UNIVERSITY OF
NEWCASTLE UPON TYNE

Spectroscopic and Process Data Fusion:

Enhanced Monitoring of an Industrial Fermentation

by
Sophia Triadaphillou

A Thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

School of Chemical Engineering and Advanced Materials
The University of Newcastle upon Tyne

December 2005

For my grandmother,

whom I will always remember for her warmth

PUBLICATIONS

CONFERENCE PAPERS

1. Triadaphillou, S., Martin, E.B., Montague, G., Jeffkins, P., Stimpson S., Nordon A. 'Industrial experiences in on-line analytical spectroscopic methods for bioreactor monitoring and control'. CPC (Chemical Process Control) 7. January 2006. Lake Louise, Alberta. Canada.

2. Triadaphillou, S., Martin, E.B., Montague, G., Jeffkins, P., Stimpson S., Nordon A., 'Spectral window selection for on-line spectroscopic analysis of fermentation progression'. 32nd FACSS (Federation of Analytical Chemistry and Spectroscopy Societies) and 51st ICASS (International Conference on Analytical Sciences and Spectroscopy). October 2005. Quebec City, Canada.

3. Triadaphillou, S., Martin, E.B., Montague, G., Jeffkins, P., Stimpson S., Nordon A. 'Enhanced modeling of an industrial fermentation process through data fusion techniques'. ESCAPE (European Symposium on Computer Aided Process Engineering) 15 conference. May 2005. Barcelona, Spain.

4. Triadaphillou, S., Martin, E.B., Montague, G., Jeffkins, P., Stimpson S., Nordon A. 'The conjunction and analysis of multiple spectral data forms: An application to a fermentation process'. APACT 05. April 2005. Birmingham, UK

5. Triadaphillou, S., Martin, E.B., Montague, G., Jeffkins, P., Stimpson S., Nordon A. 'Monitoring of a fermentation process through the on-line spectroscopic data and the conjunction of spectroscopic and process data'. BatchPro Symposium. June 2004. Poros, Greece.

6. Triadaphillou, S., Martin, E.B., Montague, G., Morris, A. J., Nordon A., Jeffkins, P., Stimpson S., 'A novel strategy for on-line spectral data analysis: application to an antibiotic production process'. APACT 04. April 2004. Bath, UK.

7. Triadaphillou, S., Wells, I., Morris, A. J., Martin, E.B. 'Investigation of calibration-free resolution techniques and independent component analysis'. ADCHEM (International Symposium on Advanced Control of Chemical Processes). January 2004. Hong Kong

8. Triadaphillou, S. 'The Application of Independent Component Analysis to Chemical Reactions'. APACT 03. April 2003. York, UK.

9. Triadaphillou, S., Morris, A.J., Martin, E.B. 'Application of independent component analysis to chemical reactions'. Independent Component Analysis Symposium. April 2003. Nara, Japan.

10. Triadaphillou, S., Martin, E.B., Morris, A. J., Wells, I. 'Taking quality assurance from the laboratory onto the process'. "SET for EUROPE". December 2002. House of Commons, London, UK.

11. Triadaphillou, S., Morris, A.J., Martin, E.B., Wells, I., Polwert, E. 'Calibration-Free Resolution Techniques. Application to Synthetic and Real Data Sets'. Advances in Process Analytics and Control Technology (APACT) 02. April 2002. Royal Society of Edinburgh, Edinburgh, UK.

JOURNAL PAPERS

1. Triadaphillou, S., Martin, E.B., Montague, G., Jeffkins, P., Stimpson S., Nordon A. 'Improved Fermentation Process Monitoring: The conjunction of Spectral and Process Data'. To be submitted to 'Biotechnology and Bioengineering'.

2. Triadaphillou, S., Martin, E.B., Montague, G., Jeffkins, P., Stimpson S., Nordon A. 'A novel strategy for on-line spectral data interpretation: Application to an antibiotic production process'. To be submitted.

3. Triadaphillou, S., Martin, E.B., Morris, A. J., Wells, I. 'Calibration-free resolution techniques and their applications for the determination of spectral and concentration profiles in unknown mixtures'. To be submitted.

ABSTRACT

Large scale manufacturing of pharmaceutical products is a highly competitive industry in which technological improvements can maintain fine business margins in the face of competition from those with lower manufacturing overheads. Processes in which pharmaceuticals are produced via fermentation are particularly susceptible to large variability and reduced productivity due to natural variation and limited monitoring and control options. The latest monitoring methods offer the potential to understand causes of variation, improve productivity and as a result maintain the competitive edge. Unfortunately the fermentation environment is not conducive to the implementation of instrumentation. This thesis shows how signals from spectral instruments can be enhanced by other process and spectroscopic measurements, to provide on-line measurements of critical broth concentrations traditionally only available from infrequent off-line analysis.

Near infrared (NIR) and Mid infra red (MIR) spectral analysis of fermentation broth can provide measurements of key concentrations throughout a batch. The off-line analysis of broth samples is more straightforward but on-line implementation is possible. In the case of on-line implementation, the quality of information is compromised, placing greater demands on the signal interpretation methods. The objective of the thesis was to understand the causes of process variation and compensate for them during batch progression, consequently on-line implementation was essential.

The construction of a robust calibration model for individual instruments is the first step in implementation. The traditional strategy is either to use multivariate techniques such as projection to latent structures (PLS) or wavelength selection through genetic algorithms followed by PLS. An alternative approach is developed where a search strategy identifies a limited number of spectral windows (SWS) that are most descriptive of the concentrations of interest. The benefit of using SWS is that problems associated with over-fitting the calibration model construction data are minimised. This is particularly important in a development environment where the number of batches is limited. The random nature of the search strategy of the SWS

algorithm results in a range of calibration models. Multiple calibration models are 'stacked' to provide improved accuracy and robustness. It is demonstrated that stacking provides an improved prediction capability compared to selecting the single 'best' performing model. Finally, developing calibration models for sub-regions of fermentation operation is contrasted with a global model.

The improvement in accuracy of measurements from SWS and stacking is significant but errors in the determination of the concentration of some compounds remained significant. To overcome these offsets, a model relating the calibration residuals to on-line process measurements was constructed using PLS. The model was then used to correct the spectral calibration prediction to result in improved determination of broth concentrations. The fermentation monitoring methodology is demonstrated by application to an industrial antibiotic production process. Corrected predictions of product concentration and broth nutrient levels demonstrate that combining multiple information sources is advantageous from a measurement perspective.

ACKNOWLEDGEMENTS

I would like to thank my supervisors Professor Elaine Martin and Professor Gary Montague for their valuable support, encouragement, trust and more than anything for their patience, Dr Suresh Thenadil for all his valuable advice and Angela Bott for everything - above all her smile.

I am grateful to Paul Jeffkins, Sarah Stimpson and Alison Dann from GSK, Worthing who provided the industrial fermentation data and for all their help and guidance. Alison Nordon and David Littlejohn from Strathclyde University have helped considerably in terms of the analysis and interpretation of the results.

I would like to acknowledge the support of the vendor companies, Clairet Scientific, Foss and Spectaprobe for the loan of the spectroscopic instrumentation.

I am also grateful to the Engineering and Physical Sciences Research Council (EPSRC), BatchPro network and the Centre for Process Analysis and Control Technology (CPACT) for financial support.

I would like to thank all the friends I have made during the time I stayed in Newcastle for all the friendship, the strength and inspiration they offered that made these five years unforgettable. My life had been an amazing emotional roller coaster. Chris, Irene, Alex, TK, Owen, Elias, Pol, Ahmed, Marco, Effi, Bahar, Daniel, Maria Alejandra, Mahesh, Mario, Claire, Kier, Dave, Gwen, Mark, Pieter, Katarina, Vanessa, Hiromi, George, Ilia thank you from my heart. Especially I would like to thank my senseis Joe Currant, Jason Steggles, John Martinez, Charles Oughton and Steve Nash who all the lessons of life they have taught me.

Finally, I would like to thank my parents, Eleftherios and Euridiki, and brother, Asimakis, for all their support and love and for always being there for me no matter what.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS AND ACRONYMS

ANN	artificial neural networks
ATR	attenuated total reflexion
biPLS	backward interval PLS
CER	carbon dioxide evolution rate
CSR	cyclic subspace regression
DT	de-trend
DoE	Design of Experiment
DOSC	direct orthogonal signal correction
EMSC	extended multiplicative scatter correction
GA	genetic algorithm
IVS-PLS	interactive variable selection
IPW-PLS	iterative predictor weighting PLS
iPLS	interval PLS
ISE	iterative stepwise elimination
LV	Latent Variables
MIR	mid infrared
MLR	multiple linear regression
MSC	multiplicative scatter correction
mwPLS	moving window PLS
NIR	near infrared
NIPALS	nonlinear iterative partial least squares
NN	neural network
OPLS	orthogonal projection to latent squares
OSC	orthogonal signal correction
OUR	oxygen uptake rate
PAT	process analytical technology
PCR	principal component regression
PLS	partial least squares or projection to latent squares
POSC	projected orthogonal signal correction
REP	relative error of prediction

RMS	root mean square
RMSET	root mean square error training
RMSEV	root mean square error validation
RNV	robust normal variate
RQ	respiratory quotient
SIS	spectral inference subtraction
siPLS	synergy interval PLS
SG	Savitzy-Golay
SNV	standard normal variate
SPA	successive projections algorithm
SWS	spectral window selection
UVE-PLS	uninformative variable elimination in PLS modeling
UVE-a	UVE with a% cut-off threshold

NOMENCLATURE

ENGLISH CHARACTERS

α_i	coefficients in binning method, determined using least squares.
A_i	bin areas in binning method
b_i	regression coefficient for i -th latent variable in PLS
b	path length
B_{win}	matrix of the coefficients at the PLS stacking
c	speed of light
C	pure column concentration profiles in MCR-ALS method or concentration matrix in MLR
CER	carbon dioxide evolution rate
d	light diffusion spectrum at MSC equation
D	augmented data matrix in MCR-ALS method
e	residual vector
E	matrix of errors or residuals
$\mathbf{E}_{MB}$	residuals for the predictors, $\mathbf{X}_{MB}$ in multi-block approach
E_{win}	the matrix of the errors associated with the model at the PLS stacking
f_j	non-linear function in non-linear PLS
$\mathbf{F}_{MB}$	residuals for the quality variable, **y** in multi-block approach
F	residuals from **Y** matrix in PLS
G_{in}	aeration rate based on inlet flow-rate ($^{mol}/_{s}$)
h	Planck's constant
$\%i^{in}$	mole percentage of component i in the inlet air to the fermenter
$\%i^{out}$	mole percentage of component i in the exit gas from the fermenter
l	number of wavelengths at SNV_i equation
I_R	reference intensity
I_S	sample intensity
k	number of latent variables retained in non-linear PLS

m	slope
n_{win}	number of models at the stacking approach
OUR	oxygen uptake rate
p_i	i'th loading vector in PCR
$\mathbf{P}$	matrix of $\mathbf{X}$ loadings
$\mathbf{P}$	$(p \times n)$ coefficient matrix in MLR
$\mathbf{P}_{MB}$	loading for $\mathbf{X}_{MB}$ in multi-block approach
$\mathbf{P}_{stack}$	final weighted predictions at the PLS stacking
$\mathbf{P}_{win}$	matrix formed by the n individual models at the PLS stacking
$\mathbf{q}$	$\mathbf{Y}$ loadings in PLS
$\mathbf{q}_j$	weight vector of $\mathbf{Y}$ in non-linear PLS
$\mathbf{q}_{MB}$	vector of $\mathbf{y}$ loadings in multi-block approach
Q	end-use parameter in binning method
r_i	absorbance of a spectrum at at wavelength i at SNV_i equation
$\bar{r}$	average of all the absorbance values across all the wavelengths at SNV_i equation
R^2	multiple correlation coefficient
R^2adj	adjusted multiple correlation coefficient
RQ	respiratory quotient
$\mathbf{S}^T$	pure row spectral signal profiles in the MCR-ALS method
$\mathbf{S}$	$(m \times p)$ spectral response matrix in MLR
SNV_i	standard normal variate of a spectrum at wavelength i
$\mathbf{t}$	scores of $\mathbf{X}$ matrix in PLS
t	time
$\mathbf{t}_j$	latent variables of process variables in PLS
t_i	i'th score vector in PCR
T	transmittance
$\mathbf{T}_{MB}$	scores for $\mathbf{X}_{MB}$ in multi-block approach
$\mathbf{T}_k$	k principal components retained in PCR
$\mathbf{u}_j$	scores of $\mathbf{Y}$ in PLS or latent variables of quality variables in PLS

$w(MW)$	molecular weight of polymer
$\mathbf{w}_j$	weight vector of $\mathbf{X}$ in non-linear PLS
$\mathbf{w}$	weights in PLS
w_k	vector of weights in PLS
x_{new}	resulting correcting spectrum at DT equation
x_{ref}	reference spectrum at MSC equation
xc	concentration of the cells
x_{cor}	corrected spectrum at MSC equation
x	sample spectrum at MSC equation
x_i	concentration of the analyte in sample i
x_{pi}	estimated concentration of the analyte in the sample i
$\bar{x}$	mean of the concentration in the training set
$\mathbf{X}$	spectral matrix or matrix of independent variable
$\mathbf{X}_{MB}$	multiblock matrix
$\mathbf{Y}$	quality variable matrix
$\mathbf{y}$	vector of dependent variable
$\hat{\mathbf{y}}$	vector of fitted values
$\hat{\mathbf{y}}_{stack}$	stacked neural network

GREEK CHARACTERS

a	offset
a_i	polynomial coefficient in DT equation
A	Absorbance matrix
$\boldsymbol{\beta}$	vector of regression coefficients
$\hat{\boldsymbol{\beta}}$	the estimate of the regression coefficient $\boldsymbol{\beta}$
$\varepsilon(\lambda)$	the molar absorptivity of the analyte for the specific wavelength λ
E	Energy
θ	PCR parameter
λ	wavelength

μ	specific growth rate
ν	frequency of the electromagnetic radiation
ω	stacking weight vector

CHAPTER 1

INTRODUCTION

In a competitive market significant emphasis is placed on product quality as it serves to help retain and, in the longer term, increase market share. Consequently effective monitoring and control procedures have become important tools to ensure the delivery of consistent and high quality product. Even though a range of monitoring and control techniques are potentially applicable, according to Montague (1997), only those that have a high benefit to cost ratio are likely to be implemented. One example is advanced instrumentation techniques.

More specifically, batch processes play an important role in the manufacture of high quality products such as polymers, chemicals, foods and pharmaceuticals. One of the major challenges in batch processing is to determine, monitor and control current process conditions as significant changes can occur throughout the duration of the batch. Measurement limitations have led, in many cases, to control strategies based on inferential model developed from off-line assays. These approaches are compromised in terms of their applicability by their low sampling frequency.

Fermentation processes are generally manufactured using a batch or fed-batch process, thus to enhance the monitoring and control of such processes, analytical instrumentation is a core requirement. Traditionally, on-line monitoring and control has been limited to environmental parameters including pH, dissolved oxygen and temperature, with critical media concentrations only being available through off-line assays. Typically the samples are recorded infrequently with the results from the laboratory analysis being returned after some delay. As a consequence of the sampling frequency and measurement delay, control of the process is limited. Furthermore, the complexity of the underlying reactions, together with the natural biological variability, necessitates the implementation of a more responsive on-line control strategy. A shift to the on-line analysis of broth concentrations offers major control opportunities, with Process Analytical Technology (PAT) being a key enabling technology (Lopes *et al.*, 2004).

Previous research has demonstrated that infrared (IR) spectral analysis of fermentation broth can provide on-line measurements of key concentrations throughout the duration of a batch but signal interpretation remains a challenge. Near infrared and mid infrared spectroscopy have provided a rich source of information

relating to the conditions within a process through their ability to detect and determine the concentration of chemical constituents. Consequently they have been implemented for the monitoring and control of chemical and biochemical processes (for example Sivakesava *et al.*, 2001). Relating the spectra to the analyte of interest requires the construction of a robust calibration model. One approach is through the application of multivariate data analysis techniques, thereby enabling the extraction of latent features inherent within the data. Considerable research in this area has been undertaken (Jouan-Rimbaud *et al.*, 1995; Westerhuis *et al.*, 2000; Arnold *et al.*, 2002; Roggo *et al.*, 2003) with the most common solutions have been based on linear Partial Least Squares (PLS), (Geladi and Kowalski, 1986; Frank and Friedman, 1993).

One of the issues in the construction of a calibration model is whether to include all the wavelengths or a selection of them in the development of a robust calibration model. Wavelength selection is one approach to eliminating those wavelengths where descriptive information relating to the analyte concentration is uninformative. The determination of the optimal wavelength subset is challenging as a consequence of the large number of correlated spectral measurements recorded from a process in which a series of complex biochemical reactions occur. Additionally absorbance ranges of different functional groups may overlap and substances contained in the complex mixture may contribute to signals that are spread across the spectral range. Several methods have been proposed to select key wavelengths or to eliminate wavelengths (Brown *et al.*, 1991, Jouan-Rimbaud *et al.*, 1996, Swierenga, *et al.*, 1998, Bakken *et al.*, 1999, Archibald and Akin, 2000, Smith and Gemperline, 2000). It is claimed in these citations that individual wavelength selection will give better predictions than when using the full set. Wavelength selection algorithms can been grouped into three categories according to the search method employed: a) dimension-wise selection algorithms, b) model-wise elimination algorithms and c) subset selection algorithms. Genetic algorithms belong to the latter category and are one of the more commonly applied wavelength selection approaches, (Abrahamsson *et al.*, 2003).

The methods mentioned previously can be biased towards including those wavelengths with a chance correlation to the prediction property. Thus the selected wavelength subset may not be accurate in predicting future concentration values. In this thesis an alternative approach, Spectral Window Selection (SWS), is proposed.

The algorithm is based on that described in Hinchliffe *et al.* (2003) and it offers the opportunity for constructing a calibration model from windows of wavelengths, i.e. from individual wavelengths to the full set as well as limiting selection to multiple sub-sets (windows) of the full set.

Due to the nature of the algorithm, the SWS methodology will typically generate slightly different subsets of wavelengths each time a model is constructed. This feature of the methodology needs to be addressed if a robust calibration model is to be obtained. An individual model may be too specific to the model building data hence the combination of multiple models is proposed. Stacking is a methodology by which different models based on the same data set are combined to produce a final model. There are a number of stacking methods that have been reported in the literature, especially in the area of combining neural network outputs (Sharkey and Sharkey, 1997; Zhang *et al.* 1997; Zhang *et al.*, 1999).

Since significant changes in broth composition occur over the period of a batch, with spectral interactions between constituents varying as a result, a further approach investigated in the thesis to improve model robustness was local modelling. The aim was to investigate whether by sub-dividing the process and building a separate model for each region, model robustness is potentially improved. Process sub-divisions can be identified by classifying regions of process behaviour. Arnold *et al.* (2001) have previously proposed the application of local modelling and demonstrated its benefits through its application to NIR spectra from a fermentation process.

Even by adopting a variety of calibration modelling procedures, it is not always possible to generate an accurate calibration model. In such cases other sources of information can be exploited, such as process data or other forms of spectroscopic data. The integration of spectral data and process data or two forms of spectral data is termed 'data fusion'. A data fusion model building framework has been investigated previously in several application fields including fault detection and process monitoring using a multiblock model (Gurden et.al., 2002, Wong *et al.*, 2005); in calibration modelling (Pedersen, 1997, Workman, 1999) and by using an augmented matrix in multivariate curve resolution (Naves *et al.*. 2003).

The conjunction of spectral and process data or two forms of spectral data in a fermentation process is also considered in this thesis. The approach adopted is to obtain the residuals from models constructed following the application of SWS to the spectral data followed by the application of stacking algorithm. These residuals are then modelled using the process data or the second form of spectral data. It is hypothesised that the inclusion of on-line process measurements such as off-gas analysis and feed addition rates, or alternative spectroscopic measurements delivers more robust and accurate models through the removal of calibration model offsets.

In the following sections, the objectives, contributions and contents of each Chapter of the Thesis are described.

1.1 Contributions of the Thesis

The primary area of contribution of the Thesis is in the field of robust calibration model construction. More specifically the key contributions include:

(1) The development of a new wavelength selection strategy, termed spectral window selection (SWS). The methodology is based on constraining wavelength selection to a limited number of windows as opposed to allowing multiple individual wavelengths to be selected. It is observed that calibration model performance is enhanced as a result of preventing the model becoming too specific to the training information. The window centres and sizes are changed by a search algorithm, with the objective of obtaining windows that lead to the best model for the training data. The wavelengths that fall within the windows are then used to produce a calibration model. The methodology does not provide a unique set of wavelengths due to the random nature of the algorithm and thus multiple calibration models can be generated.

(2) Stacking is adopted to combine the predictions from the multiple models to provide improved accuracy on the validation data set. Two methods for combining the outputs of the calibration model are considered, averaging and partial least squares (PLS).

(3) A comparison of the traditional approach i.e. performing PLS on the whole spectra/whole batch with a new strategy based on a local model approach and the application of the SWS algorithm to select the wavelengths followed by stacking is undertaken.

(4) A comparison of the SWS wavelength selection method and genetic algorithms (GAs) for the development of a robust calibration model and interval PLS (iPLS) is investigated.

(5) A demonstration of the proposed methodology, [(1) and (2)], through its application to NIR and MIR spectral data generated from the routine operation of an industrial fed-batch fermentation antibiotic production process is considered. An investigation of the benefits of an in-situ NIR and a MIR probe as analytical tools for the monitoring of several fermentation parameters is explored using the proposed methodology. It is hypothesised that the results obtained constitute a meaningful basis for the extension of NIR and MIR monitoring to the industrial fermentation process.

(6) A comparison of the merits of non-invasive and invasive instrumentation for the development of calibration models for the analysis of the particular fed-batch fermentation antibiotic production process is investigated.

(7) The development of a new strategy for the combination of process and spectral data or two forms of spectral data as an additional step for the generation of a robust calibration model of the product concentration and other biochemical data. Although with spectral data alone robust performance is attained, the accuracy of prediction is lower than that required during certain periods of the batch. It is recognised that fermentation behaviour can change significantly over batch age materialising in calibration model offsets being observed. The application of a data fusion approach, with existing on-line process measurements or other spectral data being used to correct the model, enables the removal of the offset and delivers a robust and enhanced calibration model.

1.2 Outline of the Thesis

Chapter 1 has provided an introduction to the field where the current work is relevant and the aims, objectives and contributions of the Thesis. In Chapter 2, a literature review covering the area of statistical regression is presented that focuses specifically on Partial Least Squares (PLS), which is the traditional method for building a calibration model. Both linear and non-linear variants of PLS are introduced for the analysis of the full spectral wavelength region.

Chapter 3 is divided into two parts. Initially a number of different wavelength selection algorithms are discussed focusing specifically on genetic algorithms (GAs). Spectral window selection and stacking are then introduced prior to the application of the methodology to NIR spectra attained from a data set related to diesel fuels.

The aim of the work presented in Chapter 4 is to apply the proposed calibration methodology strategy that includes the SWS algorithm for wavelength selection followed by stacking to data from an antibiotic fermentation production process. The characteristics and measurement challenges that fermentation processes raise are described and the potential benefits of spectroscopy are explained briefly prior to introducing the process. More specifically, the calibration modelling algorithms are compared for a number of 'Design of Experiment' batches and batches manufactured under normal operating conditions. In particular traditional methods are compared with the proposed approach i.e. full-spectra PLS, genetic algorithm based wavelength selection in combination with stacking and interval PLS (iPLS) are contrasted with the spectral window selection (SWS) and stacking algorithm. Finally to take account of the changes in process behaviour throughout the duration of a batch, the development of local models is investigated.

Initially in Chapter 5, the general concept of data fusion is explained and a number of applications and techniques reported in the literature are described. A data fusion technique is then proposed which includes the conjunction of process and spectral data and two forms of spectral data in a sequential manner for the construction of the calibration model. A model relating the calibration residuals to the on-line process

measurements or the second form of spectral data was constructed using PLS. The model was then used to correct the spectral calibration prediction, thus resulting in improved determination of broth concentrations.

Finally, in Chapter 6, conclusions about the performance of the proposed strategy are drawn and suggestions for future work are made.

CHAPTER 2

TRADITIONAL METHODS

FOR SPECTRAL MODEL CONSTRUCTION

2.1 Introduction

From a process perspective, a general requirement is the availability of on-line measurements to enable the rapid detection of changes in operational behaviour and where possible, allow corrective action to be taken. In many processes, even where there are a number of on-line process measurements available, off-line assays are still used for the monitoring of final product quality. However this approach is not suitable for establishing an effective monitoring and control strategy. To address this limitation, spectral instrumentation can help realise on-line measurements of product quality, that to date have only been available via off-line assays, therefore providing a route to improved control.

A characteristic of spectral devices is that they measure the absorbance or transmittance values across a large number of wavelengths, hence the challenge is to extract from this series of measurements, information pertaining to product quality through the construction of a calibration model. The traditional method of building a calibration model is through regression analysis including multiple linear regression (MLR), partial least squares (PLS) and neural networks. The emphasis of the chapter is on the analysis of the full spectral wavelength region with the subsequent chapter focusing on approaches where only a selection of wavelengths are included in the calibration model.

2.2 Implementation Issues in Chemical Analysis

The economic value of analytical results is related to their quality, the confidence the analyst has in the final result and how representative the sample is with respect to the original system (Vandeginste, 1987). In some situations, the sample may have become contaminated, its composition may have changed due to inappropriate storage, or sample degradation. In addition to these issues, the frequency and delay in performing the analysis are important factors in process monitoring and control. For example if delays in the analysis of a sample occur in the analytical laboratory, the information obtained may be less informative in terms of making process corrections. Thus, it is

necessary to maximise the sampling frequency and minimise measurement delay to attain accurate and representative process information. One solution is through the use of on-line spectroscopic measurements. The primary drivers for the implementation of on-line process spectroscopy are: a) a reduction in costs, b) improvements in measurement and hence the potential to enhance the monitoring and control of the process, c) the enhancement of safety, and d) a reduction in environmental impact.

The implementation of process spectroscopy can be classified according to three criteria, (McLennan and Kowalski, 1995): off-line, at-line and on-line. Off-line analysis involves the manual removal of a sample and transporting it to the measurement instrument for analysis in a specialised central laboratory, by qualified technical staff. The second category, at-line analysis, also involves manual sampling but this time the measurement and analysis is carried out on a dedicated analyser by the process operators. Finally on-line analysis involves the implementation of an automated analyser system. The preferable approach is on-line implementation since the opportunities for effective process monitoring and control are realizable. The measurements are recorded rapidly, hence there is the potential for automatic feedback.

Associated with the implementation approach is the sampling strategy. No matter which measurement approach is adopted, there are real challenges with regard to attaining a representative sample. There are four sampling strategies that can be applied for process analysis:

- Grab: a sample from the process stream is manually extracted and placed in a container prior to physically transporting it to the analyser.
- Extractive: a sample from the process stream is extracted automatically and transferred directly to the analyser.
- In-situ or invasive: a probe is inserted into the vessel, and there is physical contact, with the sample when making the measurements.
- Non-invasive: non-contact measurements are recorded, i.e. the instrument is not in contact with the sample.

2.3 Molecular Spectroscopy

Molecular spectroscopy methods can be implemented on-line using either an invasive or non-invasive strategy as described in section 2.2. Molecular spectroscopy deals with the interactions of molecular species with electromagnetic radiation, i.e. it is the measurement and interpretation of electromagnetic radiation absorbed, scattered or emitted by a chemical species, (McLennan and Kowalski, 1995). These interactions involve the transition between specific energy states in molecules. Absorption occurs when the transition is from a lower energy state to a higher one, whilst emission is where there is a transition from a higher energy state to a lower one. The energy of a quantum of light, a photon, E is expressed by Planck's law.

More specifically, Planck's law states that the energy E is proportional to the frequency of the electromagnetic radiation, ν and $\nu = c/\lambda$ where c is the speed of light and λ is the wavelength, i.e.

$$E = h\nu \qquad 2.1$$

where h is Planck's constant. The units of energy are often defined in terms of wavenumber, n, and expressed in cm^{-1}, while the wavelength λ is the inverse of the wavenumber, n, and is expressed in nm :

$$E \propto 1/\lambda \propto n \qquad 2.2$$

In molecular spectrometry, the objective is to understand how the molecule interacts with different types of radiation and what happens to the bonds between the component atoms of each molecule. Molecules can rotate or vibrate. When the molecules vibrate through stretching and bending motions, this requires absorption of higher energy photons of infrared radiation. There are various functional groups in a

molecule that characterise the vibrational spectra, for example $C-H$, $C=O$ and $C \equiv N$.

The transition is from a rotational level associated with one vibrational level to a rotational level at another vibrational level of different energy and identifies the type of spectroscopy that is used, McLennan and Kowalski, (1995). The transition between vibrational levels from $v=0$ to $v=1$ characterises the fundamental vibration and identifies the applicability of mid-infrared spectra while the transition between $v=0$ to $v=2$ or $v=3$ are called 'overtones' and result in the applicability of near infrared spectra.

The infrared spectrum can be considered as a fingerprint since the vibrational spectrum of the molecule is considered to be a unique characteristic of the molecule. Thus the spectrum can be used for the identification of a sample through its comparison with previously recorded spectra. In the absence of a suitable reference database, it is possible to undertake the basic interpretation of the spectrum from first principles hence the characterisation and identification of an unknown sample is possible.

A practical introduction to molecular spectroscopy and the associated instrumentation is given by McLennan and Kowalski (1995) and Faust (2001), with McKeivy *et al.* (1996) publishing a review on those aspects of infrared spectroscopy that are relevant to chemical analysis. Hassell and Bowman (1998) and Coates (2000) described process analytical chemistry for spectroscopists.

There are four main types of instrumentation used to sample processes, (Andrews and Dallin, 2003), i.e. transmittance, transflectance, reflectance and attenuated total reflexion (ATR):

(a) Transmittance probes allow light to pass through a small gap filled with the sample. Light is first carried to one probe via a fibre optic, passing through a narrow gap and is collected by a second probe, (Figure 2.1a, where the optical path length is equal to L).

(b) In tranflectance probes, the light beam is reflected by a mirror after its first pass through the sample and is collected by a lens arrangement on the same side of the sample gap as the initial illuminating fibre. After this, the collected light is returned to the analyser via a fibre optic having traversed the sample gap twice. Tranflectance probes are similar to transmittance probes (Figure 2.1b, where the optical path length is equal to 2L).

(c) In reflectance probes, the amount of light returned by diffuse reflectance is significantly less than for transmission arrangements. The fibre illuminates and collects diffuse reflectance (Figure 2.1c).

(d) In the ATR probe, the light does not physically leave the probe, but interacts with the sample via an evanescent wave at the sample/ATR crystal interface (Figure 2.1d). The effective path length depends on the probe geometry, the relative refractive indices of the ATR element and the sample, the number of internal reflections and the wavelength of light being considered.

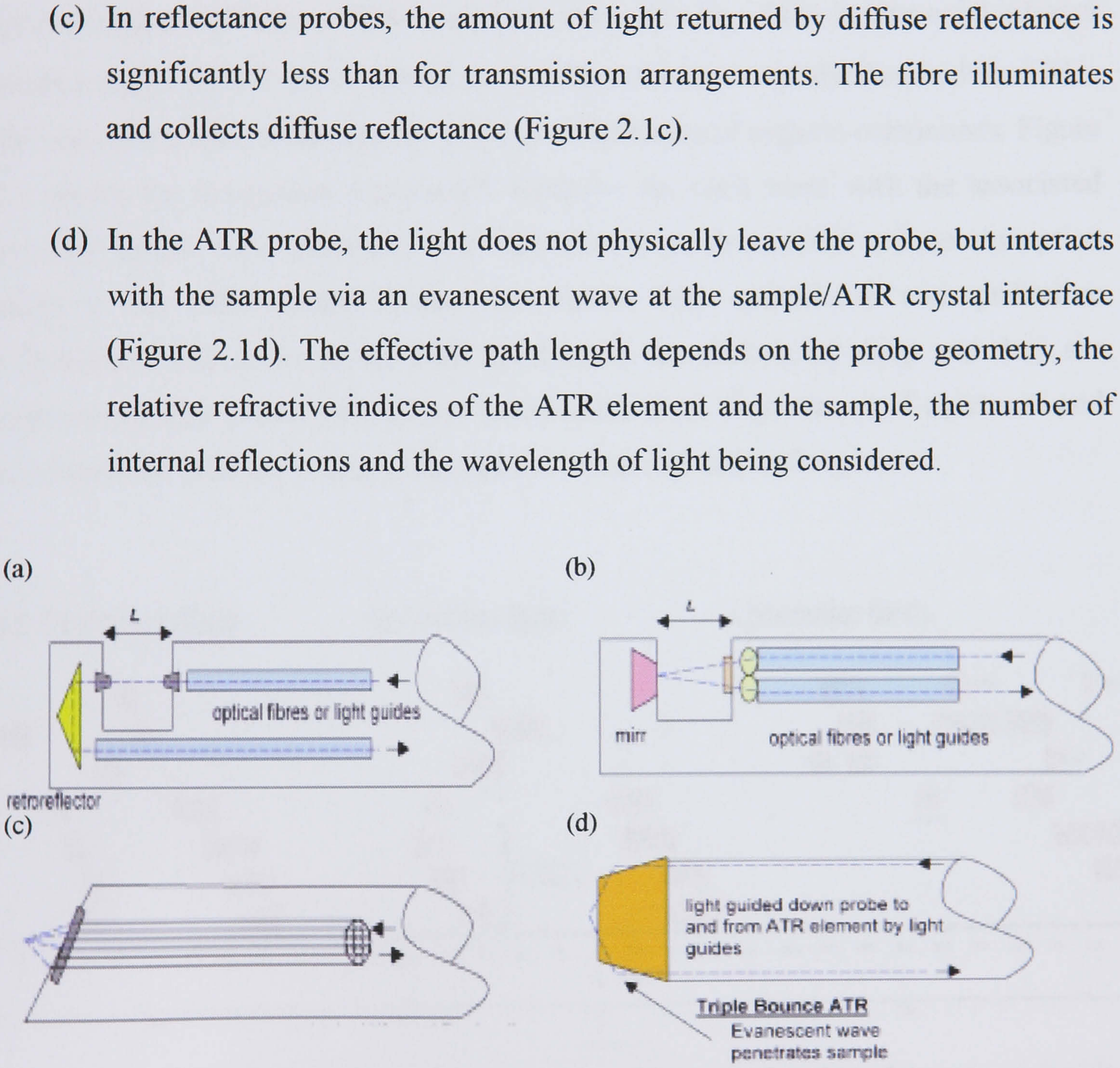


Figure 2.1. Probes for process spectroscopy (Andrews and Dallin, 2003)

The best approach i.e. transmittance, transflectance, reflectance or ATR depends on the application (Schneider and Kovar, 2003). Infrared spectroscopy has been used in a range of diverse applications from carbohydrate identification (Kacurakova and Wilson, 2001) to coral reef remote sensing (Hochberg *et al.*, 2003). Two infrared

spectroscopy methods are considered in this thesis: mid-infrared and near-infrared radiation that will be described in the following section and four different probes (three NIR and one MIR probe) that are described in greater detail in Chapter 4.

2.3.1 Near-infrared and Mid-infrared Spectroscopy

The NIR region of the electromagnetic spectrum extends from the end of the visible spectral region (780 nm or 12,800 cm^{-1}) to the beginning of the fundamental infrared spectral region (2,500 nm or 4,000 cm^{-1}). NIR radiation is mainly absorbed by -CH, -NH and -OH bonds, which are the primary constituents of organic compounds. Figure 2.2 shows the absorption wavelength windows for each bond with the associated overtone bands. Most chemical and biochemical species exhibit unique absorption bands in the NIR spectral region that deliver both quantitative and qualitative information. The peaks in the NIR spectra are not distinct or sharp and they are characteristically broad (50-100 nm bandwidth), since they consist of overtones and combinations from the primary absorptions in the mid-infrared regions.

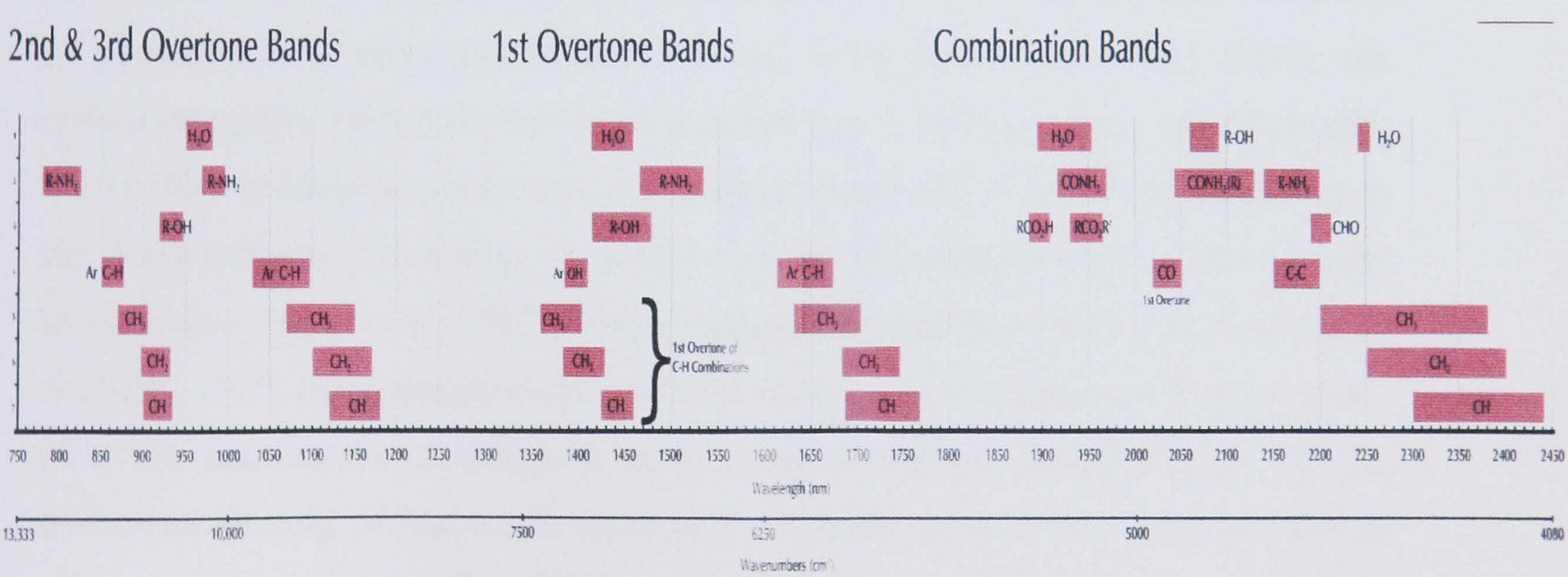


Figure 2.2. Absorption wavelength windows for each bond and overtone bands (source: Clairet Ltd website, http://www.clairet.co.uk/, 2005)

Mid-IR absorption at fundamental frequencies (4000-400 cm^{-1}) are typically 10-100 times stronger than NIR and capable of providing distinct spectral features of a compound from mixtures. Spectral bands in the MIR region are fundamental bands,

therefore the peaks are specific, sharp and sensitive. However, the strong water absorbance results in very short penetration depths for MIR radiation and for that reason very narrow transmission cells or attenuated total reflectance (ATR) elements are required. The measurement unit generally used for mid-infrared region is wavenumber while for near-infrared region wavelengths in nm are often used.

NIR applications were first reported in the 1960s in the agriculture industry, where constituents such as protein, oil and moisture were commonly measured. Since then research involving near-infrared (NIR) spectroscopy has been widely reported in the literature. Burns and Ciurczak, (1992) published a handbook on NIR while Workman, (1999) and Workman *et al.*, (1999) published a review, that included NIR spectroscopic measurements for the analysis of materials in distinct application areas including fine chemicals/chemical production, forensic science, petroleum / gas / fuel research and pharmaceutical production. Starting with only 3 publications in 1930s, there has been a significant increase in the number of publications and since 1980, the number of publications was over one thousand with the numbers continues to increase (Burns and Ciurczak, 1992). More recently, NIR papers have reported the application of neuroimage analysis (Strangman *et al.*, 2003); the monitoring of analytes in processes using biocatalysts (Bird *et al.*, 2002); the identification and the retification of pharmaceutical excipients (Candolfi *et al.*, 1999, Kramer and Ebel, 2000); the optical properties of human skin in biomedical optics analysis (Troy and Thennadil, 2001); the non-invasive prediction of blood glucose (Malin *et al.*, 1999); modelling in the food industry (including prediction of the concentration of a rice vinegar fermentation, Yano *et al.*, 1997; pattern recognition analysis of soy sauce, Iizuca and Aishima, 1999; the qualitative and quantitative analysis of green tea, Luypaert *et al.*, 2003; the multivariate classification of mayonnaise samples, Indahl *et al.*, 1999; the on-line monitoring of yoghurt fermentation, Cimander *et al.*, 2002; the prediction of oil content in instant noodles, Chen *et al.*, 2002; the modelling of oil concentrations from corn samples, Tan and Brown, 2003; and the differentiation of apple juice samples, Reid *et al.*, 2005); and in fuels (the modelling of gasoline samples Kalivas, 1997; the determination of distillation property of kerosene, Chung *et al.* 1999).

MIR has been used to monitor the fusion synthesis of resinates (Sundqvist *et al.*, 1999), to monitor batch polymerisation processes (Sprang *et al.*, 2003), to monitor

cell cultures (Rhiel *et al.*, 2002a; Rhiel *et al.*, 2002b), for clinical analysis as an alternative to the measurement of blood (Low-Ying *et al.*, 2002), for enzymatic hydrolysis monitoring (Ruckebusch *et al.*, 2001), for classification purposes in the food industry (Reid *et al.*, 2005), for the real-time monitoring of a biocatalysis (Dadd et el., 2000), for the multivariate prediction of several quality variables in fuels (Gomez-Carracedo *et al.*, 2003) and the in situ quantitative analysis of laboratory scale reactions (MacLaurin *et al.*, 1996).

2.3.2 Spectral Calibration Modelling

Spectral instruments, such as infrared devices, have the potential to measure on-line the concentration of the analyte of interest. However, in addition to the implementation issues discussed in section 2.2, the interpretation of spectral information is not straightforward as a result of the large number of variables (wavelengths) and the presence of components that exhibit overlapping features. The successful application of spectroscopic instrumentation therefore requires the application of multivariate data analysis techniques to extract the latent features from the data. According to DiFoggio (2000), the calibration model is key to chemometrics as it '*unlocks the power of the technique*' since it provides the ability to relate wavelengths to key components.

Underpinning the calibration model is Beer's law, i.e. absorbance is related to absorptivity and concentration by the following linear relationship:

$$A = \varepsilon(\lambda) \cdot b \cdot con \qquad 2.3$$

where $\varepsilon(\lambda)$ is the molar absorptivity of the analyte for the specific wavelength λ, b is the path length and con is the concentration of the analyte. The objective of a calibration model is to predict con from a measure of A.

Absorbance A is the most familiar measurement in spectroscopy and is calculated based on the 'transmitted' intensity. To calculate the absorbance, a reference spectrum is first obtained by measuring the background signal and then subsequent sample spectrum are corrected using this reference spectrum:

$$A = \log[1/T] \qquad 2.4$$

where, the transmittance, T, is calculated as the ratio of the reference intensity, I_R, to the sample intensity, I_S:

$$T = \frac{I_S}{I_R} \qquad 2.5$$

$$\text{i.e. } A = \log[\frac{I_R}{I_S}] \qquad 2.6$$

In the next section, the methodology involved in the creation of a spectral calibration model will be described. More specifically the various pre-processing, model formation and validation methodologies will be explained.

2.4 Common Framework for Calibration Modelling

Calibration modelling involves a number of important steps: a) problem definition and process understanding (i.e. what is to be modelled and what are the outputs), b) data collection for calibration model development, (c) data pre-screening and pre-processing and d) model derivation and validation. These steps form a common framework for calibration modelling development strategies.

For model identification, it is important to obtain data using the most appropriate sampling frequency, for example low sampling rates may result in information being lost and whilst sampling rates may incur heavy acquisition and handling costs. After data collection, the next step is to undertake a preliminary analysis of the data, to obtain a general appreciation of the underlying structure of the data. Pre-screening involves the visualisation of the raw data to help evaluate its quality. Any spurious

observations or unusual spectra are identified and treated appropriately. Following pre-screening, pre-processing is then undertaken. A number of pre-processing techniques are described in the following section.

2.4.1 Spectral Pre-processing

The goal of pre-processing spectra is to ensure the robustness of the subsequent calibration model. Robustness according to Zeaiter *et al.*, (2004) '*is the stability of the predictive capacity of the calibration model against perturbations centered on standard conditions*'. Pre-processing methods remove the systematic variation in the experimental data unrelated to the analyte concentrations. This variation could be a base-line drift or multiplicative scatter effects. Pre-processing involves a series of treatments that are mainly mathematical manipulations used prior to model construction. It is important that the data pre-treatment methods do not remove significant information from the spectra relating to the quality variable to be predicted. The different pre-processing techniques has been evaluated by a number of researchers. (Candolfi *et al.*, 1999 ; Azzouz *et al.*, 2003 ; Pizarro *et al.*, 2004 ; Zeaiter *et al.*, 2005). All the researchers agree that it is important that a pre-processing method is applied to the spectra but their conclusions regarding which is the most appropriate method to apply, differ. A number of key methods are described in the next section.

2.4.1.1 Various Pre-processing Techniques

The techniques can be divided into three groups: smoothing and differentiation, spectral normalisation and dimensionality reduction methods.

Within the grouping of smoothing and differentiation, first and second derivatives are the most commonly applied. These methods reduce peak overlap and eliminate baseline drift. The first derivative allows the additive constant background effects to be removed whilst the second derivative removes the baseline linear slope variations and additive effects. By adopting a smoothing approach, the noise is reduced. Noise

can be a result of random changes in amplitude from point to point within a signal. In this method, a moving average replaces each spectral point, i.e. the average is based on the adjacent points that are defined by the width of the smoothing window. The most widely adopted approach is based on the Savintzy Golay algorithm that will be explained in more detail in section 2.4.1.2.

Spectral normalisation methods give the same weight to all absorbances. These methods include the standard normal variate (SNV) transformation, which is a method that reduces the multiplicative effects created from light scattering. SNV (Barnes *et al.*, 1989) is evaluated independently for each individual spectrum and at each wavelength using the equation:

$$SNV_i = \frac{r_i - \bar{r}}{\sqrt{\sum (r_i - \bar{r})^2 / (l-1)}} \tag{2.7}$$

where SNV_i is the standard normal variate of a spectrum at wavelength i, r_i is the absorbance at the same wavelength and $\bar{r}$ is the average of all the absorbance values across all the wavelengths and finally l is the number of wavelengths.

A limitation of this method is that the multiplicative effects are assumed to be uniform over the whole spectral range, which is not always the case, thus artefacts can be introduced by implementing this transformation. A variation of SNV that addresses this problem is the robust normal variate (RNV) transformation, which was introduced by Guo *et al.*, (1999). In this case a percentile is used as opposed to the mean. However a disadvantage of this method is the need to optimise the percentile level.

A further pre-processing method falling within the spectral normalisation category is the de-trend (DT) approach. DT is normally used on reflectance spectra generated from powdered and densely packed samples. Details of this method can be found in Barnes *et al.*, (1989). The aim of DT is to eliminate the curvilinear behaviour and accounts for the variation in baseline shift. A second-degree polynomial is used to

standardise the curvilinear variation with the resulting correcting spectrum being given by:

$$x_{new} = x - (a_0 + a_1\lambda + a_2\lambda^2) \tag{2.8}$$

where a_i is the polynomial coefficient and λ is the wavelength vector of the spectrum.

A third method belong to the spectral normalisation class is multiplicative scatter correction (MSC). MSC (Geladi *et al.*, 1985) gives an estimate of the relationship of the scatter of each sample with respect to the scatter of a reference spectrum as discussed below. As a result, the same level of scatter for all spectra is obtained. In this method, the sample spectrum x is considered as a sum of two spectra, one due to light diffusion, d, and the other due to the chemical absorbances c:

$$x = d + c \tag{2.9}$$

The diffusion spectrum is modelled by least squares using a reference spectrum x_{ref}.

Thus

$$d = a + mx_{ref} + e \tag{2.10}$$

where a the offset, m the slope and e the residuals and finally the corrected spectrum x_{cor} is given by:

$$x_{cor=} \frac{(x-a)}{m} \tag{2.11}$$

Extended multiplicative scatter correction (EMSC) and spectral inference subtraction (SIS), (Zeaiter *et al.*, 2005) are alternative approaches to MSC and they require the use of prior knowledge about the pure components.

A further method, offset correction, is used to correct for a parallel baseline shift (Candolfi *et al.*, 1999). This is achieved by subtracting a chosen value from each spectrum independently. In some cases, the mean absorbance of a number of wavelengths of each spectrum can be used for correction.

The third family of methods, dimensionality reduction techniques, are orthogonal projection methods and their approach is to reduce the dimensionality of the prediction space spanned by the wavelengths to find the subspace that mainly compress variations related to the quality variable. The main methods in this category are variants of orthogonal signal correction (OSC). OSC is a method developed by Wold *et al.*, (1998). In this method, the bilinear components, which are orthogonal to the property of interest, are removed from the spectral data matrix. In this way the part that is uncorrelated to the property of interest is removed.

After the paper of Wold *et al.*, (1998), many different algorithms have been developed based on a similar concept. Westerhuis *et al.*, (2001) compared the original approach of Wold *et al.*, (1998) with four alternative approaches (Sjöblom *et al.*, 1998; Wise and Gallagher, 2005; Andersson and Bro, 2000; and Fearn, 2000). They concluded that none of these algorithm is able to reduce the error of prediction compared to the one given by PLS. As a result direct orthogonal signal correction (DOSC) was proposed by Westerhuis *et al.*, (2001). DOSC calculates directions in the spectral matrix that are orthogonal to the property of interest and account for the largest variation in the spectral matrix. These directions are attained by using least squares. Another variation of OSC, in terms of a different computation of OSC components, is projected orthogonal signal correction (POSC), (Trygg and Wold, 2002).

Azzouz *et al.*, (2003) also reported that in terms of prediction ability, the results of OSC used as a pre-processing method were not optimal. According to Zeaiter *et al.*, (2005), although the OSC methods improve the ability to interpret the model by reducing its complexity, the disadvantage of these methods remain the risk of over-fitting.

Orthogonal projection to latent squares (OPLS) is a method developed by Trygg and Wold, (2002). It is a modification of the original NIPALS algorithm and aims to

eliminate the structured variations from the space of the first PLS components that include the maximum of covariance between the spectral matrix and the property of interest. Zeaiter *et al.*, (2005) also reported the risk of over fitting for this particular method.

2.4.1.2 First and Second Derivatives

Savitzky and Golay, (1964) developed simple convolutional smoothing and differentiation routines that have found major application in spectroscopy. The appropriateness of these methods in multivariate calibration methods has been evaluated by several researchers with successful results. Forbes *et al.*, (1996) used second derivatives on NIR spectra to compensate for baseline offsets, Alciature *et al.*, (1998) applied the Savitzky-Golay (SG) procedure for the curve fitting of noisy infrared spectra, Ghasemi and Niazi, (2001) compared original, first and second derivative spectra for the simultaneous determination of cobalt and nickel and Rutledge *et al.*, (2001) proposed a new algorithm, PoliSh (smoothed partial least squares regression) that combines PLS with SG. The SG algorithm is a simple, popular, and straightforward methodology to implement to remove offsets and slopes in the data.

The SG algorithm is explained fully by Gorry, (1990) and Mark and Workman, (2003):

Step 1: A spectrum containing p evenly spaced data points y_i is considered. The p data points will be smoothed or differentiated (to order s) with a $2m+1$ point filter and a polynomial of order n.

Step 2: Each group of points is converted to a temporary coordinate system in which the coordinate values range from $i=-m$ to $i=m$, i.e. the midpoint is 0. This occurs because of the repetitive fitting of a polynomial of order n, to $2m+1$ consecutive points.

Step 3: The least-squares polynomial will then have the form:

$$f_n(i) = \sum_{k\,0}^{n} b_k i^k \tag{2.12}$$

Step 3: Least-squares is then applied to calculate the polynomial coefficients:

$$\frac{\partial}{\partial b_k}[\sum_{i=-m}^{m}(f_n(i)-y_i)^2]=0 \tag{2.13}$$

This leads to $n+1$ simultaneous equations in the unknown coefficients b_k

Step 4: The SG algorithm evaluates equation 2.12 at $i=0$, i.e. the central point.

As an example, Figure 2.3 shows the NIR spectra from a fermentation process where the SG convolution method is used for calculating the first derivative absorption spectra for different size windows.

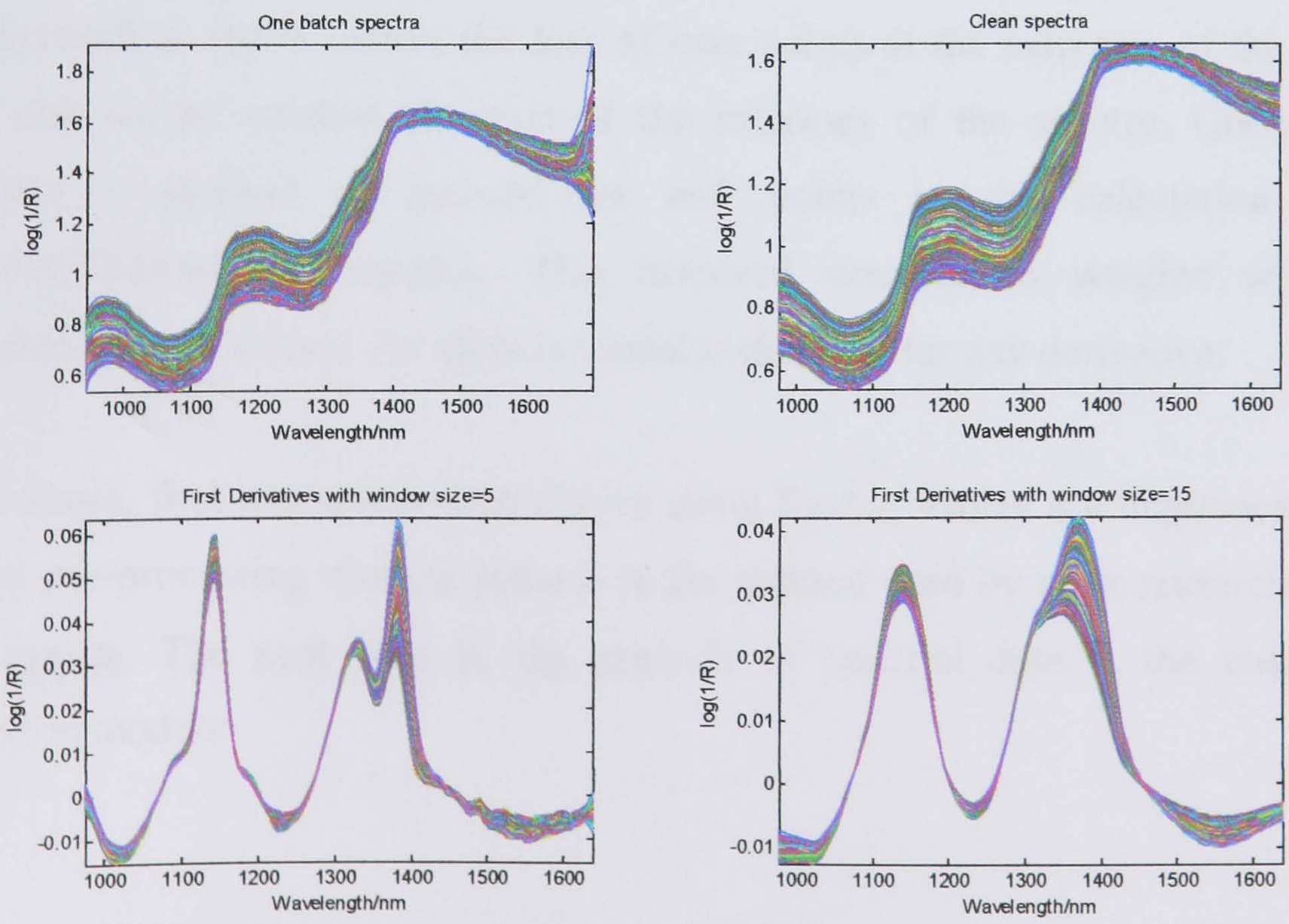


Figure 2.3. Spectra plots from the fermentation process described in Chapter 4.

The top left graph shows the raw NIR data from one batch. First the data related to the extremes of the instrument are removed since it does not provide realistic data and it is a non-informative noisy region (top right graph). Following this, the Savitzy-Golay algorithm is applied resulting in the first derivatives (bottom left graph) from a 5-point cubic smoother. The subsequent application of the Savitzy-Golay algorithm to the original data gives the first derivatives (bottom right graph) from a 15-point cubic smooth. In the second case, where the window size is bigger, there is distortion of the peaks. Thus, it can be concluded that the window size is an important parameter to ensure meaningful data are produced after pre-processing.

The Savitzky-Golay algorithm requires the specification of a number of parameters. With respect to the size of the smoothing window, it has to be large enough to reduce noise, but small enough to avoid distortion of the data. Windig (1994) studied the impact of window size and the use of second-derivatives of the spectra for curve resolution. In curve resolution, the spectra of a mixture is separated into its individual component spectra. He concluded that it is "*up to the researcher to find a compromise between the noise reduction and the loss of resolution and peak distortion when choosing the appropriate smoothing procedure*". Another problem associated with this approach is that it causes the loss of data points at the extremes of the spectral range due to the window function at the extremes of the spectra. Gorry (1990) described a method to include the end points in the calculation of the smoothed/differentiated spectra. This involved convolution weights which are calculated at all positions, for all polynomial orders and for any derivative.

In this thesis, first and second derivatives using Savitzy-Golay are implemented as a spectral pre-processing method since it is the method used by most researchers with good results. The next step in the analysis of spectral data is the building of calibration models.

2.4.2 Multivariate Linear Modelling

Multivariate linear modelling is a statistical methodology that allows the prediction of the quality variables from a number of predictor variables. It is based on least squares i.e. identifying the linear relationship that minimises the sum of vertical distances of the observations from the model. Linearity refers to the parameters of the regression model and not the variables since a polynomial model can also be viewed as a 'linear in the parameters model'. In the chemometric literature, regression modelling is often referred to as multivariate calibration (Vandeginste, 1987).

2.4.2.1 Description of Multivariate Calibration Modelling

The development of a calibration model involves taking spectral measurements, $\mathbf{X}$, and relating these to a reference, $\mathbf{Y}$, such as concentration through an underlying mathematical function. The model parameters are obtained by using experimental data. In Multiple Linear Regression (MLR), explained by Brereton, (2000), the dependent variable, $\mathbf{y}$, is a vector $(n \times 1)$ of n observations, $\mathbf{X}$ is a $(n \times p)$ matrix of p independent variables and the model is given by:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{2.14}$$

where $\boldsymbol{\beta}$ is a vector of coefficients of order $(p \times 1)$ and $\mathbf{e}\,(n \times 1)$ is a vector of residuals. The estimate $\hat{\boldsymbol{\beta}}$ of the regression coefficient vector $\boldsymbol{\beta}$, is the least squares estimate of the actual regression coefficients and is determined from:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \tag{2.15}$$

where the vector of fitted values, $\hat{\mathbf{y}}$, is given by:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \tag{2.16}$$

The difference between the actual values, $\mathbf{y}$, and the fitted values $\hat{\mathbf{y}}$, gives the vector of the residuals $\mathbf{e}$. If the resulting model is unbiased, then the residuals should satisfy the assumptions of normality and independence with their mean equal to zero and their variance equal to σ^2:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} \tag{2.17}$$

MLR can be extended to more than one dependent variable:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \tag{2.18}$$

where $\mathbf{Y}, \mathbf{B}, \mathbf{E}$ are matrices, with the number of columns equal to the number of the dependent variables.

When the application of MIR to spectral data is considered, the calibration model equation is given by:

$$C = \mathbf{S} \cdot P + \mathbf{E} \tag{2.19}$$

where

$\mathbf{C}$ is the concentration matrix of order $(m \times n)$ whose elements, c_{ij} are the concentrations of the j'th analyte for the i'th sample,

$\mathbf{S}$ is the $(m \times p)$ spectral response matrix

$\mathbf{P}$ is the $(p \times n)$ coefficient matrix and

$\mathbf{E}$ is the $(m \times n)$ matrix of errors or residuals.

The goal is to obtain an approximation of $\mathbf{P}$ that will enable the best determination of $\mathbf{C}$. The problem with applying the least squares algorithm to spectral data is that the high degree of correlation between the variables causes the models to be unstable, i.e. a small change in one value could result in the parameters being significantly affected. Also, as will be discussed in the next section, the number of samples will typically be less than the number of variables, hence this approach is not applicable since there is no unique solution for the coefficients. As a consequence, considerable research in

this area has been undertaken and reported in the literature (Jouan-Rimbaud *et al.*, 1995; Westerhuis *et al.*, 2000; Roggo *et al.*, 2003). The most common methods for building calibration models are based either on Partial Least Squares (PLS), (Geladi and Kowalski, 1986; Frank and Friedman, 1993) or neural networks (Blanco *et al.*, 2000; Zang and Friedrich, 2003).

2.4.2.2 Partial Least Squares

Partial Least Squares or Projection to Latent Structures (PLS) is one of the most frequently applied and accepted methods in analytical chemistry for the development of calibration models as it has been shown to be a reliable multivariate regression technique. The introduction of Partial Least Squares coincided with the development of the NIPALS (Non-Linear Iterative Partial Least Squares) algorithm (Wold, 1966) with the earliest application being in the field of econometrics (Geladi, 1992). The introduction of PLS to the field of chemometrics in the early 1980s (Wold *et al.*, 1983) resulted in a significant body of research both in terms of developing new PLS algorithms and also in investigating the mathematical properties of the method. Geladi and Kowalski, (1986) provided tutorials on the methodology and on the application of the NIPALS algorithm, Wold *et al.*, 1989 discussed non-linear PLS and Baffi, (1999) proposed an extension to the PLS approach. More recently, Wold *et al.*, (2001) described some recent developments in PLS modelling to handle non-linearities and the reduction of the number of variables included.

Undoubtedly, the importance of PLS in chemometrics is largely due to its ability to construct models with good prediction properties. There are a number of applications of PLS for the modelling of on-line processes. Wold, (1995) provided a definition of chemometrics, as '*how to get chemically relevant information out of measured chemical data, how to represent and display this information and how to get this information into data*' adding the undoubted importance of multivariate statistics and PLS in chemometrics. Miller, (1995) described the use of PLS in process analytical method development and operation and DiFoggio, (2000) provided guidelines for applying chemometrics to spectral data including PLS.

PLS outperforms traditional regression techniques in situations where there are more variables (inputs) than samples (observations). This is a common occurrence in multivariate data sets that result from spectroscopic applications where there is a large number of wavelengths that correspond to each sample. The fundamental idea of PLS is to reduce the dimensionality of the multivariate space by defining a set of latent variables that maximise the covariance between the input and output variables. In the next sections variants of the linear and non-linear PLS algorithms are described.

2.4.2.3 Linear PLS Modelling

Linear PLS is a multivariate regression method that projects the process variables and the quality variables, onto k latent variables, $\mathbf{t}_j$ and $\mathbf{u}_j$, (j=1,…,k) respectively. A linear regression model is then developed between the latent variables:

$$\mathbf{u}_j = \mathbf{b}_j \mathbf{t}_j + \mathbf{e}_j, \text{ where } j=1,\ldots,k \tag{2.20}$$

where the number of latent variables retained in the model is determined by cross validation (Barros and Rutledge, 2004).

PLS can also be formulated as an eigenvalue problem through the NIPALS algorithm. The algorithm is defined in Table 2.1. Consider two matrices $\mathbf{X}$ and $\mathbf{Y}$. The NIPALS based PLS algorithm performs the repeated regression of $\mathbf{X}$ on $\mathbf{u}$ to obtain an updated $\mathbf{w}$ vector (Step 2). Then $\mathbf{X}$ is regressed on the weighted vector $\mathbf{w}$ to obtain an updated score vector $\mathbf{t}$ (Step 4). The latent variables $\mathbf{u}_i$ and $\mathbf{t}_j$ are chosen in such a way that the correlation between them is maximised. Following this, $\mathbf{Y}$, is regressed on $\mathbf{t}$ to obtain an updated $\mathbf{q}$, and $\mathbf{Y}$ is regressed on $\mathbf{q}$ to provide an enhanced vector $\mathbf{u}$ (Step 7).

The matrices $\mathbf{X}$ and $\mathbf{Y}$ are decomposed as the sum of the outer products of the latent variables, $\mathbf{t}_j$, and the loadings, $\mathbf{p}_j$, and the prediction $\hat{\mathbf{u}}_j$ of $\mathbf{u}_j$ and the loadings $\mathbf{q}_j$, respectively:

$$\mathbf{X} = \sum_{j=1}^{k} \mathbf{t}_j \mathbf{p}_j^T + \mathbf{E} \tag{2.21}$$

$$\mathbf{Y} = \sum_{j=1}^{k} \hat{\mathbf{u}}_j \mathbf{q}_j^T + \mathbf{F} \tag{2.22}$$

where **E** and **F** are the residual matrices of **X** and **Y** respectively.

Table 2.1: NIPALS algorithm for PLS (Baffi *et al.*, 1999)

Step 1	Set the output scores **u** equal to any of the **Y** variables (any column)	
Step 2	Regress the columns of **X** on **u** to calculate the input weight coefficients **w**	$\mathbf{w}^T = \mathbf{u}^T\mathbf{X}/\mathbf{u}^T\mathbf{u}$
Step 3	Normalise the vector **w** to unit length	$\mathbf{w} = \frac{\mathbf{w}}{\|\mathbf{w}\|}$
Step 4	Calculate the input scores	$\mathbf{t} = \mathbf{X}\mathbf{w}/\mathbf{w}^T\mathbf{w}$
Step 5	Regress the columns of **Y** on **t**, to calculate the output loading coefficients **q**	$\mathbf{q}^T = \mathbf{t}^T\mathbf{Y}/\mathbf{t}^T\mathbf{t}$
Step 6	Normalise the vector **q** to unit length	$\mathbf{q} = \frac{\mathbf{q}}{\|\mathbf{q}\|}$
Step 7	Calculate the new output scores of **Y**, $\mathbf{u}_{new}$	$\mathbf{u}_{new} = \mathbf{Y}\mathbf{q}/\mathbf{q}^T\mathbf{q}$
Step 8	Check convergence on **u** : If $\mathbf{u}_{new}$ converges to **u** then go to step 9, else go to step 2 and replace **u** by $\mathbf{u}_{new}$	
Step 9	Calculate the input loadings **p** of **X**, by regressing the rows of **X**, on **t**	$\mathbf{p}^T = \mathbf{t}^T\mathbf{X}/\mathbf{t}^T\mathbf{t}$
Step 10	Regress the columns of **U** on **T** to find the regression coefficients $\mathbf{b}_i$ for the latent variables:	$\mathbf{b}_i = \mathbf{u}^T\mathbf{t}/\mathbf{t}^T\mathbf{t}$
Step 11	Calculate the input residual matrix **E**	$\mathbf{E} = \mathbf{X} - \mathbf{t}\mathbf{p}^T$
Step 12	Calculate the output residual matrix **F**	$\mathbf{F} = \mathbf{Y} - \mathbf{t}\mathbf{q}^T$
Step 13	If additional PLS dimensions are necessary replace **X** and **Y** by **E** and **F** and repeat steps 1 to 12	

2.4.3 Non-linear Modelling

If the relationship between the dependent and independent variables is non-linear then a linear assumption may prove to be inappropriate. In such cases it may be necessary to develop a model that is non-linear in the parameters. Non-linear modelling has been applied widely in the area of spectral calibration modelling. Several linear and non-linear modelling methods were described, compared and evaluated by Bertan *et al.* (1999) and Blanco *et al.* (1999, 2000). The focus of the papers was near-infrared spectroscopy. Methods compared were linear PLS, polynomial PLS and artificial neural networks (ANN). All the researchers agree that the most effective modelling of non-linear systems is provided by a non-linear calibration method but one should always be careful and use the simplest possible model since the high adaptability of non-linear methods can lead to overfitting. Bertran *et al.*, (1999) attained similar results for non-linear PLS and ANN while Blanco *et al.*, (2000) constructed the best model with ANN.

Two of the more commonly applied non-linear regression techniques are quadratic PLS and neural networks. These are described briefly in the following sections.

2.4.3.1 Non-linear PLS

A brief description of the background to non-linear PLS techniques is given in this section. The non-linear extension of PLS is captured by the following expression:

$$\mathbf{Y} = \sum_{j=1}^{k} f_j(\mathbf{X}_j \mathbf{w}_j)\mathbf{q}_j^T + \mathbf{E} \tag{2.23}$$

where $\mathbf{w}_j$ and $\mathbf{q}_j$ are the $\mathbf{X}$ and $\mathbf{Y}$ weight vectors, f_j is a non-linear function, $\mathbf{E}$ is the error matrix and k is the number of latent variables retained.

Table 2.2: Non-linear PLS algorithm (Baffi *et al.*, 1999)

Step 1	Set the output scores **u** equal to any of the **Y** variables (any column)	
Step 2	Regress the columns of **X** on u to calculate the input weight coefficients **w**	$\mathbf{w}^T = \mathbf{u}^T\mathbf{X}/\mathbf{u}^T\mathbf{u}$
Step 3	Normalise the vector **w** to unit length	$\mathbf{w} = \frac{\mathbf{w}}{\|\mathbf{w}\|}$
Step 4	Calculate the input scores	$\mathbf{t} = \mathbf{X}\mathbf{w}/\mathbf{w}^T\mathbf{w}$
Step 5	Fit the non-linear inner relationship	$\mathbf{u} = f(\mathbf{t}) + \mathbf{e} = \hat{\mathbf{u}} + \mathbf{e} \rightarrow \mathbf{c}$
Step 6	Calculate the non-linear prediction of **u**	$r = f(t, c)$
Step 7	Regress the columns of **Y** on **r**, to calculate the output loading coefficients **q**	$\mathbf{q}^T = r^T\mathbf{Y}/r^Tr$
Step 8	Normalise the vector **q** to unit length	$\mathbf{q} = \frac{\mathbf{q}}{\|\mathbf{q}\|}$
Step 9	Calculate the new output scores of **Y**, $\mathbf{u}_{new}$	$\mathbf{u}_{new} = \mathbf{Y}\mathbf{q}/\mathbf{q}^T\mathbf{q}$
Step 10	Update input weights **w**	
Step 11	Normalise the vector **w** to unit length	$\mathbf{w} = \frac{\mathbf{w}}{\|\mathbf{w}\|}$
Step 12	Calculate the new input scores	$\mathbf{t} = \mathbf{X}\mathbf{w}/\mathbf{w}^T\mathbf{w}$
Step 13	Check convergence on **t**: If yes then go to step 14, else go to step 5	
Step 14	Fit the non-linear inner relationship	$\mathbf{u} = f(\mathbf{t}) + \mathbf{e} = \hat{\mathbf{u}} + \mathbf{e} \rightarrow \mathbf{c}$
Step 15	Calculate the new non-linear prediction of **u**	$\mathbf{r} = f(\mathbf{t}, \mathbf{c})$
Step 16	Calculate the input loadings **p** of **X**, by regressing the rows of **X**, on **t**	$\mathbf{p}^T = \mathbf{t}^T\mathbf{X}/\mathbf{t}^T\mathbf{t}$
Step 17	Calculate the input residual matrix **E**	$\mathbf{E} = \mathbf{X} - \mathbf{t}\mathbf{p}^T$
Step 18	Calculate the output residual matrix **F**	$\mathbf{F} = \mathbf{Y} - r\mathbf{q}^T$
Step 19	If additional PLS dimensions are necessary replace **X** and **Y** by **E** and **F** and repeat steps 1 to 19	

The weight vector **w** needs to be defined and one common approach is through a recursive procedure. One of the first algorithms was proposed by Wold *et al.* (1989). Table 2.2 summarises the algorithm. They proposed a non-linear PLS algorithm where **w** was updated using a form of steepest descent optimisation by means of Newton-Raphson linearisation of the inner relationship. The inner relationship can be defined as a quadratic (second-order polynomial), cubic (third-order polynomial) or higher order polynomial. In polynomial PLS, two parameters need to be defined: the order of the polynomial and the number of latent variables to include in the PLS model. Typically, this involves the generation of different calibration models, varying the polynomial order and the number of latent variables to determine the best set of parameters. Quadratic PLS has been used successfully in several applications of near infrared spectrometry (Berntsson *et al.*, 2000; Bertan *et al.*, 1999; Blanco *et al.*, 1999; and Blanco *et al.*, 2000). Wold *et al.* (1989) recognised that this technique is fairly complicated and converges slowly and subsequently variants of this algorithm have been proposed (Baffi *et al.*, 1999).

2.4.3.2 Neural Network Modelling

Neural networks are another commonly applied non-linear modelling method. In the case of neural networks, the model is inherently non-linear and the parameters are obtained by the minimisation of an objective or error function through the application of an optimisation method. The first attempt to develop an artificial neural network was in 1943 by McCulloch, a neurophysiologist, and a mathematician, Walter Pitts, (from http://www.neurocomputing.org/History/history.html web site).

Artificial Neural Networks (ANN) generally consist of a number of layers: the *input layer*, the *output layer*, and the *hidden layers* that are between the input and output layers (Figure 2.4). The addition of hidden layers enables the network to approximate more complex non-linear behaviour. The fundamental unit or building block of the ANN is a neuron. In a neuron, a set of inputs (X_i) are weighted and a bias term, a threshold value is reached or exceeded for the neuron and then a signal is produced. A non-linear function (f_i) acts on the combined input signal (R_i) to form an output (O_i).

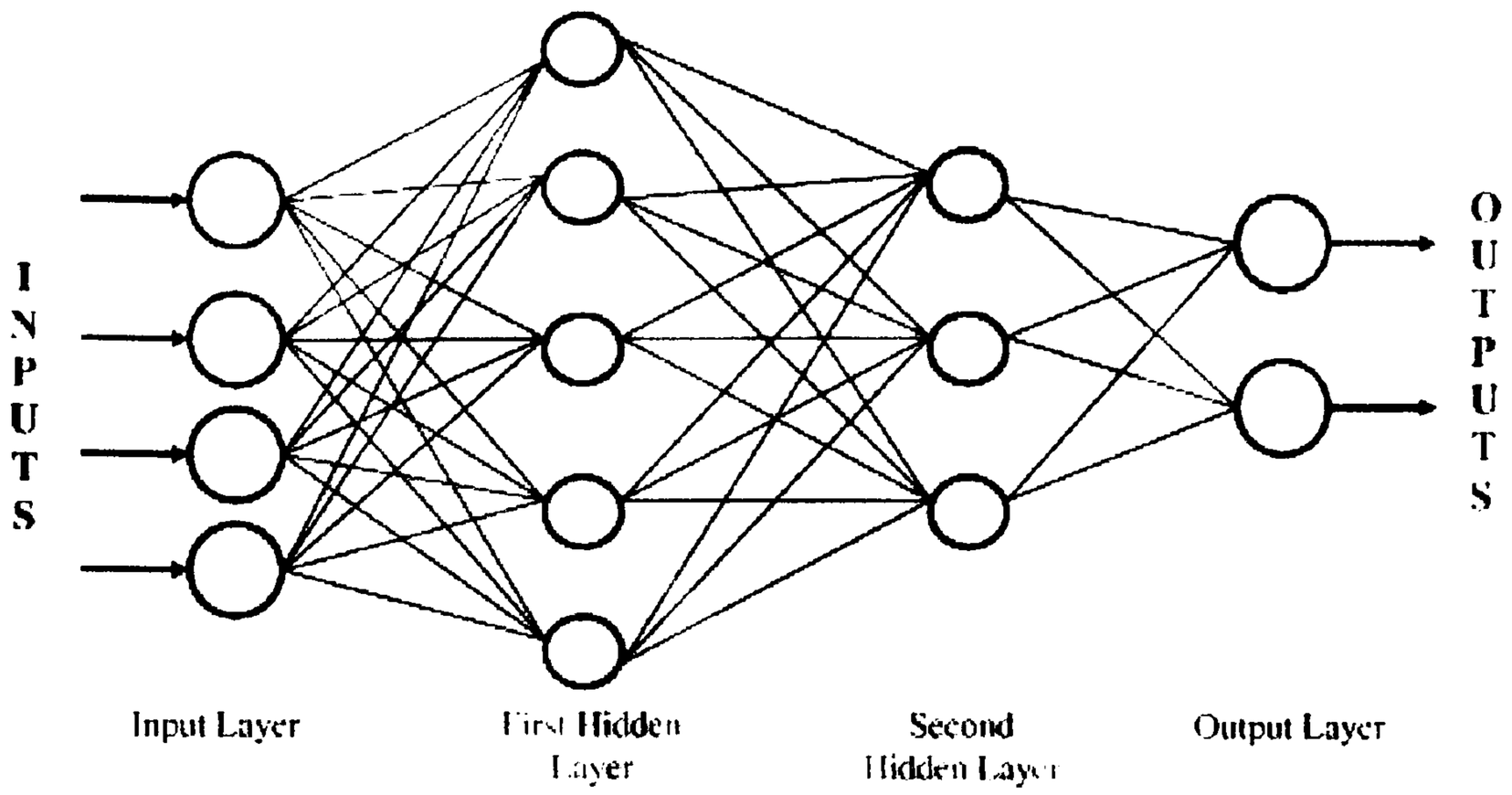


Figure 2.4. An artificial neural network, (Kadi, 2005).

The ANN structure comprising a number of interconnected neurons is illustrated in Figure 2.5 (Kadi, 2005). Each neuron implements a local computation or function. The output of each neuron is determined by its characteristics, its interconnection to other neurons, its external inputs, and its internal function.

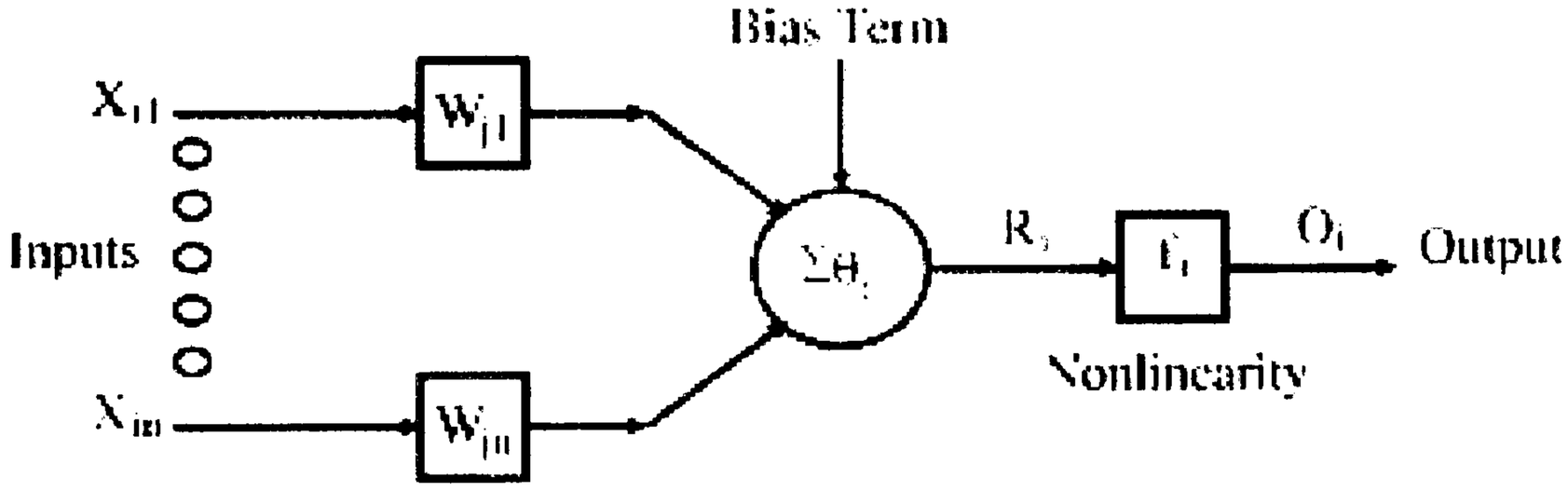


Figure 2.5. An artificial neuron, (Kadi, 2005).

Learning is a feature of neural network architectures and hence the selection of a learning algorithm is central to network development. Learning implies that a processing unit is capable of changing its input/output behaviour as a result of changes in the environment. Hence, the weights corresponding to the input vector need to be adjusted. Consequently a method is needed by which at the training stage, weights can be modified in response to the input/output pattern. A number of learning rules are available, i.e. supervised and unsupervised. Supervised learning is where the

network is provided with the target for the output during training, whilst unsupervised learning in which knowledge of the output is not utilised.

A number of studies of neural networks as applied, to chemical process monitoring and control, can be found in the literature, for example, for the estimation of polymer quality Zhang *et al.*, (1997); Zhang *et al.*, (1999); Zhang and Friedrich, (2003); Kadi, (2005) and Hussain, (1999). More specifically many papers have described the use of neural networks in the determination of calibration models for NIR spectroscopy (in fermentation analysis, Li *et al.*, 1999; in handling intrinsic non-linearity, Bertan *et al.*, 1999; in calibration Blanco *et al.*, 1999 and Blanco *et al.*, 2000 and for on-line monitoring Rantanen *et al.*, 2001 and Cimander *et al.*, 2002). A number of applications to MIR spectroscopy have also been reported (in industrial quality control, Andrade *et al.*, 1999; and for monitoring purposes Ruckebush *et al.*, 2001).

2.4.4 Model Validation

Spectral and corresponding chemical concentrations are typically divided into separate calibration and validation subsets. Only the calibration data are subjected to the modelling procedure, leaving the remaining data for use as an independent evaluation of the calibration model. This is crucial as it is important for the calibration model to be validated against unseen data.

For a model to be acceptable it must be sufficiently specific to describe the inherent relationship but also general enough to ignore chance relationships. The validation of a calibration model is extremely important since two problems may arise: Underfitting, i.e. an unsatisfactory relationship between the predicted and the actual values of the unseen data is attained and overfitting, i.e. a very good fit is obtained on the training set but poor predictions result on the validation set.

Different methods can be used to divide the data when considering the analysis of data from batch processes (Conlin *et al.*, 1998): spectra-wise and batch-wise split. Spectra-wise splitting involves the random allocation of the individual spectra to the calibration and validation sets. In contrast batch-wise splitting is based on the

assignment of all the spectra from a particular batch run to either the calibration or validation data set. The batch-wise approach is preferable when dynamic changes are to be observed and is therefore adopted in this thesis.

An alternative approach is cross-validation or 'leave-one-out'. In the case of g batches, the strategy is applied to (g-1) batches with the g'th batch used for validation. The procedure is repeated g times on each of the different combinations of the (g-1) batches and the model that gives the smallest RMS is selected.

To validate a calibration model derived by the methods described above, a number of methods have been proposed including the calculation of specific statistical metrics and the graphical methods, which mostly include residual representations.

2.4.4.1 Statistical Metrics

A number of statistical metrics can be used for the validation of the final model. A commonly used method to assess the utility of the model is the Root Mean Square (RMS) error, which gives a measure of its performance for both the training (RMSET) and the validation data set (RMSEV):

$$RMS = \sqrt{[\frac{1}{n}\sum_{i=1}^{n}(x_i - x_{pi})^2]} \tag{2.24}$$

where x_i is the concentration of the analyte in sample i,

x_{pi} represents the estimated concentration of the analyte in the sample i

n is the total number of samples used in the prediction set.

Another useful parameter is the relative error of prediction that captures the predictive ability of each component:

$$REP(\%) = \frac{100}{\bar{x}} \sqrt{[\frac{1}{n} \sum_{i=1}^{n} (x_i - x_{pi})^2]} \tag{2.25}$$

where $\bar{x}$ is the mean of the concentration in the training set

Another measure for testing the "goodness of fit" of the model (i.e. how well the model fits the data) is through the multiple correlation coefficient, R^2, which is given by the following equation:

$$R^2 = \frac{\sum_{i=1}^{n} \left(x_{pi} - \bar{x}\right)^2}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2} \tag{2.26}$$

The numerator is the residual sum of squares (RSS) and the denominator is the total sum of squares. The equation represents the portion of total variation explained by the model. The optimum value of R^2 is unity. This is one of the most common indicators for assessing the goodness of fit of the model

A modification of R^2 can also be used, the adjusted R^2, R^2adj. This can be more informative since it takes into account the number of variables p used in the regression analysis. The equation for R^2adj is given by the following equation:

$$R^2adj = 1 - \left[\frac{\frac{\sum_{i=1}^{n} x_{pi} - \bar{x}}{n-p-1}}{\frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n-1}} \right] = 1 - \left(1 - R^2\right) \frac{n-1}{n-p} \tag{2.27}$$

Whilst all the methods have found widespread use, the conclusions reached are similar. Spectral data applications have traditionally used the RMS error and therefore this metric is quoted in this thesis for validation purposes.

2.4.4.2 Graphical methods

Besides the goodness of the model fit, the residuals e_i also need to be tested. If the residuals are normally distributed, then the errors are normally distributed. The assumption of normality can be checked using a normal probability plot. Whilst a plot of the residuals e_i against the fitted values, allows the assumption of constant variance to be validated. The predicted values can also be plotted against the real values to form a parity plot, Scheilla and Junqueira, (2005).

2.5 Discussion

NIR and MIR spectroscopic measurements have been discussed in this Chapter with a general overview of multivariate calibration being given. A number of tasks in the construction of calibration models have been described including (a) data collection and the different types of process measurements, (b) data pre-screening and pre-processing, (c) calibration model derivation, and (d) model validation. The algorithms for the most common linear and non-linear regression techniques have been described, i.e. linear and non-linear PLS and neural networks. In Chapter 3, a number of variable selection methods are described that involve the elimination of wavelengths that do not contribute to describing changes in the model output rather than utilising the full set of model input data.

CHAPTER 3

BUILDING ROBUST CALIBRATION MODELS THROUGH VARIABLE SELECTION METHODS

3.1 Introduction

Chapter 2 described a number of methods that can be used for the development of a calibration model when the full set of wavelengths is included. An alternative approach is to use a sub-set of wavelengths, but the issue becomes one of identifying which wavelengths to incorporate. The fundamental theme of wavelength selection is to eliminate those wavelengths that do not contribute to capturing the behaviour thereby not attaining a model that is robust. Wavelength selection involves consideration of the correlations between the wavelengths and their relationship with the output.

However a number of alternative methods exist to help select those wavelengths/variables that give a robust regression model. These fall into four categories: all subsets, forward selection, backward elimination and stepwise approach (Xu and Zhang, 2001).

3.2 Variable Selection

The common feature of all the approaches is that a number of models with different input variables are examined and model accuracy assessed through a cost function. Consider the case where p variables are available to be included in the model. In summary the methods are (Xu and Zhang, 2001):

1. All subsets:

The simplest method for variable selection is to examine all possible subsets of the variables, i.e. if there are p initial variables then this would result in p! possible subsets. If p is large, this approach is computationally expensive and in most practical cases not viable.

2. Forward selection:

Each individual variable is tested. The one that achieves the lowest value for the cost function will be retained. The variable is then combined with each of the remaining

(p-1) variables until the pair that minimises the cost function is identified. This pair is retained and then combined with each of the remaining (p-2) variables and the model giving the minimum value for the cost function is retained. The procedure is repeated until satisfactory model performance is attained or a fixed subset of the variables is retained.

3. Backward elimination

This is the opposite of forward selection. The process begins with all p variables included in the model and each one is excluded in turn. The one that has the minimum impact on the model, is eliminated. With the remaining (p-1) variables, each again is omitted in turn and once more the subset that gives the smallest value for the cost function is retained. The procedure is repeated until the desirable size of subset is attained or the value of the cost function increases significantly.

4. Stepwise methods

These are a combination of forward selection and backward elimination. The model starts out by including only one variable and after that more variables are added but at each stage a backward elimination criterion is also applied. This approach is adopted because a variable added previously can become less important and therefore removed as a result of subsequent additions.

With respect to spectra data set modelling, the wavelength/variable selection methods are described in section 3.4, but the fundamental concepts are similar to those described above.

3.3 Wavelength Selection in Spectral Data Analysis

In building a calibration model, variable selection has the potential to produce better predictions and more informative models. For example when developing a calibration model in a simple mixture, wavelength selection would ideally select those regions most closely associated with the analyte of interest. However in practice, absorbance ranges of different functional groups may overlap and a mixture of substances contained in complex mixtures may contribute to signals across the complete spectral

wavelength range. Thus when using the complete spectra for quantitative analysis, the prediction results can be affected by those wavelengths that do not provide predictive information about the analyte of interest. In a complex mixture, using knowledge of the fundamental chemistry to select wavelengths of the pure components of interest is not necessarily sufficient to decouple the contributions from multiple chemical constituents and will not result in a good calibration model. It may be necessary to select a number of wavelength regions to decouple components with similar chemical groupings. As a consequence wavelengths selected for the mixture may contain regions with poor and good correlation to the analyte of interest from a pure component perspective.

'Automatic' wavelength selection has been used to address this issue and eliminate those wavelengths that are uninformative. A range of techniques for the selection of wavelengths from which to build a calibration model have been cited and will be described in section 3.4. It is claimed in these citations that individual wavelength selection will give better predictions than when using the full set. Selection of the most appropriate subsets of wavelengths is central to calibration model construction.

Based on first principles understanding, it is conjectured that the information of interest lies in a relatively small number of spectral regions, thus in this thesis a wavelength selection approach, Spectral Window Selection (SWS), is proposed. It offers the opportunity for constructing a model from any combination of wavelengths, i.e. from individual wavelengths to the full set as well as limiting selection to multiple sub-sets (windows) of the full set. This approach is benchmarked against genetic algorithms, GAs (Goldberg, 1989). GAs are suited to variable selection as the wavelength selection problem can be viewed as an optimisation process.

3.4 Wavelength Selection Methods

One of the issues in calibration model construction is to identify which variables to include to attain a robust and parsimonious model. The classical method is to use the basic chemical knowledge about the spectroscopic properties of the sample (Yano *et al.*, 1997; Vaidyanathan, *et al.*, 2000). Recently however, mathematical strategies for

variable selection have been utilised. A number of approaches devoted to wavelength selection for the development of a calibration model based on multivariate techniques have been presented in the literature. The basic algorithms for wavelength selection can be grouped into three categories according to the search method they employ (Forina et al., 1999; Alexandridis *et al.*, 2005): dimension-wise selection, model-wise elimination and subset selection. The first two categories will be described briefly and the latter one will be described in detail. More specifically, two of the techniques that belong in the subset selection category, interval PLS and Genetic Algorithms, will be described in details and compared with the proposed Spectral Window Selection (SWS) algorithm on a data set relating to diesel fuels.

3.4.1 Dimension-wise Selection and Model-wise Elimination Algorithms

This family of dimension-wise selection algorithms is based on the inclusion of additional predictors according to some criterion, i.e. it is similar to forward selection. The predictor may be an individual wavelength or a latent variable as determined through PLS. Algorithms that belong to this category include: interactive variable selection PLS (IVS-PLS), Abrahamsson *et al.*, (2003); cyclic subspace regression (CSR), Bakken *et al.*, (1999), Kalivas, (1999); and successive projections algorithm (SPA), Araujo *et al.*, (2000).

The general procedure followed by model-wise elimination algorithms is similar to that of backward elimination, i.e. it commence with all the wavelengths included in the model and then the uninformative variables are pruned according to some cost function. Thus it is the inverse procedure to that of dimension-wise selection. Algorithms included in this category are: iterative stepwise elimination (ISE), Forina *et al.*, (1999); uninformative variable elimination in PLS modelling (UVE-PLS), Centner *et al.*, (1996); UVE-a, Centner *et al.*, (1996); and iterative predictor weighting PLS (IPW-PLS), Forina *et al.*, (1999).

These algorithms have a number of limitations associated with the forward selection and backward elimination procedure that they employ. Their limitations are described in the following section.

3.4.2 Limitations of Dimension-wise Selection and Model-wise Elimination

According to Alexandridis et al., (2005), the main limitation of dimension-wise selection is that the methods incorporate variables until a specific criterion has been minimised or a specific number of variables have been selected. Once a variable has been included in the model, it cannot be removed. A similar issue arises with model-wise elimination, in that once a variable is removed from the model, it cannot be reinstated. As a result the search space is not explored in detail and thus the methods cannot guarantee that the subset selected is 'optimal'.

3.4.3 Subset Selection Algorithms

The distinctive feature of this family of algorithms is that they create different subsets of variables and the subset performance is evaluated. Based on this assessment, new subsets are generated from the existing ones. Algorithms that belong to this category include: interval variable selection PLS (iPLS) and genetic algorithms (GAs), which are also the most widely applied methods and for this reason these two techniques will be described in detail. Other methods that fall within this category were developed by Jouan-Rimbaund et al., (1995); Brenchley et al., (1997); McShane et al., (1997) and Höskuldsson, (2001) but these methods are not as widely applied as iPLS and GAs.

3.4.3.1 Interval Variable Selection PLS

Interval PLS (iPLS) is a methodology that calculates local PLS models based on fixed sub-intervals of the full spectral region. In this way, an overall understanding of the variation in the models between the spectral data and the quality variable is attained. The approach adopted is based on the identification of important spectral regions and hence the potential to develop good spectral local PLS model built from a limited number of wavelengths. iPLS models are built from a equal number of wavelengths. The method was developed by Nørgaard et al., (2000). The iPLS toolbox can be found at the following internet address: http://www.models.kvl.dk/source/ipls/. The

application of iPLS is examined in section 3.7 and compared with the SWS and genetic algorithms wavelength approaches for constructing a calibration model.

Some variants of iPLS include moving window PLS (mwPLS) where iPLS models are calculated based on a moving window concept and can be found in the website mention before. This approach has also been used by other researchers, including Du et al., (2004). A second form is that of synergy interval PLS (siPLS). Here the data set is split into a number of intervals (variable-wise) and all possible PLS model combinations of two, three, four, etc intervals are calculated. The computation time of this method can be significant depending on the number of intervals formed and the selected number of intervals to combine. A more recent method is that of backward interval PLS (biPLS), (Leardi and Nørgaard, 2004). It is basically a combination of iPLS and the backward elimination methodology. Similar to the interval PLS model, the data set is split into a given number of intervals, but PLS models are now calculated with each interval being ommited. The first interval omitted is the one that gives the poorest performing model with respect to the RMS error. This procedure is continued until one interval remains or the algorithm can be stopped when the number of retained intervals gives a RMS error lower than a predefined threshold. Applications of this approach can be found in the food industry, e.g. for the on-line monitoring of commercial carrageenan powders (Dyrby et al., 2004) and the quantification of the crystallinity of lactose in whey permeate powder (Nørgaard et al., 2005). It was also used to provide structural and quantitative information in protein structure analysis (Navea et al., 2005).

3.4.3.2 Genetic Algorithms

Genetic algorithms (GAs) are the most commonly applied method for wavelength selection and are treated as a benchmark against which to compare the performance of new wavelength selection strategies, Chatterjee *et al.*, (1996). GAs are optimisation tools which are powerful with respect to undertaking a global search in high-dimensional situations as found with spectral data. GAs mimic biological evolution, i.e. the work on 'Darwinian' models of population biology introduced in 1857. By applying the basic rules of life evolution, they adopt the principle of survival of the

fittest and the spreading of their genomes through reproduction. Their fitness is assessed through an objective function that characterises the performance of an individual member of the population. GAs comprise the following steps:

1. First, a decision is required regarding the mathematical representation of the population (i.e. the wavelength range). The population members are encoded as a string of binary numbers and are called Chromosomes.

2. The initial population is defined. A number of possible candidate solutions is randomly generated, with the number in the population depending on the dimension of the search space.

3. The percentage of the population retained after each evolutionary stage is defined. In some cases, the chromosomes of a given generation are completely replaced by a population of child chromosomes (Goldberg, 1989). In other cases only a specified percentage of a generation is retained as part of the next generation (Broadhurst *et al.*, 1997).

4. The maximum number of generations is defined. The GA will be terminated either when the population has converged or when the population has been through the maximum number of generations (evolutionary stages).

5. Evaluation takes place using a function, which computes a criterion, the fitness, to determine the quality of candidate solutions.

6. Reproduction occurs, where the initial population is evolved to give new solutions. Crossover is typically used for reproduction in order to recombine the fractions of two candidates (the parent strings) to obtain two new solutions (the child strings).

7. Mutation is used to randomly introduce new variability to the population of solutions and provide a means of branching out to unexplored regions of the parameter space.

Holland (1975) pioneered the development of GAs with the method now being widely applied (Chatterjee *et al.*, 1996). The first application of GAs in the chemometric literature for wavelength selection was reported by Lucasius and Kateman (1991). Since then, a number of researchers have reported the use of GAs as a tool for wavelength selection in multivariate calibration (Lucasius *et al.*, 1994; Horchner and Kalivas, 1995; Leardi *et al.*, 2000 and 2002; Ghasemi *et al*, 2003; Abdollahi and Bagheri, 2004; Alexandridis *et al.*, 2005). In addition, Diver and Ireland (1997) developed a spectral decomposition algorithm based on GAs, that reconstructs spectra to be analysed from a combination of previously measured 'library' functions by generating a number of trial solutions and 'evolving' these to obtain an optimal solution. Orthogonal projection (Gourvernec *et al.*, 2004) uses GAs as a pre-processing step to reduce the number of wavelengths included in the solution. Again, GAs were used as a pre-processing step for optimal pattern recognition (Smith and Gemperline, 2000) to reduce the misclassification of errors of similar materials. An alternative pre-processing strategy is to use GAs for wavelength selection and then apply principal component regression to build predictive models (Barros and Rutledge, 1998). Whilst the basic GA procedures are well documented, each GA based wavelength selection method reported in the literature uses variants of the algorithm. One common feature to the use of GAs is the RMS error which is used as the basis of the fitness evaluation (Broadhurst *et al.*, 1997). GAs have been applied to Raman spectral data (McShane *et al.*, 1999, Estienne *et al.*, 2000), Near infrared spectra (Pasti, *et al.*, 1998, Depczynski *et al.*, 1999) and UV-VIS (Araujo *et al.*, 2000).

GAs are not a panacea for wavelength selection. Many significant drawbacks have been reported in the literature:

- They tend to be slow to converge as they are computationally intensive, Alexandridis *et al.*, (2005).
- They present a configuration challenge because of the interaction of the adjustable factors (e.g. initial population, number of generations) that influence their outcome, Abrahamsson *et al.* (2003).
- The random nature of the algorithm results in different solutions every time the algorithm is executed, Ferreira et al., (2005).

- Wavelengths with a spurious correlation to the prediction property may be selected and the chosen wavelength subset may therefore not be appropriate for predicting future samples, Jouan-Rimbaud *et al.*, (1996).

The main issue with the genetic algorithm strategy is the final bullet point in that the opportunity exists to select specific wavelengths from many regions. Thus GAs based wavelength selection results in a model being constructed that is too specific to the model building data (i.e. includes wavelengths with a chance correlation to the property of interest) and potentially the model does not predict future samples well. The experiences of Abrahamsson *et al.* (2003) confirm this fundamental problem. Ferreira et al., (2005) compared a GA algorithm selection approach with the PLS-bootstrap (Lazraq et al., 2003) and concluded that the GA based results are not as consistent as those of the PLS-bootstrap method. Jouan-Rimbaud *et al.*, (1996) also investigated the selection of wavelengths using GAs. A number of random variables were included to the spectral matrix. If the wavelength procedures perform well, the random variables should not be selected. Whilst random correlations may result in a limited number of solutions containing a few random variables, it was observed that in spite of the use of a validation procedure, the GA approach selected a large number of irrelevant variables.

3.4.4 Spectral Window Selection Algorithm

A drawback with the wavelength selection methods, is that the selected wavelengths are typically scattered throughout the spectrum. To overcome this limitation, a new Spectral Window Selection algorithm (SWS) is proposed that focuses the search on a limited number of regions of variable width. Windows of wavelengths that may vary in size from a single wavelength to the complete spectra are automatically selected. These windows are identified with the aim of providing robust and accurate predictions of both the training and, more importantly, new unseen data sets. By constraining the wavelength selection to a limited number of windows rather than allowing multiple individual wavelengths to be selected, it is hypothesised that calibration model performance is improved by preventing the model from becoming too specific to the training information. The Spectral Window Selection method extends the philosophy of the algorithm discussed in Hinchliffe *et al.* (2003).

3.4.4.1 Binning Method

Hinchliffe *et al.*, (2003) investigated the prediction of polymer resin and end-use properties for a dual reactor solution polyethylene process. A new approach referred to as a 'binning' technique was proposed. From fundamental process understanding, it was hypothesised that relative proportions of the polymer between particular molecular weights have an influence on the polymer resin and end-use properties. The trace of polymer molecular weight distribution (MWD) is the weight fraction of polymer produced for a given molecular weight w(MW) as a function of $\log_{10}$(MW). Thus, w(MW) represents the point on a continuous distribution, with the area under the curve given by (see Figure 3.1):

$$A_1(\log_{10} MW) = \int_{P_{11}}^{P_{12}} w(MW) d \log_{10} MW \tag{3.1}$$

The difference in area from one 'bin' to the next was used as an indicator of the variations in the resin and end-use properties as it was conjectured that the relative proportions of the polymer between certain molecular weights (P_{11} to P_{12}) influences the physical properties.

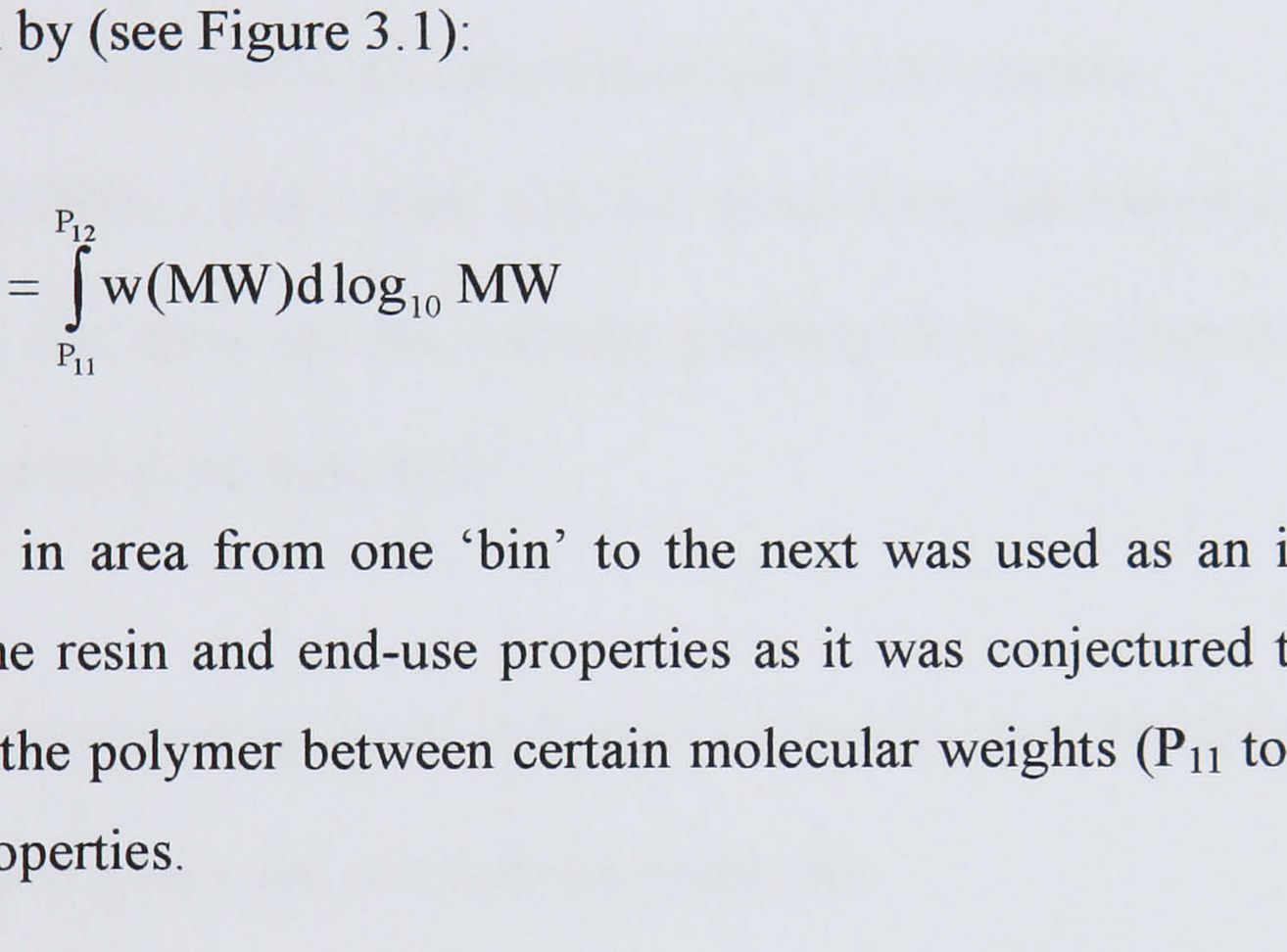


Figure 3.1. Bin model structure, where MW is the molecular weight, Hinchliffe *et al.* (2003).

The steps of the binning technique are summarised in Figure 3.2:

1. 'n' bin areas, A_1, ...,A_n are considered (where the value of 'n' is determined to obtain a minimum prediction error). Two parameters are considered, the bin centre and width enabling the calculation of the appropriate areas, A_1, ...,A_n.
2. A regression model is built between the bin areas and the end-use parameter (Q). The linear model is of the form:

$$Q = \alpha_1 A_1 + \alpha_2 A_2 + ... + \alpha_n A_n + \alpha_{n+1}$$

where the coefficients are determined using least squares.

3. The algorithm iterates with new bin areas being generated and new models relating the areas to the end-use property being constructed in order to minimise the prediction error.

The final model structure is shown in Figure 3.1. In this case two bin areas with a bias α_3, are selected to infer resin and end-use properties.

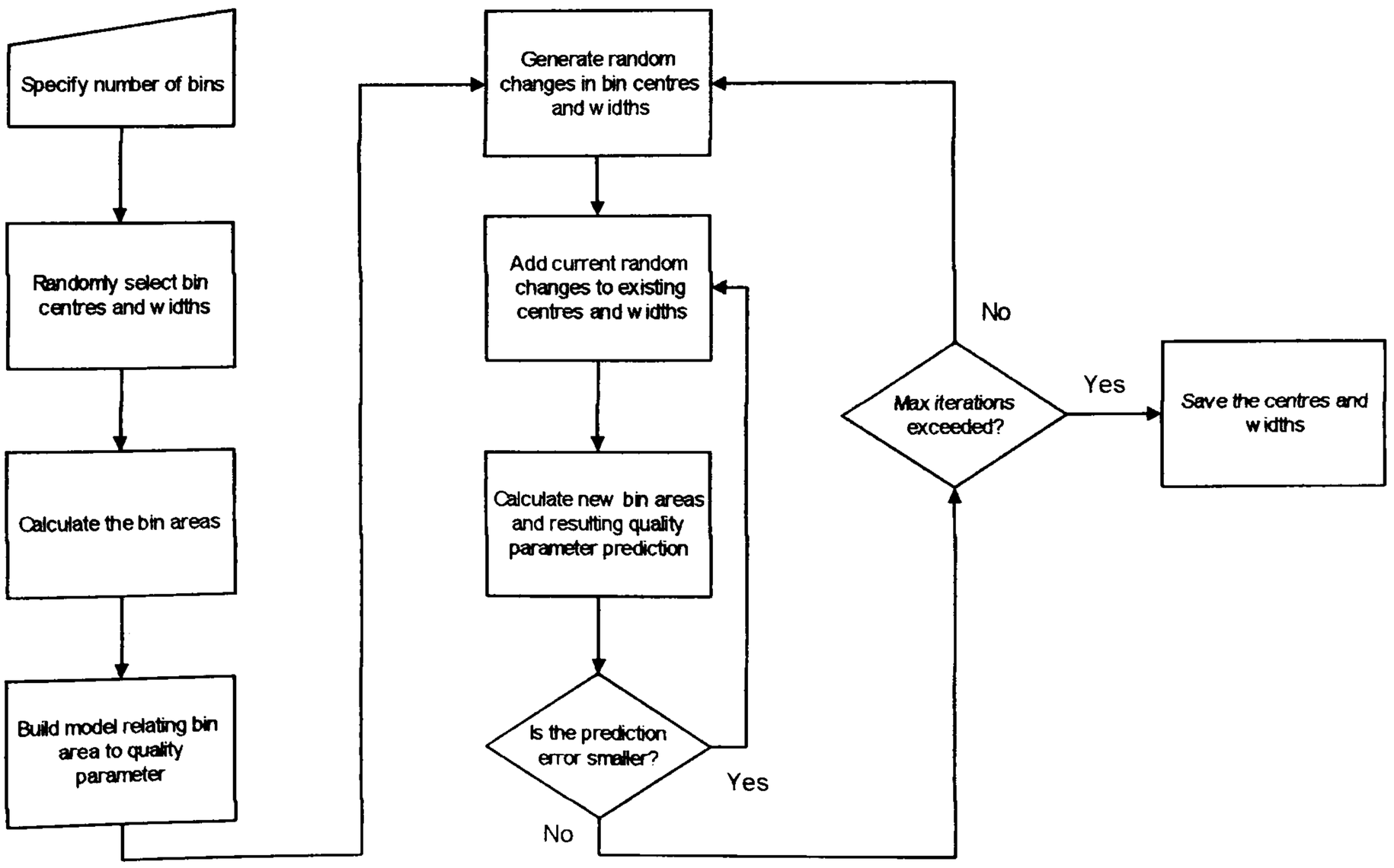


Figure 3.2. Linear bins program flowchart, Hinchliffe *et al.* (2003).

3.4.4.2 SWS Method

In the case of the SWS algorithm instead of bins, windows in the wavelength domain are selected .The implementation of SWS wavelength selection is as follows:

1. The number of windows that form the basis of the model for the inference of analyte concentration is defined. This involves a search commencing with a single region and then extending it to multiple regions, terminating when no additional predictive capability is found on the training data set. A limited number of windows is typically appropriate to capture the inherent chemical spectral information.
2. The initial centres and widths of the individual windows are randomly selected. The wavelengths in these regions are extracted, with common wavelengths in overlapping regions being removed. In Figure 3.3, two examples are presented for two windows i.e. in Figure 3.3a region 22 to 35 and 58 to 60 is selected whilst in Figure 3.3b region 43 to 62 is selected.

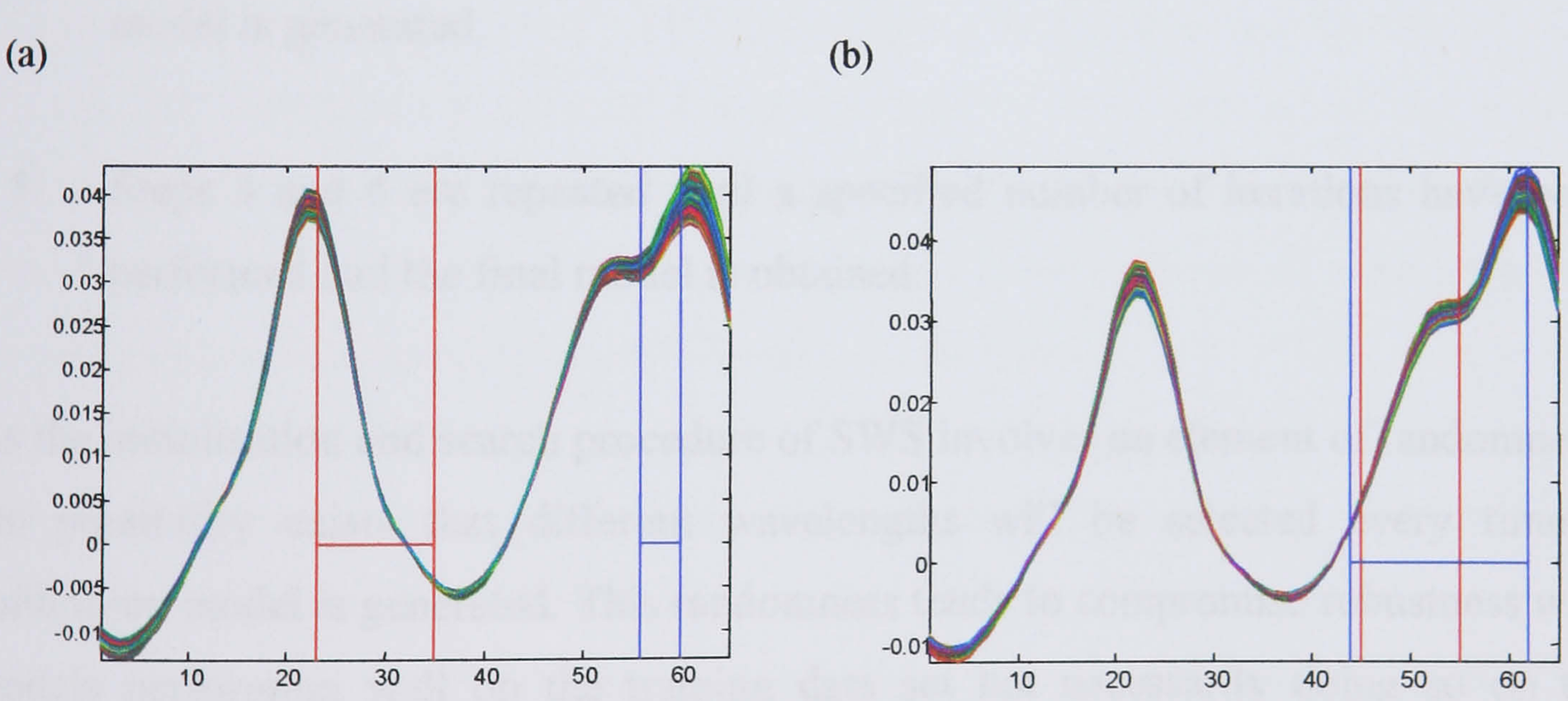


Figure 3.3. Example of the wavelength window selection for two distinct areas for the first derivatives of NIR spectra.

The resulting wavelengths, from the windows, are then used to build a PLS calibration model and the root mean square error (RMS) is calculated.

3. Increments are then generated from a uniform distribution, $U_n[0,1]$, and these are then applied to make random changes, to the centres and widths of each of the windows. The existing centres and widths are updated with the increments resulting in the windows centres shifting and changing in magnitude. Using the updated/revised windows of wavelengths, a new PLS calibration model is built and the RMS error is calculated.

4. The new RMS error is compared with the previous value. If the new RMS error is smaller, then the increments in window size and position (step 3) have materialised in an improvement in the calibration model. If this is the case then further improvements to the model may result through a further shift in the centres and widths by the same increment. If the RMS error increases, it can be concluded that the shift in the centres and widths of the windows was detrimental and the previous centres and widths are thus retained as the 'best'. New random increments are then generated and applied to the current 'best' centre and widths. The resulting subset of wavelengths is extracted and a new model is generated.

5. Steps 3 and 4 are repeated until a specified number of iterations have been performed and the final model is obtained.

As the initialisation and search procedure of SWS involves an element of randomness, the possibility exists that different wavelengths will be selected every time a calibration model is generated. This randomness tends to compromise robustness with models performing well on the training data set not necessarily doing so on the validation data sets. This issue can be addressed through the application of a stacking approach that is explained in the following paragraph.

3.5 Combining the models

As with any optimisation method, repeated runs of the proposed strategy may converge to different wavelength combinations. Recently, a number of new strategies have been proposed to improve the robustness of models in such cases. More specifically, the family of models based on the same data sets are combined to produce a final model. Several methods for model combination have been proposed including stacking.

3.5.1 Stacked Neural Networks

Stacking is a methodology whereby the predictions from multiple models are combined to provide improved accuracy. Stacking has been used in many applications, in particular in the area of neural networks (Sharkey and Sharkey, 1997; Zhang *et al.* 1997; Zhang *et al.*, 1999). According to Sharkey and Sharkey, (1997) through the combination of neural network models, one can avoid the loss of information that may result as a consequence of selecting the best performing networks. Moreover the idea of stacking is based on the exploitation, rather than the losing, of the information contained in imperfect estimators.

The idea of combining neural network models to improve modelling performance is not new and can be traced back to Nilsson, (1965). A review of neural net model combination can be found in Sharkey, (1996) and Sharkey and Sharkey, (1997). For the combination of neural network models, two approaches are considered: (a) ensemble-based approaches which are the most commonly used and where a set of neural networks are derived on the same data and then the outputs of the nets are combined, and (b) modular based approach, which is wider and where the exploitation of the specialist capabilities of different modules is taken into account.

For ensemble-based approaches, there are different methods for creating the ensemble members and different methods for combining the ensemble members, Sharkey,

(1996). A set of network models can vary in terms of their architecture (e.g. the number of hidden units), their weights and the time it takes them to converge to the solution, although they comprise the same solution. Thus, to create an ensemble member, a number of different effects can be varied: (a) the set of initial random weights, (b) the topology while keeping the training data set constant, (c) the algorithm used to train the networks, and (d) the data used for training. The later method is the one most frequently used and includes the 'bagging' and 'boosting' techniques.

'Resample and combine' or 'bagging' is an acronym for 'bootstrap aggregating'. The bootstrap approach is a popular method for resampling (Morgan, 1984). A bootstrap sample is created by drawing repeatedly with replacement, N samples from the training data set. As a result, it may contain replicates of some samples, and other samples may be omitted. The approach was originally proposed by Breiman (1993, 1996). Boosting is another successful method based again on the bootstrap sampling algorithm and is similar to bagging. It was proposed by Freund and Schapire (1996, 1997). In this case to estimate each single approximating function, a different set of weights is used for the case of the training data set.

When the set of neural net models has been created, there are a number of methods for the combination of ensemble members. These methods include averaging and weighted averaging (Perrone and Cooper, 1993), and stacked generalisation (Wolpert, 1992).

3.5.2 Weighted Stacking using Principal Component Regression

Based on the methods discussed in the previous section, Zhang *et al.* (1997, 1999) proposed a stacked neural network methodology. As each neural network representation can behave differently in different regions of the input space, representational accuracy over the entire input space can be improved by aggregating several neural network models instead of selecting a perceived 'best' single neural network for prediction purposes. The aggregated predictor is the one used for the final representation. This is particularly beneficial when the amount of training data is

limited, since the network tends to over-fit and hence exhibit significant generalisation errors.

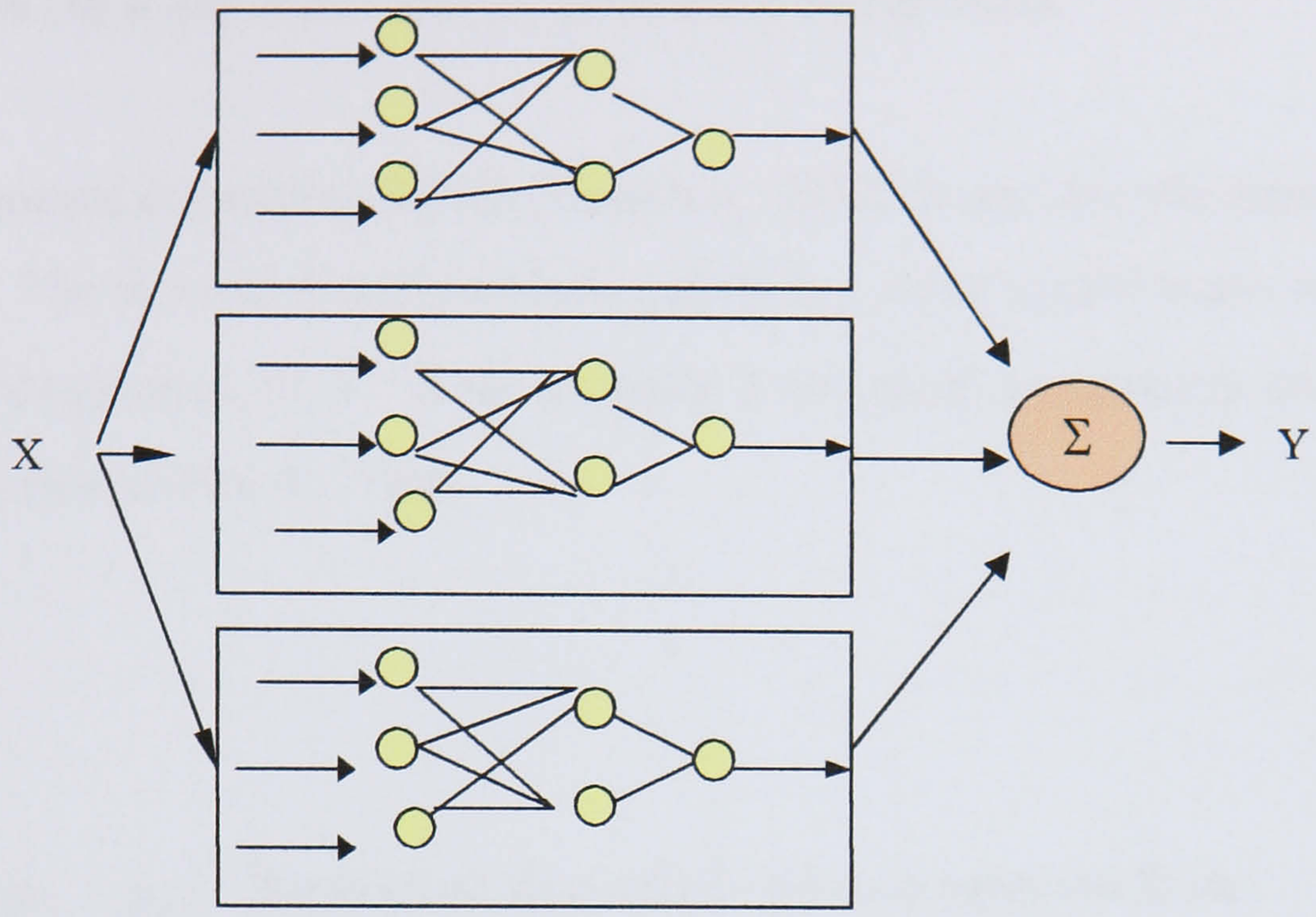


Figure 3.4. A stacked neural network where Σ is the weighted summation of all the individual neural networks.

In Zhang *et al.* , (1997, 1999), the stacking methodology is as follows (Figure 3.4): y is a vector of the expected model outputs and $\hat{\mathbf{y}}_i$ a vector of the predictions from the i'th neural network predictor. The predictions from a set of n predictors is described by the following notation:

$$\hat{\mathbf{Y}} = [\hat{\mathbf{y}}_1 \hat{\mathbf{y}}_2 \ldots \hat{\mathbf{y}}_n] \tag{3.2}$$

where each column corresponds to an individual predictor. The vector of predictions from the stacked neural network, $\hat{\mathbf{y}}_{stack}$, is represented by:

$$\hat{\mathbf{y}}_{stack} = \hat{\mathbf{Y}}\boldsymbol{\omega} = \boldsymbol{\omega}_1\hat{\mathbf{y}}_1 + \boldsymbol{\omega}_2\hat{\mathbf{y}}_2 + \ldots + \boldsymbol{\omega}_n\hat{\mathbf{y}}_n \tag{3.3}$$

where $\boldsymbol{\omega}$ is the stacking weight vector that needs to be identified. The matrix $\hat{\mathbf{Y}}$ can be decomposed into a sum of series of rank one matrices through principal component decomposition.

$$\hat{\mathbf{Y}} = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T + \ldots + \mathbf{t}_N\mathbf{p}_N^T \tag{3.4}$$

where t_i is the i'th score vector and p_i is the i'th loading vector.

Principal component regression (PCR), (Brereton, 2000), is used for the determination of the weights. The stacked neural network output is a linear combination of the first few principal components of $\hat{\mathbf{Y}}$. If for example k principal components are retained in PCR and are denoted by $\mathbf{T}_k$, then

$$\mathbf{T}_k = \hat{\mathbf{Y}}\mathbf{P}_k \tag{3.5}$$

where $\mathbf{P}_k = [\mathbf{p}_1\mathbf{p}_2 \ldots \mathbf{p}_k]$. The stacked network model is represented from:

$$\hat{\mathbf{y}}_{stack} = \mathbf{T}_k\theta = \hat{\mathbf{Y}}\mathbf{P}_k\theta \tag{3.6}$$

The least squares estimation of θ is given by:

$$\theta = (\mathbf{T}_k^T\mathbf{T}_k)^{-1}\mathbf{T}_k^T\mathbf{y} = (\mathbf{P}_k^T\hat{\mathbf{Y}}^T\hat{\mathbf{Y}}\mathbf{P}_k)^{-1}\mathbf{P}_k^T\hat{\mathbf{Y}}^T\mathbf{y} \tag{3.7}$$

and the stacking weight vector after the application of PCR is:

$$\boldsymbol{\omega} = \mathbf{P}_k\theta = \mathbf{P}_k(\mathbf{P}_k^T\hat{\mathbf{Y}}^T\hat{\mathbf{Y}}\mathbf{P}_k)^{-1}\mathbf{P}_k^T\hat{\mathbf{Y}}^T\mathbf{y} \tag{3.8}$$

By adopting this method, it was shown that although a single neural network may have poor generalisation capabilities, by combining networks, the performance of the model increases.

3.5.3 Average and Partial Least Squares Stacking

As the initialisation and search procedures of SWS and GAs involves stochastic effects, the possibility exists that different wavelengths will be selected every time a calibration model is generated. Building on previous studies, this thesis investigates whether the combination of models can improve the robustness of the calibration models.

It is postulated that a more robust prediction, $\mathbf{P}_{stack}$, will be realised through stacking than the solution obtained from a single model. The stacked prediction is generated through a weighted combination of the predictions, $\mathbf{P}_{win}$, from the n_{win} individual models:

$$\mathbf{P}_{stack} = \frac{\sum_{k=1}^{n} \mathbf{P}_{win_k} \mathbf{W}_{win_k}}{n_{win}} \tag{3.9}$$

with the weights, $\mathbf{W}_{win_k}$, being apportioned to a particular model k. Two methods were considered for the calculation of the weights. Firstly, the models were equally weighted, that is the mean of the predictions at each time point was calculated. A more sophisticated method is possible and in the second method, linear PLS was used to calculate a weighted average. In this case, the prediction of the individual models form the input matrix to the PLS model whilst the measured values of the analyte define the model outputs. The stacked model is constructed from the training data set. The PLS stacking method is based on the same principle as used in neural network stacking with PLS used instead of PCR. This variation was made to bias the weighting of those models that are more predictive of the output as opposed to only considering the variability in the inputs. The stacking model is constructed through the training data set. The model in which PLS is applied is given by the following equation:

$$\mathbf{P}_{stack} = \mathbf{B}_{win}\mathbf{P}_{win} + \mathbf{E}_{win} \tag{3.10}$$

where $\mathbf{P}_{stack}$ is the final weighted predictions

$\mathbf{P}_{win}$, is the matrix formed by the *n* individual models

$\mathbf{B}_{win}$ is the matrix of the coefficients and

$\mathbf{E}_{win}$ is the matrix of the errors associated with the model

A PLS model is then formulated using the methodology described in section 2.4.2.3 and equation (2.20).

3.6 Summary of Proposed Calibration Strategy

The steps of the proposed calibration strategy are summarised in Figure 3.5.

Prior to implementing the SWS algorithm, the following steps are performed:

1. The data is partitioned into two data sets, i.e. a training and a validation (unseen) data set.

2. The spectral and analyte data is pre-processed by taking first or second derivatives as discussed previously.

After the initial configurations, the SWS algorithm, as was described in section 3.4.4, is run. Finally, due to the random nature of the algorithm, different models can be obtained every time the algorithm is executed. To obtain a robust model, multiple models are combined as in Section 3.5.3. The number of models that are combined in the final stacked model is pre-specified. This procedure is repeated until the desired number of models required for stacking have been built and the final model is calculated using the stacking procedure discussed in the previous section.

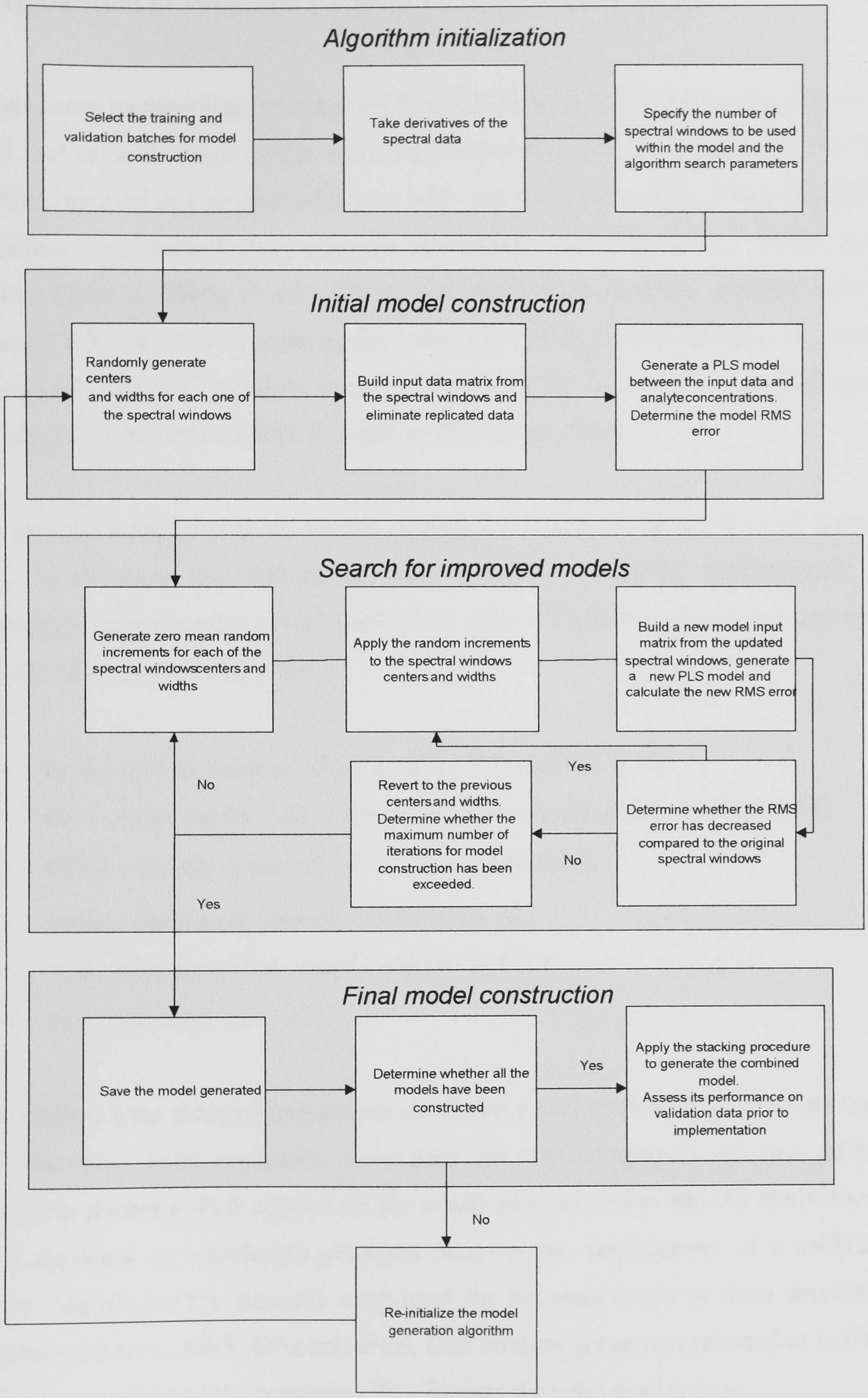


Figure 3.5. Flow diagram of the spectral window selection (SWS) algorithm in this study.

3.7 Application of Proposed Calibration Model Methodology

The proposed methodology was applied to a NIR data set attained from the analysis of diesel fuel samples (Eigenvector Research Corporation, 2005). Analysis of the data has been reported in a number of papers with respect to the testing of new modelling algorithms and spectral pre-processing techniques, including: Fuzzy Partial Least Squares (FPLS), (Bang *et al.*, 2003); artificial neural network analysis (ANN), (Boger, 2003); direct orthogonal signal correction (DOSC), (Westerhuis et. al., 2001); Orthogonal Signal Correction (OSC), (Svensson et al., 2002); and Orthogonal Wavelet Correction (OWAVEC), (Esteban-Diez et. al., 2004).

The data set comprised three groups of samples based on the analysis of summer fuels. In all cases, the outliers had been previously removed. Unfortunately the wavelength numbers were not available. The value of 6 different properties associated with the NIR spectra were given:

- bp50 - boiling point at 50% recovery, ${}^{0}C$ (ASTM D 86)
- CN - cetane number (like Octane number only for diesel, ASTM D 613)
- d4052 - density, g/mL, @ $15^{0}C$, (ASTM D 4052)
- freeze - freezing temperature of the fuel, ${}^{0}C$
- total - total aromatics, mass% (ASTM D 5186)
- visc - viscosity, cSt, $40^{0}C$

Two outputs were selected to demonstrate the proposed methodology: total aromatics and viscosity. Total aromatics were used for the comparison of two different modelling methods: PLS applied on the whole spectral region and the application of SWS approach to wavelength selection prior to the development of a calibration model. Models for the viscosity were used for the comparison of three wavelength selection methods: SWS, GAs and iPLS. One random group was selected to build the calibration model and the remaining data formed the validation data set.

The plot of the NIR first derivatives spectra for the validation data set, on Figure 3.6, gives an indication of the spectral variations.

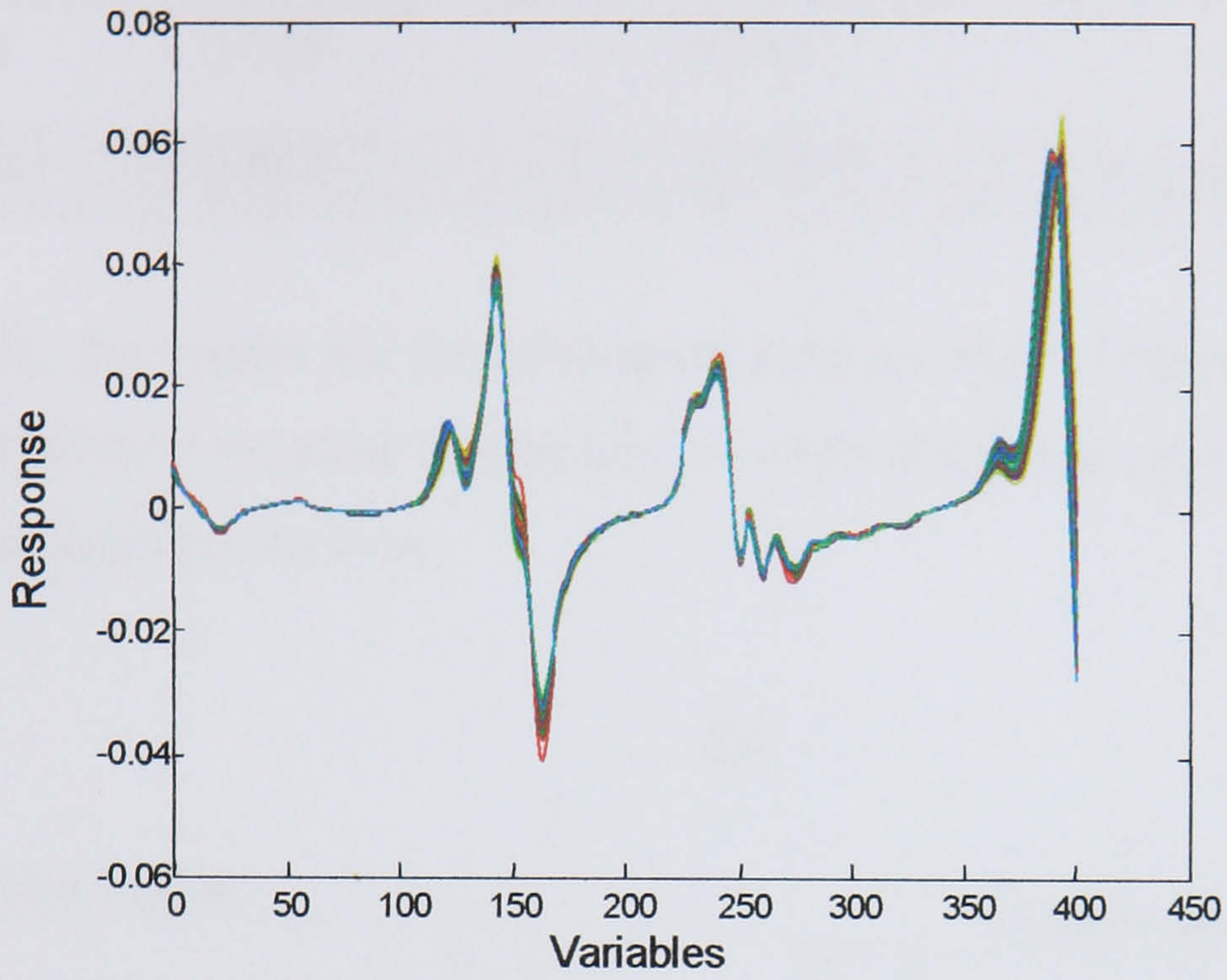


Figure 3.6. NIR validation spectra.

For the SWS based calibration algorithms, two spectral windows and thirty models were used to generate the final model. Eight latent variables was used in the PLS stacking step as indicated by cross validation.

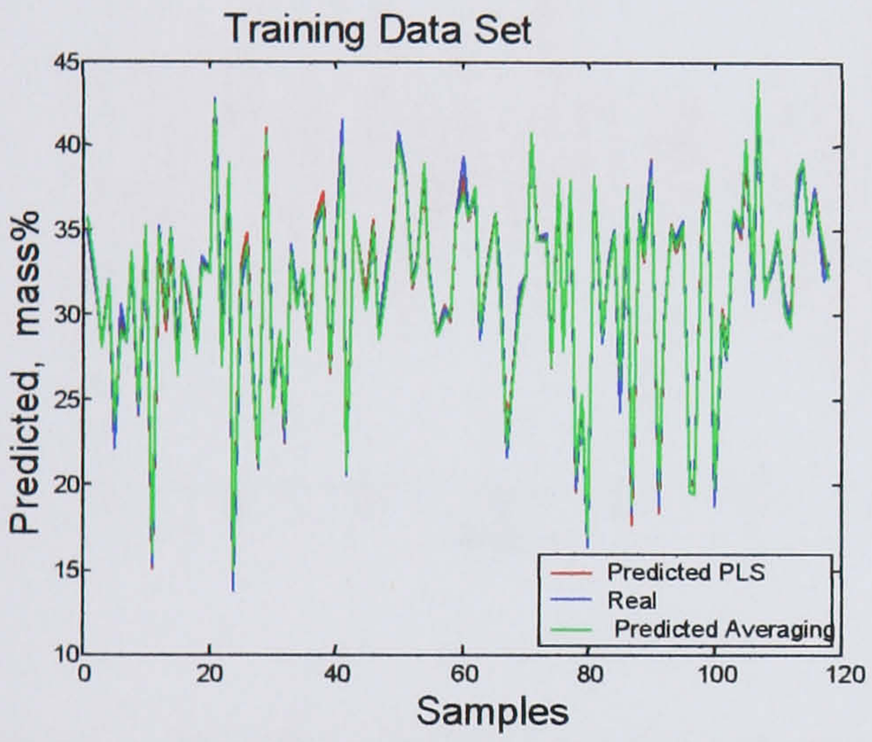
The results from the various approaches for the prediction of the total aromatics are summarised in Figures 3.7 and 3.8 and Table 3.1. Table 3.1 reports the performance of the models, summarising the RMS errors for modelling both without SWS, i.e. applying PLS to the complete spectra, and through the application of SWS and average and PLS based stacking. The SWS strategy with PLS stacking outperforms both average stacking with SWS and PLS modelling. This is a consequence of the fact that no poor models are generated thereby weighting the final model negatively. PLS stacking is a more sophisticated method to calculate a weighted average, where the predictions of the individual models form the input matrix, whilst the measured values of the analyte define the model output. Thus, there is a different weight for each model created by SWS method. On the other hand in average stacking the models are equally weighted, that is the mean value of the predictions at each time point was calculated and as a result even the poor models are weighted equally to the good ones.

Table 3.1. Results for the total aromatics

	Without SWS	**SWS with Average Stacking**	**SWS with PLS Stacking**
RMSET	0.703	0.744	0.538
RMSEV	0.829	0.843	0.557

In Figures 3.7a, the results for the calibration data set and in Figures 3.7b the results for the validation data set after the application of PLS stacking (red line) and average stacking (green line) can be seen.

(a)



(b)

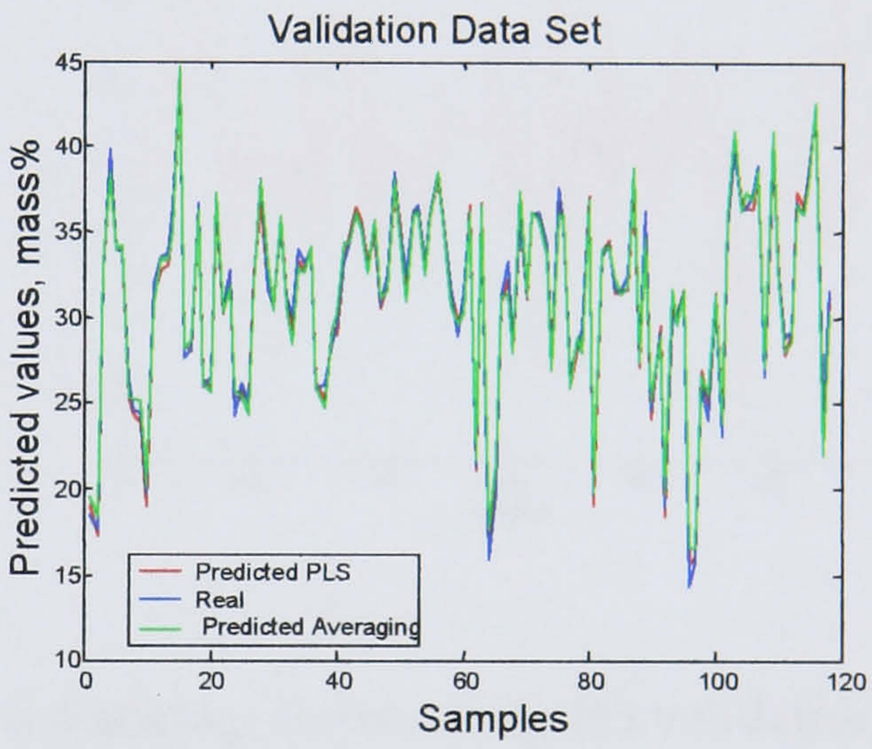


Figure 3.7. Results for the total aromatics: (a) training and (b) validation

The PLS stacking predictions can be seen more clearly in Figure 3.8. For the training data set, the fitted values versus the real values are given in Figure 3.8a and the validation data set predicted values versus the real values are given in Figure 3.8b. The predictions have a high accuracy as evident by the small residuals (Figures 3.8c and 3.8d).

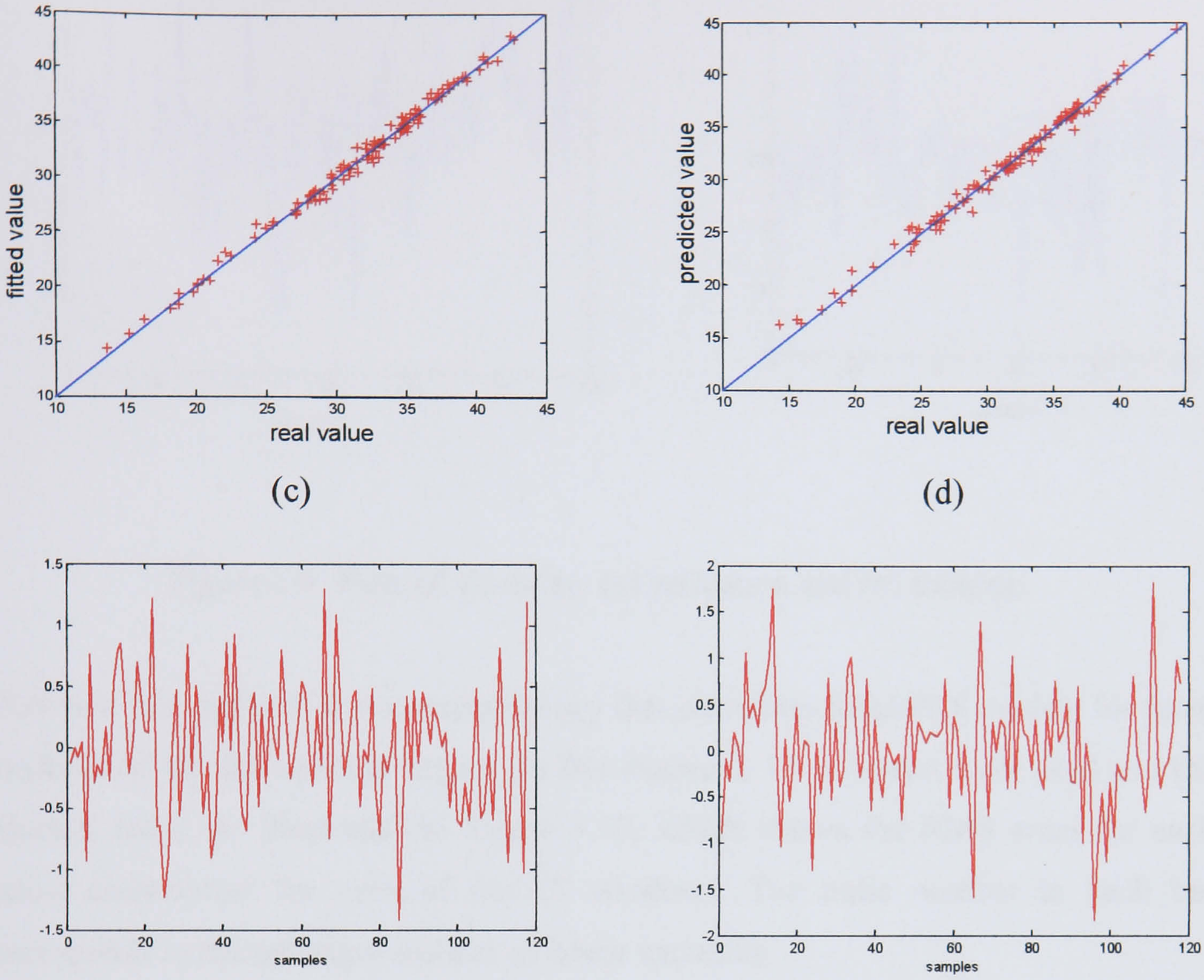


Figure 3.8. Results for total aromatics for PLS stacking: (a) training, (b) validation, (c) training residuals and (d) validation residuals.

To further investigate the performance of SWS, it is also applied and compared with the wavelength selection approaches of iPLS and GAs to predict viscosity. Figure 3.9 shows the time series plots for viscosity for the training and the validation values.

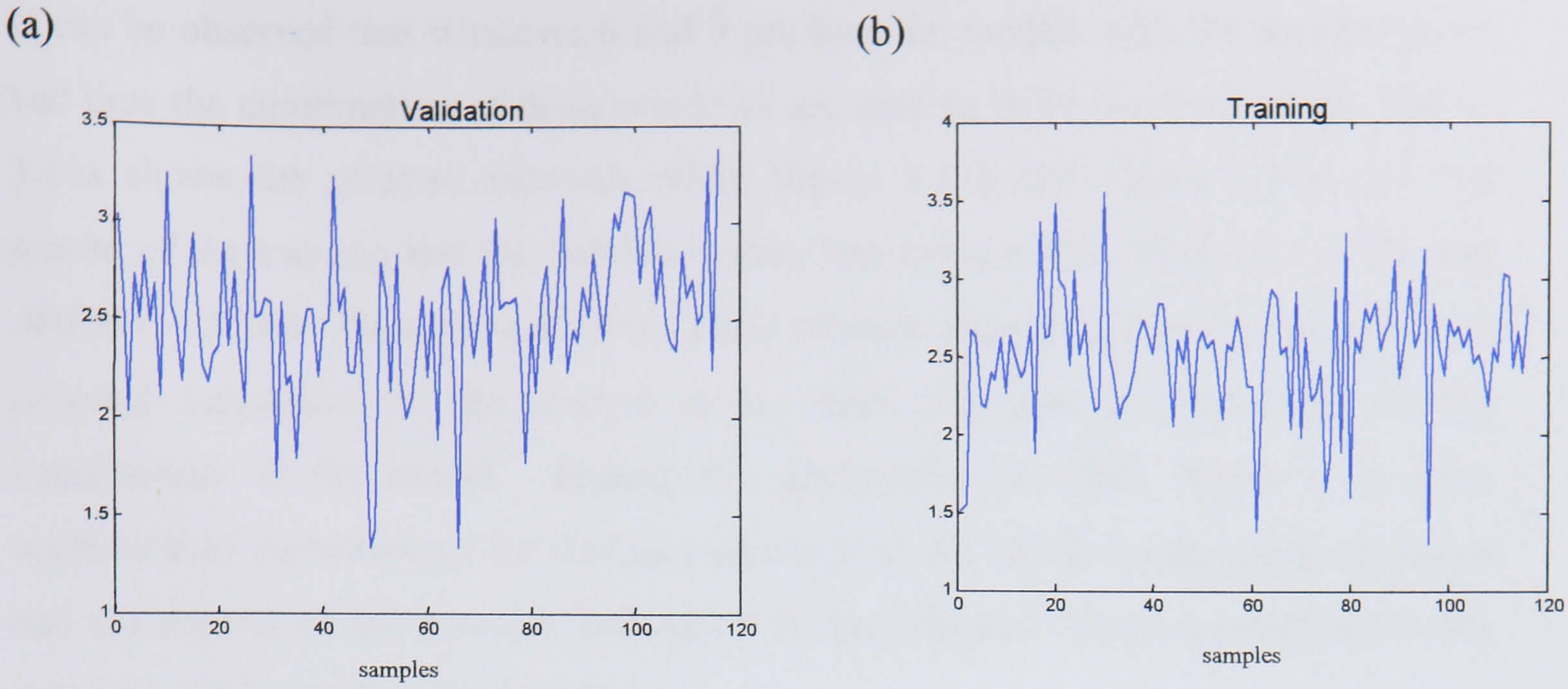


Figure 3.9. Plots of viscosity: (a) validation and (b) training.

iPLS was first applied. It is a methodology that calculates local PLS models for fixed windows of the full spectral region. In this example, 15 windows were used and the selected areas are illustrated in Figure 3.10, which shows the RMS error for each model constructed for each of the 15 windows. The italic number in each bar corresponds to the optimum number of latent variables

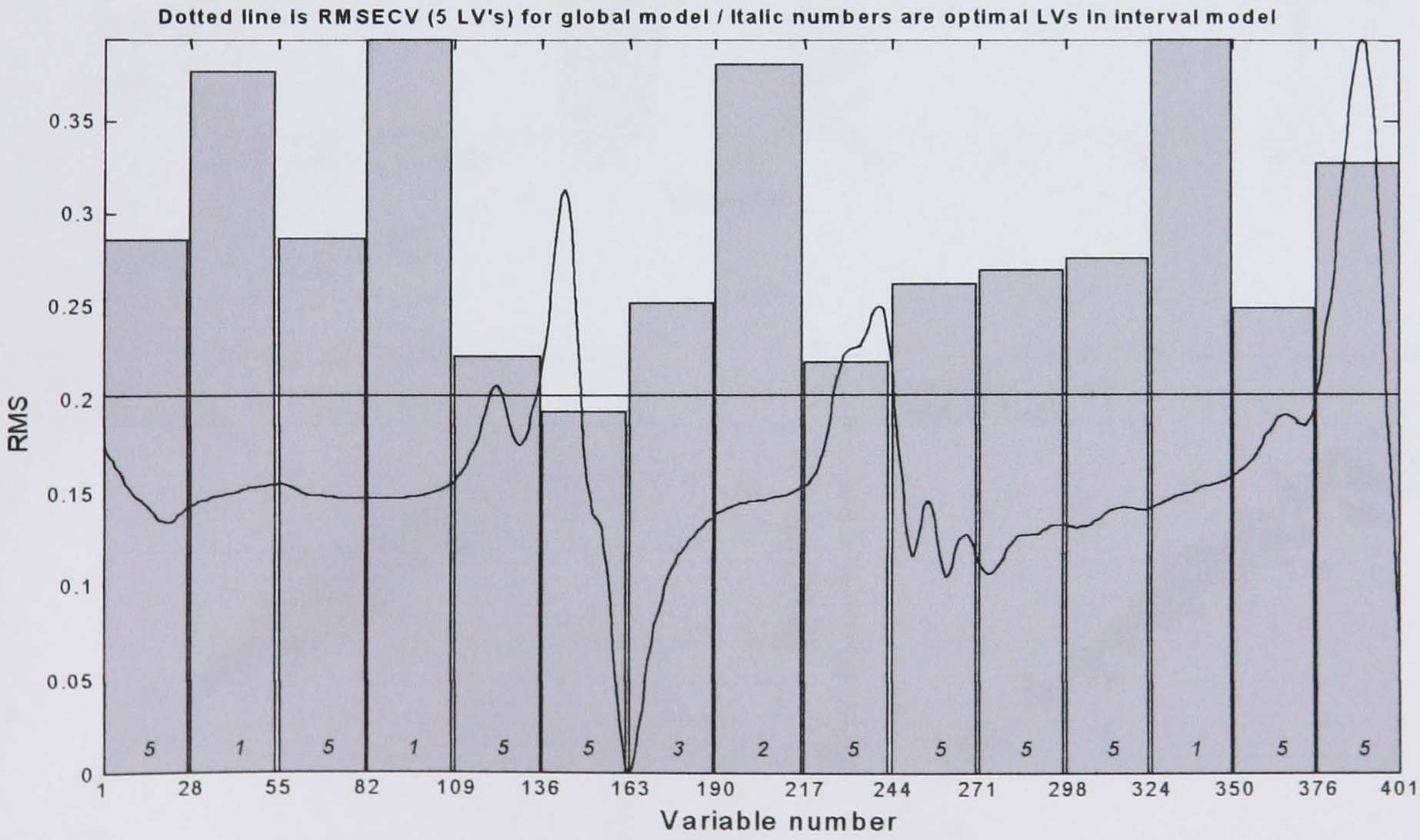


Figure 3.10. RMS error for each model constructed on each of the 15 windows after the application of iPLS.

It can be observed that windows 6 and 9 produce the models with the smallest error and thus the combination of these windows are used to build the final model. Figure 3.11a shows the selected intervals while Figure 3.11b and Figure 3.11c give the results of the training and the validation data sets respectfully: $RMSET = 0.170$ and $RMSEV = 0.160$. Thus since iPLS is a fixed window algorithm, there is an element of personal judgement by the analyst as to which windows are important for the construction of the model. Hence, the application of SWS, which is a more sophisticated methodology for finding optimal intervals for the construction of modes and not depend on the personal judgement by the analyst is more appropriate in this case and will be used in the next step.

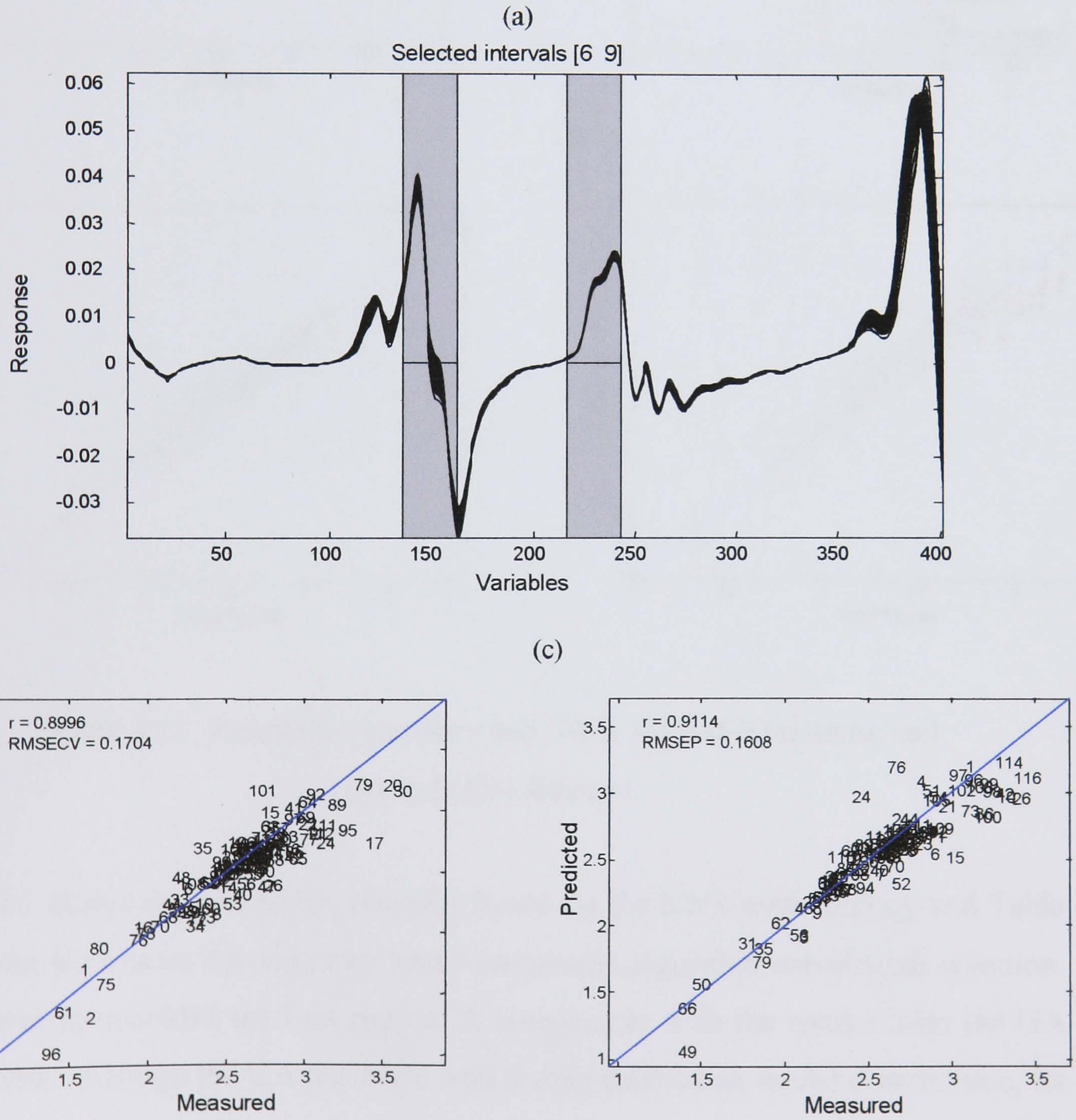
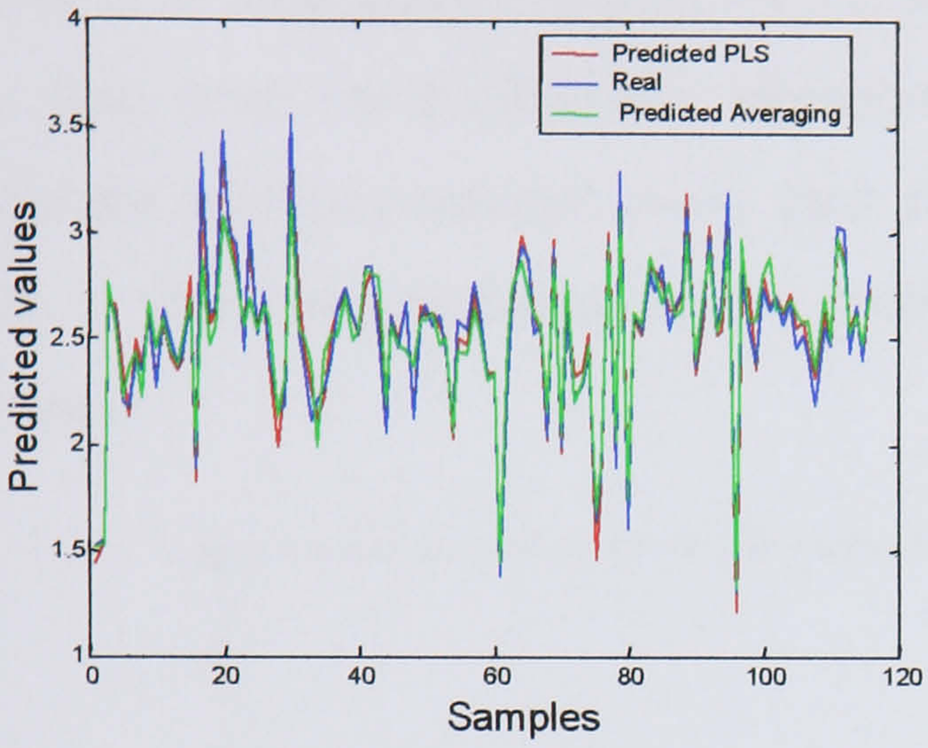


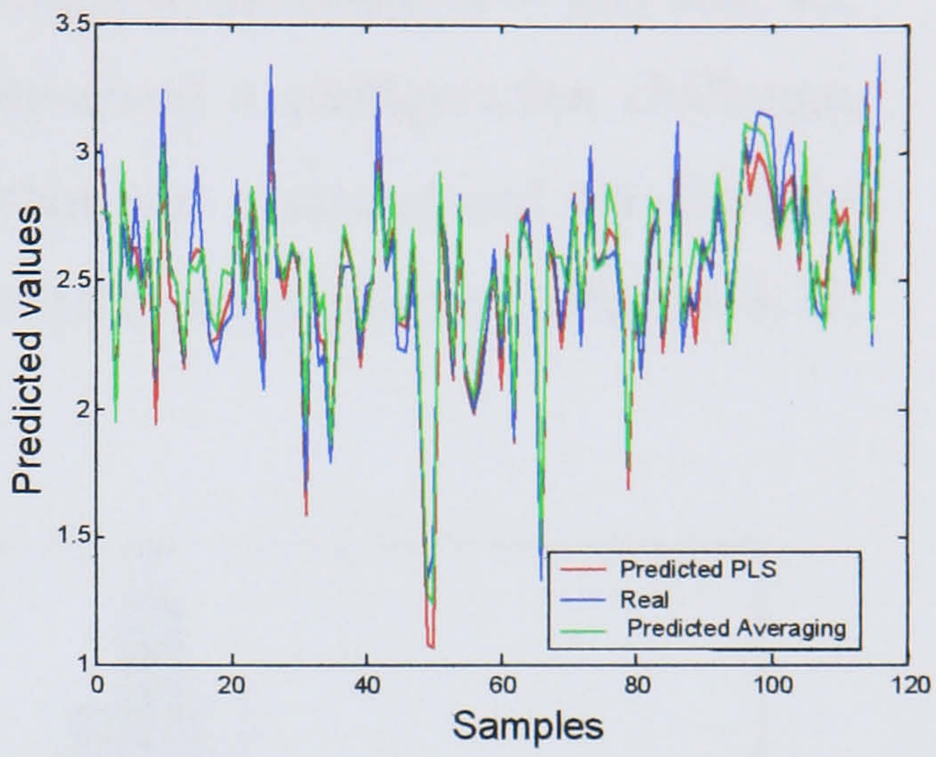
Figure 3.11. Construction of a model based on windows 6 and 9.

On the other hand, SWS is a random search based algorithm. Figure 3.12 shows the results after the application of the SWS algorithm followed by stacking. The predictions are very close to the real values.

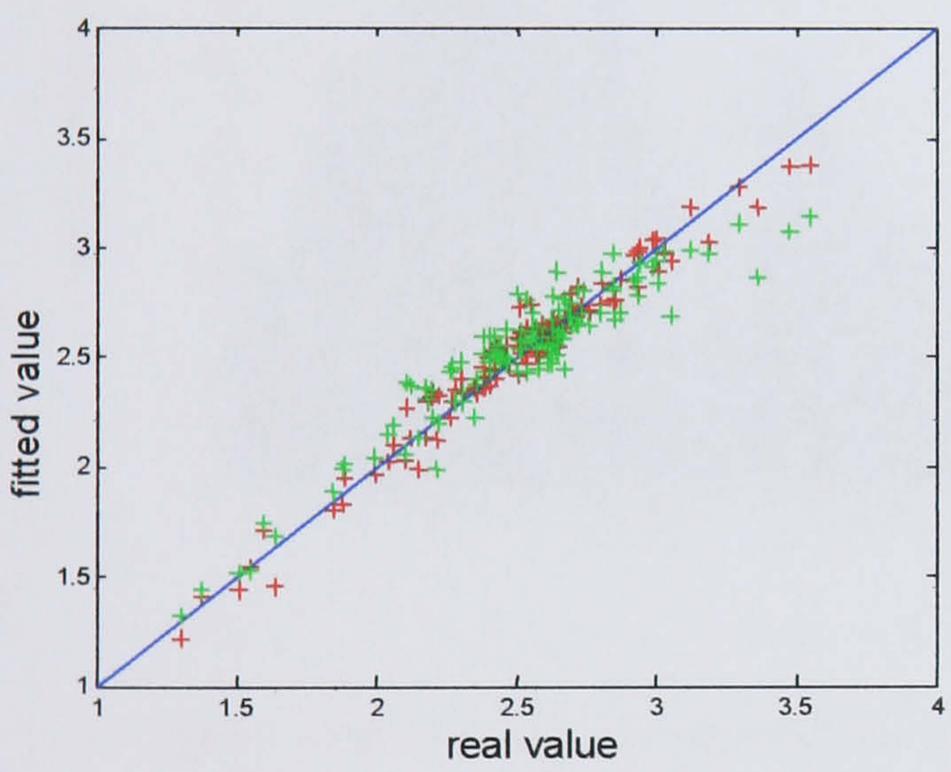
(a)



(b)



(c)



(d)

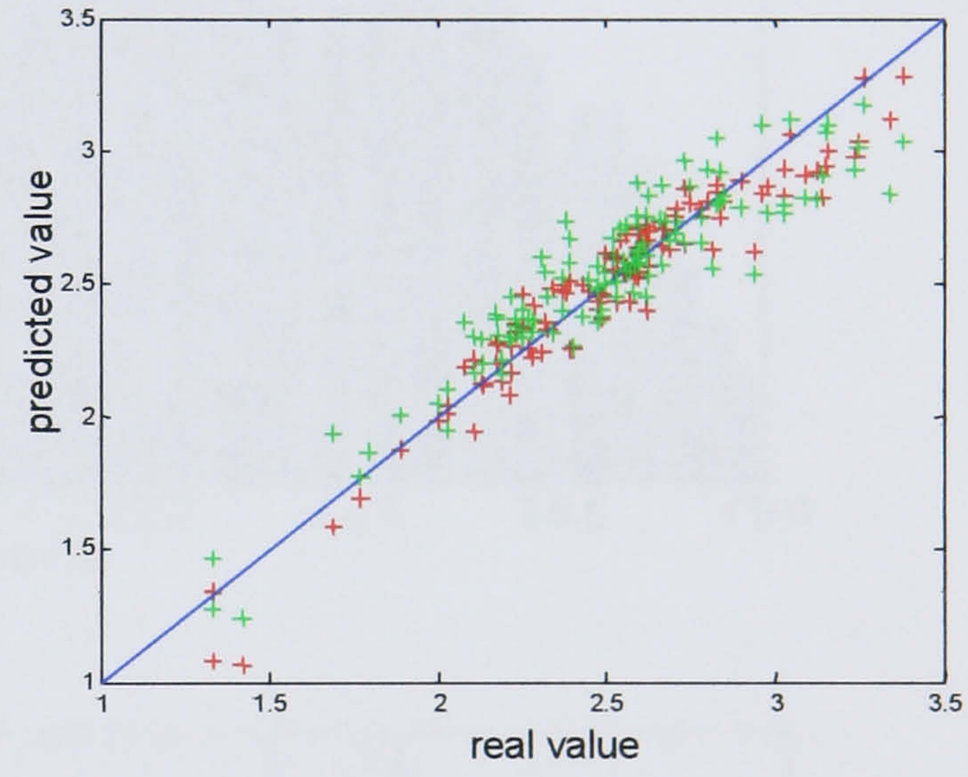


Figure 3.12. Results for viscosity and SWS: (a) and (c) training and (b) and (d) validation.

Table 3.2 shows the results for viscosity based on the SWS methodology and Table 3.3 shows the results for viscosity based on genetic algorithm wavelength selection. PLS stacking provided the best results. A comparison with the results from the GA reveals that although the GA performs well during calibration model construction, its performance during validation is slightly worse. This behaviour can be explained by examining the frequency distribution of the wavelengths selected following the application of the SWS algorithm (top) and the GA (bottom), Figure 3.13.

The wavelengths in the region around 250 were selected the most frequently by SWS (Figure 3.13, top). On the other hand the GA did not indicate any special critical regions (Figure 3.13, bottom). This could be a result of overfitting as indicated by this greater spread of wavelength selection. SWS results in this case provided a slightly improvement compared to the GA results. Moreover, the drawbacks associated with the GAs as described in section 3.4.3.2 were noticeable in the analysis of this data set, i.e. they were computationally intensive, they presented a configuration challenge, different solutions resulted every time the algorithm was executed and wavelengths with a spurious correlation to the prediction property were selected (Figure 3.13, bottom).

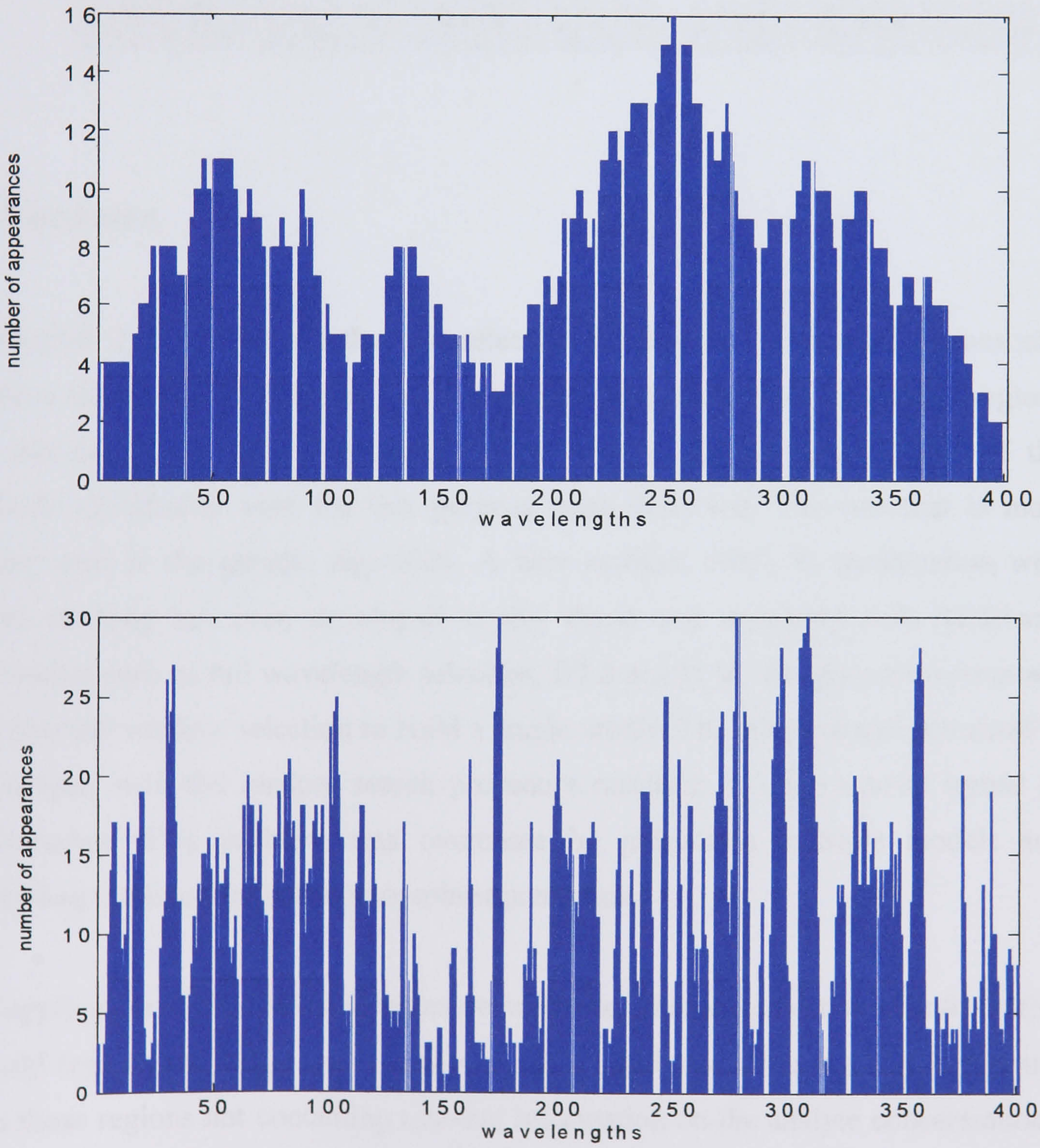


Figure 3.13. Frequency distribution of the wavelengths selected for viscosity.

Table 3.2. Results for viscosity and SWS algorithm.

	SWS with Average Stacking	**SWS with PLS Stacking**
RMSET	0.107	0.069
RMSEV	0.126	0.105

Table 3.3. Results for viscosity and genetic algorithm.

	GAs with Average Stacking	**GAs with PLS Stacking**
RMSET	0.091	0.067
RMSEV	0.131	0.111

3.8 Discussion

In Chapter 3 it was shown that the selection of informative spectral regions can improve the results by reducing the contribution of overall noise from those regions not containing relevant information on the analyte concentrations. Some of the methods specifically used for this purpose were described. The one that is most widely used is the genetic algorithm. A new method, SWS, in combination with model stacking has been developed in this thesis and compared with traditional approaches such as full wavelength selection, iPLS and GAs. The algorithm proposed uses spectral window selection to build a single model. The single model generated is not unique, with the random search procedure resulting in quite a wide spread of performance. This problem was overcome by generating multiple models and combining these to produce a more robust prediction.

The application to diesel fuel samples demonstrated that the selection of informative spectral regions can improve the results by reducing the contribution of overall noise from those regions not containing relevant information on the analyte concentrations. For this purpose, SWS in combination with stacking, has been applied and compared with full spectra PLS, iPLS and GAs. iPLS is a fixed window algorithm and there is

an element of personal judgement by the analyst as to which windows are important for the construction of the model and as a result SWS method is a superior compared to iPLS. The GA did not indicate any critical wavelength regions, they were computationally intensive and they presented a configuration challenge because of the interaction of the adjustable factors (e.g. initial population, number of generations) that influence their outcome. Thus, in this case SWS was also preferable.

To verify the performance of the SWS algorithm, it will be applied to a fermentation industrial process in the subsequent chapters.

CHAPTER 4

CALIBRATION MODELLING FOR BIOPROCESS SPECTRAL ANALYTICAL MEASUREMENTS

4.1 Introduction

Batch processing is a common strategy in the manufacture of high quality products such as polymers and pharmaceuticals with the traditional approach to ensuring high quality and consistent end-product being achieved through the following of a batch recipe. A common issue is that although the recipe is consistently applied, process differences will materialize since batch processes are complex and subject to disturbances, thus it is not straightforward to maintain consistent operating conditions. Consequently undesired variations occur between batches, resulting in the need to monitor and control an industrial batch process to maintain operation within the desired specifications. One of the challenges in batch processing is thus to track the changes in process behaviour throughout the duration of the batch (Westerhuis *et al.*, 2000, Louwerse and Smilde, 2000, Albert and Kinley, 2001, Martin and Morris, 2002; Lennox *et al.*, 2001, 2002, Sprang *et al.*, 2002, Lee *et al.*, 2004).

The case study described in this Chapter is based on an antibiotic fermentation production process from which NIR and MIR measurements are gathered. Fermentation is a process by which microorganisms convert chemical species to products of higher value. The diversity of product is significant ranging from beer and wine through to fine chemicals and antibiotics. The motivation for the choice of application is that obtaining information from fermentation processes has traditionally been an operational challenge and improvements could offer significant industrial benefits. Whilst chemical processes such as polymerisations can be challenging from a measurement perspective, biological based processes are typically more complex both in terms of measurement and the process itself.

Recently, NIR and MIR spectroscopy has been shown to provide a rich source of information with respect to the identification of chemical constituents present within a bioprocess as will been seen in section 4.3. However, the interpretation of spectral information is not straightforward as a result of the large number of variables (wavelengths) and the presence of components that exhibit overlapping absorbance features and the underlying correlation structure present between wavelengths. The successful implementation of spectroscopic instrumentation therefore requires the

application of multivariate data analysis techniques to construct a robust and reliable calibration model. Significant research in this area, although not specific to bioprocesses has been undertaken and reported in the literature (Jouan-Rimbaud et. al., 1995; Westerhuis et. al., 2000; Roggo et. al., 2003).

The objective of Chapter 4 is to apply the Spectral Window Selection (SWS) approach to the antibiotic process, where windows of wavelengths are automatically selected. The results are contrasted with more traditional approaches obtained using full-spectra PLS. It is shown that the proposed procedure outperforms full-spectra PLS, as well as enabling the identification of the critical regions of the spectra. Finally, to take account of the changes in process behaviour throughout the duration of a batch, the development of local models is also investigated.

4.2 Key Biotechnology Processes: Antibiotic and Fermentation

'*Biotechnology is the application of knowledge of living systems in order to use those systems or their components for industrial purposes*' (Bains, 1998). The term was first proposed in 1919 by the Hungarian agricultural economist Ereky and it originally meant '*all lines of work by which products are produced from raw materials with the aid of living organisms*'. Fermentation is used for the commercial growth of microorganisms and encompasses a wide range of technologies; consequently it is a key component of biotechnology today. One of the main motivations for the use of biotechnological processes is the production of drugs through fermentation processes. One class of drugs are antibiotics.

4.2.1 Antibiotic Processes

The traditional antibiotics, such as *penicillins*, were the first drug products to be developed through biotechnology. The *penicillins*, *streptomycins*, and a host of other antibiotics provided a significant breakthrough in the 1940s and 1950s in terms of healthcare. Since then, the pharmaceutical industry has built on this base to develop a

range of new antibiotics. There are four routes to developing new antibiotics (Bains 1998):

1. Hybrid antibiotics: where the synthesis of the antibiotic is a result of a number of enzyme steps in a bacterium or fungus.

2. Novel metabolites: these are produced by microorganisms and plants.

3. Animal antibacterials: where animals, especially invertebrates, produce a wide range of materials (proteins or peptides) that kill bacteria.

4. Pathogenesis-targeted antibiotics: where the goal is to identify what makes a bacterium a pathogen and then the molecular mechanism is blocked as opposed to killing the bacterium.

The chemical complexity and composition of different kinds of substrates or nutrients of the fermentation media varies considerably between processes and this is a challenge in terms of deriving robust generic methods of analysis and the modelling of the constituents of the process. An important consideration for the control of growth at the production scale is to obtain a consistent inoculum. Multivariate analysis provides important information on the state of the inoculum and insight into the progression of the production batch. Hence multivariate analysis is gaining popularity for the simultaneous determination of compounds in antibiotic processes (Pena *et al.*, 2002a, 2002b). The main application in Chapter 4 is a fermentation antibiotic process, therefore an overview of fermentation processes is given in the following section.

4.2.2 An Overview of Fermentation Processes

Fermentation may be carried out as a batch, continuous or fed-batch process. In a batch fermentation, the inoculated culture will pass through a number of phases, including the lag, exponential, stationary and death phase. The length of the different phases varies significantly with different cell types. The lag phase is a period after

inoculation in which no significant growth takes place. This is a time of adaptation for the culture as it becomes accustomed to a new environment following inoculation. It is important in a commercial process to keep the length of the lag phase as short as possible. This can be achieved through the use of a suitable inoculum and batch media, Stanbury and Whitaker, (1984).

In the exponential phase, the growth rate of cells gradually increases with the cells growing at a constant rate until they become substrate limited. This phase is described by the equation:

$$\frac{dx}{dt} = \mu \cdot xc \qquad 4.1$$

where xc is the concentration of the cells, t is time and μ is the specific growth rate.

Consequently

$$x_t = x_0 e^{\mu t} \qquad 4.2$$

and finally

$$\ln x_t = \ln x_0 + \mu t \qquad 4.3$$

A plot of log of concentration ($\ln x_t$) versus time (t) would be a straight line with slope equal to μ. During the exponential phase, the cells grow at a rate approaching maximum specific growth rate, μ_{max}. The general objective of this phase is to produce a high concentration of cells as rapidly as possible.

The stationary phase in many systems involves a balance between some cells dying and others that continue to grow. In most systems, it is during this phase that product accumulates, whilst finally in the death phase, an exponential decrease in the number of living individuals is often observed. The fermentation process terminates with the harvest of the broth and the extraction of the product in downstream processing.

Fed-batch fermentation processes are described by Bungay, (1993) as the type of system where "*nutrient is added when its concentration within the broth falls below some set point*" following which, nutrient is added in a controlled manner. To control the addition of the feed, the best strategy is to monitor the concentration of the nutrient itself in the fermenter and take appropriate action to maintain concentration levels. Through controlling nutrient addition, the reaction can proceed at a high rate of production by avoiding nutrient limitations. Another advantage of fed-batch processes is that since growth and product formation occur during different times within the batch, they can be optimised independently, (Perry *et al.*, 1997).

Moreover, depending on their need for oxygen, two main groups can be distinguished for the microrganisms, aerobes and anaerobes: (a) Anaerobes are the organisms that cannot use oxygen as they lack the respiratory system, (b) Aerobes use oxygen and extensive aeration is required in most aerobic processes (Madigan and Martinko, 2000).

Some other indicators of fermentation performance are explained in section 4.2.3, while those factors that influence fermentation performance will be described in section 4.2.4.

4.2.3 Indicators of Fermentation Conditions

Three indicators of the state of any aerobic fermentation process (i.e. fermentation in the presence of air) are: oxygen uptake rate (OUR), carbon dioxide evolution rate (CER) and respiratory quotient (RQ). These variables are not measured directly but are calculated from the inlet and outlet gas concentrations and air flowrate.

According to Carr-Brion (1991), OUR is given by the following relationship:

$$OUR = \frac{G_{in}}{100}(\%O_2^{in} - \%O_2^{out} \times \frac{\%N_2^{in}}{\%N_2^{out}}) \qquad 4.4$$

where

OUR is measured in mol/s,

G_{in} is the aeration rate based on inlet flow-rate (mol/s)

$\%i^{in}$ is the mole percentage of component i in the inlet air to the fermenter

$\%i^{out}$ is the mole percentage of component i in the exit gas from the fermenter

and for CER:

$$CER = \frac{G_{in}}{100}(\%CO_2^{out} \times \frac{\%N_2^{in}}{\%N_2^{out}} - \%CO_2^{in}) \qquad 4.5$$

According to Bailey and Ollis (1986), RQ is defined as:

$$RQ = \frac{moles \quad CO_2 \quad formed}{moles \quad O_2 \quad consumed} \quad \text{or} \quad RQ = \frac{CER}{OUR} \qquad 4.6$$

RQ is an important physiological parameter that indicates if an aerobic culture is balanced or not in terms of the consumption of substrates. The substrates are foods for growing microorganisms and the formation of product. The substrates can be categorised according to what they provide:

- Carbon substrates: mollases, malt extract, starch and dextrins, cellulose, sulphite liquor, whey, methanol, oil, gas
- Nitrogen substrates: ammonia, corn steep liquor, soy protein, yeast extracts etc.

The value of RQ depends on which food source is being used. For example when carbohydrate is used, RQ is one (e.g. MDX-Maltodextrin into glucose components):

$$C_6H_{12}O_6+6O_2 \longrightarrow 6CO_2+6H_2O$$

hence RQ=6/6=1

When fat is used (e.g. RSO-Rape Seed Oil):

$$2C_{51}H_{98}O_6+145O_2 \longrightarrow 102CO_2+98H_2O$$

hence RQ=102/145=0.70

RQ is therefore a useful indicator of the fermentation conditions since it can indicate when a particular food source becomes exhausted and growth switches to the use of an alternative nutrient. The values of RQ will typically lie within the range 0.70 to 1.

4.2.4 Influencing Features for the Fermentation Performance

These are several factors that influence the fermentation performance. The impact of the environment is critical. To design the optimum bioreactor system, according to Smith (1988), the following guidelines should be closely followed:

1. The reactor should be designed to exclude the introduction of contaminating organisms but simultaneously contain the desired organisms.
2. The dissolved oxygen level must be maintained above critical levels by aeration and culture agitation for aerobic organisms.
3. Environmental parameters such as temperature and pH must be controlled.
4. The culture volume must be well mixed.

There tends to be a lack of knowledge regarding the environmental conditions that will generate optimal yield of product. This, to an extent, is due to the limited measurements opportunities. There are a number of important parameters in fermentation process control that should be considered:

- sterilization, (Stanbury and Whitaker, 1984),
- operating temperature, (Carr-Brion, 1991),
- good mixing, (Carr-Brion, 1991) and
- dissolved oxygen (Vogel, 1997).

The impact of these parameters is considered as undesirable changes and could have a serious impact on measurement capabilities and fermentation performance. The measurement of biomass is an important aspect in bioprocess monitoring and control. Biomass is a key analyte whose determination is useful in terms of assessing the progress of a submerged culture bioprocess and it provides a reliable indicator with regard to the timing of inoculum addition and cell harvest. Several variables for the maximisation of process efficiency inherently depend on biomass determination, (Roubos *et al*., 2001).

There are many aspects that influence a fermentation process such as environmental conditions and the fact that the metabolic processes of the microorganisms are very complicated. The limited understanding and the presence of unpredictable disturbances from the operational environment, complicates the modelling of a bioprocess and they are the reason that a fermentation process exhibits both non-linear and dynamic properties. In particular for fed-batch fermentation processes, significant research has been undertaken by a lot of researchers with respect to real time optimisation and control (Zuo and Wu, 2000, Chung *et al*., 2006) and to process monitoring using multivariate statistics (Lennox *et al*. 2000; Lennox *et al*. 2001, Gregersen and Jorgensen, 1999).

4.3 NIR and MIR Analysis in Fermentation Processes

Spectroscopy has the potential for the real-time monitoring of the progress of a fermentation process. NIR and MIR spectroscopy in particular have been shown to provide a rich source of information. The most common method applied, to date, has been NIR. The motivation for using NIR spectroscopy to determine the individual component concentrations in a fermentation process, considered in this thesis, was the research reported by other workers in biotechnology and in particular fermentation processes.

McShane and Cote, (1998) applied NIR spectroscopy for the determination of glucose, lactate and ammonia in cell culture media. 'Cell culture' is the process where

cells, when removed from animal tissue, will continue to grow if provided with the apropriate nutrients and growth factors. In this application, a calibration model that gives reasonable prediction errors was developed using a combination of spectra from cell culture media samples and aqueous mixtures of glucose, lactate, ammonia, glutamate and glutamine.

Yeung *et al.*, (1999) used NIR for the measurement of selected contaminants in a complex biological process stream and to aid their selective removal. The recovery of a yeast intracellular enzyme, alcohol dehydrogenase (ADH), produced in *Saccharomyces cerevisiae* (Baker's yeast), was chosen for this study. The NIR spectra were obtained using a low cost in-house-built spectrophotometer where NIR radiation was collected between 1900 and 2500 *nm* .

Forbes *et al.*, (2001) used NIR to quantify potency and lipids in monensin fermentation broth that is produced by the fermentation of the microorganism *Strepromyces cinnamonensis*. Especially for potency, the NIR calibration model, was stable as the model was established six months prior to this study and no calibration adjustments were required for more than a year.

Macedo *et al.*, (2002) demonstrated the applicability of NIR to predict lactic acid and exopolysaccharide production and lactose consumption during a *Lactobacillus rhamnosus* fermentation.

Cimander and Mandenious, (2002) developed calibration models for the monitoring of an *Escherichia coli* fed-batch process for tryptophan production. The results showed that for both biomass and tryptophan, the detection of process deviations is possible from the different models developed.

Tamburini, et. al., (2003) determined the concentration of glucose, lactic acid, acetic acid and biomass in liquid cultures of microorganisms of the genera *Lactobacillus* and *Staphylococcus*. They concluded that satisfactory predictions were achieved and that the predictive ability of the model was better at intermediate concentrations than at very low or very high values.

Tosi *et al.*, (2003) developed calibration models for the determination of biomass, glucose and lactic and acetic acids during the fermentation of *Staphylococcus xylosus*. The models were successfully extended to other strains that were differently shaped but grow in the same medium and fermentation conditions, i.e. to *Lactobasillus fermentum* and *Streptococcus thermophylus*.

Lopes *et al.*, (2004) reported a NIR application on an industrial multistage antibiotic-like molecule (API-active ingredient product) production process. A fed-batch cultivation of a *Streptomycetes* strain formed the basis of the study. For this, a complex medium was used that contained soybean flour and a carbon source. In this work, the feasibility and benefits of using NIR was discussed and effectively applied to several process stages including (a) the quality assessment of fermentation raw materials, (b) fermentation process monitoring, and (c) downstream API purification process monitoring.

Kasprow *et al.*, (1998) used NIR for the characterisation of yeast extract compositions; and Riley *et al.*, (1998) for the quantification of nutrients and byproducts in insect cell (*Spodoptera frugiperda*) bioreactors. Navratil *et al.*, (2005) used NIR for the monitoring of biomass, glucose and acetate for growth rate control of *Vibrio cholerae* fed-batch cultivation; and Blanco *et al.*, (2005) for the identification of glucose, ethanol and biomass in an alcoholic fermentation that was run using *Saccharomyces cerevisiae* yeast.

Harvey and McNeils' group have published a number of papers in the area of fermentation monitoring with NIR including a tutorial on employing NIR methods for fermentation modelling and control, (Arnold *et al.*, 2002 part 1 and part 2, 2003). In particular they have focused on various microbial and antibiotic processes, for example:

- the filamentous bacterium *Streptomyces fradiae* process: for the modelling of four key analytes (methyl oleate, glucose, glutamine and ammonium), Arnold *et al.*, (2000); for the prediction of tylosin and oil measurements, Vaidyanathan *et al.*, (2000); for the monitoring of tylosin concentration, Arnold *et al.*, (2001); for the monitoring of a process using two types of media

and a number of replicate runs for each media, Vaidyanathan *et al.*, (2001a); and for the monitoring of the biomass to obtain morphological-related information, Vaidyanathan *et al.*, (2003),

- the *Penicillium chrysogenum* process: for the modeling of mycelian biomass, total sugars (lactose and sucrose) and ammonia where variations were introduced, Vaidyanathan *et al.*, (2001b); additionally for the modelling of penicillin and extracellular protein Vaidyanathan *et al.*, (2001c),

- the *Escherichia coli* process: for biomass modeling, Arnold *et al.*, (2002a),

- the *Sphingomonas paucimobilis* process: for the monitoring of polysaccharide and biomass in the biopolymer formation, Giavasis et. al., (2003),

- mammalian cell cultivation: for the modelling of glucose, lactate, glutamine and ammonia, Arnold *et al.*, (2003), and

- the *Pichia pastoris* fed-batch bioprocess, which is a highly complex and high cell density process: for the modelling of the key process analytes including biomass, glycerol and methanol, Crowley *et al.*, (2005).

Finally, Vaidyanathan *et al.*, (1999) investigated the validity of measuring biomass using NIR spectra for five different microorganisms: *Escherichia coli, Streptomyces fradiae, Penicillium chrysogenum, Aspergillus niger and Aureobasidium pullulans.*

Overall observations indicate that NIR spectroscopy can improve fermentation process operation by providing rapid, non-destructive, multiconstituent analyses of the fermentation broth and media. NIR spectroscopy can be implemented with minimal or no sample preparation and pre-treatment to provide a direct on-line measurement that can be utilised in the fermentation monitoring and control strategy. NIR is potentially a useful methodology for application in both support laboratories and directly in the manufacturing environment.

MIR has not been as widely applied as NIR, possibly due to the instrumentation being more expensive (McClure, 1994). However, the MIR region may offer certain benefits compared to NIR as there are stronger absorbances and more distinct spectral features. The strong water absorbance, however, results in very short penetration depths of the MIR radiation. Nevertheless, in recent years MIR has shown promise as an analytical tool in bioprocessesing.

Fayolle *et al.*, (1997) demonstrated the applicability of MIR for the monitoring of lactose, galactose, lactic acid and biomass concentration in samples of *Lactobacillus Bulgaricus*.

Doak and Phillips, (1999) monitored the *Escherichia coli* fermentation using MIR. They demonstrated the utility of MIR to quantitatively characterise both glucose and acetate in process samples. The system demonstrated excellent stability properties that were unaffected by agitation or aeration rates or shutdowns imposed by routine maintenance procedures.

Sivakesave *et al.*, (2001) developed calibration models for the monitoring of biomass, glucose and lactic acid in the lactic acid fermentation *Lactobacillus casei* using NIR, MIR and Raman spectra and concluded that the MIR predictions were better compared to the other approaches.

Kansiz *et al.*, (2001) monitored the acetone-butanol (ABE) fermentation process using MIR. The ABE process is performed by solvent-producing bacteria of the genus *Clostridium beijerinckii*. MIR was used for the prediction of the pure analytes (acetone, n-butanol, glucose, acetic acid and butyric acid) with the accuracy and precision demanded for process monitoring.

Kormann *et al.*, (2003) applied MIR to monitor simultaneously the major metabolites (fructose, acetic acid, ethanol) in a *Gluconacetobacter xylinus* fermentation for the production of gluconacetan. The final models were obtained by the addition of a number of off-line samples. Kornmann *et al.*, (2004a) also used the same fermentation for the monitoring of fructose, acetate and gluconacetan using a new methodology where a number of standard samples are periodically updated by spectra collected on-

line. Moreover for the same fermentation, phosphate and ammonium concentrations were also monitored (Kornmann *et al.*, 2004b).

Finally, Crowley *et al.*, (2000) used MIR for the monitoring of *Pichia pastoris* fed batch process successfully; and Macauley-Patric *et al.*, (2003), used MIR for the quantification of D-sorbitol and L-sorbose during a *Gluconobacter suboxydans* fermentation and the model performed well. This system is especially challenging since the two analytes are very close in terms of chemical similarity and often large quantities of one analyte are associated with low quantities of the other during the process.

4.4 Application to an Antibiotic Fermentation Process

It was clear from previous studies that spectroscopy offers potential improvements in terms of fermentation operation. However, improved methods of spectral data interpretation may offer greater insight into the monitoring of the progression of a fermentation process. To investigate the improvements that are possible, the application and interpretation of spectral data from an industrial antibiotic fermentation process is now considered. It should be noted that all values have been scaled for confidentiality reasons to mask the true values.

4.4.1 Experimental procedure

The process of interest was an industrial pilot-plant scale fermentation involving two stages, a seed stage and a production stage. Biomass is grown in the seed stage before being transferred to the final stage for the production of the desired product. The final stage is a fed batch process and lasts approximately 140 hours. NIR and MIR measurements were employed to quantify the concentration of the key constituents of the broth. NIR and MIR measurements were collected on-line from the final stage of the process and both experimental design data and standard operating data form the basis of the subsequent analysis. Multiple analyte concentrations, such as product, sugar, phosphate, lipids, ammonia, pH, viscosity and urea, were measured by off-line

assay during the course of the fermentation but the initial studies focused on two key parameters, product concentration and ammonia concentration in the broth.

In fermentation applications, the risk of contamination needs to be considered. For this reason, non-invasive measurements have significant benefits if they can be made. At small-scale operation it is common to have a glass window in a vessel, thereby enabling non-invasive NIR measurements to be recorded. Larger scale vessels do not have windows hence non-invasive measurement is not possible. However the relative merits of invasive and non-invasive spectroscopy were considered at the small scale. Due to the path length of MIR spectroscopy, it can only be implemented invasively.

Within the study, a number of probes were investigated for the monitoring of the process, Table 4.1.

Table 4.1. Retails of investigated study

	NIR	MIR
Non invasive	Zeiss, Foss	-
Invasive	ABB	Linx 5-10

Two non-invasive NIR instruments were considered to understand the advantages of a wider wavelength range. The first was a Zeiss Corona 45 NIR (Figure 4.1), which operates in the reflectance mode with the wavenumbers lying in the range 950 – 1700 *nm* with a resolution of 6 *nm*. The instrument was equipped with a diode array detector of focal length 13 *nm* with a sampling area diameter of 15 *nm* and 15 detection fibres were situated around the inner edge of the lens. The NIR data were recorded every 15 minutes.

Figure 4.1. Zeiss Corona 45.

Foss also provided a non-invasive NIR probe (Figure 4.2). Its scan time was approximately 40 seconds with an average of 32 scans. The scanning range was 400-2500 *nm*. Two detectors were used: (a) Silicon(Si): 400 to 1100 *nm* (25,000 cm^{-1} - 9091 cm^{-1}), and (b) Lead sulphide (Pbs): 1100-2500 *nm* (9091 cm^{-1} to 4000 cm^{-1}) The data interval was 2.0 *nm* and the scan speed 1.8 scans/second. The acquisition of spectroscopic data was carried out using the VISION spectral analysis software (FOSS NIRSystems, Silverspring, MD). The wavenumbers below 700 *nm* and over 2200 *nm* were considered to be outside the region of the instrument and were therefore excluded from the analysis. Scans were taken every hour.

Figure 4.2. Foss non-invasive NIR instrument.

The invasive NIR instrument utilised was a Fourier transform (FT) near infrared analyser manufactured by ABB (Figure 4.3). The spectrometer was fitted with a fibre optic launcher that sends modulated light to the probe and which received the spectral information back from the probe. The software used to collect the data was FTSW100 process control software from ABB. A diffuse reflectance probe was inserted into the

fermentation vessel and connected to the spectrometer via a fibre optic bundle. The spectrometer is capable of collecting data with a resolution of $2\,cm^{-1}$ over the range 12000 to 4000 cm^{-1}. Each spectrum consisted of the average of 256 scans. The sampling frequency was approximately 15 minutes.

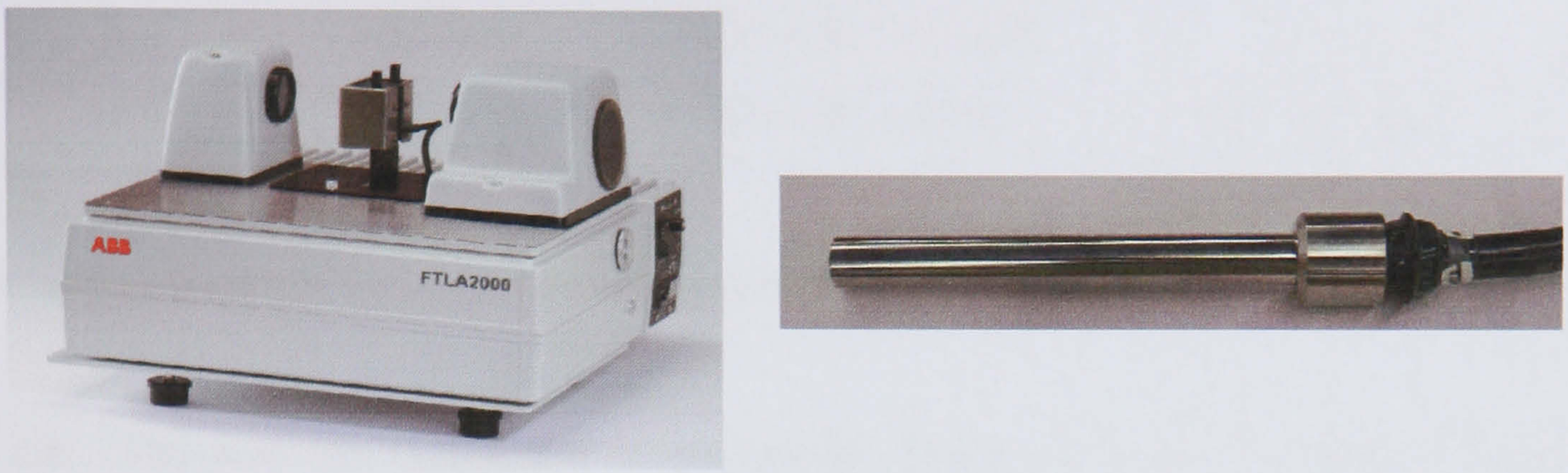


Figure 4.3. Invasive FT NIR ABB.

The MIR spectra were collected from a Linx 5-10 process development instrument manufactured by Spectraprobe (Figure 4.4). The probe and spectrometer were integrated into a single unit. A Hastelloy probe with a silicon attenuated total reflectance (ATR) crystal and a 128 element pyro-electric array detector was used. Chalcogenide fibres transported light to and from the sample. The wavenumber range was 1000-2000 cm^{-1}. The sampling frequency was approximately 18 minutes.

Figure 4.4. Linx 5-10 instrument.

To extract quantitative information from the spectra and to demonstrate the potential applicability of the instrument for subsequent in situ monitoring, experiments were performed and grouped into four data sets that corresponded to four separate data collection periods, each of which lasted several weeks (Table 4.2).

Table 4.2. Fermentation experiments performed

Data Set	Type of Batches	Number of Batches	Instrumentation
1	Standard	Seven: S1 to S7	Zeiss
2	DOE	Eight: E1 to E8	Zeiss & Linx 5-10
3	Standard	Five: SNI1 to SNI5	Foss
4	Standard	Six: SI1 to SI6	FT NIR ABB

Initially two sets of fermentation experiments were carried out:

- The first set comprised seven batches (referred to as 'standard batches'), batches S1 to S7. The seven standard batches were run under similar conditions and only NIR measurements from the Zeiss instrument were available. Natural variation resulted in terms of the degree of variability in the resulting data.

- The second data set was formed from eight batches (referred to as 'Design of Experiment batches'), E1 to E8. Both Zeiss and Spectraprobe instruments were used simultaneously. Experimental design techniques provide information rich data from a relatively limited number of experiments and thus is an excellent way to develop and apply fault detection analysis methods. However in this particular application, it was used for the construction of robust calibration models. The DoE strategy used is discussed in detail in section 4.4.5.

Subsequently, for comparison purposes two more sets of fermentation experiments were carried out in different time periods (additional standard batches).

- For the first data set, the Foss instrument was used, batches SNI1 to SNI5, and for the final group, the invasive ABB instrument was considered, batches SI1 to SI6.

The results of the last two data sets will be reported in Appendix A (for the Foss non-invasive instrument) and Appendix B (for the ABB invasive instrument).

4.4.2 Data Pre-treatment

The strategy for data pre-treatment described in Chapter 2 was adopted. First the batches were partitioned into two sets: training and validation. The assignment of whether a batch was used for training or validation was through a combination of process knowledge and observation, to ensure that representative information from the widest possible set of conditions was captured in both the training and validation data sets. This approach is preferable to a random selection of batches which when there are a limited number could result in loss of coverage of the operating region. The product and ammonia concentrations were measured using an off-line assay. The number of values recorded throughout the duration of a batch was of the order of ten. A cubic spline algorithm (Syam, 2003) was applied to obtain values of concentration at the same sampling times as the spectral measurements. Figure 4.5 shows an example of the values of batch S3 after the cubic spline algorithm was applied, where the blue points correspond to the assay values and the red line corresponds to the splined values.

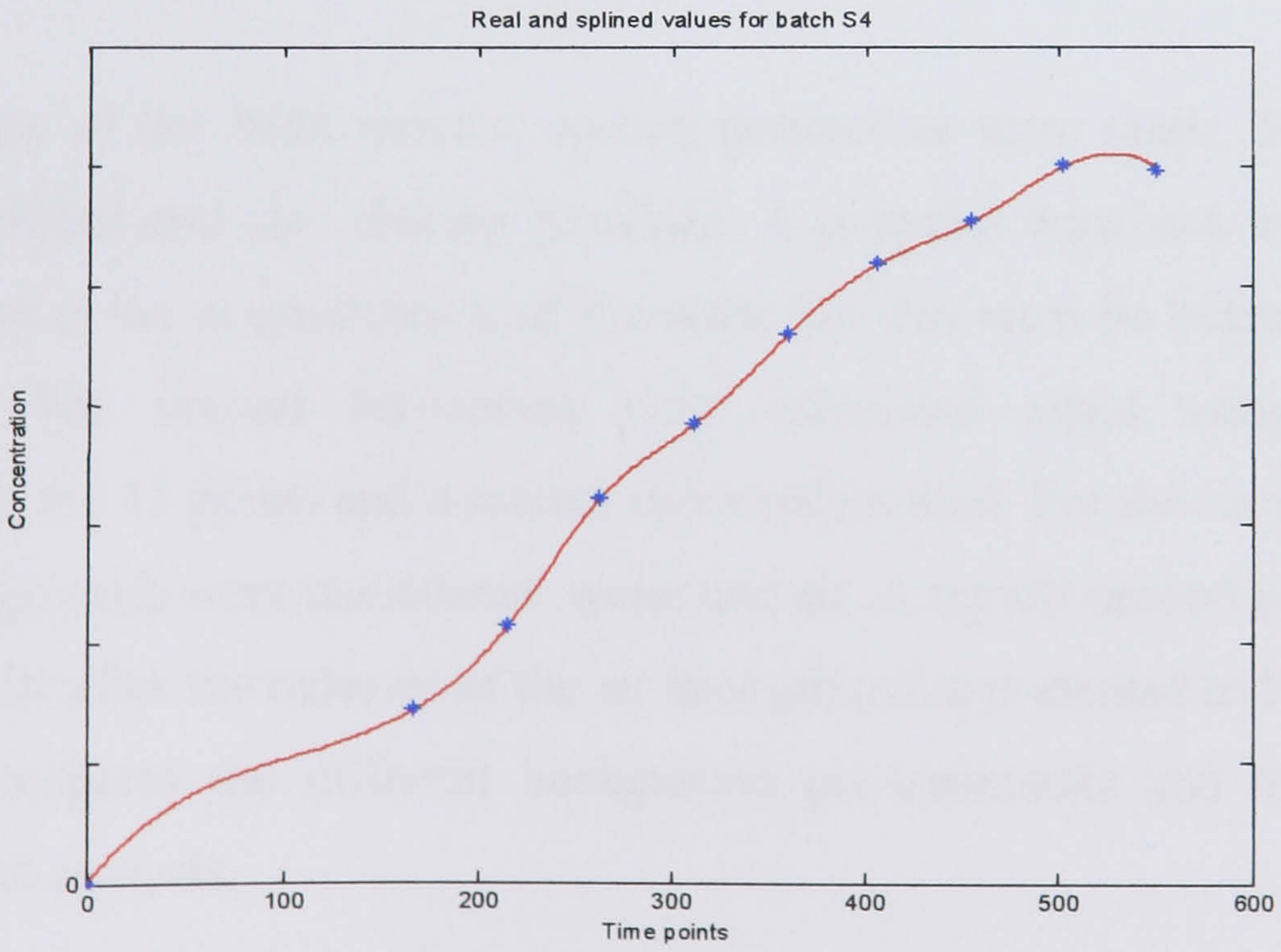


Figure 4.5: Splined values for batch S3, where '*' correspond to the assay values.

An example of the NIR spectra logged over the course of a batch is shown in Figure 4.6 for data set 1. The raw NIR spectra for one batch (Figure 4.6, left) are displayed. The Zeiss instrument was calibrated before each run and readings from both a black background and from a white tile were taken. After that, the instrument was positioned next to the fermenter, and scanning was initiated before the final stage medium was inoculated. The next step in the analysis was to take first derivatives of the NIR spectra (Figure 4.6, right) to remove the baseline offset. The first derivatives were calculated using Savitsky-Golay smoothing (Gorry, 1990) over 11 points and utilising a second order polynomial.

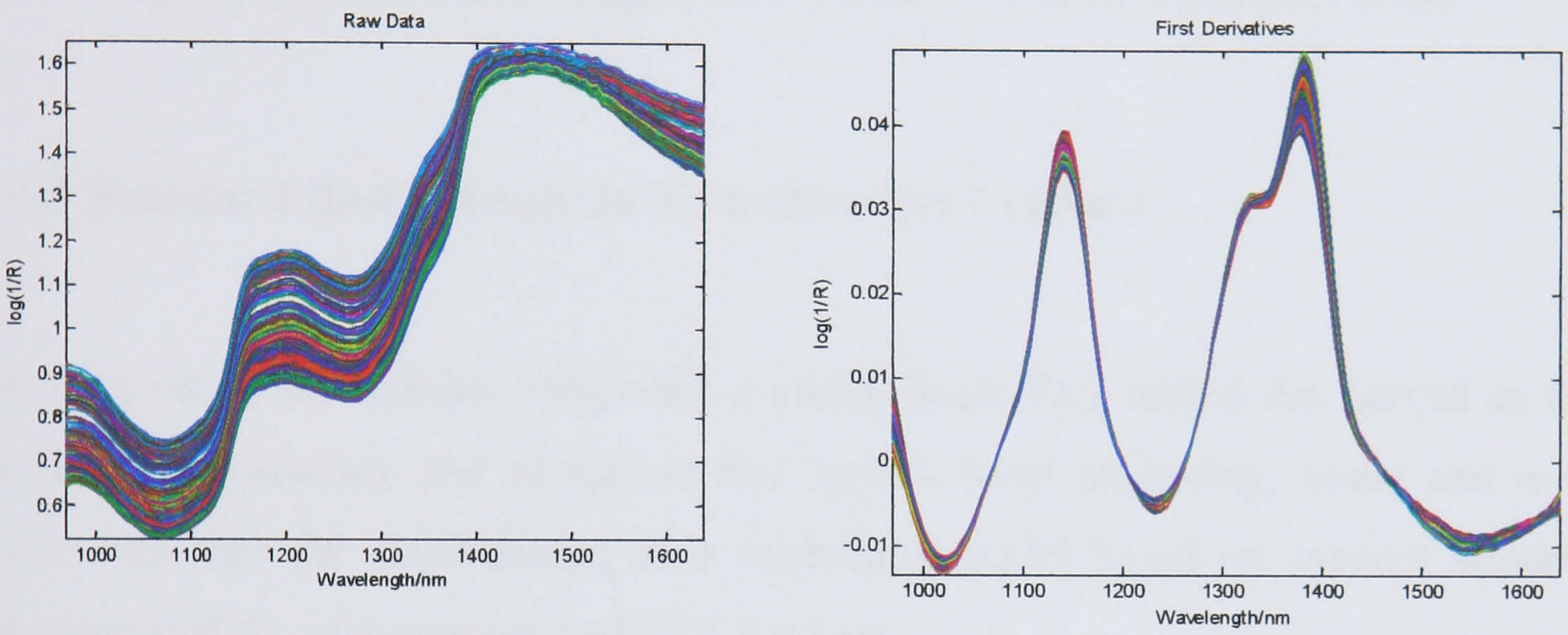


Figure 4.6. Example for raw NIR spectra (left) and first derivative spectra (right) for the standard Zeiss batches S1 to S2.

In the case of the MIR spectra, second derivatives were taken. Second derivatives remove offsets and also sloping baselines. A potential drawback when using second derivatives is the magnification of the noise but this must be balanced against offset removal. The second derivatives were calculated again using Savitsky-Golay smoothing for 11 points and a second order polynomial. For the Linx 5-10 instrument, two backgrounds were considered: water and air. A typical second derivatives spectral MIR profile after the removal of the air background is presented in Figure 4.7. Section 4.4.5.3 compares the different background pre-treatments and the impact on the subsequent analysis.

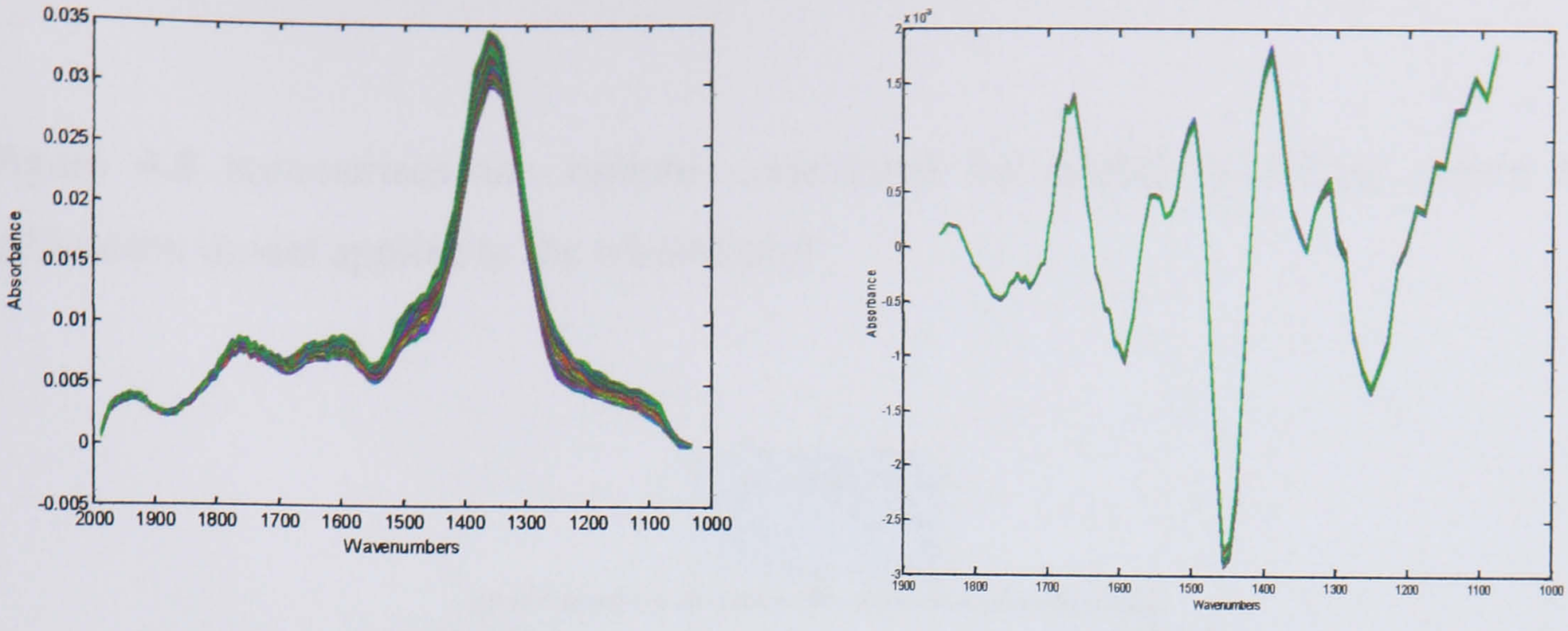


Figure 4.7. Example of MIR spectra after the removal of the air background (left) and their second derivatives (right) for the Linx 5-10 DoE batches E1 to E8.

4.4.3 Standard Batch Analysis from Data Set 1 (Zeiss)

For data set 1, the analysis composed a global linear PLS model that served as the baseline methodology and compared this against local modelling, linear and non-linear PLS and the development of a calibration model based on spectral window selection (SWS), with average and PLS stacking.

4.4.3.1 Global Modelling Approach

Initially global modelling analysis was investigated. In global modelling, all the samples of the process are used for the construction of the models, i.e. the calibration model is developed across the whole duration of the batch and for all wavelengths. For the global modelling approach the following approach was considered:

a) Global models based on linear and non-linear PLS (Neural Network PLS).

b) Global models where the SWS algorithm, with two windows, was applied to select the wavelengths. Linear and non-linear PLS (Neural Networks PLS) were used to develop the calibration model based on the wavelengths selected using the SWS algorithm. Averaging and linear PLS were used for the stacking step. Thirty models were generated and combined.

Figure 4.8 summarises the options considered for global modelling where the calibration model applies to the whole batch.

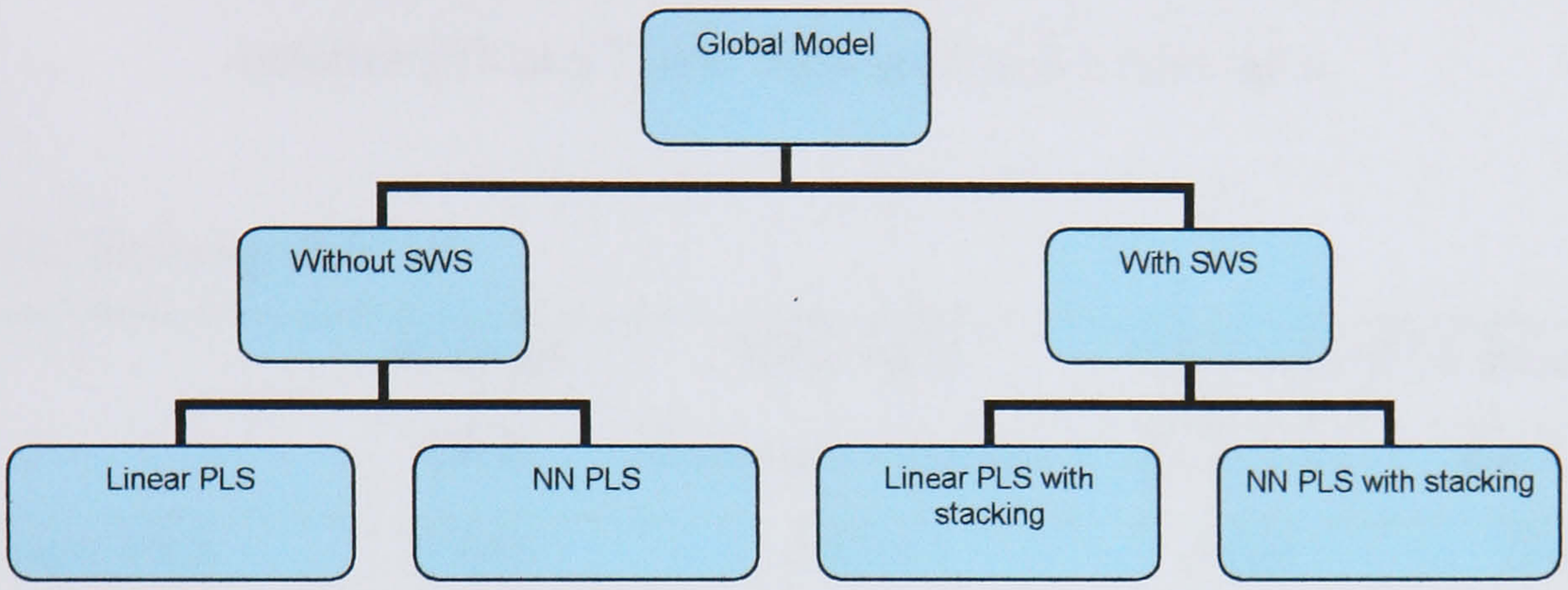


Figure 4.8. Summary of the different methods investigated for the modelling of the standard batches of data set 1 for global modelling.

Table 4.3 presents the results for the modelling of product concentration for global modelling with and without the application of SWS. The RMS error is used as a measure of comparison. Table 4.3 indicates that if a global modelling approach is implemented there are important benefits to be gained from the wavelength selection algorithm. This is apparent from the fit to the validation data (Table 4.3b). The deterioration in the quality of fit between the training and validation data for the model using all the wavelengths is due to the over-fitting of the training data. This problem is avoided when models with SWS and stacking are generated.

The results in Table 4.3 which compare average and PLS stacking demonstrate a further problem related to over-fitting. PLS stacking performed on the training data for 6 LVs heavily weights those models with the best fit and these models are those that will most severely over-fit the data. As a result the performances of the PLS stacking models degrade on the validation data compared with the average stacking Both PLS and neural network based PLS approaches suffer from the same over-fitting problem, with the problem exacerbated in the latter case due to the additional parameters within the neural network. This can be addressed when a smaller number of latent variables is chosen as suggested with the use of cross validation approach. If

a smaller number of latent variables are selected, i.e. 4 variables, the performances of the PLS stacking and average stacking are similar. No additional benefit is gained by adopting the non-linear neural network based PLS over standard linear PLS when the SWS algorithm is applied.

Table 4.3. Results for global modelling of the product concentration for the standard batches (S1 to S7) and NIR spectra for data set 1

Table 4.3a. Training data set

	Without SWS	SWS with Average Stacking	SWS with PLS Stacking	
			for 6 LVs	for 4 LVs
Linear PLS	0.056	0.059	0.045	0.048
Neural Network PLS	0.037	0.061	0.040	0.045

Table 4.3b. Validation data set

	Without SWS	SWS with Average Stacking	SWS with PLS Stacking	
			for 6 LVs	for 4 LVs
Linear PLS	0.092	0.068	0.096	0.067
Neural Network PLS	0.102	0.069	0.103	0.067

In summary it can be concluded that for a global model, SWS with stacking (either Average or PLS) performs better than the PLS based approach. In a further analysis it was decided local modelling to be tested. Local modelling could offer benefits compared to global modelling. This claim will be tested in the next two sections.

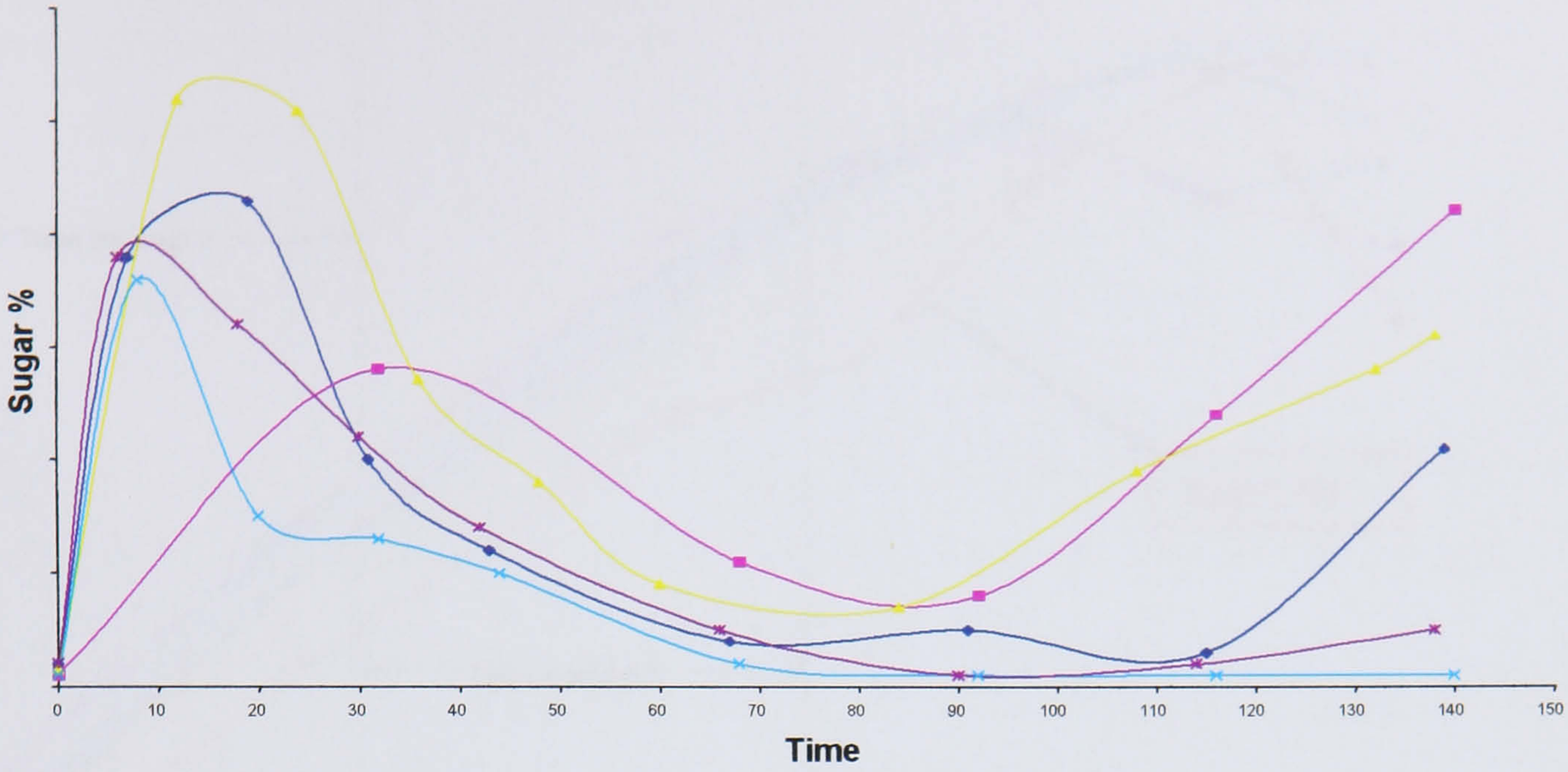
4.4.3.2 Motivation for Local Modelling Approach

Significant changes in broth composition occur over the period of the batch, with spectral interactions between constituents varying considerably as a result. Constructing a calibration model that applies over the whole fermentation can thus be problematic. An alternative approach to potentially improve model robustness is through local modelling where the process is sub-divided, as determined by the process state, and a separate model is built for each region. Arnold *et. al* (2001) have previously proposed the application of local modelling using NIR spectra for a fermentation process and stated that local modelling offers improved performance over a global model. This hypothesis is investigated for the prediction of product concentration and ammonia concentration in the fermentation process. The sub-divisions are identified by considering changes in broth composition.

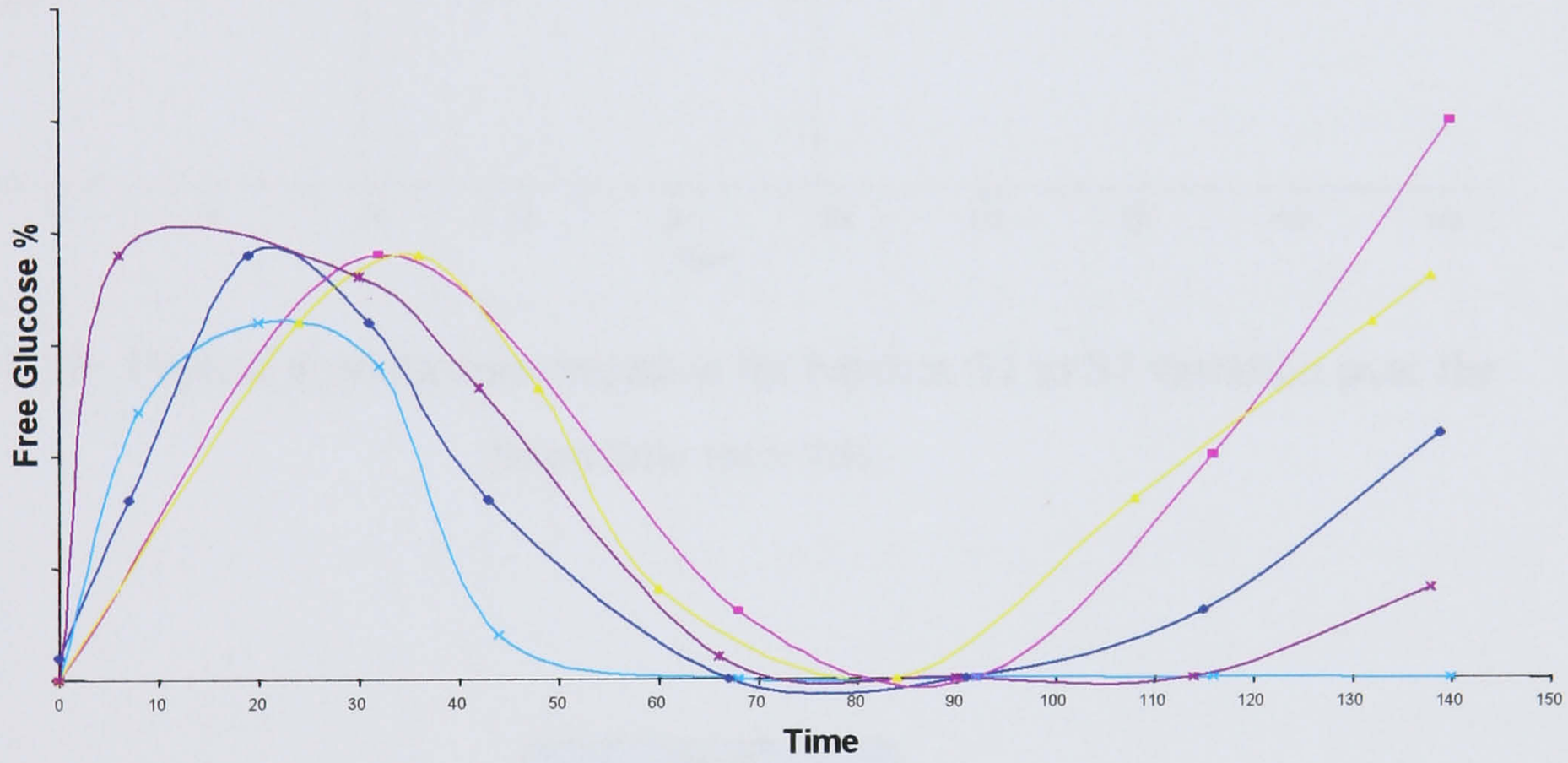
Considering the trends from five typical fermentation batches (as shown in Figure 4.9), it is possible to identify three distinct operating regions. The first region is where sugar and phosphate are present in excess and high growth rates occur. In the second region, sugar falls to limiting levels and high product formation rates are attained. In the third region, sugar again accumulates and phosphate reaches low concentrations as growth and product formation slow down. Using these observations it is possible to identify the distinct operating regions. However the measurements of broth concentrations used to identify the sub-divisions are not available on-line. Using standard operational information, significant changes between batch profiles do not occur and thus time is used as a surrogate variable to switch between models.

Using the information given in Figure 4.9, three sub-divisions were identified. Time interval 1 was chosen to be from the first available sample to 70 hours, Time interval 2 from 70 hours to 100 hours and Time interval 3 from 100 hours to the final observation. Figure 4.10 shows typical variation in product concentration over the seven final stage standard batches. It can be observed that the greatest variation in behaviour occurs in Time interval 3 where for some batches, product accumulation continues and others where formation ceases and drops off.

(a)



(b)



(c)

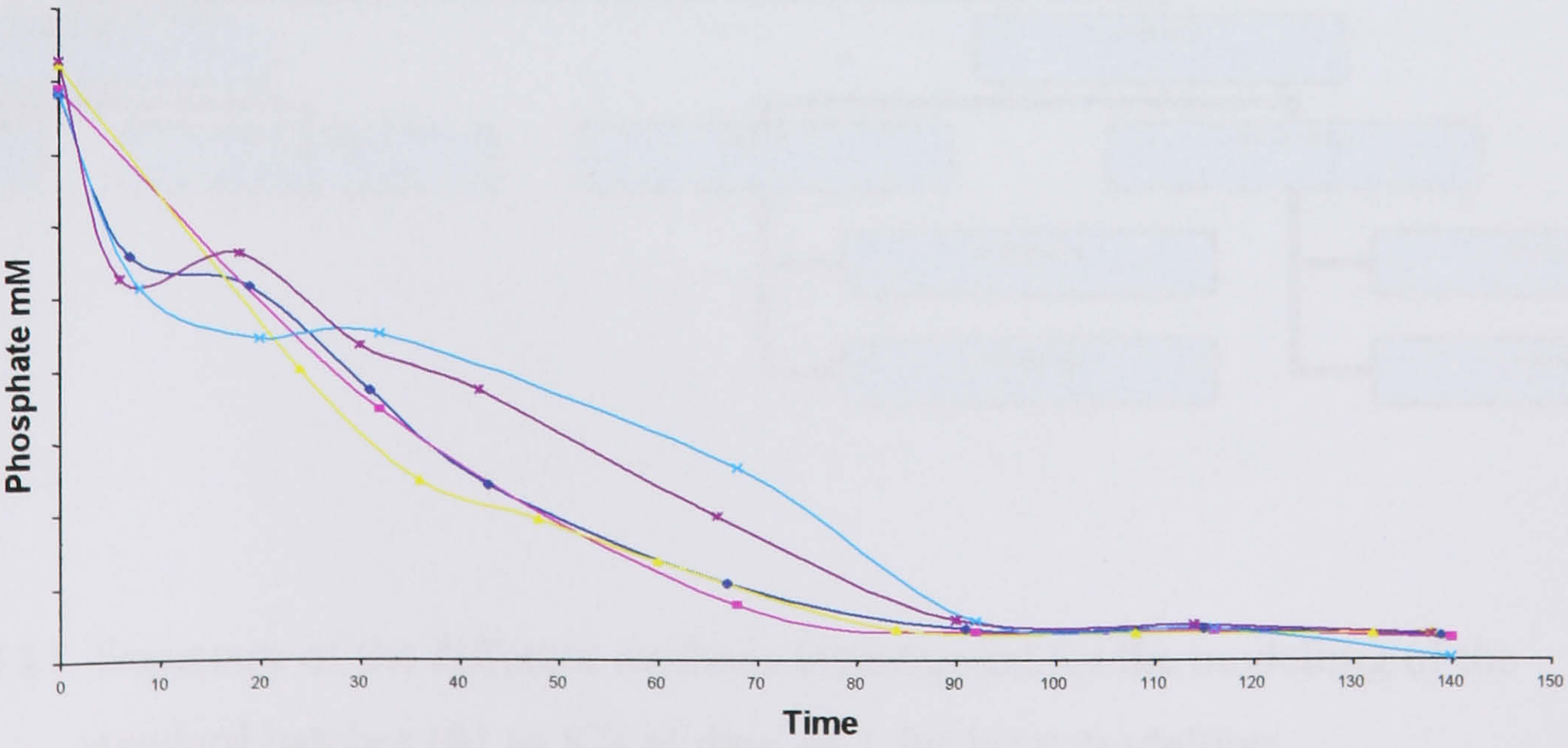


Figure 4.9. Biochemical component concentrations for five batches: (a) sugar, (b) free glucose and (c) phosphate concentration

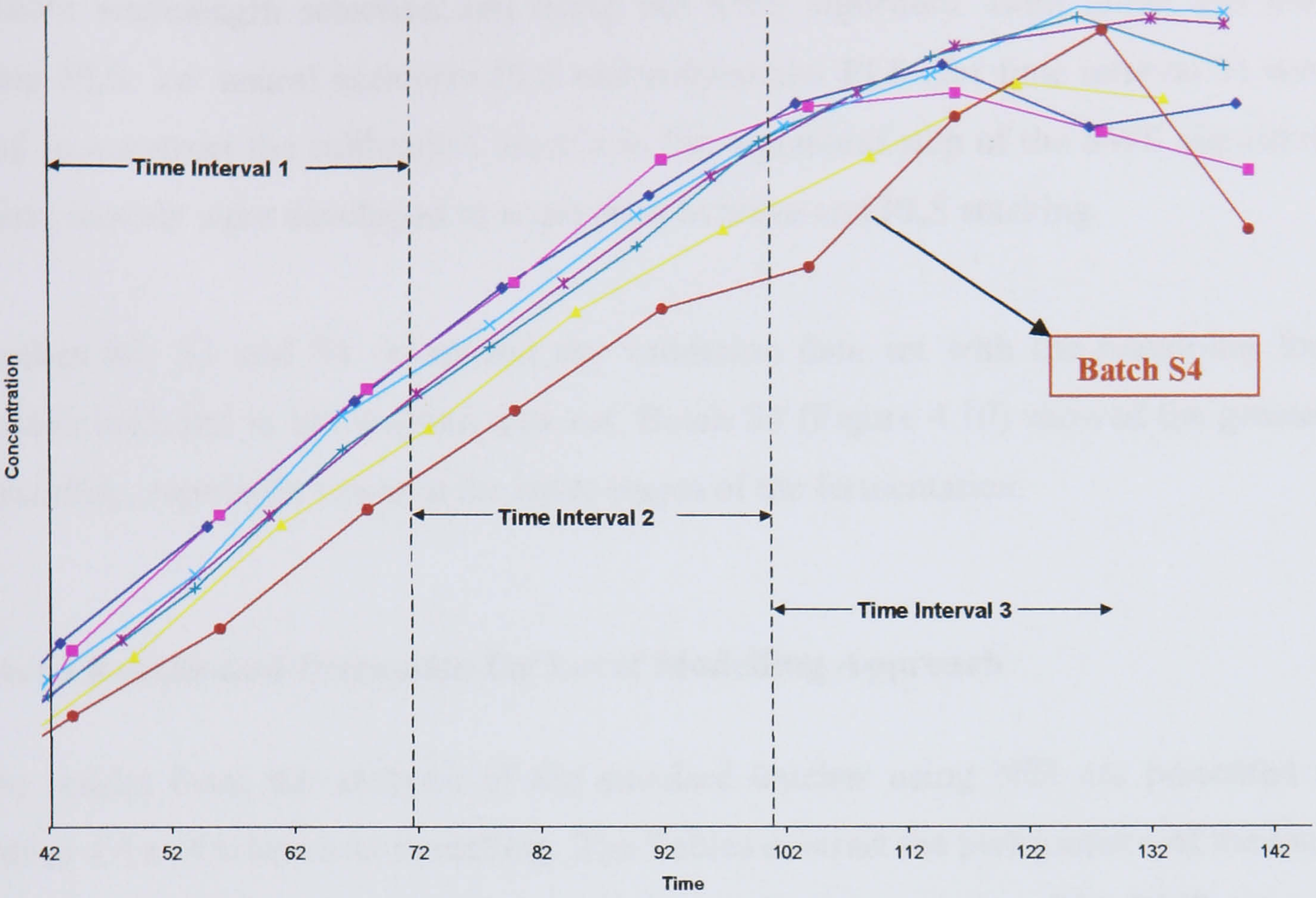


Figure 4.10: Typical product concentration for batches S1 to S7 variation over the three time intervals

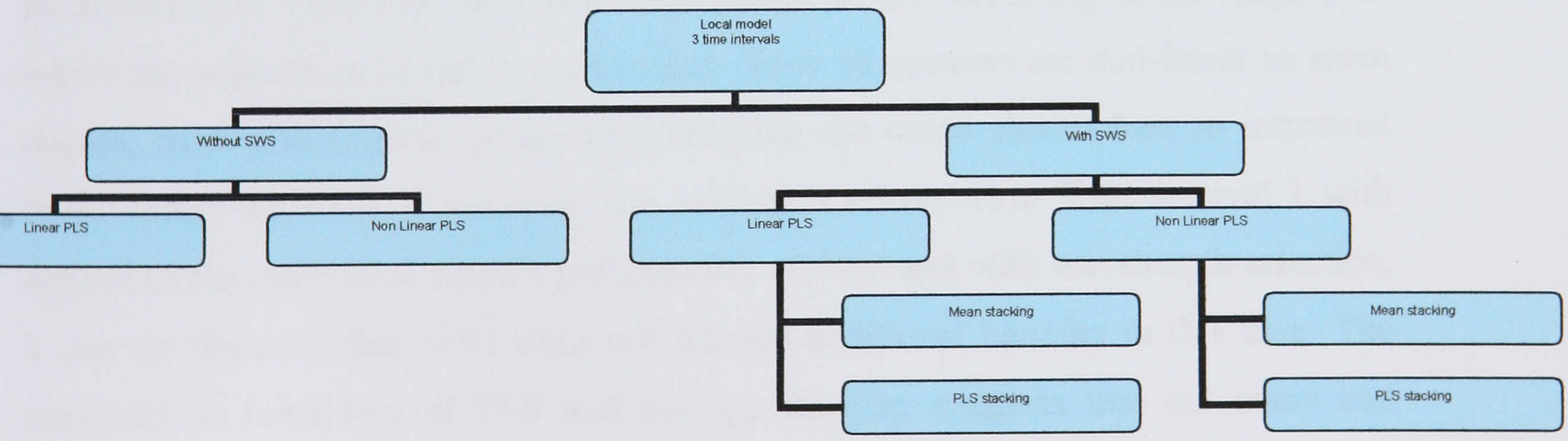


Figure 4.11. Summary of the different methods investigated for the modelling of the standard batches (S1 to S7) of data set 1 for local modelling

The approaches adopted for local modelling to analyse the standard batches is shown schematically in Figure 4.11. Local models for three time intervals were built both without wavelength selection and using the SWS algorithm. Both linear and non-linear PLS: i.e. neural networks PLS and polynomial PLS (for time interval 3) were used to construct the calibration models in the regression step of the SWS algorithm. Thirty models were developed to implement average and PLS stacking.

Batches S2, S3 and S4 comprised the validation data set with the remaining four batches included in the training data set. Batch S4 (Figure 4.10) showed the greatest variability, especially towards the latter stages of the fermentation.

4.4.3.3 Results and Discussion for Local Modelling Approach

The results from the analysis of the standard batches using NIR are presented in Tables 4.4 to 4.6 for local modelling. The Tables contrast the performance of the local models constructed both with and without wavelength selection. The RMS error is used as a means of comparison.

First, from Table 4.4 can be seen that model accuracy is significantly improved in Time Interval 1 compared to the global model results in Table 4.3. The enhanced performance in Time Interval 1 is a consequence of the narrowing of the range over which the calibration model is constructed. Since all systems are non-linear to some degree, from a theoretical perspective reducing the range should lead to improved linear model accuracy. Contrasting the validation results from Time Interval 1 with respect to the calibration models performance without and with wavelength selection, it can be observed that SWS does not provide additional benefits in this case. The similarity in behaviour of PLS and average stacking suggests that not many bad models are produced to weight negatively the outcome of the average stacking.

Table 4.4. Results for the local modelling for the training data set of the product concentration for the standard batches (S1 to S7) and NIR spectra for Time Interval 1

Table 4.4a. – Training data set - Time Interval 1

	No wavelength selection	SWS	
		PLS Stacking (for 6 LVs)	Average Stacking
Linear PLS	0.025	0.018	0.034
NN PLS	0.025	0.018	0.031

Table 4.4b. – Validation data set - Time Interval 1

	No wavelength selection	SWS	
		PLS Stacking (for 6 LVs)	Average Stacking
Linear PLS	0.042	0.045	0.049
NN PLS	0.043	0.045	0.046

In Time Interval 2, the local model results (Table 4.5) are comparable to the global model when SWS is not used for the training data set. In contrast to Time Interval 1, the validation data indicates that there is improvement in performance in Time Interval 2, by using SWS and stacking. Possible explanation for this is that towards the later stages of the process there is much greater variation between batches and more significant changes in media concentrations occur compared to Time Interval 1.

Table 4.5.Results for the local modelling for the training data set of the product concentration for the standard batches (S1 to S7) and NIR spectra for Time Interval 2

Table 4.5a. – Training data set - Time Interval 2

	No wavelength selection	SWS	
		PLS Stacking (for 8 LVs)	Average Stacking
Linear PLS	0.042	0.025	0.034
NN PLS	0.041	0.027	0.033

Table 4.5b. – Validation data set - Time Interval 2

	No wavelength selection	SWS	
		PLS Stacking (for 8 LVs)	Average Stacking
Linear PLS	0.059	0.058	0.048
NN PLS	0.060	0.051	0.053

For the third time interval, there are benefits to be gained from adopting a polynomial approach. Table 4.6b indicates that for the validation data set both polynomial PLS and neural network PLS with average stacking give a good fit to the validation data set, offering an improvement over linear PLS. In these regions, fermentation changes are significant thus a non-linear approach is appropriate. These benefits must be balanced against the fact that the increase in the number of parameters to be estimated, gives greater opportunity for over-fitting and the time taken for model parameterisation is significantly greater.

Table 4.6.Results for the local modelling for the training data set of the product concentration for the standard batches (S1 to S7) and NIR spectra for Time Interval 3

Table 4.6a. – Training data set - Time Interval 3

	No wavelength selection	SWS	
		PLS Stacking (for 6 LVs)	Average Stacking
Linear PLS	0.038	0.039	0.058
NN PLS	0.037	0.036	0.048
Poly PLS	-	0.027	0.041

Table 4.6b. - Validation data set -Time Interval 3

	No wavelength selection	SWS	
		PLS Stacking (for 8 LVs)	Average Stacking
Linear PLS	0.084	0.095	0.060
NN PLS	0.093	0.079	0.060
Poly PLS	-	0.075	0.043

In summary for the local modelling approach:

- Overall it was shown that for the case considered, (a) for time intervals 2 and 3: the application of SWS together with average stacking is the best approach for calibration model construction, (b) for time intervals 1: the application of SWS together with average stacking provided comparable results.
- Local modelling provided better results than global modelling.
- There are some improvements to be gained by adopting a non-linear SWS approach but the time taken for model parameterisation is significantly greater.

As an example of typical model behaviour of those models considered in Table 4.5, Figure 4.12 shows the results from Time Interval 1 for the training and validation batches after the application of SWS and PLS Stacking. Multiple batches are concatenated in the figure and the large decreases are the breaks between batches. It can be observed in the residual plot (Figure 4.12d) that an off-set exists which is most significant for the second validation batch. The issue of offset removal is discussed in Chapter 5.

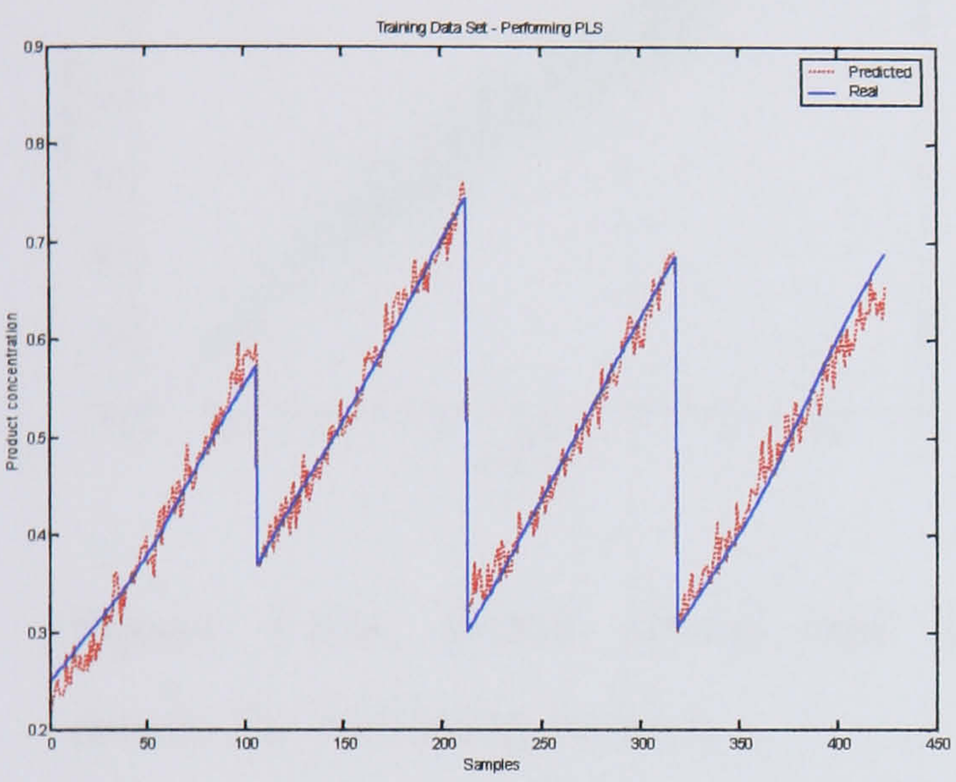


Figure 4.12a. Training data results for 4 batches.

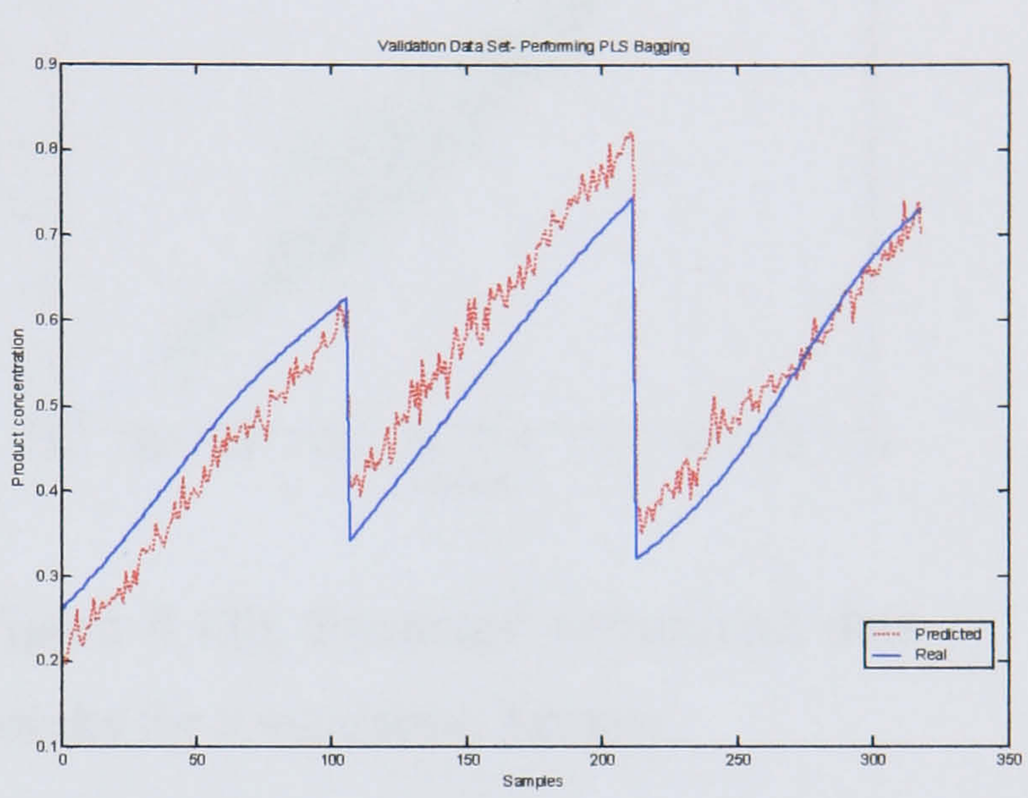


Figure 4.12b. Validation data results for 3 batches.

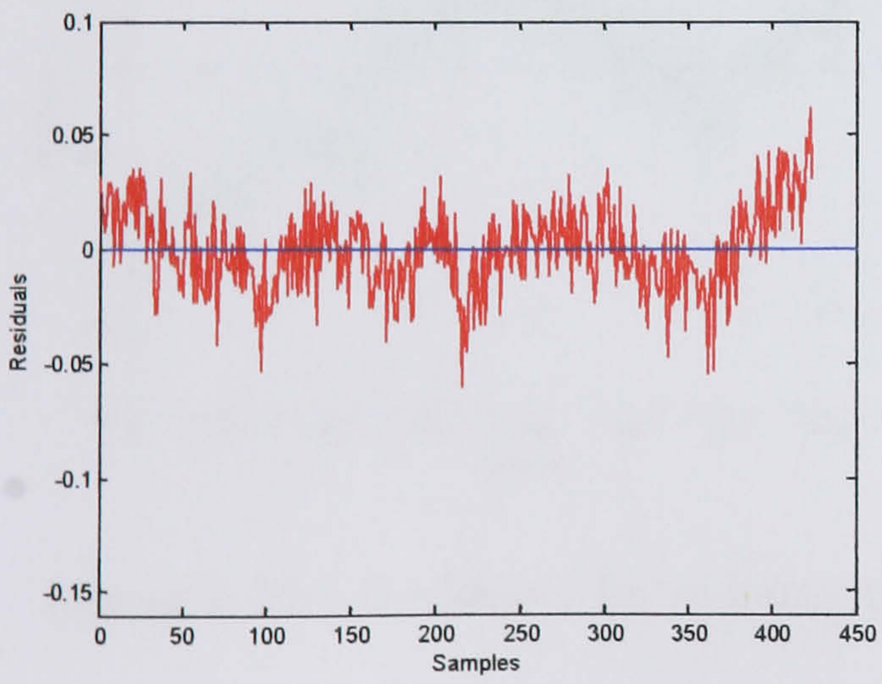


Figure 4.12c. Residuals for training data.

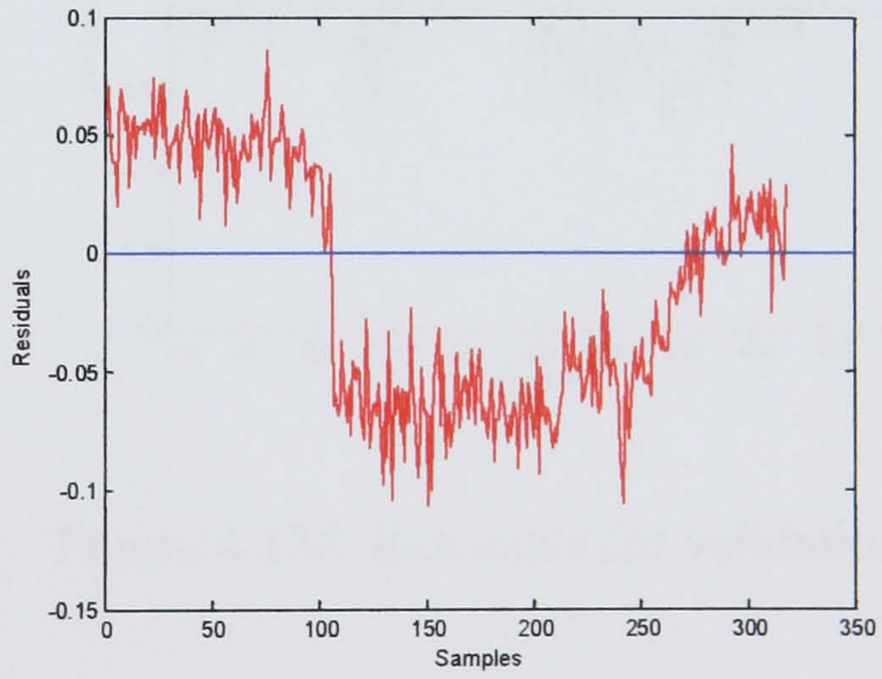


Figure 4.12d. Residuals for validation data.

Figure 4.12. First time interval: Results for the standard batches for the modelling of product concentration with average stacking.

Moreover for the behaviour of the models considered in Table 4.5, Figure 4.13 shows the results from Time Interval 2 for the training and validation batches after the application of SWS and Average Stacking. These plots also provide an indication of the magnitude of the RMS error (Figure 4.13c and 4.13d) and observed behaviour.

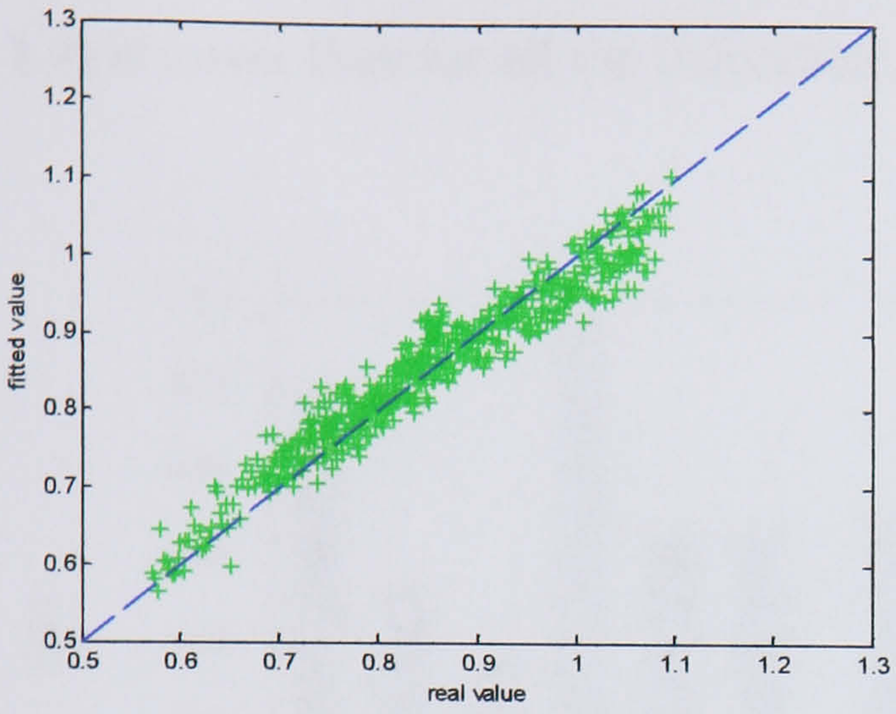


Figure 4.13a. Fitted versus real data results for 4 training batches.

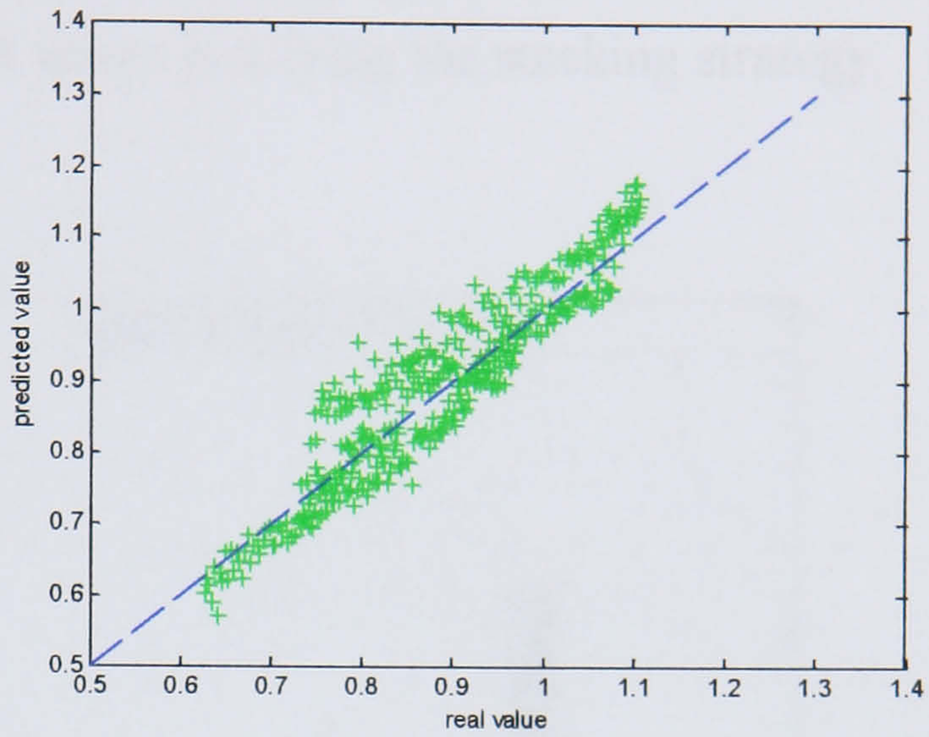


Figure 4.13b. Predicted versus real data results for 3 validation batches.

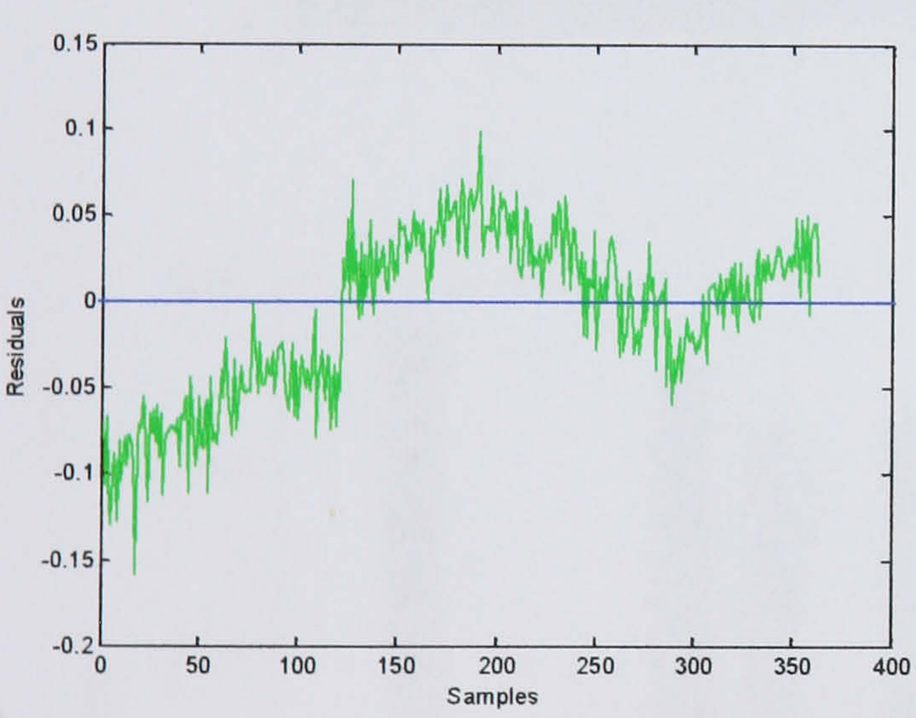


Figure 4.13c. Residuals for training data.

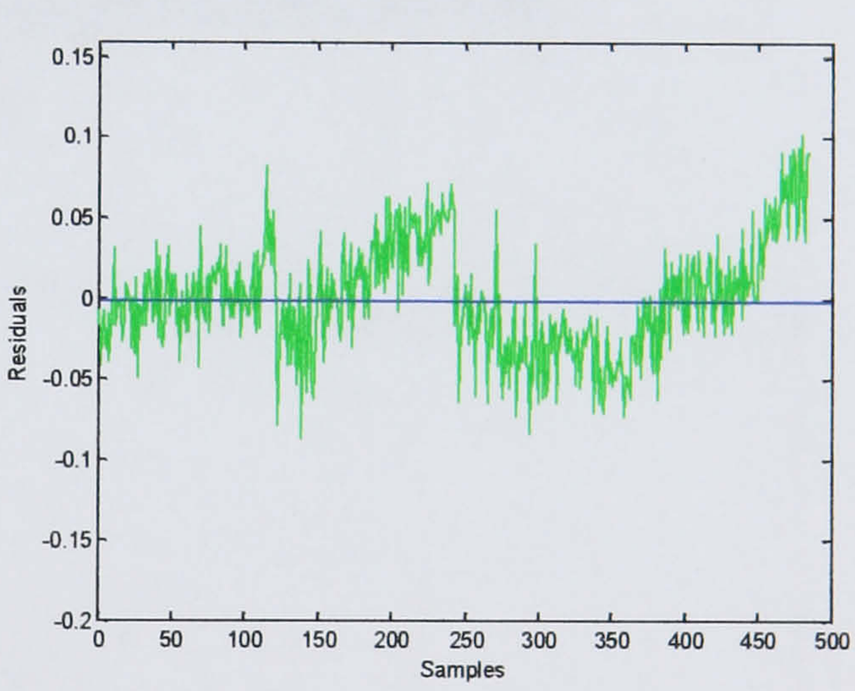


Figure 4.13d. Residuals for validation data.

Figure 4.13. Second time interval: Results for the standard batches for the modelling of product concentration with average stacking.

The importance of stacking can be observed in Figure 4.14 where the results of the thirty individual model errors are presented for Time interval 1. Considering the training data, models 13, 22 or 26 would be selected as the best individual models (Figure 4.14 top). In these cases, the models also perform well in validation. It is also

possible to obtain good training performance and poor validation (e.g. model 18). Quite varied performance between models is also possible, for example the RMS error for training for models 1, 7 and 29 is comparatively large (Figure 4.14 top) but the RMS error of the validation data set for models 1, 7 and 29 is small (Figure 4.14 bottom). Most notably, the RMS error for the PLS stacked model (presented in Table 4.4) is lower than for all the individual model errors justifying the stacking strategy.

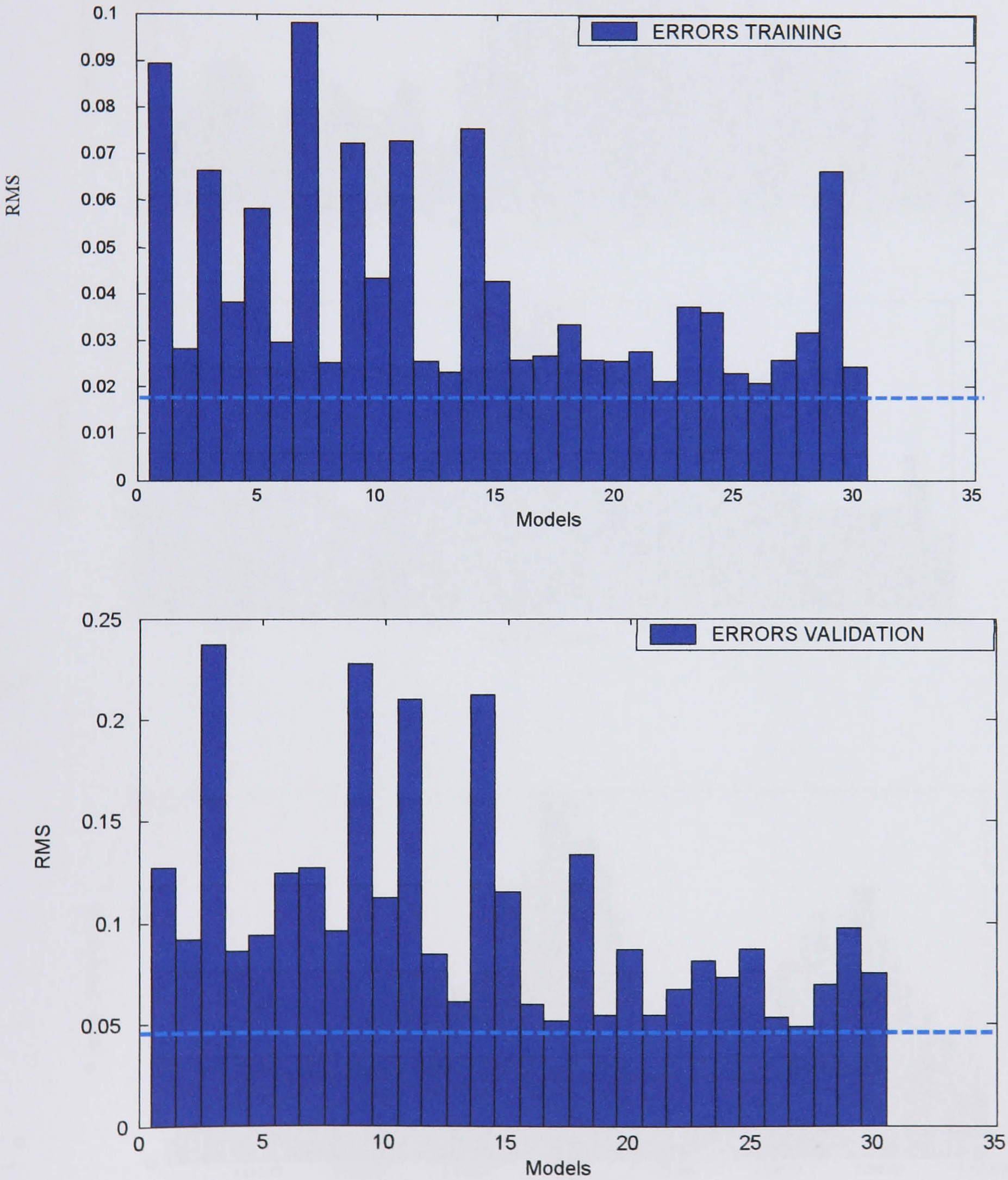
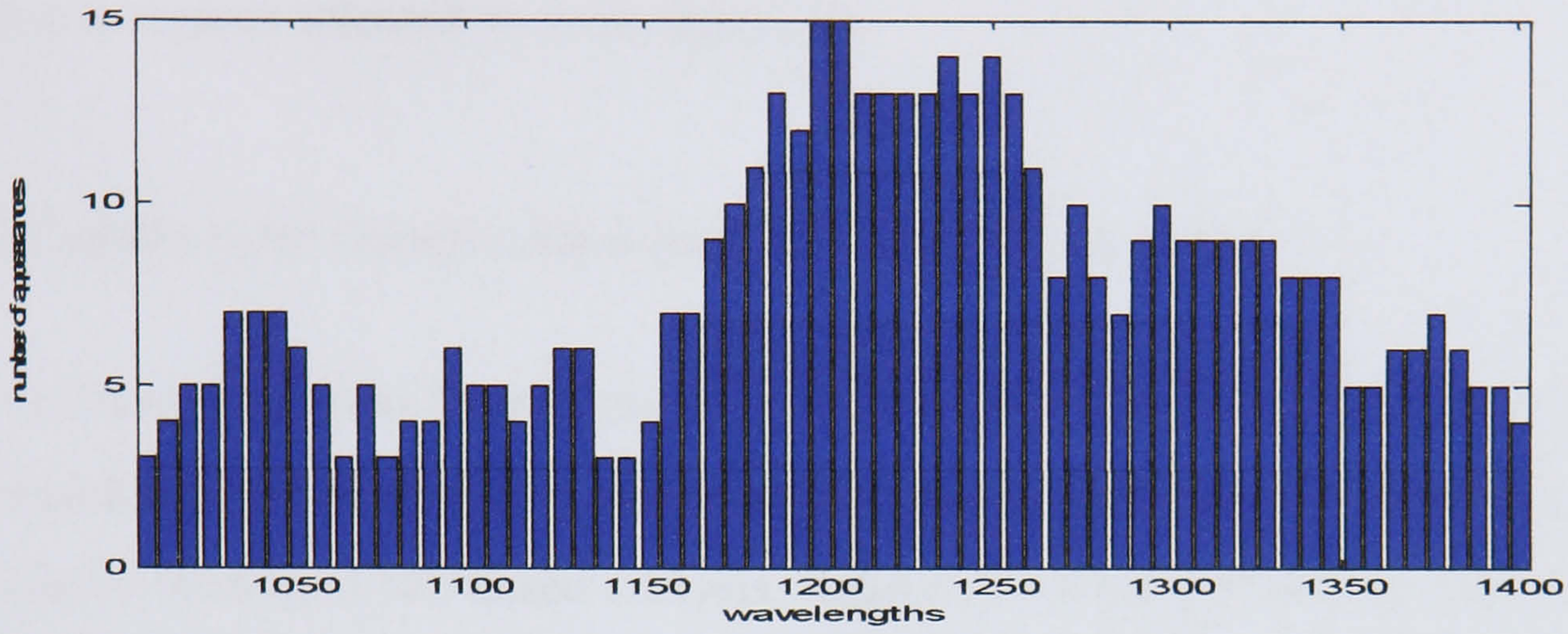
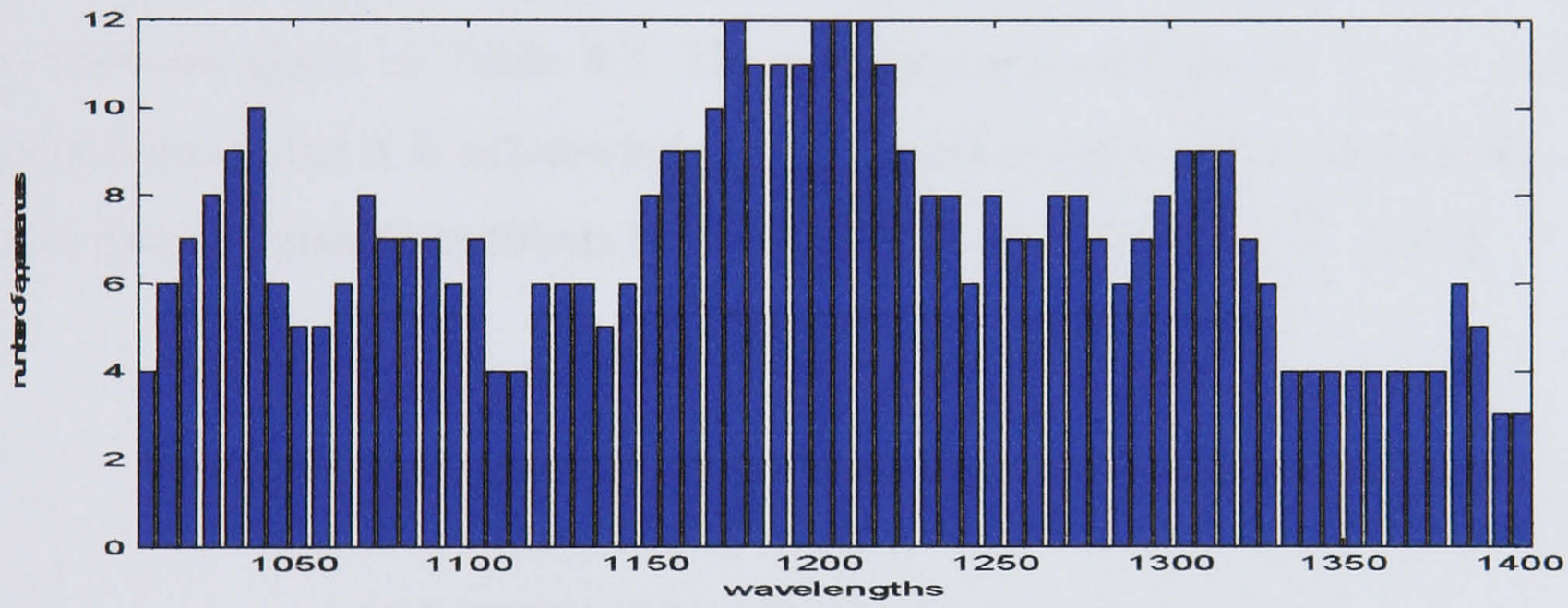


Figure 4.14. Errors for the 30 models for the first time interval for the standard batches (S1 to S7) for the experimental and the validation data set. ----- RMS error after PLS Stacking

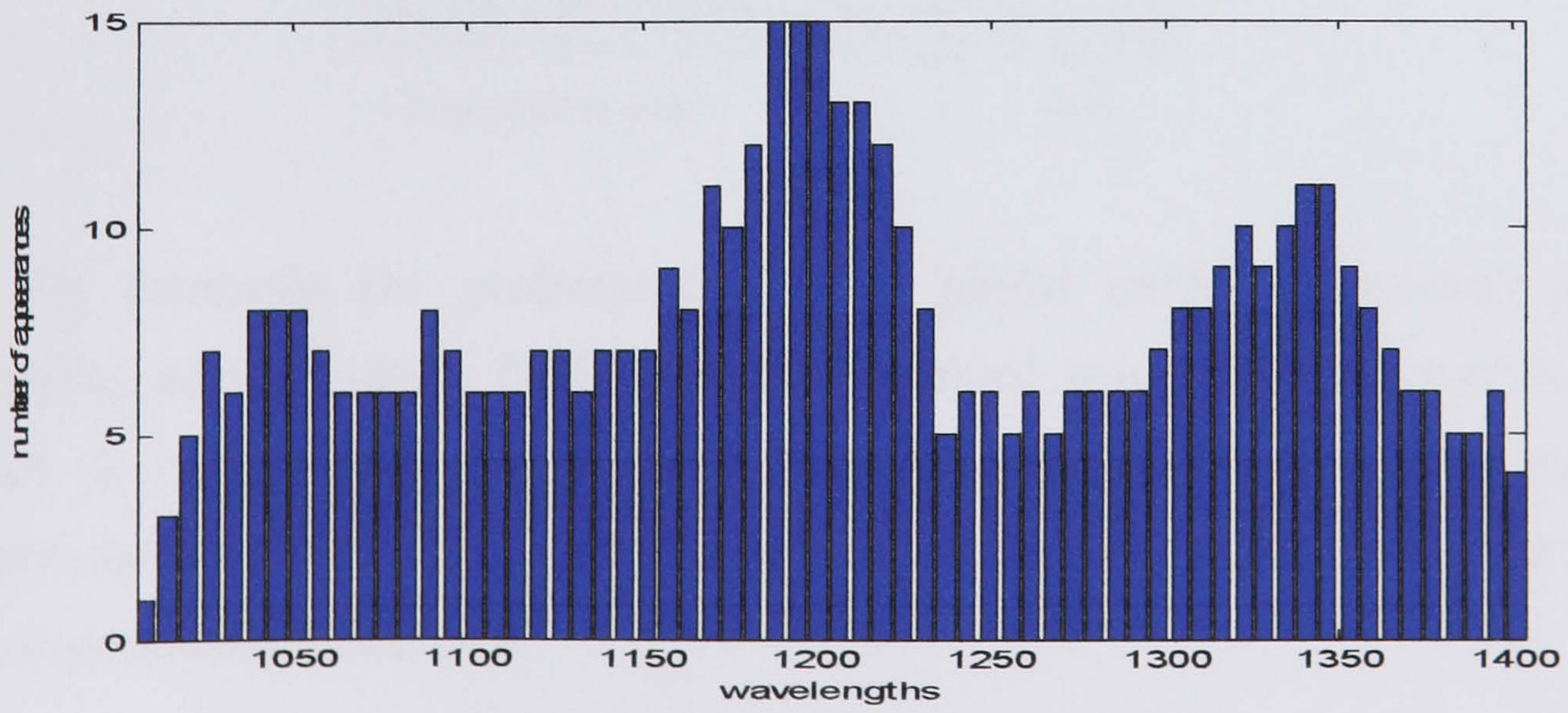
Figure 4.15 shows the frequency distribution of the wavelengths selected by the SWS algorithm for each of the three time intervals. Wavelengths in the region 1188.2 to 1250 *nm* were selected the most frequently. This range of wavelengths aligns closely with those identified by the analytical chemistry process specialists for the product under consideration.



(a)



(b)



(c)

Figure 4.15. Frequency of wavelength selection by SWS for batches S1 to S7: (a) for time interval 1, (b) for time interval 2, (c) for time interval 3.

4.4.4 Comparison of SWS with other Wavelength Selection Methods

In this subsection, traditional wavelength selection methods i.e. genetic algorithm and interval PLS (iPLS) are compared against the new wavelength selection algorithm, SWS. The comparison is preformed in terms of prediction error (RMS) and wavelength regions selected by each algorithm.

4.4.4.1 Results from Genetic Algorithm Wavelength Selection

In this section, the results from applying a genetic algorithm for wavelength selection are presented. The RMS error was used as the fitness function of the GA in order to be consistent with the SWS based analysis. Reproduction was performed with a single point crossover with probability 0.7 followed by mutation. The remainder of the GA parameters are given in Table 4.7. These parameters were found to give acceptable GA performance but it is acknowledged that selecting optimal parameters for a GA is difficult due to interaction effects (Massart et al., 1997, Zeaiter et al., 2005).

Table 4.7. Values for GA

Number of individuals	100
Number of generations	100
Number of variables used	65
Bit representation	1
Generation gap	0.5

Table 4.8 compares the performance of the global model constructed using wavelengths selected using SWS with that obtained using GAs for wavelength selection. It can be seen that SWS offers comparable results with GAs on the validation data set, but on the other hand one major drawback of the GAs is that they were computationally intensive.

Table 4.8. Results for global modelling of the product concentration for batches S1 to S7 contrasting SWS with GAs

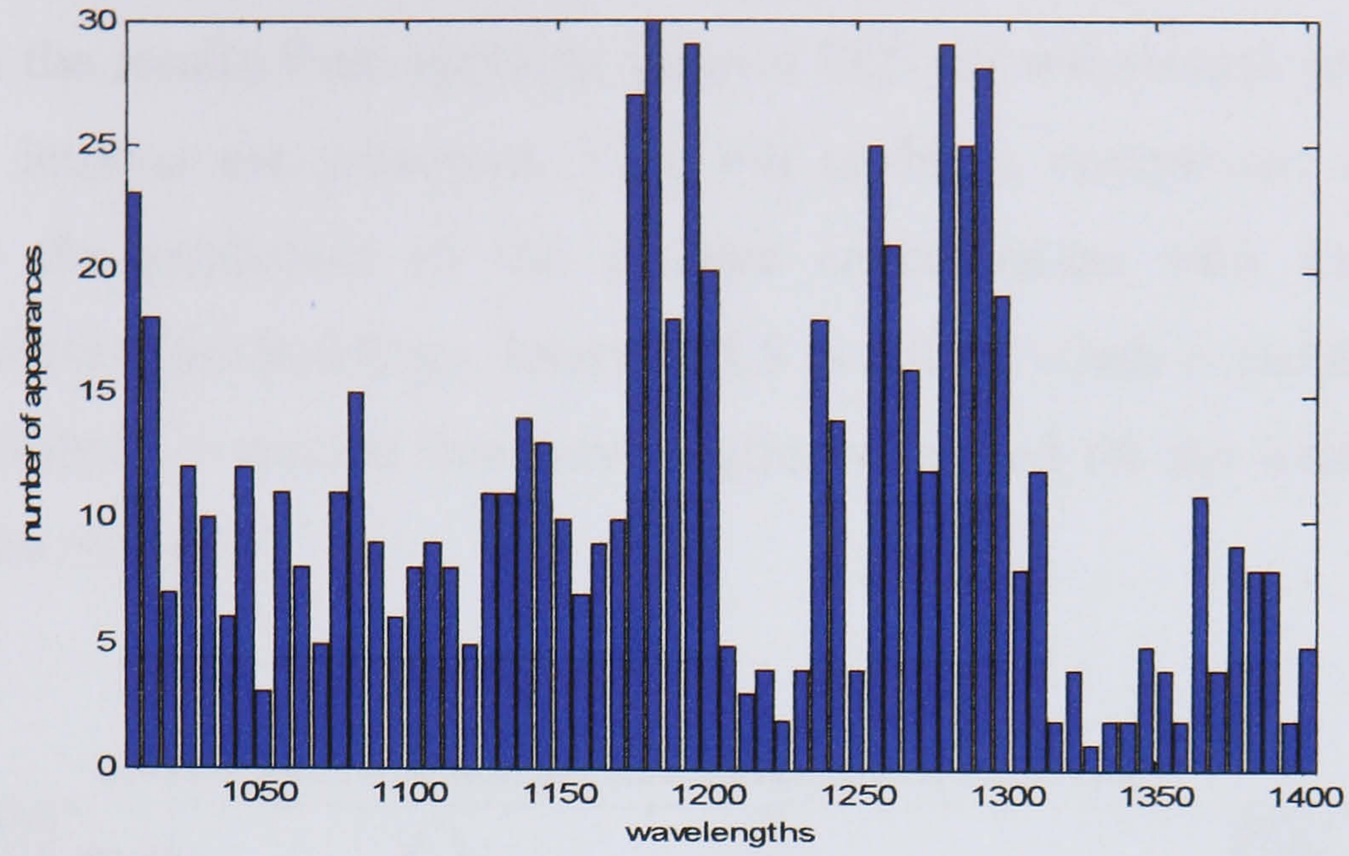
	Average Stacking		PLS Stacking	
	GAs	SWS	GAs	SWS
Training data set	0.047	0.059	0.044	0.048
Validation data set	0.066	0.068	0.085	0.067

Table 4.9 summarises the performance of the local models constructed following the application of SWS and GAs to the validation data set. SWS gave a better model than GA selection in all but the first time interval where the results were comparable. Moreover, SWS with Average stacking produced better results than PLS without wavelength selection in time intervals 2 and 3.

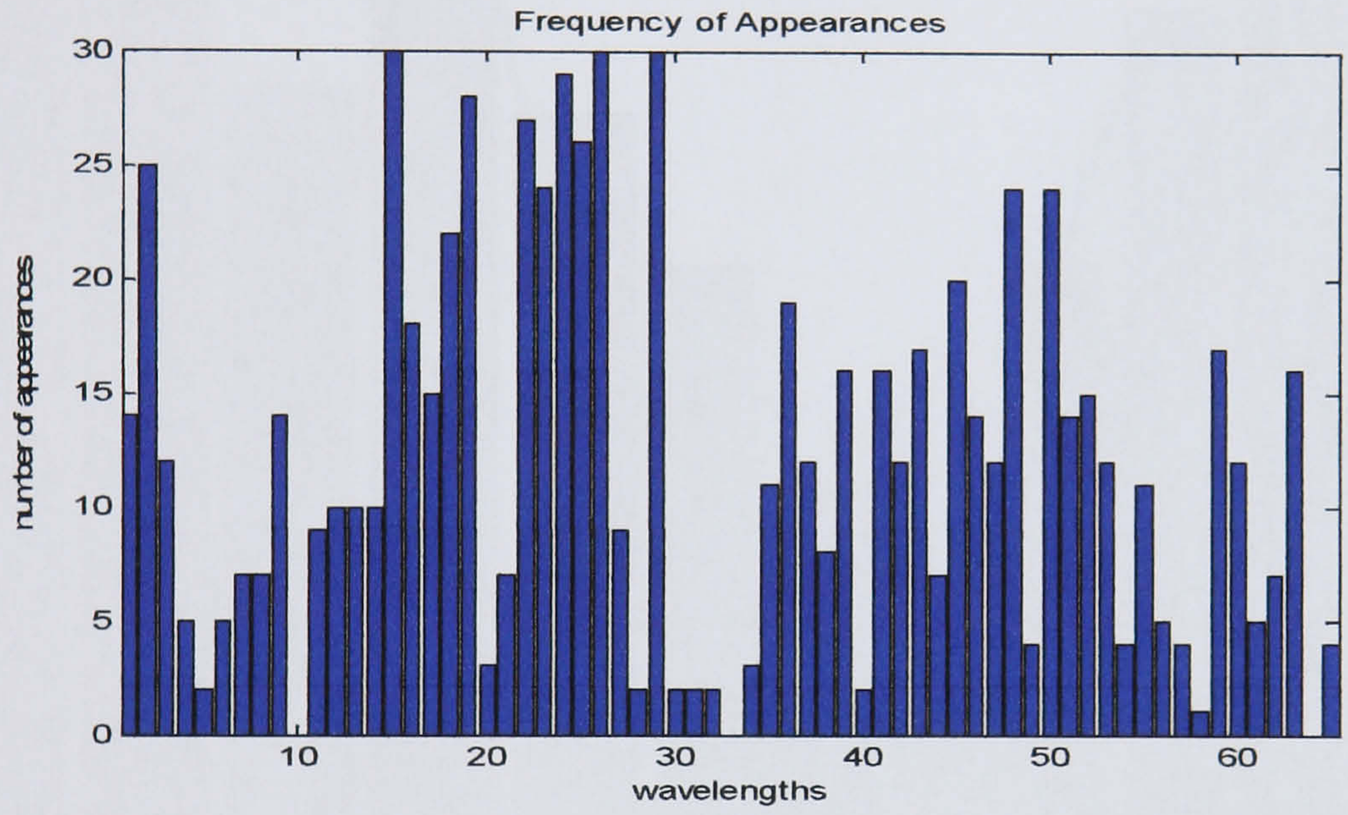
Table 4.9. RMS for the product concentration from the NIR spectra for the validation data set contrasting SWS with GAs

	Time Interval 1		Time Interval 2		Time Interval 3	
	PLS Stack	Average Stack	PLS Stack	Average Stack	PLS Stack	Average Stack
SWS with Linear PLS	0.045	0.049	0.058	0.048	0.095	0.060
GAs with Linear PLS	0.045	0.043	0.067	0.069	0.177	0.139
PLS (without wavelength selection)	0.042		0.059		0.084	

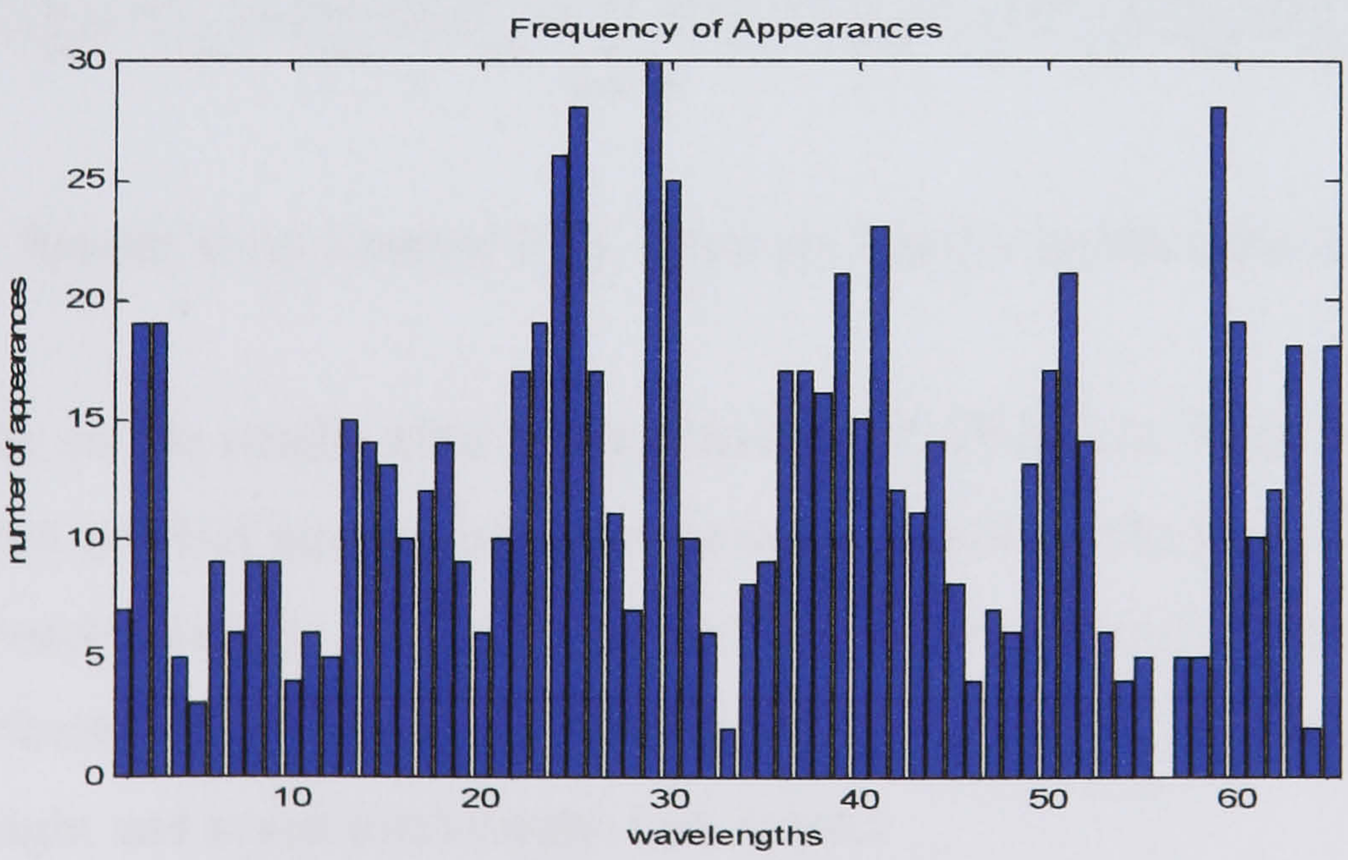
Figure 4.16 shows the frequency distribution of the wavelengths selected by the GA selection for the three time intervals, respectively. Compared with Figure 4.15 and the SWS results, the GA did not indicate any critical regions for the three time intervals and the important wavelengths, as identified by the analytical chemist, were not selected preferentially. The major issue is that genetic algorithms provide the possibility of selecting individual wavelengths which allow the models to become too specific to the training data and thus the models do not predict validation samples well.



(a)



(b)



(c)

Figure 4.16. Frequency distribution of the wavelengths selected by GA for batches S1 to S7: (a) for time interval 1, (b) for time interval 2 and (c) for time interval 3.

4.4.4.2 Results from Interval PLS Wavelength Selection

In this section the results from applying interval PLS for wavelength selection from the first time interval are presented. This will enable a comparison of the SWS algorithm for the prediction of the product concentration with an alternative wavelength selection methodology. Interval PLS is a fixed window technique. In this example 12 windows comprise five wavelengths were used for the training data set for the first time interval.

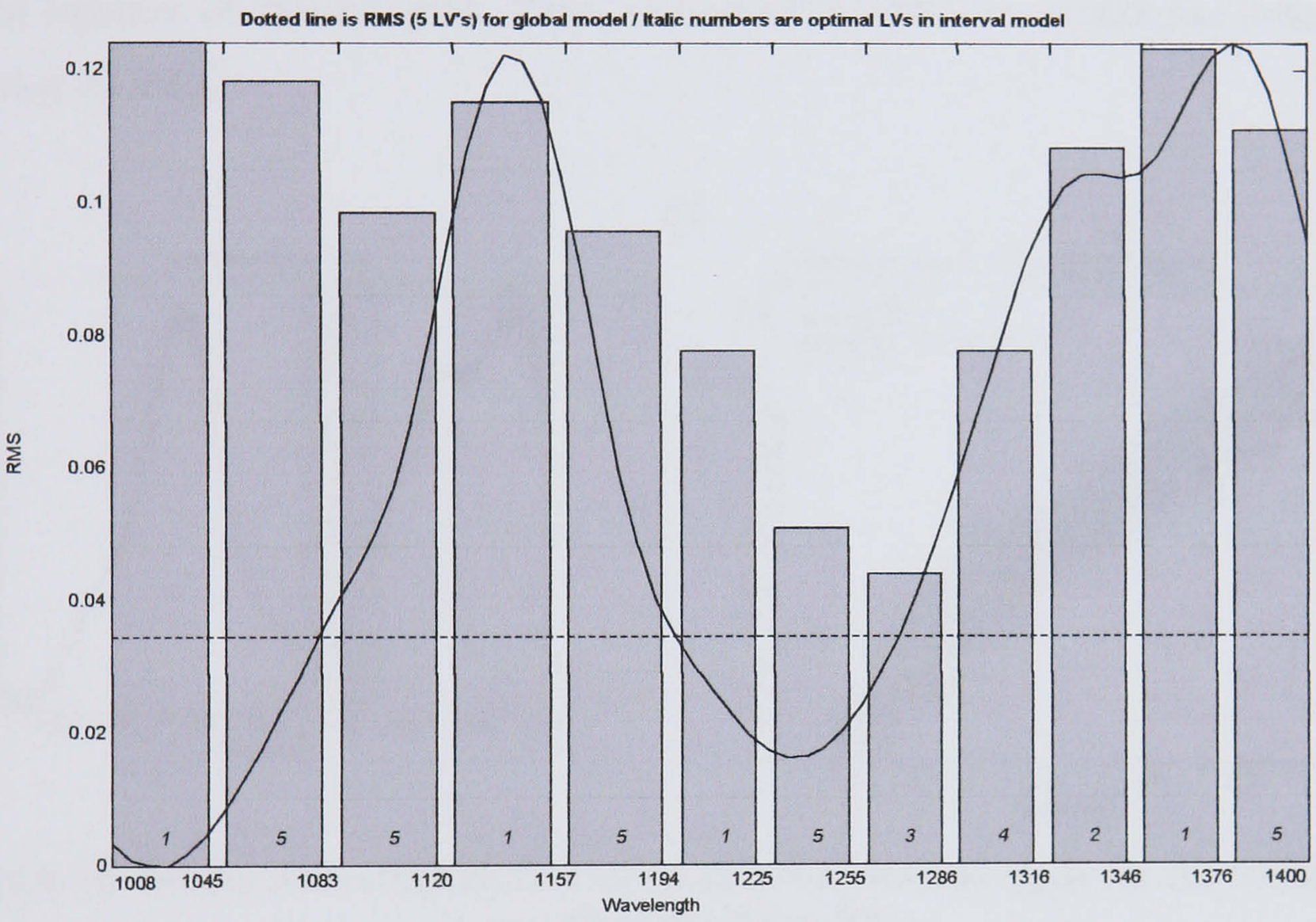


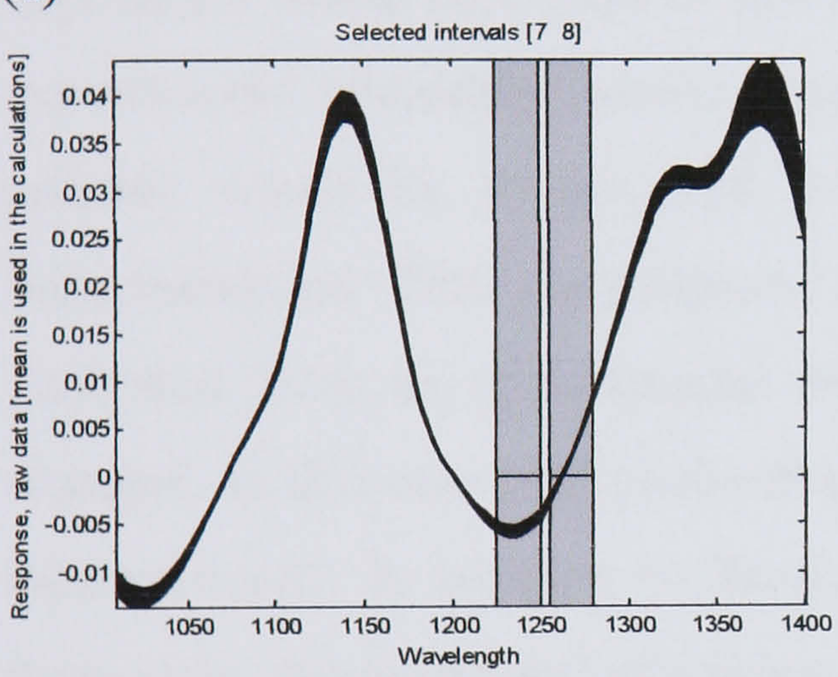
Figure 4.17. Results from Interval PLS. Intervals 7 and 8 produce the lowest error.

Figure 4.17 shows the results after the application of iPLS. The italic value on each bar indicates the optimal number of latent variables based on the minimisation of the RMS error in each interval. The visual part of the spectrum which is systematic does not contain information about the real extract. From Figure 4.17, it can be concluded that window eight and seven produce the best results.

Thus, a model is calculated using the wavelengths captured by windows seven and eight (Figure 4.18a and 4.18b). This model validation results can be seen in Figure 4.19a and 4.19b. The results are not good, the expected interval was not chosen and hence the results from the SWS algorithm were better.

The results of iPLS for the selected intervals are summarised in Table 4.10. iPLS is a methodology that calculates local PLS models on fixed sub-intervals of the full spectral region and its use includes the identification of important or varying spectral regions and the possibility of developing a good spectral local PLS model built on a limited number of wavelengths. Thus, compared to SWS, no additional benefits are provided by iPLS.

(a)



(b)

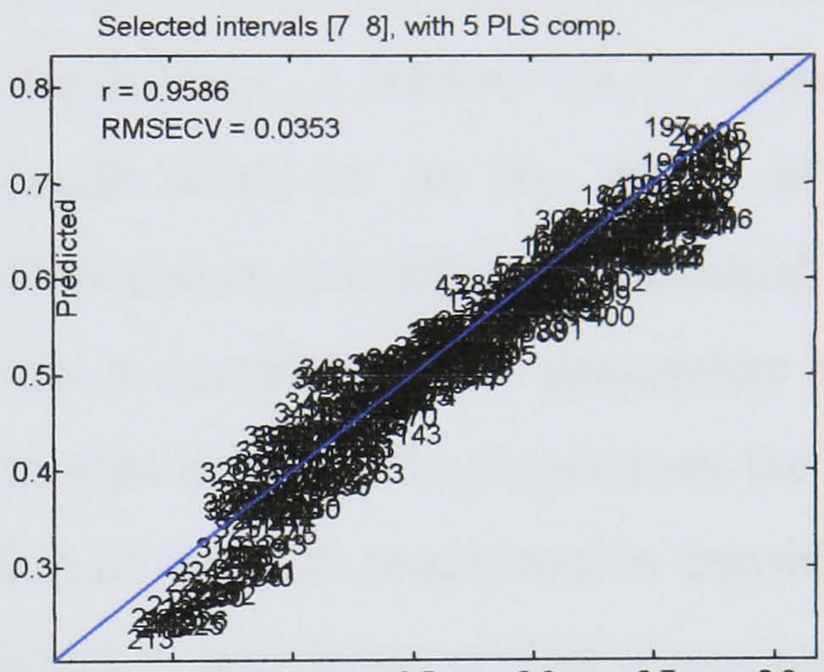
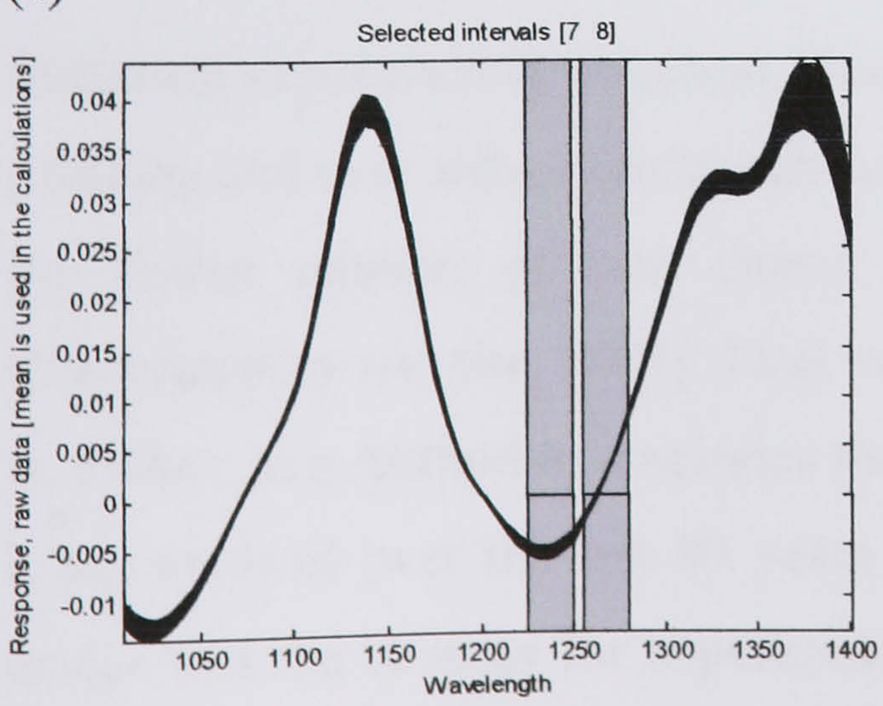


Figure 4.18. Model with combination of windows seven and eight for the training data set of batches S1 to S7.

(a)



(b)

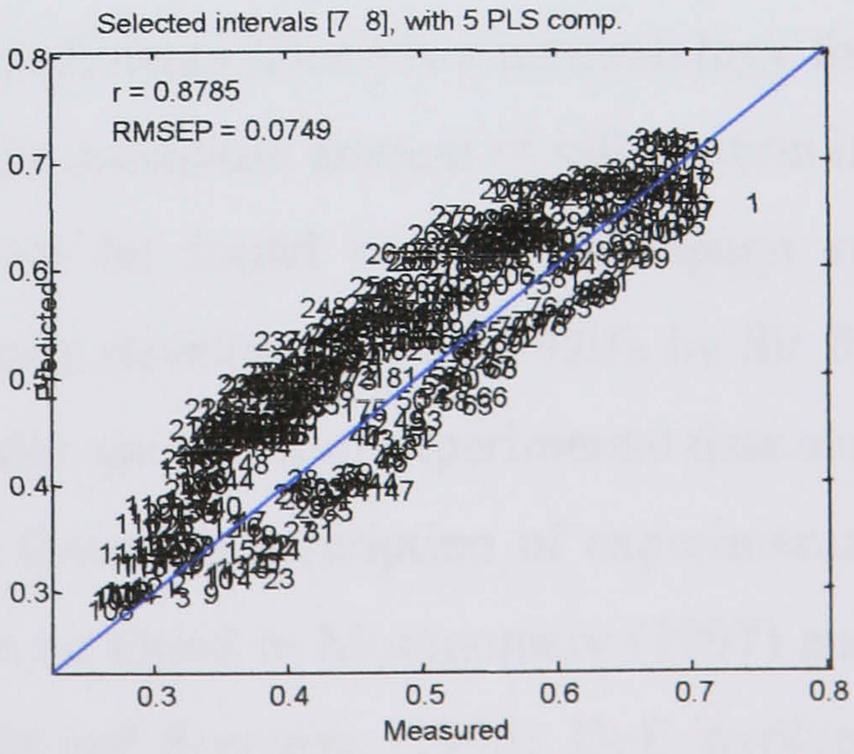


Figure 4.19. Model with combination of windows seven and eight for the validation data set of batches S1 to S7.

Table 4.10. iPLS results for combinations of intervals

	Just interval 8	Intervals 7 and 8
Training	0.044	0.035
Validation	0.085	0.074

4.4.5 Experimental Design Batch Analysis from Data Set 2 (Zeiss and Linx 5-10)

The previous results were obtained using standard fermentation operating policy with the variation in behaviour being a consequence of natural variability. Such an application would be typical of that required for the full-scale production plant. In the experimental laboratory where development takes place, a broader set of operating policies would be investigated, placing greater demands on the utility of the instrumentation. Thus the proposed methodologies were also applied to spectral data generated from an experimental design where a number of key parameters were changed. In this case MIR instrumentation was also available to complement the NIR measurements. In addition to the determination of product concentration considered previously, the prediction of ammonia concentration, which is a critical component in the fermentation process, was also investigated.

4.4.5.1 Design of Experiments

Statistical experimental design or Design of Experiments (DoE) is a methodology for planning and executing experiments to extract the maximum amount of information in the fewest number of runs (more details can be found on the Homepage of Chemometrics website, 2005). DoE was originally developed in the 1920's by Sir R. A. Fisher, as a method to maximize the knowledge gained from experimental data and it has evolved over the last 80 years. A more thorough description of experimental design and the strategy for experimentation can be found in Montgomery (1997) and three papers have also been published by Araujo and Brereton, (1996). DoE involves specifying a set of experiments that are likely to be the most informative with regard

to a specific issue (e.g. maximisation of product), consequently the strategy is problem dependent. A common approach in DoE is to define a standard reference experiment (centre-point) and then perform new, representative experiments about this point in a symmetric manner. Most experimentation involves studying the impact of the change of several process variables (factors) to optimise processes and/or investigate and understand the relationship between factors and characteristics of the process responses of interest (Figure 4.20).

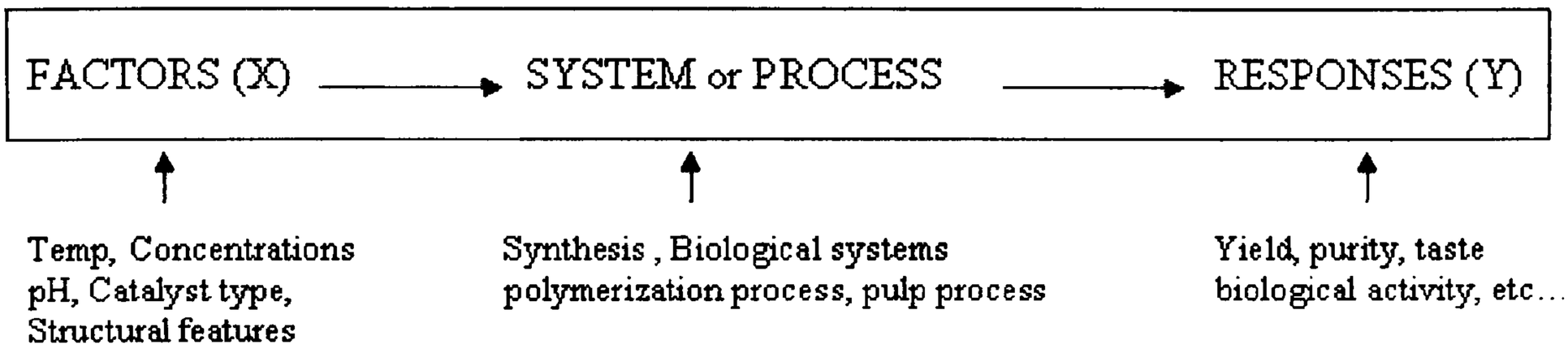


Figure 4.20. Relationship between the factors and the responses.

With DoE, the factors that have an influence on the response are identified along with the goal of achieving optimal conditions. Instead of varying one factor at the time, the different factors are varied simultaneously over a set of experimental runs. The most common approach is that of a two-level factorial design at two levels due to its simplicity with regard to both preparation and analysis of the results. All input factors are set at two levels. These levels are termed 'high' and 'low' or '+1' and '-1', respectively. A design with all possible high/low combinations of all the input factors is termed a full factorial design at two levels. If there are k factors, each at 2 levels, a full factorial design has 2^k runs. The factorial design is extendable to three or more levels.

Typical examples where DoE is informative include the development of new products and processes or the optimisation of existing manufacturing processes such as those producing chemicals, polymers, drugs, pharmaceuticals and foods and in the development of fermentation processes. The benefits of applying DoE in a fermentation development environment were outlined by Sircar *et al.* (1998). They applied a full factorial design to study the interaction of each of the components of the media and the optimisation of their composition in batch cultures of *Streptomyces*

clavuligerus. Wang *et al.* (2005) also optimised the media composition of *Streptomyces clavuligerus* by using a fractional factorial DoE and by screening a large number of experimental factors. From their research, they concluded that the systematic methods had the advantage of identifying the most significant media components and their optimal levels in a rabid and economic way.

4.4.5.2 Application of the DOE Analysis to the Antibiotic Process

A number of critical process variables were identified by the process engineers and an experimental design was conducted. For the experimental design batches, a partial factorial design was performed (Table 4.11). The factorial design was used to investigate the interaction of the environmental conditions (pH and temperature) and feed rates (sugar feed and oil feed) on the product concentration behaviour.

Table 4.11. Fermentation variations in the DoE study (L - Low value; H – High value)

Batch	Temperature	pH	Sugar	Oil	Analytical Method
E1	L	L	H	H	NIR
E2	H	H	H	H	NIR/MIR
E3	L	H	H	L	NIR/MIR
E4	H	H	L	L	NIR/MIR
E5	H	L	L	H	NIR/MIR
E6	L	L	L	L	NIR/MIR
E7	L	H	L	H	MIR
E8	L	L	H	H	MIR

Major operating changes were implemented and as a result significant variations in batches were observed (Figure 4.21). In the standard operating policy experiments it was possible to identify three distinct time intervals where significant changes in operating policy were introduced. In this case, it was not possible to identify common time intervals and thus local modelling was not feasible. An alternative approach would be to adopt a fermentation state by which to undertake the determination of

intervals but this was not a practical approach from an application perspective as the key state measurements are only available by off-line assay. Hence in the DoE based fermentation study only a global calibration model was considered.

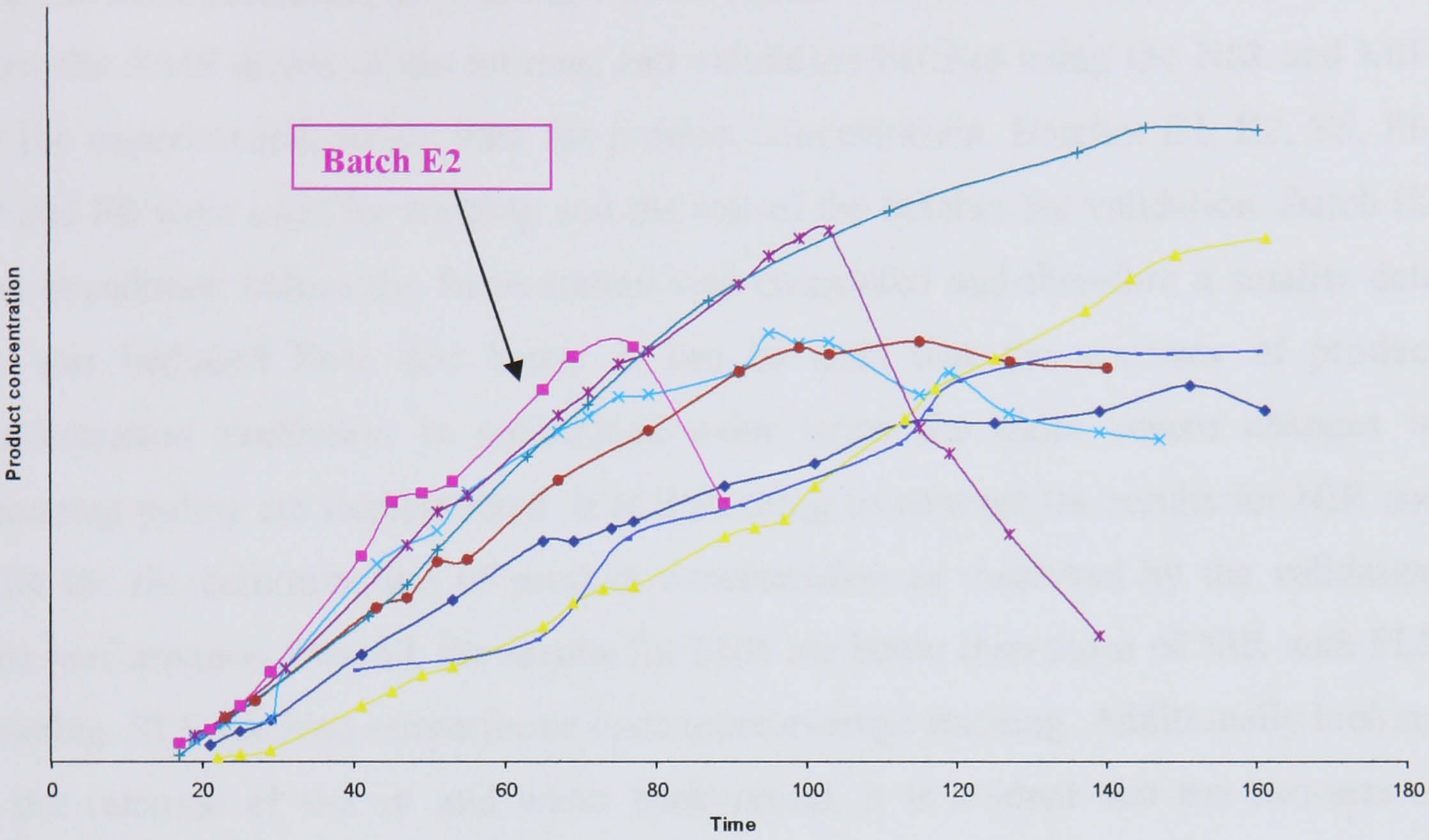


Figure 4.21. Product concentration for the DoE batches E1 to E8.

The considered calibration modelling approaches were:

a) Global models using linear and non-linear PLS.
b) Global models using the wavelengths selected by the SWS algorithm. Thirty models were developed and averaging and PLS stacking was also applied.

The approaches were applied to both the NIR and MIR measurements and in the case of MIR both water and air background removal was considered. Ideally, NIR and MIR would have been applied in all cases but instrument availability limited the experiments to which they could be applied.

4.4.5.3 Results of Experimental Design Batch Analysis

For the experimental design batches, the results from product and ammonia concentration prediction are summarised in Tables 4.12 to 4.15. Tables 4.12 and 4.13 show the RMS errors of the training and validation batches using the NIR and MIR for the experimental design data for product concentration. Batches E1, E2, E5, E6, E7 and E8 were used for training and the rest of the batches for validation. Batch E2 was abandoned before the fermentation was completed and therefore a smaller data set was included from this batch. It can be seen that the accuracy of product concentration prediction is maintained even when the more severe changes in operating policy are implemented. It is interesting to contrast the results for NIR and MIR for the determination of product concentration as measured by the validation data performance. Overall, the results for MIR are better than those of NIR with PLS stacking. PLS stacking outperforms once more average stacking. Additionally looking at the removal of the air and water background, it is evident that the two-sets of results are comparable for MIR data set analysis when determining product concentration.

Table 4.12. Results for global modelling for the DOE batches E1 to E6 using NIR spectra for the product concentration

	Without SWS	SWS	
		Average Stacking	PLS Stacking with 6 LVs
Training	0.034	0.093	0.055
Validation	0.290	0.092	0.067

Table 4.13. Results for the global modelling of product concentration using DoE batches E2 to E8 and MIR data

Table 4.13a. Training data set

	Without SWS	With SWS	
	Linear PLS	PLS Stacking with 6 LVs	Average Stacking
Subtract Water Background	0.067	0.058	0.089
Subtract Air Background	0.067	0.057	0.086

Table 4.13b. Validation data set

	Without SWS	With SWS	
	Linear PLS	PLS Stacking with 6 LVs	Average Stacking
Subtract Water Background	0.082	0.059	0.117
Subtract Air Background	0.082	0.057	0.110

The results from the prediction of ammonia concentration using MIR are given in Table 4.15 and can be contrasted with the NIR predictions from Table 4.14. Considering the MIR results, again there is little difference between the results for water and air background removal.

Table 4.14 Results for global modelling for the DOE batches E1 to E6 using NIR spectra for ammonia

	Without SWS	SWS	
		Average Stacking	PLS Stacking with 8 LVs
Training	0.080	0.043	0.051
Validation	0.248	0.077	0.060

Table 4.15. Results for the global modelling of ammonia using DoE batches E2 to E8 and MIR data

Table 4.15a. Training data set

	Without SWS	SWS	
	Linear PLS	PLS Stacking with 8 LVs	Average Stacking
Subtract Water Background	0.054	0.048	0.069
Subtract Air Background	0.054	0.048	0.058

Table 4.15b. Validation data set

	Without SWS	SWS	
	Linear PLS	PLS Stacking with 8 LVs	Average Stacking
Subtract Water Background	0.078	0.077	0.068
Subtract Air Background	0.078	0.073	0.066

Figure 4.22 shows the results for the prediction of ammonia concentration using MIR for the training and validation data set. Whilst training accuracy is high, there is a small offset in the prediction for one of the validation batches. The results from the NIR determination of ammonia concentration presented in Table 4.14 demonstrate marginally better performance than the MIR analysis, with average stacking proving to be most effective. Finally, the MIR results demonstrate that it is advantageous to adopt the SWS strategy, confirming the findings of the NIR data analysis.

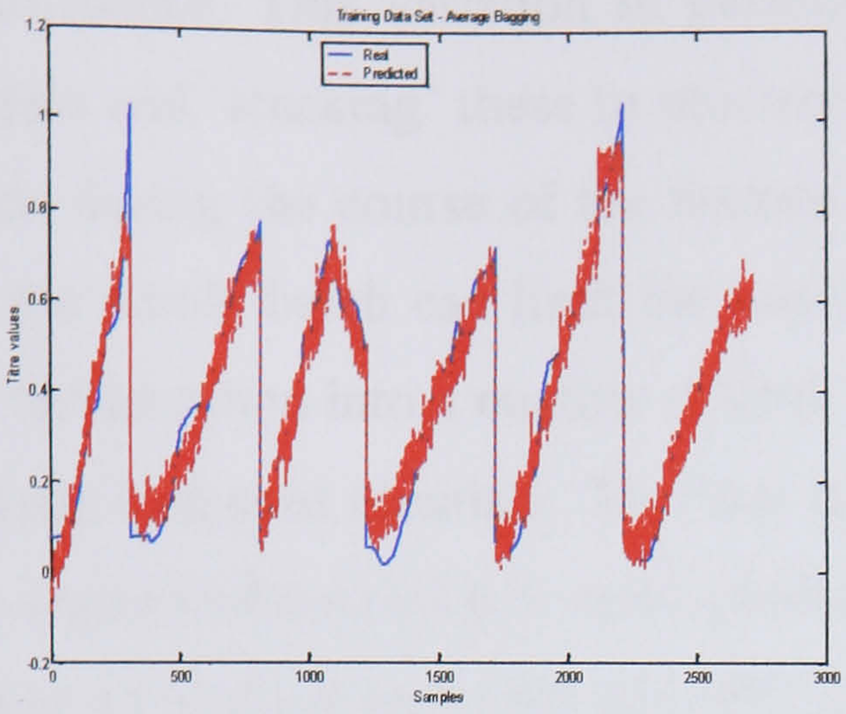


Figure 4.22a. Training data results for 6 batches.

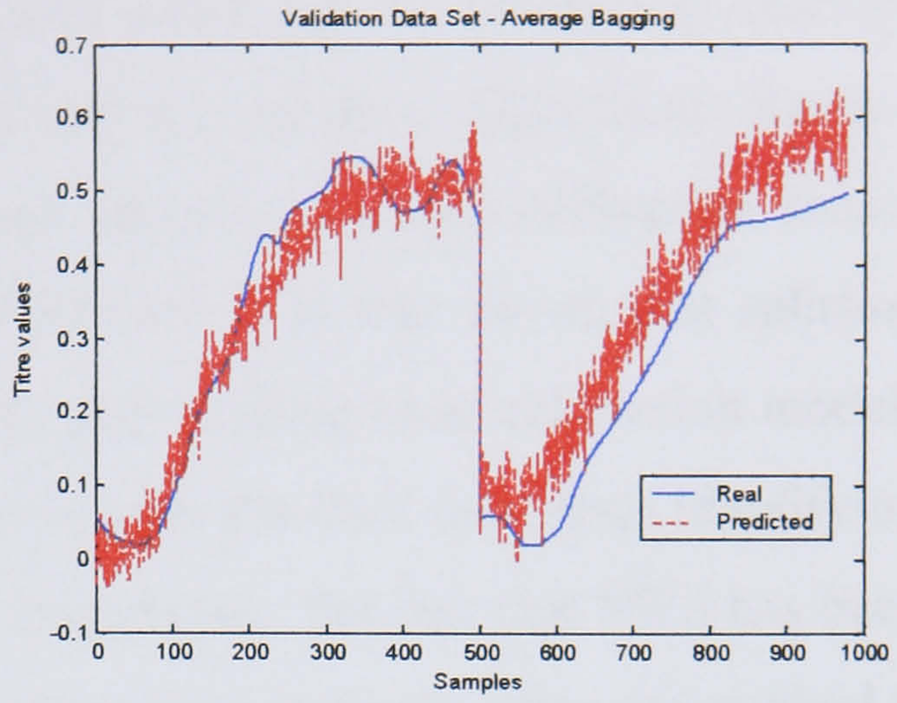


Figure 4.22b. Validation data results for 2 batches.

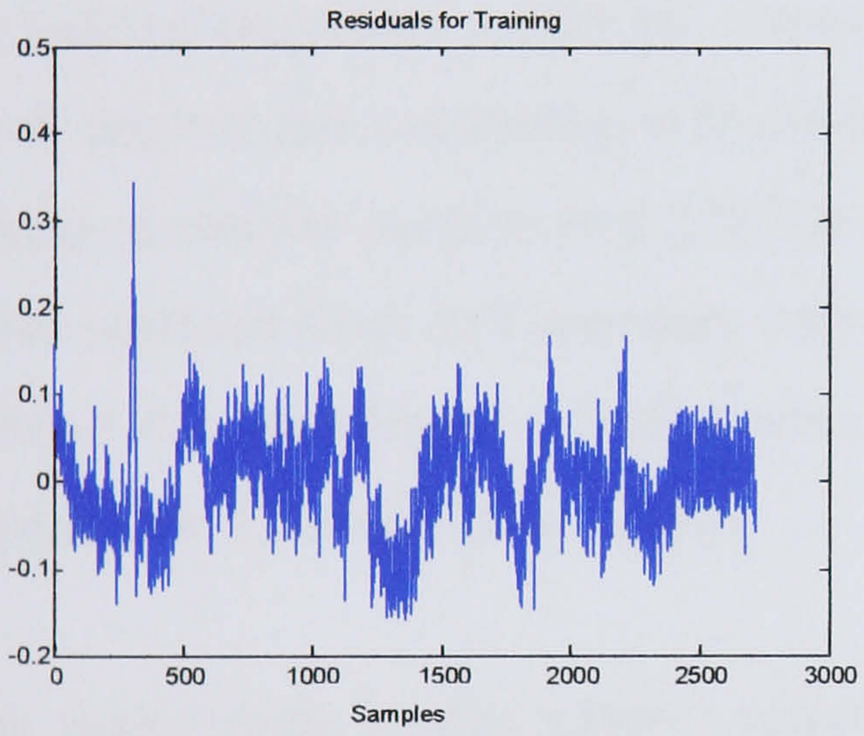


Figure 4.22c. Residuals for training data.

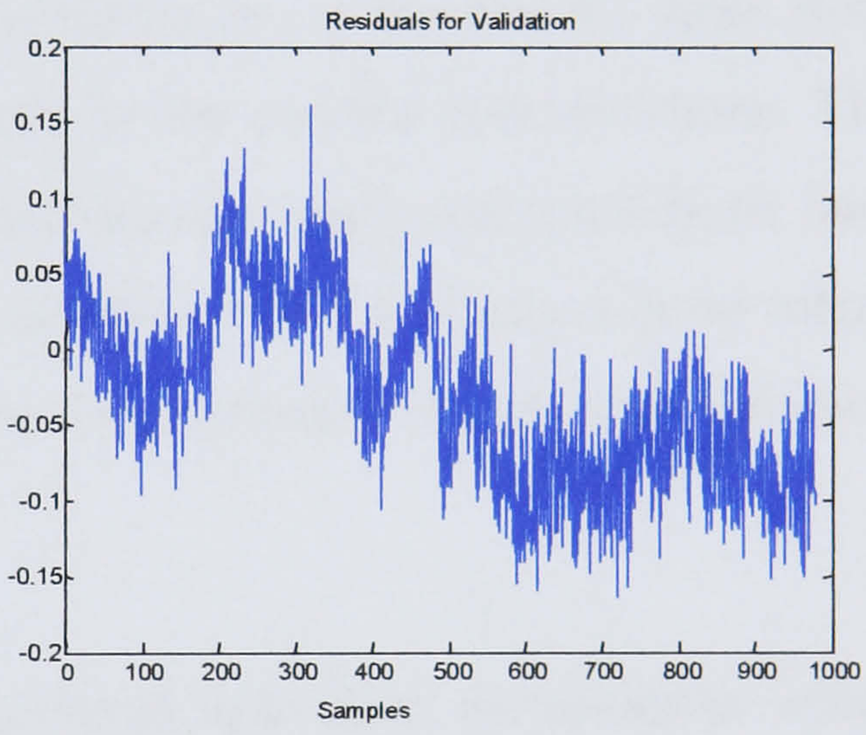


Figure 4.22d. Residuals for validation data.

Figure 4.22. Results for the DOE batches E1 to E8 for the experimental and the validation data set for the modelling of ammonia after removal of the air background.

4.5 Discussion

This Chapter has demonstrated that infrared spectroscopic techniques can be used on-line to predict the concentration of the selected key components (product concentration and ammonia) in an industrial fermentation broth. Improvements over the current calibration model approaches were sought. The algorithm proposed uses

spectral window selection (SWS) to build a single model. The single model generated is not unique, with the random search procedure resulting in a wide spread in terms of performance. This variation in performance was overcome by generating multiple models and 'stacking' these to produce a more robust prediction. Significant changes occur during the course of the fermentation and adopting a single calibration model for the whole batch can limit the accuracy of prediction. It was shown that splitting the fermentation into a number of time intervals and building local calibration models offered increased accuracy. This was not possible with the DoE data since identifying sub-regions of operation is more challenging. Importantly, the fact that SWS has been shown to work on both NIR and MIR spectra gives some indication that the method is not instrument specific.

The postulation at the outset was that selection of the spectral regions would improve the calibration model results by reducing the contribution to the overall noise from those regions not containing information related to the analyte concentrations. The results presented confirm that SWS does indeed provide improved predictions over those obtained from full spectrum analysis. In addition, SWS provides a more robust method for wavelength selection than the two most popular wavelength selection approaches: GAs and Interval PLS.

The major issue is that genetic algorithms selected individual wavelengths which allowed the models to become too specific to the training data. GAs were did not indicate any critical regions for the three time intervals and the important wavelengths, as identified by the analytical chemist, were not selected preferentially. iPLS calculated local PLS models on fixed sub-intervals of the full spectral region and thus compared to SWS, no additional benefits were provided by iPLS.

A number of issues still remain to be addressed. The results above indicate that in certain cases offsets may be present. In most instances the offset could be largely eliminated by a simple bias. This bias could be determined from a single off-line sample. In other instances offset removal is more complex and a simple bias would not suffice. Offset removal is the key to obtaining reliable information and by using other process information such as off-gas measurements it may be facilitated. The approaches for offset removal are discussed in the next Chapter.

CHAPTER 5

PROCESS AND SPECTRAL DATA INFORMATION FUSION

5.1 Introduction

In Chapter 4 it was demonstrated that the improvement in accuracy of the inference of the broth concentration as a consequence of applying SWS for wavelength selection and stacking for calibration model construction was significant. However a degree of error remained in the determination of the concentration of some broth concentrations. To address this, a data fusion strategy, which realises a model relating the calibration residuals to the on-line process measurements (e.g. off-gas concentrations, temperature, pH) is proposed. The model was used to update the spectral calibration inference, thereby giving improved determination of the broth concentrations. The general concept of data fusion is first explained in the context of spectral analysis with techniques and previous applications reported in the literature being described. Following this a new data fusion technique is proposed, which involves the conjunction of process and spectral data or two forms of spectral data in a sequential manner to improve the accuracy of the calibration model. Finally, the proposed method is applied to the design of experiment fermentation data.

5.2 Motivation for data fusion

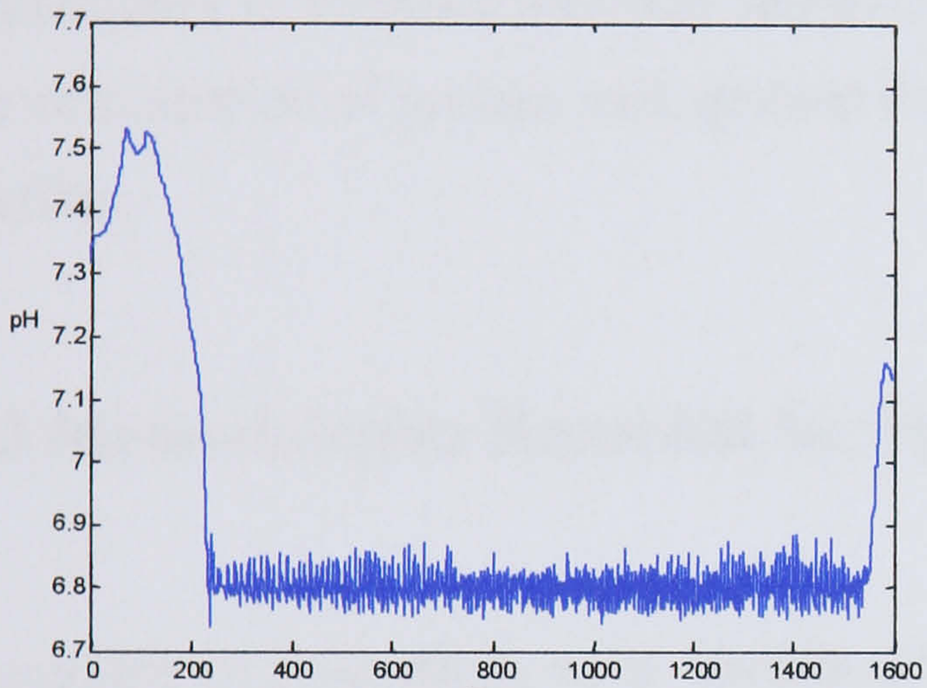
Process measurements have primarily formed the basis of modelling and monitoring strategies in the process industries, but more recently there has been an increase in the implementation of process spectroscopic instrumentation and the application of spectral data for the same tasks. However, in most studies, the spectral and process data are analysed independently and it is thus hypothesised that improved modelling can be achieved through the conjunction of different data forms.

In a fermentation process, on-line process data is typically measured at reasonably frequent time intervals (e.g. every five minutes) throughout the duration of the batch with typical process data including measurements of flow rates, levels and temperatures as well as pH and off-gas data measurements being recorded. In addition to the on-line process data, the quality/concentration of the product and other critical

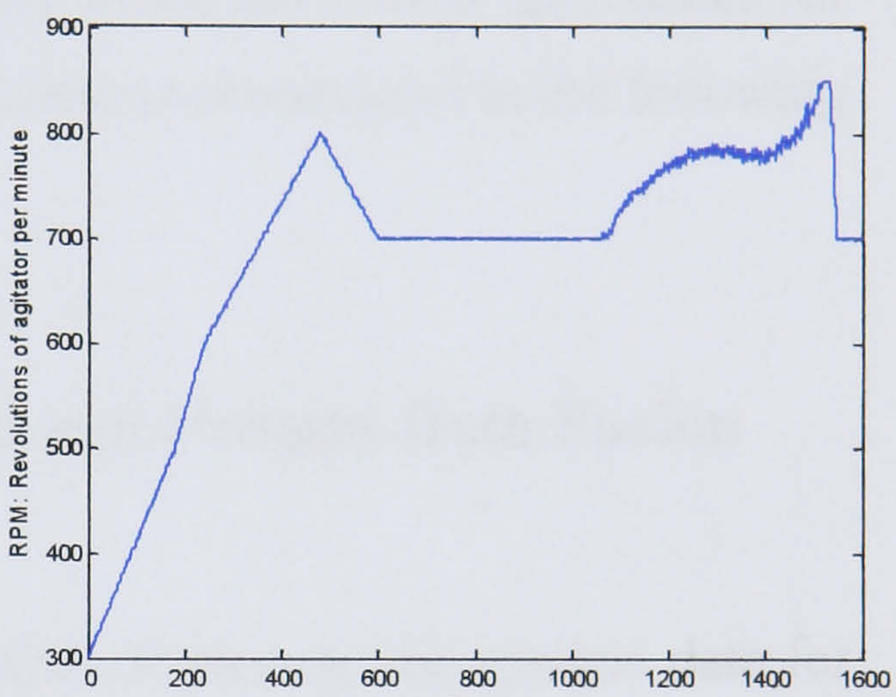
concentrations throughout the progress of the batch are measured through off-line analysis, (Sprang, 2004).

The process data according to Gurden *et al.* (2002) is characterised by heterogeneity since the process variables are measured in different units and hence scaling is necessary. On-line process data characteristics can be observed from univariate time series plots of the process variables in Figure 5.1. Clearly, a diverse range of measurements are recorded all of which are corrupted by noise, to a greater or lesser extent, as a consequence of the process and/or the measurement devices.

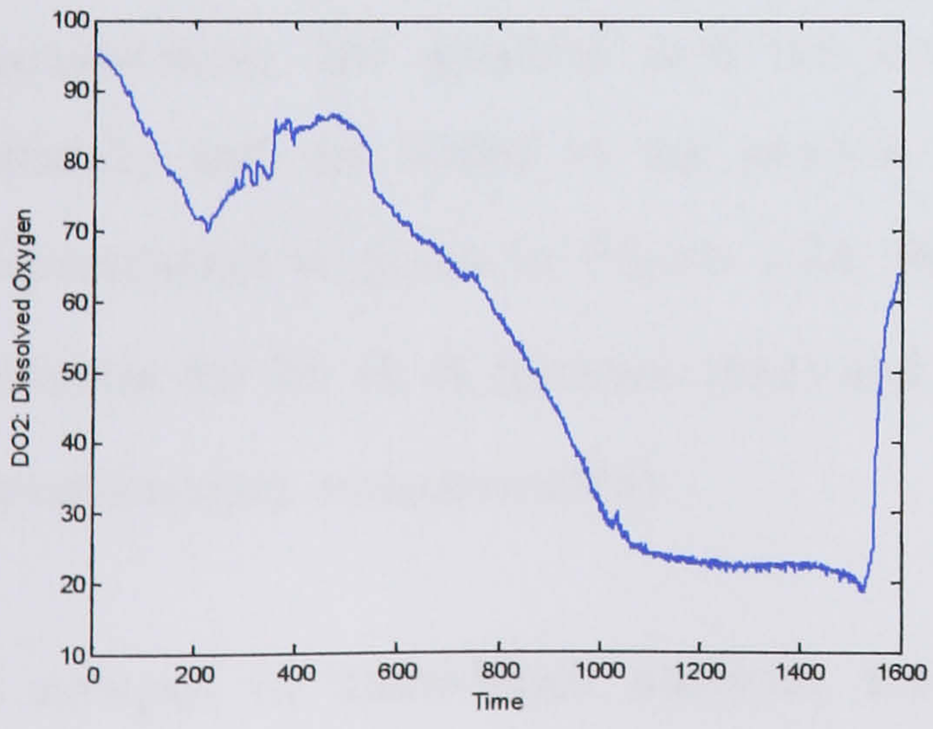
(a)



(b)



(c)



(d)

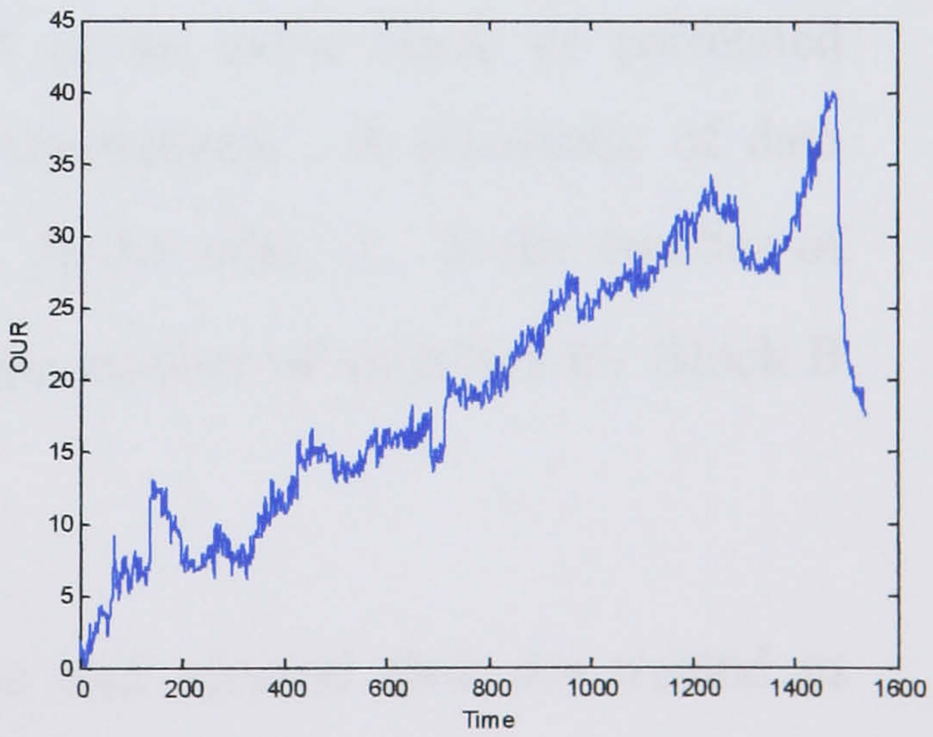


Figure 5.1. On-line process measurements: (a) pH, (b) RPM (Revolution of agitation per minute), (c) DO_2 (Dissolved oxygen), (d) OUR (Oxygen Uptake Rate).

Although the process data presented in Figure 5.1 is a valuable source of information for the construction of a process model, the use of alternative measurements strategies

such as the spectral data, can be very useful. There are fundamental and practical differences between the analysis of process and spectral data. Spectral data according to Gurden *et al.* (2002) are characterised a) by the chemical-state information as the spectra describe the chemistry relating to the molecular nature of the species, and b) homogeneity as all the wavelengths are measured in the same units and for this reason data scaling is less of an issue.

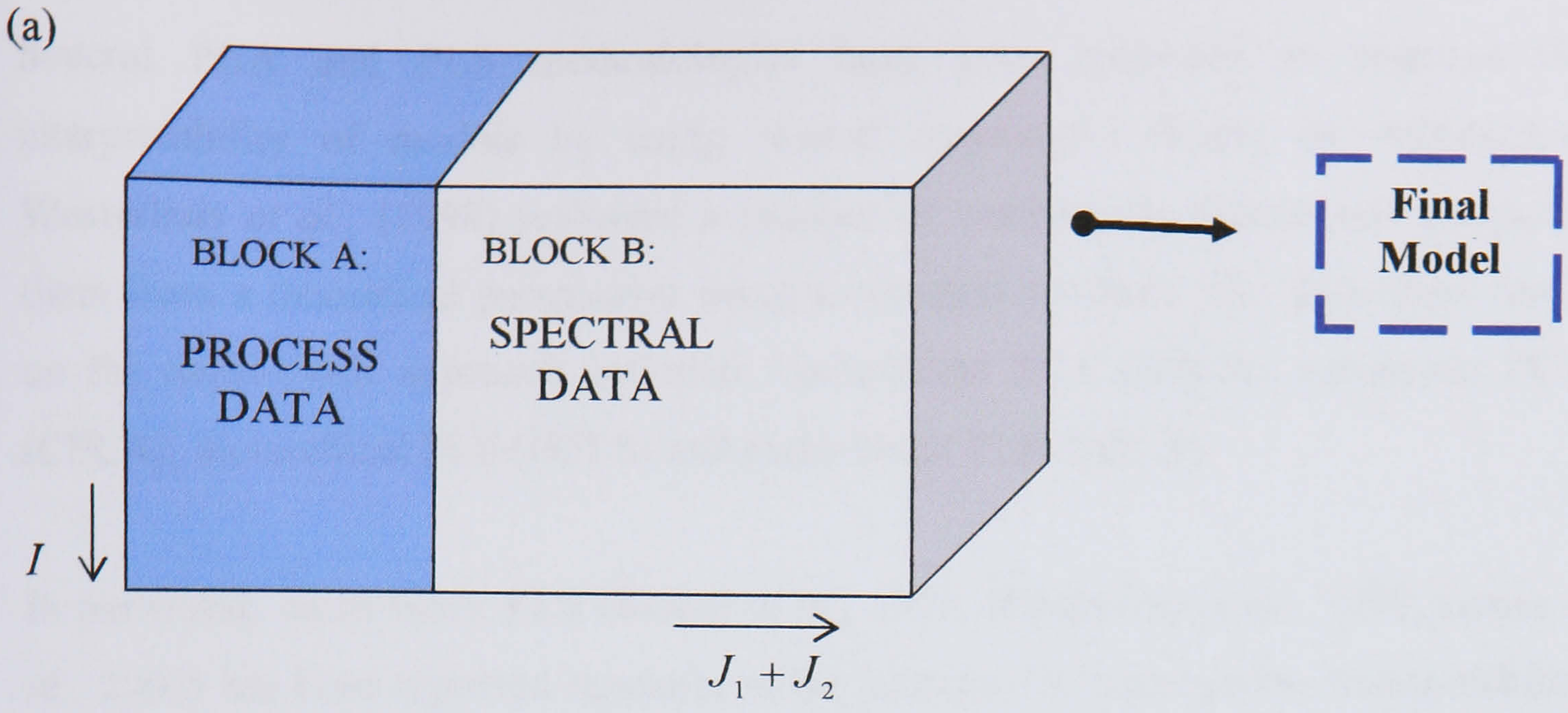
To maximise the amount of knowledge extracted about a process from data, one approach is to combine different forms of data. This will help release enhanced process control and fingerprint batches to identify poor and good production. In this study, the process and spectral data and two forms of spectral data are applied sequentially to enhance the final calibration model. There are several approaches for the combination of process and spectral data and these are considered in the following sections.

5.3 Methodologies Reported for Spectral and Process Data Fusion

A number of researchers have considered the fusion of process and spectral data for the purpose of process monitoring. To date two approaches have been reported in the literature (Gurden *et al.*, 2002): data augmentation and multi-block analysis. In data augmentation, the spectral data are considered as an extra block of correlated variables and are added to the process data measurements. A schematic of data augmentation is given in Figure 5.2a, where I, is the time, J_1, is the number of variables for Block A (process data) and J_2 is the number of variables for Block B (spectroscopic measurements).

In contrast in multi-block analysis, the process and spectral data are treated as separate blocks of information that are combined using a multi-block model (Smilde et. al., 2000). Figure 5.2b shows a schematic of the procedure.

(a)



(b)

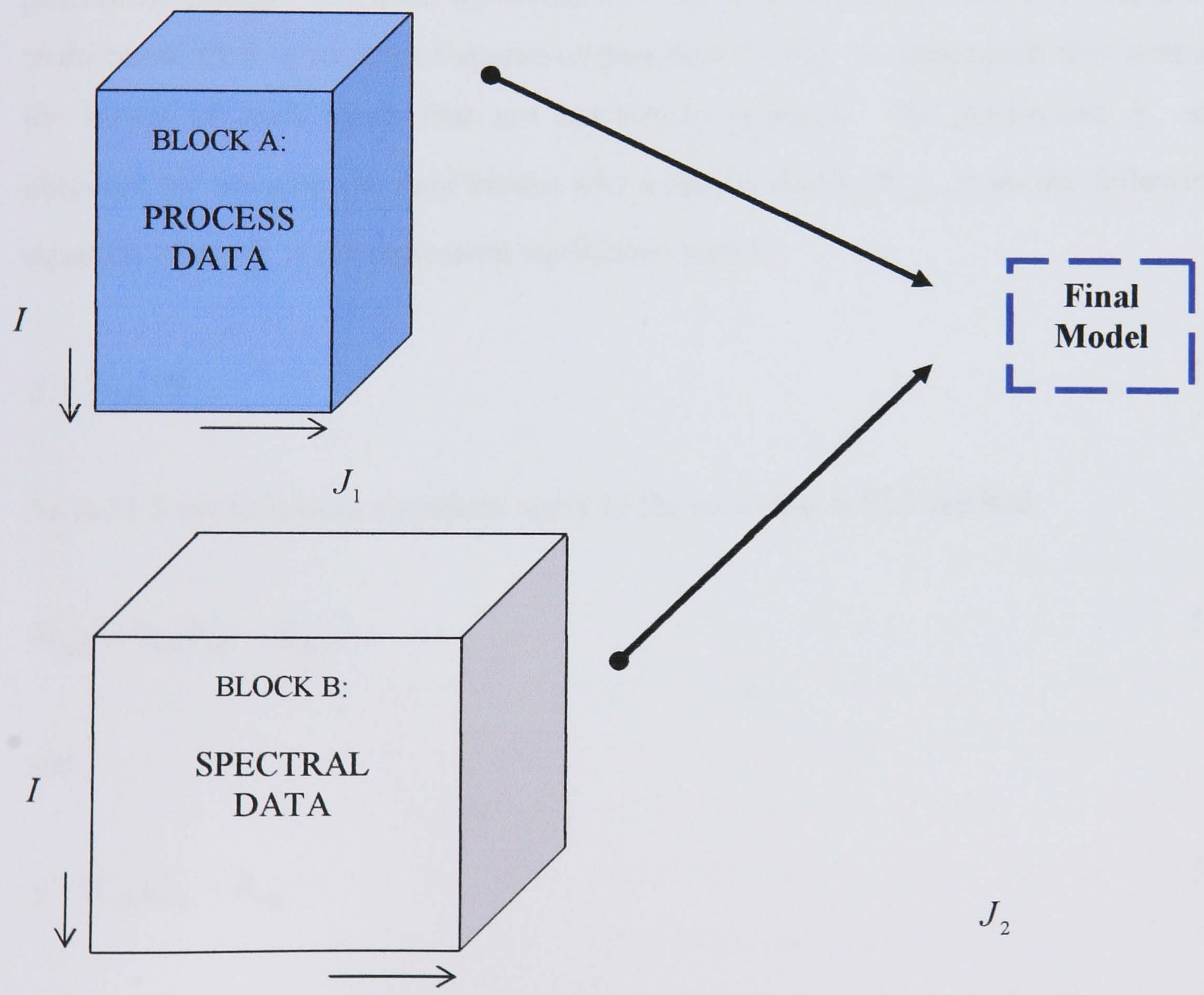


Figure 5.2. (a) Process and spectroscopic data form one set of data (Data augmentation), (b) Process and spectroscopic data fusion using a multi-block approach, (Gurden et.al., 2002).

Several PCA and PLS methodologies have been proposed to improve the interpretability of models by using several meaningful blocks of information. Westerhuis *et al.*, (1998) reviewed a number of existing algorithms and compared them from a theoretical perspective using a common notation. The algorithms based on the multi-block approach included: hierarchical PCA (HPCA), consensus PCA (CPCA), hierarchical PLS (HPLS) and multi-block PLS (MPLS).

In particular, multi-block PLS (Kourti *et al.*, 1995, Westerhuis *et al.*, 1998, Lopes *et al.*, 2002) has been reported regularly in the literature to improve the interpretability of multivariate models. In multi-block PLS, the variables are split according to process knowledge. The main difference between multi-block PLS and PLS is that for multi-block PLS, a matrix of scores termed super-scores is determined that contain the scores of each block that are computed separately. The predictions $\hat{\mathbf{y}}$, are obtained, by merging the data blocks into a single matrix $\mathbf{X}_{MB}$, from the following equation where $\boldsymbol{\beta}$ is the regression coefficient vector:

$$\hat{\mathbf{y}} = \mathbf{X}_{MB} \cdot \boldsymbol{\beta} \tag{5.1}$$

As in PLS the following equations apply to the multi-block PLS method:

$$\mathbf{X}_{MB} = \mathbf{T}_{MB}\mathbf{P}_{MB}^{T} - \mathbf{E}_{MB} \tag{5.2}$$

and

$$\mathbf{y} = \mathbf{T}_{MB}\mathbf{q}_{MB}^{T} + \mathbf{F}_{MB} \tag{5.3}$$

where $\mathbf{T}_{MB}$ and $\mathbf{P}_{MB}$ are the scores and loading matrices respectively for $\mathbf{X}_{MB}$, $\mathbf{q}_{MB}$ is the vector of $\mathbf{y}$ loadings and $\mathbf{E}_{MB}$ and $\mathbf{F}_{MB}$ are the residuals for the predictors, $\mathbf{X}_{MB}$ and for the quality variable, $\mathbf{y}$. The intermediate steps of the multi-block algorithm are described in the paper of Westerhuis *et al.*, (1998).

A number of applications based on data matrix augmentation and multi-block techniques and their limitations are described in the next two sections.

5.3.1 Applications of Process and Spectral Data Fusion

The first reported application of spectral and process data fusion using data augmentation for final product quality modelling was in food manufacturing, (Pedersen, 1997). In their initial studies, they failed to address the issue of heterogeneity i.e. scaling was not appropriately performed. For their study, 59 samples were considered. The augmented data matrix was constructed as follows: (a) the chemical analysis of 5 raw materials, (b) the NIR measurements of 3 raw materials, (c) 43 process variables and the NIR measurements of 2 intermediate products. If the original spectra were included in the augmented data matrix in full scale, they would number more than 5000 variables. Thus, it was decided to apply PCA to the NIR spectra. Based on the analysis, 10 principal components were included in the data matrix for the analysis and this gave an overall balance between the number of process variables and spectral variables. The best models for the quality parameter were built using the combined data set, although the calibration models were developed on only 18 samples and maybe cannot be stable.

Wong *et al.* (2005) proposed two different approaches based on multi-block analysis. They combined chemical and physical data to develop an enhanced model for performance monitoring and fault diagnosis based on UV-visible and process data. In the first approach, consensus PCA (Westerhuis *et al.*, 1998) was used for the fusion of the data. The spectroscopic and process data formed two blocks and was integrated using multiblock analysis (Figure 5.3). In this approach, a starting super score t_T is selected as the first column for one of the blocks and it is then regressed on both blocks to give block variable loadings. The block scores t_B, $B = 1,2$, are then calculated and combined into a super block T.

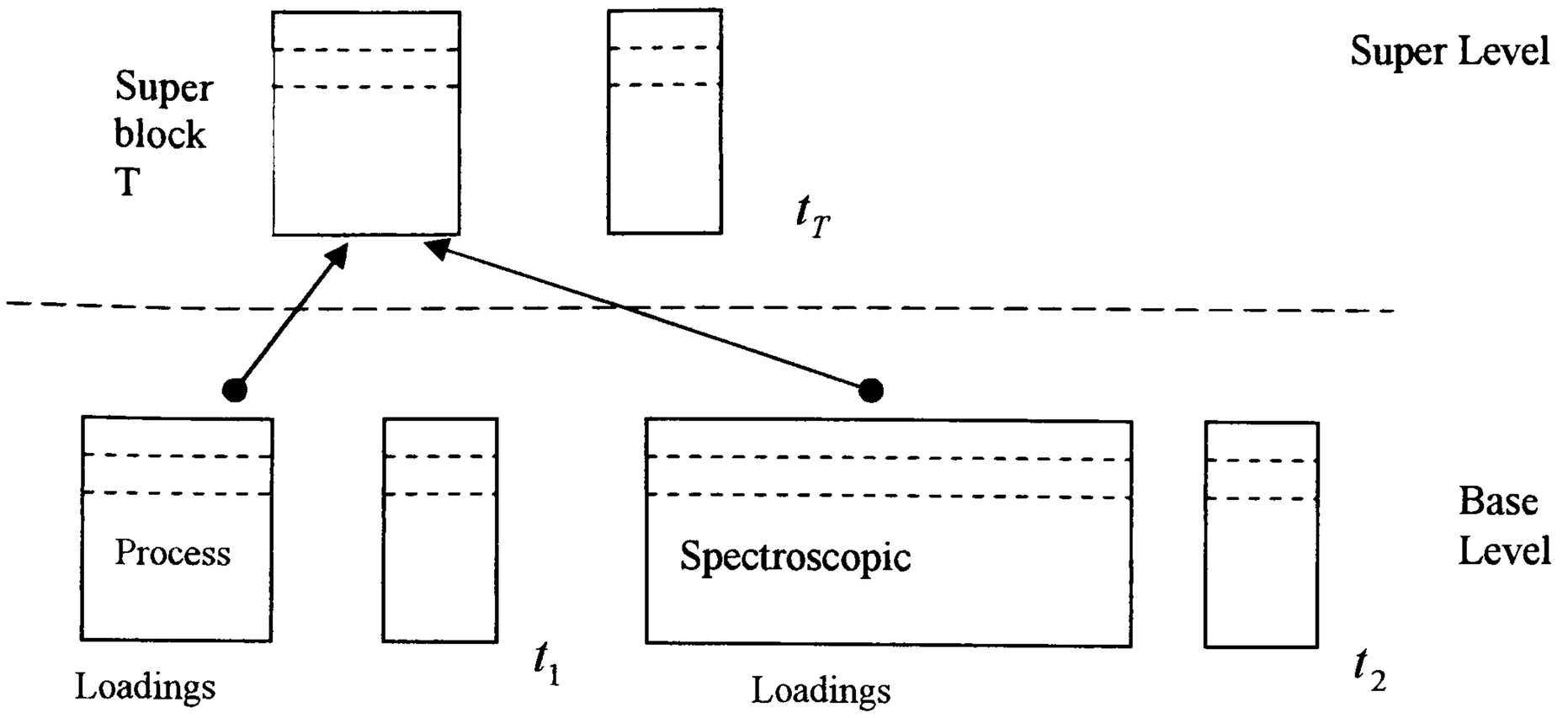


Figure 5.3: Process and spectral data integration with consensus PCA, Wong *et al.* (2005).

The only issue is that the number of process variables was seven while the number of spectroscopic variables was 216 and the weighting between these two different blocks of information was not clear from the paper. The second approach was similar to the first but wavelet analysis (Trygg and Wold, 1998) was applied initially as a pre-processing step for the spectral data since spectral data comprises a large number of wavelengths and the reduction of its dimensionality is beneficial. The multiblock-wavelet approach provided a suggestion to the above issue by reducing the spectral data from the original number of 216 wavelengths to 14 wavelet coefficients. The results of the analysis were compared with those attained from separate analysis on the spectral and process data. The authors concluded that the development of an integrated framework can help in the understanding of the process more than that attained from a model based on one block of data. Moreover the interpretation is made simpler in a single representation from the integrated model.

The current data fusion algorithms have a number of limitations and they are primarily related to the scaling of the two blocks and the weighting between the blocks. A weighting factor for each block is necessary since the number of wavelengths greatly exceeds the number of process variables. The scaling issue is also important. If the scaling of the spectral and process data is not performed appropriately, the heterogeneity of the problem may fail to be addressed. It is necessary to scale the two data sets according to their relative significance in terms of describing the process.

5.3.2 Fusion of Different Spectroscopic Measurements

In some cases, data from a number of spectral instruments may be collected on a process at the same time, for example NIR and MIR or NIR and Raman. The procedures used to combine data from these different spectral devices are the same as those discussed in section 5.3.1. For example for the augmented and the mulitblock approach illustrated schematically in Figure 5.2, the NIR data would be Block A and the MIR data or Raman data, would be block B. When different forms of spectral information are combined, the values of the variables (i.e. absorbances) are similar in magnitude, thus heterogeneity is not an issue. However the information that needs to be combined comes from different sources that had been pre-processed differently and the weighting between blocks may still be an issue.

Workman, (1999) demonstrated the benefits of spectral data augmentation by achieving the successful quantitative analysis of polymers by combining NIR and Raman spectra. More specifically various results from the multivariate regression modelling of the blend composition of poly blends were compared using the RMS error. A number of different pre-processing methods such as mean centering and autoscaling were used for the augmented spectra but no weighting considerations were taken into account for the two different kinds of spectral measurements. The results included the use of (a) only NIR measurements, (b) only Raman measurements and (c) the augmented NIR and Raman spectra values. According to Workman, (1999), data augmentation can be used to increase the information content in an analytical situation.

Following the work of Workman, (1999), Cuadrado *et al.*, (2005) studied the joint use of NIR and MIR spectra though data augmentation for the determination of wine parameters for a wide range of wines and grapes varieties. They concluded that the determinations that were carried out using only the NIR region gave better results than those from the MIR region and that this behaviour may be due to instrumental and technical characteristics. Finally, the combination of the two spectral zones (NIR and MIR) was successful and improved the determination of the studied parameters,

although no weighting was considered between the two spectral zones. Thus, a much larger data set was used that included 2100 variables/wavelengths. The combination of NIR and MIR measurements can be seen in Figure 5.4.

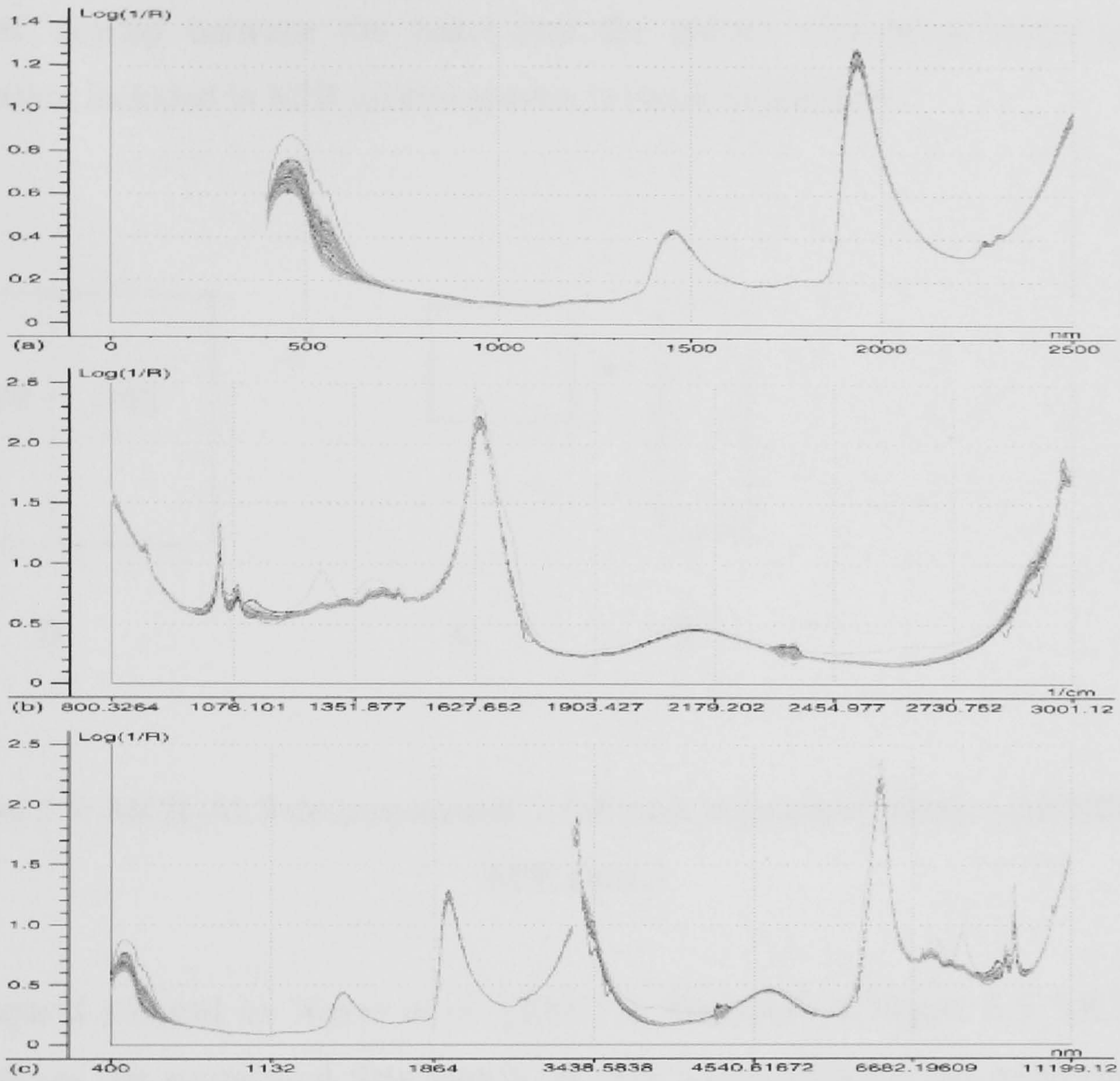


Figure 5.4: (a) NIR spectra, (b) MIR spectra, (c) combination of NIR and MIR spectra, Cuadrado *et al.*, (2005).

The use of the augmented matrix of NIR and MIR measurements has also been reported for the purposes of multivariate curve resolution (Brown *et al.*, 1996; Vandeginste, 1985). Naves *et al.*, (2003) incorporated the fusion of spectral data into multivariate curve resolution alternate least squares (MCR-ALS), which is one of the most commonly applied curve resolution techniques, (Tauler *et al.*, 1993, 1994, 1995) for the monitoring of temperature-dependent protein structural transitions. MCR-ALS is a method that allows the mathematical resolution of concentration and spectral profiles of pure species recorded in multi-component mixtures and more importantly, it can be applied simultaneously to several matrices generated from the same process

that has been monitored using several spectroscopic techniques. The methodology has been successful applied for on the modelling of other protein processes. Moreover, the combined use of the MIR and NIR spectra can be a new and improved way to study structural changes in protein processes for a variety of reasons. In NIR there is less apparent overlap between the water and the protein absorption bands and the information included in MIR protein spectra is better established.

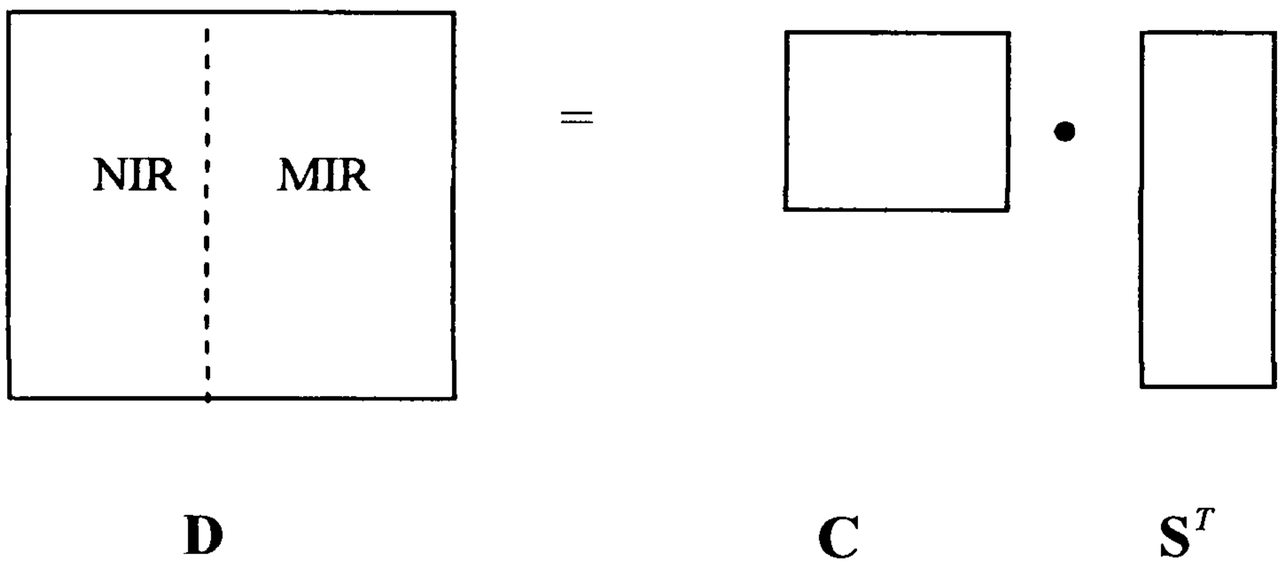


Figure 5.5: MCR-ALS decomposition: a full rank augmented matrix with NIR and MIR data.

The method adopted by Naves *et al.* (2003) is illustrated in Figure 5.5. MCR-ALS decomposes the augmented data matrix $\mathbf{D}$ into the product of two small matrices, $\mathbf{C}$ and $\mathbf{S}^T$, containing the pure profiles of all the chemical contributions present in the raw mixed experimental measurements:

$$\mathbf{D} = \mathbf{C} \cdot \mathbf{S}^T + E \qquad 5.4$$

where $\mathbf{C}$ and $\mathbf{S}^T$, contain the pure column concentration profiles and the pure row spectral signal profiles respectively, and matrix $\mathbf{E}$ is the experimental error, i.e. the residual variation of the data set that is not attributable to any chemical contribution. The results from the above application looked promising since only the combined analysis of the NIR and MIR data managed to detect and model the protein conformations involved in the process.

Recently, the multi-block approach was used for the combination of NIR and MIR data. Bras *et al.*, (2005) combined NIR and MIR spectra through multi-block PLS for the prediction of protein, moisture, fat and fiber content of soybean flour utilized in animal feeds and compared the results with those from PLS. They concluded that the outcome from multi-block PLS gave significantly better results than those from the PLS models based exclusively on the MIR data. A measure of model performance was the RMS error. On the other hand when the NIR results were compared with the combined data model, it was concluded that the combination of NIR and MIR spectra was only moderately advantageous as the RMS errors were of similar magnitude for the two methodologies. However even this improvement showed that the modeling power could be enhanced through the inclusion of some new information captured in the MIR spectra that was not evident in the NIR block. The combination of NIR and MIR measurements can be seen in Figure 5.6.

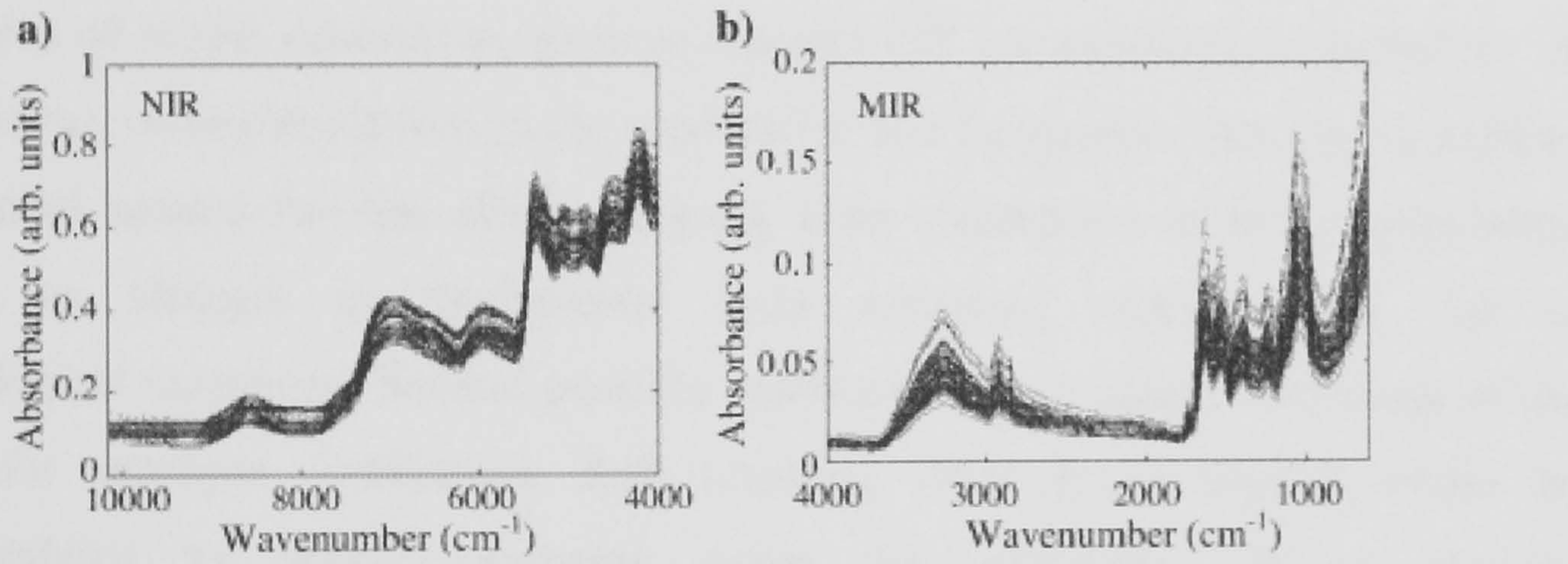


Figure 5.6. NIR spectra (left) and MIR spectra (right) of the soybean flour samples, (Bras *et al.*, 2005).

Felicio *et al.*, (2005) combined NIR and MIR spectra for the prediction of the flash point in gas oil, benzene and research octane number in gasoline, using the multi-block method and serial PLS (S-PLS), (Berglud and Wold, 1999). S-PLS is an alternative and time consuming multi-block PLS algorithm. In this particular example, for the studied parameters, S-PLS only gave marginally better results. The only pre-processing applied was dividing the MIR and NIR spectra by their maximum absorbance value as the magnitude of absorbance in the NIR spectra was much higher than that for the MIR spectra and this according to Felicio *et al.*, (2005) '*could lead to erroneous conclusions* '. After comparing the result with multi-block PLS, it was concluded that the best calibration model was that attained by applying PLS based on

one block of information. More precisely, according to the RMS error, both flash point and benzene were better modelled with the MIR data, while S-PLS provided similar results to those for the NIR spectra. Research octane number was better modelled using NIR data, with S-PLS providing the same results as for the NIR spectra.

Following the limitations identified from the previous applications reported, a novel sequential data fusion strategy is proposed and is explained in detail in the following section. In the proposed methodology, each piece of information is used sequentially; thereby remove the issues of scaling and weighting.

5.4 Sequential Data Fusion Modelling

In the area of model calibration, accuracy targets will not necessary be satisfied. In certain cases, offsets/deviations in the predictions will be present. There are a number of potential causes for the offsets ranging from calibration model extrapolation through to changes in fundamental light scattering characteristics due to morphological variations. Several possible solutions exist to address the cause of the offset, for example, calibration drift resulting from probe fouling could be accommodated by model up-dating using the infrequent off-line analysis measurements. Whatever the cause, from a user perspective, the existence of offsets limits the effectiveness of the measurement and significantly compromises the quality of control achievable. To reduce the magnitude of the offsets, a new strategy is proposed based on the modelling of the calibration model residuals.

The sequential three-step data fusion modelling procedure is summarised in Figure 5.7. It involves the conjunction of two blocks of data for the modelling of the quality variable. Step one involves the calculation of the calibration model residuals. The next step is to model the residuals using Block B data, thereby attaining the innovations, i.e. the residuals of the residuals are calculated. Figure 5.7c illustrates the final step where the predictions of the quality variable from Block A and the residuals predicted from the process data are combined to generate the final prediction value.

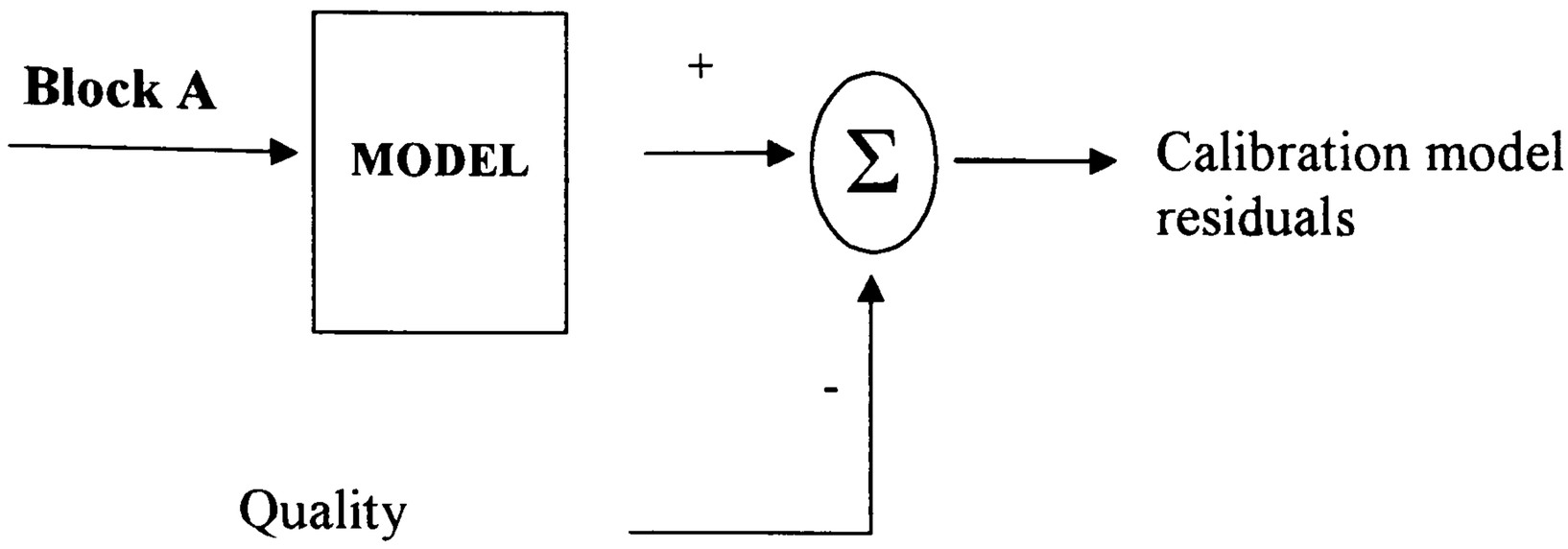


Figure 5.7a. Calculation of the calibration model residuals.

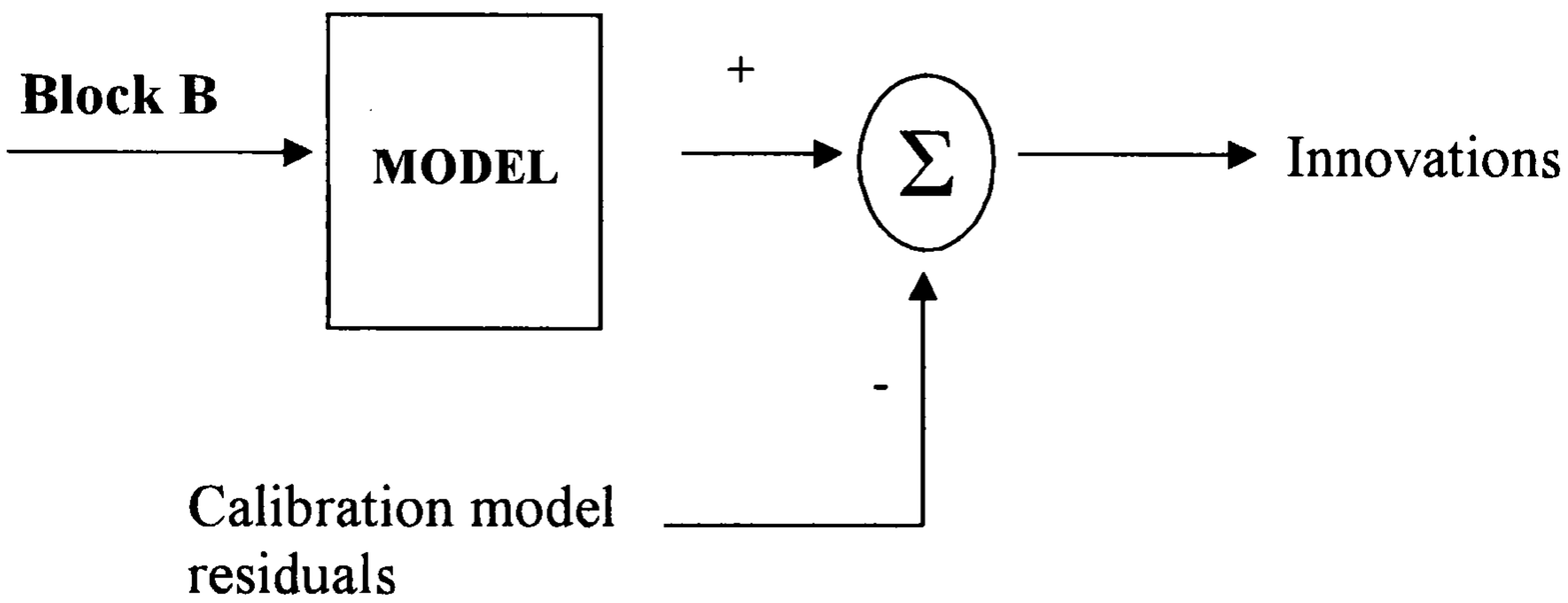


Figure 5.7b. Modelling of the calibration model residuals from Block B and the generation of the innovations.

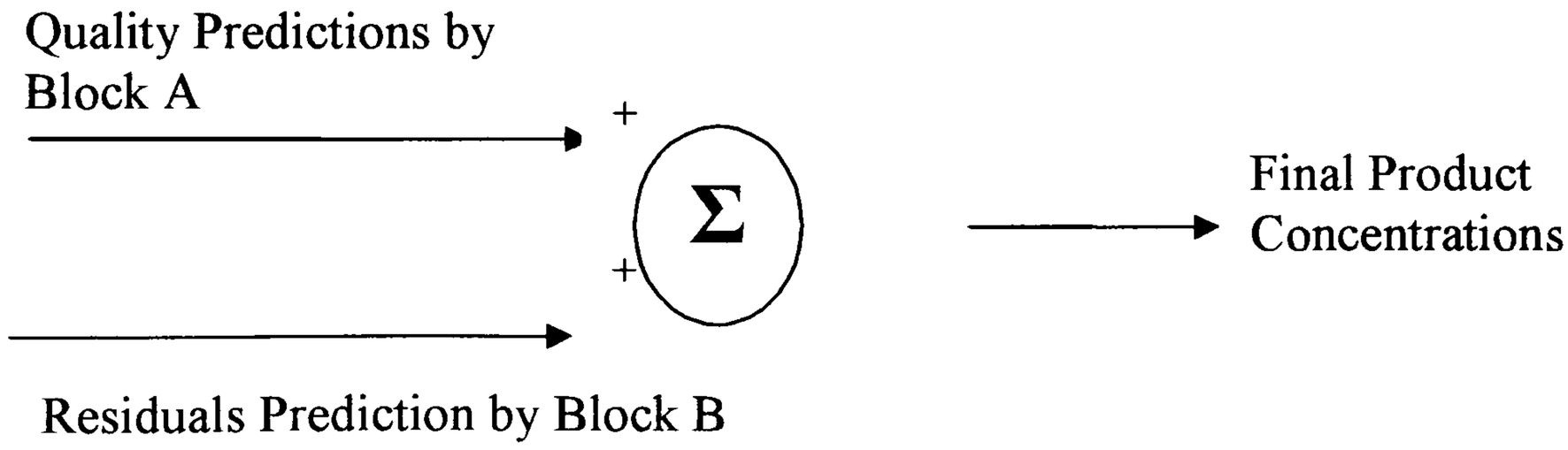


Figure 5.7c. Prediction of the quality variable.

Figure 5.7: Schematic of the proposed data fusion algorithm.

For the fermentation process, the availability of on-line process measurements or a second set of spectroscopic measurements, offers a route for the development of a data fusion strategy. More specifically, offsets in the model predictions (residuals) are modelled using a subset of the available process or spectroscopic measurements through a PLS based linear regression. The final prediction is thus attained by combining the original calibration model prediction with the residual prediction from the second set of measurements. This strategy offers the opportunity to take due account of process changes observed through alternative measurements, not captured by a unique set of spectroscopic information.

A number of different methods of data fusion are proposed in this thesis. The method selected depends on the information available from the process. In this study three different sources of on-line information were available, i.e. NIR spectra, MIR spectra and traditional process on-line measurements. Thus the following modelling cases were examined:

When one block of information is used for modelling the quality variables, the following cases were considered:

a) Model each quality variable using NIR data.
b) Model each quality variable using MIR data.
c) Model each quality variable using the most influential process variables.

When two blocks of information were considered, the following combinations of data were utilised:

a) Model the quality variables using NIR data with PLS (Figure 5.7a) followed by residuals modelling with the process data. (Figure 5.7b).
b) Model the quality variable using MIR data and model the residuals using process data.

The final and third step for each of these three cases is a summation of the quality variable prediction from the first step and the residual prediction from the second step (Figure 5.7c).

Moreover another combination was considered: the quality variables were modelled using NIR data and the residuals were modelled using MIR data.

5.5 Industrial Case Study

The motivation for the fusion of different information sources in the fermentation application comes primarily from the fact that offsets were evident in the prediction of the broth concentrations from the spectral calibration models. In this application, the experimental design data rather than standard operating data was used since it spanned a wider operating range. To demonstrate the underlying concepts of the methodology, the development of calibration models for product concentration and the concentration of ammonia was investigated. The aim of the analysis was to examine the potential of improving the calibration models through the fusion of different information sources. Prior to modelling spectral data, pre-processing was performed as described in Chapter 4 and process data pre-processing was carried out as described in the following section.

5.5.1 Process Data Pre-screening and Pre-processing

The pre-processing of the process variables is an important step in the modelling of batch processes. Batch data pre-processing has been considered and the issues associated with it have been tackled by a number of researchers including Bro and Smilde, (2003) and Kourti, (2003).

The graphical techniques associated with batch process data pre-screening (time series plots, trend plots and scatter plots) help in the understanding of the data characteristics. In this way spurious observations can be treated and relationships between variables identified. There are occasions where by examining the univariate

time series of variables, process problems can be identified immediately without resorting to more complex methods. If univariate visualisation does not identify any outliers or spurious points, it is still necessary to perform multivariate visualisation to identify more subtle data problems.

Prior to performing even the visualisation step, missing data requires to be in-filled. Missing data may be due to random events such as a manual sample entered incorrectly but more commonly they are non-random and are a consequence of instrument failure. In general, values missing at random can be treated without taking into consideration the reason why they are missing, as they tend to be over a short time scale and are not feasible process changes. Alternatively values missing over a longer period need to be considered with more care and the analysis of the data must address the cause of failure as it may be indicative of a real process deviation.

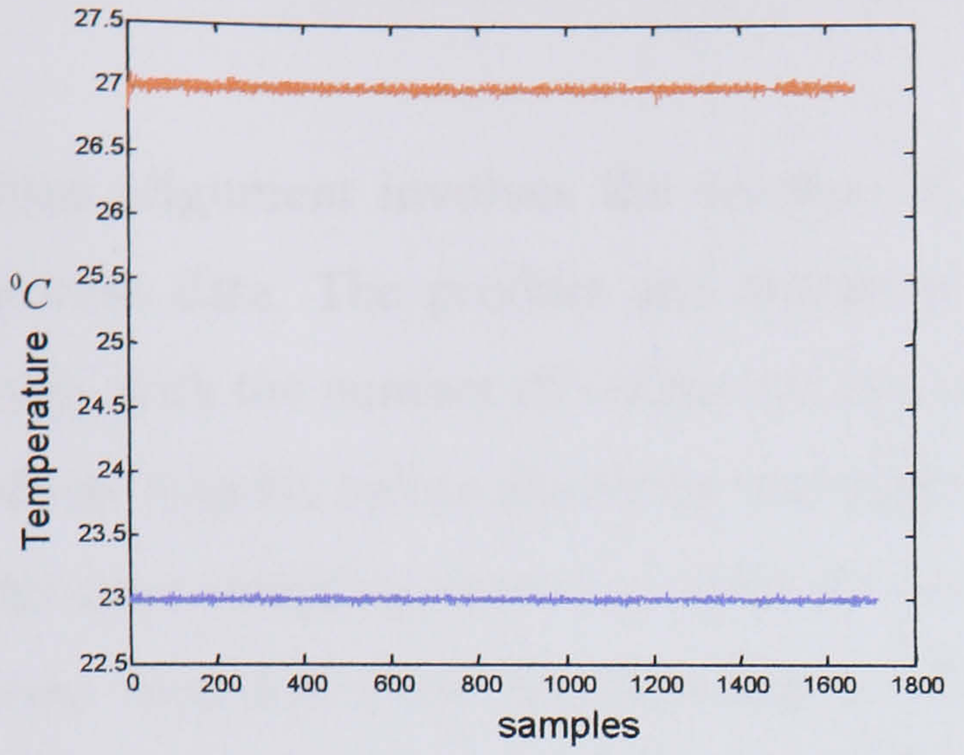
The simplest method for treating missing data is to delete time samples over which missing data is identified or alternatively remove a variable totally from the analysis or to adopt a combination of the two, thereby obtaining a data set that is complete. When there are a large number of samples missing for a specific variable, variable deletion can be used. The concern here is that important process variables may be removed from the data set and are thus not included in the final model. In most cases, in-filling techniques are used for the handling of incomplete data. There are a number of methods that have been used for in-filling (Little and Rubin, 1987). These include zero-order or first-order linear interpolation, which is the most common method, zero order regression, local averages, univariate and multivariate interpolation and auto-regressive time series models.

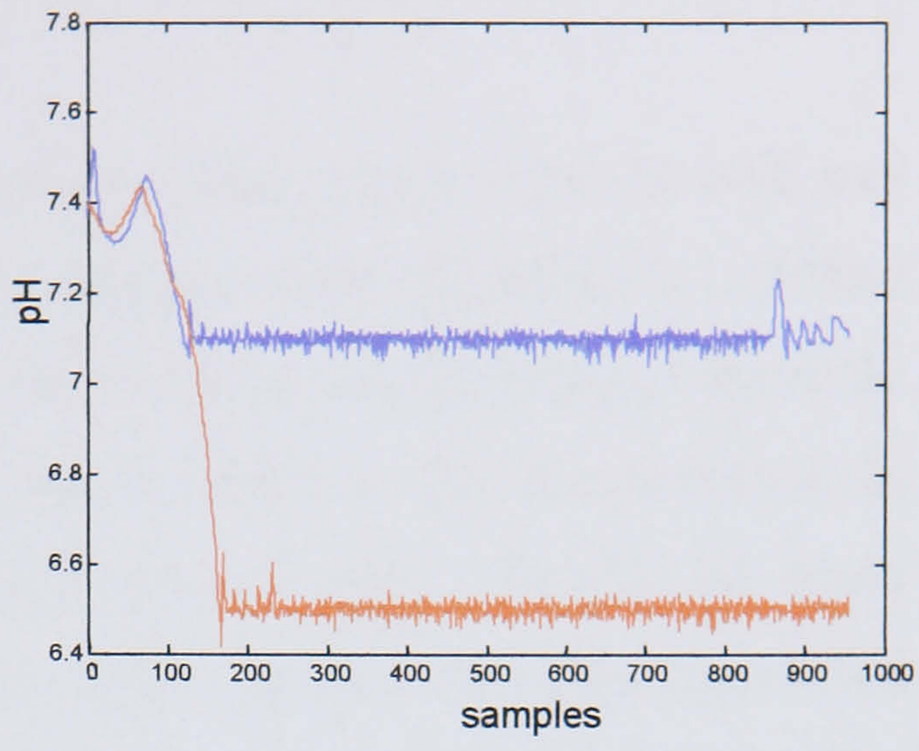
The removal of outliers is also important in the analysis. Outliers tend to be infeasible changes that occur over short time periods and are inconsistent with the rest of the data observations. They can be genuine observations or they could be caused by recording errors. Basic time series plots can help detect univariate outliers but sometimes it is necessary to look at the data in a multivariate manner to detect the outliers. Their treatment is essentially the same as that used for missing data.

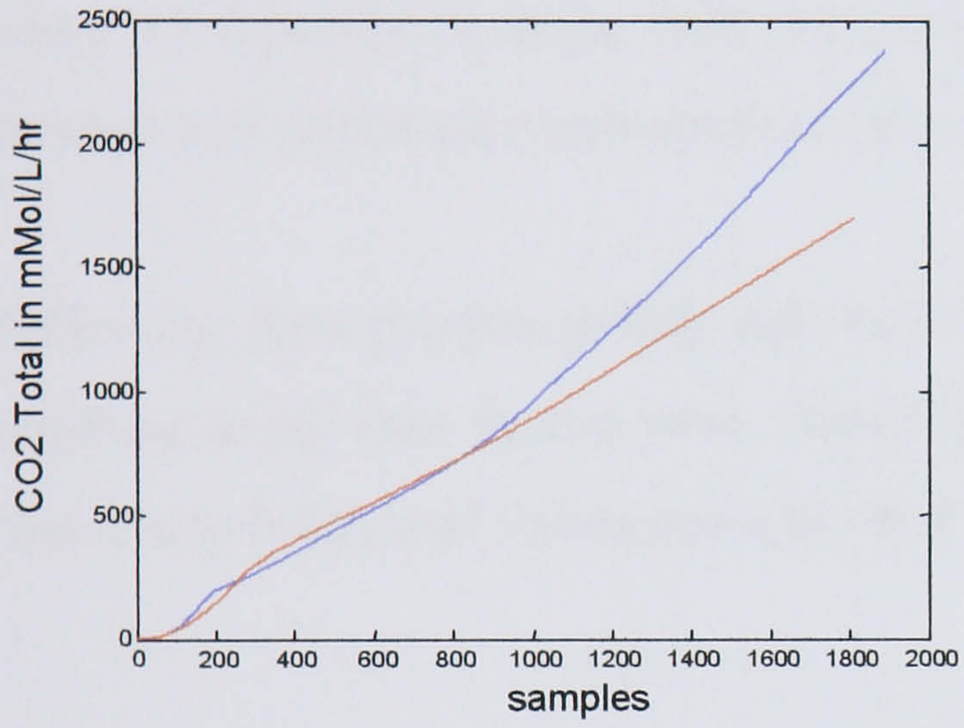
5.5.2 Selection of the Fermentation Process Variable Set

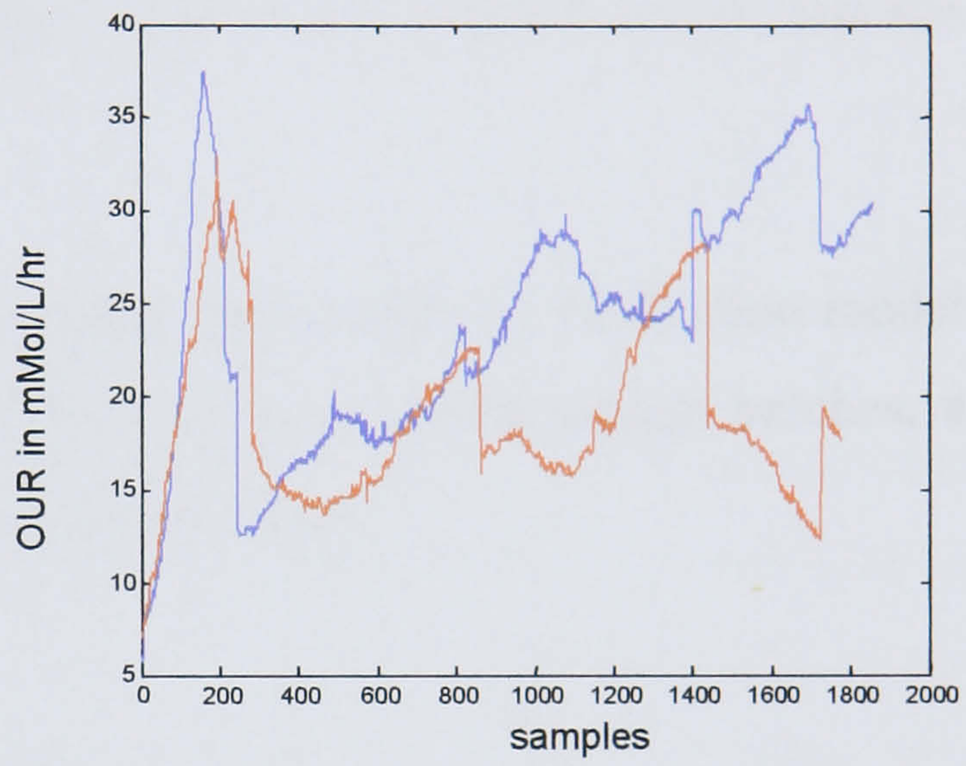
The series of fermentations undertaken as a consequence of the experimental design procedure were described in Chapter 4. The objective was to obtain better operating region coverage. A factorial design was used to investigate the interaction of the environmental conditions (pH and temperature) and the feed rates (sugar feed and oil feed) for fault diagnosis for the resulting final product concentration. All the reported results are from batches E1 to E6. Batches E7 and E8 were not used because only MIR results were available for these batches. The on-line process variables available included off gas concentration measurements (CO_2, N_2, O_2, Argon), carbon evolution rate (CER), oxygen uptake rate (OUR), respiratory quotient (RQ), dissolved oxygen (DO_2), pH, stirrer speed (in rates per minute - RPM), and temperature (^0C).

Five process variables were identified as those most indicative of product concentration: CER, CO_2 Total, OUR, pH and temperature. Although pH and temperature are control variables, their values in the experimental design vary from batch to batch as it was described in section 4.4.5.2 and more specifically in Table 4.11. The motivation for selecting this set of variables was based on bioprocess understanding. CER is included as it gives a strong indication of growth and product formation, whilst total CO_2 provides information concerning the accumulation of product over the course of the batch. OUR gives additional information over and above that provided by CER in that the ratio of CER to OUR provides insight into microbiological changes. Finally, pH and temperature are known to impact on behaviour and were manipulated within the DOE. An example of typical batch trends for these variables are shown for two batches in Figure 5.8., where the time series plots of the five chosen variables and their variations within each batch are illustrated.









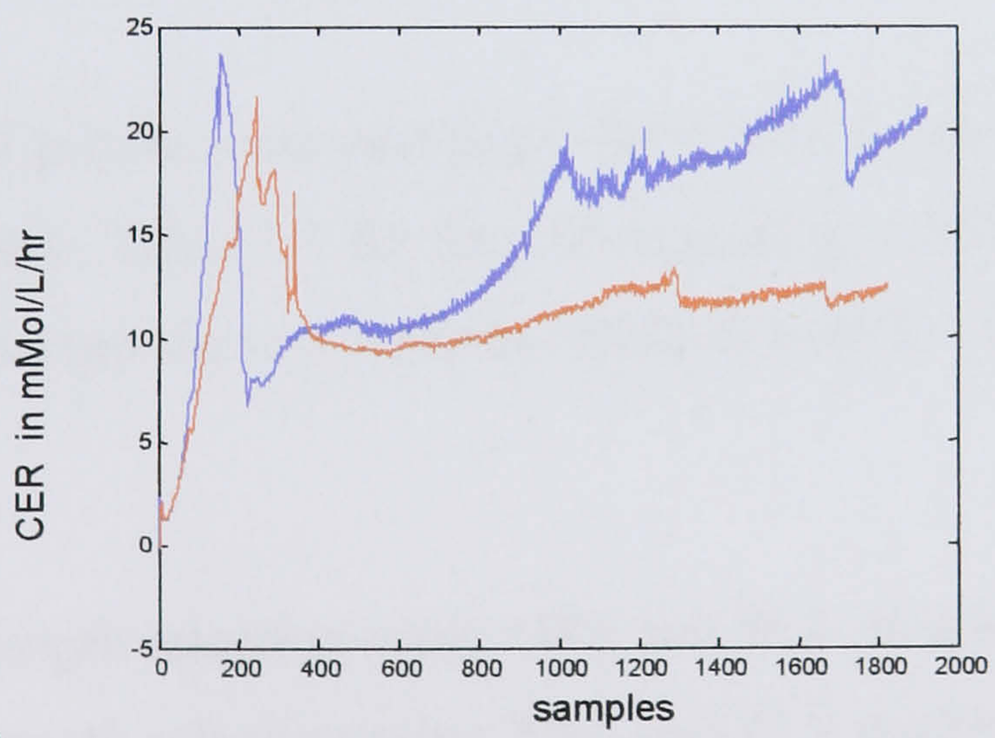


Figure 5.8. Example from two batches (blue –E3- and orange –E5-) of the five variables that were considered to be the most influential for the prediction of product concentration.

In addition to the data pre-processing and pre-screening procedures described in section 5.5.1 (treating missing data and removal of outliers), it is necessary to time align the process data and the spectral measurements.

5.5.3 Time-alignment for the Fermentation Application

Time alignment involves the creation of a common time axis for the spectral and process data. The product and ammonia concentrations were measured by off-line assay with the number of values recorded over the period of a batch being of the order of ten. A cubic spline algorithm was applied to obtain values for the concentration for the same sampling interval as used for the spectral measurements. The spectral values were recorded every fifteen minutes with the on-line process variables monitored every five minutes. Thus, it was decided to sub-sample the on-line process variables every 15 minutes to align with the sampling times of the spectral values and the product and ammonia concentration values.

Following data pre-processing and time alignment, the benefits of calibration model building using data fusion were considered. For the experimental design batches, a 'leave one batch out' validation approach was implemented.

5.5.4 Results for Product Concentration

For the prediction of product concentration, the results for five experimental design batches are presented in Table 5.1 for the corresponding validation batches. In Table 5.1 each column corresponds to a different method used for the construction of the calibration model:

a) Column 2: Wavelength selection using SWS and PLS stacking on the NIR spectra.
b) Column 3: Wavelength selection using SWS and PLS stacking on the MIR spectra
c) Column 4: linear PLS applied to the process variables.
d) Column 5: Wavelength selection using SWS and PLS stacking on the NIR spectra followed by residuals modelling using the process variables.
e) Column 6: Wavelength selection using SWS and PLS stacking on the MIR spectra followed by residuals modelling using the process variables.
f) Column 7: Wavelength selection using SWS and PLS stacking on the NIR spectra followed by residuals modelling using the MIR spectra.

Table 5.1. Results for the product concentration after the application of the traditional and proposed methods for the validation batches.

	NIR	MIR	Process Variables	**NIR + Process Variables**	MIR+ Process Variables	NIR +MIR
Batch E1	0.107	-	0.058	**0.025**	-	-
Batch E2	0.123	0.073	0.176	**0.052**	0.071	0.072
Batch E3	0.292	0.095	0.172	**0.079**	0.091	0.080
Batch E5	0.108	0.091	0.066	**0.042**	0.056	0.077
Batch E6	0.195	0.082	0.068	**0.053**	0.068	0.082

Five batches were chosen for the cross-validation purposes. The first row identifies the adopted method. The first column identifies the batch that was used for validation, with the remaining four batches used for training purposes. For example, the second row in Table 5.1 represents the validation analysis carried out on batch E1 after using training batches 2, 3, 4, 5 and 6 and the RMS results. MIR data were not available for batch 1 due to instrument availability, thus the training and validation techniques that include MIR measurements were constrained by data availability. From the analysis presented in Chapter 4, the removal of the air and water backgrounds give comparable results for the MIR data set. Thus in this case only the removal of air background was considered.

Table 5.1 shows that the application of NIR based calibration modelling followed by residuals modelling through the application of process data gave the best and most consistent results. Furthermore, whilst spectroscopy measurements provided insight into product concentration variations, the fusion with process data provided additional information regarding process changes. In all the cases, it was observed that data fusion provides better results than when using only one block of information.

To obtain an overall assessment of the performance observed in Table 5.1, a stacked bar chart of individual batch errors is presented in Figure 5.9. It should be noted that batch 1 has been omitted for comparison purposes. Although the MIR calibration results were better than the NIR calibration results, the superiority of NIR calibration modelling with SWS wavelength selection followed by residual prediction using process data is evident.

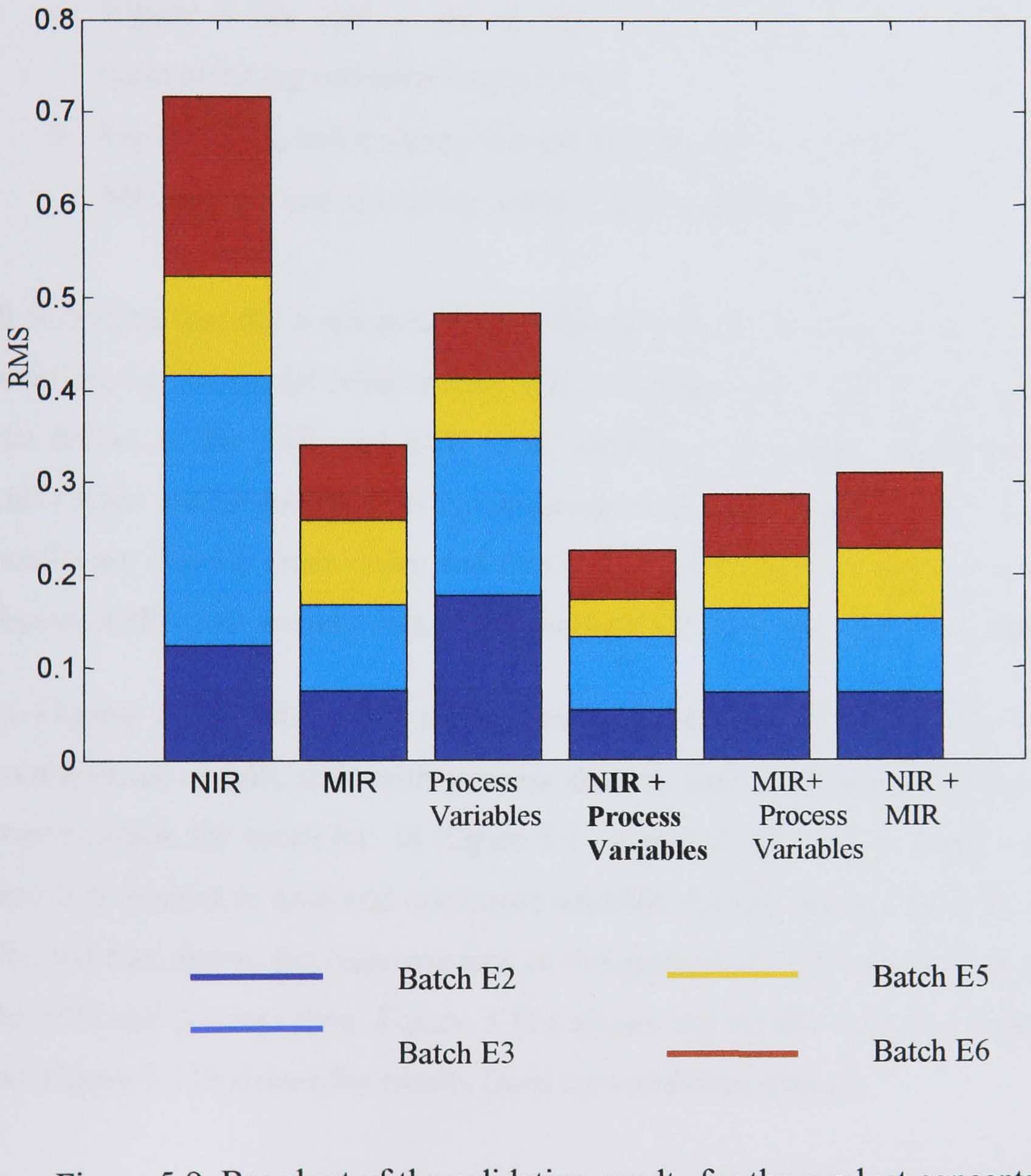


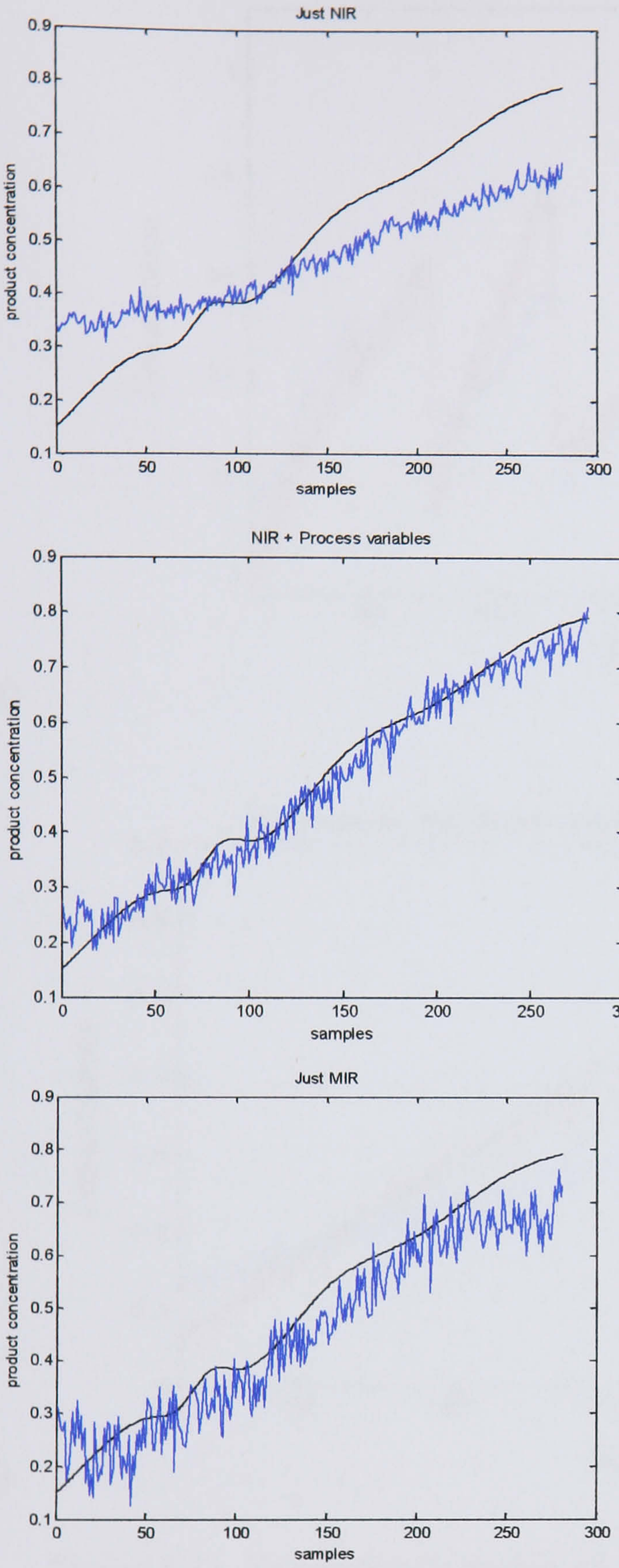
Figure 5.9. Bar chart of the validation results for the product concentration.

In Figure 5.10, the six final models for validation batch E5 are shown (i.e. row 4 in Table 5.2). More specifically on the left hand side are the predictions for the validation batch and on the right hand side the corresponding residuals:

- Figure 5.10a and b shows the model from the NIR spectra and the corresponding residuals respectively.
- Figure 5.10c and d shows the model from the conjunction of NIR spectra and the process data and the corresponding residuals respectively.
- Figure 5.10e and f shows the model from the MIR spectra and the corresponding residuals respectively.
- Figure 5.10g and h shows the model from the conjunction of NIR spectra and MIR spectra and the corresponding residuals respectively.

It is evident that the combination of different sources of information provide the best solution, i.e. the model created with NIR and process data and the one generated from the fusion of the NIR and MIR data. Moreover, it can be concluded that for the calibration model for MIR in combination with NIR and for MIR on its own, the results are slightly more noisy and this is the reason that the model created from the fusion of NIR and process data has a smaller RMS error and thus is preferable.

In Figures 5.11 and 5.12 a more detailed examination of a model based on the combination of NIR data with process data is shown for the prediction of product concentration for batch E1. In Figure 5.11 a global linear PLS model based on NIR data is presented in blue and compared with the splined assay values shown in black. The red line shows the improvement in the predictions that result from the fusion of the NIR and process data. Figure 5.11a shows the results from the training data set and Figure 5.11b shows the results from the validation data set.



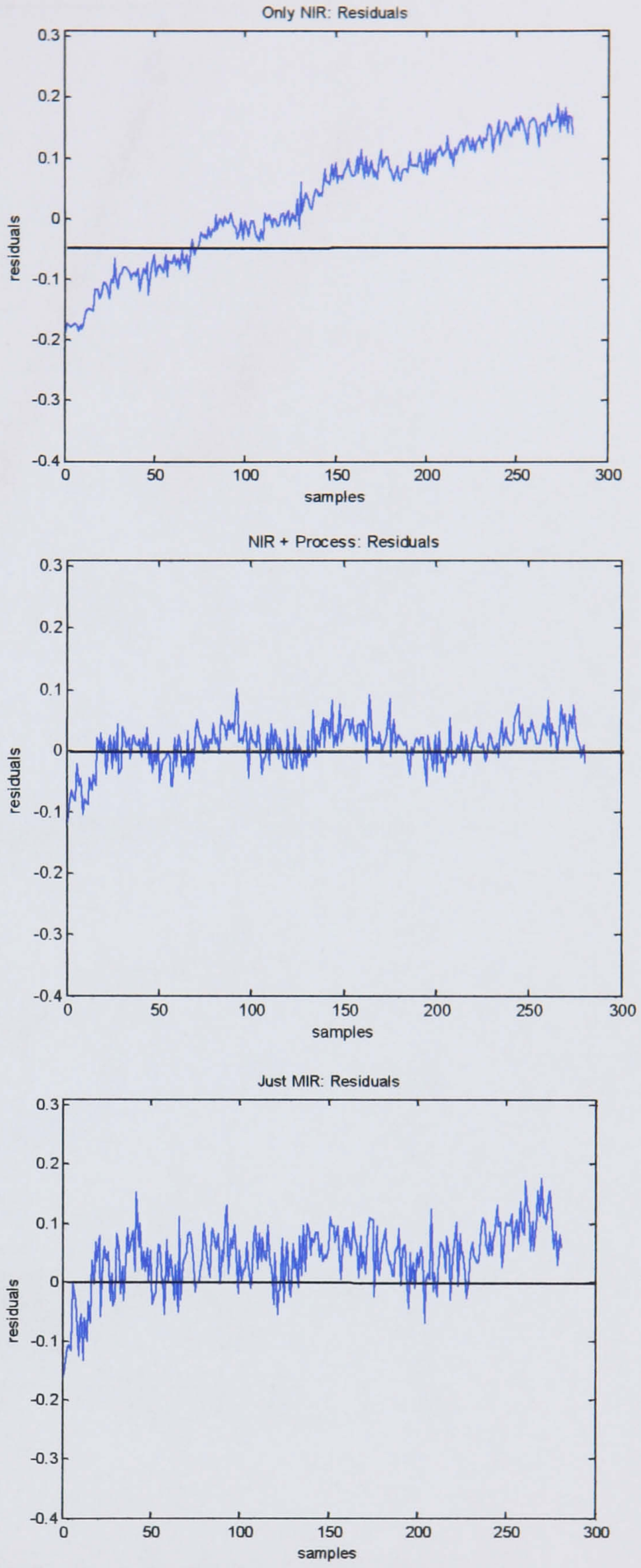
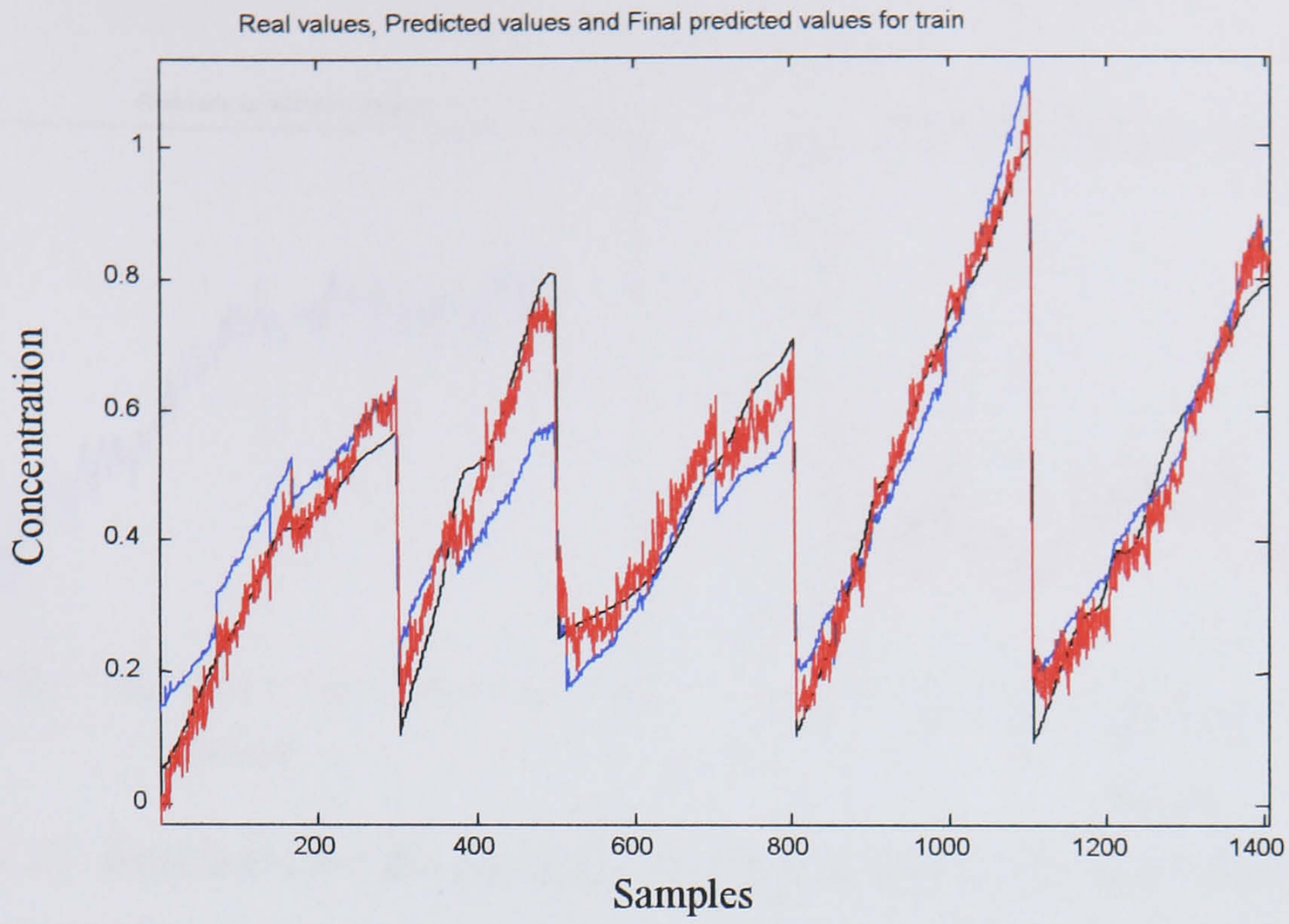
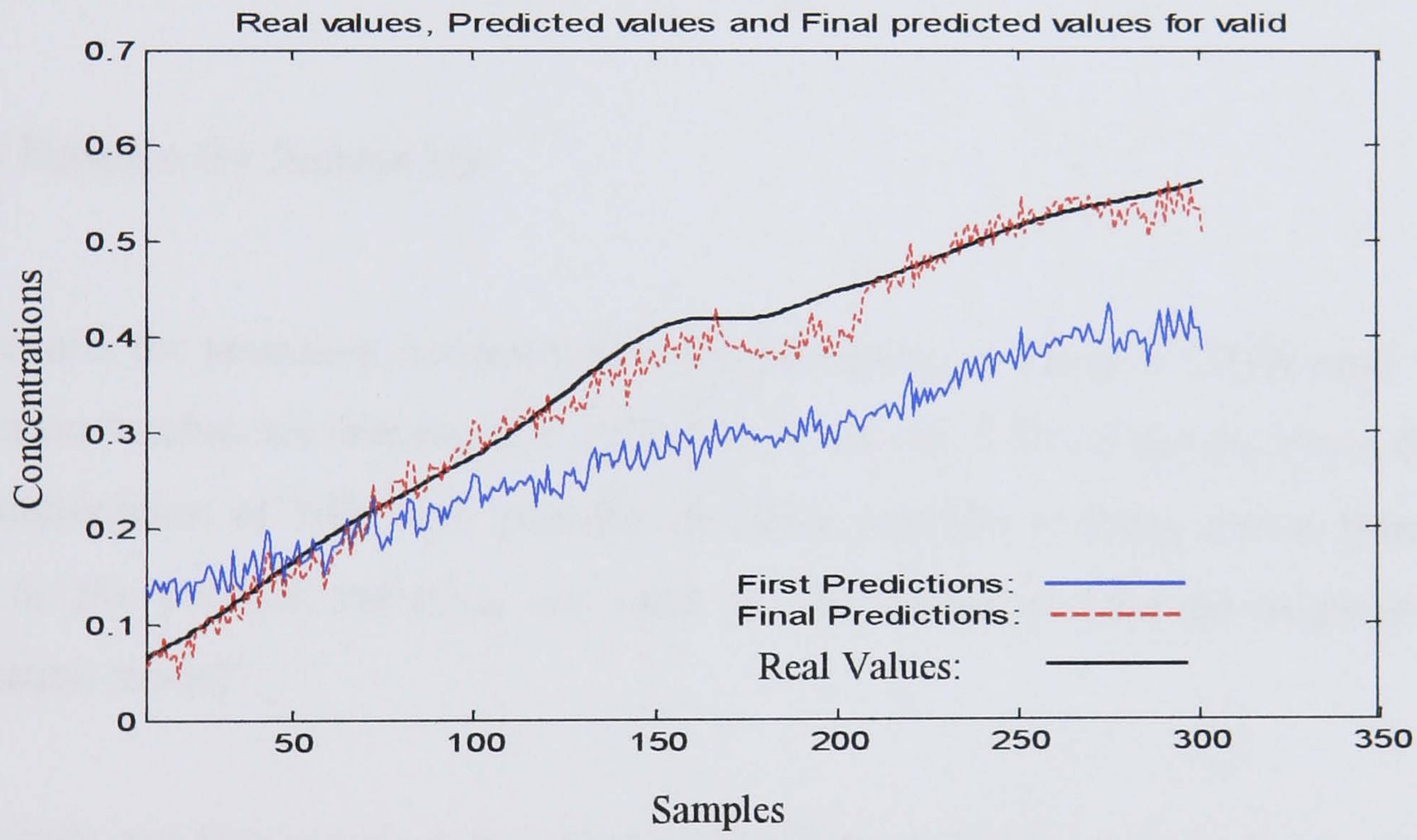


Figure 5.10. Final models, for batch E5 and the corresponding residuals.



(a)



(b)

Figure 5.11. Example of a model for the product concentration based on the combination of on-line process variables and NIR data for batch E1, where: real values (black line), calibration model based on the NIR data (blue line), calibration model based on the NIR and process data fusion (red line).

In Figure 5.12 the results are presented for the model residuals when only NIR data is used to develop the calibration model (left plot) and the residuals from the calibration model generated from the combination of data (right plot). For the NIR predictions only, the residuals have clear structure, whilst for the data fusion model the residuals exhibit less structure indicating a more accurate model.

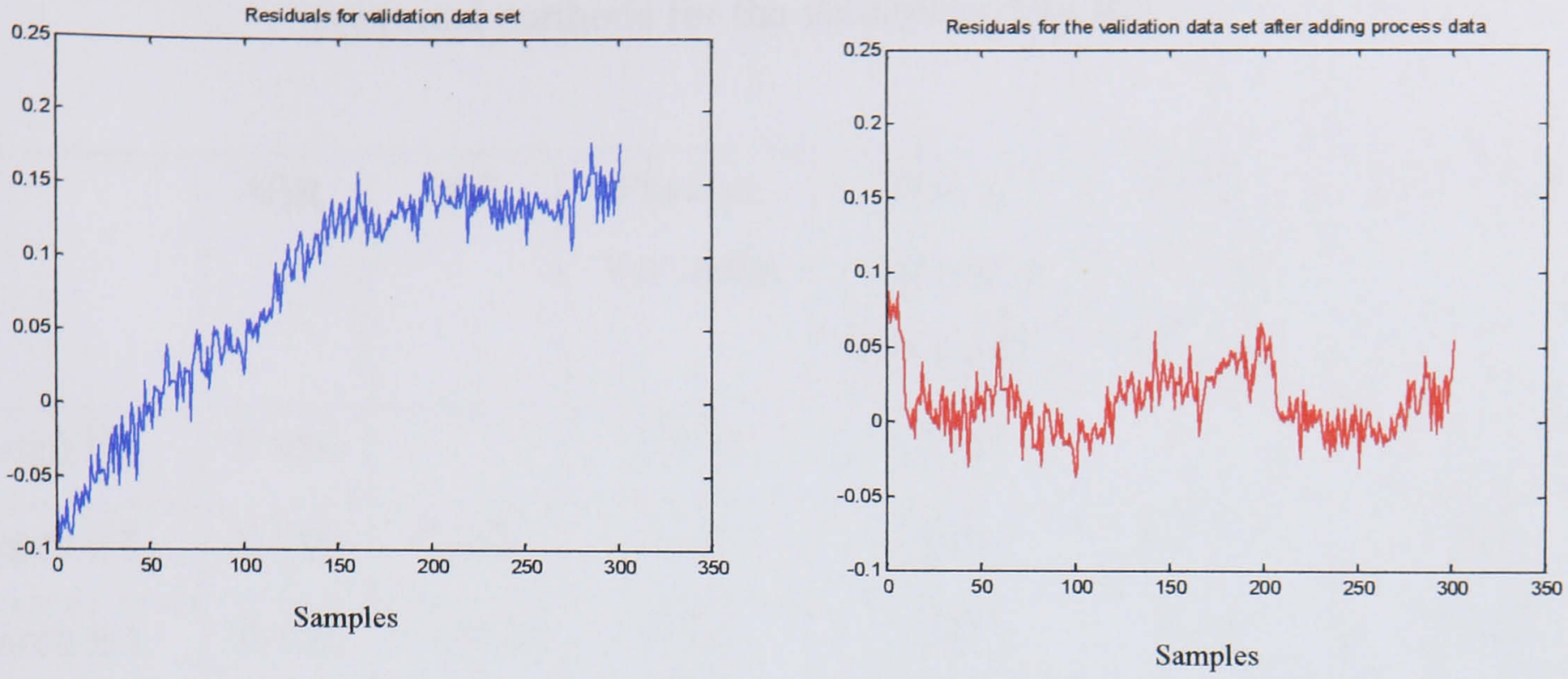


Figure 5.12. Residuals for the product concentration based on the calibration model from (a) the NIR data (left plot) and (b) process variables and NIR data fusion (right plot) for batch E1.

5.5.5 Results for Ammonia

The results for ammonia concentration determination in terms of RMS error for the validation batches are presented in Table 5.2. From this Table it can be concluded that the combination of NIR with process variables provides a better model than when MIR or the process variables are used on their own for the development of a calibration model.

The results are also reported as a stacked bar chart in Figure 5.13. It is interesting to observe from Table 5.2 that the NIR calibration model can be enhanced with MIR information. Again as for the product concentration, the combination of NIR and MIR gave better results than just using NIR or MIR alone.

Table 5.2. Results for the ammonia after the application of the traditional and proposed methods for the validation data set.

	NIR	MIR	Process Variables	**NIR + Process Variables**	MIR + Process Variables	NIR +MIR
Batch E1	0.054		0.088	**0.052**	-	
Batch E2	0.059	0.062	0.151	**0.041**	0.062	0.058
Batch E3	0.123	0.109	0.161	**0.089**	0.082	0.082
Batch E5	0.073	0.077	0.117	**0.036**	0.058	0.059
Batch E6	0.131	0.117	0.132	**0.067**	0.089	0.089

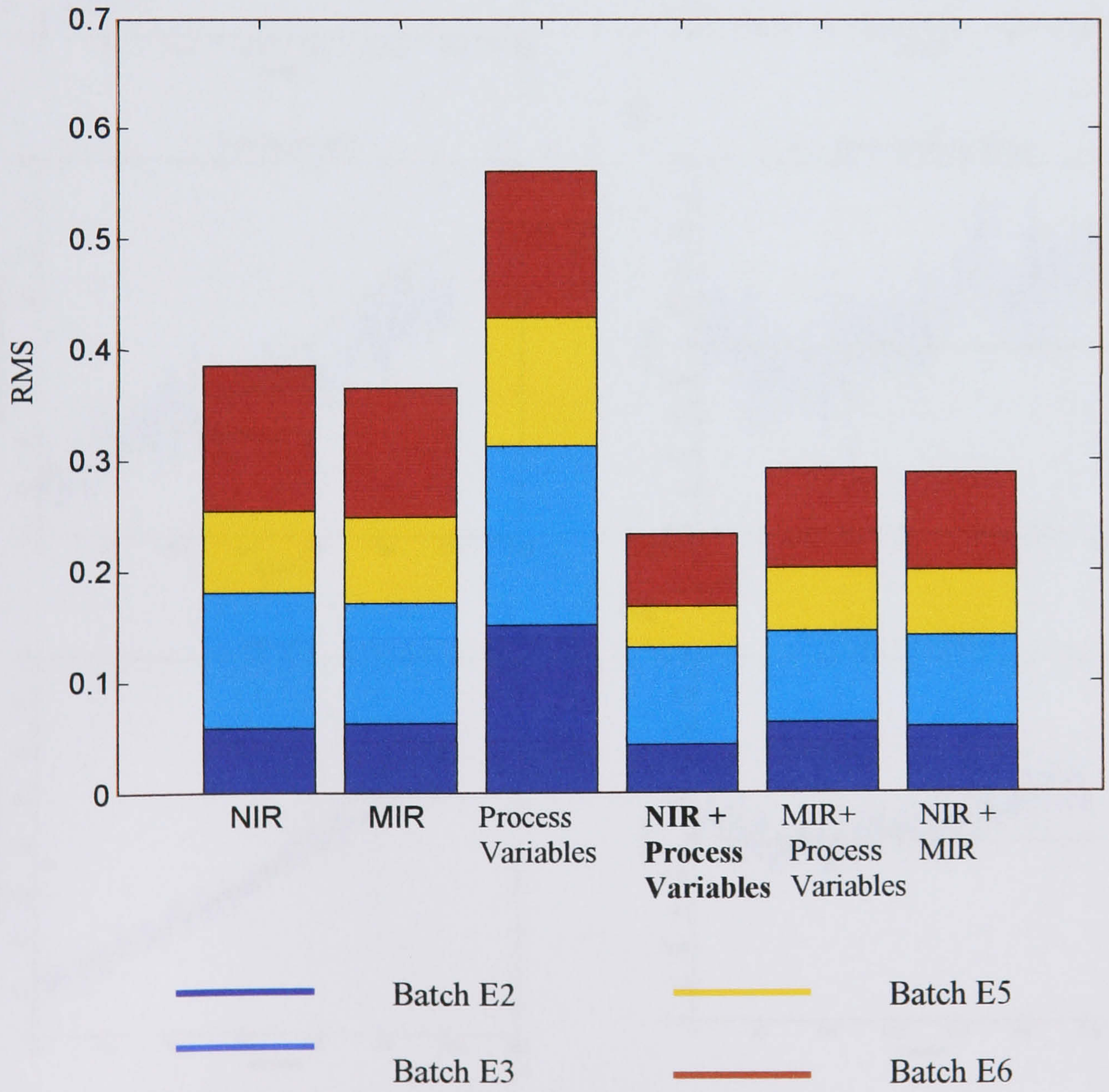
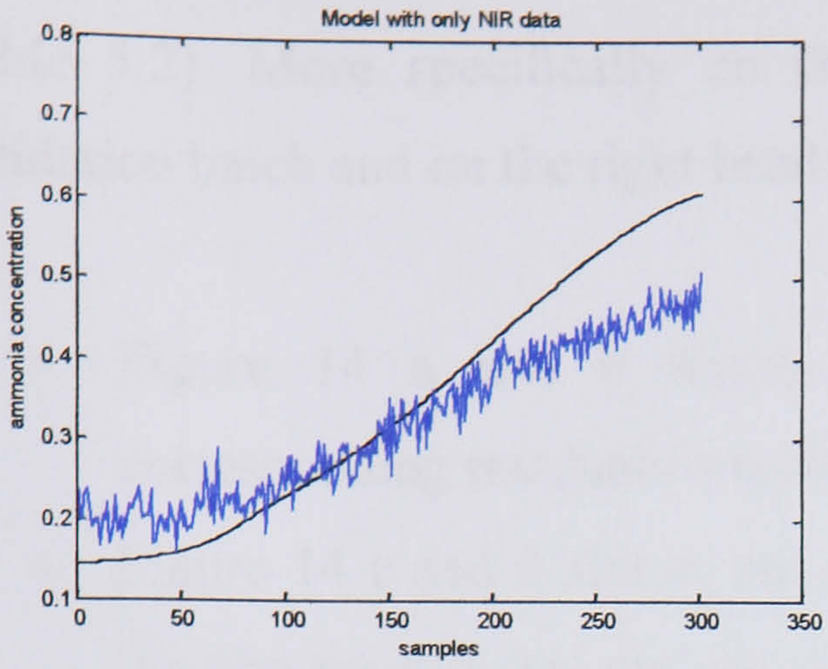
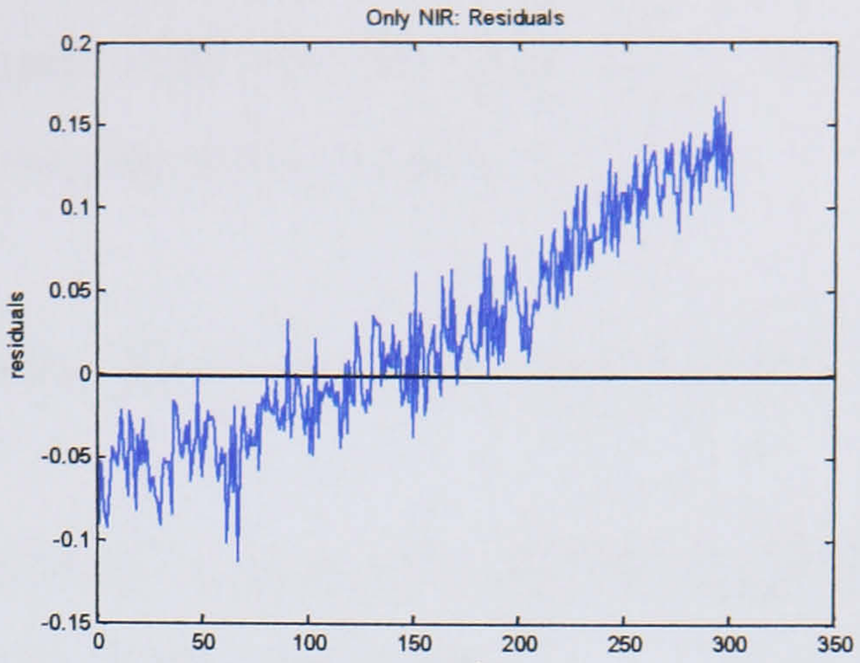


Figure 5.13. Bar chart of the validation results for ammonia concentration.

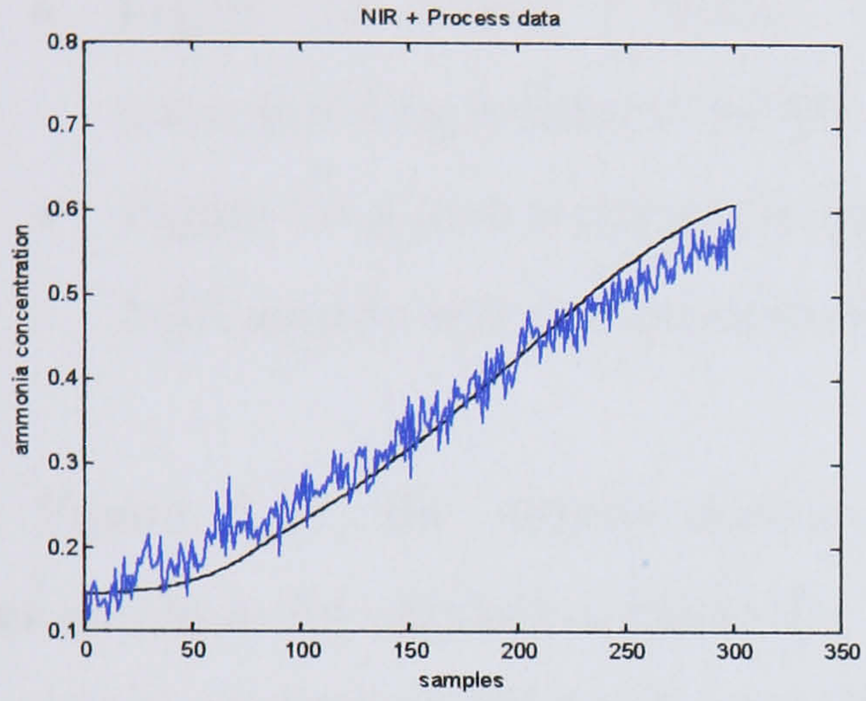
(a)



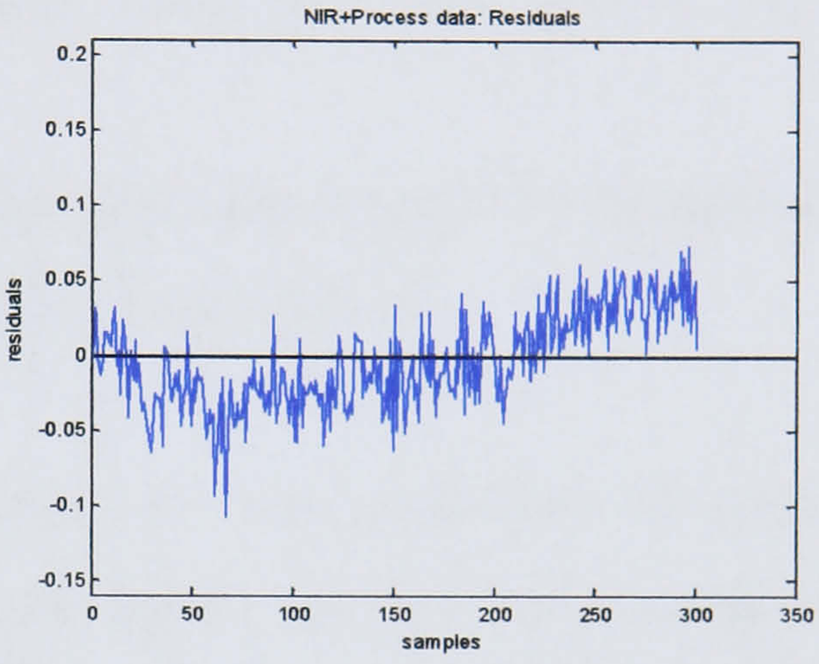
(b)



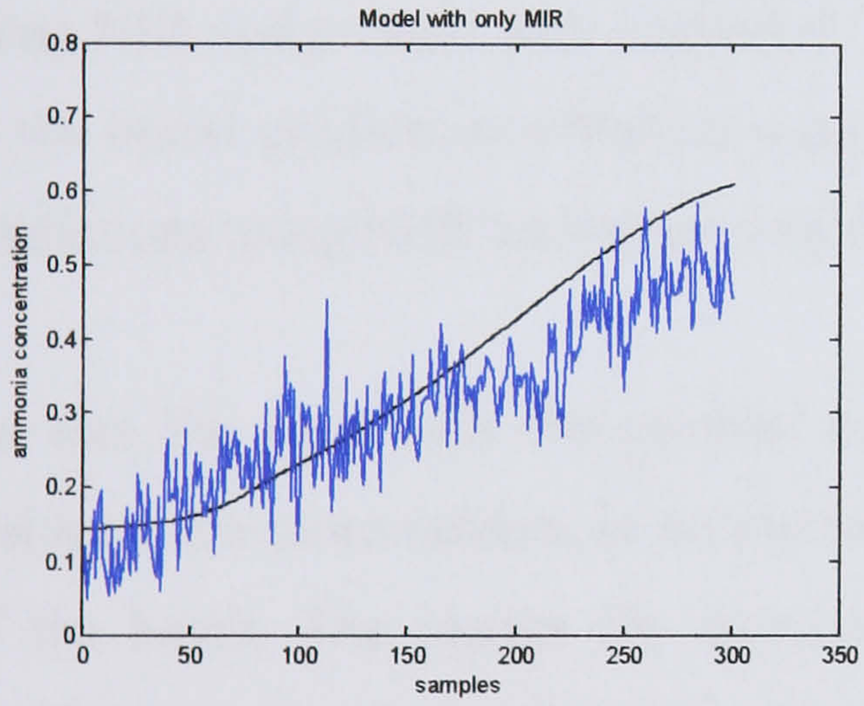
(c)



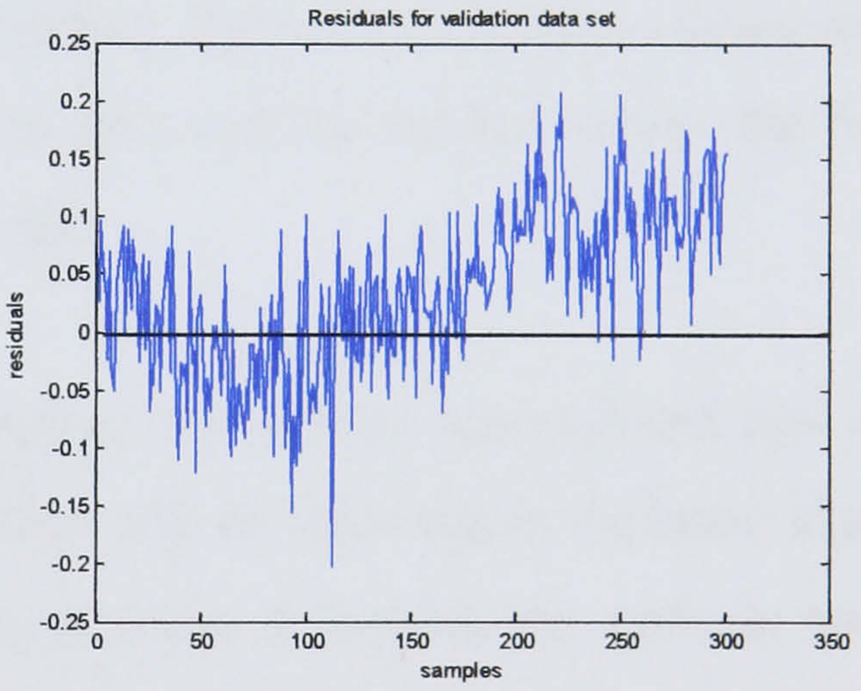
(d)



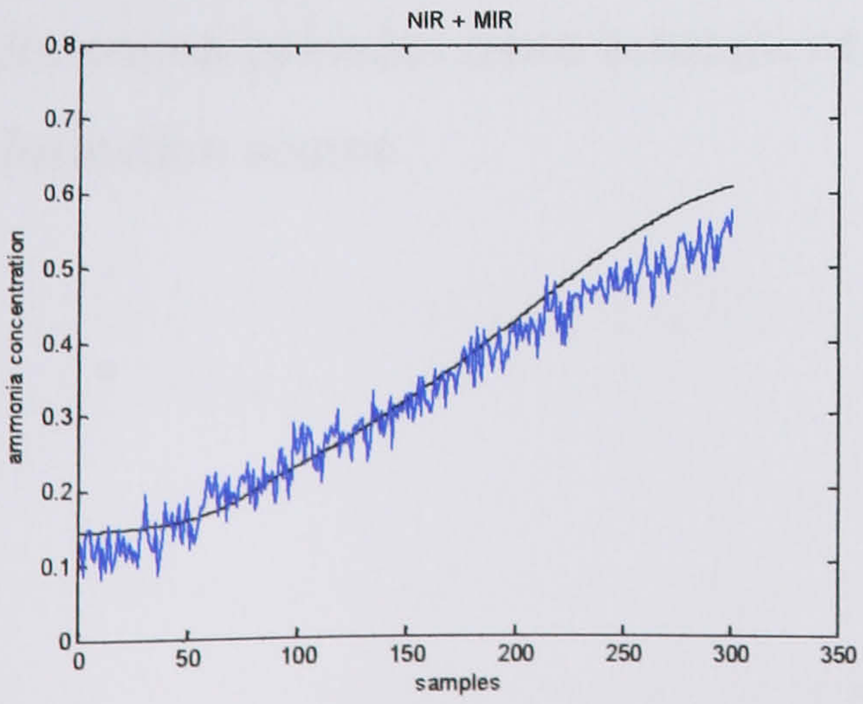
(e)



(f)



(g)



(h)

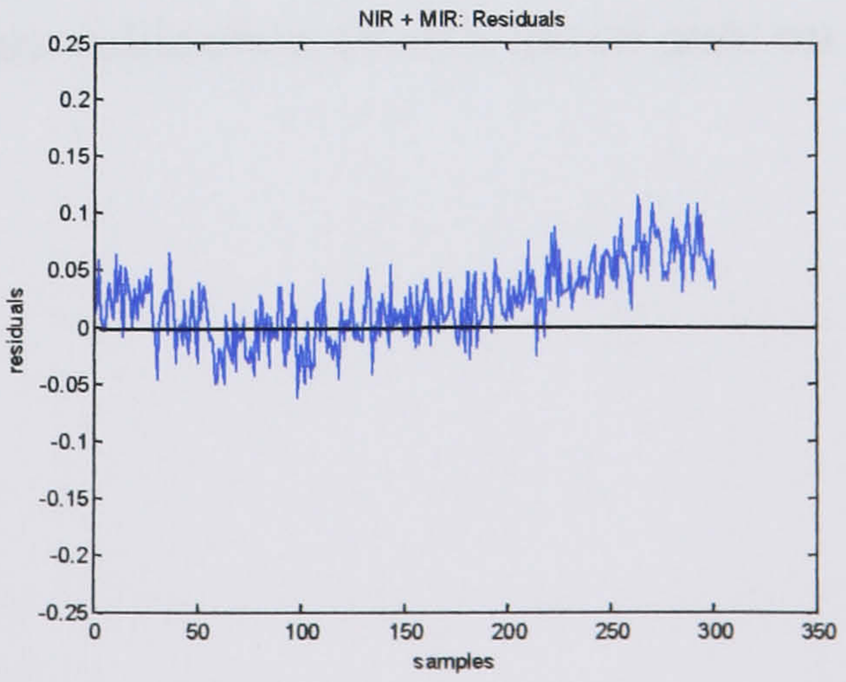


Figure 5.14. Final models for ammonia for batch E5 and the corresponding residuals.

In Figure 5.14, the six final models for validation batch E5 are shown (i.e. row 4 in Table 5.2). More specifically on the left hand side are the predictions for the validation batch and on the right hand side the corresponding residuals:

- Figure 14 a and b shows the model from the NIR spectra and the corresponding residuals respectively.
- Figure 14 c and d shows the model from the conjunction of NIR spectra and the process data and the corresponding residuals respectively.
- Figure 14 e and f shows the model from the MIR spectra and the corresponding residuals respectively.
- Figure 14 g and h shows the model from the conjunction of NIR spectra and MIR spectra and the corresponding residuals respectively.

In Figure 5.15, the improvements are shown for the prediction of ammonia concentration for validation batch E5. In the top figure, the blue line corresponds to the initial predictions obtained using NIR and the red line shows the final predictions using NIR and process data combined. In the bottom figure, the blue line corresponds to the initial predictions obtained using Process data and the red line shows the final predictions using MIR and process data combined.

As was the case with the product concentration, the offsets are reduced and the residuals are more random in nature but an offset still exists towards the latter stages of the batch. The results for ammonia concentration determination provide more justification for the claim that the combination of different sources of process information provides more accurate results than calibration models based only on one information source.

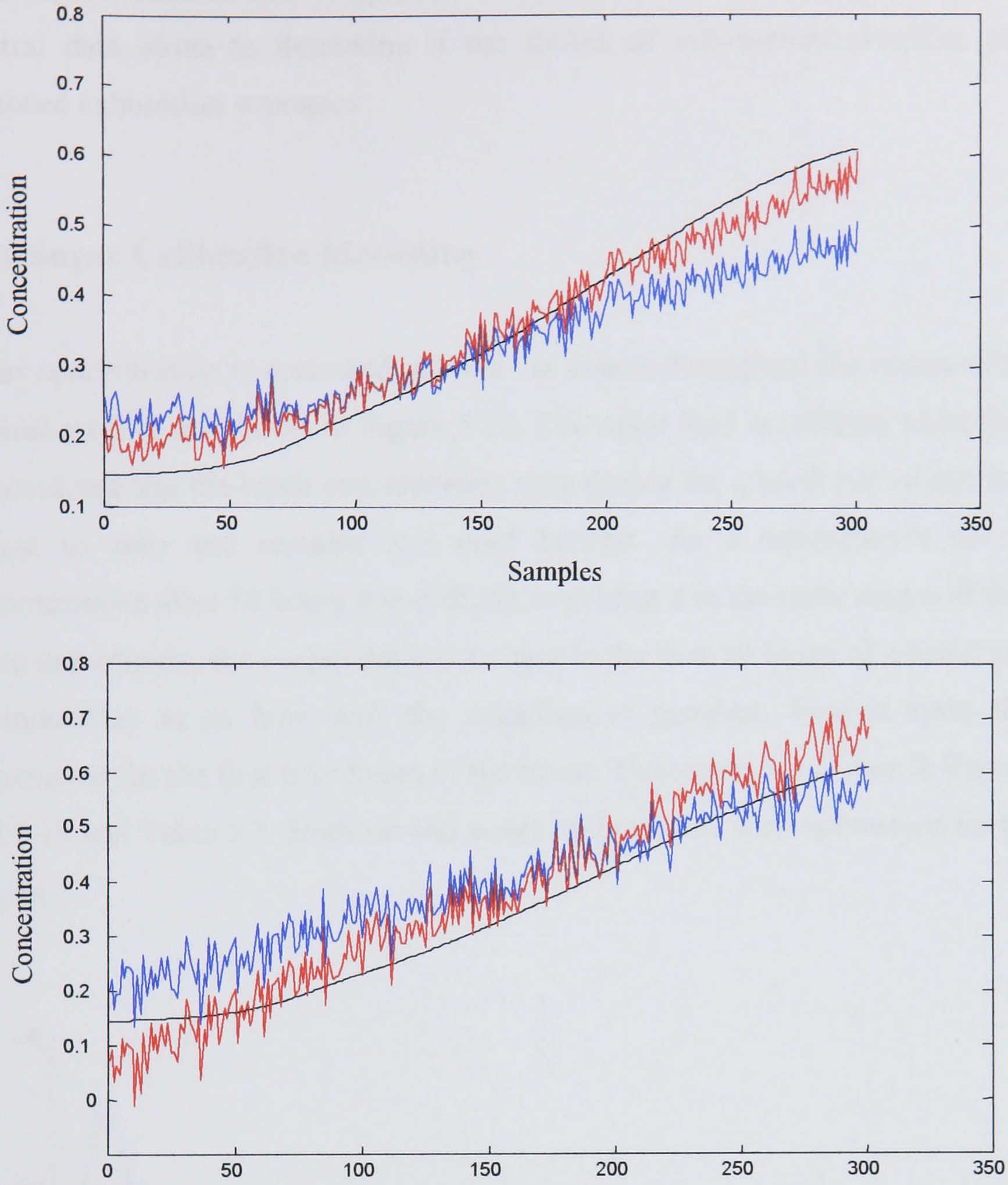


Figure 5.15. Example of a model for the ammonia concentration based on the combination of NIR and process data (top figure) and MIR and process data (bottom figure).

5.6 Application of the Calibration Model Strategy to the Biochemical Components

The generation of calibration models for three additional biochemical concentrations (sugar, lipids and phosphate) was considered using the proposed algorithm. They were applied to the same DOE data as used in section 5.5. For these components, the removal of the background was an important component of the analysis composed. In

addition, the combination of spectral and process data is compared with using the spectral data alone to determine if the fusion of information provides generally enhanced calibration strategies.

5.6.1 Sugar Calibration Modelling

Sugar concentration is measured off-line via assays throughout the course of a batch. Typical profiles are shown in Figure 5.16. The sugar feed is initiated when the vessel is inoculated and the batch concentration rises during the growth period and then falls almost to zero and remains low until harvest. As a consequence of the low concentrations after 50 hours it is difficult to predict it in the latter stages of the batch. More importantly, the accumulation of sugar in the first 50 hours of a batch may give an indication as to how well the organism is growing. Models were therefore constructed for the first fifty hours of the batch. The results are shown in Figures 5.17 and 5.18 and Table 5.3. Both air and water backgrounds were subtracted for the MIR results.

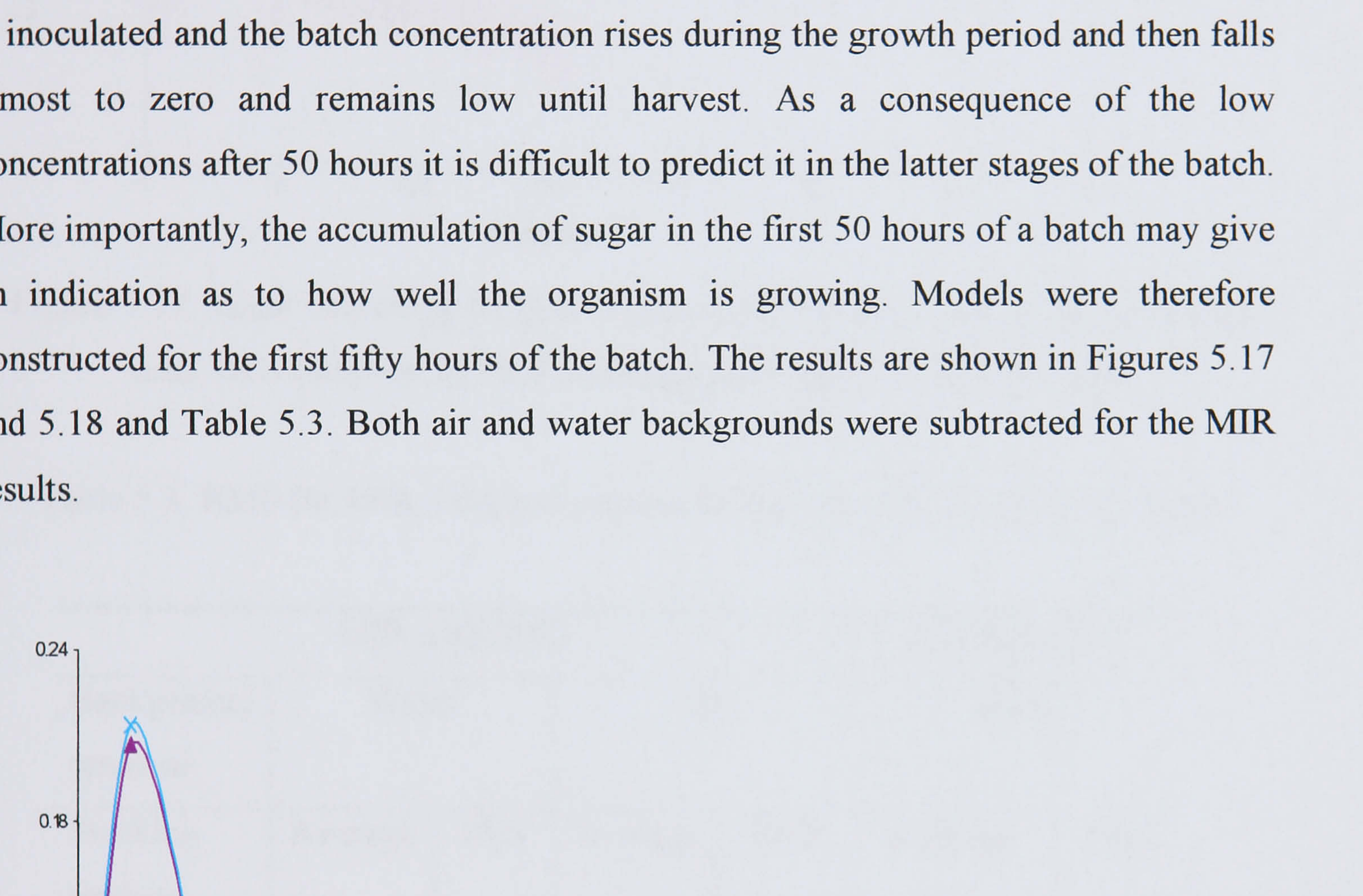


Figure 5.16. Typical sugar profile for the DOE data.

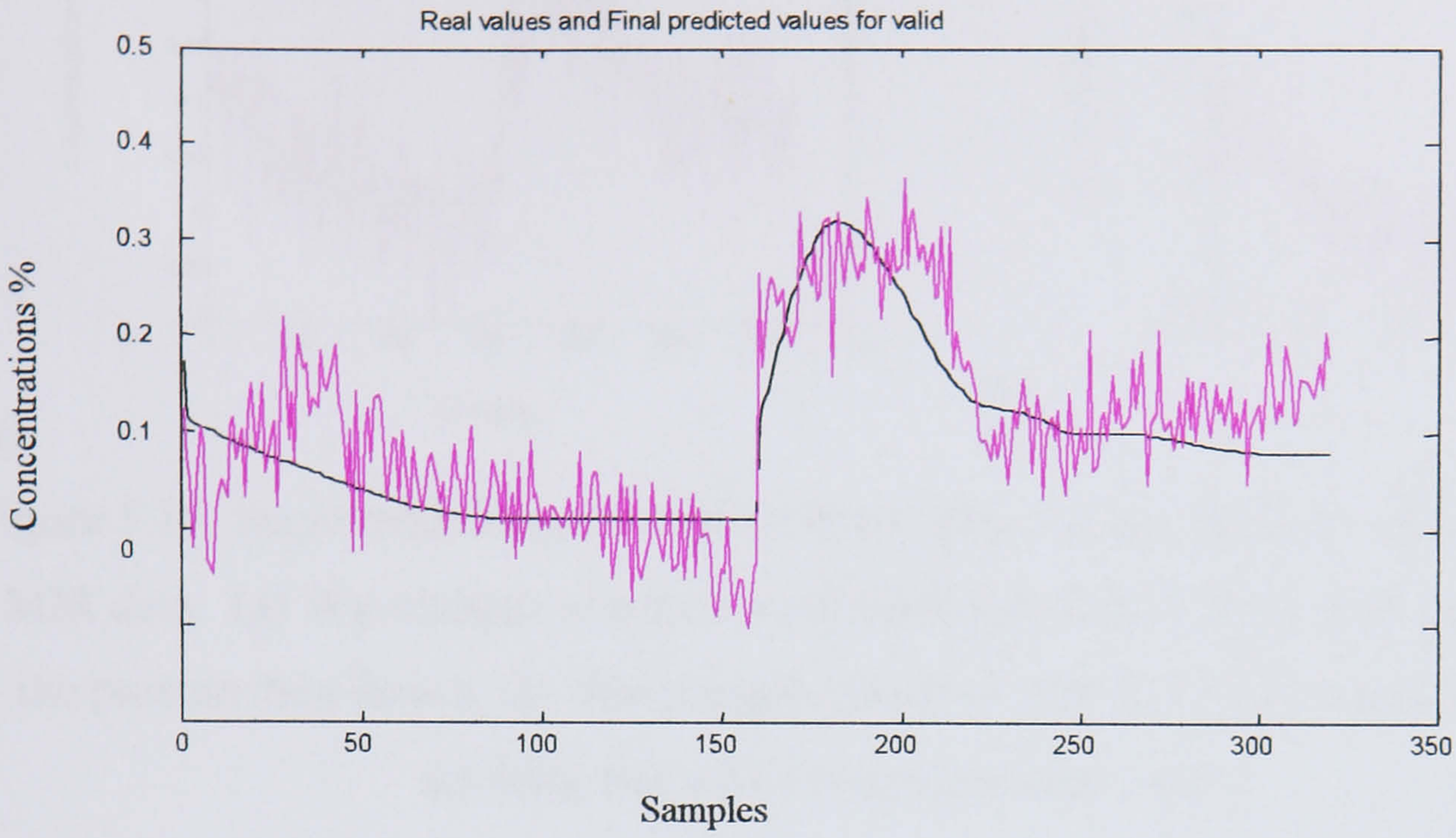
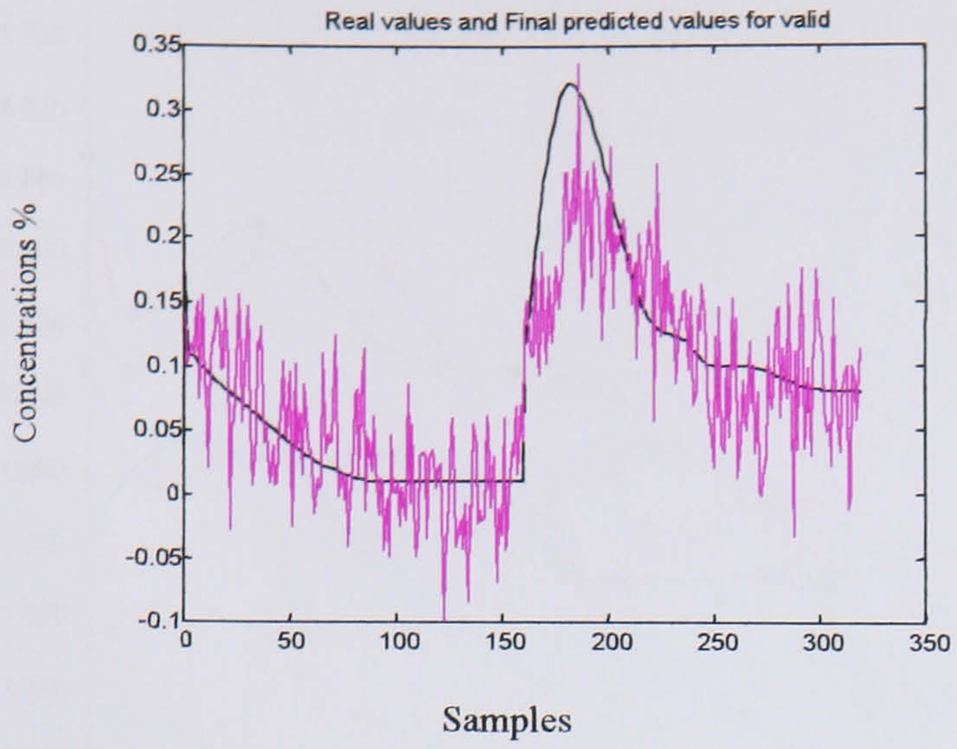


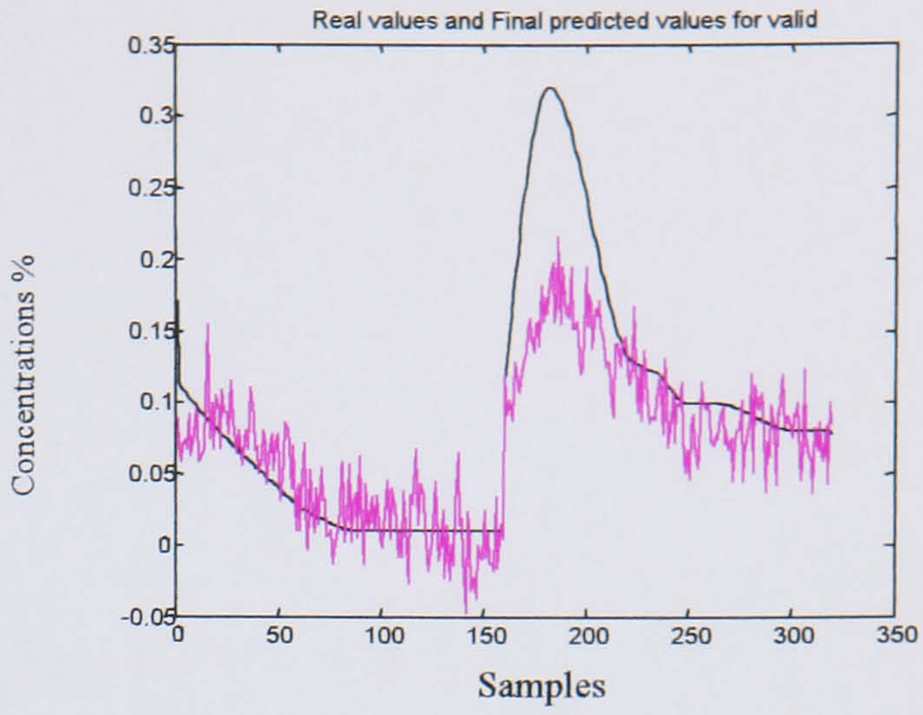
Figure 5.17. Sugar modelling for first 50 hours for NIR data (wavelength selection with SWS followed by PLS stacking) and with process data fusion.

Table 5.3. RMS for MIR, NIR and process data for the first 50 hours for sugar.

MIR with SWS					NIR with SWS	
Background removal	Water		Air		Black	
Stacking method	Average	PLS	Average	PLS	Average	PLS
Train	0.100	0.090	0.100	0.088	0.097	0.086
Valid	0.056	0.066	0.059	0.069	0.096	0.073
MIR + Process					NIR + Process	
Train	0.086	0.080	0.087	0.075	0.087	0.081
Valid	0.050	0.052	0.050	0.053	0.062	0.057



(a)



(b)

Figure 5.18. Sugar modelling for first 50 hours after Air Background Subtraction with MIR data. (a) Wavelength selection with SWS followed by PLS stacking and with the process data fusion, (b) Wavelength selection with SWS followed by Average stacking and with the process data fusion.

Table 5.3 confirms the previous observations. Benefits are to be found in all cases by enhancing the spectral data prediction with process data. Again there is no significant difference between the use of air or water background subtraction for the MIR data analysis. PLS and average stacking are comparable due to the consistency of performance across both the training and validation batches. Finally, in all cases, the prediction of sugar concentration can be seen to be reasonable with MIR slightly outperforming NIR but the predictions are corrupted by significant levels of noise. In a practical application, this noise could be reduced by data filtering.

5.6.2 Lipid Modelling

Lipid is batched in with the pre-inoculation medium and its utilisation is slow initially. Lipid concentration generally falls throughout the course of the batch. Typical lipids profiles are shown in Figure 5.19. Results from the construction of the calibration models are presented in Tables 5.4 and Figure 5.20 and the comparison of model building decisions follow the same procedures as those for sugar determination.

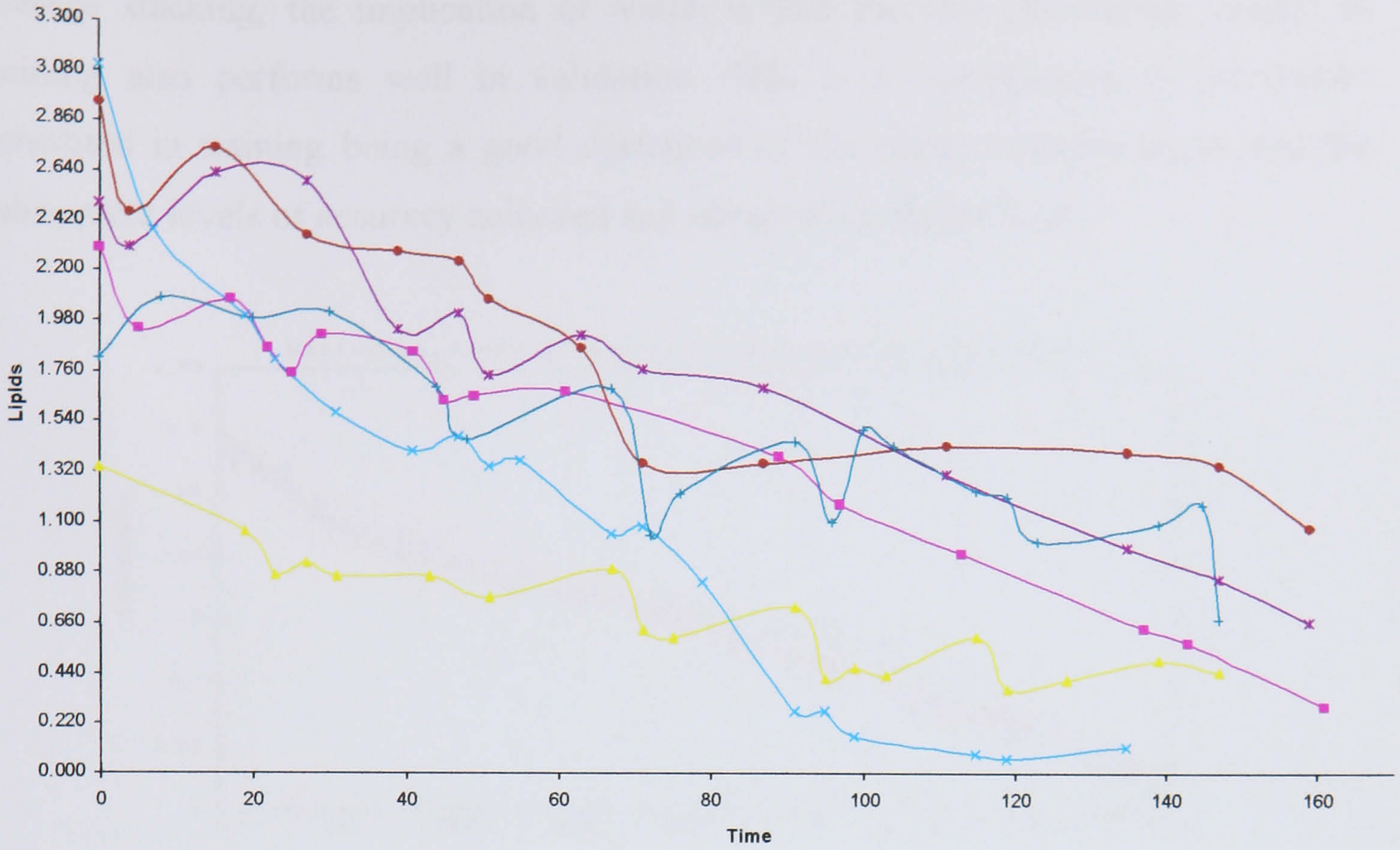


Figure 5.19. Typical Lipids profile for the DOE data.

Table 5.4 RMS for MIR, NIR and Process data for Lipids.

MIR with SWS					NIR with SWS	
Background removal	Water		Air		Black	
Stacking method	Average	PLS	Average	PLS	Average	PLS
Train	0.37	0.25	0.32	0.25	0.20	0.14
Valid	0.46	0.37	0.60	0.42	0.47	0.29
MIR + Process					NIR + Process	
Train	0.40	0.27	0.35	0.26	0.18	0.14
Valid	0.50	0.37	0.61	0.41	0.42	0.26

A number of conclusions can be drawn from Table 5.4. Firstly, NIR gives more accurate results than MIR and this accuracy is not further enhanced by the incorporation of process data. This is contradictory to the previous findings but is due to the fact that there are minimal offsets in the NIR model hence the addition of process data does not give further benefit. In this case PLS stacking is preferable to

average stacking, the implication of which is that the best performing models in training also performs well in validation. This is a consequence of the model generated in training being a good descriptor of the system characteristics and the subsequent levels of accuracy achieved and observed in Figure 5.20.

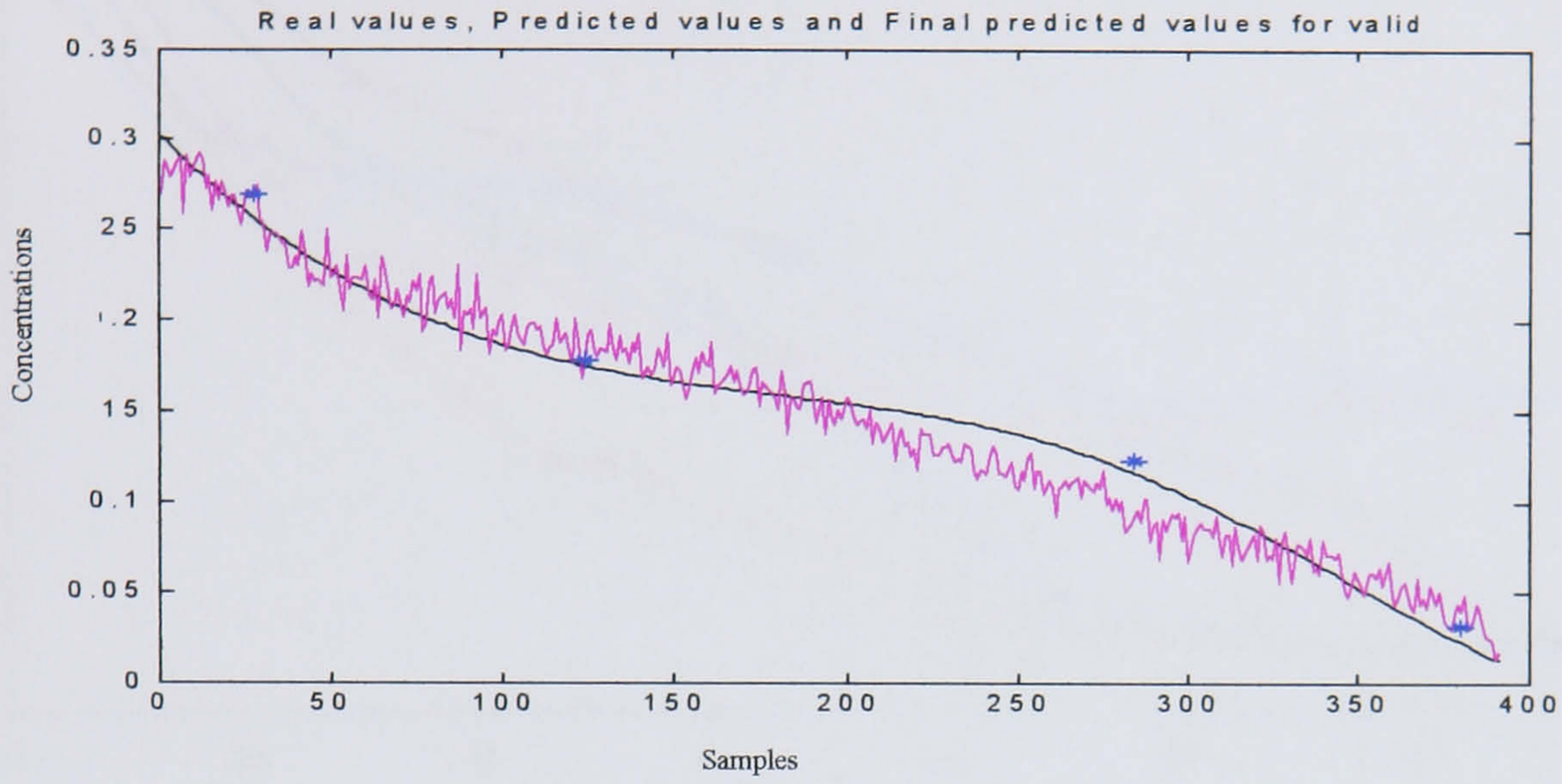


Figure 5.20. Predictions with NIR data after splining and SWS and Process data sequential addition where '*' the assay values.

5.6.3 Phosphate Modelling

Phosphate is batched in with the pre-inoculation medium. It is utilised quickly during the growth phase of the fermentation and the concentration falls to low levels typically around 100 hours into the batch. Typical phosphate profiles can be seen in Figure 5.21. Results from calibration modelling using MIR and process data are shown in Figure 5.22 and Table 5.5. The results from phosphate determination using NIR data confirmed that it could not be detected in the NIR region.

In this case, the calibration model generated with the MIR data can be improved through the incorporation of process data. Average and PLS stacking perform in a comparable manner and again the water and air background removal gave similar results. The prediction of concentration are corrupted by noise arising from the second derivatives being taken in the spectral data pre-treatment and data filtering would be required for on-line implementation.

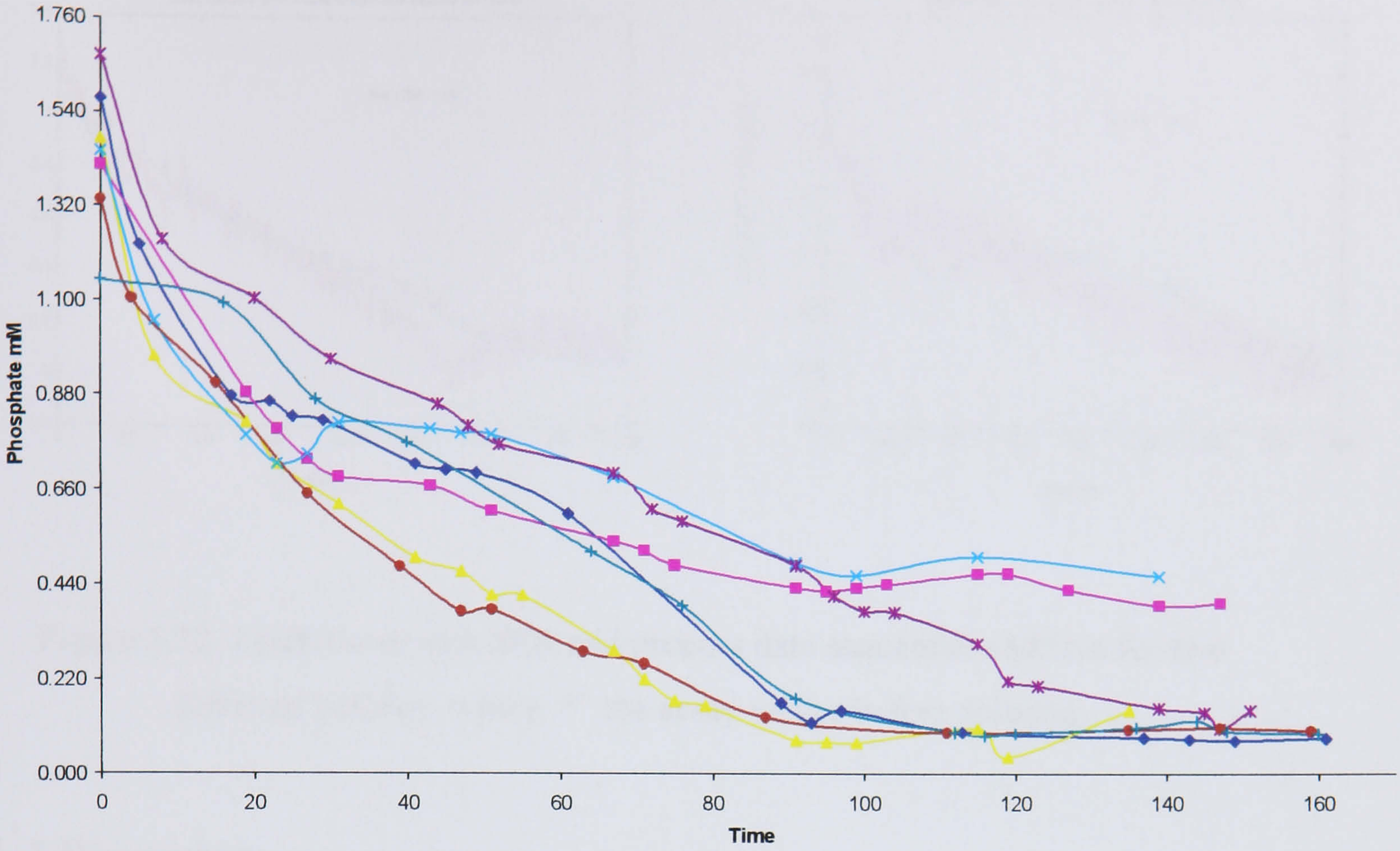
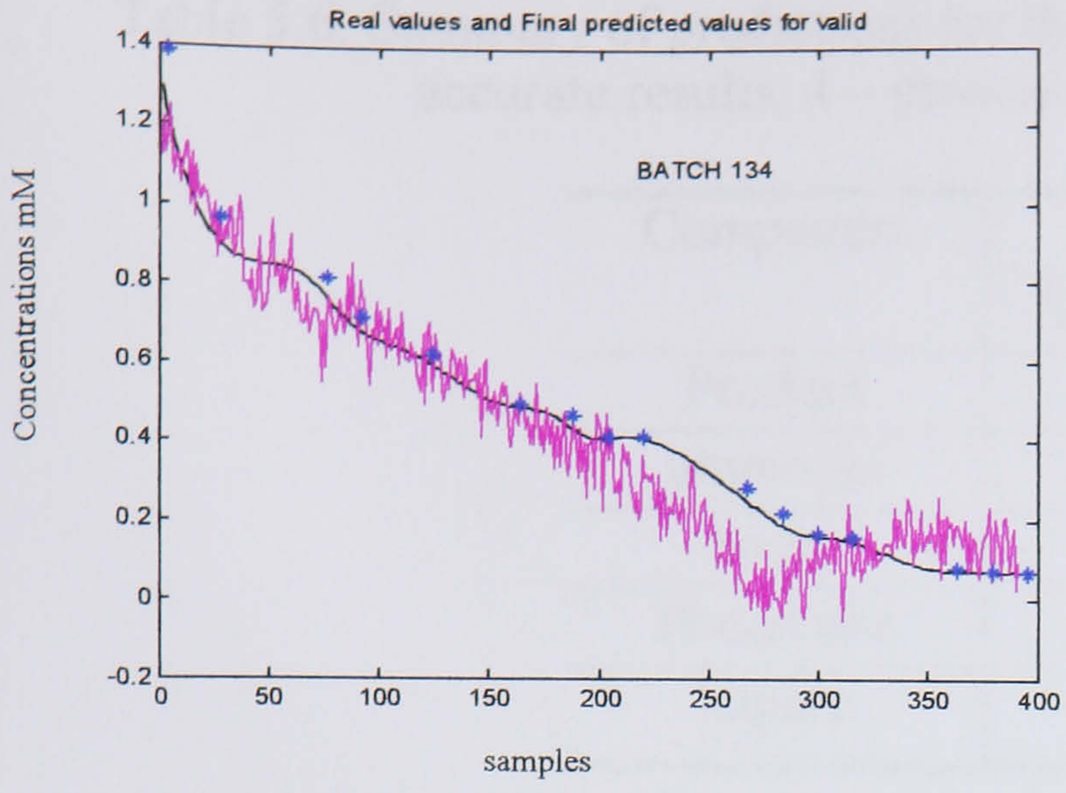


Figure 5.21. Typical phosphate profile for the DOE data.

Table 5.5. RMS for MIR and Process data for Phosphate.

MIR with SWS				
Background removal	Water		Air	
Stacking method	Average	PLS	Average	PLS
Train	0.17	0.12	0.16	0.13
Valid	0.18	0.27	0.18	0.27
MIR + Process				
Train	0.11	0.12	0.11	0.12
Valid	0.20	0.19	0.20	0.19



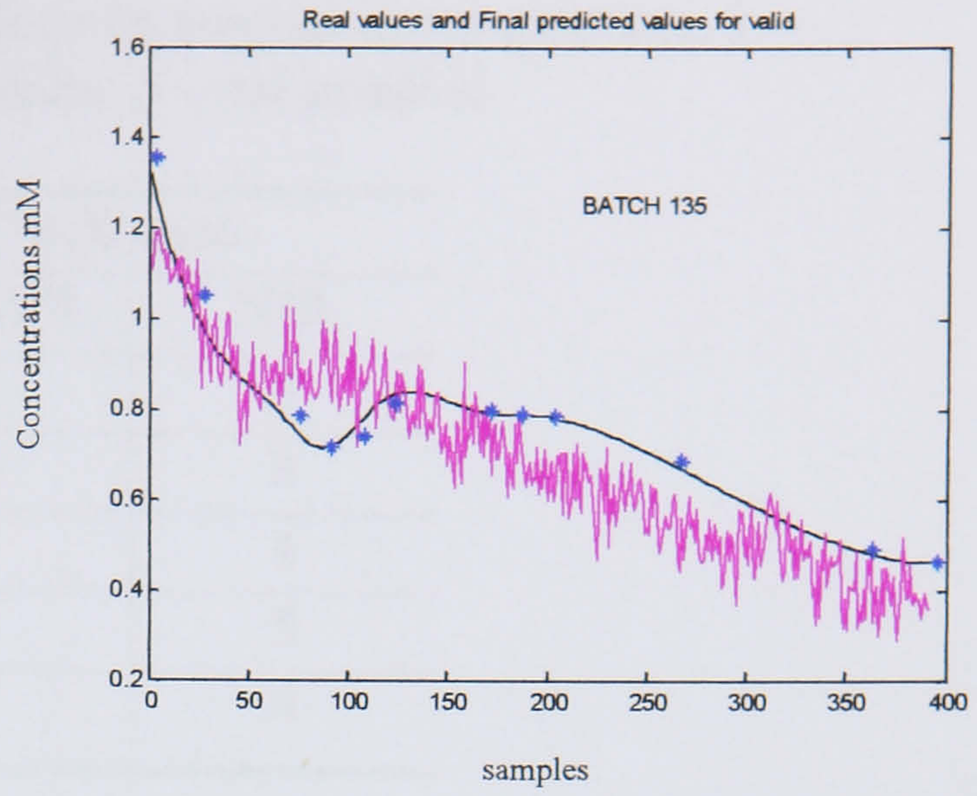


Figure 5.22. Predictions with MIR and process data sequential addition for two different batches, where '*' the assay values before splining.

5. 7 Discussion

In Chapter 5 the complementary use of spectroscopy and process data for statistical process monitoring was considered. Signals from spectral instruments can potentially be enhanced by other process measurements, to provide on-line indications of critical broth concentrations. Such components are traditionally only available from infrequent off-line analysis. In this Chapter it has been shown that it is indeed possible to improve spectral instrument calibration models through the incorporation of other process measurements. Data from different sources contain different information and together they produce a better calibration model. The traditional approaches for the combination of different sources of information together with their limitations were described. A new strategy was developed that combines the different types of information in a sequential manner to avoid any heterogeneity issues for the generation of robust calibration models.

The spectral process monitoring methodology was demonstrated by an application to the industrial antibiotic production process of a DoE data set for the determination of the concentration of a number of different biochemical components. A summary of results can be seen in Table 5.6.

Table 5.6. Summary of predictions for the various fermentation components, 3 - accurate results; 4 – promising results; 5 - not possible.

Component	Technique	
	ATR MIR	NIR
Product	3	3
ammonia	3	3
Sugar	4	4
Phosphate	4	5
Lipids	4	4

In all cases the results from data fusion were better than those using only one source of information:

a) For product and ammonia accurate results were obtained. The application of NIR based calibration modelling followed by residuals modelling through the use of process data gave the best and most consistent results.

b) For sugar concentration, MIR fused with process data slightly outperformed NIR and process data but the predictions were corrupted by significant levels of noise.

c) For lipids, NIR fused with process data gave more accurate results than MIR although in the case when the calibration model is offset free, the fusion of spectral and process data did not improve the results.

d) Finally, for phosphate the calibration generated with MIR data can be improved by the incorporation of process data.

In summary it was concluded that signals from spectral instruments can be enhanced by other process measurements by adopting the new sequential strategy described in Chapter 5. With the traditional methodology described in section 5.3, the heterogeneity of the problem may fail to be addressed and it is necessary to scale the two data sets according to their relative significance in terms of describing the process. Sequential data fusion modelling approach described in section 5.4 managed to overcome these issues.

Chapter 6 will report the contributions and the summary of findings described in all the previous Chapters.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

The aim of the thesis was to investigate strategies for the on-line application of spectroscopic measurements in bioprocess applications. As a consequence, contributions have been made to the fields of fermentation monitoring and control and robust spectral calibration modelling. The use of spectroscopy as a potential source of high-quality chemical information for process analysis and on-line monitoring offers major opportunities. However, the interpretation of spectral information is not straightforward as a result of the large number of variables (wavelengths) and the presence of components that exhibit overlapping absorbance features. The successful application of spectroscopic instrumentation therefore requires the application of multivariate data analysis techniques to extract the latent features from the data. A new approach based on spectral window selection (SWS) followed by model stacking was proposed that offers the opportunity for constructing a model from a sub-set of wavelengths. As an individual model can be too specific to the model building data, stacking, that is where multiple models are combined, was exploited to provide enhanced robustness and repeatability.

Process data can be viewed as complementary to the spectral measurements, potentially providing a more comprehensive description of the overall system variations. Whilst research into statistical methods for calibration model improvement have not been dismissed, the inclusion of additional process measurements offers an alternative approach for attaining improved and more robust models. A novel strategy has been developed that combines the spectral and process information in a sequential way. After the construction of the first calibration model from the spectral data, subsequent corrections to the spectral based prediction can be made thought the application of process data or alternative spectral measurements.

The application focus of the thesis was on fermentation and microbial antibiotic fermentation production processes in particular. For the accurate control of such fermentations, it is important to obtain frequent and precise product and biochemical components concentrations during the progression of the batch. Spectral instrumentation, such as infrared devices, offers the opportunity to measure on-line broth constituents and more importantly product concentration levels. Both invasive and non-invasive MIR and NIR probes were investigated and the utility of the calibration modelling strategy was assessed. The proposed methodology has been

shown to be effective for both NIR and MIR spectra and is not instrument supplier specific.

6.1 Overview of Findings

Whilst each Chapter reviewed specific findings, the following brings them together to draw overall conclusions:

Chapter 2: Spectroscopic measurements, in particular NIR and MIR, were discussed in this Chapter with a general overview of multivariate calibration being provided. The description of common linear and non-linear regression techniques such as linear and non-linear PLS and neural networks was included as a mathematical basis for the following Chapters. The methods utilise the full set of model input data in contrast to the approaches described in Chapter 3, which involved the elimination of wavelengths that do not contribute to describing the changes in the concentrations of the analytes of interest.

Chapter 3: The selection of informative spectral regions was discussed in this Chapter. Previous research has shown that wavelength selection can improve the results by reducing the contribution of overall noise from those regions not containing relevant information on the analyte concentrations. Existing methods were reviewed and their characteristics considered. The Spectral Window Selection (SWS) algorithm was proposed as a means of overcoming limitations of existing approaches. The key advantage of the new method is that it uses a search based spectral window selection algorithm to build a calibration model. The single model generated is not unique, with the random initialisation and search procedure resulting in quite a wide spread of performance. This problem can be alleviated by generating multiple models and combining these using stacking to produce a more robust prediction. The application to a benchmark case study of NIR spectra from diesel fuels demonstrated the functionality of the new strategy. The methods were compared with traditional approaches such as conventional PLS and wavelength selection algorithms such as genetic algorithms and iPLS. It was demonstrated that the selection of informative spectral regions can improve the results by reducing the contribution of overall noise

from those regions not containing relevant information on the analyte concentrations. Furthermore the GA did not indicate any particular regions and the selected wavelengths were scattered throughout the whole wavelength region.

Chapter 4: In this Chapter, it was demonstrated that infrared spectroscopic techniques can be used on-line to assess the concentration of key components in an industrial fermentation broth. The focus of Chapter 4 was a comparison between the different analytical methods and the various approaches for the interpretation of signals for calibration model construction. An industrial fermentation application for the production of an antibiotic formed the basis of this study. Non-invasive and invasive NIR and invasive MIR measurements were used for this investigation. The process is described and the data available from 'Design of Experiment' batches and 'Standard' batches outlined. Traditional calibration model building methods were compared with the new strategy. Developing a calibration model for the whole batch proved problematic, particularly towards the latter stages of the fermentation. A local modelling strategy where the batch was portioned into operating regions was investigated as a means to counter the limitations of global model with respect to capturing late stage batch performance. An issue with the local modelling strategy is the on-line partitioning of the fermentation into regions of 'common' characteristics. This was achieved by using process knowledge. The results of both global and local calibration modelling confirmed that the new calibration modelling strategy does indeed reduce the root mean square error of validation over that obtained using full spectrum analysis. Additionally it was observed that it outperformed interval PLS and the PLS calibration model developed from the wavelength selection using genetic algorithms.

Chapter 5: To overcome the limitations of the spectral calibration models this Chapter investigated whether it was possible for signals from spectral instruments to be enhanced by other process measurements. A new strategy was developed that combined the spectral and process information in a sequential manner for the generation of robust calibration models. To obtain more precise indicators of the analyte concentrations of interests, it was necessary to combine data from different sources. Signals from spectral instruments and other process measurements were combined, to provide on-line measurements of critical broth concentrations of a

number of biochemical components. Specifically, the data fusion strategy was applied for the construction of calibration models for product, ammonia, sugar, lipids and phosphate. In all cases the results from data fusion were better than those using only one source of information. For product and ammonia accurate results were obtained. The application of NIR based calibration modelling followed by residuals modelling through the use of process data gave the best and most consistent results. For sugar concentration, MIR fused with process data slightly outperformed NIR and process data but predictions were corrupted by significant levels of noise. For lipids, NIR fused with process data gave more accurate results than MIR. Finally, for phosphate the calibration generated with MIR data can be improved by the incorporation of process data.

6.2 Recommendations for Future work

This thesis has demonstrated that monitoring is a major challenge in the fermentation process industries. A strategy for calibration model construction to interpret spectral probe information was derived and it was shown that accuracy could be improved by process data fusion. Opportunities for further improvements and exploitation of the techniques are summarised below:

- For the proposed algorithms:

  (1) The calibration model identification method could be modified to take account of the dynamic system trends using for example dynamic PLS. Additionally, the stopping criterion in SWS could be improved using internal cross validation to supervise the convergence of the testing data set.

  (2) A comprehensive study of the impact of configurable parameters and strategies for SWS and the stacking approaches is required. For instance, the most appropriate method to weight the individual models in the stacked model requires further investigation. An optimisation approach for weight determination may be preferable to the methods proposed.

(3) The use of other process information including process data to enhance prediction accuracy has also been proven to be useful. However, this is still a relatively new area of research with only few papers reporting such applications. Research still needs to be carried out into the most effective means of fusing spectral/process and spectral/spectral data.

- For the fermentation application:

(1) The seed stage and the optimum transfer time of the seed to the final stage could be an additional factor to be considered in the final stage analysis. Multivariate curve resolution (MCR) techniques could be used for the identification of the optimum transfer time. MCR techniques can be adapted under specific constraints and offer a methodology to estimate reaction rate constants from obtained spectroscopic data. The work of Bijlsma and Smilde (1999), and Bijlsma *et al.* (2000) was a major step in the estimation of reaction rate constants. The result of their research was the estimation of the end point or optimal point of a reaction and this could have considerable benefit in fermentation process. Moreover Independent Component Analysis (ICA) as described in Hyvarinen *et al.*, (2001) is a method designed to offer a solution to the Blind Source Separation problem, i.e. separate the source signals from the observations of their mixtures and could also be used for the construction of kinetic profiles from which the end point of the reaction can be obtained.

(2) Biomass could also be considered as a very important additional factor for calibration modelling. The biomass concentration is not monitored routinely but on-line indicators are possible, exploiting for instance the biomass probe from Aber Instrument. Measurement of biomass is an important aspect in bioprocess monitoring and control. It is a key analyte whose determination is useful in assessing the progress of a submerged bioprocess culture and importantly has a critical impact on the transmittance/reflectance of the broth and thus the validity of the calibration model.

(3) The development of local models was investigated. Time was used as a surrogate variable to switch between models. Further research could be

performed to identify divisions that are meaningful descriptors of the behaviour of the fermentation process taking for instance critical characteristics determined from other process indicators such as off-gas concentrations.

- For other applications:

(1) The utility of the SWS/Stacking approach has been demonstrated on a fermentation process. Its application to other process systems and the benefits it provides over and above existing strategies requires further investigation.

(2) This thesis focused on developing monitoring and modelling tools for batch processes. Analogous methodologies could be developed for continuous processes.

(3) The calibration modelling procedures have been used to obtain indicators of broth concentrations that can be acted upon by process operators. A potentially useful extension would be to consider how available information could be used as indicators of the onset of process deviations/fault conditions. Here, the incorporation within multivariate statistical process control procedures could be the way forward.

APPENDICES

Additional data sets (Foss and ABB)

Two sets of additional instrument data from the fermentation process were also used to demonstrate the applicability of SWS for wavelength selection and stacking methodologies to attain a robust calibration model. Chapter 4 demonstrated that SWS variable selection generated calibration models that outperformed other calibration approaches. In this section, the ability of the SWS algorithm to identify the most relevant wavelength regions is assessed for different instrumentation, i.e., Foss and ABB. The data from the first 110 hours for both instruments was used and thus the last time interval was almost removed.

APPENDIX A

Analysis of Standard Batches from the Non-Invasive NIR Foss Probe

Data from seven standard batches was available and Figure A.1 shows the first derivative NIR spectral profile.

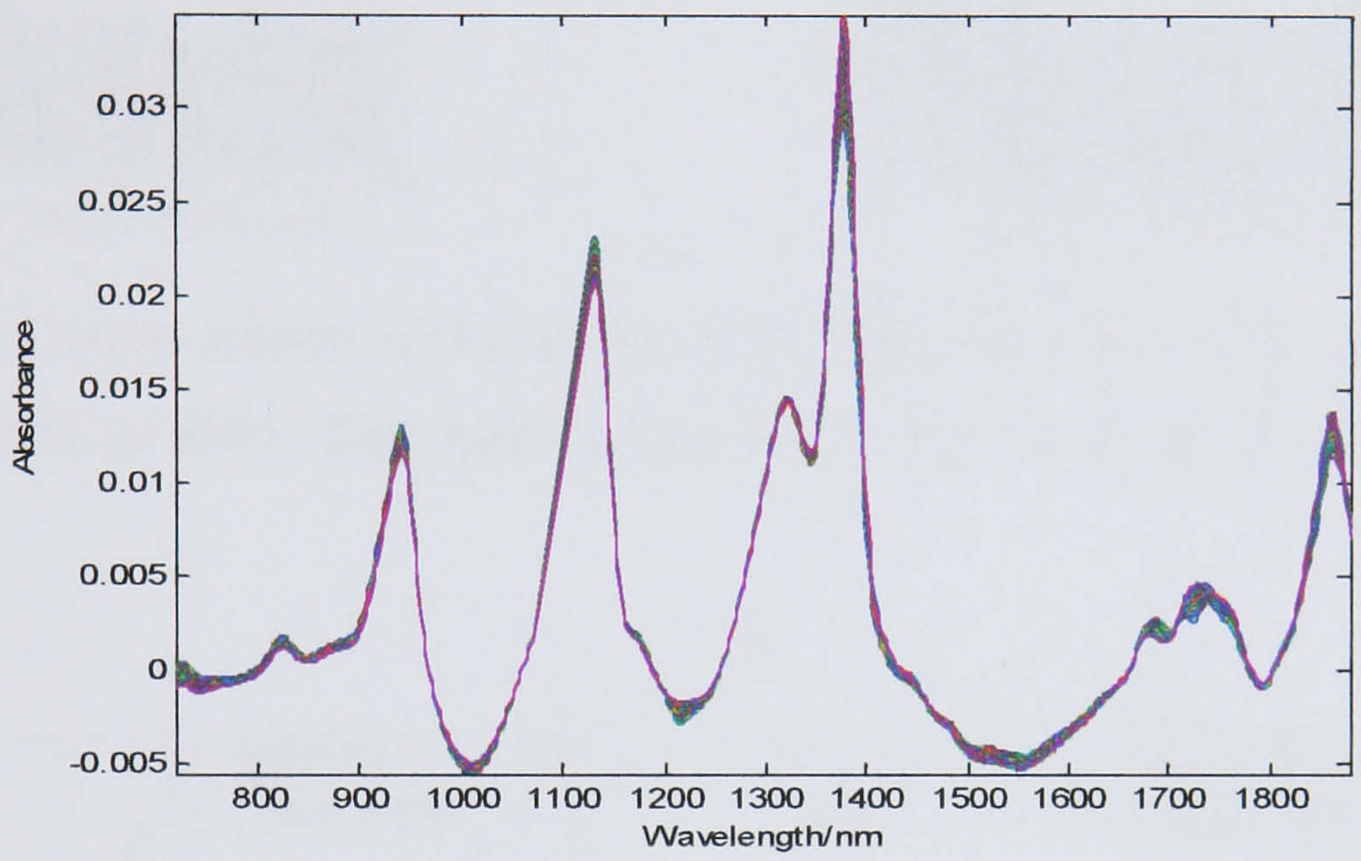


Figure A.1. Derivatives from batch SNI1 for spectra generated from Foss instrument, i.e. batches SNI1 to SNI5.

A smaller window size could also be used. With respect to the number of wavelengths included in each window in the SWS algorithm, the algorithm was run for windows up to 200 wavelengths (Figure A.2). The training data set indicates that a fixed error materialises after the 50th wavelengtht. For the validation data set, the error exhibits greater variation and increases after the 120th wavelength.

The results of SWS selection of both PLS and average stacking modelling can be seen in Figure A.3 and Table A.1. In this case PLS stacking slightly outperforms average stacking but only one validation batch is available. Ideally, the comparison should be made using a larger number of batches but only limited data was available from the experimental trials. Figures A.3a and A.3b show the results for the training and validation data sets for the Foss NIR instrument and in Figures A.3c and A.3d, the

predicted versus actual values are observed and the SWS wavelength algorithm performs well.

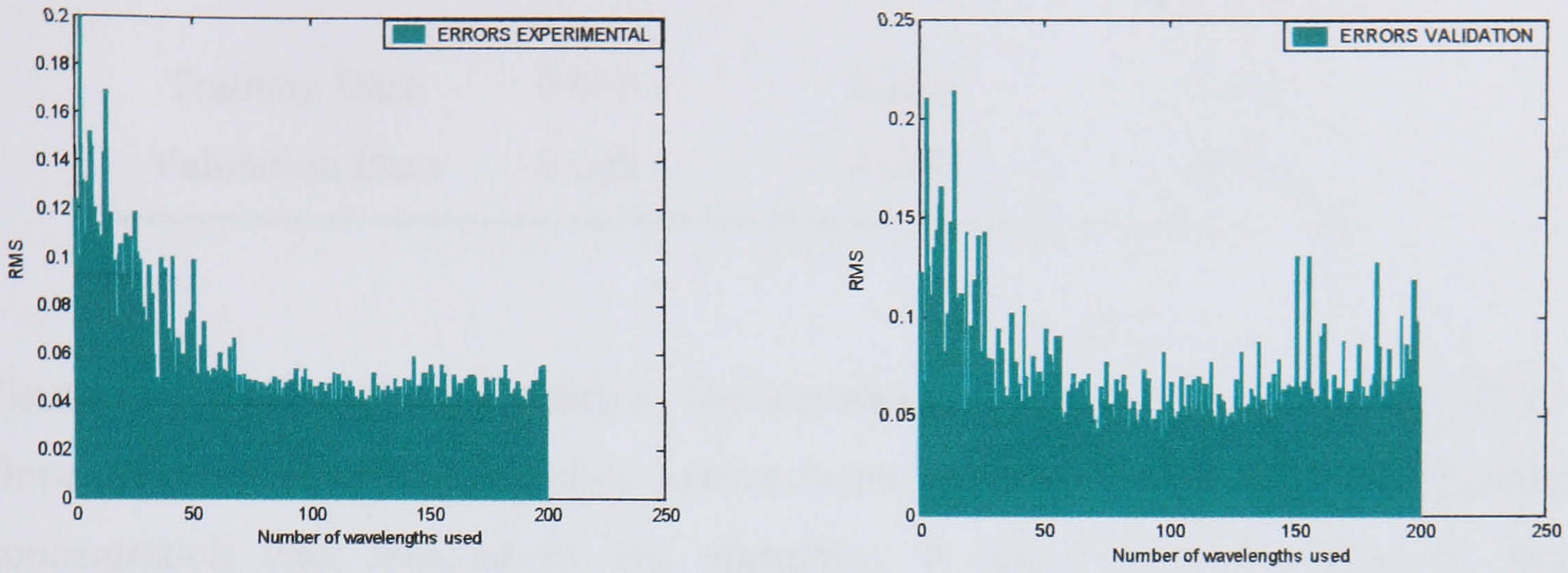
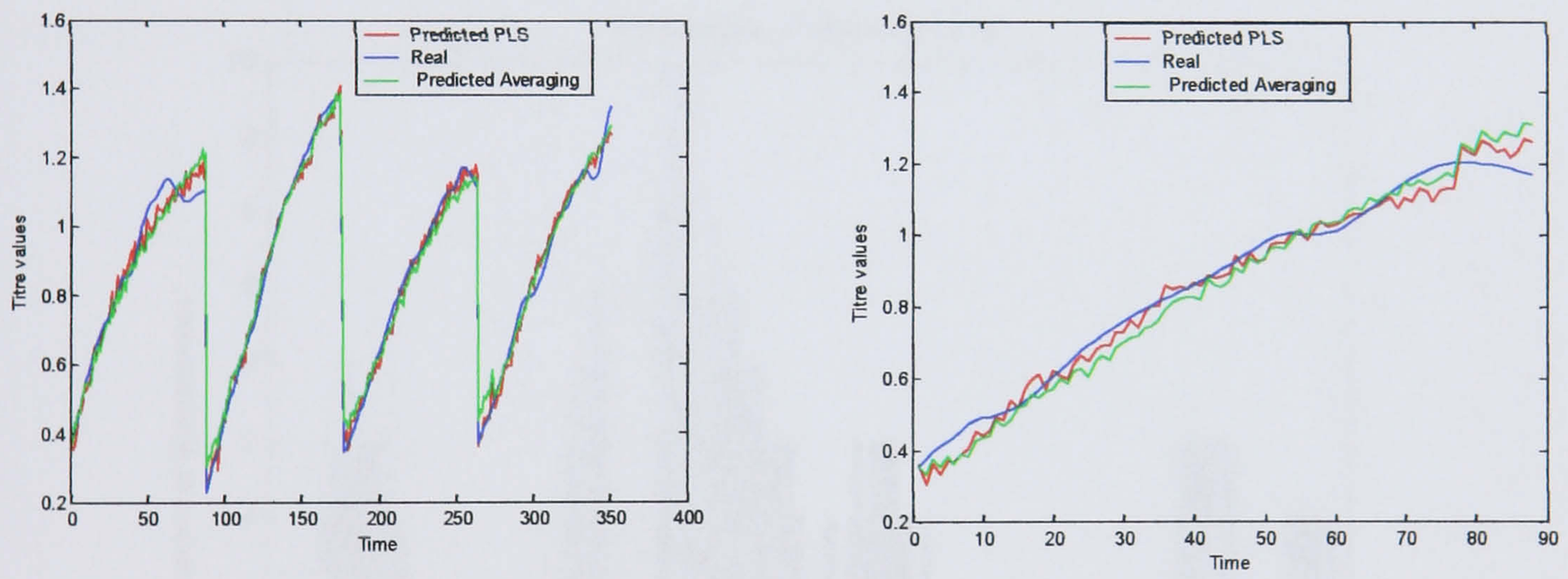


Figure A.2: Determination of the number of wavelengths used in each window for batches SNI1 to SNI5. The algorithm ran for a window up to 200 wavelengths.

(a)

(b)



(c)

(d)

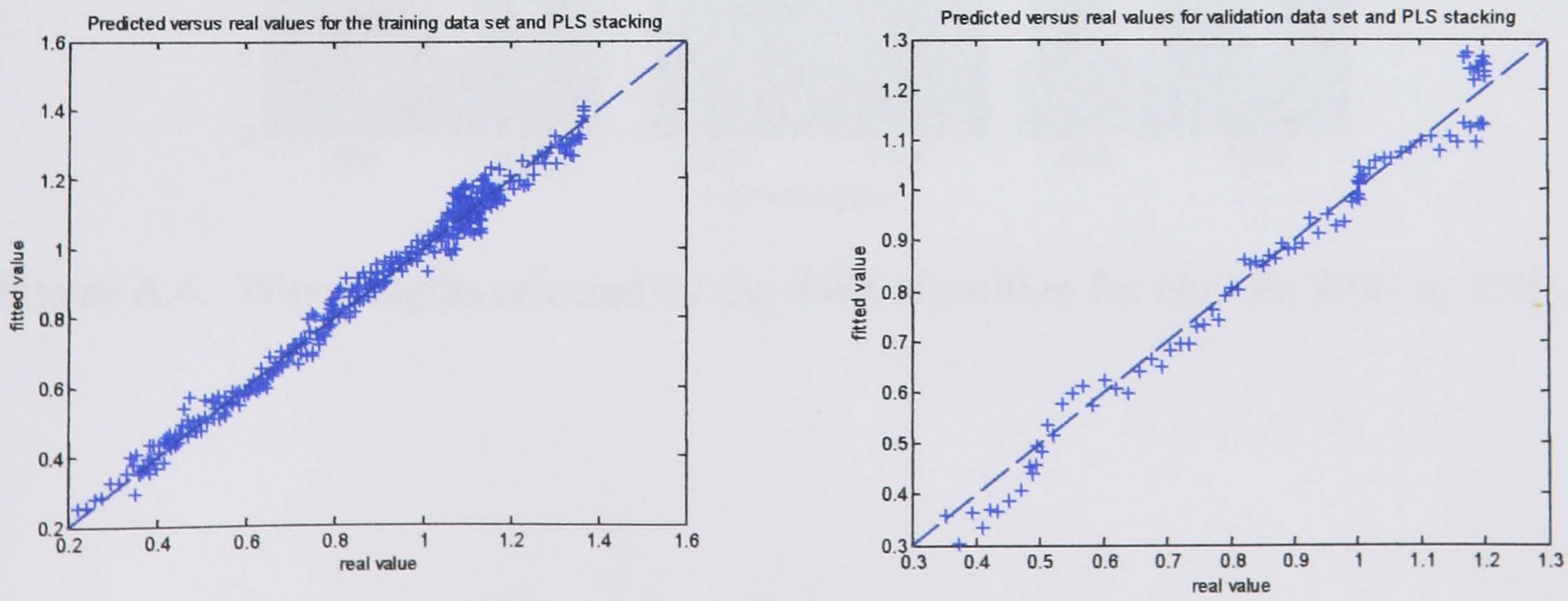


Figure A.3. Results from the Foss spectra modelling for batches SNI1 to SNI5.

Table A.1. Results from the Foss spectra modelling for batches SNI1 to SNI5

	Linear	SWS	
	PLS	Average Stacking	PLS Stacking for 6 LVs
Training Data	0.036	0.044	0.034
Validation Data	0.062	0.050	0.043

Figure A.4 presents the frequency of the wavelengths chosen by the SWS algorithm. Once again the wavelength region known to be informative with regard to product concentration was selected by the algorithm. It should also be observed that wavelengths in the region 1400-1500 nm were omitted by the algorithm. This range corresponds to a water peak, thus correctly the algorithm did not select these wavelengths.

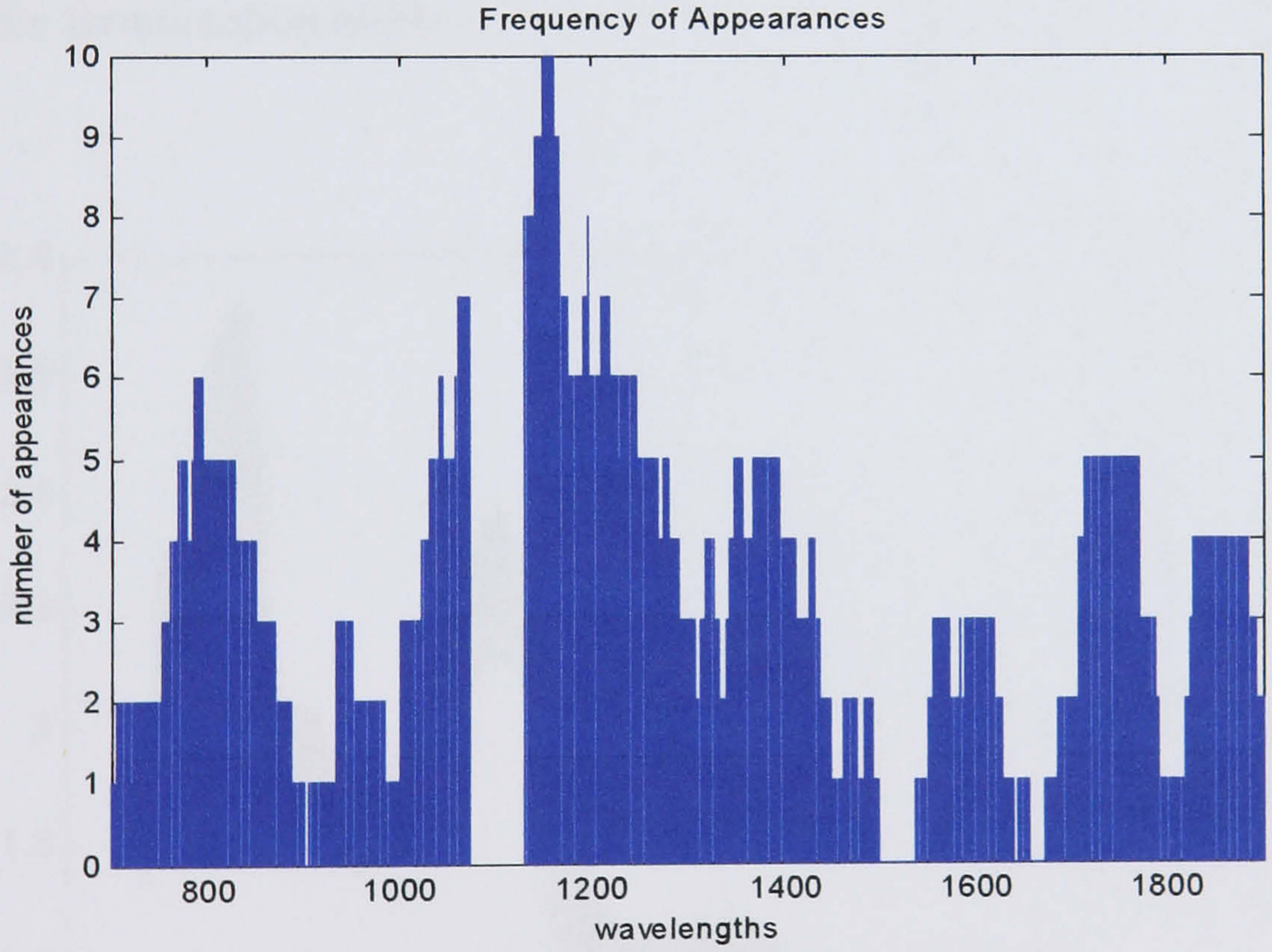


Figure A.4. Wavelengths selected by the SWS algorithm for batches SNI1 to SNI5

APPENDIX B

Analysis of Standard Batches from the Invasive NIR ABB probe

Whilst non-invasive probes can be applied at the pilot scale, larger vessels require the use of invasive probes as there are no glass windows against which to place the probe. The performance of invasive NIR probes was compared with the non-invasive results obtained previously. Final stage data from 6 standard batches using an ABB NIR invasive probe formed the basis of the study. Figure B.1 shows the raw spectra from one batch. The data pre-treatment strategy was the same as that used for the non-invasive systems.

The results are reported in Table B.1. It can be seen that the results are not as precise as for the non-invasive probe from Foss (Table A.1). This could be expected given the nature of the fermentation broth.

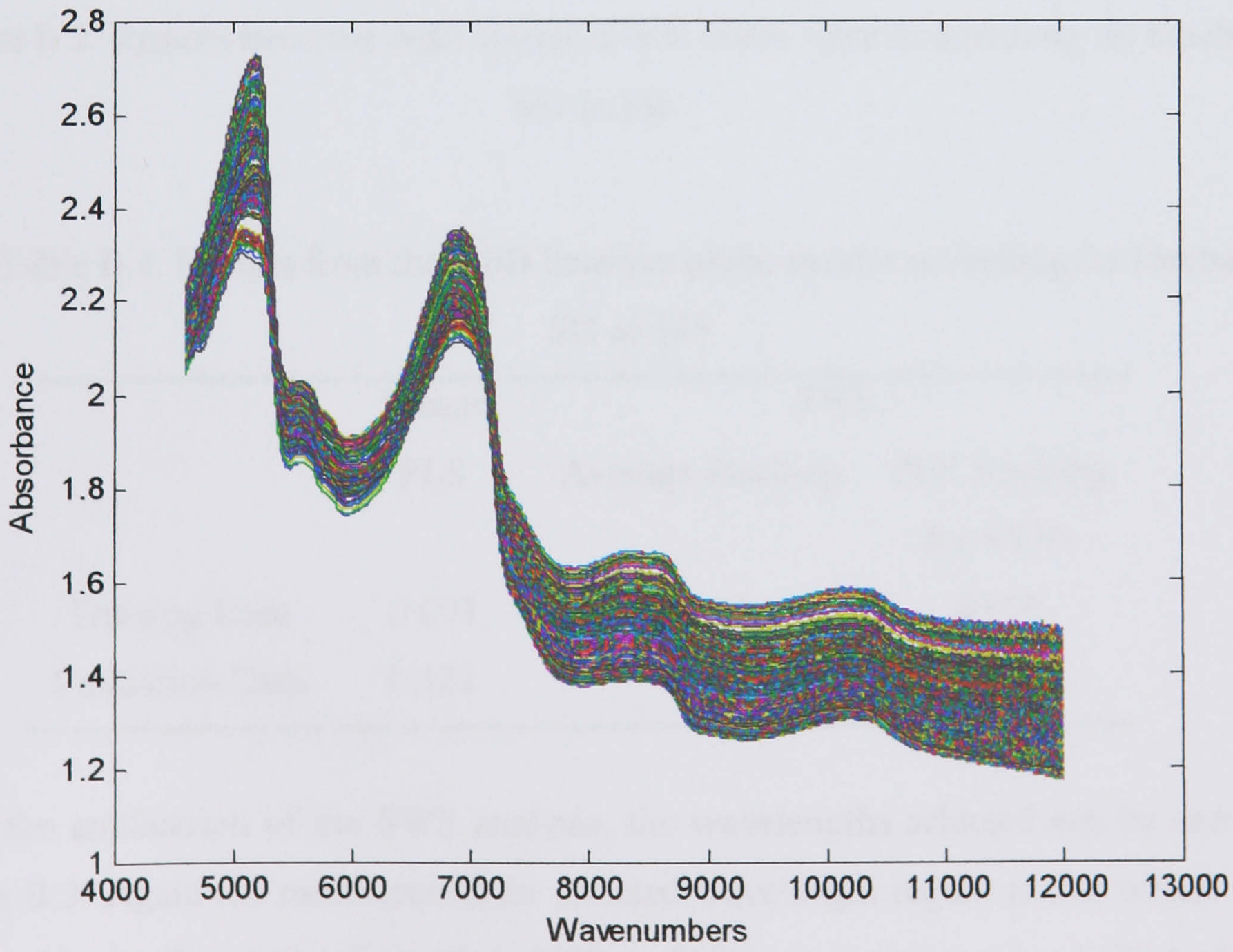


Figure B.1. Raw spectra of batch SI1 from the invasive probe, i.e. batches SI1 to SI6.

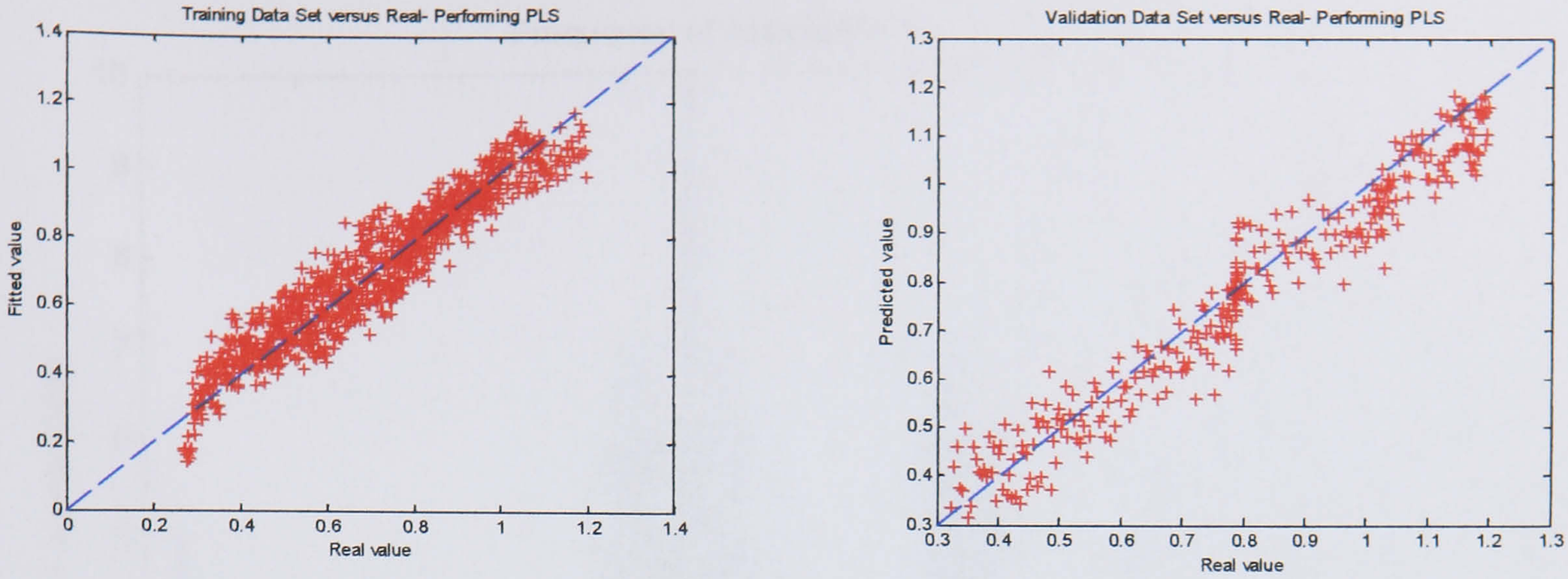


Figure B.2. Results from the ABB Invasive NIR probe spectra modelling for batches SI1 to SI6.

Table B.1. Results from the ABB invasive probe spectra modelling for batches SI1 to SI6

	Linear	SWS	
	PLS	Average Stacking	PLS Stacking for 6 LVs
Training Data	0.093	0.090	0.070
Validation Data	0.121	0.105	0.068

After the application of the SWS analysis, the wavelengths selected can be seen in Figure B.3. Again the most commonly selected wavelength region is that which was expected to be the most informative, 8000 to 8500 wavenumbers, i.e. 1250 to 1176 nm approximately.

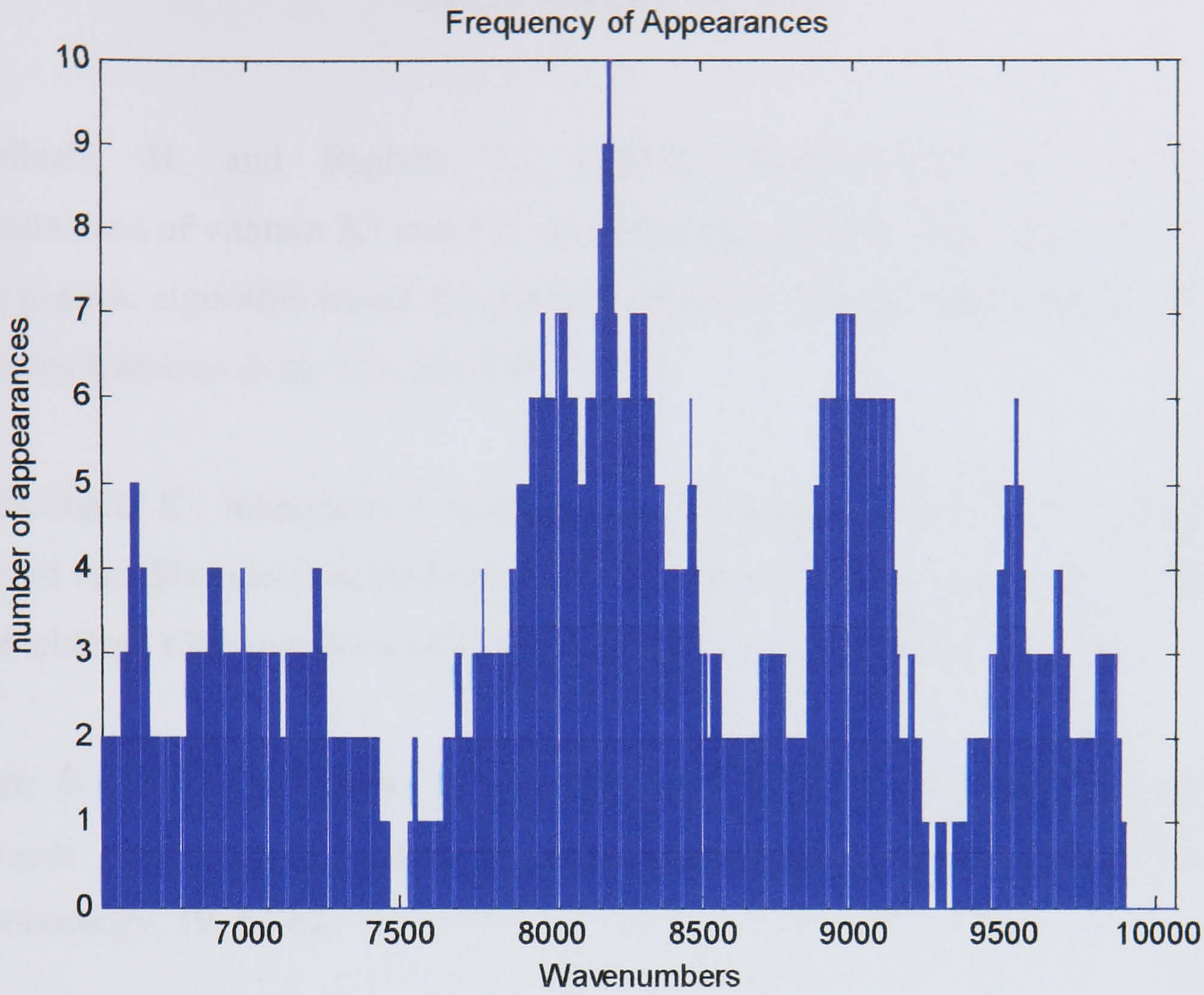


Figure B.3. Wavelengths selection frequency from the application of the SWS algorithm for batches SI1 to SI6.

REFERENCES

Abdollashi, H. and Bagheri, L., (2004), 'Simultaneous spectrophotometric determination of vitamin K3 and 1,4 - naphthoquinone after cloud point extraction by using genetic algorithm based wavelength selection - partial least squares regression, *Analytica Chimica Acta*, 514, 211-218.

Abrahamsson C., Johansson, J, Sparén A. and Lindgren, F., (2003), 'Comparison of different variable selection methods conducted on NIR transmission measurements on intact tablets', *Chemometrics and Intelligent Laboratory Systems*, 69, 3-12.

Albert, S. and R.D. Kinley, (2001), Multivariate statistical monitoring of batch processes: an industrial case study of fermentation supervision, *Trends in biotechnology*, 19, 53-62.

Alciature, C.E., E.E. Marcos, and De La Cruz, C., (1998), 'A numerical procedure for curve fitting of noisy infrared spectra', *Analytica Chimica Acta*, 376, 169-181.

Alexandridis, A., Patrinos, P., Sarimveis, H. and Tsekouras, G., (2005), 'A two-stage evolutionary algorithm for variable selection in the development of RBF neural network models. *Chemometrics and Intelligent Laboratory Systems, 75, 149-162*

Andersson, C. A. and Bro, R., (2000), 'The N-way Toolbox for Matlab', *Chemometrics and Intelligent Laboratory Systems*, 52, 1-4.

Andrade, J.M., Sanchez, M.S. and Sarabia, L.A. (1999), 'Applicability of high-absorbance MIR spectroscopy in industrial quality control of reformed gasolines', *Chemometrics and Intelligent Laboratory Systems*, 46, 41-55.

Andrews, J. and Dallin, P., (2003), 'Getting to grips with the process: extractive and remote sampling', *Spectroscopy Europe*, 15, 27-30.

Araujo, M.C.U., Saldanha, T.C.B., Galvao, R.K.H., Yoneyama, T., Chame, H.C. and Visani, V., (2000), 'The successive projections algorithm for variable selection in spectroscopic multicomponent analysis', *Chemometrics and Intelligent Laboratory Systems*, 57, 65-73.

Araujo, P. W. and R. G. Brereton (1996), 'Experimental design III. Quantification', *Trends in analytical chemistry,* 15: 156-163.

Araujo, P. W. and Brereton, R. G. (1996), 'Experimental design II. Optimisation', *Trends in analytical chemistry* , 15: 63-70.

Araujo, P. W. and Brereton, R. G. (1996), 'Experimental design I. Screening.' *Trends in analytical chemistry,* 15: 26-31.

Archibald D.D. and Akin D.E., (2000), 'Use of spectral window pre-processing for selecting near-infrared reflectance wavelengths for determination of the degree of enzymatic retting of intact flax stems', *Vibrational Spectroscopy,* 23, 169-180.

Arnold, S.A., Crowley, J., Woods, N., Harvey, L.M. and McNeil, B., (2003), 'In-Situ near infrared spectroscopy to monitor key analytes in mammalian cell cultivation', *Biotechnology and Bioengineering*, 84(1), 13-19.

Arnold, A., Harvey, L.M., McNeil, B. and Hall, J.W., (2003), 'Employing near-infrared spectroscopic methods of analysis for fermentation monitoring and control. Part 2, implementation strategies.', *BioPharm International*, 2003. 16(1), 47-49.

Arnold, A., Harvey, L.M., McNeil, B. and Hall, J.W., (2002), 'Employing near-infrared spectroscopic methods of analysis for fermentation monitoring and control. Part 1, method development.' *BioPharm International*, 15(11), p. 26-34.

Arnold, S.A., Gaensakoo, R., Harvey, L.M. and McNeil, B., (2002a). 'Use of at-line and in-situ near-infrared spectroscopy to monitor biomass in an industrial fed-batch *Escherichia coli* process', *Biotechnology and Bioengineering,* 80(4), 405 - 413.

Arnold, A.S., Matheson, L, Harvey, L M. and McNeil, B, (2001), 'Temporally segmented modelling: a route to improved bioprocess monitoring using near infrared spectroscopy?', *Biotechnology Letters*, 23, 143-147.

Arnold, S.A., Crowleya J., Vaidyanathan, S., Matheson, L., Mohan, P., Hall, J.W., Harvey, L.M. and McNeil, B., (2000), 'At-line monitoring of a submerged filamentous bacterial cultivation using near-infrared spectroscopy', *Enzyme and Microbial Technology*, 27(9), 691-697.

Azzouz, T., Puigdomenech, A., Aragay, M. and Tauler, R., (2003), 'Comparison between different data pre-treatment methods in the analysis of forage samples using near-infrared diffuse reflectance spectroscopy and partial least-squares multivariate calibration method', *Analytica Chimica Acta*, 484, 121-134.

Baffi, G., Martin, E.B. and Morris, A.J., (1999), 'Non-linear projection to latent structures revisited: the quandratic PLS algorithm', *Computers and Chemical Engineering*, 23, 395-411.

Bailey, J.E. and Ollis, D.F., (1986), 'Biochemical engineering fundamentals', second edition, McGraw-Hill international editions.

Bains,W., (1998), 'Biotechnology from A to Z', Oxford university press.

Bakken, G.A., Houghton, T.P. and Kalivas, J.H., (1999), 'Cyclic subspace regression with analysis of wavelength–selection criteria', *Chemometrics and Intelligent Laboratory Systems*, 45, 225-239.

Barnes, R.J., Dhanoa, M.S. and Lister, S.J., (1989), ' Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra', Applied Spectroscopy, 43, 772-777.

Barros, A.S. and Rutledge, D.N., (2004), 'Principal components transform-partial least squares: a novel method to accelerate cross-validation in PLS regression', *Chemometrics and Intelligent Laboratory Systems*, 73(2), 245-255.

Barros, A.S. and Rutledge, D.N., (1998), 'Genetic algorithm applied to the selection of principal components', *Chemometrics and Intelligent Laboratory Systems*, 40, 65-81.

Bank, Y.H., Yoo, C.K. and Lee, I.B., (2003), 'Nonlinear PLS modelling with fuzzy inference system', *Chemometrics and Intelligent Laboratory Systems*, 64, 137-155.

Berglund, A. and Wold, S. (1999), 'A serial extension of multiblock PLS', *Journal of Chemometrics*, 13, 461-471.

Bertran, E., Blanco, M., Maspoch, S., Ortiz, M.C., Sanchez, M.S. and Sarabia, L.A., (1999), 'Handling intrinsic non-linearity in near-infrared reflectance spectroscopy', *Chemometrics and Intelligent Laboratory Systems*, 49, 215-224.

Berntsson, O., Danielsson, L.G., Johansson, M.O. and Folestad, S., (2000), 'Quantitative determination of content in binary power mixtures using diffuse reflectance near infrared spectrometry and multivariate analysis', *Analytica Chimica Acta*, 419, 45-54.

Bijlsma, S., Boelens, H. F. M., Hoefsloot, H.C.J. and Smilde, A.K., (2000), 'Estimating reaction rate constants: comparison between traditional curve fitting and curve resolution', *Analytica Chimica Acta*, 419,197-207.

Bijlsma, S. and Smilde, A. K., (1999), 'Application of curve resolution based methods to kinetic data', *Analytica Chimica Acta*, 396, 231-240.

Bird, P.A., Sharp, D.C.A. and Woodley, J.M., (2002), 'Near-IR spectroscopic monitoring of analytes during microbially catalysed baeyer-villiger bioconversions', *Organic Process Research and Development*, 6, 569-576.

Blanco, M., Peinado, A.C. and J. Mas, (2005), 'Elucidating the composition profiles of alcoholic fermentations by use of ALS methodology' *Analytica Chimica Acta,* 544, 199-205.

Blanco, M., Coello, J., Iturriaga, H., Maspoch, S., Pages, J., (1999), 'Calibration in non-linear near infrared reflectance spectroscopy: a comparison of several methods', *Analytica Chimica Acta*, 384, 207-214.

Blanco, M., Coello, J., Iturriaga, H., Maspoch, S. and Pages, J., (2000), 'NIR calibration in non-linear systems: different PLS approaches and artificial neural networks', *Chemometrics and Intelligent Laboratory Systems*, 50, 75-82.

Boger, Z., (2003), 'Selection of quasi-optimal inputs in chemometrics modelling by artificial neural network analysis', *Analytica Chimica Acta*, 490, 31-40.

Bras, L. P., Bernardino, S.A., Lopes, J.A., Menezes, J.C., (2005), 'Multiblock PLS as an approach to compare and combine NIR and MIR spectra in calibrations of soybean flour', *Chemometrics and Intelligent Laboratory Systems,* 75, 91-99.

Breiman, L., (1993), 'Fitting additive models to regression data. Diagnostics and alternative views', *Computational Statistics and Data Analysis*, 15(1), 13-46.

Breiman, L., (1996), 'Bagging predictors', *Machine Learning J.*, 24(2), 123-140.

Brenchley, J.M., Horchner, U. and Kalivas, J.H., (1997), 'Wavelength selection characterization for NIR spectra', *Applied Spectroscopy*, 51(5), 689-699.

Brereton, R.G., (2000), 'Introduction to multivariate calibration in analytical chemistry', *Analyst*, 125, 2125-2154.

Bro, R. and Smilde, A.K., (2003), 'Centering and scaling in component analysis', *Journal of Chemometrics*, 17, 16-33.

Broadhurst, D., Goodacre, R., Jones, A., Rowland, J.J. and Kell, D.B., (1997), 'Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with applications to pyrolysis mass spectrometry', *Analytica Chimica Acta*, 348, 71-86.

Brown, S.D., Sum, S.T. and Despagne, F., (1996), 'Chemometrics', *Anal.Chem.*, 68, 21R-61R.

Brown, P.J., Spiegelman, C.H., Denham, M.C., (1991), 'Chemometrics and spectral frequency selection', *Phi., Trans., R., Soc. Lond.*, 337, 311-322.

Bungay, H.R., (1993), 'Basic biochemical engineering', BiLine Associates.

Burns, D.A. and Ciurczak, E.W., (1992), 'Handbook of near-Infrared analysis. Practical spectroscopy series', Vol. 13, Dekker.

Candolfi, A., Maesschalck, R., Massart, D.L., Hailey, P.A. and Harrington, A.C.E., (1999), 'Identification of pharmaceutical excipients using NIR spectroscopy and SIMCA', *Journal of Pharmaceutical and Biomedical Analysis*, 19, 923-935.

Carr-Brion, K.G., (1991), 'Measurement and control in bioprocessing', Elsevier Applied Science.

Centner, V., Massart, D.L., Noord, O.E, Jong, S., Vandeginste, B.G.M. and Sterna, C., (1996), 'Elimination of uninformative variables for multivariate calibration', *Analytical Chemistry*, 68, 3851-3858.

Chatterjee, S., Laudato, M. and Lynch, L.A., (1996), 'Genetic algorithms and their statistical applications: an introduction. Computational statistics and data analysis', 22, 633-651.

Chen B., Fu, X.G., Lu, D.L., (2002), 'Improvement of predicting precision of oil content in instant noodles by using wavelet transforms to treat near-infrared spectroscopy', *Journal of Food Engineering*, 53, 373-376.

Chung, H, Ku, M.S. and Lee, J.S., (1999), 'Comparison of near-infrared and mid-infrared spectroscopy for the determination of distillation property of kerosene', *Vibrational Spectroscopy,* 20, 155-163.

Chung, Y.C., Chien, I.L. and Chang, D.M., (2006), 'Multiple-model control strategy for a fed-batch high cell-density culture processing', *Journal of Process Control*, 16, 9-26.

Cimander, C. and Mandenius, C.F., (2002), 'Online monitoring of a bioprocess based on a multi-analyser system and multivariate statistical process modelling', *Journal of Chemical Technology and Biotechnology*, 77, 1157-1168.

Cimander, C., Carlsson, M. and Mandenius, C.F., (2002), 'Sensor fusion for on-line monitoring of yoghurt fermentation', *Journal of Biotechnology*, 99, 237-248.

Coates, J., (2000), 'Interpretation of infrared spectra. A practical approach', in Encyclopedia of Analytical Chemistry, Wiley, 10815-10837.

Conlin, A.K., Martin, E.B. and Morris, A.J., (1998), 'Data augmentation: An alternative approach to the analysis of spectroscopic data', *Chemometrics and Intelligent Laboratory Systems*, 44, 161-173.

Crowley, J., Arnold, S.A., Wood, N., Harvey, L.M. and McNeil, B., (2005), 'Monitoring a high cell density recombinant *pichia pastoris* fed-batch bioprocess using transmission and reflectance near infrared spectroscopy', *Enzyme and Microbial Technology*, 36, 621-628.

Crowley, J., McCarthy, B., Nunn, N.S., Harvey, L.M. and McNeil, B., (2000), 'Monitoring a recombinant *pichia pastoris* fed batch process using fourier transform mid-infrared spectroscopy (FT-MIRS)', *Biotechnology Letters*, 22, 1907-1912.

Cuadrado, M. U., de Castro, M.D.L., Juan, P.M.P. and Gomez-Nieto, M.A., (2005), 'Comparison and joint use of near infrared spectroscopy and fourier transform mid-

infrared spectroscopy for the determination of wine parameters', *Talanta*, 66, 218-224.

Dadd, M.R., Sharp, D.C.A., Pettman, A.J. and Knowles, C.J., (2000), 'Real-time monitoring of nitrile biotranformations by mid-infrared spectroscopy', *Journal of Microbiological Methods*, 41, 69-75.

Depczynski, U., Jetter, K., Molt, K. and Niemoller, A., (1999), 'Quantitative analysis of near infrared spectra by wavelet coefficient regression using a genetic algorithm', *Chemometrics and Intelligent Laboratory Systems*, 47, 179-187.

DiFoggio, R., (2000), 'Guidelines for applying chemometrics to spectra: Feasibility and error propagation', *Applied Spectroscopy*, 54(3), 94A- 113A.

Diver, D.A. and Ireland, D.G., (1997), 'Spectral decomposition by genetic algorithm. nuclear instruments and methods in physics research', 399, 414-420.

Doak, D.L. and Phillips, J.A., (1999), 'In situ monitoring of an *escherichia coli* fermentation using a diamond composition ATR probe and mid-infrared spectroscopy', *Biotechnol.Prog.*, 15, 529-539.

DoE overview, Engineering Statistics Handbook, (2004), http://www.itl.nist.gov/div898/handbook/pri/pri_d.htm.

Du, Y. P., Liang, Y.Z., Jiang, J.H., Berry, R.J., Ozaki, Y., (2004), 'Spectral regions selection to improve prediction ability of PLS models by changeable size moving window partial least squares and searching combination moving window partial least squares', *Analytica Chimica Acta*, 501, 183-191.

Dyrby, M., Petersen, R.V., Larsen, J., Rudolf, B., Nørgaard, L. and Engelsen, S.B., (2004), 'Towards on-line monitoring of the composition of commercial carrageenan powders', *Carbohydrate Polymers*, 57, 337-348.

Eigenvector Research, NIR spectra of diesel fuels, (2005),

http://software.eigenvector.com/Data/SWRI/.

Esteban-Diez, I., Gonzalez-Saiz, J.M. and Pizarro, C., (2004), OWAVEC: A combination of wavelet analysis and an orthogonalization algorithm as a pre-processing step in multivariate calibration. *Analytica Chimica Acta*, 515, 31-41.

Estienne, F., Massart, D.L., Zanier-Szydlowski, N. and Marteau, Ph., (2000), 'Multivariate calibration with raman spectroscopic data: A case study', *Analytica Chimica Acta*, 424, 185-201.

Faust, C.B., (2001), 'Modern chemical techniques', 5th ed., Royal Society of Chemistry.

Fayolle, P., Picque, D., Corrieu, G., (1997), 'Monitoring of fermentation processes producing lactic acid bacteria by mid-infrared spectroscopy', *Vibrational Spectroscopy*, 14, 247-252.

Fearn, T., (2000), 'On orthogonal signal correction', *Chemometrics and Intelligent Laboratory Systems*, 50, .47–52

Felicio, C. C., Bras, L.P., Lopes, J.A., Cabrita, L. and Menezes, J.C., (2005), 'Comparison of PLS algorithms in gasoline and gas oil parameter monitoring with MIR and NIR', *Chemometrics and Intelligent Laboratory Systems* , 78, 74-80.

Ferreira, A.P., Alves, T.P. and Menezes, J.C., (2005), 'Monitoring complex media fermentations with near-infrared spectroscopy: Comparison of different variable selection methods', Biotechnology and Bioengineering, 91, 474-481.

Forbes, R.A., Luo, M.Z. and Smith, D.R., (2001), 'Measurement of potency and lipids in monensin fermentation broth by near-infrared spectroscopy', *Journal of Pharmaceutical and Biomedical Analysis*, 25, 239-256.

Forbes, R.A., Persinger, M.L. and Smith, D.R. (1996), 'Development and validation of analytical methodology for near-infrared conformance testing of pharmaceutical intermediates', *Journal of Pharmaceutical and Biomedical Analysis*, 15, 315-327.

Forina, M., Casolino, C. and Millan, C.P., (1999), 'Iterative predictor weighting (IPW) PLS: A technique for the elimination of useless predictors in regression problems', *Journal of Chemometrics*, 13, 165-184.

Frank, I. and Friedman, J., (1993), 'A statistical view of some chemometrics regression tools', *Technometrics,* 35, 109-135.

Freund, Y. and Schapire, R., (1997), 'A decision-theoretic generalization of on-line learning and an application to boosting', *Journal of Comput. System Sci.*,. 55(1), 119-139.

Freund, Y. and Schapire., R., (1996), 'Experiments with a new boosting algorithm' in 13th International Conference on Machine Learning..

Geladi, P., (1992), 'Wold, Herman, the father of PLS', *Chemometrics And Intelligent Laboratory Systems,* 15, R7-R8.

Geladi P. and Kowalski, B., (1986), 'Partial least-squares regression: A tutorial', *Analytical Chimica Acta*, 185, 1-17.

Geladi, P., MacDougall, D. and Martens, H., (1985), 'Linearization and scatter-correction for near-infrared reflectance spectra of meat', Applied spectroscopy, 39, 491-500.

Ghasemi, J., Niazi, A.and Leardi, R., (2003), 'Genetic-algorithm-based wavelength selection in multicomponent spectrophotometric determination by PLS: application on copper and zinc mixture', *Talanta*, 59, 311-317.

Ghasemi, J. and Niazi, A., (2001), 'Simultaneous determination of cobalt and nickel. Comparison of prediction ability of PCR and PLS using original, first and second derivative spectra', *Microchemical Journal*, 68, *1-11.*

Giavasis, I., Robertson, I., McNeil, B., Harvey, L.M., (2003), 'Simultaneous and rapid monitoring of biomass and biopolymer production by *sphingomonas paucimobilis* using fourier transform-near infrared spectroscopy', *Biotechnology Letters*, 25, 975-979.

Goldberg, D., (1989), 'Genetic algorithms in search, optimisation and machine learning', Addison Wesley, Reading, Mass.

Gomez-Carracedo, M.P., Andrade, J.M., Calvino, M., Fernadez, E., Prada, D. and Muniategui, S., (2003), 'Multivariate prediction of eight kerozene properties employing vapour-phase mid-infrared spectrometry', *Fuel,* 82, 1211-1218.

Gorry, P.A., (1990), 'General least-squares smoothing and differentiation by the convolution (Savitzky-Golay) Method', Anal.Chem., 62, 570-573.

Gourvenec, S., Capron, X. and Massart, D.L., (2004), 'Genetic algorithms (GA) applied to the orthogonal projection approach (OPA) for variable selection', *Analytica Chimica Acta*, 519, 11-21.

Gregersen, L. and Jorgensen, S.B., (1999), 'Supervision of fed-batch fermentations', *Chemical Engineering Journal*, 75, 69-76.

Guo, Q., Wu, W. and Massart, D.L., (1999), 'The robust normal variate transform for pattern recognition with near-infrared data', *Analytica Chimica Acta*, 382, 87-103.

Gurden S.P., Westerhuis, J.A., Smilde, A.K., (2002), 'Monitoring of batch processes using spectroscopy', *AICHE Journal* , 48, 2283-2297.

Hassell, D.C. and Bowman, E.M., (1998), 'Process analytical chemistry for spectroscopists', *Applied Spectroscopy*, 52(1), 18A-29A.

Hinchliffe, M., Montague, G.A., Willis, M., Burke, A., (2003), 'Correlating polymer resin and end-use properties to molecular-weight distribution', *AIChE Journal*, 49, 2609-2618.

Hochberg, E.J., Atkinson, M.J. and Andréfouët, S., (2003), 'Spectral reflectance of coral reef bottom-types worldwide and implications for coral reef remote sensing', *Remote Sensing of Environment*, 85(2), 159-173.

Holland, J.M., (1975), 'Adaptation in natural and artificial systems', The University of Michigan Press.

Homepage of Chemometrics, (2004), http://www.acc.umu.se/~tnkjtg/chemometrics/editorial/aug2002.html

Horchner, U. and Kalivas, J.H., (1995), 'Further investigation on a comparative study of simulated annealing and genetic algorithm for wavelength selection', *Analytica Chimica Acta*, 311, 1-13.

Hoskuldsson, A., (2001), 'Variable and subset selection in PLS regression', *Chemometrics and Intelligent Laboratory Systems*, 55, 23-38.

Hussain, M. A., (1999), 'Review of the applications of neural networks in chemical process control - simulation and online implementation', *Artificial Intelligence in Engineering*, 13: 55-68.

Hyvarinen, A., Karhunen, J., Oja E., (2001), 'Independent component analysis', John Wiley and Sons.

Iizuka, K. and Aishima, T., (1999), 'Differentiation of soy sause by pattern recognition analysis of mid- and near-IR spectra', *Journal of Food Composition and Analysis*, 12, 197-209.

Indahl, U.-G., Sahni, N.S., Kirkhus, B. and Naes, T, (1999), 'Multivariate strategies for classification based on NIR-spectra -- with application to mayonnaise', *Chemometrics and Intelligent Laboratory Systems*,. 49, 19-31.

Jouan-Rimbaud, D., Massart, D.L., De Noord, O.E., (1996), 'Random correlation in variable selection for multivariate calibration with a genetic algorithm', *Chemometrics and Intelligent Laboratory Systems*, 35, 213-220.

Jouan-Rimbaud, D., Walczak, B., Massart, D.L., Last, I.R., Prebble, K.A., (1995), 'Comparison of multivariate methods based on wavelength selection for the analysis of near-infrared spectroscopic data', Analytica Chimica Acta, 304, 285-295.

Kacuracova, M. and Wilson, R.H., (2001), 'Developments in mid-infrared FT-IR spectroscopy of selected carbohydrates', *Carbohydrate Polymers*, 44, 291-303.

Kadi, H. E., (2005), 'Modeling the mechanical behavior of fiber-reinforced polymeric composite materials using artificial neural networks - A review', *Composite Structures, In Press, Corrected Proof, Available online 26 February 2005, www.sciencedirect.com/.*

Kalivas, J.H., (1999), 'Cyclic subspace regression with analysis of the hat matrix', *Chemometrics and Intelligent Laboratory Systems*, 45, 211-219.

Kalivas, J. H., (1997), 'Two data sets of near infrared spectra', *Chemometrics and Intelligent Laboratory Systems* , 37, 255-259.

Kansiz, M., Gapes, J.R., McNaughton, D., Lendl, B. and Schuster, K.C., (2001), 'Mid-infrared spectroscopy coupled to sequential injection analysis for the on-line monitoring of the acetone-butanol fermentation process', *Analytica Chimica Acta*, 438, 175-186.

Kasprow, R.P., Lange, A.J. and Kirwan, D.J., (1998), 'Correlation of fermentation yield with yeast extract composition as characterised by near-infrared spectroscopy', *Biotechnol. Prog.*, 14, 318-325.

Kornmann, H., Valentinotti, S., Marison, I. and Stocker, U., (2004a), 'Real time update of calibration model for better monitoring of batch processes using spectroscopy', *Biotechnology and Bioengineering*, 87, 593-601.

Kornmann, H., Valentinotti, S., Duboc, P., Marison, I. and Stocker, U., (2004b), 'Monitoring and control of *gluconacetobacter xylinus* fed-batch cultures using in situ mid-IR spectroscopy', *Journal of Biotechnology*, 113, 231-245.

Kornmann, H., Rhiel, M., Cannizzaro, C., Marison, I. and Stocker, U., (2003), 'Methodology for real-time, multianalyte monitoring of fermentations using an in-situ mid-infrared sensor', *Biotechnology and Bioengineering*, 82, 702-709.

Kourti, T., (2003), 'Multivariate dynamic data modelling for analysis and statistical process control of batch processes, start-ups and grade transitions', *Journal of Chemometrics*, 17, 93-109.

Kourti, T., Nomikos, P. and MacGregor, J.F., (1995), 'Analysis, monitoring and fault diagnosis of batch processes using multiblock and multiway PLS', *Journal of Process Control*, 5, 227-284.

Kramer, K. and Ebel, S., (2000), 'Application of NIR reflectance spectroscopy for the identification of pharmaceutical excipients', *Analytica Chimica Acta*, 420, 155-161.

Lazraq, A., Cleroux, R. and Gauchi, J.P., (2003), 'Selecting both latent and explanatory variables in the PLS1 regression model', *Chemometrics and Intelligent Laboratory Systems* , 66, 117-126.

Leardi, R. and Nørgaard, L., (2004), 'Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions', *Journal of Chemometrics*, 18, 486-497.

Leardi, R., Seasholtz, M.B. and Pell, R.J., (2002), 'Variable selection for multivariate calibration using a genetic algorithm: prediction of additive concentrations in polymer

films from fourier transform-infrared spectral data', *Analytica Chimica Acta*, 461, 189-200.

Leardi, R., (2000), 'Application of genetic algorithm-PLS for feature selection in spectral data sets', *Journal of Chemometrics*, 14, 643-655.

Lee, J.M., Yoo, C.K. and Lee, I.B., (2004), 'Enhanced process monitoring of fed-batch penicillin cultivation using time-varying and multivariate statistical analysis', *Journal of Biotechnology*, 110, 119-136.

Lennox, B., Kipling, K., Glassey, J., Montague, G.A., Willis, M. and Hiden, H., (2002), 'Automated production support for the bioprocess industry', *Biotechnol. Prog.*, 18, 269-275.

Lennox, B., Montague, G.A., Hiden, H.G., Kornfeld, G. and Goulding, P.R., (2001), 'Process monitoring of an industrial fed-batch fermentation', *Biotech Bioeng*,74(2), 125-135.

Lennox, B., Hiden, H.G., Montague, G.A., Kornfeld, G. and Goulding, P.R., (2000), 'Application of multivariate statistical process control to batch operations', *Computers and Chemical Engineering*, 24, 291-296.

Li, Y., Brown, C.W., Sun, F.M., McCrady, J.W., Traxler, R.W. and Lo, S.C., (1999), 'Non-invasive fermentation analysis using an artificial neural network algorithm for processing near infrared spectra', *J.Near Infrared Spectrosc*.,. 7, 101-108.

Little, R.J.A. and Rubin, D.B., (1987), 'Statistical analysis with missing data', Wiley.

Lopes, J., Costa, P.F., Ives, T.P. and Menezes, J.C. (2004), 'Chemometrics in bioprocess engineering: Process analytical technology (PAT) applications', *Chemometrics and Intelligent Laboratory Systems*, 74, 269-275.

Lopes, J., Menezes, J.C., Westerhuis J.A. and Smilde, A.K., (2002), 'Multiblock PLS analysis of an industrial pharmaceutical process', *Biotechnology and Bioengineering*, 80, 419-427.

Louwerse, D.J. and Smilde, A.K., (2000), 'Multivariate statistical process control of batch processes based on three-way models', *Chemical Engineering Science*, 55, 1225-1235.

Low-Ying, S., Shaw, R.A., Leroux, M. and Mantsch, H.H., (2002), 'Quantitation of glucose and urea in whole blood by mid-infrared spectroscopy of dry films', *Vibrational Spectroscopy*, 28, 111-116.

Lucasious, C.B., Beckers, M.L.M. and Kateman, G., (1994), 'Genetic algorithms in wavelength selection: a comparative study', *Analytica Chimica Acta*, 286(2), 135-153.

Lucasious, C.B. and Kateman, G., (1991), 'Genetic algorithms for large-scale optimization in chemometrics: an application', *Trends in Analytical Chemistry*, 10(8), 254-261.

Luypaert, J., Zhang, M.H., Massart, D.L., (2003), 'Feasibility study for the use of near infrared spectroscopy in the qualitative and quantitative analysis of green tea, *camellia sinensis* (L.)', *Analytica Chimica Acta*, 478, 303-312.

Macauley-Patrick, S., Arnold, S.A., McCarthy, B., Harvey, L.M. and McNeil, B., (2003), 'Attenuated total reflectance fourier transform mid-infrared spectroscopic quantification of sorbitol and sorbose during a gluconobaster biotransformation process', Biotechnology Letters, 25, 257-260.

Macedo, M.G., Laporte, M.F. and Lacroix, C., (2002), 'Quantification of exopolysaccharide, lactic acid, and lactose concentrations in culture broth by near-infrared spectroscopy'. *Journal of Agricultural and Food Chemistry*, 50, 1774-1779.

MacLaurin, P., Crabb, N.C., Wells, I., Worsfold, P.J. and Coombs, D., (1996), 'Quantitative in situ monitoring of an elevated temperature reaction using a water-cooled mid-infrared fiber-optic probe', *Analytical Chemistry*, 68, 1116-1123.

Madigan, M.T. and Martinko, J.M., (2000), 'Brock biology of microorganisms', Prentice-Hall Inc. 991.

Malin, S. F., Ruchti, T.L., Blank, T.B., Thennadil, S.N. and Monfre, S.L., (1999), 'Non-invasive prediction of glucose by near-infrared diffuse reflectance spectroscopy', *Clinical Chemistry*, 45, 1651-1658.

Mark, H. and Workman, J. J., (2003), 'Derivatives in spectroscopy. Part III - computing the derivative', *Spectroscopy*, 18, 106-111.

Martin, E.B. and Morris A.J., (2002), 'Enhanced bio-manufacturing through advanced multivariate statistical technologies', *Journal of Biotechnology*, 99, 223-235.

Massart, D.L., Vandeginste B.G.M., Buydens., L., De Jong, S. and Lewi, P., (1997), 'Handbook of chemometrics and qualimetrics', Elsevier Science Publishers B.V.

McClure, F. W., (1994), 'Near-infrared spectroscopy. The giant is running strong', *Analytical Chemistry*, 66, 43A-53A.

McKeivy, M.L., Britt, T.R., Davis, B.L., Gillie, J.K., Lentz, L.A. and Leugers, A., (1996), 'Infrared spectroscopy', *Analytical Chemistry*, 68, 93R-160R.

McLennan, F. and Kowalski, B.R., (1995), 'Process analytical chemistry', Blackie.

McShane, M.J., Cameron, B.D., Cote, G.L., Motamedi, M. and Spiegelman, C.H., (1999), 'A novel peak-hopping stepwise feature selection method with application to raman spectroscopy', *Analytica Chimica Acta*,. 388, 251-264.

McShane, M.J. and Cote, G.L., (1998), 'Near-infrared spectroscopy for determination of glucose, lactate, and ammonia in cell culture media', *Applied Spectroscopy*, 52(8), 1073-1078.

McShane, M.J., Cote, G.L. and Spiegelman, C., (1997), 'Variable selection in multivariate calibration of a spectroscopic glucose sensor', *Applied Spectroscopy*, 51(10), 1559-1564.

Miller, CE., (1995), 'The use of chemometric techniques in process analytical method development and operation', *Chemometrics and Intelligent Laboratory Systems*, 30, 11-22.

Montague, G. A., (1997), 'Monitoring and control of fermenters', Institution of Chemical Engineers.

Montgomery, D.C., (1997), 'Design and analysis of experiments', 4th ed., John Wiley and Sons.

Morgan, B.J.T., (1984), 'Elements of simulation', ed. C.a. Hall.

Navea, S., Tauler, R. and de Juan, A., (2005), 'Application of the local regression method interval partial least-squares to the elucidation of protein secondary structure', *Analytical Biochemistry*, 336, 231-242.

Naves, S., De-Juan, A. and Tauler, R., (2003), 'Modelling temperature-dependent protein structural transitions by combined near-IR and mid-IR spectroscopies and multivariate curve resolution', Anal.Chem., 75(20), 5592-5601.

Navrátil, M., Norberg, L. and Mandenius C.F., (2005), 'On-line multi-analyzer monitoring of biomass, glucose and acetate for growth rate control of a *vibrio cholerae* fed-batch cultivation', *Journal of Biotechnology,* 115, 67-79 .

Nilsson, N. J., (1965). 'Learning machines: Foundations of trainable pattern-classifying systems', McGraw-Hill.

Nørgaard, L., Hahn, M.T., Knudsen, L.B., Farhat, I.A. and Engelsen, S.B. (2005), 'Multivariate near-infrared and raman spectroscopic quantifications of the crystallinity of lactose in whey permeate powder', *International Dairy Journal*, 15, 1261-1270.

Nørgaard, L., Saudland, A. Wagner, J. Nielsen, J.P. Munck, L. and Engelsen, S.B., (2000), 'Interval partial least squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy', *Applied Spectroscopy*, 54, 413-419.

Pasti, L., Jouan-Rimbaud, D., Massart, D.L. and Noord, O.E (1998), 'Application of fourier transform to multivariate calibration of near-infrared data', *Analytica Chimica Acta*, 364, 253-263.

Pena, A.M., Acedo-Valenzuela, M.I., Espinosa-Mansilla, A. and Sanchez-Maqueda, R., (2002b), 'Stopped-flow fluorimetric determination of amoxycillin and clavulanic acid by partial least-squares multivariate calibration', *Talanta*, 56, 635-642.

Pena, A.M., Espinosa-Mansilla, A., Valenzuela, M.I.A., Goicoechea, H.C. and Olivieri, A.C., (2002a), 'Comparative study of net analyte signal-based methods and partial least squares for the simultaneous determination of amoxycillin and clavulanic acid by stopped-flow kinetic analysis', *Analytica Chimica Acta*, 463, 75-88.

Pedersen, J.G., (1997), 'Combining NIR data and production data for process control', *Process Control and Quality*, 9, 153-159.

Perrone, M. and Cooper, L. N., (1993), 'When networks disagree: Ensemble methods for hybrid neural networks', London, Chapman and Hall.

Perry, R.H., Green, D.W. and Maloney, J.O., (1997), 'Chemical engineering handbook', McGraw-Hill.

Pizarro, C., Esteban-Diez, I., Nistal, A.J. and Gonzalez-Saiz, J.M. (2004), 'Influence of data pre-processing on the ash content and lipids in roasted coffee by near infrared spectroscopy', *Analytica Chimica Acta* , 509, 217-227.

Rantanen, J., Rasanen, E., Antikainen, O., Mannermaa, J.-P. and Yliruusi, J. (2001), 'In-line moisture measurement during granulation with a four-wavelength near-infrared sensor: an evaluation of process-related variables and a development of non-linear calibration model', *Chemometrics and Intelligent Laboratory Systems*, 56, 51-58.

Reid, L. M., Woodcock, T., O'Donnell, C.P., Kelly, J.D. and Downey, G., (2005), 'Differentiation of apple juice samples on the basis of heat treatment and variety using chemometric analysis of MIR and NIR data', *Food Research International*, 38, 1109-1115.

Rhiel, M., Ducommun, P., Bolzonella, I., Marison, I. and Von Stockar, U., (2002), 'Real-Time in situ monitoring of freely suspended and immobilized cell cultures based on mid-infrared spectroscopic measurements', *Biotechnology and Bioengineering*, 77, 174-185.

Rhiel, M. H., Amrhein, M.I., Marison, I.W. and Stockar, U., (2002), 'The influence of correlated calibration samples on the prediction performance of multivariate models based on mid-infrared spectra of animal cell cultures', *Anal.Chem.*, 74, 5227-5236.

Riley, M.R., Arnold, M.A., Murhammer, D.W., Walls, E.L. and DelaCruz, N., (1998), 'Adaptive calibration scheme for quantification of nutrients and byproducts in insect cell bioreactors by near-infrared spectroscopy', Biotechnol. Prog.,. 14, 527-533.

Roggo, Y., Duponchel, L., Ruckebusch, C. and Huvenne, J.P., (2003), 'Statistical tests for comparison of quantitative models developed with near infrared spectral data', *Journal of Molecular Structure*, 654, 253-262.

Roubos, J.A., Krabben, P., Luiten, R.G.M., Verbruggen, H.B. and Heijnen, J.J., (2001), 'A quantitative approach to characterizing cell lysis caused by mechanical agitation of *streptomyces clavuligerus*', *Biotechnol.Prog.*, 17, 336-347.

Ruckebusch, C., Sombret, B., Froidevaux, R. and Huvenne, J.P., (2001), 'On-line mid-infrared spectroscopic data and chemometrics for the monitoring of an enzymatic hydrolysis', *Applied Spectroscopy*, 55, 1610-1617

Rutledge, D.N., Barros, A. and Delgadillo, I., (2001), 'PoLish - smoothed partial least-squares regression', *Analytica Chimica Acta*, 446, 281-296.

Savitzky, A. and Golay, M.J.E., (1964), 'Smoothing and differentiation of data by simplified least square procedures. *Analytical Chemistry*, 36, 1627-1639.

Schneider, R.C. and Kovar, K.A. (2003), 'Analysis of ecstasy tablets: comparison of reflectance and transmittance near infrared spectroscopy.' *Forensic Science International*, 134, 187-195.

Sharkey, A. J. C. and Sharkey, N. E. (1997), 'Combining diverse neural nets', *The Knowledge Engineering Review*, 12(3), 231-247.

Sharkey, A. J. C., (1996), 'On combining artificial neural nets', *Connection science*, 8, 299-313.

Sircar, A., Sridhar, P. and Das, P.K. (1998), 'Optimization of solid state medium for the production of clavulanic acid by *streptomyces clavuligerus*', *Process Biochemistry*, 33(3), 283-289.

Sivakesava, S., Irudayaraj, J. and Ali, D., (2001), 'Simultaneous determination of multiple components in lactic acid fermentation using FT-MIR, NIR, and FT-Raman spectroscopic techniques', *Process Biochemistry*, 37, 371-378.

Sjöblom, J., Svensson, O., Josefson, M., Kullberg, H. and Svante Wold, (1998), 'An evaluation of orthogonal signal correction applied to calibration transfer of near infrared spectra', *Chemometrics and Intelligent Laboratory Systems*, 44, 229–244.

Smilde, A.K., Westerhuis, J.A. and Boque, R., (2000), 'Multiway multiblock component and covariates regression models', *Journal of Chemometrics*, 14, 301 – 331.

Smith, B.M. and Gemperline, P.J., (2000), 'Wavelength selection and optimization of pattern recognition methods using the genetic algorithm', *Analytica Chimica Acta*. 423, 167-177.

Smith, J.E., (1988), 'Biotechnology', Second edition, Edward Arnold.

Sprang, E., (2004), 'Statistical batch process monitoring', University of Amsterdam. Thesis. p. 333.

Sprang, E.N.M., Ramaker, H.J., Boelens, H.F.M., Westerhuis, J.A., Whiteman, D., Baines, D. and Weaver, I., (2003), 'Batch process monitoring using on-line MIR spectroscopy', Analyst, 128, 98-102.

Sprang, E.N.M., Ramaker, H.J., Westerhuis, J.A., Gurden, S.P. and Smilde, A.K., (2002), 'Critical evaluation of approaches for on-line batch process monitoring', *Chemical Engineering Science*, 57, 3979-3991.

Stanbury, P.F, Whitaker, A., (1984), 'Principles of fermentation technology', Pergamon press.

Strangman, G., Franceschini, M.A. and Boas, D.A. (2003), 'Factors affecting the accuracy of near-infrared spectroscopy concentration calculations for focal changes in oxygenation parameters', *NeuroImage*, 18, 865-879.

Sundqvist, S., Leppamaki, M., Paatero, E. and Minkkinen, P., (1999), 'Application of IR spectroscopy and multivariate calibration to monitor the fusion synthesis of Ca- and Ca/Mg-resinates', *Analytica Chimica Acta*, 391, 269-276.

Svensson, O., Kourti, T. and MacGregor, J.F., (2002), 'An investigation of orthogonal signal correction algorithms and their characteristics', *Journal of Chemometrics,* 16, 176-188.

Swierenga, H., de Groot, P.J., de Weijer, A.P., Derksen, M.W.J. and Buydens, L.M.C., (1998), 'Improvement of PLS model transferability by robust wavelength selection', Chemometrics and Intelligent Laboratory Systems, 41, 237-248.

Syam, M.I., (2003), 'Cubic spline interpolation predictors over implicitly defined curves', *Journal of Computational and Applied Mathematics*, 157, 283–295.

Tamburini, E., Vaccari, G., Tosi, S., Trilli, A., (2003), 'Near-infrared spectroscopy: A tool for monitoring submerged fermentation processes using an immersion optical-fiber probe', *Applied Spectroscopy*, 57, 132-138.

Tan, H. and Brown, S.D., (2003), 'Multivariate calibration of spectral data using dual-domain regression analysis', *Analytica Chimica Acta*, 490, 291-301.

Tauler, R., Izquierdo-Ridorsa,A., Gargallo, R., Casassas, E., (1995), 'Application of a new multivariate curve resolution procedure to the simultaneous analysis of several spectroscopic titrations of the copper(II)-polynosinic acid system', Chem.Int.Lab.Systems, 27, 163-174.

Tauler, R., Smilde, A.K., Henshaw, J.M., Burgess, L.W., Kowalski, B.R., (1994), 'Multicomponent determination of chlorinated hydrocarbons using a reaction-based chemical sensor. 2. Chemical speciation using multivariate curve resolution', *Anal.Chem.*, 66, 3337-3344.

Tauler, R., Kowalski, B., Fleming, S., (1993), 'Multivariate curve resolution applied to spectral data from multiple runs of an industrial process', *Anal.Chem*, 65, 2040-2047.

Tosi, S., Rossi, M., Tamburini, E., Vaccari, G., Amaretti, A. and Matteuzzi, D., (2003), 'Assessment of in-line near-infrared spectroscopy for continuous monitoring of fermentation processes', *Biotechnol. Prog.*, 19, 1816-1821.

Troy, T. and Thennadil, S. N., (2001), 'Optical properties of human skin in the near infrared wavelength range of 1000 to 2200 nm', *Journal of biomedical Optics*, 6(2), 167-176.

Trygg, J. and Wold, S., (2002), 'Orthogonal projections to latent structures (O-PLS)', *Journal of Chemometrics* , 16, 119-128.

Trygg, J. and Wold, S., (1998). 'PLS regression on wavelet compressed NIR spectra', *Chemometrics and Intelligent Laboratory Systems*, 42, 209-220.

Vaidyanathan, S., White, S., Harvey, L.M. and McNeil, B., (2003), 'Influence of morphology on the near-infrared spectra of mycelial biomass and its implications in bioprocess monitoring', *Biotechnology and Bioengineering*, 82, 715-724.

Vaidyanathan, S., Arnold, S.A., Matheson, L., Mohan, P., McNeil, B., Harvey, L.M., (2001a), 'Assessment of near-infrared spectral information for rapid monitoring of bioprocess quality', *Biotechnology and Bioengineering*, 74, 376-388.

Vaidyanathan, S., McCaloney, G., Harvey, L.M. and McNeil, B., (2001b), 'Assessment of the Structure and Predictive Ability of Models Developed for Monitoring Key Analytes in a Submerged Fungal Bioprocess Using Near-Infrared Spectroscopy.', *Applied Spectroscopy*, **55,** 444-453.

Vaidyanathan, S., Harvey, L.H., McNeil, B., (2001c), 'Deconvolution of near-infrared spectral information for monitoring mycelial biomass and other key analytes in a submerged fungal bioprocess', *Analytica Chimica Acta*, 428, 41-59.

Vaidyanathan, S., Arnold, A., Matheson, L., Mohan, P., Macaloney, G., McNeil, B. and Harvey, L.M., (2000), 'Critical evaluation of models developed for monitoring an industrial submerged bioprocess for antibiotic production using near-infrared spectroscopy', *Biotechnol. Prog.*, 16, 1098 -1105.

Vaidyanathan, S., Macaloney, G. and McNeil, B. (1999), 'Fundamental investigations on the near-infrared spectra of microbial biomass as applicable to bioprocess monitoring', Analyst, 124(2), 157 - 162.

Vandeginste, B.G.M., (1987), 'Chemometrics - general introduction and historical development', *Topics in Current Chemistry*, 141, 24-32.

Vandeginste, B. G. M., Derks, W. and Kateman, G., (1985), 'Multicomponent self-modelling curve resolution in high-performance liquid chromatography by iterative target transformation analysis', *Analytica Chimica Acta,* 173, 253-264.

Vogel, H.C., (1997), 'Fermentation and biochemical engineering handbook', Second edition, Noyes publications.

Wang, Y.H., Yang, B., Ren, J., Dong, M.L., Liang, D. and Xu, A.L., (2005), 'Optimization of medium composition for the production of clavulanic acid by *Streptomyces clavuligerus*', *Process Biochemistry,* 40, 1161-1166.

Westerhuis, J.A., Jong, S. and Smilde, A.K., (2001), 'Direct orthogonal signal correction', *Chemometrics and Intelligent Laboratory Systems*, 56, 13-25.

Westerhuits, J.A., Gurden, S.P. and Smilde, A.K., (2000), 'Spectroscopic monitoring of batch reactions for on-line fault detection and diagnosis', *Anal.Chem*, 72, 5322-5330.

Westerhuis, J.A., Kourti, T. and MacGregor, J.F., (1998), 'Analysis of multiblock and hierarchical PCA and PLS models', *Journal of Chemometrics*, 12, 301-321.

Windig, W., (1994), 'The use of second-derivative spectra for pure-variable based self-modeling mixture analysis techniques', *Chemometrics and Intelligent Laboratory Systems*, 23, 71-86.

Wise, B. and Gallagher, N.B. (2005), 'Orthogonal signal correction', http://www.eigenvector.com/MATLAB/OSC.html

Wold, S., Trygg J., Berglund, A. and Antti, H., (2001), 'Some recent developments in PLS modeling', *Chemometrics and Intelligent Laboratory Systems*, 58, 131-150.

Wold, S., Antti, H., Lindgren, F. and Ohman, J., (1998), 'Orthogonal signal correction of near-infrared spectra', *Chemometrics and Intelligent Laboratory Systems*, 44, 75-185.

Wold, S., (1995), 'Chemometrics; What do we mean with it, and what do we want from it?', *Chemometrics and Intelligent Laboratory Systems*, 30, 109-115.

Wold, S., Kettaneh-Wold, N. and Skagerberg, B., (1989), 'Nonlinear PLS modelling', *Chemometrics and Intelligent Laboratory Systems*, 7, 53-65.

Wold, S., Martens, H. and Wold, H., (1983), 'The multivariate calibration problem in chemistry solved by the PLS method', *Proceedings Conference Matrix Pencils*. Springer Verlag, Heidelberg.

Wold, H., (1966), 'Non-linear iterative partial least squares (NIPALS) modelling: some current developments', Multivariate Analysis, Academic Press, New York.

Wolpert, D. H., (1992), 'Stacked generalization', *Neural Networks*, 5, 241-259.

Wong C. W. L., Escott , R. E. A., Morris, A. J. and Martin, E. B., (2005), 'The integration of process and spectroscopic data for enhanced knowledge extraction in batch processes', *ESCAPE (European Symposium on Computer Aided Process Engineering)*, Barcelona, Spain.

Workman, J.J., (1999), 'Quantification of LDPE [low density poly(ethylene)], LLDPE and HDPE in polymer film mixtures 'as received' using multivariate modelling with data augmentation (data fusion) and infrared, raman and near-infrared spectroscopy', *Spectroscopy Letters*, 32(6), 1057-1071.

Workman, J., Veltkamp, D.J., Doherty, S., Anderson, B.B., Creasy, K.E., Koch, M., Tatera, J.F., Robinson, A.L., Bond, L., Burgess, L.W., Bokerman, G.N., Ulman, A.H., Darsey, G.P., Mozayeni, F., Bamberger, J.A. and Greenwood, M.S., (1999), 'Process analytical chemistry', *Anal.Chem*, 71(12), 121R-180R.

Workman, J.J., (1999), 'Review of process and non-invasive near-infrared and infrared spectroscopy: 1993-1999', *Applied Spectroscopy*, 34, 1-89.

Xu, L. and Zhang, W.J. (2001), 'Comparison of different methods for variable selection', *Analytica Chimica Acta*, 446, 477-483.

Yano, T., Aimi, T., Nakano, Y. and Tamai, M., (1997), 'Prediction of the concentration of ethanol and acetic acid in the culture broth of a rice vinegar fermentation using near-infrared spectroscopy', *Journal of Fermentation and Bioengineering*, 84, 461-465.

Yeung, K.S.Y., Hoare, M., Thornhill, N.F., Williams, T. and Vaghjiani, J.D., (1999), 'Near-infrared spectroscopy for bioprocess monitoring and control', *Biotechnology and Bioengineering*, 64, 684-693.

Zeaiter, M., Roger, J. M. and Bellon-Maurel, V., (2005), 'Robustness of models developed by multivariate calibration. Part II: The influence of pre-processing methods', *Trends in analytical chemistry*, 24, 437-444.

Zeaiter, M., Roger, J. M., Bellon-Maurel, V. and Rutledge, D. N., (2004), 'Robustness of models developed by multivariate calibration. Part I: The assessment of robustness ', *Trends in Analytical Chemistry*, 23, 157-170.

Zhang, Z. and Friedrich, K., (2003), 'Artificial neural networks applied to polymer composites: a review', *Composites Science and Technology*, 63, 2029-2044.

Zhang, J., Martin, E.B., Morris, A.J. and Kiparissides, C., (1997), 'Inferential estimation of polymer quality using stacked neural networks', *Computers chem. Eng.*, 21, S1025-S1030.

Zhang, J., Martin, E.B., Morris, A.J. and Kiparissides, C., (1999), 'Estimation of impurity and fouling in batch polymerisation reactors through the application of neural networks', *Computers and Chemical Engineering*, 23, 301-314.

Zuo, K. and Wu, W.T., (2000), 'Semi-realtime optimisation and control of a fed-batch fermentation system', Computers and Chemical Engineering, 24, 1105-1109.