Chemometrie draagt in laboratoria bij aan kwaliteitszorg en aan de beheersing van Big Data

Deel 1

Dr. Dr. J.P.M. Andries

Geautomatiseerde analysesystemen in klinische chemische laboratoria, en moderne analyse-instrumenten in chemische en biomedische laboratoria zoals apparatuur voor NMR-, UV/VIS-, IR-, MS-, AAS-, FTIR-spectroscopie en gas- en vloeistofchromatografie genereren voor elk monster een groot aantal meetsignalen. Bovendien komen er steeds meer grote hoeveelheden data beschikbaar voor onderzoek in online databases. Dat heeft in de laboratoria een *'tsunami'* aan data veroorzaakt, soms ook aangeduid als een *'Big Data'* probleem.

Chemometrie, kwaliteitszorg en Big Data

De data zijn vaak sterk verschillend van aard omdat ze afkomstig zijn van veel verschillende soorten monsters en instrumenten. Meestal is een deel van de gegenereerde data niet informatief. Het probleem is dan om in de overvloed aan beschikbare data de gewenste informatie te vinden. Laboratoria moeten daarvoor gebruik maken van technieken waarmee het Big Data probleem kan worden beheerst. Daarnaast is het van belang dat de onderzoeksresultaten moeten voldoen aan strenge nationale en internationale normen om de opdrachtgevers te kunnen garanderen dat de resultaten voldoen aan vooraf gestelde eisen. Daarom moeten laboratoria beschikken over een goed kwaliteitszorgsysteem.

Chemometrie kan zowel bijdragen aan goede kwaliteitszorg als aan de beheersing van 'Big Data'. Chemometrie wordt gedefinieerd als de chemische discipline die mathematische en statistische methoden gebruikt om *(i)* optimale experimentele procedures te ontwerpen en te selecteren, en *(ii)* maximale chemisch relevante informatie te verschaffen uit chemische data. Statistiek vormt samen met wiskunde en chemie de fundamenten voor chemometrie.

Chemometrie is een nieuw vakgebied dat dateert van de jaren zeventig en tachtig van de vorige eeuw. Het vak is ontstaan door de behoefte aan kwalitatief goede data en aan oplossingen voor het Big Data probleem. De ontwikkeling van chemometrie is mede mogelijk geworden door het beschikbaar komen van snelle computers met krachtige chemometrische software waardoor de toepassing van chemometrische technieken eenvoudiger is geworden.

Chemometrie kan worden toegepast in elke fase van een analytisch proces zoals monstername en monstervoorbehandeling, kalibratie, methodeontwikkeling en methodevalidatie, experimentele optimalisering, dataverwerking en data-interpretatie. Chemometrische technieken worden tegenwoordig toegepast op tal van onderzoekgebieden zoals chemisch en biomedisch onderzoek, milieuonderzoek, levensmiddelenonderzoek, biotechnologie, bio-informatica en metabolomics. Het succes van de chemometrie is gebaseerd op het grote aantal toepassingen en de goede resultaten die daarmee bij laboratoriumonderzoek kunnen worden gerealiseerd. Het aantal laboratoriummedewerkers dat gebruik maakt van chemometrische technieken groeit daarom gestaag. In het hoger beroepsonderwijs en het wetenschappelijk onderwijs wordt hier in toenemende mate op ingespeeld. Er is daardoor in de praktijk en in het onderwijs een behoefte ontstaan aan een bijdetijds leerboek voor chemometrie.

Recent is een geheel herziene vierde druk van het boek 'Chemometrie' bij Syntax Media verschenen waarin de basisprincipes worden beschreven van veelgebruikte chemometrische technieken die zich bewezen hebben, zie referentie (1).

De behoefte aan kennis in het werkveld van het hbo Applied Science onderwijs hoe om te gaan met Big Data is ook beschreven in het recente rapport van het Domein Applied Science (DAS) met een inventarisatie en advies op het gebied van data-onderwijs, zie Ref. [2]. De auteur hoopt met het boek Chemometrie bij te dragen aan de kennisontwikkeling op dit gebied in het onderwijs en in het werkveld van het domein Applied Science.

Hierna wordt een overzicht gegeven van bijdragen die de chemometrie levert aan de kwaliteitszorg van laboratoriumonderzoek en de beheersing van Big Data, toegelicht met voorbeelden.

Regressie en kalibratie

De meeste analysemethoden zijn tegenwoordig gebaseerd op instrumentele technieken. Het is daarbij noodzakelijk om door kalibratie het verband te bepalen tussen de meetresultaten die worden geproduceerd door een analyse-instrument en de concentraties van te bepalen componenten in monsters. Kalibratie vormt de basis van elke kwantitatieve instrumentele methode. Chemometrie kan helpen om een goed kalibratiedesign te ontwikkelen. Dit omvat de keuze van het concentratiebereik, de verdeling van de standaarden over het concentratiebereik en het aantal standaarden op elk concentratieniveau. Het kalibratiedesign is van invloed op de grootte van het betrouwbaarheidsinterval van de bepaalde concentratie en op de validatie van het kalibratiemodel. Bij de validatie van een model wordt getest of het berekende model goed past bij de meetpunten door een test op lack of fit. Daarmee kan worden bepaald of bijvoorbeeld gekozen moet worden voor een eerste- of tweedegraadsmodel, en of dat al dan niet door de oorsprong moet gaan.

Bij veel toepassingen in het laboratorium wordt gebruik gemaakt van lineaire regressie. Daarbij moet voldaan worden aan een aantal voorwaarden: *(i)* er komen alleen fouten voor in de afhankelijke variabele y en niet in de onafhankelijke variabele *x, (ii)* de residuen moeten normaal verdeeld zijn en *(iii)* moeten onafhankelijk van elkaar zijn, en *(iv)* de variantie van *y* moet constant zijn over het gehele bereik van *x*. Er kunnen zich echter situaties voordoen waarbij aan één of meer van deze voorwaarden niet is voldaan. In dat geval kan een bijzondere regressietechniek worden toegepast zoals *(i)* een *robuuste regressietechniek* waarbij de meetpunten niet normaal verdeeld hoeven te zijn, bijvoorbeeld voor het testen van uitbijters of bij methodevergelijking, *(ii) gewogen regressie* als de variantie van de meetpunten niet constant is over het gehele meetbereik, *(iii)* regressietechnieken wanneer er fouten voorkomen *in x én y*, bijvoorbeeld bij methodevergelijking, *(iv)* het *lineariseren* van niet-lineaire modellen, *(v) niet-lineaire regressietechnieken* voor niet-lineaire relaties tussen *y* en *x*, en *(vi) meervoudige lineaire regressie* voor relaties tussen *y* en meerdere *x*-variabelen.

Een voorbeeld van een toepassing van niet-lineaire regressie in de enzymkinetiek is de bepaling van de maximale vormingssnelheid v_{max} en de constante K_M in de Michaelis-Menten-vergelijking. Dit is een niet-lineaire vergelijking in de parameters v_{max} en K_M, zie Figuur 1A. Er kan daarbij geen lineaire regressie worden toegepast. De Michaelis-Menten-vergelijking wordt daarom vaak omgezet in een lineaire vergelijking door $1/v$ uit te zetten tegen $1/[S]$ in de Lineweaver-Burk-plot, zie Figuur 1B. Hiervoor kan wel lineaire regressie worden toegepast. Door de omzetting in een lineaire vergelijking verandert echter de structuur van de experimentele fouten in de vormingssnelheid *v*. Als vóór het lineariseren is voldaan is aan de voorwaarde voor constante variantie dan hoeft dat niet meer te gelden ná lineariseren. De parameters v_{max} en K_M kunnen echter ook worden bepaald door middel van niet-lineaire regressie die eenvoudig kan worden uitgevoerd met de Oplosser in Excel, zie het voorbeeld in Figuur 2.

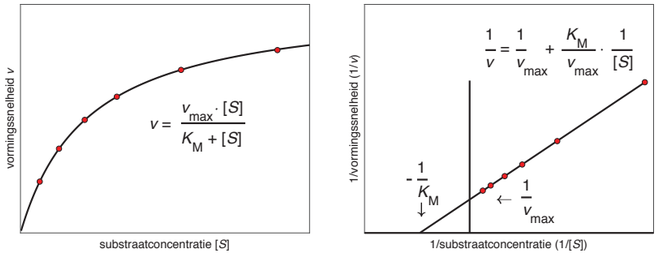
	A	B	C	D	E	F
1	**Michaelis Menten curve**					
2	Niet-lineaire regressie met Oplosser					
3						
4	enzym	snelheid	snelheid	kwadraat		
5	[S] in mmol/L	v (mmol/(L.min))	geschat	residu		
6	30,0	3,37	3,384	2,06E-04	v_max=	15,14
7	60,0	5,53	5,552	5,36E-06	KM=	104,24
8	100,0	7,42	7,415	2,80E-05		
9	150,0	8,94	8,935	2,76E-05		
10	250,0	10,70	10,688	1,56E-04		
11	400,0	12,00	12,013	1,73E-04		
12		kwadratensom residuen		5,96E-04		

Figuur 2. Voorbeeld van niet-lineaire regressie voor een Michaelis-Menten-curve met de Oplosser in Excel.

Methodevalidatie en methodevergelijking

De laatste stap in de ontwikkeling van analysemethoden is de methodevalidatie. Bij validatie wordt aangetoond dat de methode geschikt is voor een beoogde toepassing. De methode is gevalideerd voor een bepaalde toepassing als aan de gestelde eisen wordt voldaan. De eisen zijn vastgelegd in de prestatiekenmerken waarmee een analysemethode kan worden gekarakteriseerd. Zie bijvoorbeeld de Nederlandse Norm 7777 'Milieu en Voedingsmiddelen – Prestatiekenmerken van meetmethoden' die ook toegepast kan worden op andere gebieden, de richtlijn van de Nederlandse Vereniging voor Klinische Chemie 'Validatie en verificatie van onderzoeksprocedures in medische laboratoria' en de CLSI evaluatieprotocollen.

Voorbeelden van prestatiekenmerken zijn precisie, juistheid, aantoonbaarheidsgrens, bepalingsgrens en lineariteit. De prestatiekenmerken kunnen ook worden gebruikt om analysemethoden met elkaar te vergelijken.



Figuur 1. Enzymkinetiek; (A) Michaelis-Menten-vergelijking; (B) Lineweaver-Burk-plot.

Daarmee kan een betere keuze worden gemaakt voor een methode met betrekking tot een bepaalde toepassing. Methodevalidatie is een belangrijk onderdeel van de kwaliteitszorg in het laboratorium.

In een laboratorium kunnen zich verschillende situaties voordoen waarbij het noodzakelijk is de resultaten van twee verschillende analysemethoden met elkaar te vergelijken. Bijvoorbeeld na de ontwikkeling van een nieuwe methode die een bestaande methode moet gaan vervangen vanwege operationele voordelen zoals kosten of analysetijd. Of als eenzelfde bepaling op twee verschillende instrumenten of in twee verschillende laboratoria moet worden uitgevoerd. Het doel van de methodevergelijking is om vast te stellen of twee verschillende methoden overeenkomstige resultaten leveren of dat er systematische verschillen bestaan. In het ideale geval wordt een testmethode vergeleken met een referentiemethode. De referentiemethode is in de praktijk vaak de bestaande routinebepaling. Door vergelijking van de resultaten van beide methoden wordt nagegaan of de testmethode de bestaande routinebepaling kan vervangen.

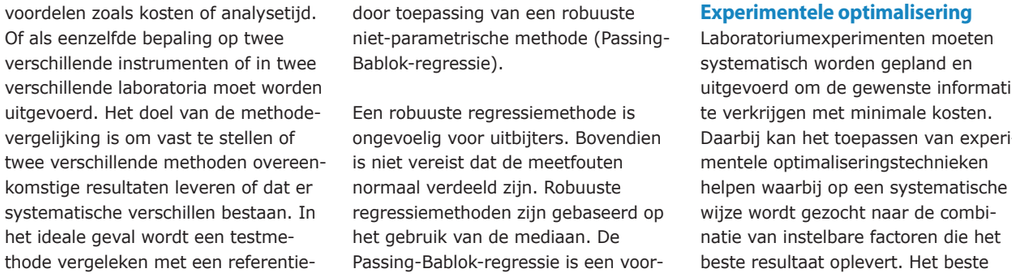
Bij de vergelijking van twee analysemethoden X en Y bestaat er onzekerheid in zowel de x- als in de y-variabele omdat er fouten voorkomen in de resultaten van beide methoden. Daarvoor kunnen verschillende technieken worden toegepast. In alle gevallen wordt daarbij een reeks praktijkmonsters geanalyseerd met beide methoden. De resultaten kunnen met elkaar worden vergeleken op basis van een gepaarde t-test, de Bland-Altman methode of op basis van regressie. Als de correlatiecoëfficiënt r ≥ 0,99 dan kan gewone lineaire regressie worden toegepast. Als r ≤ 0,975 dan moet een alternatieve regressiemethode worden toegepast. Bij een alternatieve regressiemethode kan een keuze worden gemaakt uit verschillende regressietechnieken zoals *(i)* minimalisatie van de kwadratensom van de loodrechte afstanden van de meetpunten tot de regressielijn (orthogonale regressie), *(ii)* gelijktijdige minimalisatie van de kwadratensom van de residuen in de *x*- én *y*-richting (Deming-regressie), en *(iii)* door toepassing van een robuuste niet-parametrische methode (Passing-Bablok-regressie).

Een robuuste regressiemethode is ongevoelig voor uitbijters. Bovendien is niet vereist dat de meetfouten normaal verdeeld zijn. Robuuste regressiemethoden zijn gebaseerd op het gebruik van de mediaan. De Passing-Bablok-regressie is een voorbeeld van een robuuste regressiemethode. Vanwege de voordelen wordt deze methode veel toegepast in de klinische chemie voor methodevergelijking. Bij de Passing-Bablok-regressie worden $n\cdot(n-1)/2$ hellingen berekend tussen alle mogelijke paren van n meetpunten. Daarna worden alle berekende hellingen gesorteerd op grootte en wordt een aangepaste mediaan bepaald. Vervolgens wordt door elk meetpunt een rechte lijn getrokken waarvan de helling gelijk is aan de aangepaste mediaan van alle mogelijke hellingen en worden de bijbehorende as-afsneden berekend. Van deze as-afsneden wordt vervolgens ook weer de mediaan bepaald. Dit levert de vergelijking van de Passing-Bablok-regressielijn. Details van de Passing-Bablok-regressie zijn beschreven in [3,4].

Figuur 3A geeft een voorbeeld voor een methodevergelijking met 10 meetpunten waarbij 45 hellingen zijn berekend tussen alle mogelijke meetparen. In Figuur 3B is een histogram afgebeeld voor deze hellingen waarbij, zoals verwacht mag worden, de klasse met hellingen rond de waarde 1,0 het meeste voorkomt. In Figuur 3C is de Passing-Bablok-regressielijn afgebeeld met de bijbehorende 95% betrouwbaarheidsintervallen.

Experimentele optimalisering

Laboratoriumexperimenten moeten systematisch worden gepland en uitgevoerd om de gewenste informatie te verkrijgen met minimale kosten. Daarbij kan het toepassen van experimentele optimaliseringstechnieken helpen waarbij op een systematische wijze wordt gezocht naar de combinatie van instelbare factoren die het beste resultaat oplevert. Het beste resultaat of het optimum kan zijn: 'hoogste opbrengst', 'kortste analysetijd', 'beste scheiding', 'sterkste meetsignaal', 'kleinste experimentele fout'. Experimentele designs vormen een belangrijk onderdeel van experimentele optimalisering. Ze bestaan uit een set van technieken waarmee een goed design voor experimenten kan worden opgezet om een maximale hoeveelheid informatie te verkrijgen met een minimum aantal experimenten. Experimentele optimalisering kan een onderdeel vormen van methodeontwikkeling. Door de systematische aanpak kunnen tijd en kosten worden bespaard. Daarmee kan bijvoorbeeld de opbrengst bij de productie van een enzym in een bioreactor worden verhoogd, een chromatografische scheiding worden verbeterd of de extinctie voor een spectrofotometrische bepaling worden verhoogd.

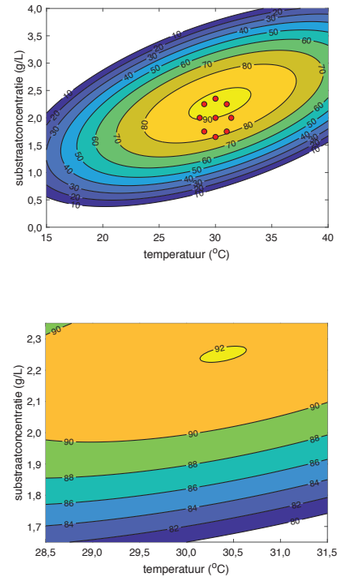


Figuur 3. Methodevergelijking op basis van de Passing-Bablok-regressie; (A) Rechte lijnen door alle mogelijke paren van meetpunten; (B) Histogram voor alle mogelijke hellingen; (C) Regressielijn met bijbehorende 95%-betrouwbaarheidsintervallen.

Experimentele optimaliseringstechnieken zijn breed toepasbaar bij laboratoriumonderzoek op veel toepassingsgebieden. Ze vormen een belangrijk onderdeel van de chemometrie. Als de klassieke optimaliseringstechniek wordt toegepast waarbij afwisselend één factor gevarieerd terwijl de overige factoren constant worden gehouden dan wordt het juiste optimum niet gevonden als de factoren interactie vertonen, wat meestal het geval is.
Om het juiste optimum efficiënt te kunnen vinden is een stapsgewijze strategie nodig. Eerst moet worden onderzocht welke factoren van invloed zijn op het optimum. Daarna moet een sequentiële methode zoals een simplexoptimalisering of de methode van het steilste pad worden toegepast om stapsgewijs de buurt van het optimum te vinden. Bij een sequentiële optimaliseringsmethode wordt gestart met een klein aantal experimenten bij vooraf bepaalde instelniveaus van de factoren. Op basis van het resultaat worden vervolgens de instelniveaus van factoren voor een volgend experiment gepland.
De sequentiële optimalisering stopt als er een voldoende respons is verkregen of als de verandering in respons klein is geworden. Als de positie van het optimum precies moet worden bepaald, dan kan daarna in de buurt van het optimum een simultane optimaliseringsmethode worden toegepast. Bij simultane optimalisering wordt vooraf, op basis van een experimenteel design, vastgesteld bij welke combinaties van factorinstellingen responsiemetingen uitgevoerd moeten worden. Op basis van de factorinstellingen en de gemeten responsies kan daarna de positie van het optimum worden berekend met behulp van een responsiemodel.

Een voorbeeld van een toepassing van deze strategie voor de optimalisering van de productie van een enzym in een bioreactor is weergegeven in Figuur 4. De belangrijkste instelbare factoren zijn daarbij de temperatuur en de substraatconcentratie. De respons is een percentage van de maximale opbrengst van het enzym. In Figuur 4A is de sequentiële optimalisering afgebeeld met het startdesign voor de methode van het steilste pad en de zoekrichting voor het optimum. De combinaties van temperatuur en druk waarbij de experimenten zijn uitgevoerd zijn aangegeven op het responsievlak dat overigens niet bekend is voor de onderzoeker. In Figuur 4B is de simultane optimalisering afgebeeld met een centraal composiet design dat in de buurt van het optimum is opgezet. Op basis van de meetresultaten voor dit design kan vervolgens het responsievlak worden berekend in de buurt van het optimum, zie Figuur 4C. Voor dit berekende responsievlak kan de positie van het optimum en de bijbehorende instellingen voor temperatuur en substraatconcentratie worden berekend. Het optimum ligt binnen de ellips met 92% van de maximale enzymopbrengst.

Dit artikel wordt vervolgd in Analyse 2 van 2020.



Figuur 4. Optimalisering van de productie van een enzym in een bioreactor; (A) Sequentiële optimalisering met het startdesign voor de methode van het steilste pad, gepositioneerd op het responsievlak, en de zoekrichting voor het optimum; (B) Simultane optimalisering met een centraal composiet design in de buurt van het optimum; (C) Berekend responsievlak met equiresponsielijnen in de buurt van het optimum

Referenties

1. *J.P.M. Andries, Chemometrie, 4e druk, Syntax Media, Utrecht, 2019, www.syntaxmedia.nl; ISBN 978 94 91764 332.*
2. *DAS rapport, Naar een datavaardige professional binnen de Applied Sciences, https://appliedscience.nl/publicaties/doi:10.1001/jamainternmed.2018.3519*
3. *J.P.M. Andries, H.M.J. Goldschmidt, L. Karreman, D.A.S. Pladdet, Methodevergelijking volgens Passing en Bablok, Analyse, september 2007, 208-210.*
4. *J.P.M. Andries, H.M.J. Goldschmidt, L. Karreman, D.A.S. Pladdet, Methodevergelijking volgens Passing en Bablok (deel 2), Analyse, november 2007, 267-270.*