Chemometrie

Dr.dr. J.P.M. Andries

Vierde, geheel herziene druk, 2019

Syntax Media – Utrecht

Voorwoord

Chemometrie is een nieuwe discipline binnen de chemie die zich in de analytische chemie een belangrijke plaats heeft weten te veroveren. Door toepassing van chemometrische technieken kan de kwaliteit van chemisch onderzoek aanzienlijk worden verbeterd en kan er ook meer informatie worden verkregen uit de grote hoeveelheden data die tegenwoordig door moderne analyse-instrumenten in de laboratoria worden gegenereerd.

De basis van het boek *Chemometrie* is gelegd in leerplanprojecten op het gebied van chemometrie in het hoger laboratoriumonderwijs. Sinds het verschijnen van de drie voorgaande drukken van het boek – die ik eerder samen met dr. A.B. de Vries schreef – hebben er belangrijke ontwikkelingen plaatsgevonden. Hier is in deze geheel herziene, vierde druk op ingespeeld.

Dit boek is geschreven voor studenten in het hoger beroepsonderwijs, het wetenschappelijk onderwijs en voor afgestudeerden in de beroepspraktijk die chemometrische technieken toepassen. De nadruk ligt daarbij op veelgebruikte technieken die zich hebben bewezen.

Het boek is geschreven op basis van ervaringen van de auteur met de verzorging van chemometrieonderwijs in het hoger laboratoriumonderwijs en de uitvoering van wetenschappelijk onderzoek op verschillende toepassingsgebieden van chemometrie. Bestaande onderwerpen zijn in het boek geactualiseerd en belangrijke nieuwe onderwerpen worden geïntroduceerd.

Lineaire regressie en univariate kalibratie zijn nu geïntegreerd. Methodevalidatie is gericht op de validatie binnen één laboratorium omdat gebruikers daar het meeste mee te maken zullen krijgen. Bij het onderwerp experimentele optimalisering worden de fundamenten van de meest toegepaste designs en sequentiële en simultane optimaliseringstechnieken beschreven.

Door de toepassing van moderne instrumentele analysetechnieken in de laboratoria is multivariate data-analyse voor de chemometrie een belangrijk toepassingsgebied geworden. Daarom zijn nieuwe hoofdstukken geschreven over de ontwikkeling van multivariate modellen voor

patroonherkenning en voorspelling van afhankelijke variabelen in nieuwe monsters met Multivariate Lineaire Regressie, Principale Componenten Analyse, Principale Componenten Regressie en Partial Least Squares.

In de uitgewerkte voorbeelden worden vanwege de leesbaarheid afgeronde tussenresultaten gepresenteerd. De eindresultaten zijn echter altijd berekend zonder tussentijds afronden.
De in het vakgebied gebruikelijke Engelse termen zijn vaak gehandhaafd om een goede aansluiting op de vakliteratuur te houden.

Matrixrekening is voor een groot deel niet meer geïntegreerd in de tekst zodat lineaire regressie bij kalibratie en optimalisering nu indien gewenst zonder matrixrekening kan worden toegepast. Matrixrekening wordt wel gebruikt in de hoofdstukken over multivariate data-analyse omdat de technieken hier voor een belangrijk deel op gebaseerd zijn. De wiskundige aspecten zijn echter zo veel mogelijk geconcentreerd in een beperkt aantal paragrafen. Daarbij kan software het werk voor de gebruiker aanzienlijk vereenvoudigen.

Het boek is onafhankelijk geschreven van specifieke software. In de eerste plaats omdat de ervaring heeft geleerd dat daar in de loop van de tijd veranderingen in kunnen optreden. In de tweede plaats om de lezer vrij te laten in de keuze daarvan. Een uitzondering wordt gemaakt voor toepassingen met Excel omdat de veranderingen daarin relatief gering zijn en het beschikbaar is voor alle gebruikers.

Tot slot wil ik mijn vrouw Corrie bedanken voor haar ondersteuning bij het schrijven van dit boek, ir. Caroline van der Meulen voor haar accurate redactiewerk en AlphaZet prepress voor de fraaie vormgeving van dit boek.

Etten-Leur, mei 2019

Dr.dr. J.P.M. Andries

Inhoud

Voorbeeld Bereken voor de ijzergehalten van de steekproef uit tabel 2.1 welk percentage van de metingen naar verwachting zal liggen tussen de grenzen 7,05% en 7,15%, als wordt aangenomen dat deze afkomstig zijn uit een normaal verdeelde populatie met een verwachtingswaarde $\mu = 7{,}10$ en een standaarddeviatie $\sigma = 0{,}0352$.

Eerst moeten de grenswaarden met (2.9) worden geschaald naar z-waarden.

$$z_1 = \frac{x_1 - \mu}{\sigma} = \frac{7{,}05 - 7{,}10}{0{,}0352} = -1{,}42$$

$$z_2 = \frac{x_2 - \mu}{\sigma} = \frac{7{,}15 - 7{,}10}{0{,}0352} = +1{,}42$$

Met behulp van tabel 1 en 2 van bijlage 1 wordt gevonden:

$P(-1{,}42 < z < +1{,}42) = P(z_2 < +1{,}42) - P(z_1 < -1{,}42) = 0{,}9222 - 0{,}0778 = 0{,}8444$.

Dus op theoretische gronden ligt 84,44% van de metingen tussen de grenzen 7,05% en 7,15%. In werkelijkheid liggen er 88 van de 100 ijzergehalten in de steekproef binnen deze grenzen.

2.4 Standaarddeviatie van het gemiddelde

Als in een steekproef een reeks monsters wordt geanalyseerd, zal er door toevallige fouten een zekere variatie optreden in de metingen. Als de meetreeks wordt herhaald, is te verwachten dat het gemiddelde van de nieuwe reeks door toevallige fouten iets zal afwijken van het eerder gevonden gemiddelde. Omdat binnen elke meetreeks de variaties zullen worden uitgemiddeld, zal de spreiding in de gemiddelden van verschillende reeksen kleiner zijn dan die in de afzonderlijke metingen.

De spreiding in de gemiddelden van verschillende meetseries wordt gegeven door de *standaarddeviatie van het gemiddelde* $s_{\bar{x}}$:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} \tag{2.12}$$

De standaarddeviatie van het gemiddelde $s_{\bar{x}}$ is kleiner dan de standaarddeviatie in de afzonderlijke metingen s door het delen door $\sqrt{n}$. Hoe groter de waarde van n is, des te kleiner is de standaarddeviatie van het gemiddelde $s_{\bar{x}}$ en des te kleiner is dus ook de spreiding in het gemiddelde.

Voorbeeld In de dataset van tabel 2.1 is de standaarddeviatie van het gemiddelde van 100 meetpunten:

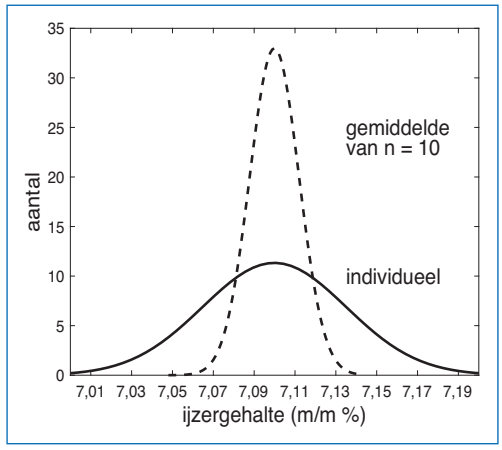
$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{0{,}0352}{\sqrt{100}} = 0{,}00352 = 3{,}52 \cdot 10^{-3}$$

Als in tabel 2.1 elke kolom als een meetserie wordt beschouwd en als daarvan per kolom het gemiddelde wordt berekend, worden de volgende 10 kolomgemiddelden gevonden:

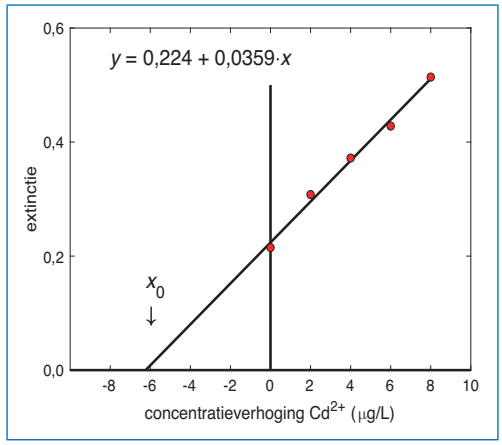
7,107 7,113 7,101 7,079 7,103 7,112 7,078 7,102 7,100 7,104.

Het gemiddelde van deze kolomgemiddelden is 7,0999. Deze waarde is gelijk aan het algemeen gemiddelde van de oorspronkelijke dataset. De standaarddeviatie van deze 10 kolomgemiddelden is 0,012096. Dat is belangrijk kleiner dan de standaarddeviatie voor de gehele set met afzonderlijke waarnemingen. Daarvoor was gevonden 0,0352.

In afbeelding 2.5 zijn de normale verdelingen gegeven voor de individuele resultaten en voor het gemiddelde van groepen van 10 metingen. Uit de grafiek blijkt dat de spreiding in de gemiddelden belangrijk kleiner is dan voor de individuele resultaten.



Afbeelding 2.5
Normale verdeling voor de ijzergehalten uit tabel 2.1 voor de individuele resultaten en het gemiddelde van groepen van 10 metingen.



Afbeelding 4.20
Kalibratielijn voor de standaard-additiereeks uit het voorbeeld.

Opgaven

1. Voor een spectrofotometrische cadmiumbepaling in water werd een test uitgevoerd op constante variantie. Daarvoor werden de extincties van een lage (x_L = 2 µg/L Cd^{2+}) en hoge (x_H = 18 µg/L Cd^{2+}) standaard tien maal gemeten. De resultaten daarvan zijn:

x_L	0,090	0,092	0,075	0,083	0,090	0,066	0,118	0,116	0,107	0,093
x_H	0,609	0,612	0,593	0,601	0,610	0,582	0,641	0639	0,629	0,614

   Test met behulp van de Hartley-test of de varianties constant te beschouwen zijn.

2. Voor een spectrofotometrische bepaling van een component werden bij een kalibratie de volgende resultaten gevonden waarin x de concentratie in mol/L is en y de gemeten extinctie.

x	$5{,}00.10^{-3}$	$10{,}00.10^{-3}$	$15{,}00.10^{-3}$	$20{,}00.10^{-3}$	$25{,}00.10^{-3}$
y	0,075	0,182	0,248	0,357	0,401

   a. Zet de meetpunten uit in een x-y grafiek.
   b. Bereken met behulp van lineaire regressie de helling en de as-afsnede van de kalibratielijn op basis van het model $y = \beta_0 + \beta_1 x$ en teken de kalibratielijn.
   c. Bereken de standaarddeviatie van de residuen en het 95%-betrouwbaarheidsinterval voor de modelparameters.
   d. Bereken de kwadratensommen, met bijbehorend aantal vrijheidsgraden, voor SS_T, SS_{Reg} en SS_{Res}.

e. Bereken de concentratie, met het 95%-betrouwbaarheidsinterval, voor een enkelvoudige extinctiemeting van 0,20.

3. a. Fit de kalibratiegegevens van opgave 2 met het tweedegraadsmodel $y = \beta_0 + \beta_1 x + \beta_2 x^2$.
   Bereken de modelparameters, geef de vergelijking van de kalibratielijn en teken deze.
   b. Bereken de standaarddeviatie van de residuen en het 95%-betrouwbaarheidsinterval voor de modelparameters.
   c. Bereken de kwadratensommen, met bijbehorend aantal vrijheidsgraden, voor SS_T, SS_{Reg} en SS_{Res}.
   d. Bereken de concentratie voor een enkelvoudige extinctiemeting van 0,20.
   e. Test of dit model significant beter is dan het eerstegraadsmodel.

4. Voor een spectrofotometrische bepaling van een component werd een kalibratie uitgevoerd in 1-cm-cuvetten met de volgende resultaten, waarin x de concentratie in mol/L is en y de gemeten extinctie.

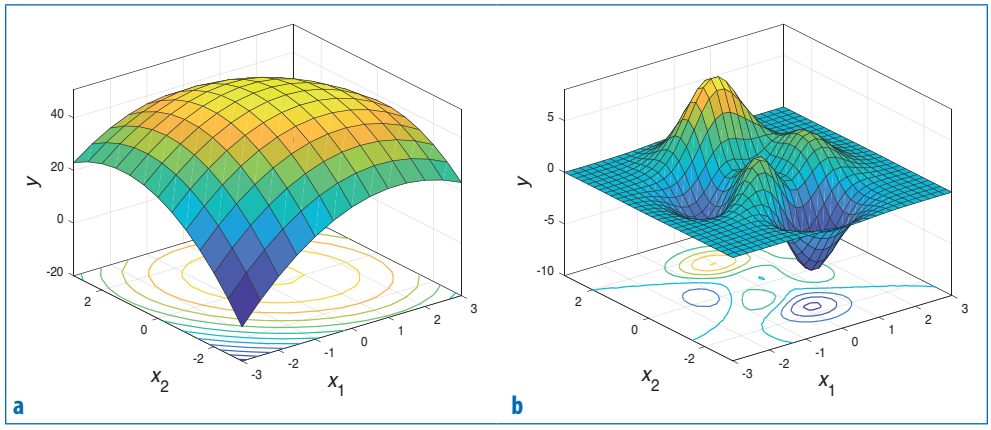
x	0,005	0,005	0,010	0,015	0,020	0,025	0,025
y	0,215	0,198	0,380	0,472	0,571	0,645	0,642

   a. Zet de meetpunten uit in een x-y grafiek.
   b. Bereken met behulp van lineaire regressie de helling en de as-afsnede van de kalibratielijn op basis van het model $y = \beta_0 + \beta_1 x$ en teken de kalibratielijn.
   c. Bereken de standaarddeviatie van de residuen en het 95%-betrouwbaarheidsinterval voor de modelparameters.
   d. Bereken de kwadratensommen, met bijbehorend aantal vrijheidsgraden, voor SS_T, SS_{Reg}, SS_{Res}, SS_{Lof} en SS_{Pe}.
   e. Bereken de concentratie, met het 95%-betrouwbaarheidsinterval, voor een enkelvoudige extinctiemeting van 0,30.
   f. Test op significantie van lack of fit.

5. a. Fit de kalibratiegegevens van opgave 4 met het tweedegraadsmodel $y = \beta_0 + \beta_1 x + \beta_2 x^2$, bereken de modelparameters, geef de vergelijking van de kalibratielijn en teken de kalibratielijn.
   b. Bereken de standaarddeviatie van de residuen en het 95%-betrouwbaarheidsinterval voor de modelparameters.
   c. Bereken de kwadratensommen, met bijbehorend aantal vrijheidsgraden, SS_T, SS_{Reg}, SS_{Res}, SS_{Lof} en SS_{Pe}.

Voor een model met twee factoren x_1 en x_2 kan de responsie y als functie van x_1 en x_2 grafisch worden weergegeven in drie dimensies (3D) (afb. 7.4). Dit wordt een *responsievlak* genoemd. Voor meer dan twee factoren kan het responsievlak wel worden berekend maar niet grafisch worden weergegeven. De projectie van een 3D-responsievlak op het vlak van de twee instelbare factoren levert een *contourplot* waarin lijnen met een gelijke responsie (*equiresponsielijnen*) kunnen worden getekend (afb. 7.4). De equiresponsielijnen zijn vergelijkbaar met hoogtelijnen op een geografische kaart. In dit hoofdstuk wordt er in het vervolg van uitgegaan dat naar de maximale responsie wordt gezocht.

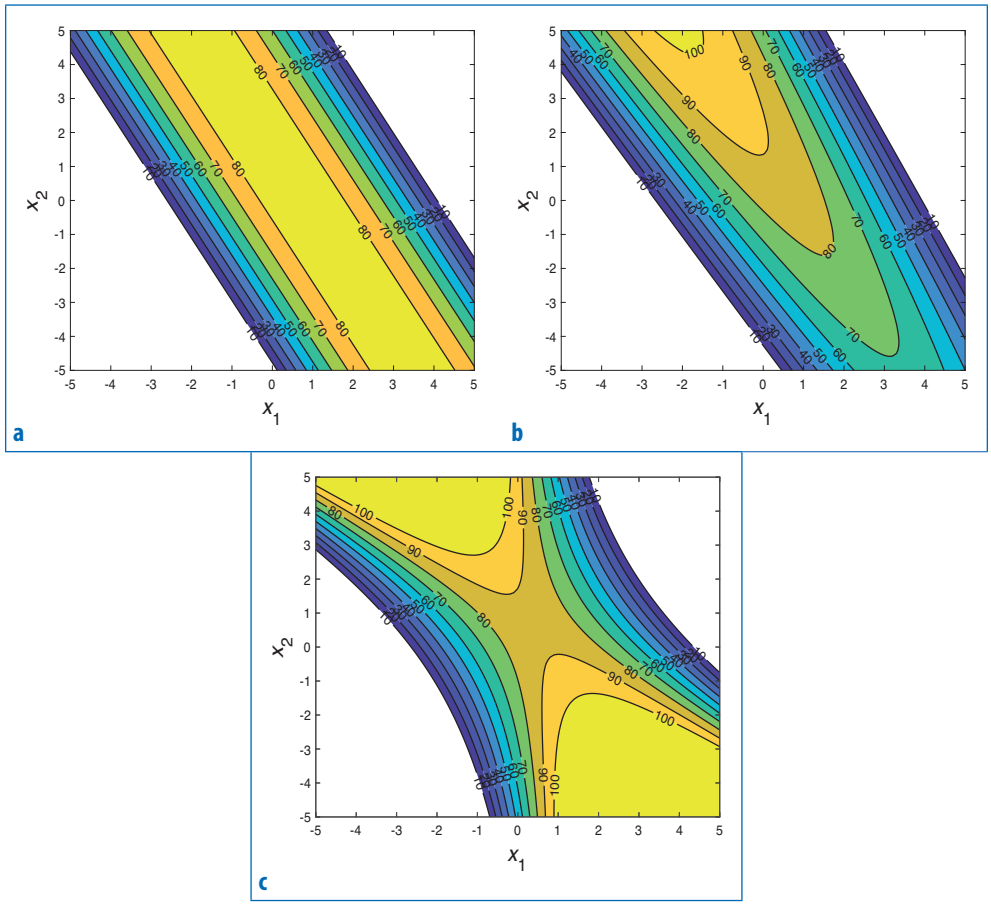
Als er één optimum in het responsievlak zit, is het responsievlak *unimodaal* (afb. 7.4a). Als er meerdere optima zijn, is het responsievlak *multimodaal*. Bij een *multimodaal* responsievlak is er naast één globaal optimum nog minstens één lokaal optimum aanwezig (afb. 7.4b).



Afbeelding 7.4 (a) Unimodaal responsievlak $y = f(x_1,x_2)$ met contourplot; (b) Multimodaal responsievlak $y = f(x_1,x_2)$ met contourplot.

Responsieoppervlakken kunnen ook andere vormen aannemen zoals blijkt uit de contourplots in afbeelding 7.5 voor (a) een *stationaire rug*, (b) een *rijzende rug* en (c) een *zadelvlak*. Een responsievlak voor een stationaire rug heeft vele optima, terwijl een rijzende rug één randoptimum heeft. Het bijzondere van een zadelvlak is dat het in de ene richting een maximum heeft terwijl dit een minimum is in de richting die daar loodrecht op staat.

In de praktijk is de vorm van het responsievlak en de bijbehorende contourplot met equiresponsielijnen niet bekend aan het begin van het optimaliseringsproces.



Afbeelding 7.5 Contourplots; (a) Stationaire rug; (b) Rijzende rug; (c) Zadelvlak.

De respons kan ook een functie zijn die meerdere criteria combineert zoals bij de *chromatografische responsiefunctie* (CRF) die wordt gebruikt in de chromatografie. Met de CRF kan een chromatogram worden geoptimaliseerd naar zo veel mogelijk pieken in een zo kort mogelijke analysetijd met tevens een korte retentietijd voor de eerste piek. De CRF is als volgt gedefinieerd [8]:

$$\mathrm{CRF} = \sum_{i=1}^{m-1} R_s(j, j+1) n^a + b(t_{\max} - t_m) - c(t_{\min} - t_1) \tag{7.4}$$

waarin m het aantal verwachte pieken is, n het aantal gedetecteerde pieken, $R_s(j, j + 1)$ de resolutie tussen piek j en $(j + 1)$, $t_{\min}$ en $t_{\max}$ de gewenste retentietijden voor de eerste en laatste piek, t_1 en t_m de retentietijden voor de eerste en laatste piek en a, b en c weegfactoren. De eerste term optimaliseert de resolutie tussen de pieken én het aantal pieken.

De *euclidische afstand* tussen object 1 en 2, berekend met (11.1) is:

$$d_{12} = \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2} = \sqrt{(1164 - 1020)^2 + (404 - 360)^2} = 150{,}6$$

Alle euclidische afstanden tussen de objecten worden op een analoge manier berekend. De 6×6 *euclidische afstandsmatrix* ***D*** voor de Fe-Ti-dataset staat in tabel 11.2. Daarin zijn de afstanden afgerond op gehele getallen.

Tabel 11.2 Euclidische afstandsmatrix ***D*** voor de Fe-Ti-dataset

	1	*2*	*3*	*4*	*5*	*6*
1	0	151	71	395	315	335
2	151	0	81	274	226	230
3	71	81	0	330	258	274
4	395	274	330	0	111	73
5	315	226	258	111	0	40
6	335	230	274	73	40	0

De afstandsmatrix ***D*** is symmetrisch ten opzichte van de diagonaal.

De *city-block-afstand* tussen object 1 en 2, berekend met (11.2) is:

$$d_{12} = |1164 - 1020| + |404 - 360| = 188$$

Alle *city-block*-afstanden tussen de objecten worden op een analoge manier berekend. De *city-block-afstandsmatrix* voor de Fe-Ti-dataset staat in tabel 11.3.

Tabel 11.3 City-block-afstandsmatrix voor de Fe-Ti-dataset

	1	*2*	*3*	*4*	*5*	*6*
1	0	188	95	559	433	470
2	188	0	93	371	245	282
3	95	93	0	464	338	375
4	559	371	464	0	126	89
5	433	245	338	126	0	43
6	470	282	375	89	43	0

De *similariteit* tussen object 1 en 2, berekend met (11.3) op basis van de euclidische afstand, met d_{max} = 395,3 (tabel 11.2), is:

$$s_{12} = 1 - \frac{d_{12}}{d_{max}} = 1 - \frac{150{,}6}{395{,}3} = 0{,}619$$

De similariteiten tussen de objecten worden op een analoge manier berekend. De *similariteitsmatrix* ***S*** voor de Fe-Ti-dataset staat in tabel 11.4.

Tabel 11.4 Similariteitsmatrix *S* voor de Fe-Ti-dataset

	1	*2*	*3*	*4*	*5*	*6*
1	1,000	0,619	0,820	0,000	0,203	0,152
2	0,619	1,000	0,795	0,307	0,429	0,418
3	0,820	0,795	1,000	0,166	0,347	0,308
4	0,000	0,307	0,166	1,000	0,719	0,817
5	0,203	0,429	0,347	0,719	1,000	0,899
6	0,152	0,418	0,308	0,817	0,899	1,000

De *correlatiecoëfficiënt r* tussen de variabelen Fe en Ti kan met de gegevens in tabel 11.1 en vergelijking (11.4) worden berekend, $r = 0{,}865$.

11.3 Hiërarchische agglomeratieve clustering

Er zijn verschillende methoden ontwikkeld voor het clusteren van objecten. Een van de meest toegepaste methoden is *hiërarchische clustering*. Hierbij kan worden gekozen tussen *agglomeratieve* en *divisieve hiërarchische clustering*.

In de *agglomeratieve clustering* worden de clusters gevormd door objecten of groepen van objecten steeds toe te voegen aan andere clusters totdat er één grote cluster is gevormd. Bij *divisieve clustering* wordt er gestart met één grote cluster die stapsgewijs wordt opgesplitst totdat er clusters overblijven die elk bestaan uit één object. In de chemie wordt het meest de hiërarchische agglomeratieve clustering toegepast.

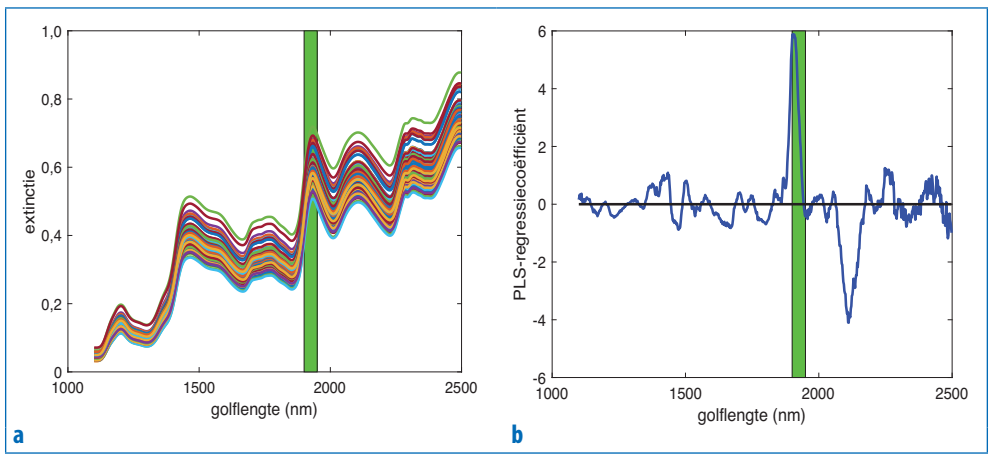
Voordat de regels voor het vormen van clusters worden gegeven, worden eerst de verschillende stappen bij het combineren van objecten in een cluster bij hiërarchische agglomeratieve clustering beschreven.

De *hiërarchische agglomeratieve clustering* verloopt in principe in de volgende stappen:

1. Elk individueel object wordt beschouwd als een cluster. Bij n objecten zijn er dus ook n clusters. De afstandsmatrix heeft n rijen en n kolommen.
2. Twee objecten die het dichtst bij elkaar liggen, worden gecombineerd tot één cluster. Er zijn daarna $(n - 1)$ clusters en de afstandsmatrix krijgt een rij en een kolom minder.

Een voorbeeld van variabelenselectie met de FCAM-methode op basis van de eigenschap REG (FCAM-REG) is beschreven in [61]. Daarbij wordt het vochtgehalte in mais bepaald met behulp van NIR-spectra. Dat is een snelle methode met een gemiddelde analysetijd van 1 minuut per monster. De dataset is gedownload van [62]. Deze bestaat uit NIR-spectra en de bijbehorende waterpercentages van 80 maismonsters met 700 meetgolflengten van 1100 tot 2498 nm met 2 nm verschil. De spectra zijn opgesplitst in een kalibratie- en testset met respectievelijk 60 en 20 monsters.

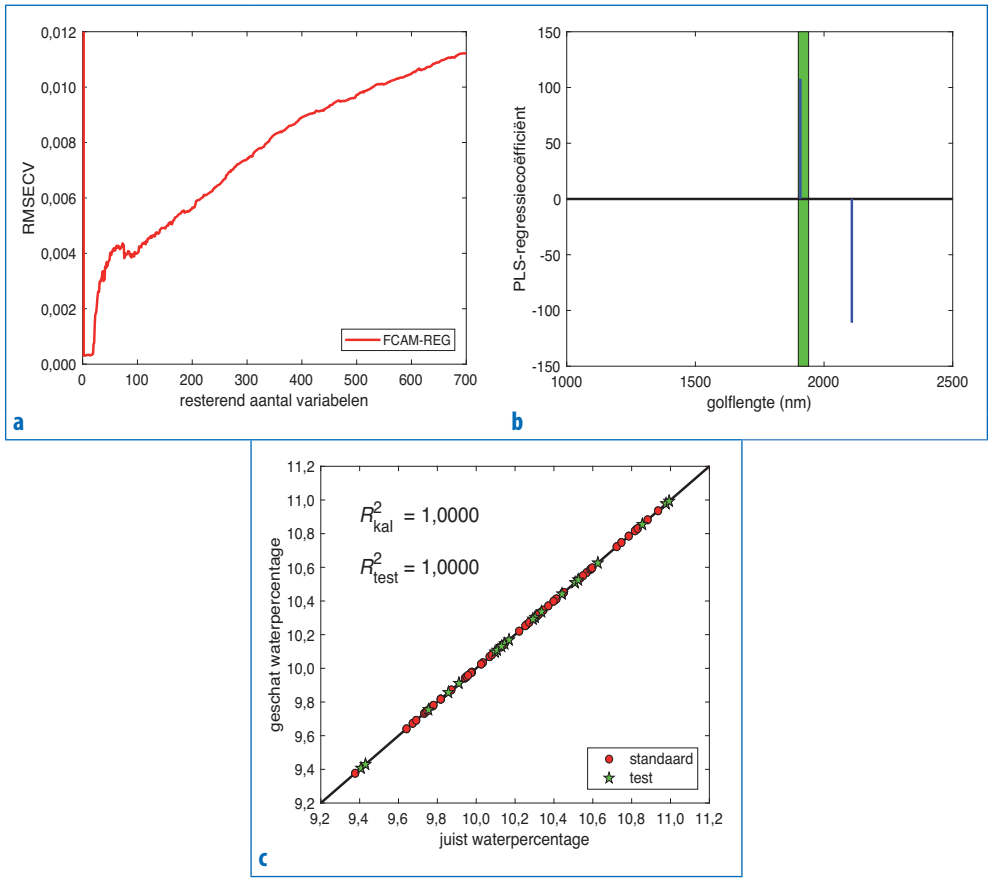
In afbeelding 15.5a zijn de NIR-absorptiespectra gegeven van de kalibratiemonsters van mais. In afbeelding 15.5b zijn de regressiecoëfficiënten van het PLS1-model voor de waterpercentages in mais weergegeven dat is ontwikkeld met de volledige spectra. De groene band is de absorptieband (1900 tot 1940 nm) van water die vaak wordt gebruikt voor de spectroscopische bepaling van water met NIR. De PLS-regressiecoëfficiënten hebben een positieve piek die gedeeltelijk overlapt met de waterband en een negatieve piek rond 2110 nm. De golflengten in de positieve en negatieve piek zijn belangrijk voor het model.



Afbeelding 15.5 (a) NIR-spectra van mais; (b) PLS1-regressiecoëfficiënten; groene band is waterabsorptieband.

Met de kalibratieset is, na centrering van $\boldsymbol{X}$ en $\boldsymbol{y}$, variabelenselectie toegepast met behulp van de FCAM-REG-methode. Het verloop van de RMSECV-curve als functie van het resterend aantal variabelen bij de variabelenselectie is weergegeven in afbeelding 15.6a. De RMSECV bij de start is $1{,}12 \cdot 10^{-2}$ op basis van het model met de volledige spectra. Op het einde, met de geselecteerde golflengtenset, is RMSECV = $3{,}04 \cdot 10^{-4}$. De curve laat zien dat (*i*) er tijdens de variabelenselectie een sterke afna-

me optreedt in RMSECV die een maat is voor de *voorspelfout*, en (*ii*) dat de variabelenselectiemethode zeer selectief is omdat er een minimum bestaat met een klein aantal resterende golflengten. Op basis van deze RMSECV-curve is een set met slechts twee variabelen geselecteerd bij 1908 en 2108 nm (afb. 15.6b). De golflengte van 1908 nm correspondeert met de positieve piek in afbeelding 15.5b en ligt binnen de waterband. De golflengte van 2108 nm correspondeert met de negatieve piek in afbeelding 15.5b. Met de geselecteerde golflengten 1908 en 2108 nm is een PLS-1 model ontwikkeld met een complexiteit $A = 2$. De voorspelfout van dit model voor de testset is RMSEP = $3{,}00 \cdot 10^{-4}$ terwijl deze vóór variabelenselectie $1{,}19 \cdot 10^{-2}$ was. De voorspelfout voor de testset is dus door variabelenselectie op basis van de RMSEP-waarde met een factor 40 gedaald.



Afbeelding 15.6 FCAM-REG-variabelenselectie; (a) RMSECV uitgezet tegen resterend aantal variabelen; (b) PLS-regressiecoëfficiënten van de twee geselecteerde golflengten; (c) PLS1-model voor de bepaling van het waterpercentage in mais op basis van twee geselecteerde golflengten bij 1908 en 2108 nm.