

Reihe 17

Biotechnik/  
Medizintechnik

Nr. 293

M.Sc. Dipl.-Ing. (FH) Jan-Patrick Voß,  
Hamburg

## Anwendung spektroskopischer Messverfahren und Multivariater Datenanalyse zur Bewertung und Beobachtung von Bioprozessen



# **Anwendung spektroskopischer Messverfahren und Multivariater Datenanalyse zur Bewertung und Beobachtung von Bioprozessen**

Von der Naturwissenschaftlichen Fakultät der  
Gottfried Wilhelm Leibniz Universität Hannover

zur Erlangung des Grades  
Doktor der Naturwissenschaften (Dr. rer. nat.)

genehmigte Dissertation  
von  
Jan-Patrick Voß, M.Sc. Dipl.-Ing. (FH)  
geboren am 26.04.1984 in Hamburg

2017

Referent: Prof. Dr. rer. nat. Thomas Scheper

Koreferent: Prof. Dr. rer. nat. Bernd Hitzmann

Tag der Promotion: 31. Juli 2017

# Fortschritt-Berichte VDI

Reihe 17

Biotechnik/  
Medizintechnik

M.Sc. Dipl.-Ing. (FH) Jan-Patrick Voß,  
Hamburg

Nr. 293

Anwendung spektroskopischer  
Messverfahren und  
Multivariater Datenanalyse  
zur Bewertung und  
Beobachtung von Bioprozessen

VDI verlag

Voß, Jan-Patrick

## **Anwendung spektroskopischer Messverfahren und Multivariater Datenanalyse zur Bewertung und Beobachtung von Bioprozessen**

Fortschr.-Ber. VDI Reihe 17 Nr. 293. Düsseldorf: VDI Verlag 2017.

178 Seiten, 111 Bilder, 36 Tabellen.

ISBN 978-3-18-329317-9, ISSN 0178-9600,

€ 67,00/VDI-Mitgliederpreis € 60,30.

**Für die Dokumentation:** Bioprozessmonitoring – Spektroskopie – Multivariate Datenanalyse – Chemometrie – Process Analytical Technology – *Pichia pastoris* – Malariaimpfstoff

Die vorliegende Arbeit befasst sich mit dem Bioprozessmonitoring unter Verwendung spektroskopischer Messverfahren und multivariater Datenanalyse nach den Grundsätzen von PAT – *Process Analytical Technology*. Mit NIR-Spektroskopie und dem Verfahren *Soft Independent Modelling of Class Analogy* (SIMCA) wurde eine Qualitätsbewertung von Hefeextrakten realisiert. Im Vordergrund stand jedoch die Quantifizierung nicht direkt messbarer Größen aus NIR-, Raman- und 2D-Fluoreszenzspektren in pharmazeutischen Produktionsprozessen mit *Pichia pastoris*. Eine entsprechende Online-Bestimmung mit der Methode *Partial Least Squares Regression* (PLSR) kam weiterführend zur Regelung der Glycerolkonzentration zum Einsatz. Darüber hinaus wurde die Verwendung nichtspektraler Online-Daten zur Prozessbeobachtung erprobt. Dabei gelang mit Hilfe des nichtlinearen Verfahrens *Support Vector Regression* (SVR) unter anderem die Bestimmung zellspezifischer Reaktionsraten.

### **Bibliographische Information der Deutschen Bibliothek**

Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet unter <http://dnb.ddb.de> abrufbar.

### **Bibliographic information published by the Deutsche Bibliothek**

(German National Library)

The Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliographie (German National Bibliography); detailed bibliographic data is available via Internet at <http://dnb.ddb.de>.

© VDI Verlag GmbH · Düsseldorf 2017

Alle Rechte, auch das des auszugsweisen Nachdruckes, der auszugsweisen oder vollständigen Wiedergabe (Fotokopie, Mikrokopie), der Speicherung in Datenverarbeitungsanlagen, im Internet und das der Übersetzung, vorbehalten.

Als Manuskript gedruckt. Printed in Germany.

ISSN 0178-9600

ISBN 978-3-18-329317-9

Referent: Prof. Dr. rer. nat. Thomas Scheper

Koreferent: Prof. Dr. rer. nat. Bernd Hitzmann

Tag der Promotion: 31. Juli 2017

## Vorwort

Die vorliegende Arbeit entstand in der Zeit von Januar 2013 bis März 2017 an der Hochschule für Angewandte Wissenschaften (HAW) Hamburg im Forschungs- und Transferzentrum für Bioprozess- und Analysetechnik unter der Leitung von Prof. Dr.-Ing. Reiner Luttmann.

Ihm gilt mein herzlichster Dank für das Vertrauen, dass er in mich gesetzt hat. Für seinen ansteckenden Enthusiasmus und die wertvollen Hilfestellungen bei fachlichen Fragen sowie dafür so viel von ihm und durch die Arbeit in seinem Labor gelernt zu haben.

Herrn Prof. Dr. Thomas Scheper von der Leibniz Universität Hannover sowie Herrn Prof. Dr. Bernd Hitzmann von der Universität Hohenheim danke ich für die Übernahme der Referate.

Für finanzielle Unterstützung möchte ich mich bei der Leitung der Fakultät Life Sciences der HAW Hamburg bedanken. Außerdem wurde diese Arbeit durch das Bundesministerium für Bildung und Forschung gefördert (Förderkennzeichen 1756X09 und 031567).

Den Firmen Polytec und Kaiser Optical Systems danke ich für die Leihgabe von Spektrometersystemen sowie für technische Unterstützung. MKS data analytics solutions (ehemals Umetrics) gilt mein Dank für die Bereitstellung benötigter Softwarelizenzen.

Den Mitarbeitern des Labors BPA danke ich für die schöne Zeit und das freundschaftliche Arbeitsklima. Frau Prof. Dr. Gesine Cornelissen stand mir jeder Zeit mit konstruktivem Rat zur Seite. Die Laboringenieure Hans-Peter Bertelsen und Ulrich Scheffler haben durch ihre kollegiale und einsatzfreudige Unterstützung maßgeblich zum Erfolg der Arbeit beigetragen.

Besonderer Dank gilt Nina Mittelheuser, Vignesh Rajamanickam und Roman Lemke für den Einsatz im Rahmen ihrer Masterarbeiten und für die dabei erzielten Ergebnisse. Ohne Euch wäre diese Arbeit nicht möglich gewesen.

Allen weiteren lieben Kollegen, vor allem Roman, Sanja, Christian, Jessica, David, Sven, Jens, Kristof und Sarah danke ich auch dafür, dass der Spaß an der Arbeit nie verloren ging.

Zum Abschluss möchte ich meiner Familie und vor allem meiner Freundin Janet für das Verständnis, die mentale Unterstützung sowie jederzeit offene Ohren danken.

*I've done the math enough  
to know the dangers of our second-guessing*

Tool

**Inhaltsverzeichnis**

<b>1. Einleitung und Zielsetzung</b> .....	<b>1</b>
1.1 PAT – Ein Werkzeug moderner pharmazeutischer Produktion .....	1
1.2 Zielsetzung dieser Arbeit.....	2
<b>2. Mikro- und molekularbiologische Grundlagen</b> .....	<b>4</b>
2.1 Das potentielle Malariavakzin D1M1H als Zielprodukt.....	4
2.2 Das eingesetzte Expressionssystem .....	5
2.2.1 Die methylotrophe Hefe <i>Pichia pastoris</i> .....	5
2.2.2 Transformation des Organismus.....	6
<b>3. Eingesetzte Bioprozesstechnik</b> .....	<b>7</b>
3.1 Kultivierung von <i>Pichia pastoris</i> .....	7
3.1.1 Der klassische Herstellungsprozess rekombinanter Proteine .....	7
3.1.2 Intensivierung der Produktion durch zyklische Prozessführung .....	8
3.2 Die verwendete Bioreaktoreinheit .....	9
3.3 Erweiterte MSR- und Automatisierungstechnik .....	10
3.3.1 Automatisierungsaufgaben .....	10
3.3.2 Der substratlimitierte Glycerol-Fed-Batch .....	11
3.3.3 Online-Estimierung der Zelldichte .....	12
3.3.4 Inline-Messung und Regelung der Methanolkonzentration .....	13
3.3.5 Atline-Quantifizierung des Zielproduktes .....	13
3.3.6 Realisierung der zyklischen Fahrweise.....	14
3.3.7 Abgasanalyse und Gasbilanzen.....	14
3.4 Prozessbegleitende Offline-Analysentechnik.....	15
3.4.1 Probenahme.....	15
3.4.2 Zelldichtebestimmung .....	15
3.4.3 Glycerol- und Methanolanalytik.....	16
3.4.4 Ammoniummessung.....	16
3.4.5 Gesamtproteinbestimmung .....	16
3.4.6 Enzymatischer Alkoholoxidasenachweis .....	16
3.4.7 Berechnung von Konzentrationen in der Flüssigphase .....	17

<b>4. Spektroskopische Messverfahren als erweiterte PAT-Werkzeuge .....</b>	<b>18</b>
4.1 Eine Übersicht über PAT-Analysatoren.....	18
4.2 Eingesetzte Spektroskopieverfahren.....	19
4.2.1 Nahinfrarotspektroskopie.....	19
4.2.2 Raman-Spektroskopie.....	22
4.2.3 2D-Fluoreszenzspektroskopie.....	24
4.3 Extraktion relevanter Informationen aus Spektren.....	26
<b>5. Einführung in die Multivariate Datenanalyse (MVDA) .....</b>	<b>27</b>
5.1 Zielsetzung bei der Anwendung der MVDA.....	27
5.2 Datenvorbereitung.....	28
5.2.1 Struktur und Modifizierung multivariater Datensätze.....	28
5.2.2 Zentrierung und Skalierung.....	28
5.2.3 Datenvorverarbeitung für Spektren.....	29
5.3 Die Hauptkomponentenanalyse (PCA).....	31
5.3.1 Dimensionsreduktion durch Hauptkomponenten.....	31
5.3.2 Das mathematische Modell der PCA.....	32
5.3.3 Berechnung der Hauptkomponenten.....	33
5.4 Partial Least Squares Regression (PLSR).....	34
5.4.1 Multivariate Kalibrierung mittels PLSR.....	34
5.4.2 Das mathematische Modell der PLSR.....	35
5.4.3 Berechnung der PLS-Komponenten.....	37
5.5 Erstellung multivariater Modelle.....	39
5.5.1 Generelle Anforderungen an das Datenmaterial.....	39
5.5.2 Ausreißerdetektion für multivariate Daten.....	39
5.5.3 Variablenselektion.....	42
5.5.4 Validierung multivariater Modelle.....	43
5.6 Multivariate Klassifizierung.....	45
5.6.1 Allgemeine Informationen.....	45
5.6.2 Soft Independent Modelling of Class Analogy (SIMCA).....	45
5.6.3 Validierung von Klassifikatoren.....	47

5.7	Support Vector Machines (SVM).....	48
5.7.1	SVM als multivariates Klassifizierungsverfahren .....	48
5.7.2	Berechnung einer optimalen Trennebene .....	50
5.7.3	Kernel-Funktionen zur Abbildung nichtlinearer Beziehungen .....	52
5.7.4	Erweiterung zur Support Vector Regression (SVR).....	53
5.8	Eingesetzte MVDA-Software .....	55
<b>6.</b>	<b>Qualitätsbewertung von Hefeextrakten mit NIR-Spektroskopie .....</b>	<b>56</b>
6.1	Motivation und Zielsetzung .....	56
6.2	Stand der Wissenschaft .....	57
6.3	Der gewählte Messaufbau .....	57
6.4	Vorstellung des Probenpools .....	58
6.5	Vorverarbeitung der Spektraldaten .....	58
6.6	Explorative Datenanalyse und Probenselektion .....	58
6.7	Entwicklung eines SIMCA-Klassifizierungsmodells .....	61
6.8	Externe Validierung des Modells.....	64
<b>7.</b>	<b>Offline-Prädiktion relevanter Variablen in Bioreaktionsprozessen .....</b>	<b>66</b>
7.1	Eine Machbarkeitsstudie anhand von Offline-Analysen .....	66
7.2	Stand der Wissenschaft .....	66
7.3	Untersuchte Prozessgrößen und vorhandenes Probenmaterial.....	67
7.4	Entwicklung von PLSR-Modellen .....	68
7.4.1	Der komplexe Prozess der PLSR-Modellentwicklung.....	68
7.4.2	Eine exemplarische Darstellung bei der PLSR-Modellentwicklung .....	69
7.4.3	Schwierigkeiten bei Betrachtung der zellhaltigen Flüssigphase L .....	76
7.5	Prädiktion der Zelldichte.....	78
7.6	Prädiktion der Glycerolkonzentration .....	80
7.7	Prädiktion der Ammoniumkonzentration .....	85
7.8	Prädiktion der Gesamtproteinkonzentration.....	87
7.9	Prädiktion des zellinternen Alkoholoxidasegehaltes .....	90
<b>8.</b>	<b>Online-Monitoring mit spektroskopischen Verfahren.....</b>	<b>93</b>
8.1	Der untersuchte zweistufige Produktionsprozess .....	93
8.1.1	Die verwendete Bioreaktoranlage .....	93

8.1.2	Die parallel/sequentielle Prozessführung .....	94
8.1.3	Die erweiterte Prozess-EDV zur Anwendung der MVDA .....	96
8.2	Skizzierung des Versuchsaufbaus .....	98
8.3	Prädiktion der Zelldichte und der Glycerolkonzentration .....	99
8.3.1	Bereitstellung geeigneten Datenmaterials .....	99
8.3.2	Ergebnisdarstellung .....	100
8.4	Prädiktion des zellinternen Alkoholoxidasegehaltes .....	103
8.4.1	Bereitstellung geeigneter Kalibrierdaten .....	103
8.4.2	Ergebnisdarstellung .....	103
8.5	Prädiktion der Gesamtproteinkonzentration .....	104
8.6	Anwendung der nichtlinearen SVR .....	105
<b>9.</b>	<b>Regelung der Glycerolkonzentration .....</b>	<b>110</b>
9.1	Das Regelungskonzept .....	110
9.2	Theoretische Betrachtung des Regelungsproblems .....	111
9.2.1	Elemente des Regelkreises .....	111
9.2.2	Charakterisierung des Streckenverhaltens .....	111
9.2.3	Einführung des linearisierten Streckenmodells .....	114
9.2.4	Vernachlässigung der Dynamik des Messsystems .....	116
9.2.5	Regel- und Stellverhalten .....	117
9.3	Untersuchung der Dynamik des Regelungsproblems .....	118
9.3.1	Übertragungsfunktionen des Regelkreises .....	118
9.3.2	Eigenwerte des geschlossenen Regelkreises .....	119
9.3.3	Schwingungsverhalten des Regelkreises .....	121
9.3.4	Vorgabe des Regelkreisverhaltens .....	121
9.4	Technische Vorgaben der Glycerolregelung .....	122
9.4.1	Prozesstechnische Umsetzung .....	122
9.4.2	Bereitstellung der Regelgröße durch ein PLSR-Modell .....	123
9.4.3	Durchführung notwendiger Online-Berechnungen .....	124
9.4.4	Vernetzung beteiligter Softwaresysteme .....	124
9.5	Experimentelle Erprobung der Substratregelung .....	125
9.5.1	Regelung im aperiodischen Grenzfall .....	125
9.5.2	Regelung der Glycerolkonzentration im Schwingfall .....	127

<b>10. Anwendung der MVDA auf nichtspektroskopische Daten .....</b>	<b>129</b>
10.1 Prädiktion nicht direkt messbarer Prozessgrößen .....	129
10.2 Vorbereitung der Modellerstellung .....	130
10.2.1 Gewählte Zielgrößen .....	130
10.2.2 Berechnung zellspezifischer Reaktionsraten .....	131
10.2.3 Bereitstellung idealisierter Kalibrierdaten .....	132
10.2.4 Auswahl und Bereitstellung der Prädiktorvariablen .....	138
10.2.5 Erzeugung benötigter Datensätze und Datenvorverarbeitung .....	138
10.3 Ergebnisdarstellung .....	139
10.3.1 Prädiktion von Zustandsgrößen .....	139
10.3.2 Prädiktion zellspezifischer Reaktionsraten .....	143
<b>11. Zusammenfassung .....</b>	<b>147</b>
<b>12. Anhang .....</b>	<b>149</b>
12.1 Kulturmedien .....	149
12.2 Kultivierungsbedingungen .....	150
12.3 Reaktionskinetische Parameter .....	151
12.4 Offline-Messungen .....	151
<b>13. Quellenverzeichnis .....</b>	<b>153</b>
<b>Veröffentlichungen des Autors .....</b>	<b>160</b>

## Nomenklatur

### Allgemeine Abkürzungen

AOX	Alkoholoxidase
BMBF	Bundesministerium für Bildung und Forschung
BPRC	<i>Biomedical Primate Research Centre</i>
D1M1H	Potentiell Malariavakzin, Fusionsprotein aus <i>PfAMA1</i> und <i>PfMSP1</i>
DCU	Bioreaktor-Kontrolleinheit ( <i>Digital Control Unit</i> )
EMA	<i>European Medicines Agency</i>
FDA	<i>Food and Drug Administration</i>
HPLC	<i>High Performance Liquid Chromatography</i>
ICH	<i>International Conference on Harmonization</i>
IMAC	<i>Immobilized Metal chelate Affinity Chromatography</i>
MVDA	Multivariate Datenanalyse
NIR, NIRS	Nahinfrarot, Nahinfrarotspektroskopie
OPC	<i>Open Platform Communications</i>
PAT	<i>Process Analytical Technology</i>
PCA	Hauptkomponentenanalyse ( <i>Principle Component Analysis</i> )
<i>PfAMA1</i>	<i>Plasmodium falciparum</i> Apical Membrane Antigen 1
<i>PfMSP1</i>	<i>Plasmodium falciparum</i> Merozoite Surface Protein 1
PLSR	<i>Partial Least Squares Regression</i>
POD	Peroxidase
QbD	<i>Quality by Design</i>
RI	Brechungsindex ( <i>refractive index</i> )
RTR	<i>Real-Time Release</i>
SCADA	<i>Supervisory Control and Data Acquisition</i>
SIMCA	<i>Soft Independent Modelling of Class Analogy</i>
SNV	<i>Standard Normal Variate</i>
SVM, SVR	<i>Support Vector Machines, Support Vector Regression</i>
UV/Vis	Ultraviolett/Visuell, Wellenlängenbereich des elektromagnetischen Spektrums
WHO	<i>World Health Organization</i>

### Prozessgrößen

$A_K$	UV/Vis- oder NIR-Absorptionsmessung im Teilsystem K	[AU]
$a_{P1}$	Eigenwert der Produktbildungsrate	[h <sup>-1</sup> ]
$C_{IK}$	Stoffmengenkonzentration von Komponente I in Teilsystem K	[mol <sup>-1</sup> ]
$C_{P2K}$	AOX-Aktivität in Teilsystem K	[U <sup>-1</sup> ]
$c_{IK}$	Massenkonzentration von Komponente I in Teilsystem K	[g <sup>-1</sup> ]
$D_A$	Gesamtverdünnungsfaktor eines Aufschlussansatzes	[-]
d	Schichtdicke	[cm]
$E_{ABTS}$	Extinktionsmessung des Chromophors ABTS	[-]

$e_{S1L}$	Regeldifferenz der Glycerolkonzentration in der Flüssigphase	$[g\ l^{-1}]$
$E_L$	Trübungsmessung in der Flüssigphase	$[AU]$
$F_K$	Volumenstrom in/aus Teilsystem K	$[l\ h^{-1}]$
$F_{ni}$	Begasungsrate der Komponente I unter Normbedingungen	$[l\ h^{-1}]$
$g_{P2/X}$	zellspezifische AOX-Aktivität	$[Ug^{-1}]$
$I_K$	Raman-Intensitätsmessung in Teilsystem K	$[IU]$
$J_I$	Gütekriterium bei Anpassung einer Komponente I	$[div.]$
IUR	volumetrische Aufnahme­rate der Komponente I	$[g\ l^{-1}\ h^{-1}]$
$K_J$	Verstärkungsfaktor eines Teilsystems J	$[div.]$
$k_{La}$	volumetrischer Sauerstofftransferkoeffizient	$[h^{-1}]$
$k_{S1}$	Monod-Limitierungskonstante für Glycerol	$[g\ l^{-1}]$
$M_I$	Molare Masse von Komponente I	$[g\ mol^{-1}]$
$m_{IK}$	Masse von Komponente I in Teilsystem K	$[g]$
$m_K$	Masse von Teilsystem K	$[g]$
$N_{St}$	Rührerdrehzahl	$[min^{-1}]$
$p_G$	Überdruck in der Gasphase	$[bar]$
pH	pH-Wert	$[-]$
$pO_2$	relativer Gelöstsauerstoffpartialdruck	$[%]$
$Q_I$	volumetrischer Massenstrom von Komponente I	$[g\ l^{-1}\ h^{-1}]$
$q_{I/X}$	zellspezifische Reaktionsrate der Komponente I	$[h^{-1}]$
R	Korrelationskoeffizient	$[-]$
RQ	Respirationsquotient	$[mol\ mol^{-1}]$
$S_K$	Fluoreszenzmessung in Teilsystem K	$[RFU]$
s	Eigenwert	$[h^{-1}]$
t	Zeit	$[h]$
$T_J$	Zeitkonstante eines Teilsystems J	$[h]$
$V_K$	Volumen des Teilsystems K	$[l]$
$V_{nM}$	Molares Normvolumen	$[l\ mol^{-1}]$
$x_{IG}$	Stoffmengenanteil von Komponente I in der Gasphase	$[-]$
$y_{I/J}$	Ausbeutekoeffizient der Komponente I aus Komponente J	$[g\ g^{-1}]$
$\alpha_{Z/X}$	Massenverhältnis von feuchten zu trockenen Zellen	$[g\ g^{-1}]$
$\epsilon_{ABTS}$	molarer Extinktionskoeffizient von ABTS	$[l\ \mu\ mol^{-1}\ cm^{-1}]$
$\vartheta$	Dämpfungsgang	$[-]$
$\vartheta_L$	Temperatur in der Flüssigphase	$[^{\circ}C]$
$\lambda$	Wellenlänge	$[nm]$
$\mu$	Zellwachstumsrate	$[h^{-1}]$
$\nu$	Wellenzahl	$[cm^{-1}]$
$\rho_K, \rho_Z$	Dichte des Teilsystems K, Dichte feuchter Zellen	$[g\ l^{-1}]$
$\sigma$	Abklingkonstante	$[h^{-1}]$
$\omega$	Kreisfrequenz	$[h^{-1}]$

**Prozessrelevante Indizes****Komponenten**

Ac	Säure ( <i>acid</i> )	P1	D1M1H (Zielprodukt)
AF	Antischaum ( <i>anti foam</i> )	P2	Alkoholoxidase (AOX)
AIR	Luft	Ptot	Gesamtprotein
Al	Base ( <i>alkali</i> )	S1	Glycerol (Substrat 1)
C, CO2	Kohlenstoffdioxid	S2	Methanol (Substrat 2)
N2	Stickstoff	X	Biotrockenmasse
O, O2	Sauerstoff	Z	Biofeuchtmasse

**Teilsysteme**

A	Analyse, Aufschlussansatz	R1	Glycerolvorlage
B	Puffersystem	R2	Methanolvorlage
G	Gasphase	R3	Medienvorlage
H	Ernte ( <i>harvest</i> )	S	Probe ( <i>sample</i> )
L	Flüssigphase (Kulturbrühe)	T1	Säurevorlage
M	Medienphase (Überstand)	T2	Basevorlage

**Laufindizes**

i	Substrate, Zeitpunkte	p	Reaktoren
j	Zeitpunkte	u	Zeitpunkte
k	Batches/Zyklen, Zeitpunkte		

**Zustände, Orte**

0	Anfangsbedingung	n	Normbedingungen
ap	aperiodisch	nir	Nahinfrarot
at	atline	off	offline
cdw	Biotrockenmasse	on	online
em	Emission	op	Arbeitspunkt ( <i>operating point</i> )
est	estimiert	P	proportional, Periodendauer
ex	Anregung ( <i>excitation</i> )	pls	vorhergesagt mit PLSR
fia	Fließinjektionsanalyse	R	Regler
flu	2D-Fluoreszenz	ram	Raman
gr	Wachstumsanteil ( <i>growth</i> )	rel	relativ
hplc	gemessen mit HPLC	S	Strecke
l	Integration	svr	vorhergesagt mit SVR
in	Eingang ( <i>inlet</i> )	tot	total
m	Erhaltungsstoffwechsel	turb	Trübung ( <i>turbidity</i> )
max	maximal	w	Sollwert
min	minimal	zi	Störgröße

**Matrizen und Vektoren in der MVDA**

B	$(m \times v)$	PLS-Regressionskoeffizientenmatrix
$b_j$	$(m \times 1)$	Spaltenvektor mit PLS-Regressionskoeffizienten
D	$(n \times m)$	Datenmatrix (Spektren oder Prozessdaten)
$d_j$	$(n \times 1)$	Spaltenvektor der Datenmatrix D
E	$(n \times m)$	Residuenmatrix des X-Datenraums
F	$(n \times v)$	Residuenmatrix des Y-Datenraums
G	$(n \times v)$	Residuenmatrix des PLS-Regressionsansatzes
M	$(n \times v)$	Messdatenmatrix (Zielgrößen bei der PLSR)
P	$(m \times r)$	<i>Loading</i> matrix des X-Datenraums
$p_l$	$(m \times 1)$	<i>Loading</i> -Spaltenvektor der Komponente l (X-Daten)
Q	$(v \times r)$	<i>Loading</i> matrix des Y-Datenraums
$q_l$	$(v \times 1)$	<i>Loading</i> -Spaltenvektor der Komponente l (Y-Daten)
T	$(n \times r)$	<i>Score</i> matrix des X-Datenraums
$t_l$	$(n \times 1)$	<i>Score</i> -Spaltenvektor der Komponente l (X-Daten)
U	$(n \times r)$	<i>Score</i> matrix des Y-Datenraums
$u_l$	$(n \times 1)$	<i>Score</i> -Spaltenvektor der Komponente l (Y-Daten)
w	$(m \times 1)$	Wichtungs-Spaltenvektor der Entscheidungsfunktion (SVR)
W	$(m \times r)$	<i>Weight</i> matrix eines PLSR-Modells
$w_l$	$(m \times 1)$	<i>Weight</i> -Spaltenvektor der Komponente l
X	$(n \times m)$	modifizierte Datenmatrix (Spektren oder Prozessdaten)
$x_j$	$(n \times 1)$	Spaltenvektor der Datenmatrix X
Y	$(n \times v)$	modifizierte Messdatenmatrix (Zielgrößen bei der PLSR und SVR)
$y_h$	$(n \times 1)$	Spaltenvektor der Messdatenmatrix Y

**MVDA-relevante Bezeichnungen**

b	Bias der Entscheidungsfunktion (SVR)
C	Güteparameter zur Wichtung der Fehler bei der SVR
c	Koeffizient der inneren Beziehung eines PLSR-Modells
DModX	Distanz zum Modell im X-Datenraum (Ausreißerdetektion)
$E_C$	Klassifizierungsfehler [%]
F	Wert einer F-Verteilung
g	Trennebene, Entscheidungsfunktion (SVR)
K	<i>Kernel</i> -Funktion (SVR)
$L, L_d$	Lagrange-Funktional (SVR)
m	Anzahl an Variablen/Spalten in der Datenmatrix X
n	Anzahl an Zeilen in der Datenmatrix X
$n_K$	(allgemein) Anzahl an Beobachtungen in der Gruppe K
r	Anzahl an Komponenten/Spalten in der <i>Score</i> matrix T
$R_{cv}^2$	Güte der Vorhersage der Kreuzvalidierung ( <i>cross validation cv</i> )

$R_p^2$	Güte der Vorhersage der externen Validierung ( <i>prediction P</i> )
$R_x^2$	Güte der Anpassung/erklärter Anteil der Varianz (Datenraum X)
$R_y^2$	Güte der Anpassung/erklärter Anteil der Varianz (Datenraum Y)
RMSEcv	Vorhersagefehler bei der Kreuzvalidierung [div.]
RMSEP	Vorhersagefehler bei der externen Validierung [div.]
$r_{Tol}$	Radius einer Hotelling $T^2$ -Ellipse in Richtung der Komponenten l
S	Spanne zwischen den Eingangsdaten (SVR)
$s_{di}$	Zeilenstandardabweichung bezogen auf Datenmatrix D
$s_{dj}$	Spaltenstandardabweichung bezogen auf Datenmatrix D
Se	Sensitivität eines Klassifikators
Sp	Spezifität eines Klassifikators
SSY	Maß für den Anteil erklärter Varianz des Y-Datenraums
$T^2$	Hotelling $T^2$ -Wert (Ausreißerdetektion)
v	Anzahl an Variablen/Spalten in der Datenmatrix Y
VIP	Variable Importance on Projection (Variablenselektion)
$\alpha$	Irrtumswahrscheinlichkeit einer statistischen Hypothese
$\alpha_i, \beta_i$	Lagrange-Multiplikatoren der Beobachtung i (SVR)
$\beta_K$	Wahrscheinlichkeit der Zugehörigkeit zu einer Klasse K
$\gamma$	Parameter des Gauß'schen RBF-Kernels (SVR)
$\varepsilon$	Fehlertoleranz-Parameter bei der SVR

### MVDA-relevante Indizes

#### Laufindizes

i	Beobachtungen, Objekte	l	Haupt-/PLS-Komponenten
j	X-Variablen	z	Beobachtungen (SVR)
h	Y-Variablen		

#### Weitere Indizes und Zustände

–	Arithmetrischer Mittelwert	opt	optimal
^	Modellschätzwert	p	positiv (Klassenzugehörigkeit)
abs	absolut	nor	normiert
av	Mittelwert ( <i>average</i> )	n	negativ (Klassenzugehörigkeit)
crit	kritischer Wert	mc	mittenzentriert
CS	Kalibrierdatensatz	f	falsch (Klassifizierung)
cv	Kreuzvalidierung	de1	erste Ableitung
P	Vorhersage ( <i>prediction</i> )	t	wahr (Klassifizierung)
PS	Vorhersagedatensatz	sv	Stützvektor ( <i>support vector</i> )
rel	relativ	tot	total
snv	SNV-gefiltert	uv	autoskaliert ( <i>unit variance</i> )
te	temporär	VS	Validierdatensatz

## Kurzfassung

Jan-Patrick Voß

### Anwendung spektroskopischer Messverfahren und Multivariater Datenanalyse zur Bewertung und Beobachtung von Bioprozessen

Die *Process Analytical Technology* (PAT) Initiative empfiehlt den Einsatz fortschrittlicher Analysensysteme in der pharmazeutischen Produktion. Die vorliegende Arbeit beschreibt wichtige Schritte bei der Etablierung von spektroskopischen Messverfahren in Bioprozessen.

Ein erstes Ziel war die Entwicklung eines Qualitätsbewertungsverfahrens für Hefeextrakte, basierend auf NIR-Spektroskopie und multivariater Klassifizierung. Die gewählte Methode *Soft Independent Modelling of Class Analogy* (SIMCA) führte mit einem Klassifizierungsfehler von 1,5 % zu einem sehr guten Ergebnis.

Ein überwiegender Teil dieser Arbeit widmete sich der Beobachtung zyklischer Kultivierungsprozesse der methylophilen Hefe *Pichia pastoris* zur Herstellung des potentiellen Malaria-vakzins D1M1H. Hierbei wurde die Quantifizierung von Zelldichte und AOX-Gehalt sowie Glycerol-, Ammonium- und Produktkonzentration mit *Partial Least Squares Regression* (PLSR), basierend auf NIR-, Raman- und 2D-Fluoreszenzspektren zunächst offline erprobt.

Im Anschluss erfolgte eine Übertragung der erarbeiteten Methoden auf den Online-Betrieb. Dabei kam neben der PLSR auch das nichtlineare Verfahren *Support Vector Regression* (SVR) zum Einsatz. Dieses verbesserte unter anderem die Bestimmung der Glycerolkonzentration mit Raman-Spektroskopie und erreichte einen Vorhersagefehler von ca. 3 %.

Zur Regelung von Glycerol wurden Raman-Spektren und PLSR erfolgreich eingesetzt und damit die technische Relevanz der multivariaten Kalibrierung über eine reine Prozessbeobachtung hinaus demonstriert. Die Realisierung erfolgte mit einer Störgrößenaufschaltung und einer ergänzenden adaptiven Regelung der Abweichungen vom Sollwert.

Bei der technischen Umsetzung kamen mit SIMATIC SIPAT, Umetrics SIMCA und MATLAB® eine Reihe industrierelevanter Softwarepakete innerhalb einer komplexen Prozess-EDV in einem zweistufigen Produktionsprozess zum Einsatz.

Den Abschluss dieser Arbeit bildet die Anwendung multivariater Kalibrierverfahren auf nicht-spektroskopische Online-Prozessgrößen. Hierbei wurden die obigen biotechnologischen Variablen über 15 klassische Online-Messgrößen eines Bioreaktorprozesses ermittelt. Eine Bereitstellung geeigneter Trainingsdaten erfolgte durch Simulation und Parameteridentifikation von Modellen mit Massenbilanzen zur Glättung von Offline-Analysen.

Die daraus resultierende Bestimmung der Zielproduktkonzentration erreichte mit der nicht-linearen SVR einen Vorhersagefehler von 3,2 %. Die Quantifizierung zellspezifischer Reaktionsraten für Zellmasse, Glycerol, Methanol, Sauerstoff und Produkt war auf diesem Wege ebenfalls erfolgreich. Mit einem Fehler von 2,7 % war die Prädiktion der Aufnahme von Methanol am besten.

Schlüsselworte: Bioprozessmonitoring; Spektroskopie; Multivariate Datenanalyse; Chemo-metrie; Process Analytical Technology; *Pichia pastoris*; Malariaimpfstoff

## Abstract

Jan-Patrick Voß

### **Application of spectroscopic measurement methods and multivariate data analysis for evaluation and monitoring of bioprocesses**

The *Process Analytical Technology* (PAT) initiative proposes the application of advanced analysis systems in pharmaceutical production. This work describes important steps in establishing spectroscopic measurement methods in bioprocesses.

A first objective was the development of a quality evaluation system for yeast extracts based on NIR-spectroscopy and multivariate classification. The selected method *Soft Independent Modelling of Class Analogy* (SIMCA) gave excellent results by obtaining a classification error of 1.5 %.

Most of the work was dedicated to the monitoring of cyclic cultivation processes of the methylotrophic yeast *Pichia pastoris* for production of the potential malaria vaccine D1M1H. First, quantification of cell density, AOX content as well as glycerol, ammonia and product concentration with *Partial Least Squares Regression* (PLSR) based on NIR-, Raman- and 2D-fluorescence spectra was tested off-line.

A transfer of developed methods to the on-line operation took place subsequently. Besides PLSR, the non-linear method *Support Vector Regression* (SVR) was also used on-line and enabled the improvement of glycerol determination with Raman-Spectroscopy. A prediction error of approx. 3 % was achieved.

For glycerol control, Raman spectra and PLSR were successfully applied, thus demonstrating the technical relevance of multivariate calibration beyond mere process monitoring. The implementation was carried out with a feed-forward control and an additional adaptive compensation of the deviations from the setpoint.

A complex data processing system, involving software with industrial relevance like SIMATIC SIPAT, Umetrics SIMCA and MATLAB<sup>®</sup>, was used for the technical implementation in a two-stage production process.

In conclusion of this work, multivariate calibration methods were applied to non-spectroscopic on-line process data. Here, the biotechnological variables mentioned above were determined using 15 classical on-line measurements of a bioreactor process. Suitable training data was provided by smoothing off-line measurements via simulation of models with mass balances and parameter identification.

The resulting determination of target protein concentration with the non-linear SVR obtained an error of 3.2 %. In addition, quantification of cell specific reaction rates for cell mass, glycerol, methanol, oxygen and product was also successful in this context. With an error of 2.7 %, prediction of methanol uptake performed the best.

**Keywords:** bioprocess monitoring; spectroscopy; multivariate data analysis; chemometrics; Process Analytical Technology; *Pichia pastoris*; malaria vaccine