waarin wordt gestipuleerd hoe de afhankelijke variabele afhangt van de onafhankelijke variabelen. In dit geval spreekt men wel van *parametrische* regressie. Bij parametrische regressie is het verband tussen afhankelijke variabele en onafhankelijke variabelen bekend op een klein aantal parameters na. Deze parameters worden geschat op basis van de waarnemingen. Binnen deze parametrische aanpak zijn de wiskundige eigenschappen van deze schatters goed bekend.

Het nadeel van deze parametrische aanpak is gelegen in de aanname van het verband tussen de afhankelijke en de onafhankelijke variabelen. Mogelijke discrepantie tussen aangenomen verband en werkelijkheid zorgt ervoor dat de modellen slecht voorspellen en dat de fraaie wiskundige analyse van de schatters niet van toepassing is. Daarnaast is het formuleren van plausibele modellen een tijdrovende zaak of ontbreekt de benodigde voorkennis om tot een bevredigend model te komen.

Deze problemen inspireerden *niet-parametrische* benaderingen van het regressievraagstuk. Binnen de niet-parametrische regressie wordt het modelleringsprobleem omzeild en wordt de regressiefunctie $f$ in (1) geconstrueerd op basis van alleen de waarnemingen. Traditioneel worden binnen deze niet-parametrische benadering neurale netwerken van statistische methoden onderscheiden. Ripley (1993) geeft een uitmuntend overzicht van de vele soorten neurale netwerken als bijvoorbeeld het meerlaags perceptron, de Boltzmann machine, het Hopfield netwerk en het recurrente netwerk. Wij beperken ons in dit verhaal tot het meerlaags perceptron met foutenpropagatie als leeralgoritme.

Bekende statistische technieken zijn CART, ACE, LOESS, projection pursuit regression, nearest neighbourhood regression, generalized additive models, MARS en PIMPLE. Wij beschrijven CART (Breiman et. al. 1984), een relatief vroege en inzichtelijke methode, en MARS (Friedman, 1991), een mijns inziens zeer veel belovende methode gebaseerd op CART. Voor de beschrijving van de overige methoden verwijzen wij naar de technische literatuur (Annals of Statistics, JASA of Technometrics) en in het bijzonder naar Friedman (1991).

## 2.1 Het meerlaags perceptron

Het werk op het gebied van de neurale netwerken wordt geïnspireerd door de formidabele prestaties van het menselijke brein. Enerzijds poogt de neurowetenschap de werking van de menselijke hersenen beter te begrijpen door een aantal essentieel geachte eigenschappen ervan in hardware na te bootsen. Anderzijds denkt men een nieuwe generatie krachtigere computers te creëren, juist door deze te modelleren naar de essenties van de menselijke hersenen.

Binnen deze visie is een neuraal netwerk een imitatie van de menselijke hersenen. De bouwsteen voor een dergelijk neuraal netwerk vormt de neuron, een imitatie van de zenuwcel in de hersenen. Net als de zenuwcel kan zo'n neuron beschreven worden als een niet-lineaire, 'sigmoïdale', bewerking op een input. Deze bewerking wordt vaak beschreven met de logistische functie

$$h(x) = \frac{1}{1 + \exp\{-x\}}. \tag{2}$$

Door deze neuronen op verschillende manieren met elkaar te verbinden ontstaan diverse klassen neurale netwerken. Wij beperken ons tot het *meerlaags perceptron*. Het meerlaags perceptron is een neuraal netwerk dat bestaat uit verschillende lagen van neuronen. In zo'n meerlaags perceptron vloeit informatie in één richting: ieder neuron uit een gegeven laag heeft als input de gewogen som van outputs van de neuronen uit de direct voorafgaande laag. In een regressiecontext fungeren de waarden van de onafhankelijke variabelen als input voor de onderste laag (laag 0); de voorspelde waarde voor de afhankelijke variabele is de output van de neuron in de bovenste laag (laag L). Wij beperken ons hier tot het geval van regressie met slechts één afhankelijke variabele.

Formeel kan je een dergelijk netwerk dus beschrijven als een niet-lineaire transformatie van de onafhankelijke variabelen:

$$y = N(x_1, \ldots, x_p), \tag{3}$$

waarbij $N$ beschreven wordt door

$$N \;=\; g^{(L)} \circ g^{(L-1)} \circ \ldots \circ g^{(1)}$$

$$g^{(l)}(x) \;=\; h(W^{(l)}x), l = 1, \ldots, L.$$

Hierbij is $\circ$ de functie samenstelling, $W^{(l)}, l = 1, \ldots, L$ zijn gewichtsmatrices en $h$ is de logistische functie (2) met componentsgewijze evaluatie van zijn vectorargument.

Gegeven de gewichtsmatrices $W^{(l)}, l = 1, \ldots, L$ implementeert het meerlaags perceptron een parallelle berekening van de niet-lineaire functie (3). Deze gewichten moeten dus zo gekozen worden dat de geïmplementeerde functie, $N$, een geschikte benadering vormt van de regressiefunctie. Het bepalen van de gewichten gebeurt in de zogenaamde *leerfase*. In deze fase worden de gewichten zo bepaald dat een foutcriterium

$$E = \sum_{i=1}^{m} (y_i - N(x_i))^2 \tag{4}$$

geminimaliseerd wordt voor de waarnemingen uit de de *leerverzameling*. Deze optimalisering wordt vaak uitgevoerd met het *foutenpropagatie algoritme*; een gradient descent-algoritme dat parallel geïmplementeerd kan worden met het netwerk dat ook de functie-evaluatie uitvoert.

Door het aantal lagen en het aantal neuronen per laag (de architectuur van het netwerk) te variëren verkrijgt men een klasse van meerlaags perceptrons. Van deze klasse kan men het volgende bewijzen (zie White, 1989).

- Iedere 'redelijke' functie valt willekeurig goed te benaderen met behulp van een neuraal netwerk met drie lagen ($L = 3$). Eigenlijk zijn er in dit geval vier lagen in het netwerk, maar het is gebruik de onderste (input) laag niet mee te tellen. In dit geval spreekt men ook over twee *verborgen* lagen.

- Iedere continue functie valt willekeurig goed te benaderen met behulp van een neuraal netwerk met twee lagen.

- De benadering valt te leren op basis van een groeiend aantal waarnemingen (consistentie).

Het zijn deze eigenschappen die de weg vrij maken voor de toepassingen van neurale netwerken binnen de statistiek. Voor elke regressiefunctie is er een feedforward netwerk dat deze regressiefunctie goed benadert. Bovendien valt deze functie te leren op basis van waarnemingen.

Zijn er dan nog problemen voor een dergelijke toepassing? Ja, enerzijds liggen er problemen in het numerieke vlak. Het gradient descent-algoritme is een benaderingsalgoritme dat convergeert naar een lokaal minimum, dat niet noodzakelijk een globaal minimum is. Voor de consistentiestellingen is het echter nodig dat er een globaal minimum gevonden wordt van het criterium (4). Bovendien is convergentie van het algoritme niet altijd gegarandeerd.

Naast dit numerieke probleem is het volgende essentieel. Het neurale netwerk benadert iedere willekeurige functie willekeurig goed voor een alsmaar toenemend aantal neuronen in het netwerk. De keuze van de grootte van het netwerk is daarmee een factor van belang bij een gegeven probleem. In paragraaf 3 zullen wij beargumenteren dat die keuze zelfs essentieel is voor een succesvolle toepassing van neurale netwerken in de statistiek.

## 2.2 CART

CART (Classification And Regression Trees) is een binnen de statistiek ontwikkelde methode voor het schatten van een algemene regressiefunctie (1). Breiman, Friedman, Olshen en Stone (1984) geven een prachtige beschrijving van de methode en Clark en Pregibon (1991) beschrijven de implementatie in S-plus.

CART zoekt de benadering van de regressiefunctie binnen de functieklasse van de lokaal-constante functies van de vorm

$$C(x) = \sum_{B \in \mathcal{B}} c_B \text{Ind}(x \in B). \tag{5}$$

Hierbij is $\mathcal{B}$ een partitie van de ruimte opgespannen door de predictoren, Ind is de indicator-functie en de $c_B, B \in \mathcal{B}$ zijn nader te bepalen constantes. CART beperkt zich tot eenvoudige partities van de predictorruimte bestaand uit alleen 'hyper'-rechthoeken. In dit geval zijn de indicatorfuncties te schrijven als producten van univariate stapfuncties $H$

$$\text{Ind}((x_1, \ldots, x_p) \in B) = \prod_{j=1}^{p} H(x_j - a_j^{(l)}) H(-(x_j - a_j^{(r)})). \tag{6}$$

Hierbij zijn $a_j^{(l)} \in \Re$ en $a_j^{(r)} \in \Re$ de te kiezen linker- en rechtergrens van de hyper-rechthoek $B$; de afhankelijkheid van deze grenzen van $B$ is in de notatie onderdrukt. Beide grenzen kunnen $\infty$ zijn. De functie $H$ is een univariate stapfunctie

$$H(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0. \end{cases} \tag{7}$$

Net als het meerlaags perceptron implementeert CART een niet-lineaire functie, gegeven de constanten $c_B$ en de grenzen $a_j^{(l)}$ en $a_j^{(r)}$ van de partitie van de predictorruimte. Wederom worden deze gevonden in een zogenaamde leerfase waarin CART een 'goede' partitie en constantes $c_B$ zoekt zodat een foutcriterium

$$E = \sum_{i=1}^{m} (y_i - C(x_i))^2 \tag{8}$$

'klein' wordt voor de waarnemingen uit de leerverzameling. Er wordt hier van 'goed' en 'klein' gesproken, en niet van optimaal respectievelijk minimaal, omdat er hier geen sprake is van echte minimalisering. Dat is namelijk een $NP$-lastig probleem. CART gebruikt een 'greedy' algoritme waarmee een redelijke benadering wordt gevonden.

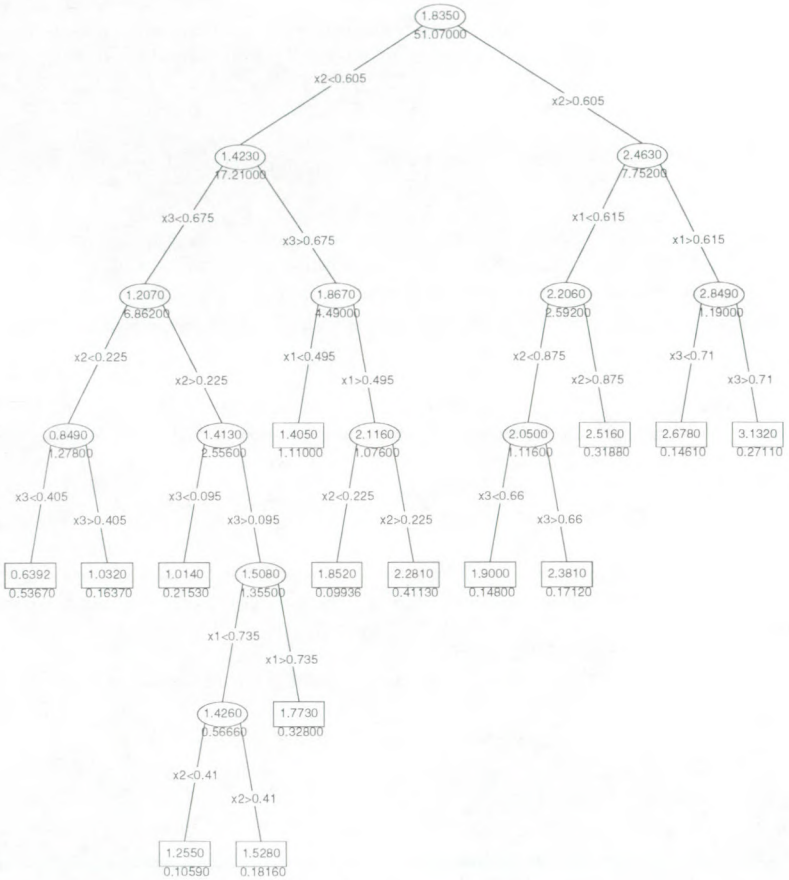
In Figuur 1 geven wij een illustratie, het is de met CART gevonden benadering van de functie

$$y = f(x) = x_1^2 + 2x_2 + x_3, \tag{9}$$

voor $0 \leq x_1 \leq 1$, $0 \leq x_2 \leq 1$ en $0 \leq x_3 \leq 1$. Een dergelijke benadering wordt wel een regressieboom genoemd. De bladeren van de boom (de rechthoeken) zijn de deelgebieden in de opsplitsing van de predictorruimte. Het getal binnen het vierkant is de benaderende waarde op een dergelijk deelgebied. De paden naar de eindgebieden, de labels bij de lijnen, geven aan hoe de opsplitsing in deelgebieden wordt gevonden. Dus in het predictorgebied $\{x_2 > .605, x_1 > .615, x_3 > .71\}$ (rechtsonder) wordt de afhankelijke variabele benadert met 3.132.

Dit voorbeeld illustreert tevens de zwaktes van de regressiebomen, waarbij men moet bedenken dat het zwaktes zijn van een zeer nuttig algoritme met zeer succesvolle toepassingen, zie bijvoorbeeld Geman, Bienenstock en Doursat (1992). Ten eerste is de benadering met behulp van een regressieboom altijd discontinu. In veel praktijkproblemen zijn de te benaderen functies echter continu. Ten tweede maskeert de CART benadering eenvoudige additieve structuren. Wij spreken van een additieve structuur indien er functies $f_i : \Re \to \Re, i = 1, \ldots, p$ bestaan waarvoor de te benaderen functie $f$ te schrijven is als

$$f(x_1, \ldots, x_p) = \sum_{i=1}^{p} f_i(x_i). \tag{10}$$




Figuur 1: CART benadering van functie (9): labels bij verbindingslijnen geven aan hoe CART de predictorruimte stapsgewijs opsplitst: na stap 1 is de ruimte opgedeeld in de twee deelgebieden $\{x_2 \leq .605\}$ en $\{x_2 \geq .605\}$. Ovalen geven deelgebieden aan van de voorlopige opsplitsing; rechthoeken de deelgebieden van de definitieve opsplitsing. Het getal binnen een ovaal of rechthoek is de benaderende waarde op het corresponderende deelgebied; het getal onder ovaal of rechthoek is de gemiddelde kwadratische afwijking van de waarnemingen binnen dat deelgebied ten opzichte van deze benaderende waarde.

$$\begin{aligned}\hat{M}(x) \;=\; & .0 + \\ & .4(x_1 - .15)^+ + .5(x_1 - .35)^+ - .1(-(x_1 - .35))^+ + \\ & .3(x_1 - .5)^+ + .5(x_1 - .7)^+ + \\ & 2x_2 + \\ & x_3\end{aligned}$$

Figuur 2: MARS benadering van functie (9); de $^+$ functie is het "positieve deel" van: $(x)^+ = \max\{0, x\}$. MARS vindt het lineaire stuk van de oorspronkelijke functie expliciet terug en geeft een stuksgewijs lineaire benadering voor de kwadratische term $x_1^2$.

Lineaire modellen zijn de bekendste voorbeelden van additieve modellen. CART werkt juist door 'interacties' tussen de predictoren te creëren en maskeert daardoor hoofdeffecten. In veel praktijkproblemen zijn de effecten van de predictoren onafhankelijk en daardoor komen dergelijke additieve structuren veel voor.

## 2.3 MARS

De zwaktes van CART inspireerden Friedman (1991) tot twee zeer eenvoudige observaties. Ten eerste is het duidelijk dat de discontinuïteit van de regressiebomen wordt veroorzaakt door de discontinuiteit van de stapfuncties, $H$. Ten tweede worden additieve effecten gemaskeerd doordat men een partitie van de predictorruimte kiest. Het verschil van MARS (Multivariate Adaptive Regression Splines) ten opzichte van CART is met name gelegen in deze twee punten.

MARS benadert een regressie functie met een element van de functieklasse

$$M(x) = \sum_{J \in \mathcal{J}} c_J \prod_{j \in J} (\text{sign}_J^j (x_j - a_j^J))^+. \tag{11}$$

Hierbij zijn $c_J \in \Re$ en $a_j^J \in \Re$ onbekende en te bepalen coefficienten en is $\text{sign}_J^j$ een te bepalen teken. Verder is $^+$ de functie

$$(x)^+ = \max(0, x) \tag{12}$$

en vervangt de $H$-functies, (7), van het CART-algoritme. Tenslotte is $J$ een deelverzameling van de verzameling van predictoren en $\mathcal{J}$ is een (te bepalen) verzameling van dergelijke deelverzamelingen. Net als bij CART en het meerlaags perceptron wordt er in de leerfase een heuristisch algoritme gebruikt om een element uit de functieklasse (11) te kiezen dat goed presteert voor de waarnemingen uit de leerverzameling.

Het MARS-algoritme heeft als aantrekkelijke bijkomstigheid dat er eenvoudig gemanipuleerd kan worden met de maximaal toegestane cardinaliteit van de deelverzamelingen $J$ in (11). Neemt men bijvoorbeeld de waarde 1 als de maximale cardinaliteit, dan beperkt het MARS-algoritme zich tot additieve benaderingen (10). Neemt men bijvoorbeeld de waarde 2, dan beperkt MARS zich tot de klasse van bivariate benaderingen

$$f(x) = \sum_{i \neq j} f_{ij}(x_i, x_j)$$

enzovoorts.

Ter illustratie toont Figuur 2 de MARS-benadering van functie (9). De benadering is additief en continu en de voordelen ten opzichte van de CART-benadering zijn evident.

# 3 Onzuiverheid versus variantie

De boven beschreven functieklassen zijn dermate flexibel dat redelijke functies er goed door benaderd worden. Het lijkt alsof er helemaal geen probleem is en dat niet alleen neurale, maar ook statistische methoden als CART en MARS 'black-box-'methoden zijn die complexe samenhangen leren uit door ons verzamelde gegevens. Hier schuilt echter een probleem aangezien optimalisering binnen een grotere functieklasse niet automatisch een betere benadering impliceert. Deze paradox wordt verklaard door het zogenaamde 'onzuiverheid-variantie-'dilemma.

Het basisprobleem is het construeren van een benadering $f(x)$ in (1) op basis van verzamelde gegevens $(y_i, x_i) = (y_i, x_{i1}, \ldots, x_{ip}), i = 1, \ldots, m$. De benadering is in het algemeen niet exact gelijk aan de ware, onbekende, functie $F(x)$, ongeacht de niet-parametrische techniek waarmee de benadering berekend is. Het verwachte kwadratische verschil tussen ware en benaderende functie

$$E\{(f(x) - F(x))^2\}, \tag{13}$$

is een maat voor de kwaliteit van de benadering. Hierbij is $f$ een functie van de verzamelde waarnemingen $(y_i, x_i), i = 1, \ldots, m$ en de verwachting, $E$, middelt over dergelijke steekproeven ter grootte $m$.

Het is makkelijk in te zien dat het verwachte kwadratische verschil in (13) in twee delen kan worden gesplitst:

$$\begin{aligned} E\{(f(x) - F(x))^2\} &= E\{(f(x) - E\{f(x)\})^2\} \\ &\quad + (E\{f(x)\} - F(x))^2. \end{aligned} \tag{14}$$

De eerste term in het rechterlid van (14) wordt de *variantie* term genoemd. Deze term geeft uitdrukking aan de variatie in de geconstrueerde functies bij herhaalde toepassing van de niet-parametrische methode op steeds nieuwe gegevensverzamelingen. De tweede term is de *onzuiverheid*. De onzuiverheid is een maat voor de systematische afwijking in de benaderingen ten opzichte van de ware functie bij herhaalde toepassing van de niet-parametrische methode op steeds nieuwe gegevensverzamelingen.

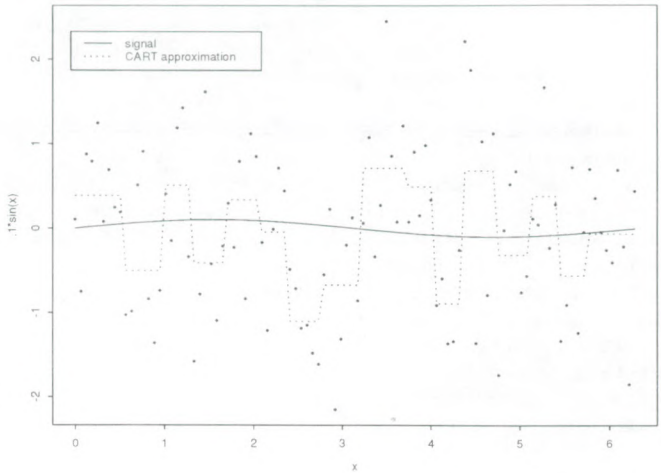
Deze maten zijn puntmaten. Zij kwantificeren de kwaliteit van de niet-parametrische procedure in één punt $x$. Globale maten verkrijgt men door integratie over het domein van de predictor-variabelen, maar dit technische aspect laten wij hier onbesproken.

Een goede niet-parametrische regressiemethode heeft een geringe verwachte kwadratische afwijking en dus zowel een geringe variantie als een geringe onzuiverheid. Het probleem, of dilemma, is nu dat deze twee eisen elkaar in zekere zin tegenspreken. Stel dat een niet-parametrische procedure wordt toegepast, waarbij wordt geoptimaliseerd binnen een kleine functieklasse. Herhaalde toepassing van de procedure op nieuwe gegevens zal benaderingen opleveren die weinig variëren. De procedure heeft in dit geval een kleine variantie. Anderzijds is het echter niet waarschijnlijk dat de gezochte functie $F$ goed benaderd wordt binnen de kleine functieklasse. Juist omdat de functieklasse klein is, heeft de methode een grote systematische fout.

De situatie wordt omgekeerd door de functieklasse zeer groot te kiezen. In dit geval is het waarschijnlijker dat er functies in de functieklasse zijn die een goede benadering geven van de gezochte functie en is de onzuiverheid van de procedure klein. Echter, omdat de functieklasse groot is, is het tevens waarschijnlijk dat de variatie groter wordt.

Een eenvoudige illustratie van dit dilemma wordt gegeven in Figuur 3. Er is slechts één predictor, $x$, en de afhankelijke variabele wordt gegenereerd als

$$y = 0.1 \sin(x) + \epsilon$$




Figuur 3: Het onzuiverheid-variantie probleem voor CART

met $\epsilon \sim N(0,1)$. Herhaalde toepassing van het CART-algoritme met een te kleine functieklasse, bijvoorbeeld met alleen constante functies, levert benaderingen die weinig variëren. In dit geval is echter de onzuiverheid substantieel. Toepassing van het CART-algoritme met een te grote functieklasse levert benaderingen die gemiddeld goed zijn. De Illustratie 3 maakt echter duidelijk dat in dit geval de variantie van de procedure groot zal zijn. In dit geval past de procedure de benadering te veel aan aan de waarnemingen.

Dus voor CART moet de omvang van de partitie $\mathcal{B}$ in (5) zo gekozen worden dat simultaan onzuiverheid en variantie geminimaliseerd worden. Bij MARS moet een soortgelijke keuze gemaakt worden. Hier moeten in (11) zowel het aantal deelverzamelingen in $\mathcal{J}$ bepaald worden als de maximale cardinaliteit van de deelverzamelingen $J \in \mathcal{J}$.

Ook bij neurale netwerken speelt dit onzuiverheid-variantie-probleem een grote rol, zoals Geman, Bienenstock en Doursat (1992) aan de hand van een aantal voorbeelden demonstreren. Ik beperk mij tot een eenvoudig klassificatieprobleem uit hun artikel. De predictoren zijn $x_1 \in [-6, 6]$ en $x_2 \in [-1.5, 1.5]$ en de afhankelijke variabele is

$$y = \begin{cases} -.9 & \text{als } x_2 < \sin(x_1\pi/2) \\ .9 & \text{als } x_2 \geq \sin(x_1\pi/2). \end{cases}$$

Een meerlaags perceptron moet $y$ leren klassificeren op basis van een leerverzameling van 100 waarnemingen. Door dit experiment 1000 maal te herhalen kunnen de verwachtingen in Vergelijking (14) benaderd worden met de gemiddelden over de herhalingen. Zo is het mogelijk de verwachte kwadratische fout experimenteel te splitsen in de onzuiverheid en de variantie. Geman, Bienenstock en Doursat herhalen dit onzuiverheid-variantie experiment voor een aantal netwerk architecturen, d.w.z. voor verschillende aantallen neuronen in de verborgen laag. Zo kunnen zij de onzuiverheid en variantie bij dit probleem uitzetten als functie van de omvang van het netwerk. Hun conclusie is dat de onzuiverheid afneemt met de omvang van het netwerk, maar dat de variantie toeneemt. Als gevolg hiervan neemt de totale fout eerst af en vervolgens

toe.

De omvang van een net is derhalve de essentiële parameter in het onzuiverheid-variantie-probleem en een kritische factor voor het functioneren van het net. De omvang van het netwerk speelt daarmee een gelijke rol als het aantal basisfuncties bij het CART-algorithme. Evenzo is het aantal iteraties van het foutenpropagatie-algoritme relevant voor de totale fout. Geman, Bienenstock en Doursat geven voorbeelden waarbij de totale fout eerst daalt met het aantal iteraties, en vervolgens weer stijgt.

Hiermee is geïllustreerd dat het onzuiverheid-variantie-probleem een rol speelt voor toepassing van niet-parametrische regressiemethoden, neuraal en statistisch. De kritische factoren voor het meerlaags perceptron zijn de omvang van het netwerk en het aantal iteraties van het foutenpropagatie-algoritme. Bij CART en MARS is de omvang van de functieklasse waarbinnen men een optimum zoekt de kritische factor.

Er zijn een aantal standaardoplossingen voor het onzuiverheid-variantie dilemma, gebaseerd op bijvoorbeeld asymptotiek, data splitsing, kruisvalidatie of regularisatie. Daarmee zijn echter een aantal heuristieken cruciaal voor het succesvol toepassen van de niet-parametrische technieken. Dit geeft al aan dat een zuiver theoretische vergelijking ons geen antwoord levert op de vraag welke methode de beste is. Daarom is een praktische vergelijking, het onderwerp van de volgende sectie, van belang.

# 4 Vergelijking

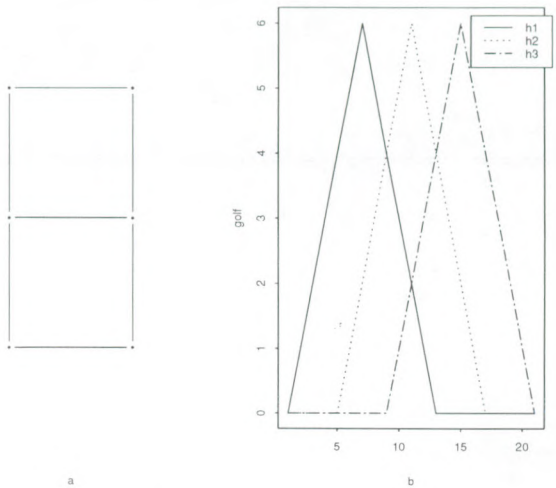
Wij vatten het voorafgaande samen. Niet-parametrische regressietechnieken beogen een model te vinden waarin een afhankelijke variabele wordt benaderd met een niet-lineaire functie van een groot aantal onafhankelijke variabelen. Zij onderscheiden zich van parametrische regressietechnieken door het ontbreken van a priori aannames over de functionele vorm van de te benaderen regressiefunctie. Er bestaat een groot aantal van dergelijke regressietechnieken. Een gedeelte hiervan valt onder het kopje 'neurale methoden'. Een andere gedeelte is ontwikkeld door statistici en is wellicht alleen bij statistische specialisten bekend.

De methodes verschillen in details. Zo is er verschil in de functieklasse waarbinnen een optimum gezocht wordt. Grof gesproken onderscheiden de statistische technieken zich door intuïtief plausibele functies van de neurale netwerken. De benaderingseigenschappen van de CART-modellen bijvoorbeeld zijn duidelijk en de functies zijn eenvoudig voor te stellen. De benaderingseigenschappen van het meerlaags perceptron zijn verrassend en liggen wiskundig 'diep'.

Een verder verschil is dat de statistische methoden vaak *adaptief* zijn. Adaptief wil hier zeggen dat de verzameling basis-functies voor het model stapsgewijs wordt uitgebreid, parameters worden toegevoegd, op basis van de waarnemingen. Kenmerkend voor de statistische methoden is ook dat er sprake is van een voorwaartse stap waarin een model wordt opgesteld met te veel parameters. Dit te verfijnde model wordt daarna onderworpen aan een terugwaartse reductieslag.

Zoals gezegd, deze technische verschillen mogen niet de essentiële overeenkomst verhullen die gelegen is in het schatten van de regressiefunctie. De te prefereren methode is uiteraard die methode die deze doelstelling het best verwezenlijkt. Is er geen uniforme beste, en waarom zou die er zijn, dan ligt er een belangrijke onderzoekstaak in het formuleren van richtlijnen wanneer welke methode de voorkeur verdient. Een dergelijk onderzoek zal haar startpunt moeten vinden in de empirie, in een vergelijkend warenonderzoek waarbij binnen één onderzoek verschillende niet-parametrische technieken worden toegepast.

Er wordt dan ook driftig vergeleken en Ripley (1993) geeft een uitmuntende inleiding. Om




Figuur 4: (a) Illustratie van het cijferprobleem. Een intuïtieve representatie van de cijfers door een rijtje nullen en enen verkrijgt men door de bits te laten corresponderen met de verbindingslijnen in de bovenstaande illustratie, als bij een digitaal horloge. (b) Illustratie van de basisgolven $h_1, h_2$ en $h_3$ uit het golfprobleem

enigszins de sfeer te proeven van dergelijke vergelijkingen behandelen wij ook hier een aantal gevallen. In Paragraaf 4.1 bespreken wij twee logische problemen ontleend aan Breiman, Friedman, Olshen en Stone (1984) en aan Ripley (1993). In de Paragrafen 4.2 en 4.3 worden vervolgens twee praktijkproblemen besproken. Paragraaf 4.2 behandelt het voorkomen van de Tsetse vlieg in Zimbabwe in 1896 en is ontleend aan Ripley (1993). Paragraaf 4.3 behandelt het nut van niet-parametrische regressietechnieken bij lampontwerp.

## 4.1 Logische problemen

Als test voor het CART-algoritme beschrijven Breiman, Friedman, Olshen en Stone het cijferprobleem en het golfprobleem. Bij het cijferprobleem moeten de cijfers, 0 tot en met 9, herkend worden op basis van hun representatie als een rijtje van zeven nullen en enen, zie Illustratie 4(a). Het cijferprobleem kent een deterministische variant, waarbij geleerd wordt op basis van correcte representaties van de cijfers en de door ons gepresenteerde stochastische variant waarbij elke bit in de representatie met onafhankelijke kans p=.1 verkeerd is. Het herkennen van cijfers wordt geleerd op basis van een leerverzameling van 300 waarnemingen. Het percentage foutieve klassificaties wordt bepaald aan de hand van een onafhankelijke gegevensverzameling, die dusdanig groot is dat de variantie van het geschatte percentage verwaarloosbaar is.

In het golfprobleem worden golven gegenereerd volgens drie verschillende mechanismen:

$$
\begin{aligned}
1:&\quad uh_1(t) + (1-u)h_2(t) \quad 1 \le t \le 21\\
2:&\quad uh_1(t) + (1-u)h_3(t) \quad 1 \le t \le 21\\
3:&\quad uh_2(t) + (1-u)h_3(t) \quad 1 \le t \le 21.
\end{aligned}
$$

| probleem | MLP | CART | MARS | DA | LR |
|---|---|---|---|---|---|
| cijfer (stochastisch)* | 35 % | 30% | | | |
| golf* | 19 % | 28 % | | | |
| Tsetse vlieg* | 8.4% | 9.6% | | 13.8 % | |
| Lampontwerp** | 5.1-9.3%*** | | 7.5% | | 9.3% |

Tabel 1: Vergelijking meerlaags perceptron (MLP) met verschillende statistische methoden, zowel niet-parametrische methoden (CART en MARS) als parametrische methoden (DA = Discriminant Analyse en LR= Lineaire Regressie). * Percentage foutieve klassificaties; ** Relatief foutpercentage; *** Beste en slechtste prestatie bij 37 verschillende architecturen.

waarbij $u$ uniform verdeeld is op $[0, 1]$. De basisgolfjes $h_i, i = 1, 2, 3$ zijn weergegeven in Illustratie 4. Een golf wordt waargenomen als

$$h(m) + \epsilon_m, m = 1, \ldots, 21$$

waarbij de $\epsilon_m, m = 1, \ldots 21$ onafhankelijk zijn en een standaard-normale verdeling volgen; $h(m)$ is een golf gegenereerd volgens één van de drie mechanismen. De leerverzameling bestaat uit 300 zo waargenomen golven tesamen met het genererende mechanisme. De te leren taak is het correct klassificeren van verdere golven.

In Tabel 1 geven wij de scores van CART en het beste meerlaags perceptron voor deze twee problemen. Wij zien dat er geen uniforme beste is. De verschillen in prestaties zijn niet dramatisch, maar ook niet verwaarloosbaar.

## 4.2 De Tsetse vlieg

De Tsetse vlieg is in vrijwel geheel Zimbabwe uitgeroeid. Ripley (1993) beschrijft echter gegevens over het al dan niet voorkomen van Tsetse vliegen in Zimbabwe in 1896, voordat er aan een verdelgingsprogramma was begonnen. In deze gegevensverzameling wordt Zimbabwe opgedeeld in ongeveer 5000 vierkante regio's van ca. 7 kilometer omtrek. Van elk van deze regio's is bekend of de Tsetse vlieg er in 1896 voorkwam. Bovendien zijn er van elke regio twaalf omgevingsfactoren bekend die betrekking hebben op hoogte, temperatuur, neerslag en vegetatie. Doel is het klassificeren van de regio's als wel/niet geschikt voor Tsetse vliegen op basis van de omgevingsfactoren.

Ripley vergelijkt een groot aantal neurale netwerken en statistische methoden voor dit klassificatieprobleem. Hij doet dit door de methoden te trainen met een leerverzameling van 500 willekeurig gekozen regio's. De vergelijking vindt daarna plaats met behulp van een disjuncte validatieverzameling van eveneens 500 regio's. In Tabel 1 geven wij een klein gedeelte van zijn bevindingen.

Ripley's globale conclusie is dat neurale netwerken en statistische methoden vergelijkbaar presteren wat de percentages betreft. Hij constateert wel een duidelijk verschil in snelheid. De sequentiële implementaties van neurale netwerken zijn enorm veel trager dan de statistische methoden. Ook constateert hij regelmatig convergentieproblemen met de software voor de neurale netwerken.

Het verschil in percentage foutieve klassificaties is een relevante, maar zeker niet alles bepalende factor in het gegeven probleem. Wij zoeken immers vooral antwoord op de vraag of de gekozen omgevingsfactoren bepalend zijn voor het al dan niet voorkomen van de Tsetse vlieg in een gegeven regio. Een vervolgvraag is welk van de omgevingsfactoren de meest bepalende zijn. Het is duidelijk dat het neurale netwerk relevante informatie bijdraagt voor het beantwoor-

den van deze vragen. Het is echter even duidelijk dat het expliciete model van CART meer handvatten geeft om met name de tweede vraag aan te pakken.

### 4.3 Lampontwerp

Het volgende probleem, lampontwerp, onderscheidt zich van de voorafgaande problemen, doordat de afhankelijke variabele continu is. Er is dus sprake van regressie en niet van klassificatie. Bij het ontwerpen van een lamp zijn er een aantal onafhankelijke parameters die vrij gevarieerd kunnen worden; dit zijn onder andere de afmetingen van de buis van de lamp, de chemische samenstelling van de inhoud van de buis en de druk in de buis. Deze onafhankelijke parameters hebben gevolgen voor bijvoorbeeld de lichtopbrengst van de lamp. Voor lampontwikkeling is het van belang deze afhankelijkheidsrelatie zo goed mogelijk te kennen. Het is dan mogelijk met weinig prototypes goede lampen te ontwikkelen.

Deze afhankelijkheidsrelaties kunnen geleerd worden door middel van theoretische studies, maar ook door waarnemingen aan de hand van de prototypes die in de loop der tijd ontwikkeld zijn en niet-parametrische regressietechnieken. In een gerichte studie werd onderzocht of een dergelijke empirische aanpak mogelijk is en snel tot een goede lampenvoorspeller leidt. Binnen het project werden onder andere neurale netwerken, twee verschillende implementaties van de hierbeschreven meerlaags perceptrons, vergeleken met MARS.

De bestudeerde gegevensverzameling bestond uit 463 lampen en 14 variabelen. Er waren tien onafhankelijke variabelen en vier afhankelijke variabelen. De gegevensverzameling werd gesplitst in een leerverzameling van 232 waarnemingen en een onafhankelijke validatieverzameling van 231 waarnemingen. Tabel 1 toont de resultaten, gemiddelde percentuele fout over de vier te voorspellen variabelen voor de validatieverzameling. De weergegeven resultaten zijn die van het beste van de twee meerlaags perceptrons, van MARS en van een benadering met lineaire regressie. De verschillen zijn wederom niet groot, met een lichte voorkeur voor het neurale netwerk. Beide niet-parametrische technieken betekenen een verbetering ten opzichte van lineaire regressie.

Bij deze resulaten moeten twee kanttekeningen geplaatst worden. Ten eerste werd het MARS-algoritme niet in zijn standaardvorm gebruikt. Er werd kruisvalidatie gebruikt om uit verschillende kandidaatmodellen te selecteren. Ten tweede presteerde één van de meerlaags perceptrons, hier gepresenteerd, duidelijk beter dan het andere perceptron, dat hier niet meegenomen is. Het betrof echter in beide gevallen implementaties van een meerlaags perceptron. Voorts was het zo dat het optimale neurale netwerk werd gevonden door specialisten op het gebied van neurale netwerken, terwijl het slecht presterende netwerk werd bediend door leken. De conclusie dat de gebruiker van verregaande invloed is op de kwaliteit van het leerproces, door de instelling en validatie van de kritische parameters, is zeker niet vergezocht.

## 5 Conclusies

In dit betoog hebben wij een aantal aspecten van niet-parametrische regressiemethoden de revue laten passeren. Daarbij is speciaal onderscheid gemaakt tussen enerzijds methoden uit de statistiek, als CART en MARS, en anderzijds neurale netwerken, zoals het meerlaags perceptron met foutenpropagatie als leeralgoritme. Onze eerste conclusies betreffen deze tweedeling.

- Vergelijkingen tussen neurale netwerken en niet-parametrische statistische methoden laten weinig, en in ieder geval geen uniforme, verschillen zien in prestatie.
- Statistische methoden leveren expliciete functionele beschrijvingen van de modellen. Dit heeft een aantal verschillen tot gevolg.

- Neurale netwerken zijn geschikt als voorspellers, maar geven minder handvatten voor begrip van het model dan de statistische methoden.
- Als gevolg hiervan zijn de statistische modellen vaak deeloplossingen die in een verdere analyse nog verfijnd worden; de resultaten van de neurale netwerken zijn eindoplossingen.
- Niet alle empirische modellen zijn zinvol te interpreteren en pogingen dit toch te doen kunnen tot tijdverlies en ergernis bij de klant leiden.

Op basis hiervan zou ik de voorkeur geven aan neurale netwerken bij eenvoudige, redelijk geïsoleerde modelleringsproblemen, waarbij de nadruk ligt op handelen en minder op begrip.

De tweede reeks conclusies betreft de impliciete aanname van de subtitel van dit artikel, dat de behandelde methoden zo krachtig zijn dat de statisticus overbodig is bij de toepassing ervan.

- Het onzuiverheid-variantie-probleem zorgt ervoor dat de methoden niet zo 'black-box' zijn als wel wordt gesuggereerd. De omgang met dit probleem is niet eenvoudig.
- De traagheid van sequentiële implementaties van neurale netwerken bemoeilijkt de implementatie van automatische, maar rekenintensieve oplossingen van het onzuiverheid-variantie probleem, als bijvoorbeeld kruisvalidatie.
- Werkelijke vooruitgang in voorspellende kracht van de modellen wordt in het algemeen niet verkregen door steeds maar meer gegevens te verzamelen, maar door het grondig analyseren van die gegevens. Bij de toepassing van MARS op het lampontwerp werd de grootste vooruitgang geboekt door verschillende types lampen te onderscheiden en die afzonderlijk te modelleren. Ook Geman, Bienenstock en Doursat (1992) boeken pas dan vooruitgang met het herkennen van handgeschreven tekens, als zij constateren dat het criterium (4) niet voldoet voor hun probleem. Eindeloos meer gegevens vergaren had de oplossing van het herkenningsprobleem niet naderbij gebracht.

## Referenties

[1] Breiman, L., J.H. Friedman, R.A. Olshen, C.J. Stone (1984), Classification and regression trees, Wadsworth and Broks/Cole.

[2] Clark, L.A. and D. Pregibon (1992), Tree-based models, in Chambers, J.M. and T.J. Hastie (eds.), statistical models in S, Wadsworth and Broks/Cole.

[3] Friedman, J.H. (1991), Multivariate Adaptive Regression Splines, Annals of Statistics 19, pp1-141.

[4] Hecht-Nielsen, R. (1990), Neurocomputing, Addison-Wesley.

[5] Geman, S., E. Bienenstock, and R. Doursat (1992), Neural networks and the bias/variance dilemma, Neural Computation 4, 1-58.

[6] Ripley, B.D. (1993), Statistical aspects of neural networks, Proc. Semstat, Sandbjerg, 4/92, Chapman and Hall.

[7] White, H. (1989), Learning in statistical neural networks: a statistical perspective, Neural Computation 1, 426-464.

ontvangen 22-4-1993
geaccepteerd 18-6-1993