University of Strathclyde

Department of Chemical & Process Engineering

Spatially and Angularly Resolved Diffuse Reflectance Spectroscopy for *in-situ* Monitoring of Suspension Polymerisation Reactions

Gledson Emidio José

A thesis presented in fulfillment of the requirements

for the degree of Doctor of Philosophy

2017

Spatially and Angularly Resolved Diffuse Reflectance Spectroscopy for *in-situ* Monitoring of Suspension Polymerisation Reactions / by Gledson Emidio José.

Glasgow: University of Strathclyde, 2017.

Supervisors:

Professor Suresh Thennadil

Professor Julian Morris

Dr Paul Mulheran

This project was based at the Measurements and Analytics Team (MAT), at the Department of Chemical and Process Engineering, and was associated to the Centre for Process Analytical Technology (CPACT), Department of Pure and Applied Chemistry, University of Strathclyde.

This work was funded by the European Commission. The contents of the publication reflects only the authors view.

Declaration of author's rights

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Gledson Emidio José

June 2017

To my wife

Preface

This thesis presents the research performed during my PhD project at the University of Strathclyde. Most of the work was done within the framework of the project 'Model-Based Optimisation & Control for Process-Intensification in Chemical and Biopharmaceutical Systems' (OPTICO/ G.A. No. 280813), funded by the European Commission, as part of a wide consortium of several universities, industries and SMEs. The results presented here were developed in close collaboration between the Measurements and Analytics Team (MAT), the Centre for Process Analytical Technology (CPACT), BASF Chemicals (Ludwigshafen, Germany), and Fibre Photonics. Salary, training, travelling and outreach activities were also funded by the European Commission through a Marie Curie Research Training Fellowship, as part of the project 'Multiscale Modeling of Chemical and Biochemical Systems' (MULTIMOD ITN FP7 under grant agreement no. 238013), which brought together 8 academic research groups and 5 industrial partners from 8 European countries.

Acknowledgments

In life as in industry, the process is just as important as the product. I can certainly say this was absolutely true for my time in Scotland. Anyone at the University of Strathclyde can feel certain vibrancy in the air, and I suspect I have benefitted enormously from such environment. I would like to express my sincere gratitude to Professor Suresh Thennadil, who trusted me with freedom to follow my research and professional interests, which enhanced my experience both as a researcher and as a person, even though this sometimes came at the expense of time. I am indebted by his patience, friendship and his knowledgeable mind from which I could always tap in. Thanks to my co-supervisor, Professor Julian Morris, for opening doors, and to Dr Paul Mulheran, for taking care of all the paperwork as acting supervisor after Suresh's move to Australia. Special thanks to my studies advisor Dr Ashleigh Fletcher and to my mentor Dr Steven Ford for all the help and counselling.

I moreover owe sincere thankfulness to all past and present MAT members, especially to Dr Claudia Yi-Chieh-Chen, for being such a kind friend, and for the guidance on the instruments and probes. Likewise, I am also very grateful to Dr David Foo for coding the GUI interfaces, to Nicolau Dehanov and Dr Vítor Magueijo, who were great friends and excellent source of help at the lab and for always being ready for general discussions. Thanks to the students whose summer projects I had the pleasure to supervise: the ERASMUS French student Léa Graffin, and the Science without Borders Brazilian students Lucas Cardoso, Thomaz Campos, Felipe Ambrósio and Beatriz Abramof. The Chemical and Process Engineering Department has a great team who are always ready to help. Thanks to all technical and secretarial staff, especially to Jim Murphy, Matthew Kidd, Terry Knight, Caroline Rashid, Megan O'Donnell, and Laura Kane, and to all faculty members.

I am truly thankful to Barry Robertson, my partner during the long laboratory hours, and to Dr Alison Nordon and Professor David Littlejohn for fruitful discussions in our fortnightly meetings. Thanks to Gary Colquhoun, Trevor Whittley and David Doig, from Fibre Photonics for the Dual Reflectance probe prototype. From BASF Germany, thanks to Dr Libor Sëda for samples, analysis and guidance on the polymerisation process, and to Dr Klaus Dieter, for giving directions on the big picture. Thanks to all OPTICO and MULTIMOD partners, especially to Professor Costas Kiparissides, for leading the projects and always offering us warm Greek receptions. Thanks to all MULTIMODers for the great moments during the regular European meetings, courses and project activities.

I had a gorgeous time in Scotland, and I truly believe it was a result of the great friendships I had the pleasure to have. In special, I would like to thank Mariana Fazenda, for all the moments we shared together, either happy or sad, and for always being present. Thanks to Nicolau, Vítor, Claudia, Iain, Kaska, Rui, Tiger, Ornella and Susanne, friends who I will always carry in my heart.

Finally, thanks to my parents Emidio and Cirlei, for believing in my dreams, and to my wife, Fabiane, for walking all the steps with me.

Abstract

Despite the widespread use of optical spectroscopy for monitoring polymerisation processes, significant challenges remain for its successful application on suspension polymerisation reactions. The high heterogeneity and viscosity of such reaction media make sampling a challenging task and deteriorates the accuracy of reference and spectroscopic measurements. In this thesis, this problem is tackled by taking advantage of the stronger scattering susceptibility of the visible spectral range and the deeper penetration depth offered by the near infrared region. An investigation is carried out to evaluate whether the predictive capability of multivariate calibration models could be improved by introducing new variations in measurement configuration, in particular, spatially and angularly resolved illumination. An empirical approach based on multivariate calibration methods and chemometrics is also proposed for the extraction, pre-processing, fusing and modelling of such multi-dimensional information. How well these different measurements are integrated and how accurate the multivariate models can be, are one of the main questions of this thesis. They are first studied through a two-component system composed of polystyrene and water, followed by full suspension polymerisation reactions.

The results suggest that the accuracy of multivariate calibration models can be improved by (i) including angularly orientated fibres, (ii) fusing information from two or more source-detector separations, (iii) by how data is fused or manipulated, and (iv) by the quality of the measurements. To the best of my knowledge, this work is the first attempt to employ spatially and angularly resolved diffuse reflectance spectroscopy as a Process Analytical Technology (PAT) tool for monitoring suspension polymerisation reactions, and also the first in which a data fusion approach based on Multiblock Partial Least Squares regression (MB-PLS) is evaluated for this purpose.

Table of Contents

List of Figures

List of Tables

Notation & Units

The following notation has been used throughout the thesis. Vectors are denoted by boldface, lower-case characters, whereas matrices are denoted by boldface capital letters. A matrix inverse is denoted $\mathbf{X}^{-1}$. Vectors are assumed to be column vectors unless indicated by a transposition, e.g. $\mathbf{t}^{T}$.

B	Regression coefficients
e	Residuals vector of independent variables
E	Residuals matrix of independent variables
f	Residuals vector of dependent variables
F	Residuals matrix of dependent variables
p	Loading vector of independent variables
P	Loading matrix of independent variables
q	Loading vector of dependent variables
Q	Loading matrix of dependent variables
t	Scores vector of independent variables
T	Scores matrix of independent variables
u	Score vector of dependent variables
U	Scores matrix of dependent variables
W	Loading weights matrix of independent variables
X	Data matrix of independent variables
y **, y, Y**	Number, vector and data matrix of dependent variables

μa	Absorption coefficients
μm	Micrometres
μs	Scattering coefficients

λ	Wavelength
ε	Molar absorptivity
ℓ	Pathlength
ν	Vibrational quantum level
c	Concentration of absorbing species
cm^{-1}	Wavenumber in reciprocal centimeters
I_s, I_{is}	Sample and Reference spectra
I_{off}, I_{isoff}	Sample background and reference background spectra
k	Number of columns (or observations)
n	Number of rows (or samples)
nb	Number of blocks

Abbreviations & Symbols

API	Active Pharmaceutical Ingredient
CCD	Charge-coupled device
CI	Confidence Interval
CPACT	Centre for Process Analytical Technology
DBPO	Di-benzoyl peroxide
EMSC	Extended Multiplicative Scatter Correction
EPLS	Ensemble Partial Least Squares
GUI	Guided Interfaces
LED	Light Emitting Diodes
libPLS	Integrated Library for Partial Least Squares and Discriminant
LV	Latent Variables
InGaAs	Indium gallium arsenide
IYL	International Year of Light

ITN	Intra Networks
MAT	Measurements and Analytics Team
MATLAB	Matrix laboratory
MB-PLS	Multi-block Partial Least Squares
MIR	Medium-Infrared
MLR	Multiple Linear Regression
MSC	Multiplicative Scatter Correction
NIRS	Near infrared Spectroscopy
OD	Optical Distance
PC	Principal Components
PEEK	Polyether ether ketone
PVC	Polyvinyl chloride
PTFE	Polytetrafluoroethylene
PAT	Process Analytical Technology
PI	Process Intensification
PLS	Partial Least Squares
PCA	Principal Component Analysis
PMMA	Poly (Methyl Methacrylate)
PRESS	Predicted Residual Error Sum of Squares
PS	Polystyrene
PVA	Poly(vinyl acetate)
rdCV	Repeated Double Cross Validation
RMieS	Resonant Mie Scattering
RMS	Root Mean Square
RMSEC	Root Mean Square Error of Calibration
RMSECV	Root Mean Square Error of Cross Validation
RMSEP	Root Mean Square Error of Prediction
SARDR	Spatially and Angularly Resolved Diffuse Reflectance

SD	Source-Detector
SDBS	Sodium Dodecyl Benzene Sulfonate
SMA	SubMiniature version A
SME	Small and Medium Enterprises
SNV	Standard Normal Variate
S/N	Signal to noise
SVM	Support Vector Machines
TCP	Tricalcium phosphate
TEC	Thermo Electric Coller
Vis	Visible
WT	Wavelet Transform
UNESCO	United Nations Educational, Scientific and Cultural Organisation
UV	Ultraviolet
XRF	X-ray fluorescence
2-D	Two-dimensional

Chapter 1

1 Introduction

Despite the widespread use of optical spectroscopy for monitoring a diverse range of polymerisation processes, significant challenges remain for its successful application on suspension polymerisation reactions. The high heterogeneity and viscosity[1] in such reaction media make sampling a challenging task, and reference and spectroscopic measurements less accurate[2]. Moreover, as light is heavily scattered by polystyrene beads, spectroscopic measurements are affected in different extensions at each reaction phase. This causes a non-linear variation on their intensity, shape and baseline, which deteriorates the estimation of chemical properties by multivariate calibration models.

Spatially Resolved Diffuse Reflectance measurements (SRDR) and light propagation models based on first principles have been proposed for separating scattering effects from absorption information in particulate systems. Although such approach allows access to knowledge on the interplay of scattering and absorption, it involves complex mathematical operations[3] and intensive computational calculations[4], most of the times unsuitable for real-time applications[5], where fast and autonomous analyses are preferred.

A less rigorous approach for modelling particulate media with spatially resolved diffuse reflectance spectroscopy is based on the development of empirical multivariate calibration models on the averaged signal from a number of source-detector distances or combinations of it[6]. This strategy, however, may hinder the extraction of spatial resolution, which is the focal point of the technique.

In another slightly different approach, but which would potentially preserve the information regarding different penetration depths and scattering behaviour, instead of using the co-added measurement, the information from individual source-detector distances or combinations of it[6,7] are fused together, being such information represented by spectra or by linear combinations of it, the score values. These combined measurements may be based on raw spectra or on measurements subjected to empirical light scattering pre-processing. This approach, although not separating the absorption and scattering components, has the potential to provide real time measurements is a fast way, and may allow calibration models less affected by poor signal-to-noise measurements than the first principles based approach.

In this thesis, this problem is addressed by integrating spectral datasets collected at different angles, source-detector distances and fibre geometries. The combined information should reveal more insight into the sample than that obtained if each spectral dataset is analysed individually. A data fusion strategy based on Multiblock Partial Least Squares (MB-PLS) will be explored to yield a more accurate estimation of critical quality parameters of polymer beads, modelling the relationship between the different source-detector separations and angles of incidence, and establishing the contribution of each of them to the final model. Specifically, the aim is to evaluate the sensing abilities of both visible and Near-Infrared (NIR) spectroscopy, the possible benefits of including information from angularly orientated fibres to the combinations of normal incident ones and the impact on the predictive capability of the multivariate calibration models. How well the data fusion strategies will cope with the heterogeneity inherent to such reactions is also of great interest. To the best of my knowledge, this is the first attempt to employ spatially and angularly resolved diffuse reflectance spectroscopy (SARDR) for monitoring suspension polymerisation reactions.

1.1 Motivation

We have recently celebrated the UNESCO International Year of Light and Light-based Technologies (IYL 2015), a worldwide series of events showing how important optical technologies are in our daily lives[8]. In fact, light-based technologies are vital for existing and future advances in a range of fields of knowledge, from laser surgeries in medicine[9] to Light-Emitting Diode (LED) illumination in farming [10], from chemistry to data storage[11].

In manufacturing industries, light-based technologies are in the spotlight, among other reasons, by allowing analytical measurements to be made real-time, a condition that is central to Process Analytical Technology (PAT)[12] and Process Intensification (PI) approaches to quality.

In the polymer industry, which is the particular case studied in this thesis, the motivation for the implementation of PAT and PI tools to monitor and control of polymerisation reactions is manifold. Besides the economic incentives, new worldwide market requirements are being put in place for tighter product specifications and stricter safety and environmental constraints[13]. These incentives are moving polymer manufacturing from a post-production quality control set-up towards a quality by design approach, where the quality is directly built into process design[12]. Especially for product-by-process materials like polymers, where the quality and properties of the final product is largely dictated by the reaction profile, comprehensive process and product knowledge are important to obtain a reliable process profile that is less likely to have quality issues, supplying products with improved performance[1]. The rationale behind this evolution is that by gathering more information about a process, we can better understand it and, consequently, be better prepared to monitor and control it.

1.2 PAT in the polymer industry: spectroscopic sensors for monitoring polymerisation processes

Fibre-optic spectroscopic sensors are key enablers of PAT and PI implementations in industry. Indeed, spectroscopic techniques like near infrared and Raman are used almost as a synonym for PAT. Such techniques have been successfully implemented in a varied range of polymerisation process, among other reasons, for allowing measurements to be made without handling the reaction media outside the reaction vessel, an essential requirement for monitoring of polymerisation reactions due to their high viscous and unstable nature. Furthermore, together with advanced multivariate data analysis methods[14,15], spectroscopic sensors provide simultaneous determination of several process quality attributes, all in a fast, reliable and cost efficient way[16]. Such characteristics are especially important to allow rapid decisions to be made on time whenever deviations are found, or for fast transition between different operation conditions to reduce off-spec production[1]. This ultimately may allow real-time release of batches, saving resources by speeding up the production cycle[17].

Important quality attributes monitored in polymerisation reactions are the conversion degree of monomer into polymer and the size of the produced polymer beads. Monitoring monomer conversion is relevant because it is directly related to productivity. Particle size monitoring is also important for economic reasons, since fractions too below or above the required diameter range lead to losses of product and resources in the post-reaction processing. Furthermore, as the intended applications of polystyrene beads are dependent on their diameter, polystyrene is normally produced at tailor made diameter fractions on demand.

Monomer conversion and bead size have been successfully monitored in a range of polymerisation reactions by different spectroscopies, as example of the ultraviolet-visible (UV-VIS)[18,19], near infrared (NIR)[20], middle infrared (MIR)[21] and Raman spectroscopies. Especially, NIR and Raman have been extensively studied for monitoring monomer conversion, whereas the visible and NIR spectroscopy for mean bead size monitoring.

Emulsion polymerisation has been by far the most studied polymerisation regarding the application of spectroscopic methods. Early works[22] employed off-line or at-line measurements, and were mainly based in one or a few wavelengths. Although good models can be obtained by using univariate models, especially for Raman, such models tend to be biased due to the heterogeneity of the reactions[23].

In the last decade, several applications of on-line methods started to be proposed. Reis *et al*[27] published a comprehensive comparison of Raman and NIR for a range of homo- and co-polymerisation emulsion reactions. Although not pointing out which technique is better, they suggest that both techniques are suitable for on-line monitoring. Recently, by comparing Raman and NIR for monitoring a mini-emulsion polymerisation, Ambrogi *et al*[24] identified NIR spectroscopy as a reliable tool for monitoring simultaneously monomer conversion and particle size.

In particular, NIR spectroscopy has been widely applied to monitor conversion on several different polymerisation processes. Applications were developed for emulsion copolymerisation and polymerisation reactions[25,26], bulk[27] and solution polymerisation of methyl methacrylate[28], polyurethane step-growth polymerization[29], ethane/1-hexene solution copolymerization[30], and many others. However, applications in styrene suspension polymerisations are scarce. Lousberg *et al*[31] employed short-NIR online measurements to determine conversions up to 35% in bulk styrene batch polymerisation reactions. Although the results were good, the

conversion range was limited and a small bias in low and high monomer conversion was detected. Silva *et al* [32]used a NIR interactance probe to monitor the evolution of monomer and copolymer concentrations during batch acrylic acid and vinyl acetate suspension co-polymerisations. They were able to develop calibration models based on PLS giving good estimations.

Santos *et al*[33] developed PLS cross-validation models relating in-line measured NIR spectra and mean bead diameter. Although the authors did not validate the models using external reactions, the work was important since it was one of the first to take advantage of the NIR sensitivity to changes in particle size distributions. Later, the same authors[34] developed a control strategy for monitoring mean bead size along the reaction.

Faria Junior *et al*[35,36] devised a control strategy for control of Polyvinyl Chloride (PVC) particle size and morphological properties of its resins in suspension polymerisations. The authors successfully predicted dynamic trajectories of morphological properties based on NIR, which can be used as reference for monitoring and control during batch operation.

There is a no work in the literature that is concerned to the application of spatially resolved measurements to monitor polymerisation reactions of any kind, despite its advantages compared to the standard single point diffuse reflectance NIR measurements. In particular, its ability to probe the beads at different points and angles, improving the amount of information gathered, and potentially providing a more comprehensive assessment of the reaction components.

1.3 Objectives and Contributions

The overarching goal of the present thesis is to propose spatially and angularly resolved diffuse reflectance spectroscopy as a PAT tool for *in-situ* monitoring of polymerisation reactions. The aim is to investigate whether the individual or combined information gathered from spectroscopic measurements collected from different wavelength regions, fibre geometries, source-to-detector distances and angles of incidence could improve the predictive capability of multivariate calibration models.

In order to fulfil this objective, the aim is to develop the required multivariate calibration and data fusion strategies for the extraction, interpretation, visualisation, fusing and modelling of information collected, first, from a two-component system composed of polystyrene and water, and later from full suspension polymerisation reactions.

The following are the primary contributions of the present thesis:

- Implementation of the data acquisition routines and data processing strategies for the optimum extraction of information from spectroscopic measurements;
- Development of statistical multivariate calibration modelling strategies for the estimation of bead size and concentration in a two-component system and monomer conversion in full suspension polymerisation reactions;
- Implementation of a data fusion modelling strategy, evaluating its potential for integrating spectroscopy data, and the benefits of doing so;
- Implementation of a robust validation scheme based on the repeated double cross validation and on external data, studying the several factors which can affect the predictive capability of the models;
- Evaluation of alternative optical probe configurations and their potential for estimation of polymer quality parameters.

The secondary contributions are:

- Implementation of methods and protocols for running suspension polymerisations at the Measurements and Analytics Team Laboratory;

- Implementation of the methods and protocols for collecting reference and spectroscopic measurements from polymerisation reaction;

1.4 Thesis Overview

This thesis is organised in eight chapters. The first introduced the thesis objectives and contributions, presenting why and how spectroscopy is being adopted by the polymer industry as an important Process Analytical Technology tool.

The second chapter covers the theoretical background on optical sensors and chemometrics, whereas Chapter 3 presents the proposed spectroscopic measurement system, describing the prototype optical probes, experiments, equipments, and methodologies used to evaluate the models and sensors proposed in this thesis.

In Chapters 4 and 5 the results based on a two-component system are described, in an attempt to obtain a better understanding of the spectra behaviour and the modelling strategies studied here. Chapter 4 studies the impact of each individual source-detector distance on the predictive models developed for visible and near infrared spectroscopy, whereas in Chapter 5 the fusion of two or more source-detector distances and angles of incidence are evaluated for both spectroscopies.

In Chapter 6, the modelling strategies studied in Chapter 4 and 5 are applied to the monitoring of conversion on suspension polymerisation reactions. Models are first

evaluated by fusing individual normal incident source-detector distances, and later the impact of angularly resolved information is evaluated.

In Chapter 7, potential improvements in the spectra acquisition are evaluated by proposing alternative fibre geometry. For this, a prototype dual-ring diffuse reflectance probe is evaluated by using the two-component system, and the results compared to the ones obtained by using the SARDR probe.

Chapter 8 completes the thesis by listing the overall conclusions, making considerations based on all results and setting up directions for future work.

1.5 Bibliographic references

1. Leiza, J. R. & Pinto, J. C. in *Polymer Reaction Engineering* 315–362 (Blackwell Publishing Ltd, 2007). doi:10.1002/9780470692134.ch8

2. Santos, J. C. *et al.* Online Monitoring of Suspension Polymerization Reactions Using Raman Spectroscopy. *Ind. Eng. Chem. Res.* **43,** 7282–7289 (2004).

3. Thennadil, S. N., Martens, H. & Kohler, A. Physics-based multiplicative scatter correction approaches for improving the performance of calibration models. *Appl Spectrosc* **60,** 315–321 (2006).

4. Steponavicius, R. & Thennadil, S. N. Extraction of chemical information of suspensions using radiative transfer theory to remove multiple scattering effects: application to a model multicomponent system. *Anal. Chem.* **83,** 1931–1937 (2011).

5. Farrell, T. J., Wilson, B. C. & Patterson, M. S. The use of a neural network to determine tissue optical properties from spatially resolved diffuse reflectance measurements. *Phys. Med. Biol.* **37,** 2281–2286 (1992).

6. Igne, B., Talwar, S., Feng, H., Drennen, J. K. & Anderson, C. A. Near-Infrared Spatially Resolved Spectroscopy for Tablet Quality Determination. *Journal of Pharmaceutical Sciences* **104,** 4074–4081 (2015).

7. Bogomolov, A., Belikova, V., Galyanin, V., Melenteva, A. & Meyer, H. Reference-free spectroscopic determination of fat and protein in milk in the visible and near infrared region below 1000 nm using spatially resolved diffuse reflectance fiber probe. *Talanta* **167,** 563–572 (2017).

8. Assembly, G. & Nations, U. *Resolution adopted by the General Assembly on 20 December 2013*

9. Roxby, P. How lasers will light up the future of medicine. *BBC News* (2010).
10. Kozai, T., Fujiwara, K. & Runkle, E. S. *LED Lighting for Urban Agriculture*. (Springer, 2016).

11. Haw, M. Holographic data storage: The light fantastic. *Nature* **422,** 556–558 (2003).

12. Felizardo, P., Folque, F., Machado, J. E. & Menezes, J. C. Process analytical technology: a common approach across different industries. *NIR news* (2012).

13. Araújo, P. H. H., Sayer, C., Giudici, R. & Poço, J. G. R. Techniques for reducing residual monomer content in polymers: A review. *Polym Eng Sci* **42,** 1442–1468 (2004).

14. MacGregor, J. F., Jaeckle, C., Kiparissides, C. & Koutoudi, M. Process monitoring and diagnosis by multiblock PLS methods. *AIChE Journal* **40,** 826–838 (1994).

15. Menezes, J. C., Ferreira, A. P., Rodrigues, L. O. & Brás, L. P. *Chemometrics role within the PAT context: Examples from primary pharmaceutical manufacturing. Edited by S. Brown, R. Tauler, and B. Walczak*. (Compr. Chemom, 2010).

16. Huck, C. The future role of near infrared spectroscopy in polymer and chemical analysis. *NIR news* **27,** 17–23 (2016).

17. Chen, Z., Lovett, D. & Morris, J. Process analytical technologies and real time process control a review of some spectroscopic issues and challenges. *Journal of Process Control* **21,** 1467–1482 (2011).

18. Celis, M.-T. & Rubio, L. H. G. Spectroscopy as a Tool for Characterization of Monomer Emulsions. *Journal of Dispersion Science and Technology* **28,** 271–278 (2007).

19. Cardenas-Valencia, A. M. & García-Rubio, L. H. Reflectance Spectroscopy Probes for Continuous Process Monitoring of Particle Size, Concentration and Chemical Composition. in 14th International Congress of Chemical and Process Engineering.Prague, Czech Republic,27-31 August 2000.

20. Florenzano, F. H., Enohnyaket, P., Fleming, V. & Reed, W. F. Coupling of near infrared spectroscopy to automatic continuous online monitoring of polymerization reactions. *European Polymer Journal* **41,** 535–545 (2005).

21. Pasquale, A. J. & Long, T. E. Real-Time Monitoring of the Stable Free Radical Polymerization of Styrene via in-Situ Mid-Infrared Spectroscopy. *Macromolecules* **32,** 7954–7957 (1999).

22. Brookes, A., Dyke, J. M., Hendra, P. J. & Strawn, A. The investigation of polymerisation reactions in situ using FT-Raman spectroscopy. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **53,** 2303–2311 (1997).

23. Reis, M. M., Araújo, P. H. H., Sayer, C. & Giudici, R. Comparing near infrared and Raman spectroscopy for on-line monitoring of emulsion copolymerization reactions. *Macromol. Symp.* **206,** 165–178 (2004).

24. Ambrogi, P. M. N., Colmán, M. M. E. & Giudici, R. Miniemulsion Polymerization Monitoring Using Off-Line Raman Spectroscopy and In-Line NIR Spectroscopy. *Macromolecular Reaction Engineering* **31,** 443 (2016).

25. Dehanov, N. & Thennadil, S. Monitoring of conversion during emulsion polymerisation using near infrared spectroscopy. *NIR news* **25,** 9 (2014).

26. Reis, M. M., Araújo, P. H. H., Sayer, C. & Giudici, R. In Situ Near-Infrared Spectroscopy for Simultaneous Monitoring of Multiple Process Variables in Emulsion Copolymerization. *Ind. Eng. Chem. Res.* **43,** 7243–7250 (2004).

27. Aldridge, P. K., Kelly, J. J., Callis, J. B. & Burns, D. H. Noninvasive monitoring of bulk polymerization using short-wavelength near-infrared spectroscopy. *Anal. Chem.* **65,** 3581–3585 (1993).

28. Cherfi, A. & Févotte, G. On-line conversion monitoring of the solution polymerization of methyl methacrylate using near-infrared spectroscopy. *Macromolecular Chemistry and Physics* **203,** 1188–1193 (2002).

29. Nogueira, E. S., Borges, C. P. & Pinto, J. C. In-Line Monitoring and Control of Conversion and Weight-Average Molecular Weight of Polyurethanes in Solution Step-Growth Polymerization Based on Near Infrared Spectroscopy and Torquemetry. *Macromolecular Materials and Engineering* **290,** 272–282 (2005).

30. Kappler, B. *et al.* Real-time monitoring of ethene/1-hexene copolymerizations: determination of catalyst activity, copolymer composition and copolymerization parameters. *Polymer* **44,** 6179–6186 (2003).

31. Lousberg, H. H. A., Boelens, H. F. M., Le Comte, E. P., Hoefsloot, H. C. J. & Smilde, A. K. On-line determination of the conversion in a styrene bulk polymerization batch reactor using near-infrared spectroscopy. *J. Appl. Polym. Sci.* **84,** 90–98 (2002).

32. Silva, F. M., Lima, E. L. & Pinto, J. C. Control of the Copolymer Composition in Suspension Copolymerization Reactions. *Ind. Eng. Chem. Res.* **43,** 7312–7323 (2004).

33. Santos, A. F., Lima, E. L. & Pinto, J. C. In-line evaluation of average particle size in styrene suspension polymerizations using near-infrared spectroscopy. *J. Appl. Polym. Sci.* **70,** 1737–1745 (1998).

34. Santos, A. F., Lima, E. L. & Pinto, J. C. Control and design of average particle size in styrene suspension polymerizations using NIRS. *J. Appl. Polym. Sci.* **77,** 453–462 (2000).

35. de Faria, J. M., Jr., Machado, F., Lima, E. L. & Pinto, J. C. In-Line Monitoring of Vinyl Chloride Suspension Polymerization with Near Infrared Spectroscopy, 2 - Design of an Advanced Control Strategy. *Macromolecular Reaction Engineering* **4,** 486–498 (2010).

36. de Faria, J. M., Jr., Machado, F., Lima, E. L. & Pinto, J. C. In-Line Monitoring of Vinyl Chloride Suspension Polymerization with Near-Infrared Spectroscopy, 1 - Analysis of Morphological Properties. *Macromolecular Reaction Engineering* **4,** 11–24 (2010).

Chapter 2

2 Spectroscopy methods & multivariate data analysis

This chapter covers the theoretical background on spectroscopy and multivariate data analysis. It is divided into three parts. In the first one, after a short historical overview, the theoretical bases of visible and NIR spectroscopies are presented, introducing how light scattering in particulate systems affects the predictive capability of multivariate calibration, and describing the methodologies involved in its correction. The practical aspects for their implementation are covered at the end of this first part. The second part presents the multivariate data analysis methods and approaches used for extracting chemical and physical information from spectral measurements, whereas the procedures taken for the validation of these models are described in the third part.

"Measure what can be measured and make measurable what cannot be measured" Galileo Galilei

2.1 Visible and near-infrared spectroscopy

Throughout this thesis, I will work specifically with radiation in the wavelengths within the electromagnetic spectrum range of 500 to 1800 nm, which includes the visible, short-near infrared, and part of the near-infrared region. By choosing this region, I want to take advantage of the stronger scattering susceptibility of the visible spectral range, and the deeper penetration depth offered by the NIR region, in an attempt to simultaneously extract physical and chemical information from a particulate system. I envisage to combine spatial and angular information collected as multi-point measurements, either by individual spatially and angularly resolved fibres or as rings of fibres. The purpose is to obtain a more complete picture of the particulate system studied here, and ultimately, improve the predictive capability of multivariate calibration models.

Before presenting the theoretical and practical aspects of the visible and NIR diffuse reflectance spectroscopies and its associated modelling approaches, I will first introduce the shared historic context of NIR spectroscopy and chemometrics.

2.1.1 A bit of history

Visible spectroscopy has been used for analytical measurements well before the near-infrared was discovered for this purpose. Its origins go back to Sir Isaac Newton's experiments with a prism, but it was only around 1860 that visible spectroscopy was employed in chemical analysis, when Gustave Kirchhoff stated that any substance that was a good light emitter of some particular wavelength would also absorb light of the same wavelength[1].

While visible spectroscopy and even mid-infrared spectroscopy were gaining acceptance as analytical techniques, the NIR range was merely a not well-understood region between the visible and infrared, and of not much use to analysts[2]. The absence of definite peaks, as opposed to infrared spectroscopy (IR), was certainly a limiting factor, especially because analysts were accustomed to looking at the peaks in order to give inferences about quantity and presence of a certain chemical compound. Another reason why NIR spectral region took so long to be perceived as a relevant analytical tool was the lack of proper instrumentation[2].

Although few applications involving NIR measurements can be found in the literature in the early twentieth century, NIR spectroscopy only became a trend as analytical technique in the early 1970's, thanks to the work of Karl Norris from the United Stated Department of Agriculture (USDA), on agricultural particulate samples[3]. His group pioneered the use of diffuse reflectance, working on the modeling approaches and instrumental modifications needed for its implementation.

Phil Williams, who recently was awarded the Karl Norris award in recognition of his major contribution to the science of near-infrared spectroscopy[4], was one of the first to use a commercial NIR instrument, but also contributed with many developments in research and development of new NIR technologies[5]. Rapidly the usage of NIR

spectroscopy spread to other areas like fuel, food, polymers, and especially the pharmaceutical sector, where the technique gained maturity.

Advances in instrumentation and computational power contributed to the maturity of the field. Instrumental noise, accuracy and sensitivity were greatly reduced in modern spectrometers, especially due to better manufacturing of its components, like detectors, power supply, and circuitry. As the first NIR spectrometers were filter based, calibrations were mostly done by Multiple Linear Regression (MLR) on a few wavelengths[6]. With the development of scanning spectrometers, the Vis and NIR wavelength range was fully available for calibration, paving the way to a new revolution introduced by Harald Martens, Svante Word and Bruce Kowalski, on the PLS regression and related chemometrics techniques. Improvements on the hardware came accompanied by the development of integrated and dedicated data analysis software developed in most cases by spin-out companies from the academic world, like CAMO (Unscrambler®, from Martens), Umetrics (SIMCA®, from Word), and Infometrics (ARTHUR® and Piroutte®, from Kowalski), to name a few.

NIR development always came hand-in-hand with chemometrics. Indeed, it can be said that chemometrics was the driving force that brought NIR-based methods from theory to the industrial plant floor. The development and maturity of chemometrics and NIR spectroscopy are a prime example of a successful industry-academia partnership. NIR was reborn in industry and grew up from industry-academia partnerships, like the important initiatives of the Centre for Process Analysis & Control (CPAC-Washington) and the Centre for Process Analytical and Control Technology (CPACT-UK), from which the latter this thesis is associated. These consortia played a crucial role in spreading the NIR technology for a varied range of industries around the world, and helped such fields to reach the actual levels of maturity.

2.1.2 Principles

As with most of the spectroscopic techniques, difference in energy, frequency or wavelength, sets the visible and NIR apart[7], and dictates most of their properties. Figure 2.1, together with equations 2.1 and 2.2 show how the energy is related to wavelength and frequency along the electromagnetic spectrum from ultraviolet to infrared, and highlights the spectral region used throughout this thesis.

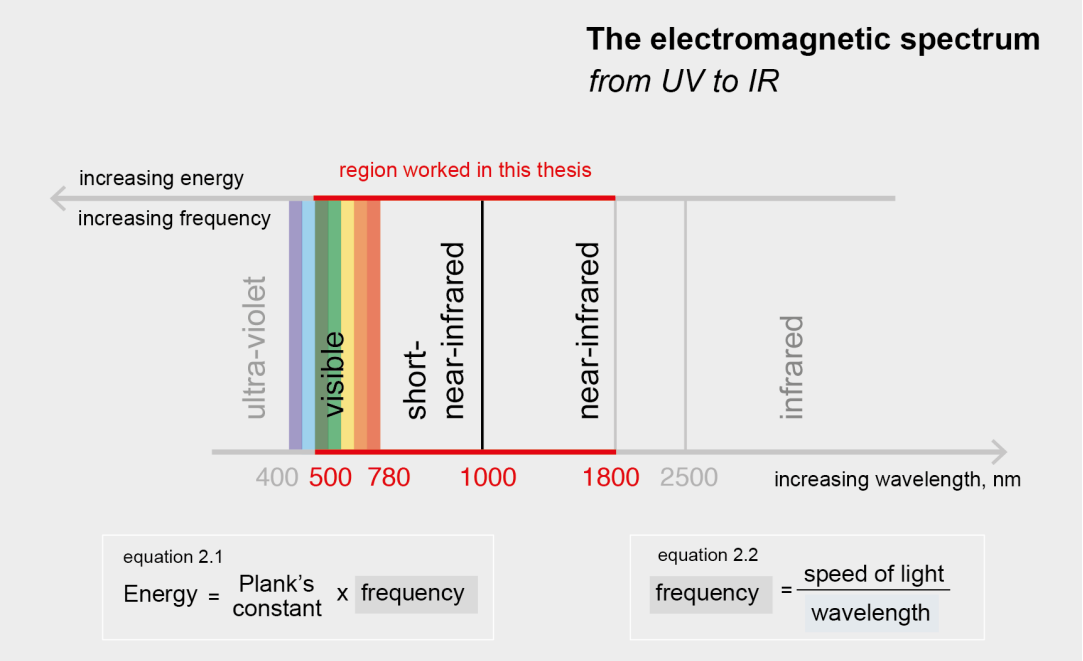

Figure 2.1 The electromagnetic region from ultra-violet to infrared. The range of interest studied in this thesis is highlighted in red. Plank's constant = 6.626×10^{-34} J.s. Speed of light = 2.998×10^{8} m/s

As can be seen from Figure 2.1, the energy in the visible region is higher than in the NIR region, which, in its turn, is higher than in the mid-infrared. The opposite happens with wavelength, as it is inversely proportional to frequency, and consequently, to energy (Equations 2.1 and 2.2). Despite being neighbour regions, the energy differences in the Vis and NIR are sufficient to provoke distinctive properties. Figure 2.2 illustrates the impact of energetic levels in the vibrational transitions of these three regions.

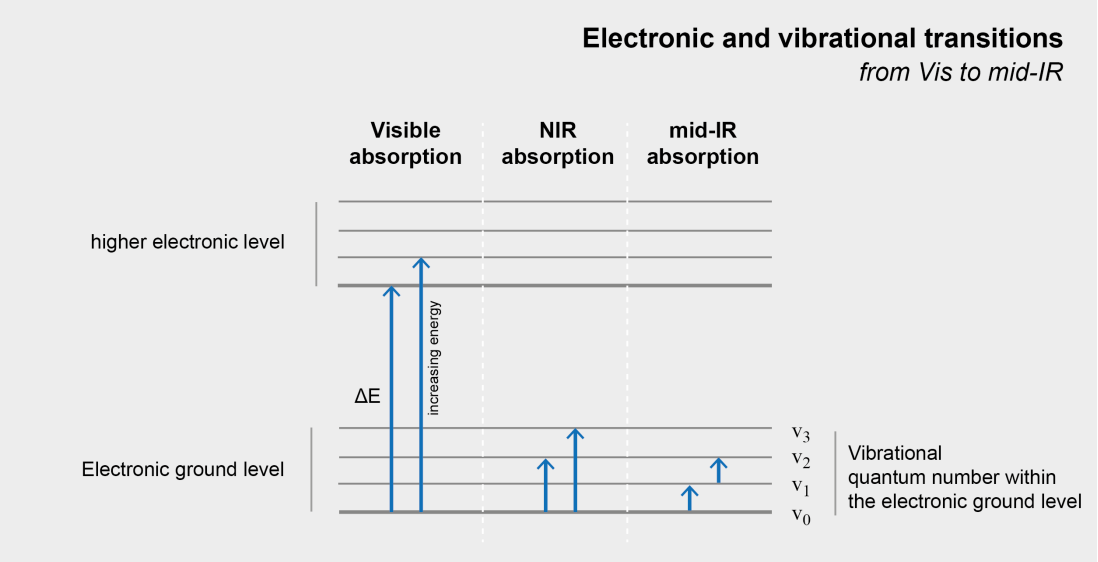

Figure 2.2 Absorption of radiation by a molecule in the visible, near-infrared and mid-infrared. Blue arrows indicate the transition between different vibrational and electronic levels.

When a molecule absorbs infrared radiation, the atoms start to vibrate by simultaneously increasing and decreasing their bond lengths. Such change in the vibrational state excites the molecular bond from a resting energy level to a discrete higher quantum level[3]. The absorption only occurs if the difference in energy between the resting level and the excited state is equal to the photon's energy.

Generally, the energy in the mid-infrared is not strong enough to excite bonds to further higher vibrational quantum levels. Consequently, the absorptions in the mid-infrared region are mainly due to fundamental vibrations, which involves a transition between the ground state and the first vibrational quantum level[8]. Transitions to higher energy levels may happen due to anharmonicity, since the energy between levels can vary. This give rise to overtone bands, which in the mid-infrared region occur at approximately twice the fundamental frequency (first overtone), but are predominantly found in both visible and near-infrared regions.

In the visible region, as the energy required to excite a molecule is much higher, transitions are not restricted only to the vibrational level. When radiation is absorbed, one of the electrons of valence is promoted to a more energetic electronic level[7,9] (Figure 2.2). The transition energy will depend on the types of orbitals these valence electrons move from and to. The superposition of rotational and vibrational transitions on the electronic transitions gives rise to combination bands, which together with overtones explain the characteristic broad and superimposed bands in the Vis range. To illustrate how the fundamental absorptions are related to overtones and combination bands, the aromatic C-H bond in styrene will to be used as example. Figure 2.3 presents the near-infrared and mid-infrared spectra of styrene. For the sake of clarity, only a short mid-infrared wavelength range is shown. For the mid-infrared, the wavenumber (units of cm^{-1}) is preserved as a frequency unit, which is usually used by the infrared community, but the correspondent wavelengths (units of nm), usually used in the NIR, are also shown. The difference in scale compared to the NIR wavelengths can also be noted. Wavenumbers and wavelengths are related by equation 2.3, shown in Figure 2.3.

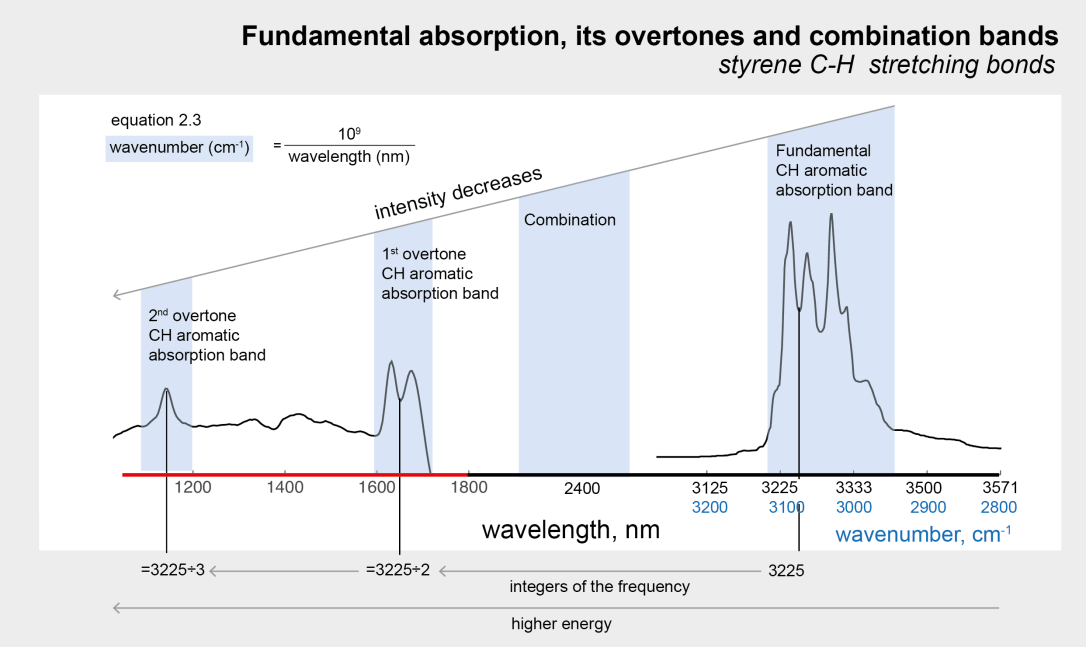

Figure 2.3 Relationship between fundamental, combination and overtone absorption. The case of styrene C-H aromatic stretching bonds.

The aromatic C-H bond vibrates by stretching, generating a few weak-to-moderate bands centred at around 3000 cm^{-1}. By exciting the same molecule with radiation in the near-infrared region, the C and H bonds will vibrate by stretching as well, but as the near-infrared light carries more energy, an absorption band (a first overtone) will appear in the NIR spectrum at 1680 nm[10], twice the wavenumber of the fundamental (2 x 3000 cm^{-1} = 6000 cm^{-1} = 1680 nm), and its second overtone at 1143 nm (3 x 3000 cm^{-1} = 9000 cm^{-1} = 1143 nm). Most of the overtone bands in the NIR spectrum arise from X-H stretching modes (O-H, C-H, S-H, and N-H). It is important to note that only the vibrations that cause a change in the dipole moment of the molecule give rise to absorptions, and the larger the dipole moment is, stronger the intensity of the absorption band will be. Therefore, the absorption intensity will usually decrease by a factor of 10 to 100 from the fundamental to the subsequent overtones[11], since

these transitions are quantum mechanically forbidden, and consequently, have a lower probability to happen[12].

Another type of absorption band found in the near-infrared region is the combination band, which are sum of difference bands due to transitions from energy states higher than the ground state[8]. A combination band happens at around 2400 nm for the C-H aromatic bond, therefore, outside the working range studied in this thesis.

The frequency or wavelength at which a molecule absorbs energy is dependent on the type of the vibrational and electronic transitions happening, and the amplitude of such absorption is determined by its absorptivity and how many molecules can be encountered within the light beam path. From that follows the Lambert-Beer Law, which states that the concentration of a substance in solution is directly proportional to the absorbance of the solution. Figure 2.4 presents the equations for the calculation of absorbance by the Beer's law and shows how the absorbance is related to the transmittance.

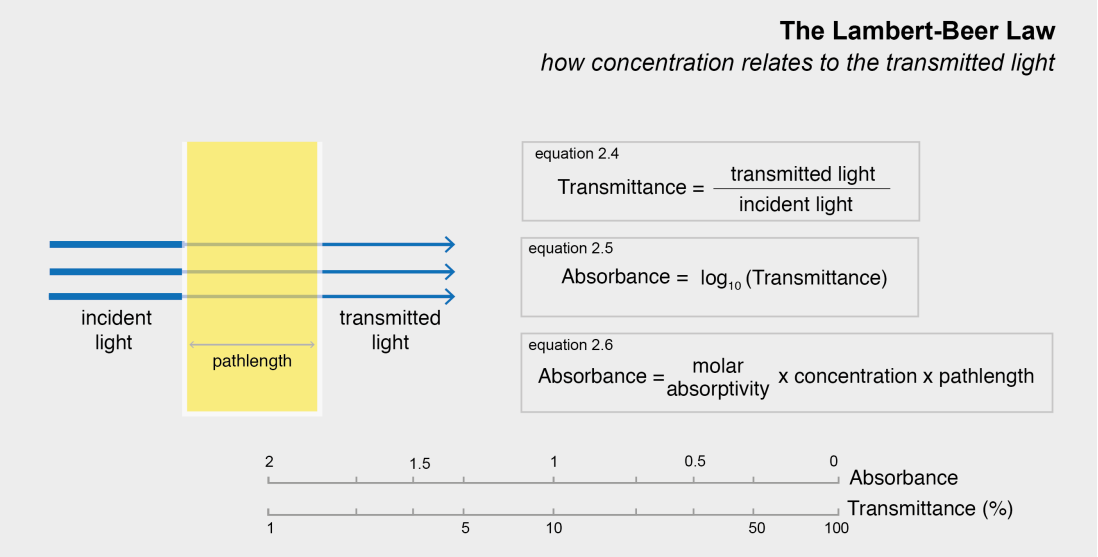

Figure 2.4 Schematic illustration of the Lambert-Beer Law and equations for converting transmittance to absorbance.

2.2 Absorption and scattering in particulate systems

As discussed in the last section, absorption of light predominantly generates electronic transitions if it happens in the visible electromagnetic region, or molecular vibrations in the wavelengths within the near-infrared region. These electronic or vibrational transitions are then translated into spectral bands, which are the visual assignment of what is happening at the electronic and molecular levels and from which the concentration of the constituents present in a sample are deduced through the use of Lambert-Beer's law.

The problem is that not all light is absorbed, some photons are scattered or transmitted. Indeed, visible and near-infrared light propagation in particulate and turbid systems like the polystyrene suspension is generally dominated by scattering, which happens due to differences in the refractive index caused by variations in bead size and concentration (packing). This introduces non-linearity in the measurements, since light may travel different paths lengths, deteriorating the predictive capability of multivariate calibration models.

When a beam of light is incident upon polystyrene beads suspended in water, for example, it penetrates the surface of the polystyrene beads for a short depth, excites the aromatic C-H bonds of the molecules of styrene, and is then scattered at all angles due to refractive index mismatches between the beads and the voids between them. The average of the number of times a photon changes it direction can be described by the scattering coefficient, μs, which measures the probability of scattering per unit length. The absorption component, similarly, can be described by the absorption coefficients, μa, which measures the probability of absorption per unit path length.

Given that the scattering and absorption optical properties contain rich information on physical and chemical characteristics of the suspension, extracting them from

diffuse reflectance measurements may improve our understanding of the suspension as a whole, providing better conditions for monitoring such systems. However, scattering and absorption properties are closely intertwined, so their extraction from diffuse reflectance spectral measurements is not a straightforward task. Most approaches aiming for it are based on light propagation models and a combination of multiple reflectance measurements.

The propagation of light in any media can be described by solving the Maxwell equations. Because it requires complex and intensive calculations, especially in situations of multiple scattering, simplified models are usually performed based on the Radiative Transfer Equation (RTE) or approximations derived from it. These implementations require less computational time and provide results accurate enough for many applications[13-15]. These methods require a combination of multiple reflectance measurements, like reflectance spectra taken at varying distances from the detector (spatially resolved), different time frames (time resolved), or multiple collection modes (diffuse transmittance, diffuse reflectance and collimated transmittance), as the ones collected by an integrating sphere[16]. The rationale is that absorption and scattering will be affected differently in each of these measurements, and that light propagation models are able to account for such differences.

The integrating sphere is the standard method because it provides accurate estimation of the optical properties. However, it is not suitable for real-time measurements. Time-resolved spectroscopy has provided interesting applications, but the instrumentation is complicated and expensive. Spatially resolved diffuse reflectance measurements have been extensively investigated for extracting optical properties in medical applications[17-20], especially due to its simple instrumentation, which is relatively inexpensive and easy to use.

2.2.1 Spatially resolved diffuse reflectance measurements

The determination of optical properties from spatially resolved spectral data is normally based on the use of at least two channels at different source-detector distances, from which the absorbance coefficients are computed on the absorbance spatial gradient[21]. Spatially Resolved Spectroscopy (SRS) can be used either as a multi-probe or imagery-sensing mode.

Although the first spatially resolved spectra were measured almost seventy years ago[22], it was only in the 90's that Vis-NIR-Spatially Resolved Spectroscopy emerged as an important on-line analytical sensor. The dissemination of the technique, however, increased after Farrell et al proposed an analytical diffusion equation solution which provided straightforward and accurate separation of absorption and scattering coefficients (μa and μs, respectively) from a single measurement of the spatially resolved diffused reflectance[23]. More recently, inspired on the developments on the medical sector, interesting applications started to be developed for quality control of food[24,25] and agricultural products[26,27,28-30], and in the pharmaceutical industry [31 32].

Although these developments allow access to knowledge on the interplay of scattering and absorption in such systems, many difficulties related to the spectroscopic measurements and due to the light propagation models[33] have been faced when evaluating then for quantitative prediction of quality attributes from particulate samples. Moreover, the measurement and processing time is yet to be reduced sufficiently for real-time applications, which is important for translating these techniques to other fields which require faster processing times, as in process analysis.

Another less rigorous approach for modelling particulate media with spatially resolved diffuse reflectance is based on the development of empirical multivariate

calibration regression models on the combined or fused measurements from the individual source-detector distances[34,35] and angles. The combined measurement may capture the information regarding different penetration depths and scattering behaviour in a more efficient way than using measurements from an individual SD distances, as in standard NIR spectroscopy. They may be based on raw spectra or based on measurements subjected to empirical light scattering pre-processing. This approach, although not separating the absorption and scattering components, can provide real time measurements is a fast way, and may allow calibration models less affected by poor signal-to-noise measurements than the first principles based approach.

2.2.2 Empirical scatter correction methods

Another commonly used approach for accessing the absorption and scattering information from spectral measurements is the application of empirical pre-processing methods, such as Standard Normal Variate (SNV), Multiplicative Signal Correction (MSC) and Extended Multiplicative Signal Correction (EMSC)[36]. As these methods do not take into account the propagation of light in the sample, they do not completely separate the absorption and scattering components. Instead, they attempt to correct the signal against the scattering component, which for particulate samples is represented by systematic variations in the spectra caused by light scattering and differences in the path length travelled by the photons. Take for example the spectra collected from polystyrene beads suspended in water at varying bead size diameters, as shown in the Figure 2.5. The spectral offset along the whole spectral region is evident, as it is the baseline inclination as the wavelength increases.

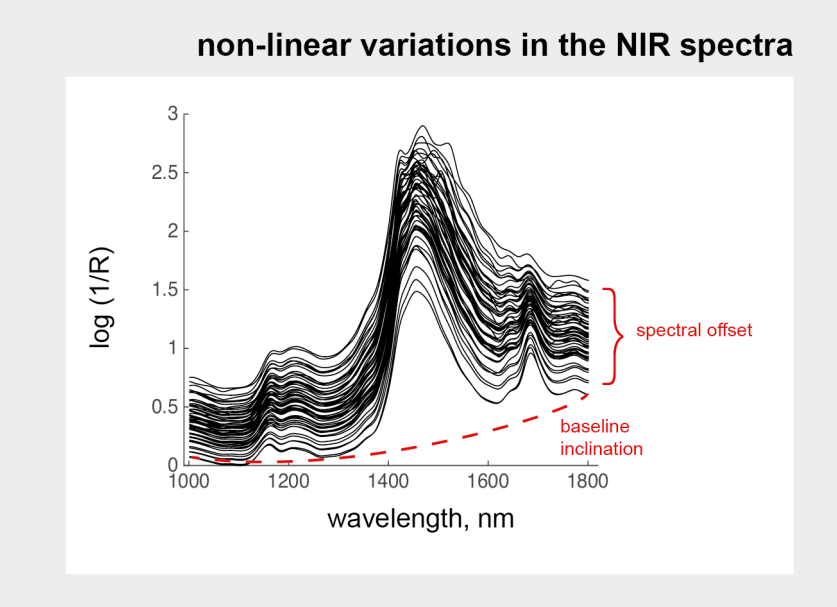

Figure 2.5 Non-linear variation in the NIR spectra from 56 samples of polystyrene beads suspended in water at varying sizes and concentrations.

As the pathlength is largely dependent on the size of the beads, it will be longer in samples containing smaller polystyrene beads, whereas shorter in samples of larger beads. This variability in pathlength causes an additive effect that manifests itself by a baseline inclination or shift of the spectra[37]. Moreover, as the probability of a photon to be absorbed increases with the pathlength, a multiplicative effect is also observed. These non-linear variations can complicate the interpretation of the spectra and the development of calibration models.

2.2.2.1 Standard Normal Variate (SNV)

Among the most applied empirical scatter correction methods, the SNV is the simplest one. It is applied to each individual spectrum by subtracting its absorbance average and dividing the difference spectrum by the standard deviation from all absorbances[38,39], as it can be seen in equation 2.7 presented in Figure 2.6.

Consequently, not all points contribute to the normalization equally, meaning that values away from the average absorbance will be weighted more than the values near the average. The procedure is then similar to an auto-scaling of the rows instead of the columns of the matrix[40].

In Figure 2.6 it can be seen that the offset variation was significantly reduced in the spectra after SNV correction, although the curvilinear trend still persists. This inclination of the baseline is a common feature on spectra from solids, and is influenced by particle size and packing density. To correct the spectrum against this curvilinear trend, the SNV is sometimes performed together with a de-trending normalisation step, which can be a second-degree polynomial[38].

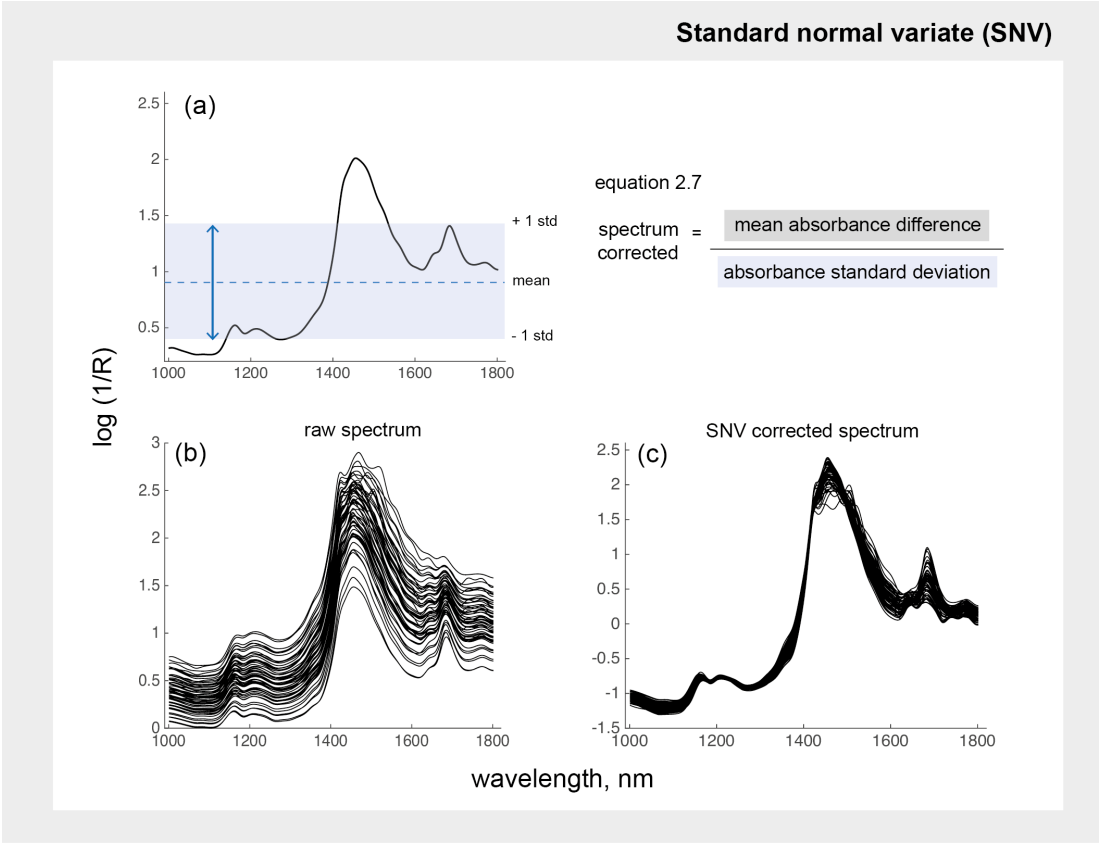

Figure 2.6 SNV pre-processing (a) mean value and standard deviation are calculated from all absorbances of an NIR spectrum. (b) Raw and (c) corrected spectra from 56 samples of polystyrene suspended in water.

The SNV and SNV-De-trending (SNV-DT) do not require a reference spectrum, since they are performed on a per-spectrum basis[41]. That simplifies their application, especially when validating models in future samples, since a reference spectrum from the calibration dataset is mandatory when using MSC and related versions.

2.2.2.2 Multiplicative signal correction (MSC) and related versions

In the MSC, the correction is performed in the whole dataset, instead of in each spectrum, as in the SNV method. It employs a two-step procedure, where for each spectrum the individual spectral absorbance values are fitted against the reference spectrum (Figure 2.7 a and Equation 2.8), and the resultant intercept and slope are used to correct the raw spectra from additive and multiplicative effects (equation 2.9), respectively. A reference spectrum is required in the MSC, but the mean spectrum from the dataset is normally sufficient and practical for most situations.

In Figure 2.7a, the spectral values from the reference spectrum are plotted against the individual spectral values from the following four polystyrene samples containing:

- Sample 1 = less beads / smaller beads = 15% w/w / 415 um
- Sample 2 = less beads / bigger beads = 15% w/w / 1115 um
- Sample 3 = more beads / smaller beads = 50% w/w / 415 um
- Sample 4 = more beads / bigger beads = 50% w/w / 1115 um

In this figure, the resultant best fit for each sample lies close to a straight line, which can be interpreted as the scattering spectrum of the sample. Thus, their differences are due to scatter effects, while the deviation on each of them are due to chemical information (water and polystyrene) contained in each sample.

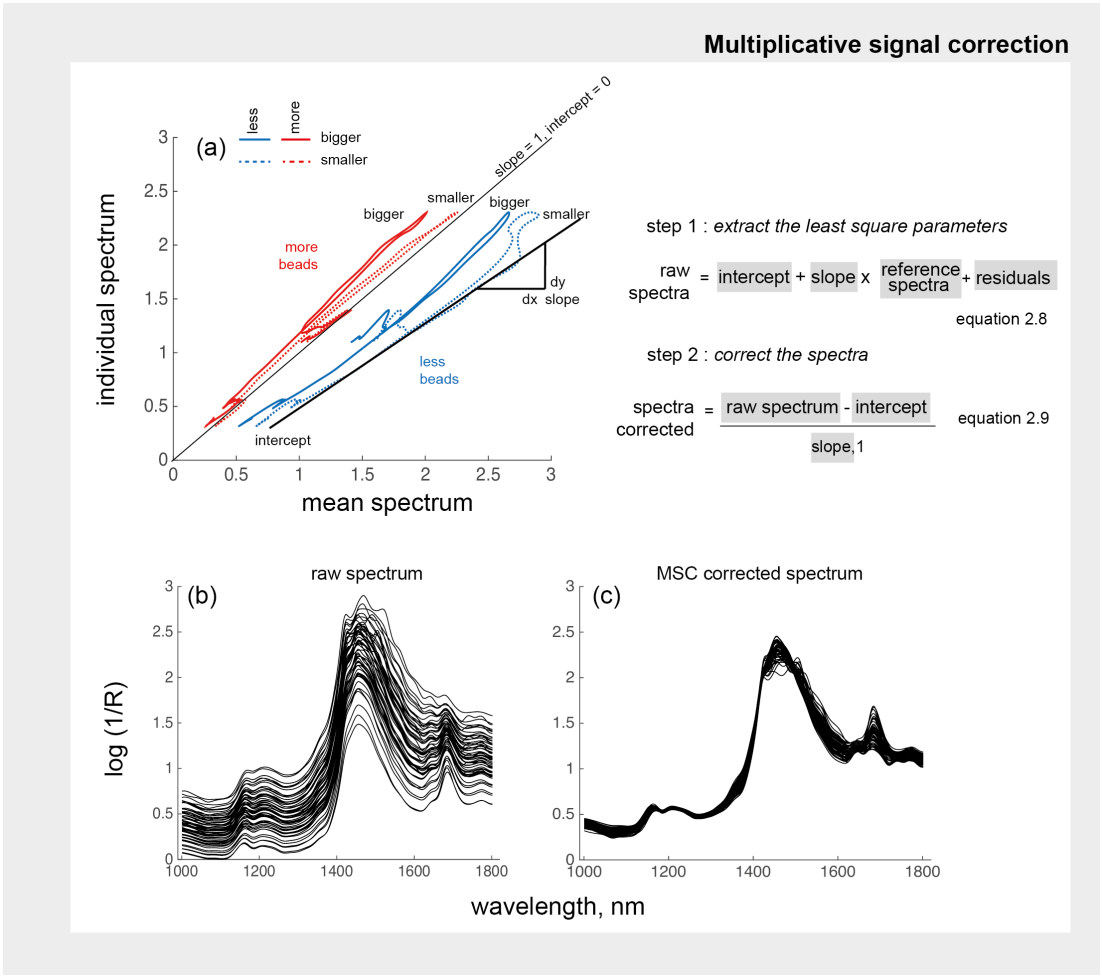

Figure 2.7 MSC pre-processing (a) individual *versus* average values obtained from NIR spectra of four samples, and the equations used for the two-step MSC calculation. (b) Raw and (c) corrected spectra from 56 samples of polystyrene suspended in water at varying concentrations and mean bead sizes.

For this example, it can then be observed from Figure 2.7 that the concentration of beads, i.e., their quantity, had a stronger impact in the scattering that their mean size, since scattering profiles from samples with higher and lower concentrations are far apart than the profiles from samples of different mean bead size.

Moreover, as expected, samples containing bigger beads were much more affected by scattering than the smaller beads, in both concentration regimes.

By inspecting Figure 2.7 b and c, it can be seen that the spectral offset was considerably reduced, and consequently, would be suitable to be used for estimation of concentration of chemical species. However, the removed part can also be useful. Recently, the nature of the information contained in the MSC coefficients was investigated aiming for determining mean particle diameter[42]. The authors found that the MSC parameters are strongly influenced not only by particle size but also by particle concentration and the refractive index of the medium.

Several variations of the MSC have been proposed, the most used being the Extended MSC, which was proposed by the same authors, and also generated several other versions. In the original version, the wavelength dependency of light-scattering effects was taken into account through a second-order polynomial. Other variations include the incorporation of physics-based models proposed by Thennadil et al[43], a version for correction of Mie scattering/dispersion artefacts in biological samples called Resonant Mie Scattering (RMieS)[44] correction, among many others. Further information about the EMSC and related versions can be found in excellent reviews [36,45].

2.3 Practical aspects

2.3.1 Sampling modes

In most applications, sample preparation is not needed when making analytical measurements by using visible or NIR spectroscopy. This is one of the main advantages of these techniques, not only due to the savings related to reagents and time involved in preparing a sample, but especially because it allows measurements to be made in systems where sample preparation is not possible, or when the errors resultant of it are too high, as it is the case of chemical reactions and biological fermentations.

Vis-NIR light can be measured in reflectance, transmittance or transflectance, depending on the sample properties (liquids, solids or gases) and on the intended application. When visible or near-infrared light strikes polystyrene beads or styrene droplets, all the incident photons are reflected, transmitted or absorbed.

In diffuse reflectance, the objective is to detect the light that is diffusely reflected from the surface of the sample, after it has undergone multiple scattering inside the sample. The diffuse reflected light is normally collected at an angle of 5 to 85 degrees. In reflectance mode the absorbance is defined as:

$$absorbance(\lambda) = \log \frac{1}{reflectance} \qquad 2.10$$

In diffuse transmittance measurements, light is detected after passing through the sample and leaving from the opposite side. Standard transmission is the conventional way of measuring a spectrum, but is normally used for analysis of transparent liquids, where light is not scattered. However, its diffuse version is also possible,

giving good results especially due to its bigger pathlength. In transmission mode, the absorbance is defined as:

$$absorbance(\lambda) = \log \frac{1}{transmittance} = \varepsilon\, c\, \ell \qquad 2.11$$

where, α is the molar absorptivity, c represents the concentration of chemical species and ℓ is the pathlength. In practical terms, an absorbance of 1 means that 10% of the photons were detected whereas an absorbance of 2 means that only 1% of them was detected, relative to the standard.

Transmittance and diffuse light can also be collected by means of an integrating sphere, allowing light collection over all reflected angles, and isolation of the light which is reflected from the one that is transmitted[46]. The basic principle is that light enters the sphere and is reflected multiple times by the highly reflective coating (over 99% reflection) and is scattered uniformly around the interior of the sphere.

Due to the high scattering nature of the polystyrene beads, diffuse reflectance is normally the method of choice for accounting for these photons, but applications can be found in diffuse transmittance as well[13,47]. Figure 2.8 shows how light travels through the sample in the diffuse reflectance and transmission setups.

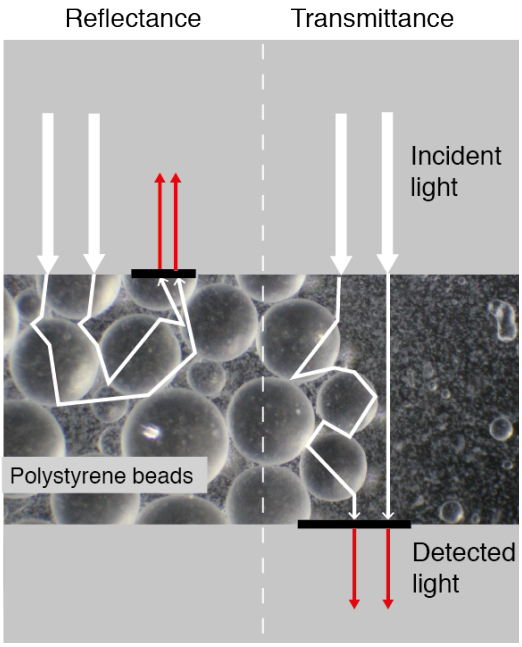

Figure 2.8 Schematic illustration of the diffuse reflectance and transmission acquisition in a polystyrene bead suspension.

2.3.2 Sampling interfaces

Sampling in polymerisation reactions is a challenging task due to the high heterogeneity and viscosity of the reaction media. Several sources of errors can deteriorate dramatically the quality of the measurements:

- Sample may not be representative of the whole media;
- Rapid evaporation of the monomer due to the high temperature of the sample;
- Reaction continues to run in the sample until quenching;
- Uncertainties related to the addition of chemicals for quenching;
- Uncertainties related to weighting of chemicals, and sample before and after evaporation, freezing, etc.

On-line or in-line measurements have the potential to optimise the monitoring and control of such reactions, since they allow non-destructive and non-invasive measurements to be acquired at faster pace, allowing a more efficient, reliable and

representative sampling. Figure 2.9 presents a schematic illustration of the sampling modes that can be chosen for polymerisation reactions.

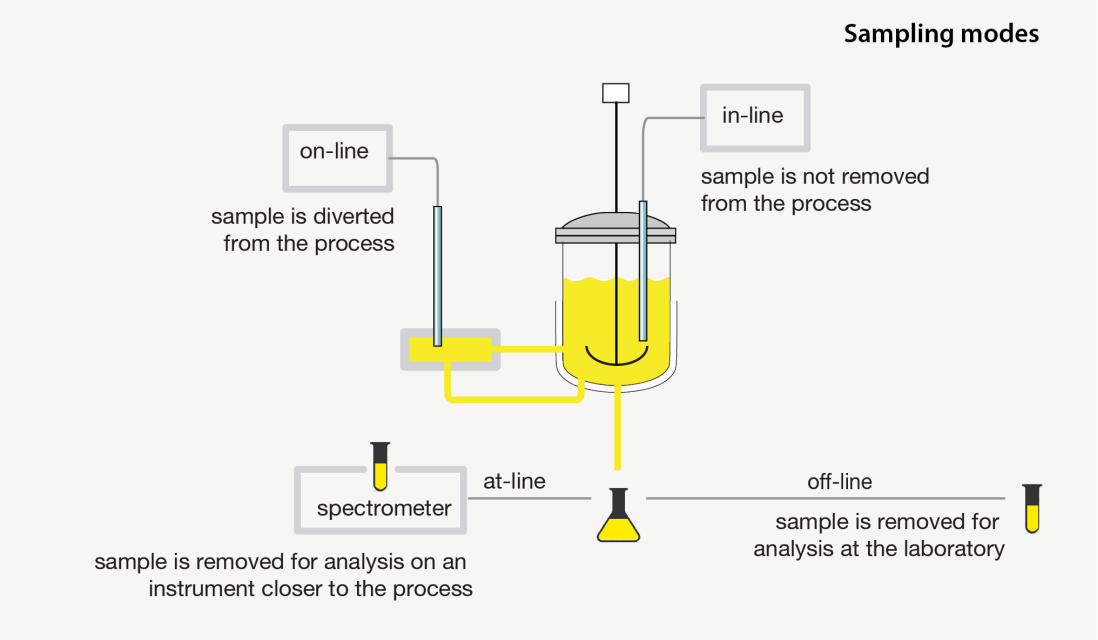

Figure 2.9 Schematic illustration of the measurement models available for Vis and NIR measurements

On-line and in-line monitoring are the fastest ways of obtaining information from polymerisation reactions, providing real-time data for process monitoring and control. Besides savings with time, transport of sample, and chemicals involved in the quenching, for example, their measurements are more accurate because the source of errors associated to manipulation and sample preparation are avoided.

Because the sample is diverted from the process in on-line measurement systems, they are non-invasive. While this may be relevant in some biotech applications, due to GMP regulations for avoiding contamination, in polymerisation reactions this type of measurement brings uncertainties related to the diversion of the sample, especially in regard to its representativeness. For these systems, in-line measurements would be

preferable, because they allow analysis to be done *in-situ*[48], with the spectroscopic sensors mounted directly inside the reactor vessel, in contact with the reaction media.

Fibre optic sensors are the responsible for the versatility of these spectroscopic process analytical measurement modes. The spectrometer can be placed remotely, and light conveyed through the fibres from the source to the sample, where it is absorbed, reflected or transmitted, and taken to the detector by another fibre, or even the same. These fibres can be multiplexed, allowing several measurements to be acquired simultaneously, and several similar or different sensors to analyse multiple properties from the process.

2.3.3 Considerations on optical probe geometry

When designing an optical probe, a few considerations need to be made for obtaining an optimum performance for the desired application. Besides physical characteristics, materials and strength of the probe, important properties are the core diameter, numerical aperture, fibre packing and source-to-detector distances.

Numerical aperture determines the range of angles at which the fibre can accept or emit light. It is defined by the difference in the refractive indices of the core and the cladding material. A fibre with higher numerical aperture will be able to collect more light than a fibre with a lower numerical aperture.

A smaller ***core diameter*** is desired in order to obtain an optical fibre as small as possible. However, bundle geometry and fibre packing usually determine the core diameter for most of the applications. Figure 2.10 below presents a few fibre bundle geometries commonly used for standard and spatially resolved Vis-NIR spectroscopy. The geometries studied in this thesis are also shown.

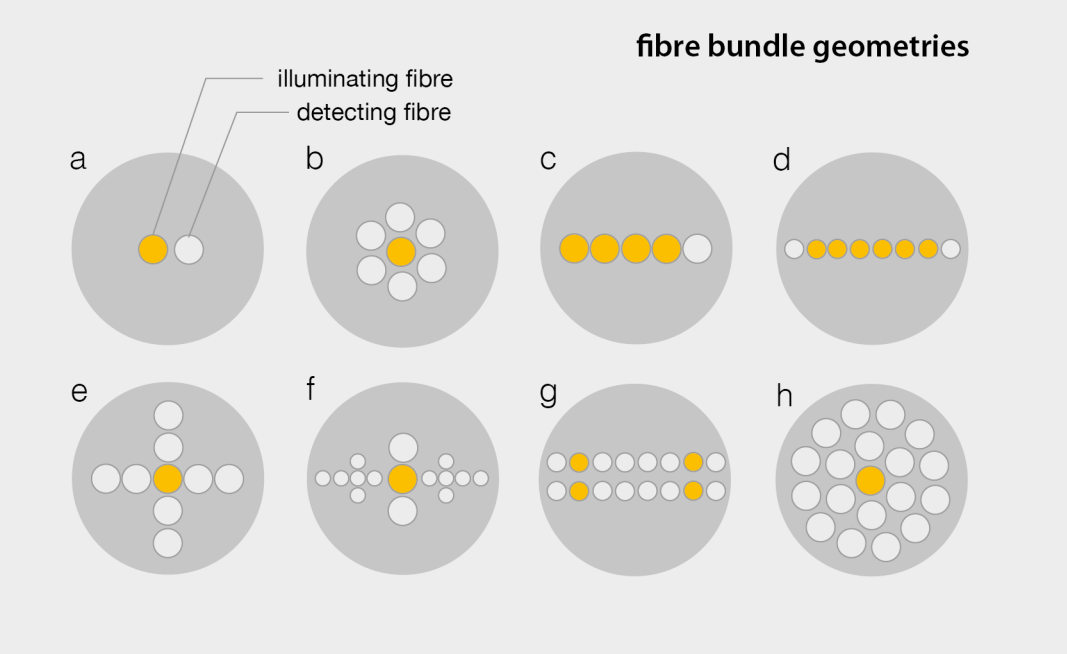

Figure 2.10 Different fibre bundle geometries.

A basic optical probe consists of two separate fibres for illumination and collection (Figure 2.10 a). A single-fibre probe would also be possible, but the measurement can be contaminated by background signals produced by the illumination[49]. One of the most common configurations for NIR optical probes is presented in b). It is a 6-around-1-bundle configuration, where six illuminating fibres circle a central collecting fibre.

In spatially resolved optical probes, fibres are normally arranged in linear arrays, where the detecting fibres are placed at varying distances from the illumination fibre. An example is the optical probe used by Nguyen Do Trong et al[50] for monitoring apples. In this configuration, as illustrated in Figure 2.10 c, five 200 mm detecting fibres were placed at increasing distances from the illuminating fibre, all with a numerical aperture of 0.22. Very recently, Bogomolov et al[35] used a similar probe, but this time consisting of 8 fibres, for monitoring fat and protein in milk. Bevilacqua et al[51] presented another example of linear array configuration, aimed at medical

applications. The probe geometry is shown in d), where two illuminating fibres are disposed symmetrically in relation to the collecting fibres. The small SD distances (<1.4 mm) adopted in this design were able to probe a smaller volume of tissue, providing better differentiation between different types of tissue.

Fibres can also be arranged as two crossing rows of fibres (Figure 2.10 e), where several illuminating fibres are placed at similar distances from a central detecting fibre, being the reverse order also possible, i.e., one central illuminating fibre surrounded by detecting fibres. With this design, more light collected at each SD distance reaches the detector, improving the S/N ratio.

A more elaborate fibre design is commercialised by Indatech (Clapiers, France) for pharmaceutical applications. In this probe, illustrated in Figure 2.10 f, the illuminating is made by three fibres placed in a row at the centre of the probe head, which are surrounded by six collecting fibres in each side. This design allows good sensitivity to local chemical heterogeneity and differences in density or mean particle size at the cost of reduced sampling area of the sample[32]. Finally, the last two examples are custom made probes proposed in the scope of this thesis, which will be discussed in the following chapters.

2.3.4 Instrumentation

As mentioned in the beginning of this chapter, the basic NIR and Visible spectrometers are not significantly different from each other. The basic components consist of a radiation source, a wavelength selection device, sample holder and a detector. Based on this, multiple possibilities of configurations are possible, and several other components can be added depending of the desired application. For research applications, for example, benchtop spectrometers like the Fourier transform (FT) based ones are commonly used due to their high spectral resolution

and better signal-to-noise ratio[3]. For real-time applications in a plant factory, miniaturised spectrometers are becoming more common due to their compactness and low costs. Furthermore, in recent years, advances have been made in their technology allowing improved precision [52] [53] [54].

There are several ways of categorising the measurement devices, but let's follow here the one recommended by Ciurczak and Igne[46] as a guide for choosing a measurement system. According to these authors, the measurement configurations can be divided in these categories:

- Filter-based instruments
- Scanning grating monochromators
- Interferometer-based instruments
- Acousto-optic tunable filters
- Photodiode arrays
- Specialty and custom instruments
- Optical parameter instrumentation

Spatially resolved spectroscopy is an example of optical parameter instrumentation, along side the integrating sphere-based measurements, time-resolved spectroscopy, and frequency domain photo migration. Such configurations are required for the separation of absorption and scattering. Spatially resolved measurement systems require a portable spectrometer which can be connected to fibre optics, or a Charge-Coupled Device (CCD) camera. Miniature photodiode arrays and CCD-based spectrometers are normally used due to its ruggedness and speed. For most of the applications, a multiplexer is also required in order to probe light separately from each of the fibres.

2.4 Multivariate latent variable modelling

The development and consolidation of chemometrics and near-infrared spectroscopy are closely intertwined[55]. Without chemometrics, the NIR region would probably still be an obscure and unexplored region, and without NIR, chemometrics would not have reached the maturity and wide acceptance it has nowadays.

This close relationship may be in part explained by the indirect nature of the NIR method. Quantitative analysis based on NIR spectroscopy cannot rely on assignments of single absorption bands, as other spectroscopies do, but rather requires the development of a regression model relating the NIR spectra and the analyte of interest, measured by the reference method. In fact, multivariate calibration based on latent variables is a major cornerstone of chemometrics[56]. But chemometrics in near-infrared spectroscopy is not restricted to multivariate calibration; its tools are present in all phases of method development and application. It starts in the design of experiments, follows to signal processing, selection of spectral ranges and samples, pattern recognition, development of predictive models, data fusion, calibration transfer, until validation and model maintenance. In each of these steps the development and research of new methods have seen an increase over the years.

2.4.1 Data organisation in chemometrics

Chemometrics relies heavily in linear algebra, hence data are organised in matrices of n rows and k columns, where rows represent samples and columns the variables describing the system. For example, near-infrared data taken during a full polymerisation reaction are organised as a matrix (**X**), where the n rows contains NIR spectra collected at different sampling intervals, and the k columns contains the

absorption intensities in the wavelengths from a specific spectral region. When quantitative analysis is involved, the readings for the quality parameter of interest, let us say the monomer conversions obtained by gravimetric analysis for each sample, are organised in a column vector (**Y**), since only one conversion percentage is measured for each sample. Throughout this thesis, we will refer to the **Y** vector as the reference matrix, or reference values. Taken together, the **X** matrix and the **Y** vector comprise a dataset.

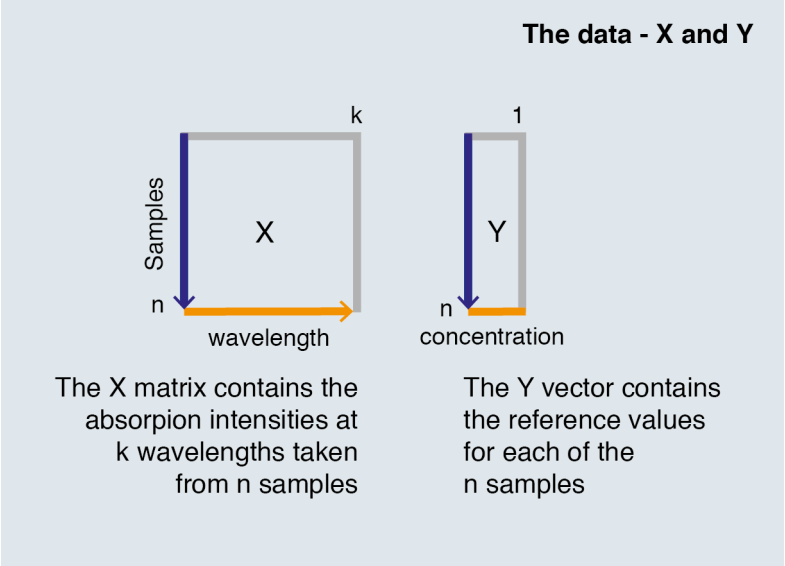

Figure 2.11 Matrix representation of spectroscopic based methods

Visible and near-infrared spectroscopic data are normally affected by multicollinearity, which happens when the absorption intensities at two or more wavelengths are not independent of each other. Multivariate latent variables methods like Principal Component Analysis (PCA) and Partial Least Squares (PLS) are widely used in spectroscopy data because they can handle very well that high multicollinearity in a straightforward manner, by converting the information contained in hundreds of wavelengths into a few latent variables, which are easily interpretable. They provide a visual guide of the wavelengths or combinations of

wavelengths that are affected by a variation on the reference values, and show how such variations are impacted by the process operating conditions.

2.4.2 Principal Component Analysis (PCA)

PCA aims to extract all relevant information from a multitude of variables, representing it as a reduced set of new orthogonal variables called principal components, or latent variables. The new coordinates are represented by the scores and loadings, which allow us to visualise multivariate spectral information in a simpler way, by uncovering the relationship among samples and variables, showing which variables are correlated between each other and which of them are the most significant in describing the complete dataset. Figure 2.12 shows the components of PCA decomposition, presenting the basic equation of multivariate latent variable analysis.

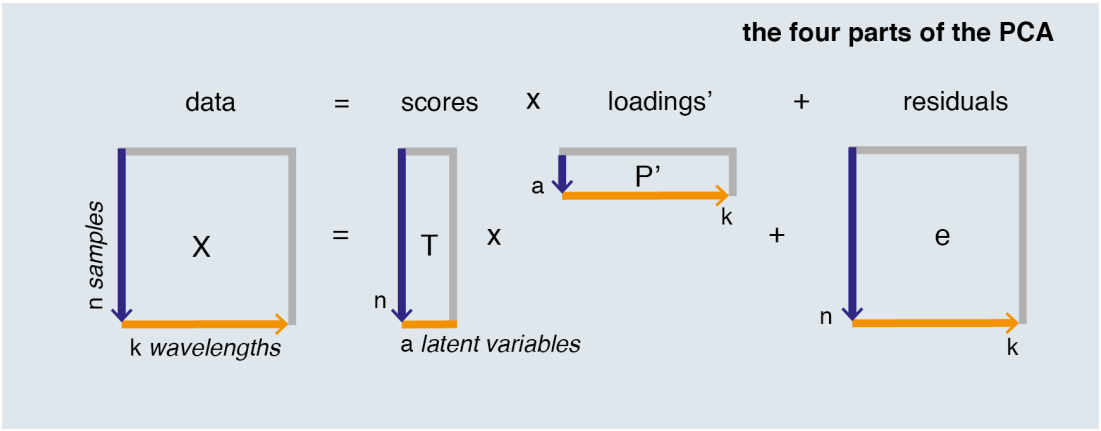

Figure 2.12 PCA decomposition of the spectra in scores, loading and residuals, represented in an equation (equation 2.12) and in matricial form.

The scores and loadings simplify the visualisation of similarities and differences within the data, which can be difficult to see on a higher dimensional set up. The scores values allow us to see the correlation between the samples whereas the

loadings values show us the correlation within the variables. As they are complementary and superimposable, patterns seen in a specific direction of the score plot can be interpreted by looking at the same direction in the loading plot.

By using the scores and loading information simultaneously, we can understand better the causes of variation and their magnitudes. Furthermore, detection of outliers is straightforward, and in a process perspective, detecting anomalous variations or unconformities is easier as well.

Each latent variable, made of one score and one loading value, is ordered according to their decreasing variance. Hence, the first latent variable describes most of the variability in the data, while the second one, being orthogonal to the first LV, carries the second highest amount of variability in the data, and so forth. Depending on the nature of the data, further latent variables may describe less structure of the original dataset, so they can be discarded without losing important information.

2.4.3 Partial least squares (PLS) regression

Projection to latent structures by means of partial least squares, or simply PLS regression, is a regression extension of the PCA where both **X** and **Y** matrices are used for modelling purposes. Its aims to extract the latent variable scores that best describe the variation in the **X** dataset that is most influential on the **Y** data set.

As opposed to other regression techniques like Principal Component Regression (PCR) and Multiple Linear regression (MLR), in PLS the dimensionality of the **X** matrix is reduced by maximizing its covariance with **Y**. Consequently, it takes in account the errors of both spectroscopic and reference methods. PCR and MLR methods model **Y** only through the correlation of **X**, so they assume the reference is error free, which is definitively not the case in analytical applications. Moreover,

PLS can analyse collinear, noisy and numerous **X**-variables[57], as opposed to MLR, which can model only less variables than samples and it is severely affected by multicollinearity and noise. For those reasons, PLS is the regression technique of reference in process analytical applications, being extensively investigated by the chemometrics and NIR spectroscopy communities, which can be seen by the innumerous version of the original algorithm already implemented.

The basic PLS algorithm used in this work is the Non-linear Iterative Partial Least Squares (NIPALS). The procedure is illustrated in the Figure 2.13 through an array scheme.

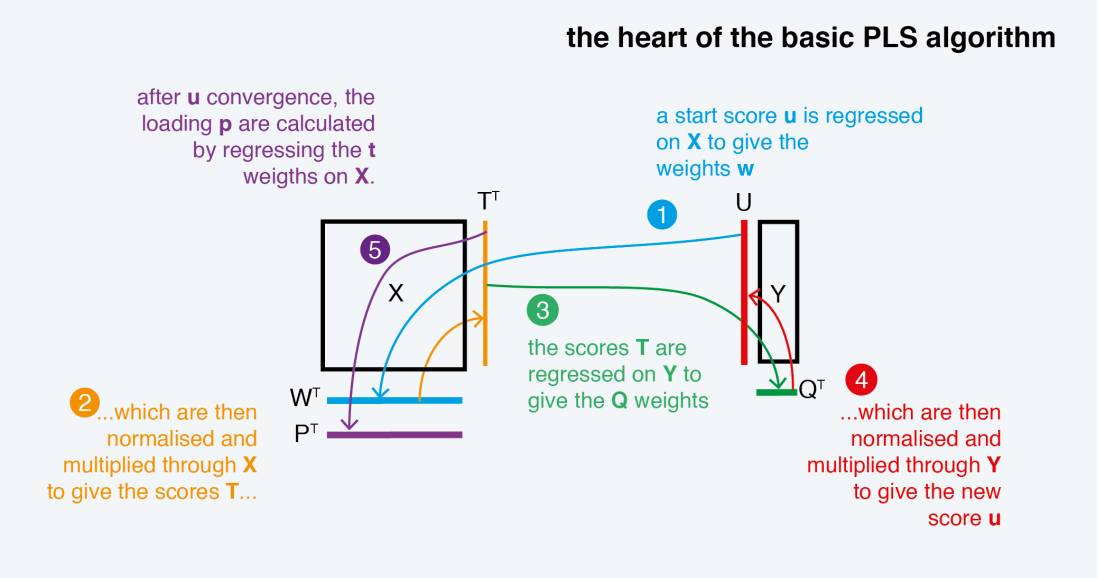

Figure 2.13 Schematic illustration of the PLS algorithm's main steps.

The algorithm starts by choosing a start score vector **u,** which can be any vector from the **Y** matrix**,** for example. This initial score vector is (1) then regressed on **X** to obtain the variable weights, **w,** which are then normalized to unity length. (2) By multiplying the normalised weights to **X**, a first approximation of the score vector **t** is obtained. The score vector is (3) regressed on **Y** to give the **q** weights, which are also normalised to unity length. A new **u** score vector can now be (4) calculated

based on the **q** weights and the **Y** vector. This cycle is repeated until the score vector converges. When a convergence is reached, the (5) loadings **P** are calculated based on **X** and the **t** scores. A regression is performed between **u** and **t**, to obtain the regression coefficients **b**. The residual matrices (**E** and **F**) are then obtained by subtracting the X loading and scores from **X**, and **Y** loadings, scores and weights from **Y**, respectively. The next set of latent variables is calculated by using **E** and **F** instead of **X** and **Y** to start the iteration.

A PLS model is represented by both loading matrices, **P** and **Q**, together with the weight matrix **W**. Based on them, the estimations for new samples are calculated. The number of latent variables that describe the model can be found by cross-validation.

2.4.4 Multiblock PLS

MB-PLS is an extension of the ordinary PLS regression method, with the difference being that it is performed relating several blocks for the purpose of prediction or interpretation. Multiblock PLS operates in two distinct levels[58]: a sub-level, which works similarly as individual PLS models, where the features of each block are modelled separately; and a super level, which aggregates this summarised information from each block to form a super block.

There are several variants of the MB-PLS method[59,60], two of them widely used in the multi-block community. They differ mainly on the chosen deflation procedure. While Wangen and Kowalski[61] proposed the use of the block scores for the deflation of the blocks, Westerhuis and Coenegracht [60] suggested the use of the super score for the deflation step, to avoid loss of information in the deflation process. This variant of the Multiblock PLS will be used in this work. More information about the MB-PLS and data fusion will be given in Chapter 5.

2.5 Validation

''With great power comes great responsibility''

As Charles Whelan said in his naked statistics book[62], ''regression analysis is the hydrogen bomb of the statistical arsenal''. It is indeed a very powerful technique, and as such, need to be used with caution. Therefore, a robust validation is essential to avoid developing models that yield misleading results, and to evaluate their usefulness in real life applications.

When validating multivariate calibration models, especially those ones based on PLS, we usually concern on three mains aspects[56]:

1 How many latent variables are enough for characterising a dataset

2 How reliable is the prediction

3 How representative are the data used to produce a model.

To properly answer these questions, a minimum of two sets of samples are needed, one for calibration, where a calibration model evaluating the model parameters is built, and another one where the predictive capability of such calibration model is evaluated. How these subsets of samples are chosen and used for validation is what sets apart the two main approaches for validation: test set and cross-validation.

2.5.1 Test set validation

The **test set validation** is a more rigorous type of validation because it is performed in an independent set of samples, not used in the calibration phase[63]. The calibration model is built solely based on samples from the calibration set, and validated against

the samples from the validation, or test set. The performance of the models can then be accessed by the calculation of the root mean square error of prediction (RMSEP), which is done according to the equation 2.13.

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n_{pred}}} \tag{2.13}$$

The RMSEP is an estimate of the variation of the reference values, $\hat{y}_i$, and their respective estimated values, y_i, divided by the number of samples found in the prediction set. Despite its advantages, it may not be easy to select these independent samples, and more importantly, not enough samples may be available for building two distinct sets of samples in real applications. In those cases, cross-validation methods are normally adopted for model evaluation.

2.5.2 Cross-validation

Cross-validation offers a more practical approach to validation, since it is iteratively performed in subsets of the calibration set. In its simplest version, the leave-one-out cross-validation, the validation set consists of just one sample that is removed from the calibration set containing *ncal* samples. The model is then built based on this *ncal*-1 calibration set, and validated upon the removed sample (validation set). This procedure is repeated by re-incorporating the deleted sample to the calibration set, being replaced by another one, until all samples are validated once. The performance of the models is accessed by the root mean square error of cross-validation (RMSECV), which is done according to the equation 2.14.

$$\text{RMSECV} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{ncal}} \tag{2.14}$$

Alternatively, the validation set can consist of block of samples, as it is in the case of the *n* fold-cross validation, where more than one sample is deleted from the calibration set and used in the validation set. This option is normally preferable than the leave-one out cross-validation, and the number of samples of each fold is usually set to five to ten samples.

It is important to highlight that besides assessing a model performance, the RMSECV estimates are also used to choose the optimum parameters for the model, as example of the number of latent variables needed to describe the data variance. It follows that the optimised errors may be an over optimistic measure, even in the case of the test set calibration, since when tuning the calibration model the test set may be used also in the training set. Solutions to the over-optimism of cross-validation have been proposed, as example of bootstrapping, Monte-Carlo cross-validation[64] and double cross validation and its repetition version, the repeated double cross-validation.

2.5.3 Double cross validation

Double cross-validation combines both model selection and model assessment by using two nested cross-validation loops[65]. The outer loop is used to estimate the prediction performance of the models. It works like an external validation, because the Y values for the test sets are estimated based on calibration sets that were not used for selecting the model parameters, as example of number of latent variables required to build the models. The model complexity is chosen based on the inner loop, which also estimates the model's errors. It consists of splitting the calibration samples defined in the outer loop in two subsets of validation and training sets (one segment as validation set and the remaining as training set). Both inner and outer

loops are repeated *n*SEG times, with *n*SEG being the number of segments used to split the datasets. Usually, sample sets in both loops are divided in 4 to 10 segments.

2.5.4 Repeated Double Cross Validation (rdCV)

Filzmoser et al[66] proposed a version of the double cross-validation where repetitions of the whole double cross validation procedure is performed many times. It was initially proposed for NIR data[67], but since has been applied in data from a variety of analytical techniques[67,68].

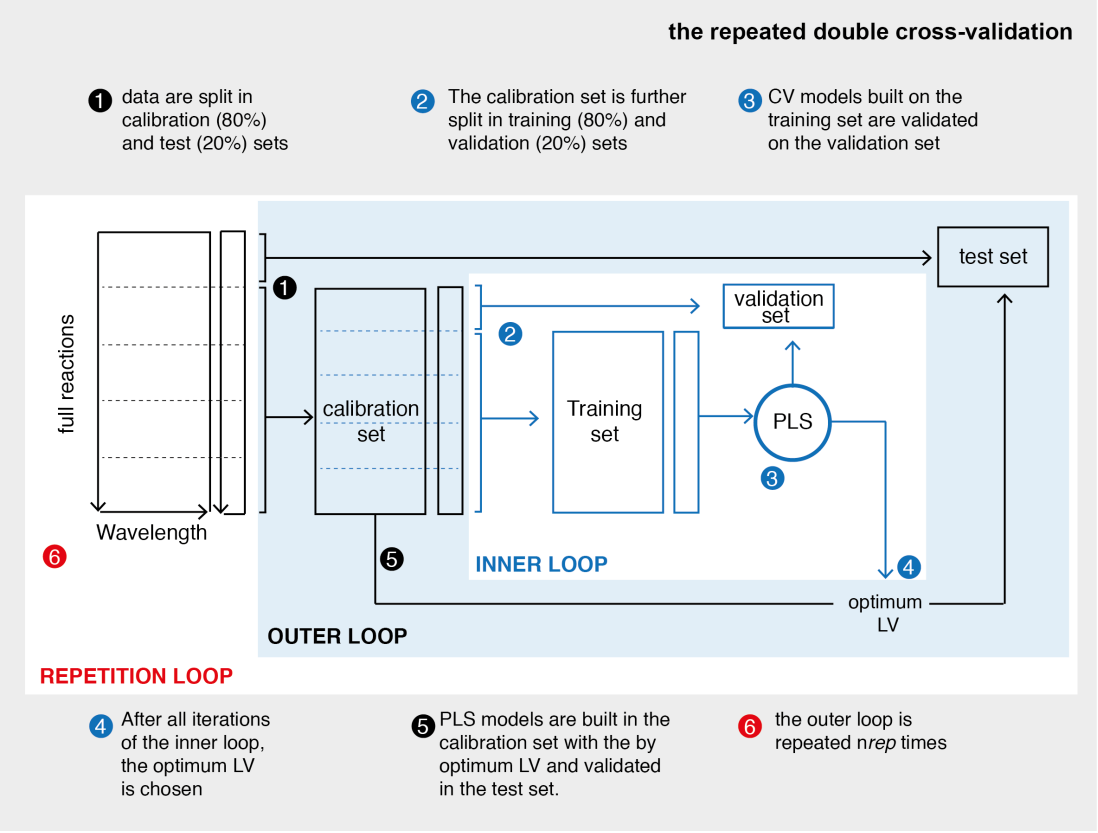

Figure 2.14 Schematic illustration of the rdCV method.

The repeated double cross validation (rdCV) advantage resides in the large numbers of estimates obtained, allowing a better estimate the prediction performance and its

variability. A central point in the rdCV approach proposed by Filzmoser is the selection of the optimum number of latent variables (step 4), which is performed based on the standard error method. In step 3 of the Figure 2.14, for each one of the inner loop's iterations, cross-validation is performed for a chosen number of latent variables (20, for example) and the models are validated upon the validation samples. For each one of these twenty latent variables, the mean squared error (MSE) is obtained from the residuals between the estimated and measured values for the quality parameter, the Y values. In the standard error method, it is assumed that these estimates of MSE values are prone to errors, thus instead of selecting the latent variable that gave the minimum MSE value, the smallest number of LV that is distant one standard deviation of the MSE value is the one selected as the optimum LV. This is done nSEG x nREP times (500 for 5 segments and 100 repetitions, for example), and the LV selected more frequently is chosen as the optimum LV to be used in the test set. The MSE and the standard error method are described by equations 2.15 and 2.16.

$$\text{mean square error } = \frac{1}{n_{MSE}} \sum_{i=1}^{n_{MSE}} (y\ measured - y\ estimated_i)^2 \tag{2.15}$$

$$mean_{\text{MSE}} < \min mean_{\text{MSE}} + \frac{parsimony\ factor. standard\ deviation_{\text{MSE}}}{\sqrt{number\ of\ samples\ in\ the\ calibration\ set}} \tag{2.16}$$

It can be seen in equation 2.15 that the mean standard error is a more conservative way of selecting the number of latent variables, aiming to avoid overfitting. How much conservative it is can be regulated by the parsimony factor. By choosing one standard error, it gives a 68% confidence interval. A bigger parsimony factor (larger confidence interval) may result in an unrealistic low number of LVs, potentially causing underfitting, but much lower confidence may lead to results similar to the minimum MSE.

Other approaches for selecting the model complexity can also be studied. The simplest way chooses the latent variable that results in the minimum MSE[69], or any other measure like RMSECV or Prediction Error Sum of Squares (PRESS). However, for the reason stated earlier, although simpler, this approach can potentially overfit the estimations. Another commonly applied method was proposed by Haaland and Thomas[70], in which the optimum LV is the lowest possible provided that it gives a PRESS value not significantly different from the minimum PRESS. For this, the F statistic is calculated to obtain the statistical significance determination.

Independent of the method for selecting the optimum number of latent variables, the rdCV is robust choice especially when dealing with small datasets. It is superior to conventional leave-out approaches for cross-validation, since it reduce chances of obtaining accidental performance results, due to the high number of splits of the data into calibration and test sets[66]. Moreover, the complexity of the models is optimised for each one of the calibration sets in the inner loop, and the prediction performance is estimated on independent test sets.

2.6 Bibliographic references

1. Thomas, N. C. The early history of spectroscopy. *Journal of Chemical Education* **68,** 631 (1991).

2. Wheeler, O. H. Near infrared spectra: A neglected field of spectral study. *Journal of Chemical Education* **37,** 234 (1960).

3. Pasquini, C. Near Infrared Spectroscopy: fundamentals, practical aspects and analytical applications. *Journal of the Brazilian Chemical Society* **14,** 198–219 (2003).

4. Vance, C. NIR-2015 Brazil—Highlighting South America. *NIR news* **26,** 17–20 (2015).

5. Williams, P. How it all began. in **I,** 36736 (Galoa, 2015).

6. Wetzel, D. L. Near-Infrared Reflectance Analysis. *Anal. Chem.* **55,** 1165A–1176A (2012).

7. Workman, J. & Workman, J. in **2,** 1–22 (World scientific, 2016).

8. Coates, J. *Interpretation of Infrared Spectra, A Practical Approach*. (John Wiley & Sons, Ltd, 2000). doi:10.1002/9780470027318.a5606

9. Ciurczak, E. W. & Fountain, M., III. *Molecular Spectroscopy Workbench*. (Applied Spectroscopy, 1998).

10. Workman, J. & Weyer, L. *Practical Guide and Spectral Atlas for Interpretive Near*. (Infrared Spectroscopy, 2012).

11. Siesler, H. W., Ozaki, Y., Kawata, S. & Heise, H. M. *Near-Infrared Spectroscopy: Principles, Instruments, Applications*. (Wiley, 2008).

12. Burns, D. A. & Ciurczak, E. W. *Handbook of near-infrared analysis*. (2007).

13. Steponavicius, R. & Thennadil, S. N. Extraction of chemical information of suspensions using radiative transfer theory to remove multiple scattering effects: application to a model multicomponent system. *Anal. Chem.* **83,** 1931–1937 (2011).

14. Dam, J. S., Yavari, N., Sørensen, S. & Andersson-Engels, S. Real-time absorption and scattering characterization of slab-shaped turbid samples obtained by a combination of angular and spatially resolved measurements. *Appl. Opt., AO* **44,** 4281–4290 (2005).

15. Kienle, A. & Patterson, M. S. Determination of the optical properties of turbid media from a single Monte Carlo simulation. *Phys. Med. Biol.* **41,** 2221–2227 (1999).

16. Thennadil, S. N. & Chen, Y.-C. Alternative Measurement Configurations for Extracting Bulk Optical Properties Using an Integrating Sphere Setup. *Appl Spectrosc* **71,** 224–237 (2017).

17. Hjalmarsson, P. & Thennadil, S. N. Spatially resolved in vivo measurement system for estimating the optical properties of tissue in the wavelength range 1000-1700 nm. *Diagnostic Optical Spectroscopy in Biomedicine IV* **6628,** 662805–662805–9 (2007).

18. Hjalmarsson, P. & Thennadil, S. N. Determination of glucose concentration in tissue-like material using spatially resolved steady-state diffuse reflectance spectroscopy. in (eds. Tuchin, V. V. & Wang, L. V.) **6855,** 685508–685508–9 (SPIE, 2008).

19. Suzuki, S., Takasaki, S., Ozaki, T. & Kobayashi, Y. Tissue oxygenation monitor using NIR spatially resolved spectroscopy. *BiOS '99 International Biomedical Optics Symposium* **3597,** 582–592 (1999).

20. Kholodtsova, M. N., Daul, C., Loschenov, V. B. & Blondel, W. C. P. M. Spatially and spectrally resolved particle swarm optimization for precise optical property estimation using diffuse-reflectance spectroscopy. *Opt Express* **24,** 12682–12700 (2016).

21. Nguyen, H.-D. & Hong, K.-S. Bundled-optode method for detection of brain activity in functional near-infrared spectroscopy. in 1112–1117 (IEEE, 2016). doi:10.1109/ICCAS.2016.7832450

22. Barer, R., Cole, A. R. H. & Thompson, H. W. Infra-Red Spectroscopy with the Reflecting Microscope in Physics, Chemistry and Biology. *Nature* **163,** 198–201 (1949).

23. Farrell, T. J., Patterson, M. S. & Wilson, B. A diffusion theory model of spatially resolved, steady‐state diffuse reflectance for the noninvasive determination of tissue optical properties invivo. *Medical Physics* **19,** 879–888 (1992).

24. Abildgaard, O. H. A. *et al.* Non-Invasive Assessment of Dairy Products Using Spatially Resolved Diffuse Reflectance Spectroscopy. *Appl Spectrosc* **69,** 1096–1105 (2015).

25. Xia, J. J., Berg, E. P., Lee, J. W. & Yao, G. Characterizing beef muscles with optical scattering and absorption coefficients in VIS-NIR region. *Meat Science* **75,** 78–83 (2007).

26. Nguyen Do Trong, N. *et al.* Spatially resolved diffuse reflectance in the visible and near-infrared wavelength range for non-destructive quality assessment of 'Braeburn' apples. *Postharvest Biology and Technology* **91,** 39–48 (2014).

27. Nguyen Do Trong, N. *et al.* Optical properties–microstructure–texture relationships of dried apple slices: Spatially resolved diffuse reflectance spectroscopy as a novel technique for analysis and process control. *Innovative Food Science & Emerging Technologies* **21,** 160–168 (2014).

28. Vaudelle, F. & L'Huillier, J.-P. Influence of the size and skin thickness of apple varieties on the retrieval of internal optical properties using Vis/NIR spectroscopy: A Monte Carlo-based study. *Computers and Electronics in Agriculture* **116,** 137–149 (2015).

29. Cen, H., Lu, R., Mendoza, F. A. & Ariana, D. P. Peach maturity/quality assessment using hyperspectral imaging-based spatially resolved technique. in (eds. Kim, M. S., Tu, S.-I. & Chao, K.) **8027,** 80270L–80270L–15 (SPIE, 2011).

30. Inagaki, T., Nozawa, D., Shimomur, Y. & Tsuchikawa, S. Three-fibre-based diffuse reflectance spectroscopy for estimation of total solid content in natural rubber latex. *Journal of Near Infrared Spectroscopy* **24,** 327–335 (2016).

31. Shi, Z. & Anderson, C. A. Application of Monte Carlo Simulation - Based Photon Migration for Enhanced Understanding of Near - Infrared (NIR) Diffuse Reflectance. Part I: Depth of Penetration in Pharmaceutical Materials. *Journal of Pharmaceutical Sciences* **99,** 2399–2412 (2010).

32. Igne, B., Talwar, S., Feng, H., Drennen, J. K. & Anderson, C. A. Near-Infrared Spatially Resolved Spectroscopy for Tablet Quality Determination. *Journal of Pharmaceutical Sciences* **104,** 4074–4081 (2015).

33. Wang, L. V. & Jacques, S. L. Source of error in calculation of optical diffuse reflectance from turbid media using diffusion theory. *Computer Methods and Programs in Biomedicine* **61,** 163–170 (2000).

34. Igne, B., Talwar, S., Feng, H., Drennen, J. K. & Anderson, C. A. Near-Infrared Spatially Resolved Spectroscopy for Tablet Quality Determination. *Journal of Pharmaceutical Sciences* **104,** 4074–4081 (2015).

35. Bogomolov, A., Belikova, V., Galyanin, V., Melenteva, A. & Meyer, H. Reference-free spectroscopic determination of fat and protein in milk in the visible and near infrared region below 1000 nm using spatially resolved diffuse reflectance fiber probe. *Talanta* **167,** 563–572 (2017).

36. Rinnan, Å., Berg, F. V. D. & Engelsen, S. B. Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry* **28,** 1201–1222 (2009).

37. Dhanoa, M., Lister, S., Sanderson, R. & Barnes, R. The link between multiplicative scatter correction (MSC) and standard normal variate (SNV) transformations of NIR spectra. *Journal of Near Infrared Spectroscopy* **2,** 43–47 (1994).

38. Barnes, R. J., Dhanoa, M. S. & Lister, S. J. Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra. *Appl Spectrosc* **43,** 772–777 (1989).

39. Kohler, A., Zimonja, M., Segtnan, V. & Martens, H. in *Comprehensive Chemometrics* 139–162 (Elsevier, 2009). doi:10.1016/B978-044452701-1.00102-2

40. Wehrens, R. *Chemometrics with R*. (Springer Science & Business Media, 2011).

41. Fearn, T., Riccioli, C., Garrido-Varo, A. & Guerrero-Ginel, J. E. On the geometry of SNV and MSC. *Chemometrics and Intelligent Laboratory Systems* **96,** 22–26 (2009).

42. Chen, Y.-C. & Thennadil, S. N. Insights into information contained in multiplicative scatter correction parameters and the potential for estimating particle size from these parameters. *Analytica Chimica Acta* **746,** 37–46 (2012).

43. Thennadil, S. N., Martens, H. & Kohler, A. Physics-based multiplicative scatter correction approaches for improving the performance of calibration models. *Appl Spectrosc* **60,** 315–321 (2006).

44. Bassan, P. *et al.* Resonant Mie scattering (RMieS) correction of infrared spectra from highly scattering biological samples. *Analyst* **135,** 268–277 (2010).

45. Afseth, N. K. & Kohler, A. Extended multiplicative signal correction in vibrational spectroscopy, a tutorial. *Chemometrics and Intelligent Laboratory Systems* 1–8 (2012). doi:10.1016/j.chemolab.2012.03.004

46. Ciurczak, E. W. & Igne, B. *Pharmaceutical and Medical Applications of Near-Infrared Spectroscopy, Second Edition*. (CRC Press, 2015).

47. Dehanov, N. & Thennadil, S. Monitoring of conversion during emulsion polymerisation using near infrared spectroscopy. *NIR news* **25,** 9–16 (2014).

48. Callis, J. B., Illman, D. L. & Kowalski, B. R. Process Analytical Chemistry. *Anal. Chem.* **59,** 624A–637A (1987).

49. Utzinger, U. & Richards-Kortum, R. R. Fiber optic probes for biomedical optical spectroscopy. *J. Biomed. Opt.* **8,** 121–27 (2003).

50. Nguyen Do Trong, N. *et al.* Spatially resolved spectroscopy for nondestructive quality measurements of Braeburn apples cultivated in sub-fertilization condition. in (ed. Kondo, N.) **8881,** (SPIE, 2013).

51. Bevilacqua, F. *et al.* In vivo local determination of tissue optical properties: applications to human brain. *Appl. Opt., AO* **38,** 4939 (1999).

52. Crocombe, R. A. Handheld spectrometers: the state of the art. in (eds. Druy, M. A. & Crocombe, R. A.) **8726,** 87260R (SPIE, 2013).

53. Correia, R. M. *et al.* Portable near infrared spectroscopy applied to fuel quality control. *Talanta* **176,** 26–33 (2018).

54. Paiva, E. M., Rohwedder, J. J. R., Pasquini, C., Pimentel, M. F. & Pereira, C. F. Quantification of biodiesel and adulteration with vegetable oils in diesel/biodiesel blends using portable near-infrared spectrometer. *Fuel* **160,** 57–63 (2015).

55. Gemperline, P. in *Practical Guide To Chemometrics, Second Edition* 509–520 (CRC Press, 2009). doi:10.1201/9781420018301.ch13

56. Brereton, R. G. *Chemometrics: Data Analysis for the Laboratory and Chemical Plant.* (2013).

57. Wold, S., Sjöström, M. & Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* **58,** 109–130 (2001).

58. Westerhuis, J. A., Kourti, T. & MacGregor, J. F. Analysis of multiblock and hierarchical PCA and PLS models. *J. Chemometrics* **12,** 301–321 (1998).

59. Jin Hong, J., Zhang, J. & Morris, J. in *19th European Symposium on Computer Aided Process Engineering* **26,** 327–332 (Elsevier, 2009).

60. Westerhuis, J. A. & Coenegracht, P. M. J. Multivariate modelling of the pharmaceutical two-step process of wet granulation and tableting with multiblock partial least squares. *J. Chemometrics* **11,** 379–392 (1997).

61. Wangen, L. E. & Kowalski, B. R. A multiblock partial least squares algorithm for investigating complex chemical systems. *J. Chemometrics* **3,** 3–20 (1989).

62. Wheelan, C. *Naked statistics: stripping the dread from the data*. **1,** (2013).

63. Westad, F. & Marini, F. Validation of chemometric models – A tutorial. *Analytica Chimica Acta* **893,** 14–24 (2015).

64. Xu, Q. S., Liang, Y. Z. & Du, Y. P. Monte Carlo cross - validation for selecting a model and estimating the prediction error in multivariate calibration. *J. Chemometrics* **18,** 112–120 (2004).

65. Fearn, T. Chemometric Space: Double cross-validation. *NIR news* **21,** 14–15 (2010).

66. Filzmoser, P., Liebmann, B. & Varmuza, K. Repeated double cross validation. *J. Chemometrics* **23,** 160–171 (2009).

67. Liebmann, B., Friedl, A. & Varmuza, K. Applicability of near-infrared spectroscopy for process monitoring in bioethanol production. *Biochemical Engineering Journal* **52,** 187–193 (2010).

68. Khalid, T. *et al.* Urinary Volatile Organic Compounds for the Detection of Prostate Cancer. *PLoS ONE* **10,** e0143283 (2015).

69. Bro, R., Kjeldahl, K., Smilde, A. K. & Kiers, H. A. L. Cross-validation of component models: A critical look at current methods. *Anal Bioanal Chem* **390,** 1241–1251 (2008).

70. Haaland, D. M. & Thomas, E. V. Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. *Anal. Chem.* **60,** 1193–1202 (1988).

Chapter 3

3 Experiments & Methods

This chapter presents the experiments, equipments, methods and data analysis procedures that will be explored throughout this thesis. First, the experiments performed for the evaluation of the optical probes, spectroscopic measurement modes and modelling strategies are discussed in detail. The spectroscopic systems and their individual components are presented, followed by the strategies involved in the extraction and manipulation of the data generated by them. Finally, the methods for multivariate calibration model building and validation are shortly described.

3.1 Experiments

Solid spherical polystyrene beads are the final product of a suspension polymerisation reaction[1]. Such very heterogeneous reaction is difficult to monitor even by the reference methods[2], hence, comparing the efficiencies of the different optical probes, source-detectors distances and angles of incidence would prove to be very difficult. It would make sense, though, to evaluate these effects separately, without the influence of the reaction media. For this, before analysing the information from full suspension polymerisation reactions, we first devised a simpler experiment in which post-production polystyrene beads were suspended in water at varying concentrations and sizes. The knowledge acquired from this controlled experiment may lead to a better understanding of the impact of physical and chemical information on spectroscopic measurements, and allow an easier comparison of the performance of different optical probes, spectroscopic modes and modeling strategies. That, ultimately, may be relevant for the analysis and interpretation of the models developed on actual suspension polymerisation reactions.

Figure 3.1 Polystyrene beads produced by suspension polymerisation reaction

3.1.1 Two-component system

Post-production expandable polystyrene (EPS) beads obtained from several batches of suspension polymerisation reactions were gathered to enable a broad range of bead sizes with sufficient quantity to explore different concentrations. After dried, the beads were sieved by using a series of stainless steel woven wire sieves (Endecotts Ltd., London) in different mesh sizes, in conjunction with a Minor M200 Automatic shaker (Endecotts Ltd., London).

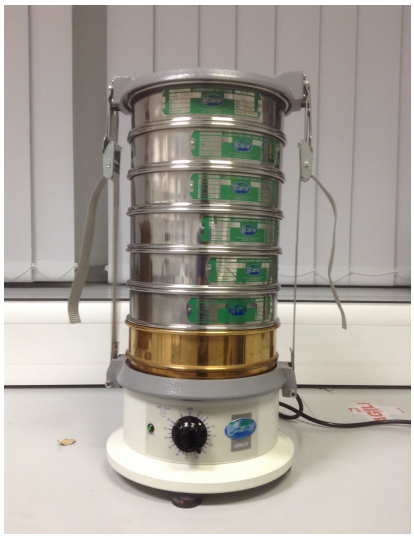

Figure 3.2 Minor M200 automatic shaker with a set of stainless steel woven wire sieves from Endecotts.

These beads were subjected to sieving in order to obtain several narrowed size ranges, yielding a total of seven portions, from 0.090 to 2.0 mm in diameter, as described in Table 3.1. The content of each sieve was then weighted and the particle size distribution calculated. For each of these portions, the mean bead diameter and its particle size distribution were also measured by using a Camsizer® dynamic image analysis system (Retsch Technology GmbH, Germany), performed by BASF. The D50 cumulative distribution was taken as a measure of the mean.

Table 3.1 Mean bead sizes and their span calculated for each diameter range

Range	**Diameter range,** mm	**D50,** mm
1	0.300:0.500	0.405 ± −
2	0.500: 0.630	0.565 ± 0.190
3	0.630: 0.800	0.699 ± 0.222
4	0.800: 1.000	0.890 ± 0.201
5	1.000: 1.250	1.108 ± 0.187
6	1.250: 1.400	1.314 ± 0.115
7	1.400: 1.600	1.540 ± 0.264

3.1.1.1 Experimental procedure

The experiments were performed in a 1 L stirred borosilicate jacketed reaction vessel (Figure 3.3 a). The temperature was controlled at 24ºC (±2º) by a thermostatic bath (Figure 3.3 b) connected to a PT-100 temperature sensor. A polytetrafluorethylene (PTFE) anchor stirrer (Radleys) was used and the agitation speed adjusted at 250 rpm.

Figure 3.3 a) Reaction vessel and the anchor impeller used in all experiments and (b) reaction system containing with heating bath, temperature controller and stirrer controller.

In order to obtain sample runs with different concentrations of polystyrene beads, the amount of polystyrene beads was maintained the same for every bead size range, and only the amount of water was changed. This procedure was chosen to have a defined mean bead size and bead size distribution for every concentration range. The following procedure was repeated for every bead size range:

- Add 200g of water into the empty reactor;
- Add 200g of polystyrene beads to obtain a concentration of 50% w/w;
- Acquire spectra;
- Add water until the concentration was lowered to 45% w/w;
- Acquire spectra;
- Lower the concentration in 5% steps by adding water until a final concentration of 15% w/w is obtained, acquiring spectra after each increment.

Previously, a similar preliminary experiment was prepared with polystyrene beads supplied by BASF. In this experiment, instead of adding water to obtain decreasing concentrations, the total volume was kept constant at 600mL. In order to obtain different concentrations, an amount of water was removed by using a pipette and the same amount of beads were added to reach the desired concentration at each increment. In this experiment, the mean bead diameter and bead diameter distribution may be different for every concentration, which could potentially add an extra source of error to the models to be developed. The results for this experiment are not shown in this thesis.

A total of 56 sample runs were performed in the system illustrated in the Figure 3.4, spanning the mean bead diameter range from 0.400 to 1.540 mm and the concentration range from 15 to 50%, at 5% intervals. The experimental runs were randomly performed over several days.

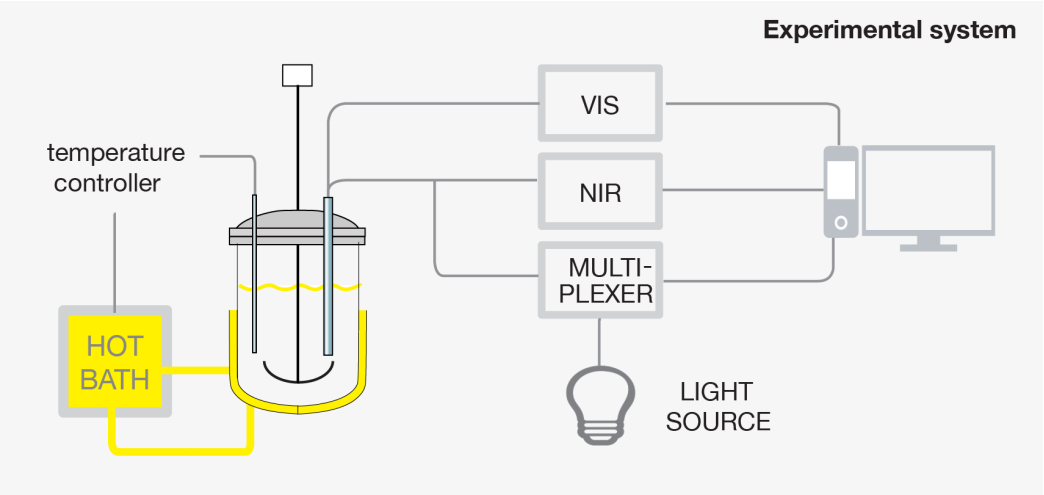

Figure 3.4 Schematic illustration of the experimental system.

Measurements in the near infrared region were obtained by a NIR spectrometer (NIRQuest-512, Ocean Optics) while the spectra in the visible region were registered by a visible range spectrometer (USB-4000, Ocean Optics), connected to the proposed SARDR probe. A multiplexer was employed for the coupling of fibres. A

detailed description of the optical system and its components will be given in the Section 3.2, later in this chapter.

The SARDR probe was positioned between the vessel wall and the impeller, deep enough into the media in order to avoid unstirred spots that could potentially result in fouling at the probe head when actual polymerisation reactions are developed[3]. Furthermore, as the probe head is in permanent contact with the stream of particles being carried by the mechanical stirrer, it allows for a representative bulk sample continuously being probed.

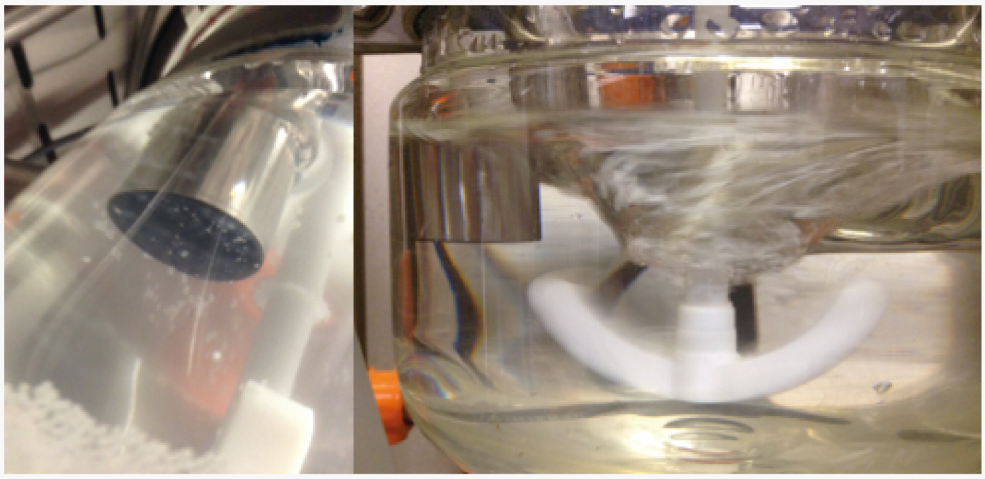

Figure 3.5 Positioning of the SARDR probe in the reactor vessel.

3.1.2 Styrene suspension polymerisation reactions

3.1.2.1 Process description

The suspension polymerisation process investigated in this work is a laboratory-scale batch reaction for the production of expandable polystyrene beads. The reaction lasts approximately seven hours, where a full conversion from styrene to polystyrene is achieved. A standard polymerisation recipe was used and the reaction conditions were obtained through assistance from BASF (Ludwigshafen, Germany). Both reaction conditions and recipes, however, are different to those employed at BASF owing to safety issues, lack of availability of equipment at the university lab to perform reactions under pressure and due to confidentiality restrictions.

Since our group had no previous experience on running such reactions, before performing experiments, all the equipment had to be designed and ordered, including the glass reactor and a glass reactor lid which could fit all probes. After the equipment was ready, a long learning time was needed to master the reaction running. The first reactions failed completely, resulting in a sticky mass of melted plastic. The first complete reactions produced beads with sizes comparable to those produced at BASF using their normal operating conditions, but there were issues with agglomeration of beads during the reaction. This caused problems in terms of sampling and also probe fouling. After further modifications of the reaction conditions and recipe it was possible to produce polystyrene beads with sizes in the expected range (0.3 – 2.0 mm), and with no agglomeration or fouling issues.

3.1.2.2 Chemicals

Deionized water was used as the suspending medium and styrene as the monomer. Luperox® benzoyl peroxide (BPO) with moisture content of 25% was used as the initiator and sodium dodecyl benzene sulfonate (SDBS) was employed as ionic emulsifier. Tricalcium phosphate (TCP) was used as the stabilizing agent and was

supplied by BASF. Toluene and *para*-benzoquinone were used in the samples withdrawn for gravimetric analysis, for dissolving the polystyrene and quenching the reaction, respectively. All reactants, unless otherwise stated, were supplied by Sigma-Aldrich and used as supplied.

3.1.2.3 Set-up

Styrene suspension polymerization reactions were performed in a 1 L stirred boro-silicate jacketed reaction vessel, especially designed for this project and supplied by Radleys. The temperature was controlled by a thermostatic bath (Huber GmbH) connected to a PT–100 temperature sensor. A PTFE Anchor stirrer (Radleys) was used and the agitation speed adjusted at the desired speed before the start of each reaction.

3.1.2.4 Procedure

Approximately 300 g of styrene was poured into the sealed vessel through an aperture in the lid and the stirrer speed set to approximately 100 rpm. Once the temperature inside the vessel had reached approximately 30°C, 8 g of solid polystyrene beads were added through the aperture to seed the reaction. When the beads had dissolved, a solution of the remaining styrene (approximately 104 g) and the initiator DBPO were added to the vessel. As the temperature of the media in the vessel reached 40°C, 650 g of water with an appropriate amount of TCP dispersed throughout was added to the vessel and the stirrer speed was set to the desired level. Spectral acquisition began 5 minutes after the temperature of the media reached 80 °C and this temperature was maintained within ± 0.5°C throughout the remainder of the reaction. Spectra were taken at every 10 minutes and samples for gravimetric analysis were taken at every 20 minutes. After 1 hour and 15 minutes from the reaction beginning, 1 mL of 2.5% SDBS solution was added to the vessel. During the reaction, samples were collected from the reactor at regular intervals for

characterisation by the conventional gravimetric analysis and particle size measurements. The reaction finished after around 7 hours from its initiation. The reaction mixture was then removed from the reaction vessel, cooled and suction filtered through a glass vacuum funnel.

3.1.2.5 Gravimetric Analysis

For offline estimation of monomer conversion, samples were taken throughout each reaction for gravimetric analysis, which is the technique routinely used throughout the polymer industry. This work used an adapted version of the sampling procedure proposed by Santos *et al*. The procedure involves removing approximately 5 mL of reaction media, immediately weighing it in an aluminium pan, and adding 5 mL of 0.1% w/w *para*-benzoquinone inhibitor in toluene to quench the reaction. Both the pan and the volume of the *para*-benzoquinone solution are weighed beforehand. To further ensure that the reaction stopped completely, the pan containing the sample is then put into a freezer for approximately 2 minutes before being weighed once more. Once the cool weight is obtained, the sample is placed in an oven at 45°C for several days to evaporate any solvents. The solid samples are taken from the oven and weighed once a day until an approximately constant weight is achieved. Then, the theoretical weights of all known solid constituents – surfactant, stabiliser, *para*-hydroquinone and initiator – are subtracted, leaving the weight of polystyrene in the sample. This is then taken as a percentage of the theoretical weight of polystyrene at 100% conversion to give the estimated percentage conversion at each sample point. The use of *p*-benzoquinone allows faster diffusion into the beads as it is more hydrophobic than hydroquinone; therefore, faster quenching of the reaction in the sample is possible and a more accurate estimation can be achieved.

3.2 Spectroscopic system

Figure 3.6 presents a diagram showing various components of the spatially and angularly resolved Vis-NIR diffuse reflectance spectroscopic system used in this thesis. It consists of two miniaturised spectrometers, a light source, a multiplexer, a multi-fibre probe, optical fibres, an integrating sphere and a computer integrated to a graphical interface. Each component of the system is described in the following sections of this chapter.

Light is propagated via an optical fibre from the light source to the multiplexer, where it is channeled to the individual fibres on the probe placed in contact with the sample. The light reflected by the sample is collected by a collecting fibre and guided towards the spectrometer, where a spectrum is obtained, further processed and stored by a computer. An integrating sphere is used for collection of the reference spectrum.

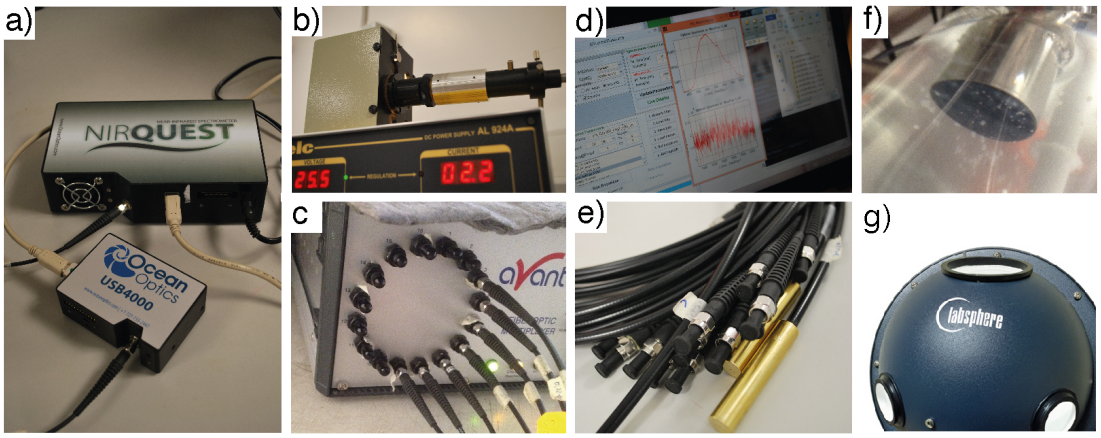

Figure 3.6 Vis and NIR SARDR spectroscopic system: a) spectrometers, b) light source, c) multiplexer, d) computer and graphical interface, e) optical fibres, f) optical probe and g) integrating sphere.

3.2.1 Spectrometers

Measurements in the near-infrared region were obtained by a NIR spectrometer (NIRQuest-512, Ocean Optics) whereas the spectra in the visible region were registered by a visible range spectrometer (USB-4000, Ocean Optics). The NIRQuest 512 spectrometer has a Hamamatsu G9206-512W InGaAs array detector and is responsive from 900-2200nm[4]. It contains a Thermo Electric Coller (TEC) to reduce noise and maintain a stable operating temperature. The USB-4000 spectrometer uses a Toshiba TCD1304AP Linear CCD array detector, responsive from 350 to 1100 nm and operated at room temperature[5]. Both spectrometers are connected to a computer via USB ports. A comparison of the specifications for both equipments is shown in the Table 3.2.

Table 3.2 Comparison of the NIR and Vis spectrometers used in this experiment

Specifications	NIR Quest	USB 4000
Detector	Hamamatsu G9206-512W InGaAs linear array	Toshiba TCD1304AP Linear CCD array
Optical Resolution	5 nm	0.34 nm
Pixels	512 pixels	3648 pixels
Pixel size	25x250 um	8 um x 200 um
Wavelength range	900-2200 nm	200-1100 nm
Integration time	1ms-1s	10us-10s
Dynamic range	7.5x106 (system); 10000:1 for a single acquisition	$3.4x10^6$; 1300:1 for a single acquisition
Signal-to-noise ratio	10000:1 @ 100ms integration	300:1 at full signal
Dark noise	6 RMS counts @ 100ms	50 RMS counts

3.2.2 Light Source

The light source used for emitting the radiation uses a 400W tungsten-halogen lamp positioned in a modified lamp house (LSH-T100, Horiba Jobin-Yvon) and powered at 29V, 10A. The lamp house couples the light into the source fibres in a ferrule which illuminates the incident light at the end of optical fibre probe. The lamp house also provides a shutter for blocking the incident light entering the optical fibre.

3.2.3 Multiplexer

A fibre-optic multiplexer is used to couple the light from a single fibre sequentially into a set of other fibres[6]. In this experiment, a 16x2 FOM-Vis-NIR optical multiplexer (Avantes Inc.) was employed for coupling the seven illuminating fibres of the SARDR probe into the light source to provide timely and sequential illumination. The multiplexer has sixteen different output channels coupled by one input channel. A controlled stepper motor and a rotator block switch the fibres according to pre-defined sequence and time intervals by the user in the software interface. The multiplexer is connected to the computer through a serial port communication, RS-232.

3.2.4 Optical Probes

The design of an optical probe affects directly the optimal illumination and light collection in spectroscopic applications[7]. Probe geometry, source-detector distance and angle of incidence are important factors to consider when building or choosing a probe with a specific application in mind.

In this work, the sensing abilities of two different probes and their intrinsic characteristics were analysed for their potential for extracting qualitative and quantitative information from spectroscopic measurements on a polymer industrial application.

3.2.4.1 Probe 1 – Spatially and angularly resolved diffuse reflectance probe

With this probe the aim is to investigate the impact of different source-detector separations and angles of incidence on the information extracted from the two spectroscopic measurement modes analysed in this thesis. The schematic diagram of the probe head can be seen in Figure 3.7.

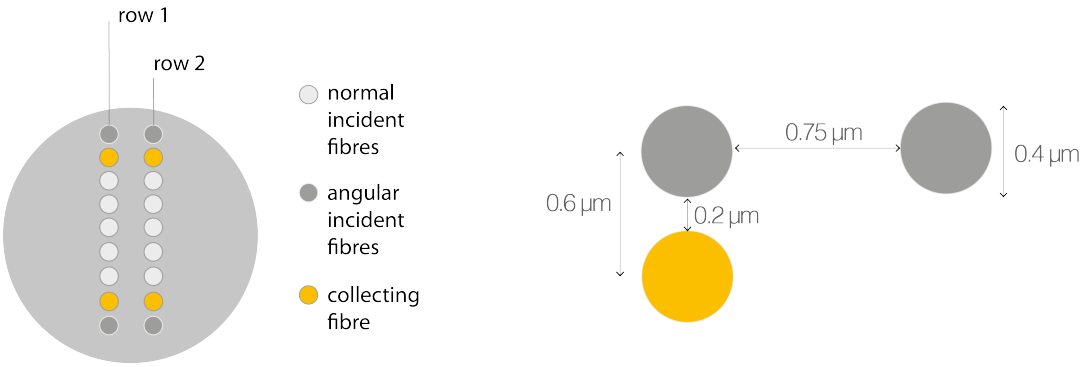

Figure 3.7 Schematic diagram of the SARDR probe head, showing the fibre's ends and their distances in micrometers.

The SARDR probe is a cylindrical probe containing two parallel rows of 400 µm

Vis-NIR low-OH fibres, each row having a different numerical aperture. Row one has two angular fibres (30° and 45°) and seven normal incident fibres with a numerical aperture of 0.22. Row two has the same number and type of fibres, but with a 0.11 numerical aperture. The fibre core-to-core separation is 600 μm and the array separation is 7.5 mm. In this thesis, only the row containing the 0.22 numerical aperture fibres was used.

The fibre length is 3 m, with a 0.4 m common end made of stainless steel tube and a black PEEK (Victrex^{TM} plc., UK) ferrule to hold all the fibres at the probe end using black epoxy for adhesion. All fibres were polished to the surface of the probe end. The bundle is made of PTFE braided stainless steel interlock conduit, 10 mm OD, and the legs with a PVC furcation tubing 3.00 mm OD. The split point is 2m from the back of the probe, encased in 40-50 mm long brass junctions. The leg ends are composed of 18 SMA 905 strain relief endings. The results obtained by this probe are presented in Chapters 4, 5 and 6.

3.2.4.2 Probe 2 – Dual-ring diffuse reflectance optical probe

This prototype probe was specifically developed for OPTICO FP7 European Project by Fibre Photonics Ltd, in partnership with the Measurements and Analytics Team at University of Strathclyde.

The probe consists of a central collecting fibre surrounded by two illumination rings. The inner ring is composed of six fibres and the outer ring has twelve fibres. All fibres have a 400 um diameter and are held together and bonded in to a ceramic ferrule. The geometry of the probe head can be found in Figure 3.8.

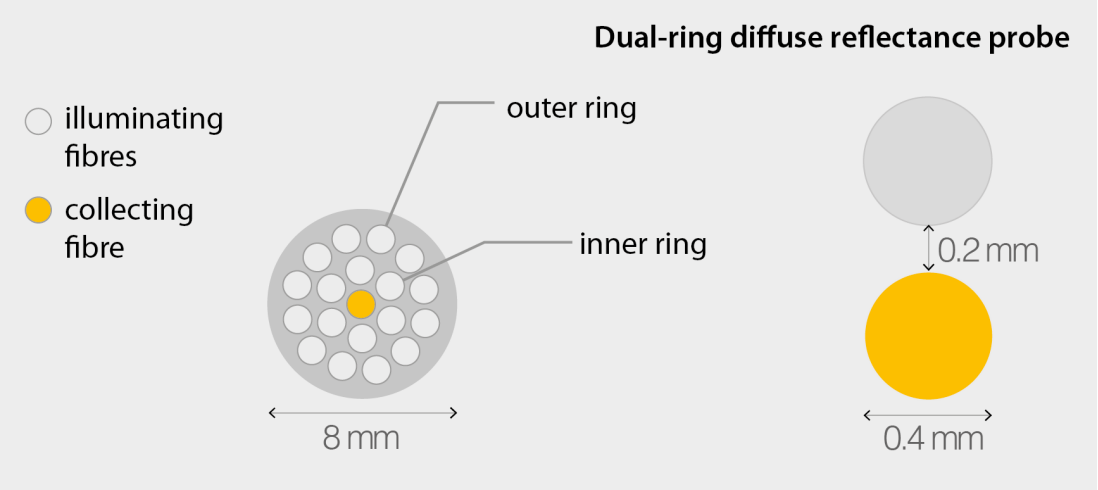

Figure 3.8 Schematic illustration of the dual-ring Diffuse Reflectance probe

3.2.5 Graphical Interface – Spatially resolved system

A graphical interface developed in MATLAB™ (R2012b, Mathworks) was used to manage the acquisition of measurements by synchronising the spectral acquisition and the switching of the multiplexer channels. It works by interfacing the spectrometers, multiplexer and computer, allowing a quick visualisation of the spectrum while it is being recorded. The step-by-step acquisition process by using the interface is shown in Figure 3.9.

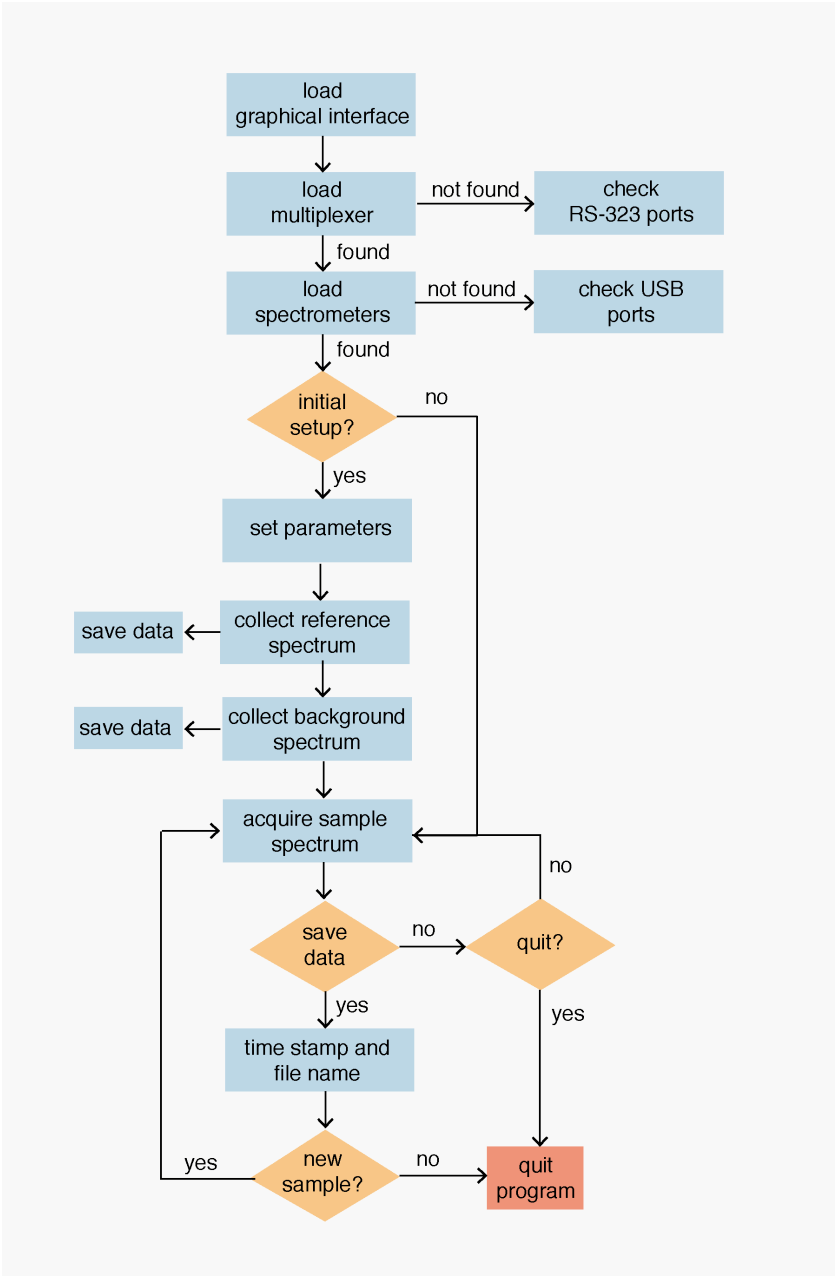

Figure 3.9 Spectral acquisition process by using the graphical interface.

After the graphical interface is loaded in MATLAB, the program initialises by identifying and loading the spectrometers and the multiplexer, which are connected to the computer through USB ports and a serial port communication, RS-232, respectively. As input for the software, the user must define the experiment name, the number of scanning repetitions and a file containing the settings for the spectra

acquisition. The settings include the multiplexer positions and integrating times to be applied for each fibre and spectroscopic mode. After the settings are uploaded, the program is ready to run. As the spectral repetitions are acquired the signal is continually updated and displayed in the computer screen. For each fibre and spectroscopy technique a *.mat* file is created containing the spectra collected for all repetitions.

3.2.6 Software

All software programs were written and implemented in MATLAB (Mathworks Inc.). Repeated double cross validation[8] was adapted from the LibPLS library developed by *Li et al*[9]. Multi-block PLS[10] was adapted from a routine kindly supplied by Dr João A. Lopes, from the Chemometrics & Process Analytical Technology Unit, University of Porto, Portugal. Wavelet denoising routines were implemented by this author by using the Wavelet Toolbox. Principal component analysis (PCA) models were implemented by using PLS Toolbox (Eigenvector Inc). A MATLAB-based graphic user interface was developed to integrate the algorithms and routines developed for spectral acquisition in the spatially and angularly resolved spectroscopy system.

3.3 Spectra acquisition and processing

3.3.1 Spectra acquisition

The SARDR system was switched on for at least an hour before collecting any measurements, in order to warm up the spectrometers and stabilise the light source. In the experimental set up used in this thesis, in addition to the spectra taken from actual samples, three other measurements are collected at each acquisition cycle:

- reference spectrum
- background reference spectrum
- sample background spectrum

The reference spectrum is taken from a highly reflecting Spectralon® 99% surface, which acts as an almost perfect Lambertian surface. In this work, a 150 mm diameter integrating sphere (Labsphere, USA) was used for this purpose. The reference spectrum and its background spectrum are collected before the beginning of the measurement cycle, by placing the probe head at the entrance port of the integrating sphere. The sample background spectrum is measured with the probe inside the reaction vessel, at the beginning of the measurement cycle.

The background spectrum is basically a spectrum taken when no light is coming from the light source. It is subtracted from the sample and/or reference spectrum to eliminate possible bias relative to the electronic noise from the system. For this, in our experimental set up the light source shutter was turned off and the room darkened. Although both background spectra are quite similar, spurious light from the experimental environment may contaminate the sample and reference spectra differently. Hence, by acquiring their background spectrum independently, it is possible to cancel out any of this contamination.

The quality of the spectrum is directly influenced by the integration time chosen for its acquisition. Thus, the choice of an adequate integration time for each

configuration is crucial for acquiring a reliable spectrum. However, such task is not straightforward, since it requires a compromise to be made between acquisition time and spectral sensitivity. Setting it too short may result in a poor signal-to-noise ratio, since less photon will be able to reach the detector, but having it too long, there may be some region of the spectrum that becomes saturated, yielding incorrect data. The spectrometers used in this work allow integrating times in the range between 3.8 ms and 10s for the USB-4000 spectrometer and 1ms to 1s for the NIR Quest spectrometer[4]. Within these ranges, several tests were performed for each application (two-component system and polymerisation reactions). The chosen integrating times for the experiments performed in this thesis are summarised in Table 3.3.

Table 3.3 Integration times and number of repetitions used for all probes

	Integration time (repetitions), milliseconds			
SARDR probe	**NIR**	**NIR reference**	**Visible**	**Vis reference**
Individual fibres	6000-8000 (8)	10000 (8)	20-100 (8)	60-300 (8)
Dual ring probe				
Inner/outer ring	4000 (8)	8000 (8)	50 (8)	100 (8)

From Table 3.3 one can notice that the NIR Quest spectrometer required an integration time that is up to 80 times longer than in the USB-4000. Consequently, the noise component is much stronger in NIR than in Vis spectrum. To improve the signal-to-noise ratio, the measurements were signal-averaged from eight measurements. The sample and reference spectra from both spectroscopy techniques were also denoised by wavelet transform.

By using a spatially resolved system, it is even more difficult to set up a standard integration time that suits all conditions, especially because each individual fibre

may respond differently with regards to spectral intensity. Moreover, the signal intensity varies considerably along the experiment due to the different concentrations and size of beads. During our experiments, the integrating time had to be shortened in a few instances to allow a reliable measurement to be obtained.

3.3.2 Signal processing

The raw measurements obtained by the spectrometers used in this experiment present considerable noise. Reducing this noise contribution is essential for extracting useful information from analytical data. The following steps were performed for each spectral data set.

3.3.2.1 Wavelet denoising

Smoothing and denoising strategies based on Wavelet transform were investigated in this work. A wavelet smoothing strategy based on the Coiflet 5 wavelet mother was employed for all measurements, with a crude threshold chosen for the removal of high-frequency components. Four decomposition levels were required for proper noise suppression.

3.3.2.2 Normalisation and background subtraction

The spectral measurements obtained from both spectrometers are recorded in digital count intensity units. After denoising, these measurements were converted to reflectance units through normalisation of the sample measurement (I_s) by the reference spectrum (I_{is}), after their respective background subtraction (I_{off} and I_{isoff}) is performed. This normalisation is done as follows:

$$R = \frac{I_s - I_{off}}{I_{is} - I_{isoff}} \qquad 3.1$$

Although the Visible and NIR reflectance spectra can be used in reflectance units, general practice recommends using it in logarithmic units, log (1/R)[11,12]. It is believed that the log (1/R) presents more linearity with regards to the analyte concentration.

3.3.2.3 Excluding non-informative regions and downsizing

Although the wavelength range of the visible spectrum (346-1010 nm) obtained by the USB-4000 is much narrower than in the NIR spectrum (899-2217 nm), it presents a higher resolution and increased wavelength sampling. This happens because while a spectral count was collected for every 2.23 nm in the NIR, in the Vis region the acquisition was performed with a 0.21 nm interval. Consequently, it has a higher number of variables (3648) than in the NIR signal (512). Considering future application of data fusion, this large difference may bring difficulties for comparison of both techniques. Consequently, it was decided to downsize the Vis spectrum by a factor of four, leaving one variable for every 0.8447 nm. It is important to highlight that this procedure has not affected the spectral character and information contained on it.

The extremities in both spectra are very noisy due to the detector insensitivity. These non-informative regions were removed to avoid degradation of the models that will be developed in the next chapters. The resultant region for NIR is 1000 to 1800 nm, and 500 to 938 nm in the visible region.

3.3.2.4 Signal processing overview

Figure 3.10 presents the spectral processing procedure adopted in this work. The process is illustrated for NIR spectroscopy, but the steps and methods are the same used for Visible spectroscopy.

The signal processing procedure starts with the sample spectrum (blue profile in Figure 3.10 a) and its background measurement (grey profile in the same figure), which are normally difficult to distinguish due to the strong stochastic noise, as well as the reference spectrum and its background spectrum (Figure 3.10 b). The background spectra are subtracted from their respective sample and reference spectra (grey curves in Figure 3.10 c and d. The resultant subtracted spectrum is denoised by using wavelet denoising (blue curves in Figure 3.10 c and d). After the normalisation is performed according to the Equation 3.1, the reflectance spectrum is obtained (Figure 3.10 e). Finally, the reflectance spectrum is converted to the absorption spectrum, shown as a grey curve in the Figure 3.10 f, and the final wavelength region represented by the blue curve.

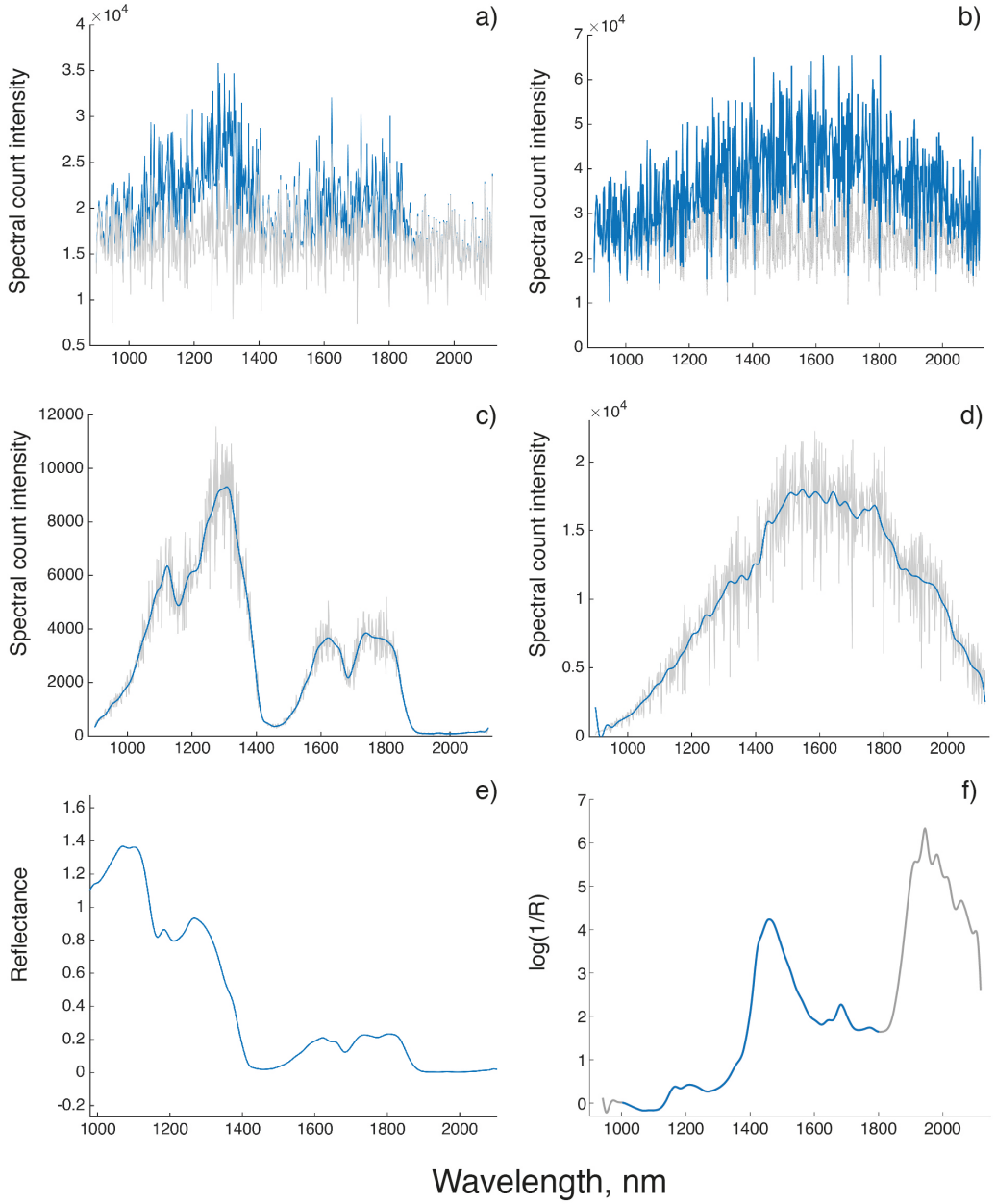

Figure 3.10 NIR spectra pre-processing procedure: (a and b) sample and reference spectra (blue) and their background spectrum (grey), (c and d) subtracted spectrum and its denoised version (blue), e) reflectance spectrum and f) full log(1/R) spectrum in grey and the wavelength selected version in blue.

3.4 Methods

All regression models developed throughout this thesis are based on partial least squares method. For the data fusion models, when more than one **X** matrix was employed, multiblock PLS was adopted as the regression method.

In order to guarantee that the models predictive capability will hold true for future unseen samples, a PLS strategy based on repeated double cross-validation was employed for model validation. Moreover, since the output of several different models will be compared in this work, this technique was also chosen to allow a reliable comparison among them[8].

The outputs for every model are the root mean squared error of cross validation (RMSECV), root mean square error of prediction (RMSEP) and their distributions, squared coefficient of correlation (R^2), optimum number of PLS components and its frequency plot, and the residuals for all estimations. The confidence intervals (CI 95%) are calculated for RMSECV and RMSEP based on their values obtained from all estimations from n_{rep} test sets.

3.5 Bibliographic references

1. Kiparissides, C. Polymerization reactor modeling: A review of recent developments and future directions. *Chemical Engineering Science* **51,** 1637–1659 (1996).

2. Kotoulas, C. & Kiparissides, C. in *Polymer Reaction Engineering* 209–232 (Blackwell Publishing Ltd, 2007). doi:10.1002/9780470692134.ch5

3. Brearley, A. M. in *Process Analytical Technology* 392–423 (Blackwell Publishing Ltd, 2005). doi:10.1002/9780470988459.ch11

4. Ocean Optics. *Ocean Optics. NIR Quest 512 Data Sheet*. (2016).

5. Ocean Optics. *Ocean Optics. USB4000 data sheet*. (2016).

6. Ciurczak, E. W. & Fountain, M., III. *Molecular Spectroscopy Workbench*. (Applied Spectroscopy, 1998).

7. Watte, R., Aernouts, B., Van Beers, R., Postelmans, A. & Saeys, W. Computational optimization of the configuration of a spatially resolved spectroscopy sensor for milk analysis. *Analytica Chimica Acta* **917,** 53–63 (2016).

8. Filzmoser, P., Liebmann, B. & Varmuza, K. Repeated double cross validation. *J. Chemometrics* **23,** 160–171 (2009).

9. Li, H., Xu, Q. & Liang, Y. libPLS: An Integrated Library for Partial Least Squares Regression and Discriminant Analysis. *PeerJ PrePrints* (2014).

10. Lopes, J. A., Menezes, J. C., Westerhuis, J. A. & Smilde, A. K. Multiblock PLS analysis of an industrial pharmaceutical process. *Biotechnol. Bioeng.* **80,** 419–427 (2002).

11. Norris, K. H. Why log (1/R) for composition analysis with NIR? *NIR news* (2005). doi:10.1177/NIRA_16_8;

12. Griffiths, P. R. & Dahm, D. J. *Continuum and discontinuum theories of diffuse reflection*. (Handbook of Near-Infrared Analysis, 2007).

Chapter 4

4 Investigating the influence of source-detector separation and angle of incidence on the predictive capability of models

The system, methods and experiments presented in the previous chapter are used here to evaluate the analytical potential of visible and near-infrared spatially and angularly resolved diffuse reflectance measurements on the estimation of quality attributes of a polymer-based system. Main interest is placed on the quality and type of information gathered from each of the different source-detectors distances and spectroscopic techniques, and their impact on predictive models. The results are based on a new SARDR optical probe developed *in house*, especially designed for this application.

4.1 Introduction

Although the first spatially resolved spectra were measured almost seventy years ago[1], it was only in the 90's that Vis-NIR-Spatially Resolved Spectroscopy (SRS) emerged as an important on-line analytical sensor. The technique gained maturity in the biomedical field[2,3], but has only recently been used in Process Analytical Technology, mostly in food[4,5] and pharmaceutical applications[6]. Despite being a well-accepted analytical tool in the polymer industry, its spatially resolved implementation has been overlooked, which can be perceived by the scarcity of publications.

The main advantage of SRS over the classical approach for Vis-NIR spectroscopy comes from the ability to retrieve information on the distance travelled by the photons through the sample. Photons that have travelled shorter distances with less penetration can be differentiated from photons that have travelled further with a deeper penetration through the sample. That extra layer of information helps us to better understand how deep light can penetrate into the sample, and which type of interaction takes place, whether mainly absorption or scattering. Consequently, this allows the extraction of chemical and physical information from the sample.

The simplicity and affordability of its instrumentation is also a great attraction of the technique. As one author mentioned, SRS sits at the interface between hyperspectral imaging and single-point spectroscopy[6], but with the advantage of being easy to implement and cost efficient. In fact, apart from the optical probe, only a spectrometer and a multiplexer, or a combination of spectrograph and CCD camera are required to acquire reliable measurements. Portable spectrometers are suitable for the task and can be found in the market at a fraction of the price of hyperspectral imaging systems.

In this chapter, we will limit our focus to explore reflectance measurements obtained from fibres with normal and angular incident illumination, evaluating their potential in estimating quality attributes of a system based on expandable polystyrene beads suspended in water at different concentrations and sizes.

The information obtained from this controlled experiment may lead to a better understanding of the various phenomena involved in such systems, allowing a clearer evaluation of the (1) new optical probe proposed, (2) the sensing abilities of each spectroscopic mode, and (3) the impact of different source-detector separations and angles of incidence on the predictive capability of the resultant models. That would potentially give the insight needed before up-scaling the strategy to monitor full polymerisation reactions.

4.2 Source-detector distances: does it really matter?

In short, spatially resolved diffuse reflectance spectroscopy (SRDR) works by collecting reflectance measurements at different distances from the illuminating fibre. The principle behind SRDR spectroscopy is illustrated in the Figure 4.1, below.

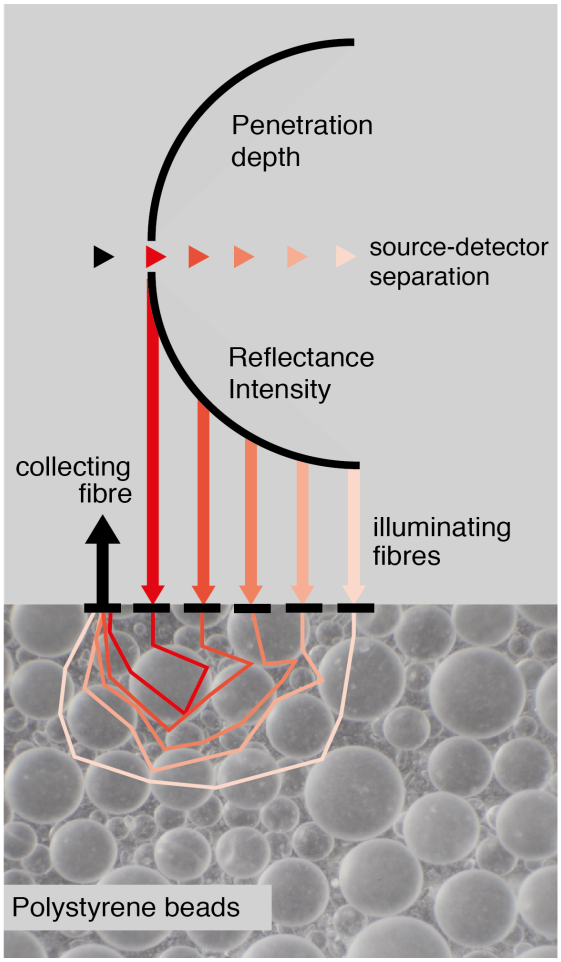

Figure 4.1 Schematic illustration of the measurement principle of spatially resolved spectroscopy in the polystyrene bead system.

For the present application, light delivered by each illuminating fibre is scattered and absorbed by the polystyrene beads and water molecules, and then detected via collecting fibre. The reflectance intensities obtained when the illumination comes

from the closest illuminating fibre are naturally stronger (red arrow and trajectory), since photons have had a shorter path to travel and, consequently, were less affected by scattering or absorption. As the separation between the collecting and the detecting fibres increases, the reflectance intensities become weaker, since now the pathlength is longer and the probability of absorption/scattering events is much higher at each increasing SD separation. Such variation of pathlength results in another important advantage of the SRS measurements: multiple source-detectors distances allow measurements at different penetration depths. As the pathlength increases by the wider SD separations, the probability of photons penetrating deeper in the sample becomes higher, hence, information from chemical or physical traits not readily available at the surface of the sample may be now uncovered.

From the above explanation, it is clear that the selection of the source-separation distances in an optical probe design demands a trade-off between penetration depth and signal quality. If the separation is too wide, it may result in a spectrum with an unsatisfactory signal-to-noise ratio, but having it too short may limit the probing to the surface of the sample, due to the shallower penetration depth.

Another less explored way to maximise the extraction of information from the differences in chemical and physical characteristics of a sample is by acquiring diffuse reflectance at different angles of incidence. Together to the collected spatial information, such measurements may offer a better representation of the sample scattering and absorption behaviour.

In order to investigate the impact of the angular incidence and source-detector separation in the Vis and NIR spectra, a new SARDR probe designed *in house* is evaluated here. Specifically, the objective is to understand the influence of the SD separation and the angle of incident on the predictive performance of multivariate regression models developed for the polymer based system in study. For this, to each

spectroscopy technique, the spectral data acquired from five normal incident fibres placed at different distances from the collecting fibre will be analysed. Moreover, spectra will also be obtained from two angularly incident fibres, 30° and 45° to the normal. The SARDR probe head is illustrated in Figure 4.2 below, depicting the configuration of optical fibres and their distances for both Visible and NIR spectral collection.

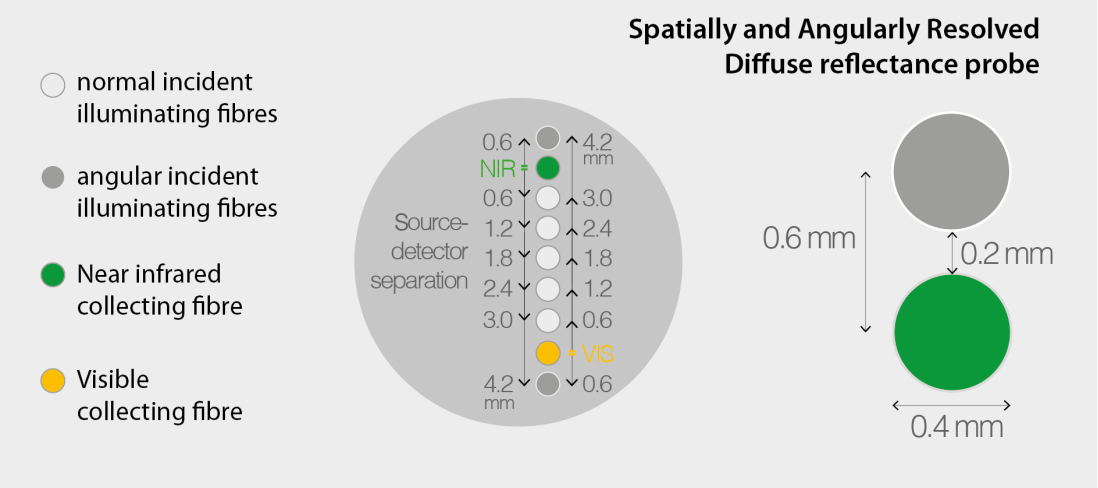

Figure 4.2 Schematic illustration of the source-detectors distances and angles of incidence in the SARDR probe head. The distances given are the ones from the Vis and NIR collecting fibres (detector) to each illuminating fibre (source).

Any of the fibres could be assigned for light collection or illumination. The configuration presented in Figure 4.2 was chosen to maximise the number of SD distances available for NIR and Visible spectra acquisition, providing similar grounds of comparison between both techniques, and to allow both spectrometers to be connected at the same time. In this way, two diametrically opposed collecting fibres were chosen, one connected to the Visible spectrometer (yellow circle) and another one to the NIR spectrometer (green circle). Each of them collects light from five normal incident fibres placed at increasing distances (0.6 to 3.0 mm) and from two angularly incident fibres (0.6 and 4.2 mm). Note that the light illuminated from a specific fibre is detected at different distances by the Visible and NIR collecting

fibres, with the exception of the central fibre, which is located at similar distance, 1.8 mm, from both Visible and NIR collecting fibres.

As it was explained earlier, a natural consequence of collecting spectra at different source-detector distances is the monotonically decrease of signal strength as the separation widens. That happens because fewer photons travelling a longer path will be able to reach the detector without being absorbed. This behaviour is exemplified in Figure 4.3 which shows a series of (a) Visible and (b) NIR spectra acquired by the SARDR probe using normal incident 400 μm fibres at five different SD distances.

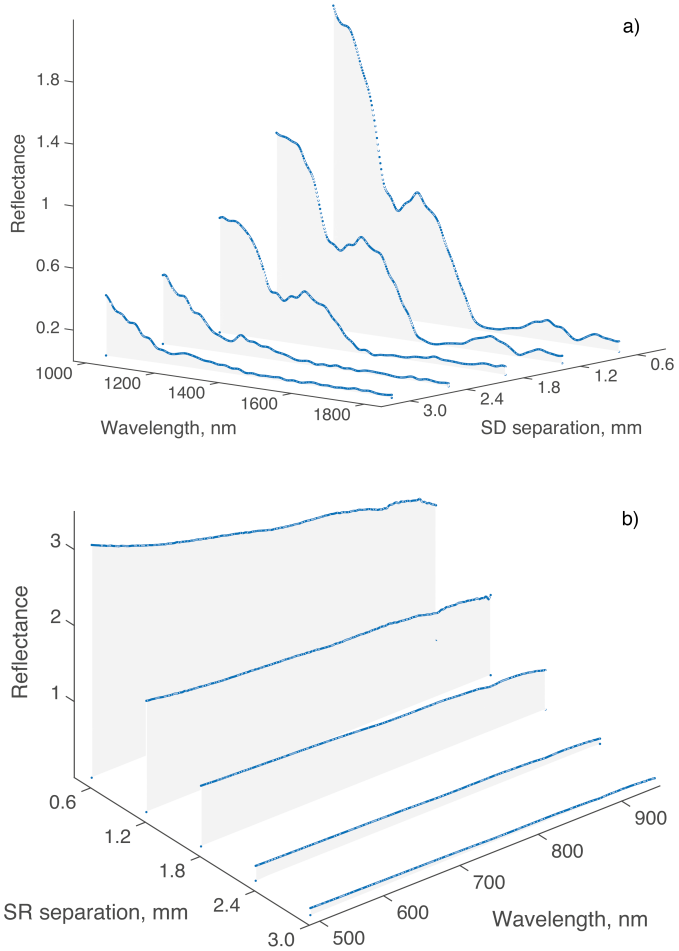

Figure 4.3 (a) NIR and (b) visible spectra taken at different source detector distances, from 0.6 to 3.0 mm. The spectra were acquired from the EPS bead system with a mean bead size of 0.69 mm and concentration of 50 % w/w.

In this figure, spectra are shown in reflectance units for better visualisation of their differences. The reflectance values are higher than 1 because the reference spectrum collected by the integrating sphere has intensities lower than the ones observed in the sample spectrum.

It can be seen that the reflectance values for all wavelengths decay rapidly as the SD separation increases. In the NIR region, the wavelength bands of water (around 1450 nm) and polystyrene (around 1670 nm) become subtler at wider separations, whereas shorter wavelengths the signal is better preserved, despite slightly decaying as well.

Nevertheless, it is important to highlight that signal strength solely may not directly translate in better or worse predictive models. Since those photons travelling farther have a generally higher chance of interacting with a larger volume of EPS beads, being scattered or absorbed, or penetrating deeper in the sample. Indeed, longer SD separations result in deeper light penetration[7], which may result in better sensitivities.

Differences in sensitivity, scattering and absorption for each distance bring extra layers of chemical and physical information from samples and may impact the predictive ability of multivariate calibration models. Thus, the SARDR probe has potential to obtain measurements under different penetration depths, absorption levels and scattering properties, amplifying the range of information available for analysis. What is relevant to understand is the type and quality of information obtained by each SD separation, screening the best situations for maximising the synergism obtained by their combination.

4.3 Experiments and methods

A total of 56 sample runs were performed in the system illustrated in the Figure 4.4 below, spanning the mean bead diameter range from 0.400 to 1.540 mm and the concentration range from 15 to 50%, at 5% intervals.

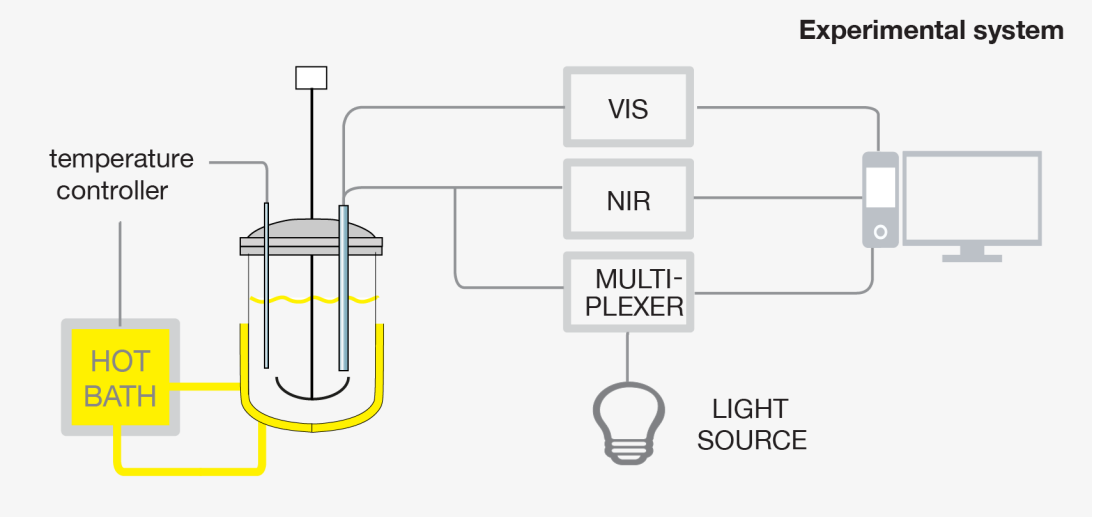

Figure 4.4 Schematic illustration of the experimental system.

For all sample runs, temperature and stirrer rate were maintained constant at 24°C and 250 rpm, respectively. Measurements in the near-infrared region were obtained by a NIRQuest-512 spectrometer and in the visible region by the USB-4000 spectrometer, both from Ocean Optics. The spectrometers were connected to the SARDR probe, and a multiplexer was employed for the coupling of fibres. Further details of the experiment can be found in Chapter 3, section 3.1.

4.4 Model development and validation

PLS multivariate regression models were developed between the two quality parameters namely, particle concentration and particle diameter and each spectroscopy mode and source detector distance. Prior to modelling, each spectral data set was subject to normalisation, background correction, de-noising and conversion from reflectance to logarithmic units, as detailed in the Chapter 3 Section 3.3. For all models, both spectra (**X** data) and the property of interest (***y*** data) were autoscaled.

To guarantee that the predictive capability of all models will hold true on future unseen samples, a PLS strategy based on repeated double cross-validation was employed for model validation. Moreover, since the output of several different models will be compared in this work, this technique was also chosen to allow for a reliable comparison among them. Within the rdCV, a 5-fold cross validation was carried out for a maximum of 20 latent variables, to determine the optimum number of PLS components. The whole iteration (inner and outer loops) was repeated 100 times. In this way, the predicted performance and its variability can be better estimated, since the number of available test set y-values is enormously increased. Moreover, by having an optimum number of components in each of these iterations, its overall variability can be estimated and a final optimum number of components can be derived for a model containing all objects [12] . The outputs for every model are the root mean squared error of cross validation (RMSECV), root mean square error of prediction (RMSEP) and their distributions, squared coefficient of correlation (R^2), optimum number of PLS components and its frequency plot, and the residuals for all estimations, as it was already explained in Chapter 2 section 2.5.

4.5 Impact of different source-detector distances and angles of incidence in the predictive ability of Vis and NIR models

The model system used for evaluation of the SARDR optical probe and modelling strategies discussed in this chapter is composed of expandable polystyrene beads suspended in water. Hence, water and polystyrene are the sole chemical entities that affect the NIR and visible spectra in this study. Chemical effects, however, are not the only manifestation in this system, since physical effects also have a strong influence over the spectrum. By varying the distance between the detector and the light source, these manifestations may be detected differently by the Vis and NIR sensors, which may contribute to a richer extraction of information from the samples. The information captured by the Vis and NIR spectra collected at each SD distance and angle of incidence will be evaluated for the determination of polystyrene concentration and EPS bead size.

4.5.1 Determination of polystyrene concentration

In order to evaluate the spectroscopic techniques and calibration models without the disturbances present in a typical reaction, fully converted EPS beads suspended in water are analysed through an experiment run under controlled temperatures and stirrer speeds. SARDR Vis and NIR spectra were acquired from 56 sample runs spanning a range of 15 to 50 weight per cent of EPS beads, from each of the seven illuminating fibres placed at different distances and angles. After spectral correction, normalisation with the integrating sphere spectrum, denoising, and conversion from reflectance to logarithmic units, the spectra were ready for the development of the PLS multivariate calibration models.

4.5.1.1 Visible region

Visible spectroscopy is not a natural choice when the goal is to monitor polystyrene concentration in polymers, especially due to the absence of related absorption bands in this region. Despite this, since the concentration of beads affects the spectral baseline, it may be possible to correlate the shifts in the spectra with the concentration of polystyrene in the sample. This is exemplified in the Figure 4.5, which shows the visible spectra from samples of same mean bead size distribution but under four different polystyrene concentrations (15, 25, 35 and 45%).

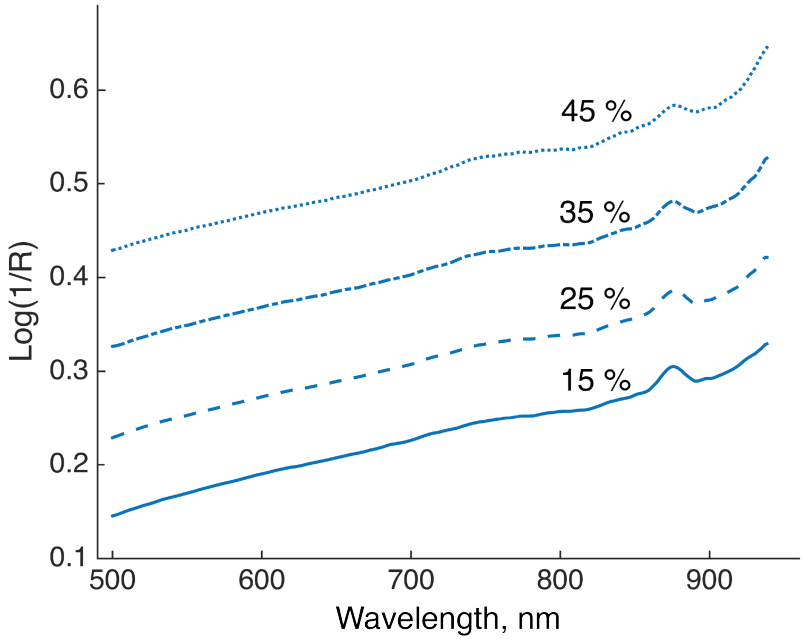

Figure 4.5 Vis spectra taken from four sample runs under different polystyrene bead concentrations and the same mean bead diameter of 0.890 mm.

Indeed, the figure clearly shows that the visible spectrum is affected by changing the concentration of polystyrene beads. The spectrum of each sample is shifted up as the bead concentration is increased and their absorbances become stronger as the wavelengths widens, since less light is scattered. The visible spectrum may not explicitly contain chemical information, but since it can be affected by the size of

beads, it may be correlated to their concentration. That happens because the number of particles per unit volume, i.e., the particle number density, can also affect the bead concentration. This can lead to a situation where it is possible to have the same bead concentration in three different situations: different bead sizes but same number density, same bead size but different number density or a combination of both.

Although the effect of changing concentration can be clearly seen on the spectra, it remains to be seen whether there is a consistent pattern and if the modelling strategies are able to capture such information. For this, regression models were developed between Vis spectra and polystyrene bead concentration for every one of the seven illuminating fibres of the SARDR probe. The statistical summary of all models is shown in Table 4.1.

Table 4.1 Summary of results of rdCV PLS regression models for the determination of EPS bead concentration in water by using the SARDR probe in the visible range. RMSECV and RMSEP are shown with their 95% confidence interval.

	PLS Repeated Double CV			
Distance mm	**RMSECV** % w/w ±CI95%	**RMSEP** % w/w ±CI95%	**R^2**	**LV**
0.6 / 45°	6.4 ± 0.06	6.5 ± 0.14	0.88	6
0.6	6.2 ± 0.08	7.7 ± 0.27	0.88	4
1.2	6.7 ± 0.05	6.7 ± 0.14	0.84	6
1.8	3.3 ± 0.02	3.4 ± 0.07	0.95	5
2.4	3.2 ± 0.04	3.5 ± 0.08	0.95	4
3.0	2.7 ± 0.02	2.9 ± 0.05	0.96	4
4.2 / 30°	4.5 ± 0.04	4.4 ± 0.10	0.93	4

Bead concentration range: 15 – 50 % w/w

In Table 4.1, a comparison of SD distances and angles of incidence based on the average RMSECV and RMSEP obtained by rdCV models, reveal a clear distinction between longer and shorter source-detector distances. While wider separations resulted in lower errors and better predictive ability, models for the shortest distances provided poorer predictions and lower coefficients of correlation, R^2. The lowest error was obtained by the illuminating fibre located 3.0 mm distant from the detecting fibre, i.e., the widest separation among the normal incident fibres, estimating bead concentration with an error lower than 3% ± 0.02. The same pattern was found for the angular incident fibres, where the farthest fibre also achieved a lower error and required a lesser number of latent variables to explain the data variance.

It is important to remember that these metrics presented in Table 4.1 were obtained from 100 repetitions of double cross validations, each of them having different random splits of the data into test and calibration sets. In this way, the optimum number of latent variables could be derived from 100 sub-models.

For the model that gave the best results (distance 3.0 mm), most of the sub-models have shown similar complexity by selecting four latent variables as the optimum number (almost 70% of them). This is shown in Figure 4.6 a and b, where the frequency distribution is depicted, followed by a scree plot showing the RMSECV values as a function of the number of latent variables (LV) for all 100 repetitions. In Figure 4.6 b, the blue curve is the average RMSECV for every latent variable number while the grey shadow represents the errors for all 100 individual models.

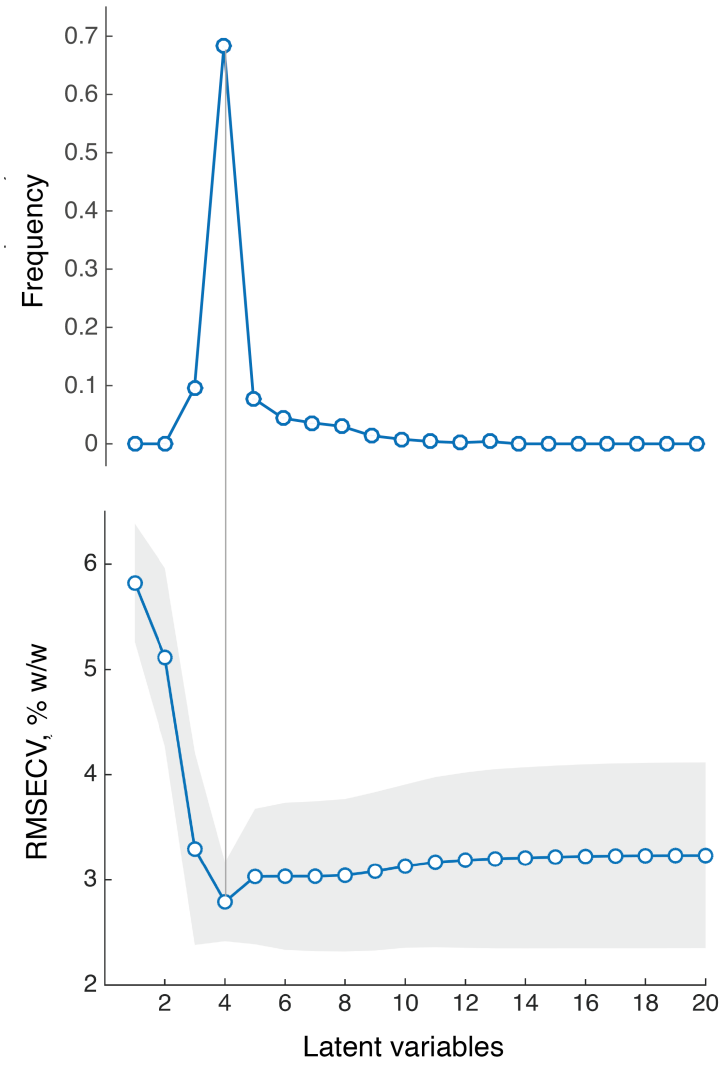

Figure 4.6 a) Frequency distribution of the optimum number of LVs chosen from 100 repetitions of double cross validation models. b) average RMSECV (blue curves) as a function of LVs obtained for the bead concentration model based on Vis spectra (SD distance = 3.0 mm). The grey shadow represents the RMSECV obtained from all repetitions.

For the same model, the estimated versus measured polystyrene concentration is shown in Figure 4.7 below, where the blue circles represent the mean estimates calculated from 100 double cross validation repetitions, and the grey crosses represent the 5600 estimated values (12 prediction samples *versus* 5 cross validation segments *versus* 100 repetitions). A good fit between the measured and estimated values can be observed, with a squared coefficient of correlation of 0.96.

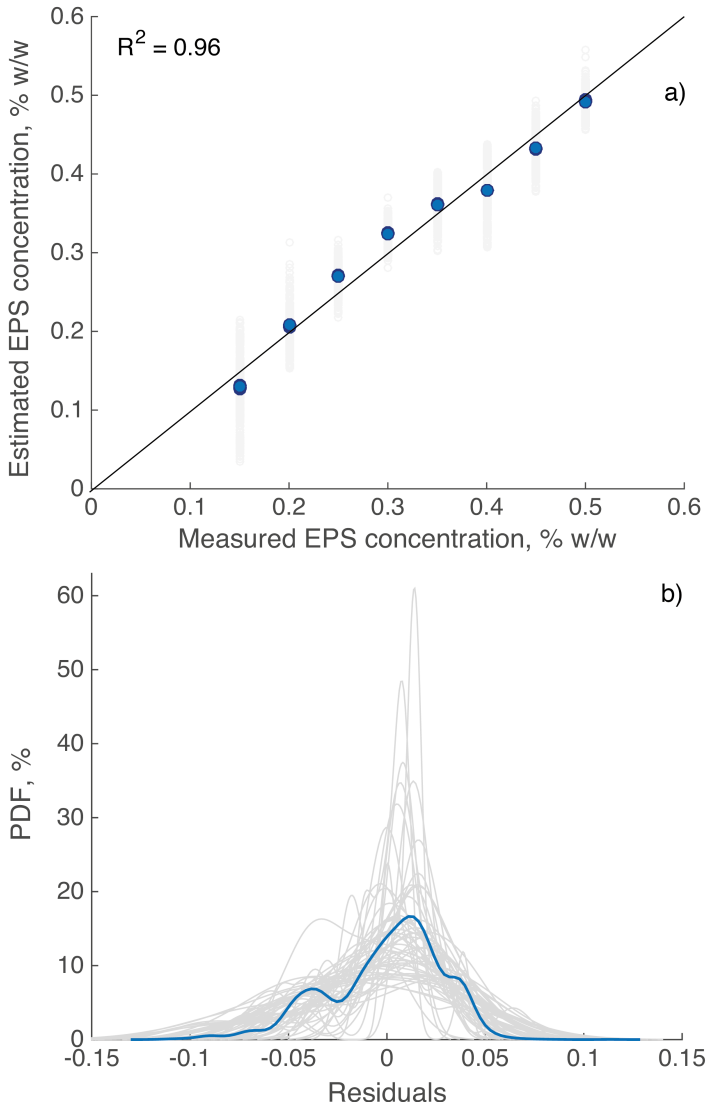

Figure 4.7 a) Measured *versus* estimated bead concentration obtained by models developed based on Vis spectra (SD distance = 3 mm) and b) probability density plot from all residuals from the rdCV models.

The distribution of the prediction errors is shown as light grey curves in Figure 4.7 b, whereas the blue curve represents the average Probability Density Function (PDF) from 100 double cross validation repetitions. These PDF curves provide important information about model performance, indicating how far the estimated values are from the reference ones.

4.5.1.2 Near-infrared region

The presence of strong polystyrene absorption bands in the NIR region makes the technique relevant for monitoring polymer concentration. Figure 4.8 presents a typical NIR spectrum taken from the system in question, spanning the wavelength region of 1000 to 1820 nm.

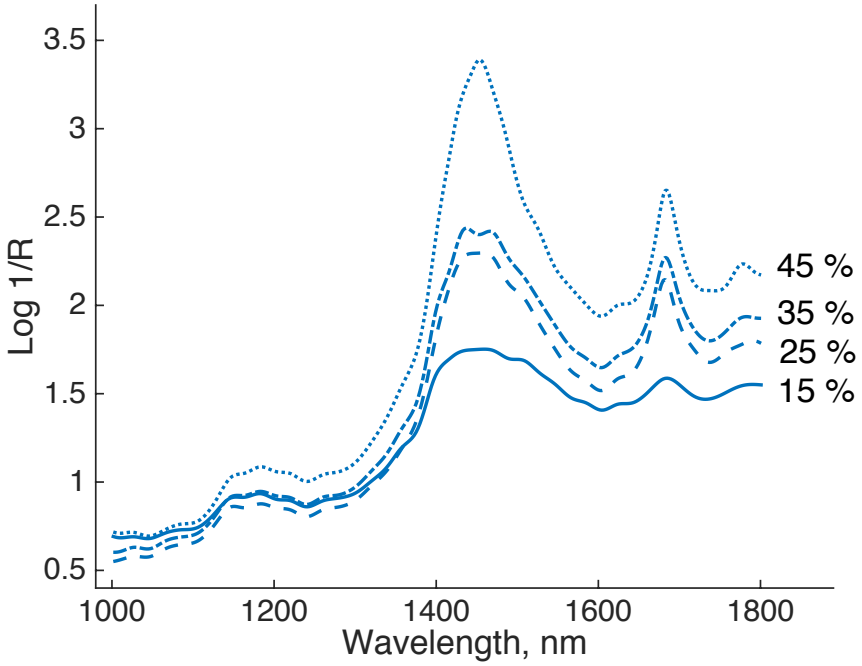

Figure 4.8 NIR spectra of EPS beads suspended in water at four different concentrations.

The most important NIR region for monomer-polymer conversion monitoring[8] is located in a window between 1580 to 1680 nm. While polystyrene absorbs mainly around 1640 and 1684 nm due to the first overtone of the aromatic C-H stretch, styrene bands can be visualised centred around 1580 nm and 1670 nm. Consequently, in a polymerisation progresses, as the styrene is continuously converted in polystyrene, the intensity of the absorption at the styrene C-H vibration bonds vanishes (1580 and 1670 nm) through out the reaction, at the same time that

the aromatic C-H stretch in the first overtone has its absorption intensity (1640 and 1684 nm) increasing continuously. A weaker, but not less important, second overtone band of the aromatic C-H stretch can also be seen around 1142 nm. The spectrum is also dominated by the overtone of water between the 1400-1600 nm, but since the strong water influence is confined to this region, the polystyrene-styrene fingerprint region is not affected.

Based on previously published applicability[9,10] of near-infrared spectroscopy for monitoring styrene-polystyrene conversion in polymerisation reactions, it was expected that models developed here for polystyrene concentration by using the NIR-SARDR probe would provide good results. However, that was not the case for most of the source-detectors distances, as can be seen in Table 4.2, which summarises the results obtained by the developed rdCV regression models.

Table 4.2 Summary of results of rdCV PLS regression models for the determination of EPS bead concentration in water by the SARDR probe in the NIR range. RMSECV and RMSEP are shown with their 95% confidence interval.

	PLS Repeated Double CV			
Distance mm	**RMSECV** % w/w (±CI95%)	**RMSEP** % w/w (±CI95%)	R^2	LV
0.6 / 30°	5.4 ± 0.04	5.6 ± 0.11	0.87	4
0.6	5.1 ± 0.06	5.4 ± 0.12	0.90	5
1.2	7.4 ± 0.08	7.7 ± 0.19	0.78	6
1.8	7.1 ± 0.04	7.1 ± 0.13	0.74	4
2.4	6.8 ± 0.05	7.0 ± 0.12	0.78	4
3.0	7.8 ± 0.05	7.9 ± 0.14	0.70	4
4.2 / 45°	7.0 ± 0.05	7.2 ± 0.13	0.74	4

Bead concentration range: 15 – 50 % w/w

All models provided errors higher than 5% for bead concentration. In fact, the best model was able to estimate bead concentration by using five latent variables with an error of 5.4%, whereas in the poorest result, the RMSEP obtained was 7.9 %, with a coefficient of determination of 0.70, indicating a poor fit of the model.

Contrary to expectations, the errors in general were higher at wider SD separations. The information gathered by the NIR spectra acquired on wider distances tend to be richer, since photons that were able to reach the detector generally have a bigger probability of travelling deeper in the sample. At the same time, this deeper penetration depth normally may come at the expense of a lower signal-to-noise ratio, since fewer photons may be able to reach the detector, per unit time, without being absorbed. Therefore, light intensity needs to be strong enough to reach all available distances on the probe with sufficient power, and that may have been one of the problems faced by the NIR models developed here. Although 400 μm fibres were employed in the SARDR probe, light strength may have not been sufficient, affecting negatively the sensitivities of the spectra collected at the wider distances.

To exemplify this argument, the NIR spectra obtained from four different bead concentrations (15, 25, 35 and 45%) are plotted for each of the five normal incident source-detector distances (Figure 4.9). The spectra from different SD distances are presented with an offset to facilitate visualization of their differences.

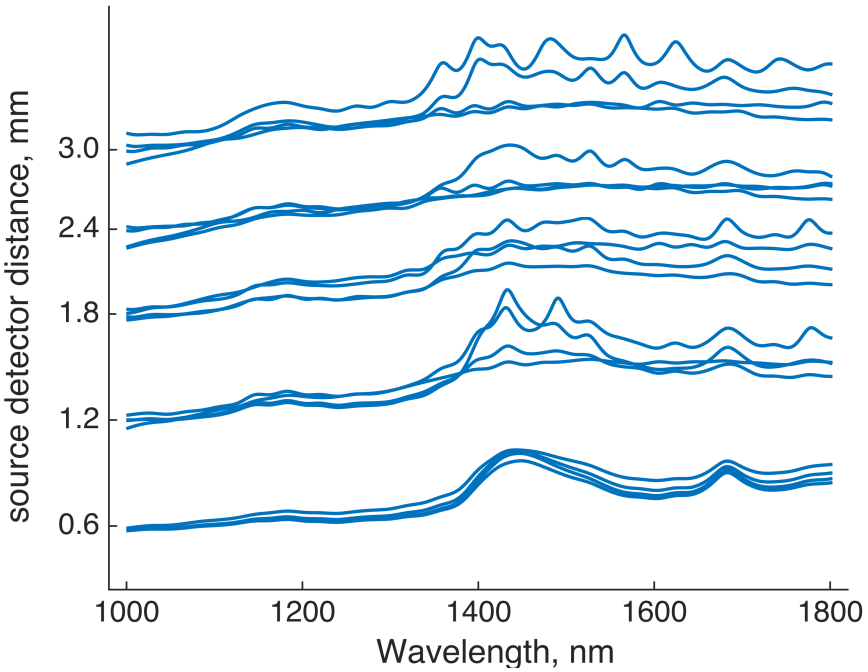

Figure 4.9 NIR spectra from four different bead concentrations taken from five source-detector distances.

By inspecting the Figure 4.9, it is clear that the spectra from short SD distances are of much better quality. As the SD distance increases, the variance of the spectra increases, and the their character become anomalous, especially in the region above 1400 nm. It is interesting to remember that longer wavelengths in the NIR spectrum penetrate the medium in a greater extent than do shorter wavelengths. This may partly explain the better quality of the signal until 1400 nm for all distances, since at shorter wavelengths the penetration depth is shallower and the scattering higher. However, the main contributing factor is the weaker light throughput observed at wider SD distances, resulting in a lower signal-to-noise ration for those measurements.

4.5.2 Determination of mean EPS bead diameter

Obtaining reliable estimates of the mean EPS size and its distribution along a reaction allows the production of a final product with fewer losses and an easier post-production processing[11]. In this section, similarly to the determination of bead concentration, the SARDR spectra from 56 sample runs spanning the mean diameter range of 0.4 to 1.54 mm were collected for all SD distances and angles of incidence. A large quantity of post-production beads was sieved to obtain several narrowed size ranges, from which the mean bead diameter and its distribution were measured, as discussed in Chapter 3 Section 3.1, and shown Table 4.3, below.

Table 4.3 Mean bead sizes and their span calculated for each diameter range

Range	**Diameter range,** mm	**D50,** mm
1	0.300:0.500	0.405 ± -
2	0.500: 0.630	0.565 ± 0.190
3	0.630: 0.800	0.699 ± 0.222
4	0.800: 1.000	0.890 ± 0.201
5	1.000: 1.250	1.108 ± 0.187
6	1.250: 1.400	1.314 ± 0.115
7	1.400: 1.600	1.540 ± 0.264

4.5.2.1 Visible region

The spectrum in the Visible range is strongly affected by light scattering coming from the polymer beads. It contains information on the scattering behaviour of the samples, and, therefore, on the particle size. It has been used extensively for particle sizing applications[12], especially due its susceptibility to light scattering. Table 4.4 presents the summary of the rdCV regression models developed for mean EPS bead diameter based on the Visible spectra of every source-detector separation and angle of incidence.

Table 4.4 Summary of results of rdCV PLS regression models for the determination of EPS bead size by the SARDR probe in the visible range. RMSECV and RMSEP are shown with their 95% confidence interval.

	PLS Repeated Double CV			
Distance mm	**RMSECV** mm ±CI95%	**RMSEP** mm ±CI95%	**R^2**	**LV**
0.6 / 45°	0.123 ±0.001	0.132 ±0.003	0.94	4
0.6	0.079 ±0.001	0.091 ±0.005	0.99	6
1.2	0.201 ±0.002	0.204 ±0.004	0.89	6
1.8	0.081 ±0.001	0.083 ±0.002	0.98	6
2.4	0.098 ±0.001	0.098 ±0.002	0.97	5
3.0	0.077 ±0.001	0.076 ±0.002	0.97	4
4.2 / 30°	0.151 ±0.001	0.155 ±0.003	0.96	7

Mean bead diameter range: 0.400 to 1.540 mm.

It can be seen from Table 4.4 that apart from one fibre (1.2 mm), all other normal incident fibres resulted in good predictive capabilities by showing low errors of prediction. Similarly to what happened with the bead concentration models, the best model was obtained for the longest distance (3.0 mm). However, in this case, the errors obtained from the other distances were not remarkable different. The 3.0 mm-distant fibre was able to estimate mean bead diameter with an error lower than 0.08 mm ± 0.002. Both angular fibres provided poorer models when compared to the normal incident fibres. Figure 4.10 a and b shows the measured versus estimated values obtained by the rdCV models for the source-detector distances which provided the best (3.0 mm) and the poorer (1.2 mm) performance. Here, the blue circles represent the mean estimated value calculated from 100 repetitions of PLS double cross validations, and the light grey circles represents the results from all those 100 models

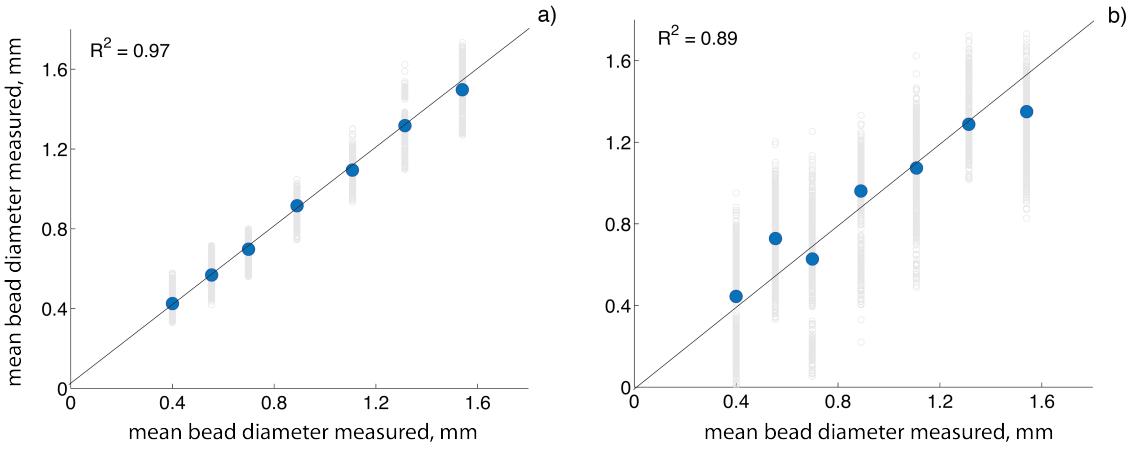

Figure 4.10 Measured *vs.* estimated mean bead size for the SD distances which resulted in the (a) best and (b) worst performance.

The model built when the illuminating fibre was placed 3.0 mm distant from the collecting fibre obtained an excellent fit. The coefficient of correlation reached 0.97 using only four latent variables to describe the data variance. Furthermore, the results for this fibre and most of the other ones were better than the method of reference, as

can be seen from Table 4.3, which shows that the reference errors for the individual sample size ranges vary from 0.115 to 0.264 mm. The fibre placed at 1.2 mm from the collecting fibre provided the poorer results, but still shows a relatively good fit, as it can be seen in the right hand side plot of Figure 4.10. The variation of results obtained by all 100-model repetitions is much wider than the one from distance 3.0 mm, but show relatively good linearity. Other evaluation tools for these two models can be seen in Figure 4.11 a to d, below.

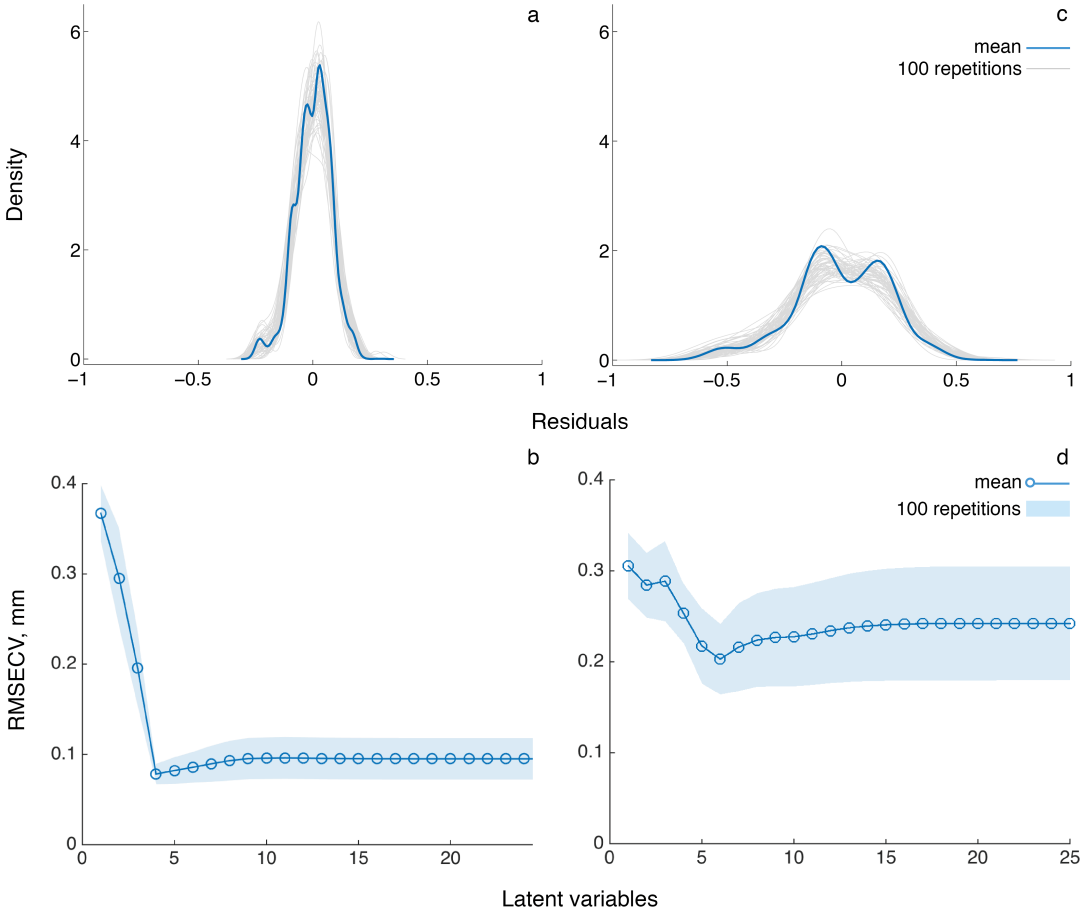

Figure 4.11 Diagnostic plots for the SD distances that provided the best (a and b) and the poorest (b and c) performance. The probability density functions calculated from 100 repetitions are shown in a and c. Blue curves represent the mean residuals and the grey ones the 5600 residuals from 100 repetitions. The RMSECV *vs* LVs selected by all models are shown in b and d. Blue circles represent the mean values and the light blue shadow all 100 models.

The residuals obtained from all models for both SD distances are presented in Figure 4.11 a and c, and corroborate the previous results by showing that the best model has a much narrower distribution. Besides, it needed a lower number of latent variables and obtained RMSECVs varying tightly around the mean errors (Figure 4.11 b and d). It is important to note that the residual distribution for the 1.2 mm distant fibre differs from a normal distribution by presenting two slightly different distributions, which may be an indicative of an outlier sample or group of samples that are skewing the estimations.

4.5.2.2 Near- infrared region

Similarly to the NIR models for polystyrene concentration, the PLS models for mean EPS bead diameter have also provided poorer results for most of the SD distances, as it can be seen in Table 4.5, below, which summarises the results.

Table 4.5 Summary of results of PLS regression models for the determination of EPS bead diameter based on the NIR-SARDR probe. Results from rdCV are compared to the ones from a PLS model built with all samples

	PLS Repeated Double CV			
Distance mm	**RMSECV** mm ±CI95%	**RMSEP** mm ±CI95%	**R^2**	**LV**
0.6 / 30°	0.153 ±0.002	0.165 ±0.004	0.92	4
0.6	0.120 ±0.002	0.144 ±0.006	0.95	7
1.2	0.207 ±0.002	0.220 ±0.005	0.85	6
1.8	0.318 ±0.004	0.362 ±0.010	0.59	4
2.4	0.302 ±0.003	0.307 ±0.006	0.64	4
3.0	0.306 ±0.002	0.311 ±0.006	0.61	4
4.2 / 45°	0.314 ±0.003	0.325 ±0.007	0.63	5

Mean bead diameter range: 0.400 to 1.540 mm.

The predictive power of the NIR models generally decreased with the widening of the SD distance, and the models were more complex at the shortest distances, i.e., they needed more latent variables to describe the data variance. The best result was obtained by the shortest distance (0.6 mm), which needed seven latent variables to describe the model with an error of 0.120 ± 0.002 mm and squared coefficient of correlation of 0.95, indicating a reasonably good fit. Although the errors are still acceptable when compared to the reference errors, they are much poorer than the ones obtained by visible spectroscopy. All the remaining SD distances did not provide sufficient predictive power, as evidenced by the high errors and poorer squared coefficients of determination, R^2.

4.6 Conclusions and Outlook

The goal of this chapter was to investigate the performance of the visible and near-infrared SARDR probe on estimating composition and size of expandable polystyrene beads. Particular interest was placed on studying the effect of source-detector separation and angle of incidence on the spectral sensibility. A controlled experiment based on polystyrene beads suspended in water under different sizes and concentrations was devised to better isolate the impact of pathlength variation and spectral sensitivities of each source-detector separation on the spectra and the models derived from them.

As discussed earlier, signal strength solely may not be a good indicator of the feasibility for predictive quality. Although the quality of the signal may be poorer when acquiring spectra at longer SD distances, those photons travelling farther have had a higher chance of interacting with a larger volume of EPS beads, being scattered or absorbed, or penetrating deeper in the sample, which may result in better sensitivities and better models.

In general, modelling based on longer SD separations in the visible range resulted in better outputs, providing lower errors and simpler models. Polystyrene beads are scatterers par excellence, and since scattering is stronger at lower wavelengths, spectrum in the visible range is naturally more affected by it. Furthermore, as the distance from the collecting fibre is increased, scattering tends to be stronger. Particle sizing applications can take full advantage of that.

On the contrary, near infrared models developed with spectra collected by the SARDR probe were not able to provide results up to the potential of the technique. Although 400 μm fibres were employed in the SARDR probe, the light throughput may have not been stronger enough to reach all detecting fibres with sufficient

signal-to-noise ratio, which affected negatively the sensitivities of the spectra collected, as shown in Figure 4.9.

It is important to remember from the experimental description in Chapter 3, that the integrating time needed to generate the NIR spectrum was 80 to 100 times higher than that required by the visible spectrum. Consequently, the NIR signal obtained before normalisation was very noisy. The background spectrum acquired in the integrating sphere was even poorer, lower than those acquired on the sample, which may have contributed to the low quality reflectance signals obtained after normalisation.

Overall, visible models provided better models for both mean bead diameter and bead concentration. For most of the SD distances, models based on visible spectra were able to provide estimates with much lower errors, tighter residual distributions, and a lesser number of latent variables needed to describe the models.

4.6.1 Outlook

In this chapter individual models were developed for each source-detector distance in order to extract the information collected from each of them. The results have shown that each distance probes a different depth of information, being it on scattering, absorption or deeper levels of penetration. But what would be the impact of integrating these different sources of information in just one model? Could we enhance the predictive ability of multivariate calibration models by just combining them? In the next chapter I will investigate ways of integrating information from different SD distances in order to improve the predictive performance of models.

4.7 Bibliographic references

1. Barer, R., Cole, A. R. H. & Thompson, H. W. Infra-Red Spectroscopy with the Reflecting Microscope in Physics, Chemistry and Biology. *Nature* **163,** 198–201 (1949).

2. Suzuki, S., Takasaki, S., Ozaki, T. & Kobayashi, Y. Tissue oxygenation monitor using NIR spatially resolved spectroscopy. *BiOS '99 International Biomedical Optics Symposium* **3597,** 582–592 (1999).

3. Cubeddu, R., Pifferi, A., Taroni, P., Torricelli, A. & Valentini, G. Noninvasive absorption and scattering spectroscopy of bulk diffusive media: An application to the optical characterization of human breast. *Appl. Phys. Lett.* **74,** 874 (1999).

4. Nguyen Do Trong, N. *et al.* Optical properties–microstructure–texture relationships of dried apple slices: Spatially resolved diffuse reflectance spectroscopy as a novel technique for analysis and process control. *Innovative Food Science & Emerging Technologies* **21,** 160–168 (2014).

5. Nguyen Do Trong, N. *et al.* Differentiation of microstructures of sugar foams by means of spatially resolved spectroscopy. in (eds. Berghmans, F., Mignani, A. G. & De Moor, P.) **8439,** 843914–843914–9 (SPIE, 2012).

6. Igne, B., Talwar, S., Feng, H., Drennen, J. K. & Anderson, C. A. Near-Infrared Spatially Resolved Spectroscopy for Tablet Quality Determination. *Journal of Pharmaceutical Sciences* **104,** 4074–4081 (2015).

7. Hajime, M. in *Application of Near Infrared Spectroscopy in Biomedicine* 59–74 (Springer US, 2013). doi:10.1007/978-1-4614-6252-1_4

8. Workman, L. & Weyer, J. *Practical Guide and Spectral Atlas for Interpretive Near-Infrared Spectroscopy, Sercond Edition Spectroscopy, Second Edition*. (CRC Press, 2012). doi:10.1201/b11894

9. Dehanov, N. & Thennadil, S. Monitoring of conversion during emulsion polymerisation using near infrared spectroscopy. *NIR news* **25,** 9 (2014).

10. Nogueira, E. S., Borges, C. P. & Pinto, J. C. In-Line Monitoring and Control of Conversion and Weight-Average Molecular Weight of Polyurethanes in Solution Step-Growth Polymerization Based on Near Infrared Spectroscopy and Torquemetry. *Macromolecular Materials and Engineering* **290,** 272–282 (2005).

11. Silva, W. K., Chicoma, D. L. & Giudici, R. In-situ real-time monitoring of particle size, polymer, and monomer contents in emulsion polymerization of methyl methacrylate by near infrared spectroscopy. *Polym Eng Sci* **51,** 2024–2034 (2014).

12. Celis, M.-T., Forgiarini, A., Briceño, M.-I. & García-Rubio, L. H. Spectroscopy measurements for determination of polymer particle size distribution. *Colloids and Surfaces A: Physicochemical and Engineering Aspects* **331,** 91–96 (2008).

Chapter 5

5. On fusing spatially and angularly resolved spectra

The previous chapter has shown that spectroscopic measurements acquired at different source-detector separations and angles of incidence capture distinct information about the composition and size of polystyrene beads, diversely influencing the predictive capability of individual models developed with such measurements.

In the present chapter, it is investigated whether predictive improvements can be obtained by combining these different pieces of complementary information. For this purpose, a sensor data fusion strategy based on multiblock PLS is proposed.

The overall results demonstrated a better performance for all models when one or more extra blocks of SD distances are fused together, and especially when an angular incident fibre is included. Enhanced predictive power and reduced errors of prediction were achieved, allowing reliable estimates for polystyrene concentration and bead size in a polymer particulate system.

5.1. Data Fusion

According to an excellent definition given by Hall and Llinas[1], "data fusion techniques combine data from multiple sensors and related information from associated databases to achieve improved accuracy and more specific inferences than could be achieved by the use of a single sensor alone." It therefore seems natural to link data fusion tools to the process analytical technology framework. In fact, multi-sensor data fusion has been widely applied in PAT, with most of the applications being concerned with combining data from multiple spectroscopic sensors[2,3], process phases[4,5], and or process parameters[6-9].

For fusing information from different spectroscopy sensors, which is the interest of the present piece of work, the literature is vast for a wide range of industrial areas. In a biotech industrial application previously published by this author[2], a data fusion strategy was employed to combine NIR, mid-infrared and 2D-fluorescence spectra of two important process raw materials for production of therapeutics protein, enabling predictive models that outperformed any other models involving either one raw material alone, or only one type of spectroscopic tool for both raw materials.

In a similar fashion, Lee *et al*[10] used Ensemble PLS (EPLS) to combine information from four different spectroscopic techniques, NIR, Raman, 2D-fluorescence, and X-ray fluorescence (XRF), to predict raw material quality in mammalian cell culture. Again, data fusion models provided better estimates when compared to any of the single spectroscopic techniques.

NIR and Raman were also combined by multiblock methods for the quantitative prediction of multiple Active Pharmaceutical Ingredient (API) solid dosage forms[11], showing that each spectroscopy may contain information not present or captured with the other spectroscopic technique.

In food industry, where data fusion has seen most of its applications, the benefits of combining NIR and mid-infrared spectra were described for estimation of soybean flour quality properties[3], cultivar identification on virgin oils[12], and coffee varietal identification[13], amongst others.

Workman Jr. was one of the first to use data augmentation for fusing spectroscopic data in the polymer/chemical industry[14]. He combined NIR and Raman spectra for measuring blend composition in polymer films. Nonetheless, although fewer polymer applications of fusing spectroscopic data could be found, it was in the chemical and polymer industry that multiblock and multi-way techniques gained attention due to the developments from Professor Macgregor's group in Canada, especially on modelling different process phases and parameters[15-17]. Together with Professor Kiparissides in Greece, they were the first to apply PLS and PLS related multiblock techniques to monitoring and control of polymerisation reactions[17].

Despite the multitude of applications of data fusion on vibrational spectroscopic sensors, limited work has been published on fusing spatially resolved spectra. Spatially resolved data presents a tri-linear relationship between the sample, the wavelength and the SD distance modes, being represented by a three-dimensional hypercube, $\mathbf{X}$ ($\mathbf{n} \times \mathbf{k} \times \boldsymbol{\lambda}$), where $\boldsymbol{n}$ refers to the sample mode, $\boldsymbol{k}$ the SD distance, and $\boldsymbol{\lambda}$ the wavelength mode, as demonstrated in Figure 5.1. Therefore, several ways of handling such data structure for modelling purposes are available.

Modelling strategies can be developed directly based on the (a) three-way array, (b) on blocks of data from different source-detector separations, (c) augmented data or (d) co-added spectra from each SD distance. For each of these data handling types, different modelling approaches may apply.

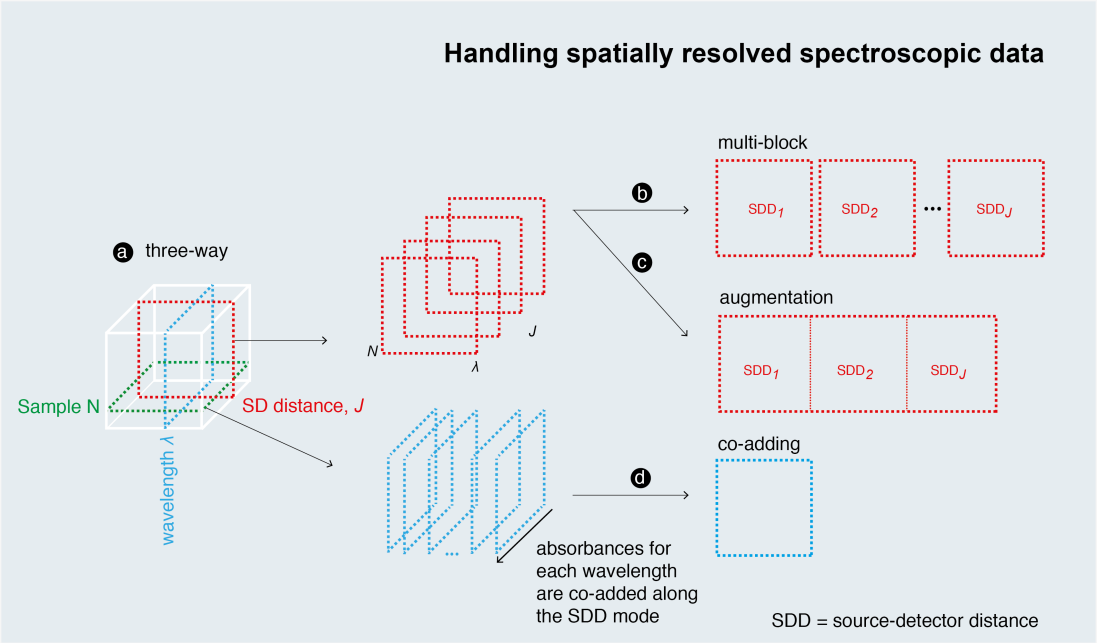

Figure 5.1 Handling SRS data: a) three-way, b) multiblock, c) data augmentation and d) co-adding approaches.

Models developed directly in the three-way array are built by using multi-way decomposition methods like Parallel Factor Analysis (PARAFAC). Tsuta *et al* [18] modelled the degradation of chlorophyll in apple skin during ripening and storage by using PARAFAC to decompose spatially resolved spectra from seven different source-detector distances. The authors were able to qualitatively differentiate between flesh and skin of intact apples based on the PARAFAC wavelength loadings, as well as differences in skin color and chlorophyll content. Bogomolov *et al*[19] applied tri-linear PLS to determine fat and protein in milk based on vis-NIR spectra collected from six SD distances. However, for their object of study, the three-way data structure did not present any advantage compared to the augmentation approach.

Alternatively, the hypercube data can be treated as a two dimensional dataset, which is accomplished by unfolding it into one large matrix (***n*k*** *x* ***λ***), where the spectral datasets for each SD distance are concatenated side-by-side Figure 5.1 c). In this way, each row of the resultant unfolded matrix represents the spectra acquired from every SD distance for a specific sample, and the columns represent the wavelengths. The advantage of using the unfolded matrix is the simplicity. In fact, handling all data blocks in a single matrix allows the use of standard partial least squares Regression (PLSR) for model building. However, the influence of each SD distance can be hidden by other SD distances' influence[20], also not allowing an evaluation of each block's influence to the overall model. Despite of this, Bogomolov *et al*[19] used this method for fusing information from different source-detector distances and obtained improved results when compared to the tri-linear PLS results.

Spatially resolved spectra, although measuring distinct locations and depths of penetrations, have similar shape and nature. This similarity allows the combination of the spectra at different SD distances by simple averaging their absorbances for each wavelength. Such procedure is commonly adopted due to its simplicity, and was employed by Benoit *et al*[21] for assessing pharmaceutical active principle ingredients within tablets by using averaging spectra collected from different multi-point fibres. However, the spatial information may be flattened and lost by averaging process, decreasing the amount of knowledge that can be extracted from the spectra.

Although applications of multiblock methods on spatially resolved measurements could not be found in the literature, the technique has potential to extract complementary information from such measurements. Multiblock PLS was first proposed for monitoring and diagnosing of separate process phases[22], or any conceptually meaningful blocks [23], but the advantages for combining data from spectroscopic datasets were soon explored.

In predictive terms, MB-PLS provides similar results than the data augmentation approach[23,24]. Its advantage over data augmentation resides in the improved interpretability of models[7] and easier diagnostic procedures[25]. This happens because in MB-PLS, the features of each block are initially modelled separately, and the summarised information is then aggregated from each block to form a super block, containing the relationship between the blocks.

Despite the cited advantages, integrating data from different sensors, however, does not necessarily guarantee that results will improve for all applications. In the literature, there are several examples of fusing strategies that did not bring improvements in the predictive performance[26], or even deteriorate the results. The same data integration may work for one analyte but not for another[27]. Thus, the type and object of fusion needs to be analysed with regard to their complementary characteristics and potential synergism, and ultimately tested with real data.

In this chapter, we will address this problem by integrating spatially and angularly resolved spectral datasets by using a strategy based on MB-PLS, which will be investigated in its potential to yield a more accurate estimation of important polymer quality attributes, modelling the relationship between the different source-detector separations, and establishing the contribution of each of them to the final model. To the best of my knowledge, this is the first attempt to employ multiblock PLS as a data fusion strategy on SARDR spectra. Moreover, the application of such data fusion approach in polymer particulate systems is also a new contribution.

5.1.1 Multiblock PLS

MBPLS is an extension of the ordinary PLS regression method, with the difference being that it is performed relating several blocks for the purpose of prediction or interpretation. Multiblock PLS operates in two distinct levels[7]: a sub-level, which works similarly as individual PLS models, where the features of each block are modelled separately; and a super level, which aggregates this summarised information from each block to form a super block.

In the sub-level, each spectral block (SD distances spectral datasets) and the response block (polystyrene concentration or mean bead diameter) are decomposed into a sum of components for every latent variable, according to the equations below.

$$\mathbf{X} = \mathbf{t}\,\mathbf{p}^{\mathbf{t}} + \mathbf{e} \tag{5.1}$$

$$\boldsymbol{y} = \mathbf{u}\,\mathbf{q}^{\mathbf{t}} + \mathbf{f} \tag{5.2}$$

This decomposition generates the scores (**t** and **u**), loading (**p** and **q**) and the residuals (**e** and **f**) for **X** and ***y***, respectively, for each block. The basic multiblock procedure is illustrated in the Figure 5.2 through an array scheme.

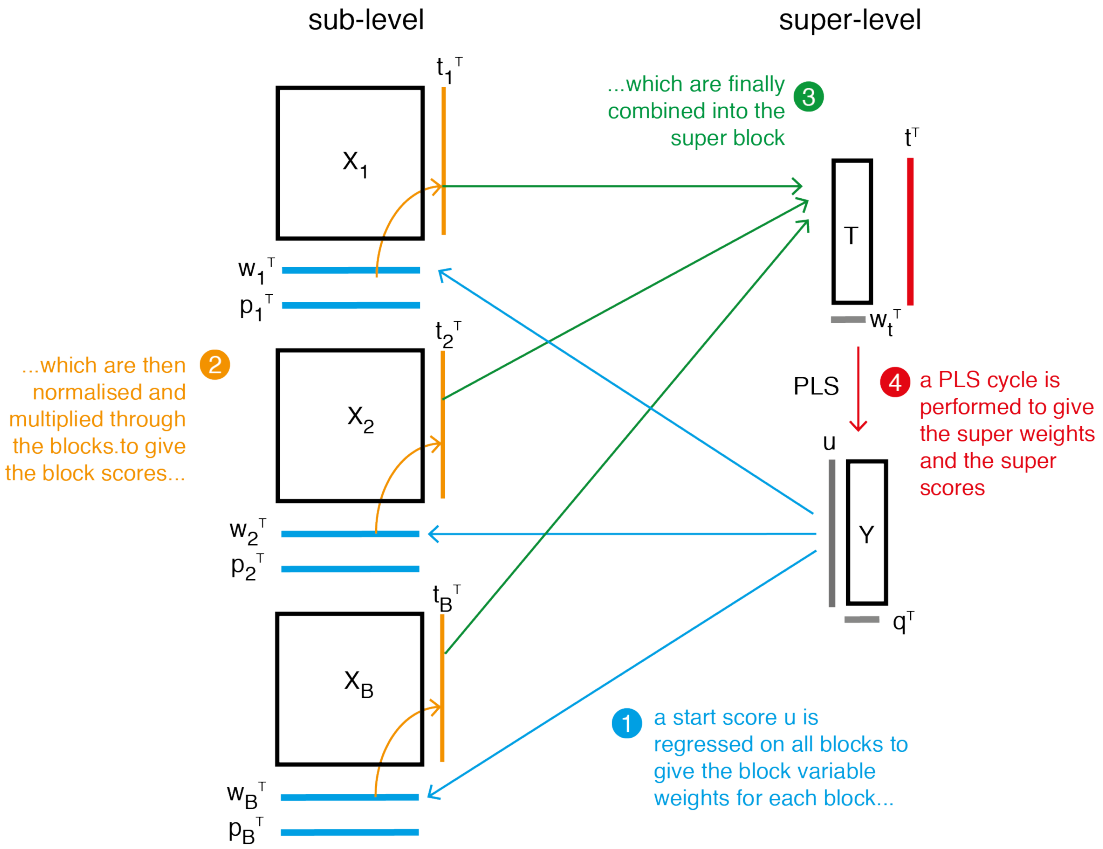

Figure 5.2 Array scheme of the basic Multiblock PLS. The procedure described is repeated until convergence of the super scores, t^T. Adapted from Westerhuis *et al* (1998).

A score *u* taken from the response vector is regressed on all blocks to give their block variance weights, which are very useful for interpreting block importance. These block weights are then multiplied through their respective **X** block to give the block score. The block scores group data features from each block, which are aggregated in a super block, obtained by unfolding all block scores side-by-side.

Predictions for new samples are obtained by applying the calibration scaling factors and the regression coefficients (b_{PLS}) to the unfolded super scores matrix, according to the equation below.

$$\hat{y}_{new} = \mathbf{X}\text{new}\ \mathbf{b}_{PLS} \qquad 5.3$$

There are several variants of the MB-PLS method[4,6], two of them widely used in the multiblock community. They differ mainly on the chosen deflation procedure. While Wangen and Kowalski[22] proposed the use of the block scores for the deflation of the blocks, Westerhuis and Coenegracht[4] suggested the use of the super score for the deflation step, to avoid loss of information on the deflation process. This variant of the multiblock PLS will be used in this work.

5.1.2 Software

All software programs were written and implemented in MATLAB™ programming language (Mathworks Inc.). For repeated double cross validation, the Integrated Library for Partial Least Squares and Discriminant Analysis (libPLS)[28] was used with adaptations made by this author. The libPLS is open source and freely available at www.libpls.net. Multiblock PLS[8] was adapted from a routine kindly supplied by Dr João A. Lopes, of the Chemometrics & Process Analytical Technology Unit, University of Porto, Portugal. Calculations and data analysis were performed in an iMac 21.5-inch with 2.7 GHz Intel Core i5 processor and 8 GB 1600 MHz memory.

5.2. Integrating diffuse reflectance spectra from different source-detector separations and angles of incidence

For the MB-PLS, each source-detector separation spectral dataset is treated as one block. For each of the spectroscopy modes, visible and NIR, seven predictor blocks were available for modelling, consisting of five normal incident SD separations (0.6, 1.2, 1.8, 2.4 and 3.0 mm), and two angularly incident SD separations (30° and 45°). Moreover, two response blocks were also available, the mean bead diameter and EPS concentration, which were used one at a time.

Spectral blocks were fused in several different combinations between them. These combinations were based on the results obtained for each individual SD distance, as seen in the previous chapter, and also on probe design requirements. First, fused models were built based on the blocks of normal incident fibres, by adding one block at a time, until all SD distances were added to the modelling strategy. The purpose was to evaluate how the increase in the SD distance can impact the predictive ability of models. Fused models were then built with both blocks of angularly resolved spectra, and finally, an angularly incident fibre was combined to all normal incident blocks to investigate whether the addition of angularly orientated information could improve the results. When building the models, every block was mean centred and divided by their standard deviation. Further normalisation was not needed since all blocks are similar, sharing the same type and quality of measurements.

Similarly to the individual models seen in Chapter 4, MB-PLS models were built and validated by using a strategy based on repeated double cross-validation. Five-fold cross validation was performed in both loops. The whole procedure was repeated 100 times in an additional loop where calibration and test sets are randomly created. The output for every model are the RMSECV, RMSEP, optimum number of PLS components and its frequency plot, the distribution of RMSECV and RMSEP errors.

5.3. Determination of polystyrene bead concentration

In Chapter 4, PLS models were developed between each individual source-detector distance spectral datasets and EPS bead concentration for both visible and NIR spectroscopies. Overall, the results indicated that NIR models were better when the illuminating fibre was placed closer to the detecting fibre, whereas visible models were better at longer separations, as can be seen in Figure 5.3, which summarises the results for polystyrene concentration.

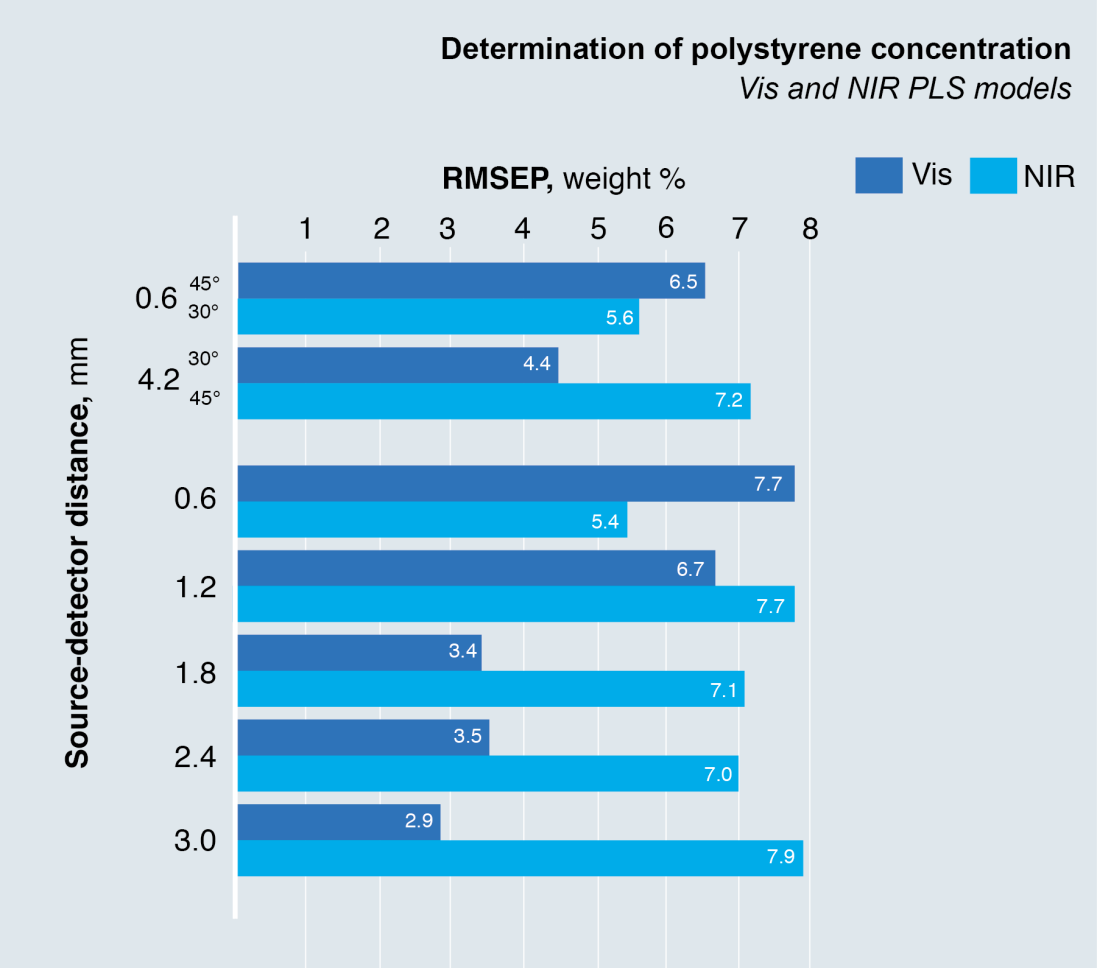

Figure 5.3 Prediction performance of rdCV PLS models developed by using visible and NIR spectra at different SD distances to estimate polystyrene concentration.

The PLS models developed with the best performing distance for NIR obtained an average error of 5.4 % w/w. For Visible, the best result yielded an average error of only 2.9 weight %. Based on these conclusions, the best performing distances for NIR and Vis will be used as a benchmark for comparison with the results obtained by data fusion. In this case, the shortest separation between the illuminating fibre and the NIR collecting fibre (SD distance=0.06 mm) will be selected for NIR and the farthest illuminating fibre from the Visible detector (SD distance =3.0 mm) for the Vis models.

5.3.1 NIR-SARDR data fusion

As discussed previously, near infrared models had a weaker performance when built using individual source-detector separations. Nevertheless, important information present in each spectral dataset, although individually not sufficient to estimate polystyrene concentration with good accuracy, may be complementary if analysed together. In order to investigate whether combining these individual datasets could improve the predictive power of the models, multiblock PLS models were developed with several combinations of fibres. First, a two-block model was built and validated by integrating spectra from just one extra fibre (SD distance = 1.2 mm) to the 0.6 mm distant fibre. A further fibre (SD distance = 1.8 mm) was then added and the MB-PLS model updated, and subsequent distances added one at a time, until all of the normal incident fibres were included in the model, ending with a five-block model.

A summary of the results obtained by the data fusion strategy for estimation of bead concentration is shown in the Table 5.1

Table 5.1 Prediction performance of MB-PLS models for determination of EPS bead concentration from fused NIR-SARDR spectra collected at different normal incident SD distances. Average RMSECV and RMSEP were obtained from rdCV models.

PLS Repeated Double CV					
SD Distance mm	**number of blocks**	**RMSECV** % w/w CI95%	**RMSEP** % w/w CI95%	**R^2**	**LV**
0.6	1	5.1 ±0.06	5.4 ±0.14	0.90	5
0.6 / 1.2	2	5.2 ±0.05	5.8 ±0.14	0.89	6
0.6 / 1.2 / 1.8	3	5.1 ±0.05	5.5 ±0.12	0.91	8
0.6 / 1.2 / 1.8 / 2.4	4	5.6 ±0.05	5.9 ±0.12	0.89	5
0.6 / 1.2 / 1.8/ 2.4 / 3.0	5	5.4 ±0.05	5.7 ±0.11	0.90	5

Bead concentration range: 15 – 50 % w/w

The integration of normal incident fibres did not significantly influence the prediction performance of the models. The best result was achieved when fusing three distances (0.6/1.2/1.8 mm), yielding similar predictive performance than the individual SD distance model, but at a cost of three extra latent variables. Instead of five latent variables, eight were needed to describe the variance with similar coefficient of determination, around 0.90. The addition of the closer angularly incident fibre, however, yielded improved results, as shown in the Table 5.2.

When the closer angularly incident fibre (0.6 mm) was combined to the normal incident fibre combinations, the errors dropped for all of them, especially the shorter SD distances.

Table 5.2 Prediction performance of MB-PLS models for determination of polystyrene concentration from fused NIR-SARDR spectra collected at different normal incident SD distances and from a 0.6 mm angularly incident fibre. Average RMSECV and RMSEP were obtained from rdCV models.

PLS Repeated Double CV					
SD Distance mm	**nb****	**RMSECV** % w/w CI95%	**RMSEP** % w/w CI95%	**R^2**	**LV**
0.6 A* / 0.6	2	4.0 ±0.03	4.1 ±0.08	0.94	6
0.6 A / 0.6 / 1.2	3	4.6 ±0.05	4.6 ±0.11	0.95	8
0.6A / 0.6 / 1.2 / 1.8	4	4.8 ±0.05	5.0 ±0.11	0.94	9
0.6A / 0.6 until 2.4	5	5.2 ±0.05	5.7 ±0.12	0.91	5
0.6 A / 0.6 until 3.0	6	5.4 ±0.04	5.5 ±0.10	0.90	5

Bead concentration range: 15 – 50 % w/w
*SDD=0.6 mm at 30° to the normal
**nb= number of blocks

This behaviour becomes even clearer when the information from Table 5.1 and Table 5.2 is compared in the Figure 5.4, which better illustrates the impact of adding the angularly incident fibre on the prediction errors. In this figure, the blue circles represent the mean RMSEPs obtained by MB-PLS models developed with the normal incident fibre combinations, starting from the individual 0.6 mm distance, towards the fusion with the subsequent distances. The grey circles represent the errors obtained when the 0.6 mm angular incident fibre is fused to the respective combinations of normal incident fibres.

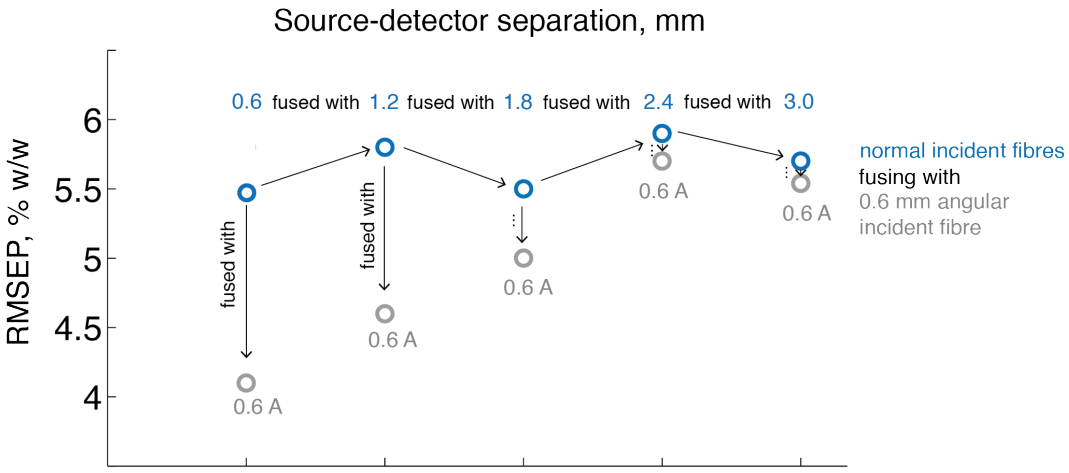

Figure 5.4 Improvement of the prediction performance of MB-PLS models when an angularly incidence fibre is fused to the normal incident combinations. Blue circles represent the average RMSEP values obtained by fusing only normal incident fibres whereas grey circles represent the inclusion of an angular incident fibre. Arrows represents the addition of blocks of spectra.

It is evident from the above illustration that the level of improvement by adding an angularly incident fibre decreases as more normal incident distances were added to the model. Other fibre combinations were also analysed, for example, by using the angular 4.2 mm distant fibre instead of the 0.6 mm one, but did not bring any improvement to the results. Those results are not shown here for sake of clarity.

In the best result, by using both normal and angular incident 0.6 mm fibres, the error dropped from 5.4 to 4.1 % by weight. Since the individual models for these fibres yielded an error of 5.4 and 5.6 % w/w, respectively (Table 5.1), the integration of information from these two fibres improved the predictive capability of the model. This is even more evident when looking at the residuals calculated for both models, as shown in Figure 5.5. The blue line represents the distribution from all 5600 residuals, while the grey lines represent the distribution for each of the 100 repetitions of the double cross validation.

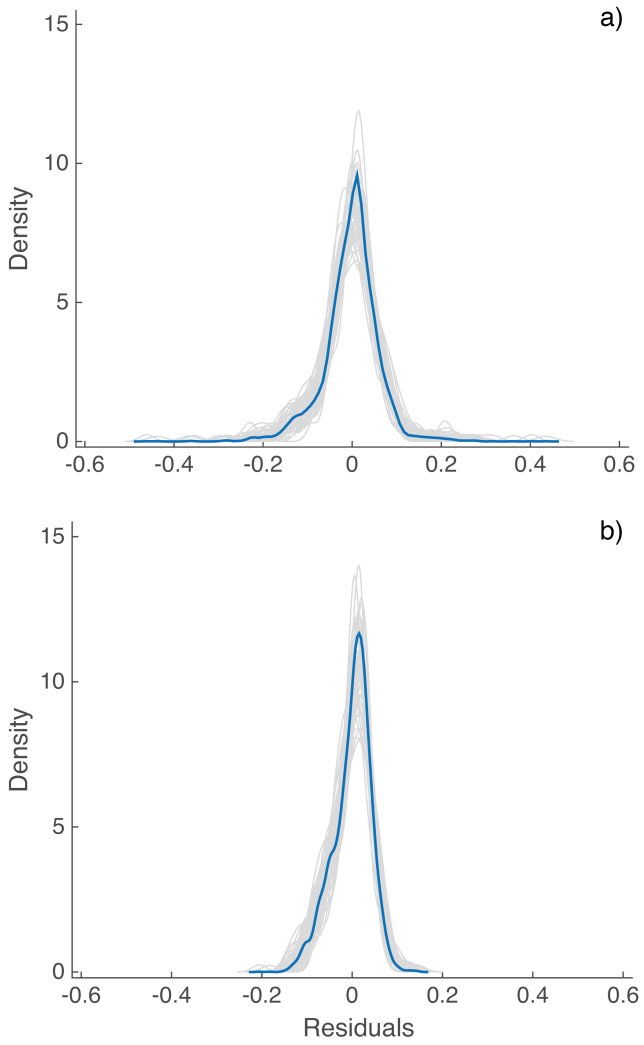

Figure 5.5 Probability density functions calculated from 100 repetitions of double cross validation models based on a) 0.6 mm distance fibre and b) when adding an angularly incident fibre (0.6mm). The blue curve represents the mean residuals while the grey curves represents the 5600 residuals from all 100 repetitions.

The residuals for the models built by using the individual 0.6 mm distant fibre clearly have a higher prediction uncertainty, shown by the wider distribution, when compared to the one observed when the angularly incident fibre is added to the model. Therefore, the data fusion strategy was able to improve the predictive power of the NIR models, providing better accuracy and stable estimations of polystyrene concentration.

5.3.2 Visible-SARDR data fusion

Visible models built with individual fibres were much better at longer SD distances. For this reason, in contrast to the methodology employed for NIR data fusion models, firstly a model was built with the longest normal incident distance (3.0 mm), and subsequently decreased distances were fused together until all normal incident fibres were contemplated by the fused model. Afterwards, the best performing angularly incident fibre was added to each of the fused models to evaluate the impact of detecting photons that were emitted at a different angular orientation. The results obtained when fusing only normal incident fibres is shown in the Table 5.3 below.

Table 5.3 Prediction performance of MB-PLS models for determination of bead concentration from Vis-SARDR spectra collected at different SD distances. Average RMSECV and RMSEP were obtained from rdCV models.

MB-PLS Repeated Double CV					
Distance mm	**Number of blocks**	**RMSECV** % w/w ± CI95%	**RMSEP** % w/w ± CI95%	**R^2**	**LV**
3.0	1	2.7 ±0.02	2.9 ±0.05	0.96	4
fused to 2.4	2	2.7 ±0.03	3.0 ±0.06	0.98	5
and 1.8	3	2.8 ±0.03	3.0 ±0.06	0.98	11
and 1.2	4	3.1 ±0.03	3.1 ±0.07	0.98	10
and 0.6	5	2.9 ±0.03	3.0 ±0.07	0.98	12
3.0 fused to 1.8 and 0.6	3	2.5 ±0.02	2.6 ±0.05	0.98	8
3.0 fused to 0.6	2	2.6 ±0.02	2.7 ±0.06	0.98	7

Bead concentration range: 15 – 50 % w/w

The errors do not seen to improve by adding consecutively shorter distances. In fact, they are similar, even slight increased for some combinations, including their confidence intervals. Furthermore, as the number of blocks increased, the models needed more latent variables to describe the data variability. However, when increasing the spacing of fibre, for example, by using only the 3.0, 1.8 and 0.6 distances, or the 3.0 and 0.6 distances, the results improved significantly. When using the three-block model, a prediction error of 2.6 ± 0.05 mm was obtained by requiring eight latent variables to explain the data variance.

In order to investigate whether adding an angularly orientated light can improve the predictive capability of the MB-PLS models, they were updated by adding one block of spectra collected by the angularly incident fibre places 4.2 mm distant to the detector. The results are shown in Figure 5.6 a and b below, by showing the average RMSEP obtained by each combination containing or not the 4.2 mm distant angular incident fibre. In this case, the blue bars represent the prediction errors for the fusion of normal incident fibres whereas the grey bars represent the addition of the angularly incident fibre. The number of latent variables needed to describe the models is shown in Figure 5.6 b.

It can be observed that the addition of an angularly incident fibre to the combinations of normal incident fibres improved the predictive capability of all models. The errors decreased considerably, although less remarkably than the ones obtained by NIR models, where the improvement was more evident. Most of the models, however, required a higher number of latent variables by adding an extra fibre.

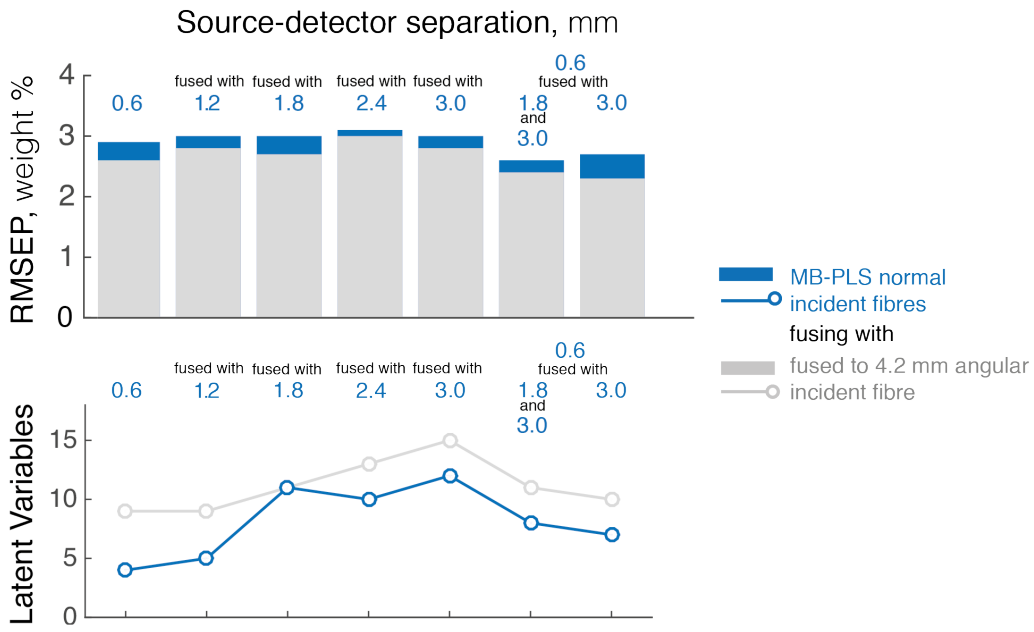

Figure 5.6 Improvement of the prediction performance of MB-PLS models when an angularly incidence fibre is fused to the normal incident combinations. The bars represent the errors obtained by fusing only normal incident fibres (blue) and when an angular incident fibre is added to the models (grey). Average RMSEP values were obtained from rdCV models. (b) number of LVs needed to describe the variability on each combination of fibres, without (blue) and with the addition (grey) of an angularly incident fibre.

When models were build with more spaced distances, i.e., 1.2 mm intervals, instead of 0.6 mm, the improvement of the models was more significant, as an example of the models built with the combinations 0.6, 1.8 and 3.0 mm, or the extreme distances (0.6 and 3.0 mm). The multiblock model developed by fusing the 0.6 and 3.0 mm distant fibre together to the 4.2 angularly incident fibre was able to provide the most accurate and stable estimations, with a mean RMSEP of 2.3% ± 0.05 and requiring ten latent variables.

5.4. Mean bead size determination

Visible spectra are more sensitive to scattering, which in its turn, is more affected by physical properties. As a consequence, models based on visible spectra tend to provide better estimates for mean bead size. In Chapter 4, we could see that this was the case, since visible models outperformed the NIR models for determining mean bead size. The overall predictive results obtained in Chapter 4 for bead size models are summarised in Figure 5.7.

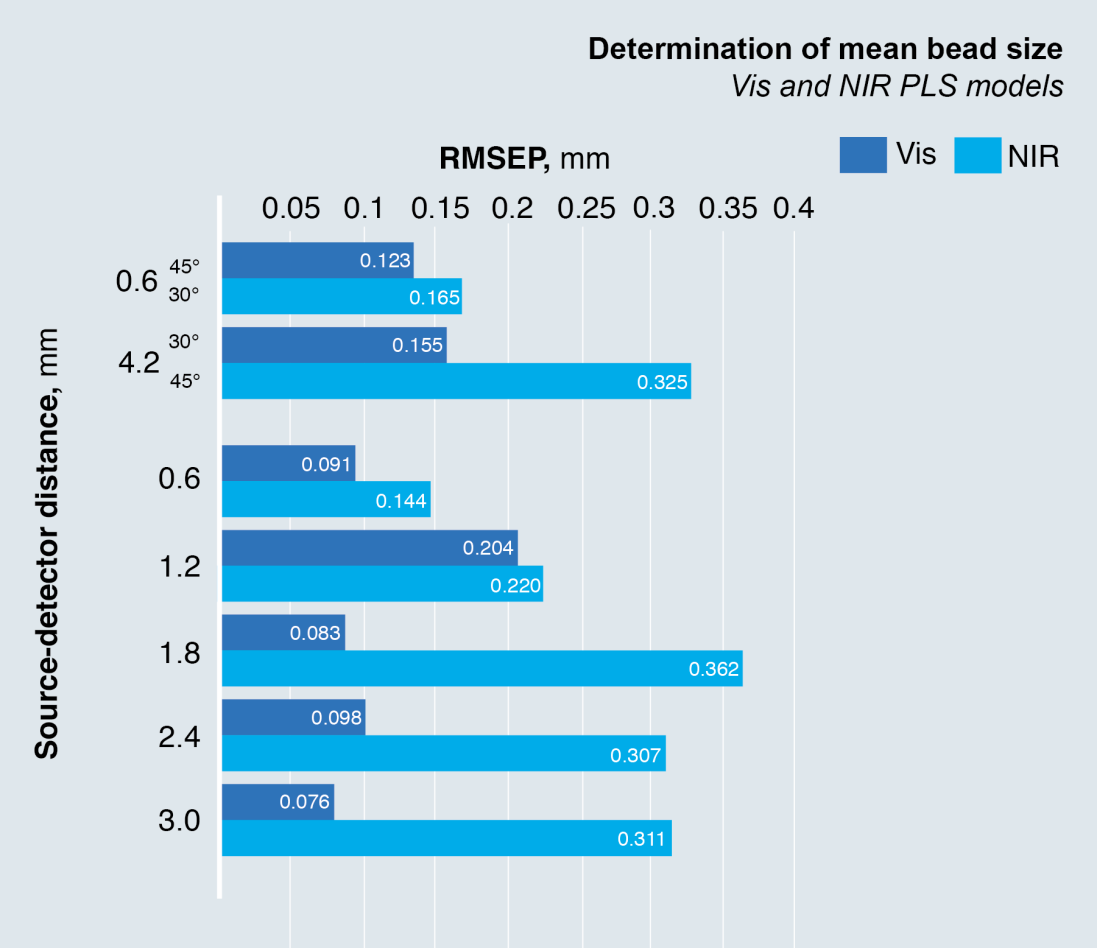

Figure 5.7 Prediction performance of rdCV PLS models developed by using Visible and NIR spectra at different SD distances to estimate mean bead diameter. The first two results are from angularly incident fibres whereas the remaining from normal incident fibres.

It can be seen that the best result obtained by Visible spectroscopy was achieved by the 3.0 mm distant fibre, yielding an error of 0.076 mm, while the best estimates for NIR were obtained by the closest of its fibres, with an error of 0.144 mm.

Although the quality of the NIR spectra from some SD distances was rather low, it would be interesting to see whether they may still hold complementary information that could improve the results if combined with each other.

5.4.1 NIR-SARDR data fusion

Similar to the bead concentration models, MB-PLS models were first developed by adding one distance at a time, starting from the shortest distance, which has also resulted in the best performance among the individual distance models. Table 5.4 below presents a summary of the statistics obtained from the MB-PLS models developed for the fibre combinations.

Table 5.4 Prediction performance of MB-PLS models for determination of mean EPS bead diameter from fused NIR-SARDR spectra collected at different normal incident SD distances. Mean RMSECV and RMSEP values were obtained from rdCV models.

		MB-PLS Repeated Double CV			
SD distances mm	**Number of blocks**	**RMSECV** mm (± CI95%)	**RMSEP** mm (± CI95%)	R^2	LV
0.6	1	0.120 ±0.002	0.144 ±0.006	0.95	7
fused with 1.2	2	0.111 ±0.002	0.114 ±0.003	0.98	11
and 1.8	3	0.129 ±0.002	0.134 ±0.003	0.97	10
and 2.4	4	0.127 ±0.001	0.125 ±0.003	0.98	11
and 3.0	5	0.143 ±0.001	0.146 ±0.004	0.97	10

Mean bead diameter range: 0.405 to 1.540 mm

Table 5.4 presents the mean cross validation and prediction errors decreasing after fusing the two closest distances from the detecting fibre, 0.6 and 1.2 mm. When more distances are added to the fused model, the improvement of prediction error becomes lesser and the cross validation error increases. The best result would then be obtained by just adding one more block to the model, yielding a prediction error of 0.114 ± 0.003 mm.

Notwithstanding the decrease in errors when performing data fusion, they are still high compared to the errors produced by the visible models, although still better than the reference method. Therefore, further MB-PLS models were developed based on these same combinations, but now adding one block of angular incident spectra to the models. The closest incident fibre (0.6 mm) was chosen due to the best results obtained when compared to further away 4.2 mm fibre. Figure 5.8 below illustrates the RMSEP errors obtaining by adding or not the 0.6 mm angular information to the MB-PLS combination of SD distances.

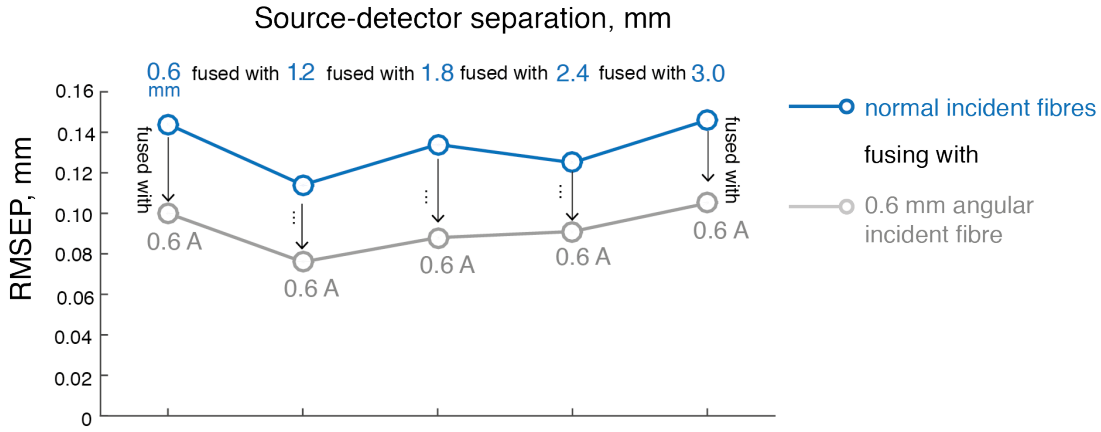

Figure 5.8 Improvement of the prediction performance of MB-PLS models for mean bead size when an angularly incident fibre is fused to the normal incident combinations. Open circles represent the errors obtained by fusing only normal incident fibres (blue circles) and when an angular incident fibre is added to the models (grey circles). Arrows represents the addition of blocks of spectra. Average RMSEP values were obtained from rdCV models.

Figure 5.8 clearly shows that adding information from angular incident light to all MB-PLS models improved their predictive capability. In the best result, the average RMSEP dramatically dropped from 0.114 ± 0.003 mm to only 0.076 ± 0.002 mm, possible with a parsimonious model by using only three blocks comprised of both 0.6 mm distances (normal and angular orientated) and the 1.2 mm distant fibre.

The addition of more blocks deteriorated the models, but still provided better results when compared to the MB-PLS models built only using normal incident fibres.

It is worth mentioning that the addition of an extra block did not increase the complexity of the model for most of the combinations. For the best result mentioned above, the model complexity slightly decreased, since the model only needed 10 latent variables, instead of nine. These results as well as the full summary of statistics are show in Table A1 in the Appendix.

5.4.2 Vis-SARDR data fusion

The models developed for mean bead diameter with visible SARDR spectra followed the same procedure adopted by the ones built for polystyrene concentration. It started by modelling the longest normal incident distance (3.0 mm), and subsequently adding decreased distances until all normal incident fibres were fused together. Table 5.5 below shows the results obtained when fusing normal incident fibres.

Table 5.5 Prediction performance of MB-PLS models for determination of mean EPS bead diameter from fused Vis-SARDR spectra collected at different normal incident SD distances. Mean RMSECV and RMSEP were obtained from 100 rdCV models.

PLS Repeated Double CV

Distance mm	**number of blocks**	**RMSECV** mm ± CI95%	**RMSEP** mm ± CI95%	**R^2**	**LV**
3.0	1	0.077 ±0.001	0.076 ±0.002	0.97	4
3.0 / 2.4	2	0.074 ±0.001	0.079 ±0.002	0.98	6
3.0 / 2.4 / 1.8	3	0.076 ±0.001	0.076 ±0.002	0.99	8
3.0 / 2.4 / 1.8 / 1.2	4	0.086 ±0.001	0.079 ±0.003	0.99	11
3.0 / 2.4 / 1.8 / 1.2 / 0.6	5	0.050 ±0.001	0.048 ±0.001	0.99	11
3.0 / 1.8 / 0.6	3	0.048 ±0.001	0.046 ±0.001	0.99	10
3.0 / 0.6	2	0.057 ±0.001	0.057 ±0.001	0.99	9

Mean bead diameter range: 0.405 to 1.540 mm

Here, the performance of fused models for the three longest distances showed no significant improvement over what had been achieved by the individual 0.6 mm fibre model. However, lower errors were obtained when all normal incident fibres were integrated together, dropping considerably from 0.076 ± 0.001 to 0.048 ± 0.001. This same level of errors, but with a more parsimonious model, was obtained when using

a fibre-to-fibre distance of 1.2 mm, by fusing the shortest, middle and largest distances, 0.6, 1.8 and 3.0 mm. Figure 5.9 a and b shows the measured versus estimated mean bead diameter regression plots for the models developed by using the a) individual fibre which presented the best results (3.0 mm) and the b) MB-PLS fused model.

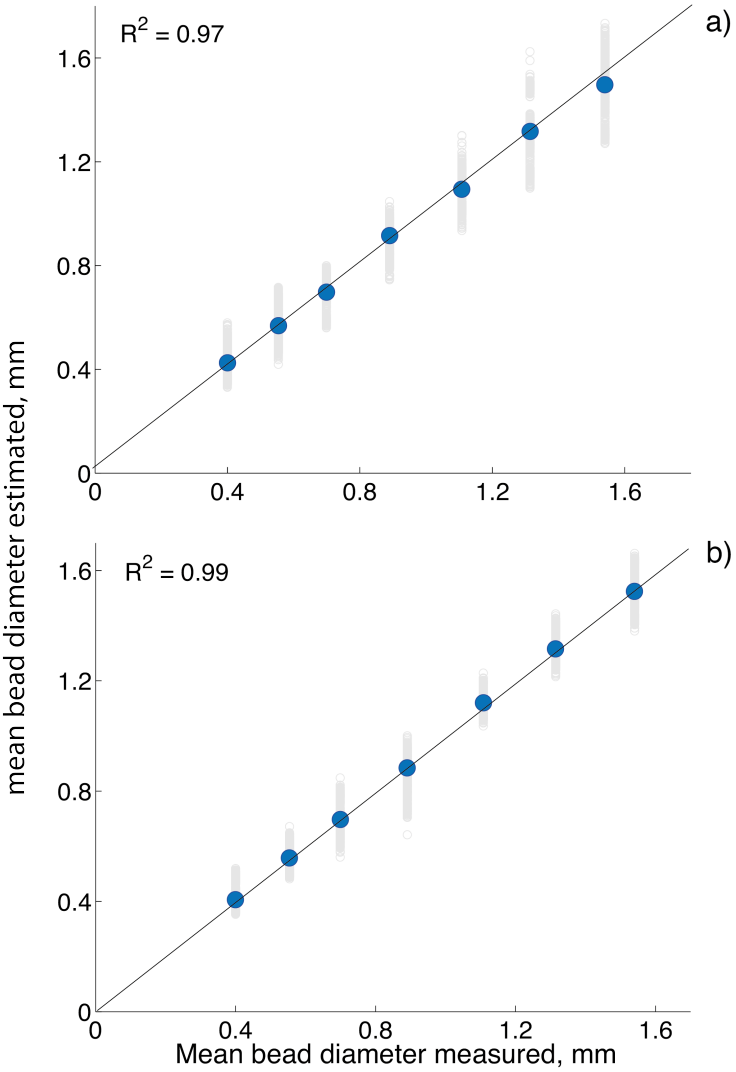

Figure 5.9 Measured *versus* estimated mean EPS bead diameter for rdCV models developed for the (a) fibre 3.0 mm distant from the detector and the (b) MB-PLS models developed by fusing three fibres, 0.6, 1.8 and 3.0 mm. The blue circles represent the mean estimation from a total of 100 model repetitions, while the grey circles represents all the estimations.

Despite the good model developed by using the individual fibre place at 3.0 mm distance for the detector, the fused model clearly provided much better results. Its measured versus estimated values shown in Figure 5.9 b shows a much better fit with a R^2 of 0.99, and a mean error of 0.048 ± 0.001 mm.

Contrary to what happened for the polystyrene concentration models, there was no gain in adding the angularly orientated fibre to the fused models, considering the parsimony of the models and magnitude of errors obtained. The results are not included here, but can be seen in the Table A2 in the Appendix.

5.5. Summary of results

Overall, for both polymer quality parameters studied herein, the MB-PLS data fusion strategy provided similar or lower errors of calibration and prediction when compared to the PLS models developed on individual source-to-detector distances.

For the determination of mean bead size, the modelling strategies allowed prediction errors lower than 0.1 mm to be reached. The improvements obtained by the data fusion strategy were more pronounced for the NIR spectroscopy models, which lead to mean bead size errors dropping by almost half, from 0.144 mm to 0.076 mm, when an angularly orientated fibre was added to the shortest SD distance. Thus, only two fibres were required, and only one extra latent variable was needed to describe the variance in the data. Although the improvement obtained by Visible models was less remarkably than the NIR ones, errors decreased from 0.076 to 0.046 mm, by needing only three blocks of normal incident fibres.

Likewise, for the determination of polystyrene bead concentration, Visible models were also better than the NIR ones, improving estimations from 2.9% to 2.6%, whereas in the NIR the RMSEP decreased from 5.4% to 4.1 %.

It becomes clear that the stronger scattering present in the Visible range contains important information related to the number of beads and their size, which also carried information about the concentration.

5.6. Conclusion and Outlook

The goal of this chapter was to investigate the impact of fusing information from different source-detector distances on the predictive capability of Visible and Near-infrared SARDR models to estimate composition and size of expandable polystyrene beads. Multiblock PLS was applied to several combinations of SD distances, and the repeated double cross validation was used to select the number of latent variables for each model and to estimate the variability of results. The controlled experiment used in Chapter 4 was again evaluated here, in order to better isolate the impact of the modelling strategies evaluated here.

Overall, when fusing normal incident fibres, all models provided good results. However, the biggest improvement happened when an angularly incident fibre was added to the fused models. The information captured for the angularly incident fibre improved the sensibility of the techniques and resulted in lower errors and better predictive power, especially for NIR models. The only exception was the bead size models built with Visible spectra, where the addition of the angularly incident fibre did not surpass the results obtained only with fused normal incident ones, but were still better than using only the individual fibre. In general, although the data fusion resulted in remarkable improvements for NIR data, the best models were the ones developed based on Visible spectra.

Visible and NIR spatially and angularly resolved spectroscopy proved to be a reliable technique for the estimation of polymer critical quality attributes, and have a great potential for online monitoring and control of polymerisation processes.

5.7. Bibliographic references

1. Llinas, J. & Hall, D. in *Handbook of Multisensor Data Fusion* **0,** 1–14 (CRC Press, 2009).

2. Jose, G. E. *et al.* Predicting mab product yields from cultivation media components, using near‐infrared and 2D‐fluorescence spectroscopies. *Biotechnol. Prog.* **27,** 1339–1346 (2011).

3. Brás, L. P., Bernardino, S. A., Lopes, J. A. & Menezes, J. C. 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 (2005).

4. Westerhuis, J. A. & Coenegracht, P. M. J. Multivariate modelling of the pharmaceutical two-step process of wet granulation and tableting with multiblock partial least squares. *J. Chemometrics* **11,** 379–392 (1997).

5. Brás, L. P., Lopes, J. A., Santos, C. R., Cardoso, J. P. & Menezes, J. C. in *European Symposium on Computer-Aided Process Engineering-14, 37th European Symposium of the Working Party on Computer-Aided Process Engineering* **18,** 601–606 (Elsevier, 2004).

6. Jin Hong, J., Zhang, J. & Morris, J. in *19th European Symposium on Computer Aided Process Engineering* **26,** 327–332 (Elsevier, 2009).

7. Westerhuis, J. A., Kourti, T. & MacGregor, J. F. Analysis of multiblock and hierarchical PCA and PLS models. *J. Chemometrics* **12,** 301–321 (1998).

8. Lopes, J. A., Menezes, J. C., Westerhuis, J. A. & Smilde, A. K. Multiblock PLS analysis of an industrial pharmaceutical process. *Biotechnol. Bioeng.* **80,** 419–427 (2002).

9. Dalle Zotte, A. *et al.* Authentication of raw and cooked freeze-dried rainbow trout (Oncorhynchus mykiss) by means of near infrared spectroscopy and data fusion. *Food Research International* **60,** 180–188 (2014).

10. Lee, H. W., Christie, A., Xu, J. & Yoon, S. Data fusion-based assessment of raw materials in mammalian cell culture. *Biotechnol. Bioeng.* **109,** 2819–2828 (2012).

11. Hertrampf, A., Sousa, R. M., Menezes, J. C. & Herdling, T. Semi-quantitative prediction of a multiple API solid dosage form with a combination of vibrational spectroscopy methods. *J Pharm Biomed Anal* **124,** 246–253 (2016).

12. Casale, M., Sinelli, N., Oliveri, P., Di Egidio, V. & Lanteri, S. Chemometrical strategies for feature selection and data compression applied to NIR and MIR spectra of extra virgin olive oils for cultivar identification. *Talanta* **80,** 1832–1837 (2010).

13. Gerard Downey, Romain Briandet, Reginald H Wilson, A. & Kemsley, E. K. Near- and Mid-Infrared Spectroscopies in Food Authentication: Coffee Varietal Identification. *J. Agric. Food Chem.* **45,** 4357–4361 (1997).

14. Workman, J. J., Jr. Quantification of LDPE [Low Density Poly(Ethylene)], Lldpe [Linear Low Density Poly(Ethylene)], and Hdpe [High Density Poly(Ethylene)] In Polymer Film Mixtures 'As Received' Using Multivariate Modeling with Data Augmentation (Data Fusion) and Infrared, Raman, and Near-Infrared Spectroscopy. *Spectroscopy Letters* **32,** 1057–1071 (1999).

15. Kourti, T. & MacGregor, J. F. Process analysis, monitoring and diagnosis, using multivariate projection methods. *Chemometrics and Intelligent Laboratory Systems* **28,** 3–21 (1995).

16. Nomikos, P. & MacGregor, J. F. Multi-way partial least squares in monitoring batch processes. *Chemometrics and Intelligent Laboratory Systems* **30,** 97–108 (1995).

17. MacGregor, J. F., Jaeckle, C., Kiparissides, C. & Koutoudi, M. Process monitoring and diagnosis by multiblock PLS methods. *AIChE Journal* **40,** 826–838 (1994).

18. Tsuta, M., Trong, N. N. D., Herremans, E., De Baerdemaeker, J. & Saeys, W. Decomposition of absorption spectra of multi-layered biological materials by spatially-resolved spectroscopy and parallel factor analysis. *Procedia Food Science* **1,** 528–535 (2011).

19. Bogomolov, A., Belikova, V., Galyanin, V., Melenteva, A. & Meyer, H. Reference-free spectroscopic determination of fat and protein in milk in the visible and near infrared region below 1000 nm using spatially resolved diffuse reflectance fiber probe. *Talanta* **167,** 563–572 (2017).

20. Kohonen, J. *et al.* Multi‐block methods in multivariate process control. *J. Chemometrics* **22,** 281–287 (2008).

21. Igne, B., Talwar, S., Feng, H., Drennen, J. K. & Anderson, C. A. Near-Infrared Spatially Resolved Spectroscopy for Tablet Quality Determination. *Journal of Pharmaceutical Sciences* **104,** 4074–4081 (2015).

22. Wangen, L. E. & Kowalski, B. R. A multiblock partial least squares algorithm for investigating complex chemical systems. *J. Chemometrics* **3,** 3–20 (1989).

23. Westerhuis, J. A. & Smilde, A. K. Deflation in multiblock PLS. *J. Chemometrics* **15,** 485–493 (2001).

24. Geladi, P. Notes on the history and nature of partial least squares (PLS) modelling. *J. Chemometrics* **2,** 231–246 (1988).

25. Kourti, T., Nomikos, P. & MacGregor, J. F. Analysis, monitoring and fault diagnosis of batch processes using multiblock and multiway PLS. *Journal of Process Control* **5,** 277–284 (1995).

26. De Groot, P. J., Swierenga, H., Postma, G. J., Melssen, W. J. & Buydens, L. M. C. Effect on the partial least-squares prediction of yarn properties combining raman and infrared measurements and applying wavelength selection. *Appl Spectrosc* **57,** 642–648 (2003).

27. Bogomolov, A., Engler, M. & Melichar, M. In‐line analysis of a fluid bed pellet coating process using a combination of near infrared and Raman spectroscopy. *J. Chemometrics* (2010). doi:10.1002/cem.1329/full

28. Li, H., Xu, Q. & Liang, Y. libPLS: An Integrated Library for Partial Least Squares Regression and Discriminant Analysis. *PeerJ PrePrints* (2014).

Chapter 6

6 Spatially and angularly resolved spectroscopy for *in-situ* monitoring of monomer conversion in suspension polymerisation reactions

In this chapter, I wish to propose a spatially and angularly resolved measurement system for *in-situ* monitoring of suspension polymerisation reactions. I propose an empirical approach for manipulating the spatially and angular resolved information based on multivariate statistical methods and chemometrics. Specifically, I am interested on the sensing abilities of each source-detector distance for both visible and near-infrared spectroscopies, and in exploring whether introducing new variations in measurement configurations could improve the predictive capability of multivariate calibration models. How well these different measurements are integrated and how accurate the multivariate models can be, are the main questions of this chapter. To the best of my knowledge, this work is the first attempt to employ spatially and angularly resolved diffuse reflectance spectroscopy on the monitoring of suspension polymerisation reactions, and also the first in which a multiblock data fusion approach will be evaluated for this purpose.

6.1 Styrene suspension polymerisation reactions

Styrene suspension polymerisation is one of the preferred methods for the production of polystyrene beads, a polymer used as raw material in a wide range of applications, from thermal insulation boards in construction industry to food packaging. The reaction to produce polystyrene beads was first proposed in 1909 [1], but the market for polystyrene and expanded polystyrene is stronger as ever has been, being among the biggest commodity polymers produced in the world. Other important polymers also manufactured by suspension polymerisation include poly(vinyl chloride) (PVC)[2], poly (methyl methacrylate) (PMMA)[3], poly(vinyl acetate) (PVA), styrene–acrylonitrile copolymers and a variety of ion exchange resins of great commercial value.

In suspension polymerisation, the reaction occurs inside of monomer droplets, which are progressively transformed in sticky monomer-polymer droplets and finally into solid spherical polymer beads [1] (Figure 6.1).

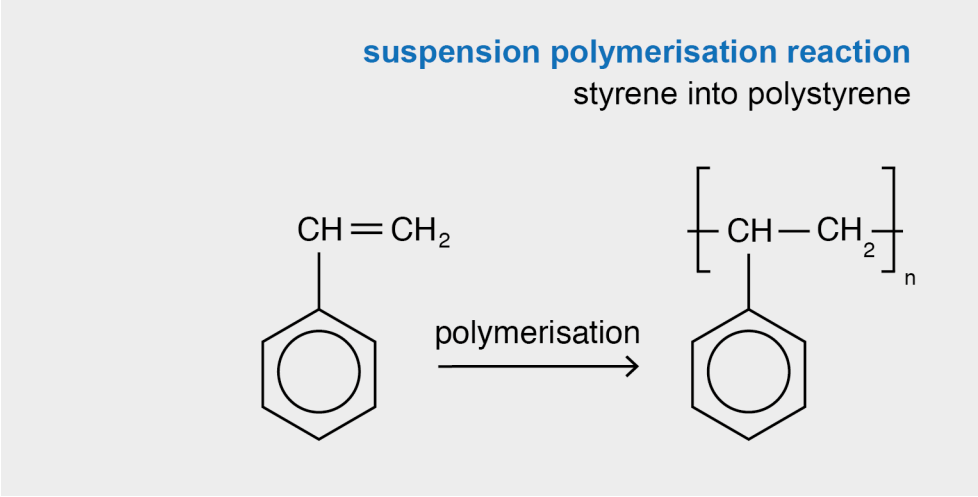

Figure 6.1 Reaction to produce the polymer polystyrene from the monomer styrene.

As the initiator is soluble in styrene, which in its turn is insoluble in water, droplets of monomer containing the initiator are formed as soon as suspended in water, the continuous phase throughout the reaction. The system is continuously dispersed by agitation and stabilised by the addition of an insoluble stabiliser [4].

Droplet stability is very important in a suspension polymerisation. As the suspended droplets collide with each other at fast speeds, they might be fused (coalesce) or divided (break-up), forming bigger or smaller droplets [5]. As the reaction progresses and the droplets viscosity increases, coalescence events become more probable than break-up ones, which may lead to coagulation if the stability is not properly controlled, especially between 25 and 75 % of conversion. The addition of the stabiliser acts by creating an interfacial layer around the droplets, reducing the rate of coalescence. After this stickier phase, breakage and coalescence are ceased, the Particle Point of Identification (PIP) [6] is reached and the droplets become solid spherical beads.

One of the main advantages of suspension polymerisation is the good thermal control[1]. As each droplet acts as polymerisation reactor, the internal viscosity of the droplets increases along the reaction, but the overall viscosity of the suspension remains low, allowing a good heat transfer. Moreover, higher conversions of monomer into polymer are achieved inside the droplets [7], and fast reaction times are obtained.

Suspension polymerisation normally produces large beads, in the range of a few micrometres to a few millimetres, although 0.2 to 2 mm is the predominant diameter for most applications. Specific bead size ranges are produced based on the intended application by manipulating several parameters like stirring speed, reactor and impeller design, the monomer-water ratio, and concentration of stabiliser and initiator [8], although the stirrer speed has the biggest impact on the bead size[9].

6.1.1 On-line monitoring of polymer quality

As demands for polystyrene beads rise, the need for highly effective and robust monitoring tools for polymerisation processes increases. Monitoring conversion of styrene to polystyrene is highly relevant because it is directly related to productivity. Moreover, new worldwide market requirements are being put in place for tighter product specifications and stricter safety and environmental constraints[10], especially concerned in limiting the residual monomer in the final product[11].

Considerable progress has been made in on-line instrumentation since Kiparissides and Morris [12] considered it to be the weakest link in polymer reactor control. Fibre optical technologies have been developed and applied to different spectroscopies for on-line monitoring of conversion since the 90's, as example of UV-VIS[13,14], NIR[15], MIR[16] and Raman[17] spectroscopies, especially because they allow non-invasive analytical measurements to be made real-time, a condition that is central to implementations of Process Analytical Technology (PAT)[18] and Process Intensification (PI).

NIR spectroscopy has been successfully used for real-time monomer conversion in several different polymerisation processes. Applications were developed for emulsion copolymerisation and polymerisation reactions[19,20], bulk[21] and solution polymerisation of methyl methacrylate[22], polyurethane step-growth polymerization[23], and ethane/1-hexene solution copolymerization[24].

Despite the widespread use of visible and near-infrared spectroscopy in polymerisation processes, significant challenges remain in its successful use for monitor and control of suspension polymerisation reactions, which may explain the shortage of applications for this specific reaction. Obtaining analytical measurements from suspension polymerisation media is difficult especially due to its high heterogeneity and viscosity, which turns sampling a challenging task and makes

reference and spectroscopic measurements less accurate. Furthermore, as the media is agitated, monomers are continually dispersed in small spherical droplets, which scatter light in different extensions at each reaction phase, adversely affecting spectroscopic measurements.

In fact, light scattering heavily affects optical based measurements in particulate media such as polymerisation reactions. Visually, it manifests in a spectrum by a non-linear variation on its intensity, shape and baseline, which can compromise the estimation of chemical and physical properties based on the spectrum. Those non-linear variations are normally corrected or at least reduced by using empirical scatter correction methods, like Standard Normal Variate (SNV)[25], Multiplicative Scatter Correction (MSC)[26] and its extended version (EMSC)[27], which are among the most used ones.

Spatially resolved diffuse reflectance measurements have been used for extracting the scattering effects from the spectra. The technique allows probing particulate media at several source-detector separations, accessing different penetration depths and scattering/absorption interactions. Such information, if efficiently extracted from the complex interplay of absorption and scattering effects, may enhance our knowledge about the process. However, despite its advantages over the classical approach for Vis and NIR spectroscopy, application in the literature on spatially resolved spectroscopy for the monitoring of polymerisation reactions could not be found. One of the reasons may be the difficulty faced when modelling the optical properties or inverting the measurements to extract them. These procedures involve complex mathematical operations[28] and intensive computational calculations[29], unsuitable for real-time applications[30].

In this chapter, the suitability of the spatially and angular resolved measurements for in-situ monitoring of suspension polymerisation reactions is investigated. Instead of the usual approach employed in Spatially Resolved Spectroscopy (SRS) data, in which the absorption and scattering spectra are extracted from multiple measurements[31] by invoking the computationally intensive radiative transfer theory[29], I propose an empirical approach for manipulating the spatially and angular resolved information based on multivariate statistical methods and chemometrics[32]. Specifically, I am interested in the sensing abilities of each SD distance for both visible and near-infrared spectroscopies, and in exploring whether introducing new variations in measurement configurations could improve the predictive capability of multivariate calibration models. To this aim, I will evaluate combinations of fibres placed at different distances from the detector, and at different angles of incidence. How well these different measurements will be integrated and how accurate the resultant multivariate calibration models will be are also of great interest in this chapter.

To the best of my knowledge, this work is the first attempt to employ spatially and angularly resolved diffuse reflectance spectroscopy on the monitoring of suspension polymerisation reactions, and also the first where a multi-block data fusion approach will be evaluated for this purpose.

6.2 Process description

The suspension polymerisation process investigated in this work is a laboratory-scale batch reaction for the production of expandable polystyrene beads. Each reaction lasts approximately seven hours, where a full conversion from styrene to polystyrene is achieved. In order to guarantee variability in the process profiles, different combinations of stirrer speed and amounts of initiator and stabiliser were employed to run eight full polymerisation batches, according to the recipes presented the Table 6.1. The amount of styrene and water were held constant for all reactions. The conditions for each parameter were varied inside the boundaries of a practical operational space designed during previous test reactions.

Table 6.1 Experimental conditions for the eight suspension polymerisation reactions.

Reaction run	**Stirrer speed,** rpm	**Initiator,** g/L	**Stabiliser,** g/L
R1	250	3.00	2.50
R6*	250	3.50	2.50
R7*	250	4.00	2.50
R5*	300	3.00	1.50
R2	300	3.00	2.50
R8*	300	3.00	3.00
R3	300	3.50	3.00
R4	300	4.00	3.00

* reactions performed after probe retrofitting.

During every reaction, samples were withdrawn from the reactor at every 20 minutes for characterisation by conventional gravimetric analysis and particle size measurements. Spectroscopic measurements were collected *in situ* at every 10

minutes with the probe head inside the reactor. At every two measurement cycles, the spectra acquisition was done just before a sample was withdrawn for reference analysis. The experimental system is illustrated in Figure 6.2, which shows how the optical probes are connected to the spectrometers and the reaction vessel. Experimental details are outlined in the next sections and are thoroughly discussed in Chapter 3

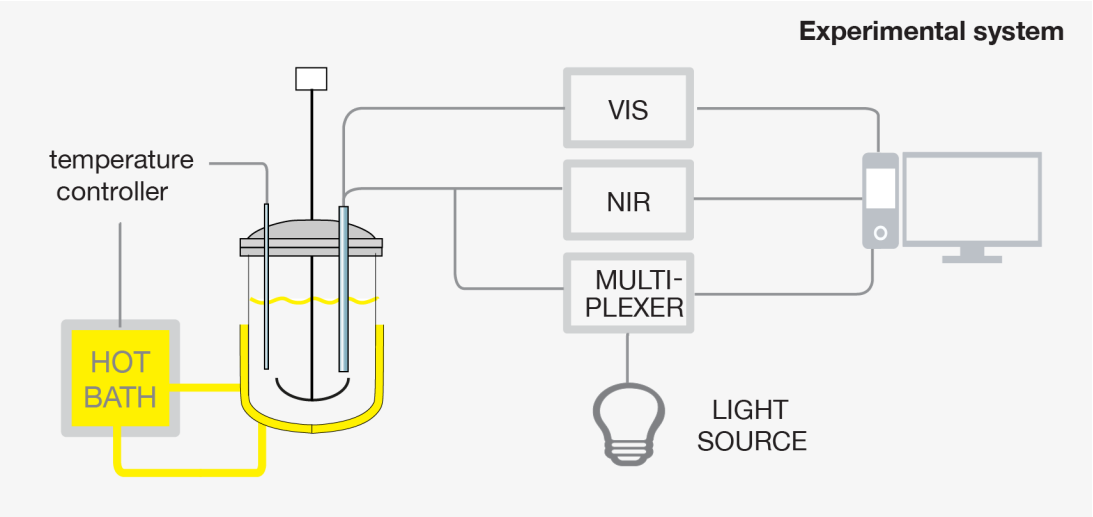

Figure 6.2 Experimental set-up for the suspension polymerisation reactions

6.2.1 Spectral acquisition

Measurements in the near infrared region were obtained by a NIR spectrometer (NIRQuest-512, Ocean Optics) while the spectra in the Visible region were registered by a visible range spectrometer (USB-4000, Ocean Optics. For interfacing the spectrometers with the reaction vessel, a prototype spatially and angularly resolved diffuse reflectance optical probe was directly immersed in the reaction media for *in situ* measurements. The probe head is illustrated in the Figure 6.3 below, which also shows the fibre optical disposition and distances.

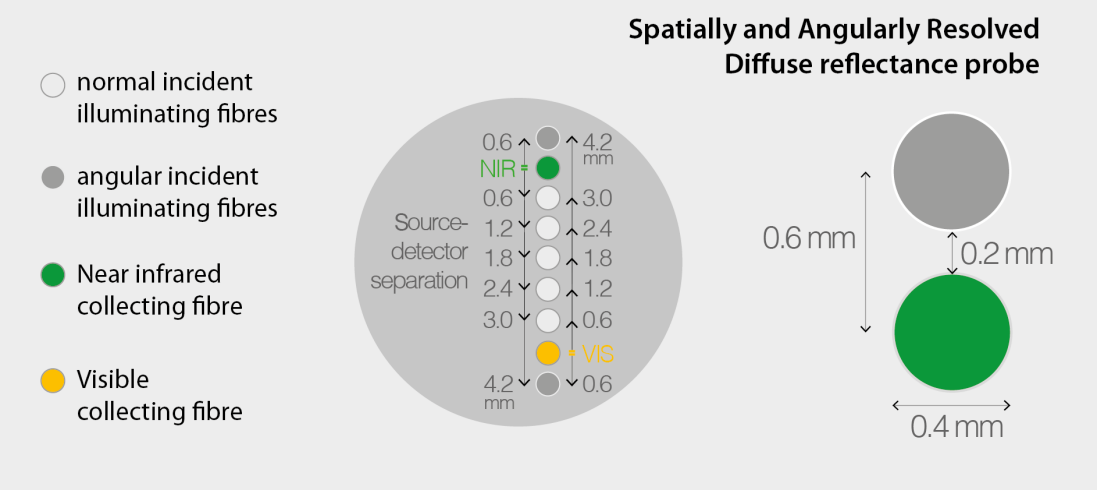

Figure 6.3 Schematic illustration of the SD distances in the SARDR probe head. The distances given are from Vis and NIR collecting fibres (detector) to each illuminating fibre (source).

6.2.2 Gravimetric Analysis

The styrene-polystyrene conversion was determined by gravimetric analysis, which is the technique routinely used throughout the polymer industry. Samples were taken from the reaction vessel at every 20 minutes, by using an automatic pipette with long length pipette tips. After weighing the extracted sample, 5mL of a solution of toluene containing 1% of the inhibitor *para*-benzoquinone was added to the sample for reaction quenching, and was then placed in a vacuum oven until a constant weight was obtained. The dry mass was weighed and the conversion determined. The use of *p*-benzoquinone allows faster diffusion into the beads as it is more hydrophobic than hydroquinone; therefore, faster quenching of the reaction in the sample is possible and a more accurate estimation can be achieved.

6.3 Data manipulation and model development

6.3.1 Data fusion methods

A large amount of data is produced when using spectroscopic techniques to monitor batch processes. The batch data itself is a three-dimensional data array containing the intensities at different wavelengths across both reactions and time dimensions. On top of that, unlike standard spectroscopic techniques that provide one spectrum per sample, in the spatially resolved data an extra dimension is added by having several spectra for the different source-detectors separations and angular incidences. A proper manipulation of such amount of data is fundamental for the successful viability of the technique.

Several ways of handling multi-dimensional data are available for quantitative analysis. Usually, SRS spectra from different fibres are subjected to co-adding, and the final averaged spectra used for regression purposes[33,34]. Data augmentation or unfolding is another way of handling this type of data. The advantage of using the unfolded and the co-adding approaches is the simplicity. In fact, handling all data blocks in a single matrix allows the use of standard Partial Least Squares Regression (PLSR) for model building. However, the influence of each block can be hidden by other block's influence[35], which means that spatial information may be flattened, causing loss of block related knowledge.

In Chapter 5, a data fusion strategy based on multi-block was employed giving improved results when compared with the models built from individual fibres. Here, besides the multi-block method, the co-adding approach will also be investigated.

6.3.2 Spectra pre-processing

Vis and NIR SARDR spectra were acquired from each one of the 7 collecting fibres. Before analysis, data were subjected to spectral correction, normalisation with the integrating sphere spectrum, smothing by wavelet transform, conversion from reflectance to logarithmic units, and autoscaling of variables.

To remove baseline distortions among the different batches, as well as scattering effects, a few empirical pre-processing techniques were investigated, namely SNV, SNV-DT[25], MSC and EMSC.

6.3.3 Model development and validation

Similar to the previous chapters, the repeated double cross-validation (rdCV)[36] was employed for both model building and validation. This approach was chosen because repeated instances of the double cross validation allow a reliable comparison of the predictive capability and parsimony of models. In the scope of this chapter this property is highly relevant, especially for comparing the output of the several different combinations of fibres, spectroscopy modes, and model parameters.

The validation in the rdCV approach is performed on test set samples not used in the calibration phase, ensuring the estimations are not biased. However, as we are dealing with process data in this chapter, the test set samples may not be completely independent from the calibration set, since they are withdrawn from the same reaction runs used in the model building phase, and consequently, make part of the same population. Hence, to increase the robustness of the models developed here, and to guarantee that estimations based on such process data are not biased by reaction or sample information, an external validation step based on full independent reactions, not individual samples, was performed. In this particular case, the

calibration set was extended to incorporate all samples from the rdCV sample sets (calibration, validation and test sets). The updated calibration set was then used to estimate the conversion from the independent full reactions, based on the number of latent variables chosen by the rdCV approach. A schematic illustration of the model calibration and validation procedure used here is shown in the Figure 6.4 below.

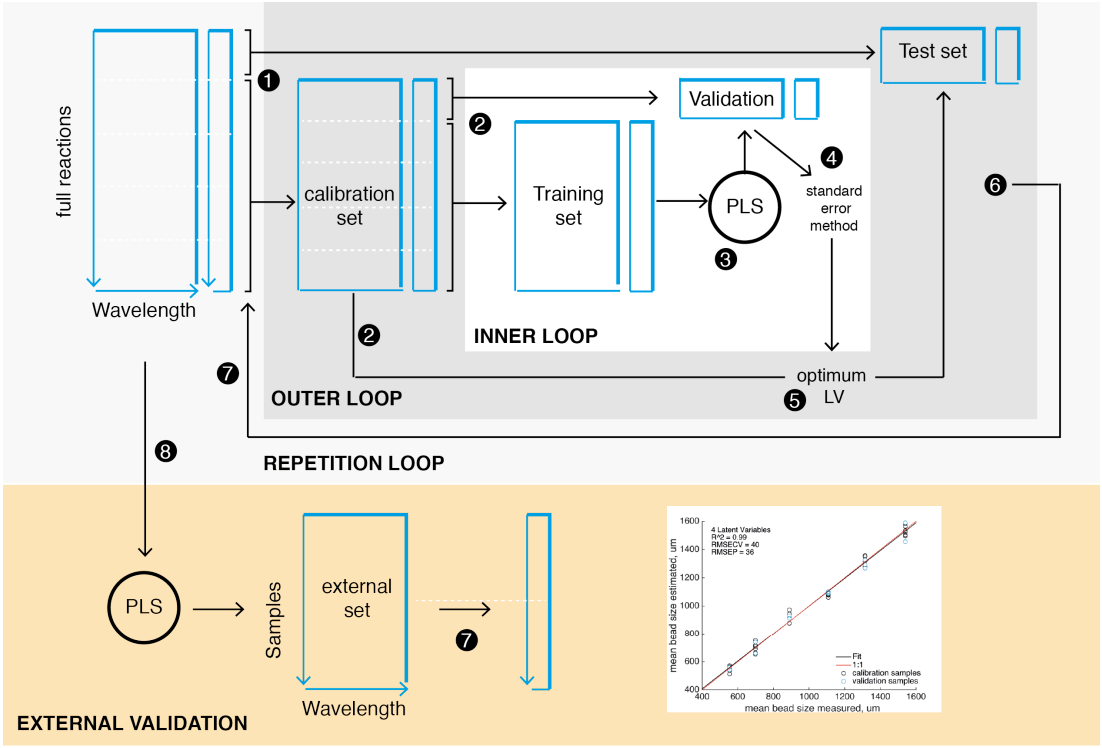

Figure 6.4 Schematic illustration of the rdCV technique.

The predictive ability in each of the calibration and validation phases is assessed by the root mean squared error of cross validation (RMSECV), root mean square error of prediction (RMSEP) and their distributions, squared coefficient of correlation (R^2), optimum number of PLS components and its frequency plot, and the residuals for all estimations. The confidence intervals (CI 95%) are calculated for RMSECV and RMSEP based on their values obtained from all estimations from 100 Test sets.

6.4 Number of segments and Latent Variables selection

In both model building and validation by using the rdCV, the number of sample segments in the outer and inner loops is important since it affects the size of the sample sets, and the number of values from which the statistics are calculated. In this work, segment sizes from 4 to 10 were evaluated in their influence on the RMSECV and in the number of latent variables used to build the models.

In the same experiment, the criterion for selecting the number of latent variables was also investigated. In the original repeated double cross validation, the optimum number of latent variables is chosen in the inner loop by using the one-standard-error rule, where the most parsimonious model is selected within one standard error of the minimum Mean Square Error (MSE), obtained from cross validation within the calibration set. In preliminary work on the suspension polymerisation reaction data, it was found that this approach for latent variable selection might be too conservative, selecting a low number of latent variables, but resulting in high prediction errors, which might be an indication of under fitting. An alternative would be to select a number of latent variables that resulted in the minimum mean square error, which can be sometimes too optimistic, over fitting the results. Another alternative is the Haaland Thomas[37] method, which consists in selecting the model with the fewest number of latent variables such that the Prediction Error Sum of Squares (PRESS) for that model is not significant greater than the PRESS for the model with the minimum PRESS.

In this work, the standard error rule was evaluated by using five different parsimony factors, namely 1.00, 0.75, 0.50, 0.25 standard errors. The results were compared to the Haaland Thomas method and to the global minimum MSE criterion (which is equivalent of using a parsimony factor of zero standard error). The evaluation was

performed on models developed based on the simplest fibre combination comprising the closest normal and angularly incident fibres (0.6 mm).

Table 6.2 **Selecting model complexity.** Optimum number of latent variables for the PLS models developed by varying the number of sample segments in the cross-validation and parsimony factors for the MSE and Haaland Thomas criteria. RMSEP values are highlighted in green.

PLS Repeated Double CV					
Segments/ parsimony factor	**zero**	**0.25**	**0.50**	**1.00**	**HT**
	RMSECV/ RMSEP, %				
4	5.1/6.3	5.9/7.5	6.6/7.5	7.8/7.3	5.6/5.9
7	4.8/6.3	5.6/7.1	6.2/7.5	7.3/5.7	5.3/5.7
10	4.8/6.3	5.5/5.7	6.1/7.5	7.1/5.7	5.2/5.5
	Number of latent variables				
4	20	11	10	7	12
7	20	11	10	8	13
10	20	12	10	8	14

Model complexity in the standard error method is regulated by the parsimony factor. As can be seen in Table 6.2, by increasing the parsimony factor from zero to 1.00 SD, the number of latent variables naturally decreases, since the selection becomes more conservative. As a consequence, the RMSECV tends to increase. However, this change in model complexity does not seem to have impacted the estimation of monomer conversion on external reactions, since the RMSEP maintained stable at around 7.5 %, regardless of the parsimony factor, as it is the case when using 4 segments. This finding may indicate the models built with low parsimony factor were overfitted. When using 7 or 10 segments, a contrary effect was found. The

RMSECV errors were higher in a few cases, potentially indicating under fitting. The best case was obtained by using 10 segments and 0.25 SD, resulting in similar predictive capability in both cross validation and external predictions.

The best overall results, in terms of model predictive capability and complexity, were obtained when applying the Haaland Thomas criteria. A reasonable choice for the number of latent variables was reached, presenting the lowest RMSECV values. By testing the model against external reactions, the RMSEP obtained were similar to the calibration errors, which may indicate a correctly assigned model complexity, avoiding under or overfitting. Furthermore, the Haaland Thomas criterion was able to select a reasonable number of latent variables, closer to the ones obtained by the standard error rule using a reduced parsimony factor of 0.25 standard deviation. This indicates that the original one standard error method is too conservative for the data in question, and that the HT method allows a proper selection without requiring additional heuristics to find the best parsimony factor.

In the context of cross-validation, higher the number of segments, higher the numbers of models to be developed, since cross validation will be iterating more times. Because the number of segments has not significantly affected the results, and taking in consideration the higher computation time needed for running the models with higher number of segments, it was decided to use the Haaland Thomas criteria and only four segments throughout this chapter.

6.5 Spectral evolution along the reaction

In chapters 4 and 5 it was demonstrated that polystyrene concentration could be estimated with good accuracy by NIR spectroscopy. In those experiments, the system used was well controlled and made of only water and polystyrene beads, facilitating the interpretation and analysis of the results. In suspension polymerisation reactions, on the other hand, the media is much more heterogeneous, which brings new challenges for the statistical methods and the spectroscopic technique itself. To evaluate whether the SARDR probe can sense styrene consumption and polystyrene formation with satisfactory accuracy, despite the heterogeneity and disturbances found in the reaction system, quantitative models were developed for the online monitoring of styrene conversion into polystyrene. Before discussing the models, an overview of the NIR spectral behaviour throughout the course of full suspension polymerisation reactions is presented.

6.5.1 Near-infrared spectral evolution along the reaction

Both styrene and polystyrene have important absorption bands in the NIR region, mostly arising from aromatic CH overtones. In the past, their overlapping nature hindered the development of on-line NIR spectroscopic monitoring of polymerisation reactions, since a clear separation of the absorbance bands was necessary [38]. Nowadays, advanced chemometrics methods for spectral pre-processing and calibration are able to tackle this problem efficiently. Figure 6.5 below shows an NIR spectrum taken at a particular time from a suspension polymerisation reaction.

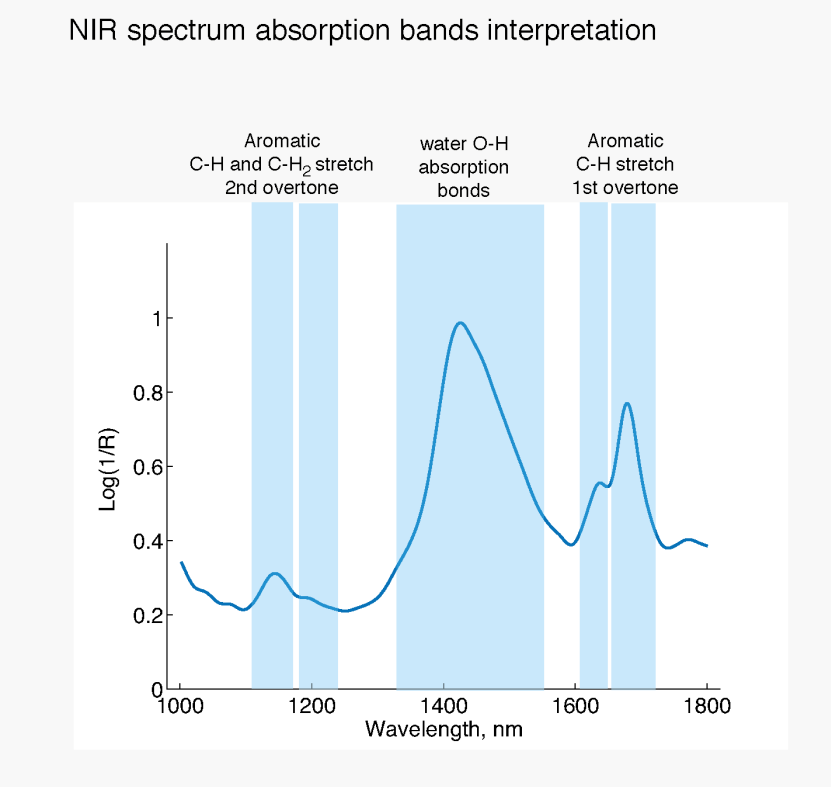

Figure 6.5 Near-infrared spectrum taken from a suspension polymerisation reaction.

In the NIR spectrum from Figure 6.5, the broad absorptions due to the O-H absorption from water[39] can clearly be identified around 1440 nm, dominating the spectrum. The regions of interest for styrene-polystyrene conversion are the absorption bands around 1200 nm, due to the aromatic CH and CH_2 stretch of the second overtone[40] and around 1640 nm, due to the aromatic CH and stretch of the first overtone[41]. The latter one presents a much stronger absorption in the collected spectrum, being able to clearly indicate qualitatively the conversion of monomer into polymer. In order to better illustrate this behaviour, the evolution of a complete seven hours long suspension polymerisation is shown in the Figure 6.6 a. In this case, the first derivative of the log (1/R) spectrum is shown for easier identification of the bands. Moreover, only the spectral region from 1580 to 1740 nm is presented, which contains the stronger bands of interest.

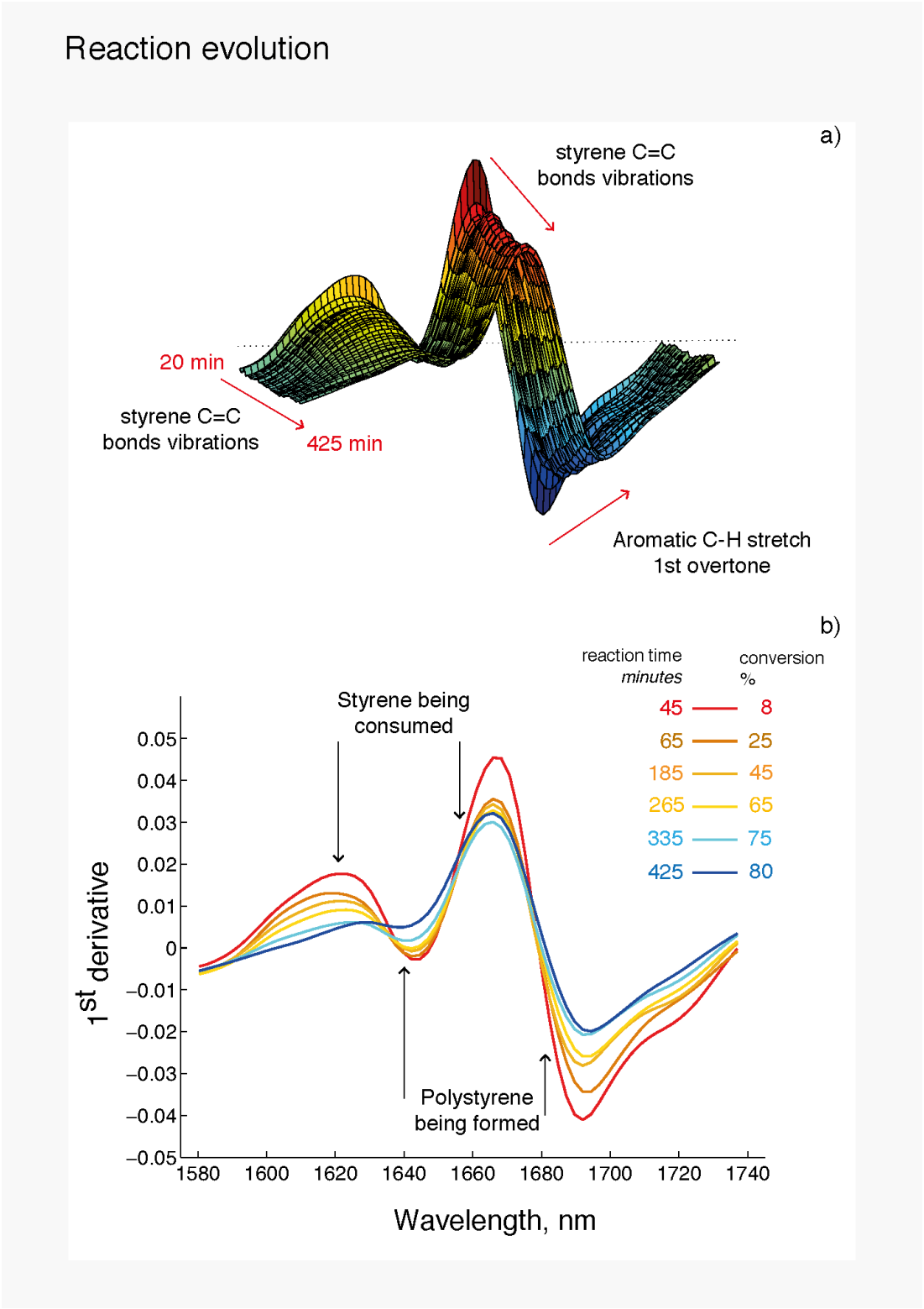

Figure 6.6 Evolution of the first derivative of near-infrared spectra collected throughout a suspension polymerisation reaction. Spectra collected with the illuminating fibre place at 0.6 mm from the collected fibre. Only the region between 1580 and 1740 nm is shown for clarity.

As the reaction progresses, the styrene is being continuously converted in polystyrene, having 8% conversion at 45 minutes and reaching 80% at 425 minutes, when the reaction finishes. It is very clear that the intensity of the absorption at the styrene C-H vibration bonds vanishes through out the reaction, at the same time that the aromatic C-H stretch in the first overtone has their absorption increasing continuously. The same plot is shown for a few samples of increasing conversion degree (8, 25, 45, 65, 75 and 80%) for clear visualisation of the polystyrene bands (Figure 6.6b). By comparing the first and last samples, 8% at 45 minutes and 80% at 425 minutes, respectively, it becomes clear that the intensity of the absorption bands centred around 1610 nm and 1670 nm decrease, as a result of the monomer depletion, while the bands around 1640 and 1690 nm have their absorptions increased due to the polymer production.

6.5.2 Visible spectral evolution along the reaction

Visible spectroscopy is not a natural method of choice for this application, especially due to the absence of absorption bands related to styrene and polystyrene in the visible spectrum. However, multivariate calibration models may capture information from possible indirect correlations between the concentration and size of beads, mainly through the susceptibility to scattering effects characteristic from the Visible region. By having multiple fibres measuring the flow of beads at different distances and sample depths, as it is the case of our SARDR probe, it may be possible to enhance the mining of the required information. This possibility was already presented and successfully executed in Chapters 4 and 5, where predictive models developed with Visible spectra for estimation of polystyrene concentration provided excellent results. Such models, though, were based in a simple system, consisting

only of polystyrene and water. In actual reactions, where the level of complexity is much higher, the validity of the predictions of these models remains to be proved.

A qualitative evaluation of the spectral features was first performed by means of Principal Component Analysis (PCA) and spectral inspection. For this example, spectra from six polymerisation reactions were gathered, comprising 211 samples and 355 wavelengths, ranging from 500 to 800 nm. Only the spectra collected from the farthest normal incident (3.0 mm) fibre were used, since it provided good results in the two-component system. Data from each block were mean centred before model building. Figure 6.7 below presents a PCA score plot from the two first latent variables. The score plot reveals the relationship between samples by providing a sample map for the reactions, allowing an easier identification of patterns and differences between samples or group of samples.

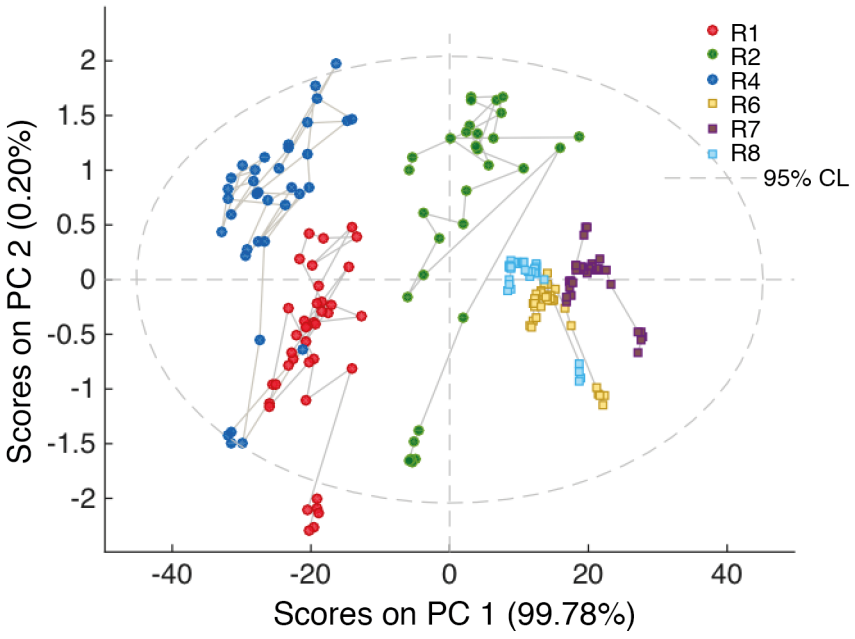

Figure 6.7 PCA score plot built with Visible spectra of 211 samples from six polymerisation reactions, represented by coloured circles and squares. Grey line connectors are plotted for sake of clarity only.

In the score plot of the Figure 6.7, the first principal component dominates the spectral information, explaining almost 100 % of the data variance. All reactions are clearly separated by the scores values in the PC1. Notwithstanding its low explained variance, PC2 is still able to describe a rough process trajectory for each reaction. The initial samples from each reaction have more negative PC2 score values, and are clearly separated from the remaining of the reaction. Those samples are from the first reaction phase, prior to the addition of stabiliser, where the droplets are in constant coalescence and breakage.

It is also evident from the score plot that the scores from samples represented by coloured circles are spread more than the scores from samples represented by coloured squares. As explained in the Experiments section, the measurement protocol changed after a few reactions were performed due to a probe retrofit. This experimental difference can be clearly seen in the score plot, as well as in the score versus reaction time plot, in the Figure 6.8. It can be found that the reactions run before (coloured circles) the probe retrofit are much more scattered and widely spread than the reactions performed after (coloured squares) the probe retrofit.

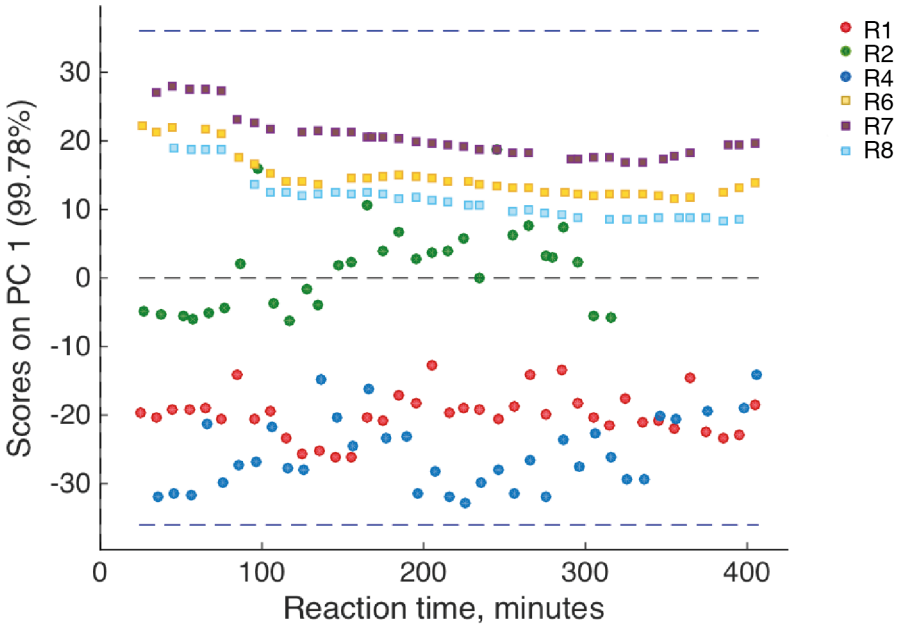

Figure 6.8 Scores versus reaction time for six polymerisation reactions. Circles and squares represent samples taken at different time from polymerisation reactions.

This behaviour can directly influence the model development and compromise the predictive capability of the resultant models if not taken in account in all phases of the validation strategy. In this experiment, this difference will work as a challenge test of the robustness of the models developed after experimental modifications.

6.6 Estimating monomer conversion on suspension polymerisation reactions

PLS-based regression models were developed based on the fused diffuse reflectance spectra taken at different source-detector separations and angular incidences. The aim was to evaluate whether spatially and angularly resolved Vis and NIR diffuse reflectance measurements could provide reliable *in situ* estimates of monomer conversion throughout suspension polymerisation reactions.

Initially, only normal incident fibres were considered. For the development of models based on different fibre combinations, two different data arrangement approaches were studied with regards to their influence on the model's predictive capability: multi-block and co-adding approaches. In the multi-block approach, data were organised in blocks of spectra, each of them containing information from one specific fibre. For the co-adding approach, a single-block data set was obtained by averaging the absorbances of each sample for more than one fibre. This approach is similar to the standard NIR or Vis spectroscopic methods, where a unique spectrum per sample is available.

For each of these approaches, PLS rdCV models were developed based on NIR and Vis spectra obtained from six full polymerisation reactions. For NIRS, the 223 samples obtained were randomly divided in calibration (192) and test sets (31), since a 4-fold cross validation was adopted based on the tests performed in Section 6.4. The calibration set was further divided in two sets for cross-validation (164 and 28 samples, respectively). For Vis models, 12 samples were discarded due to saturation in their spectra. Hence, 211 samples were randomly divided in calibration (159) and test sets (52), with the calibration set being divided in two sets for cross-validation (120 and 39 samples). In this way, predictions were made based on non-sequential samples from several different reactions. The procedure was repeated 100 times for a

reasonable evaluation of models. When building all models, every block/matrix was mean centred and divided by their standard deviation. Further normalisation for the multi-block approach was not needed since all blocks shared the same type and quality of measurements.

After rdCV calibration and validation phases, an updated calibration set containing all samples used in the rdCV (223 for NIR and 211 for Vis) was used to estimate the conversion from two independent full reactions. For this, the number of latent variables used was the ones chosen by the rdCV method.

It is worth remembering from the experimental chapter that the polymerisation reactions were run at different days and weeks and have a particular instrumental distinction. The probe needed maintenance half way through the experiment, since its fibre ends became misaligned. This unexpected event led to two different groups of reactions; the ones run before and after the probe was repaired. For the model-building phase, both reaction groups were included to the models to evaluate the robustness of the method upon small changes in the equipment, and to guarantee a higher number of samples in the modelling strategy. One reaction from each group was allocated for the external validation. Reaction R3 was run before the repair, whereas reaction R5 after it.

For the NIRS models, the fusion strategy started by integrating spectra from the two shortest source-detector distances (0.6 and 1.2 mm), to which all subsequent distances (1.8, 2.4 and 3.0 mm) were added one at the time, until a final fused model was built including all five available SD separations. Vis based models adopted the inverse order, i.e., started with the two longest SD separations (3.0 and 2.4 mm), followed by the addition of subsequent closer fibres, one at the time. The rationale behind this sequence was to guarantee that a good performing fibre was included on all fused models. It was chosen based on the findings shown in Chapter 4 on the

polystyrene-water system, where it was found that the shortest SD distance led to the best result for NIRS, whereas the longest SD distance to the lowest error among the Vis models. Such trend was confirmed on the polymerisation process data, as shown in Figure 6.9, which presents the RMSECV values obtained by rdCV PLS models developed on each individual source-detector distance, for both NIR (blue circles) and Vis (orange circles) spectra collected along six full polymerisation reactions.

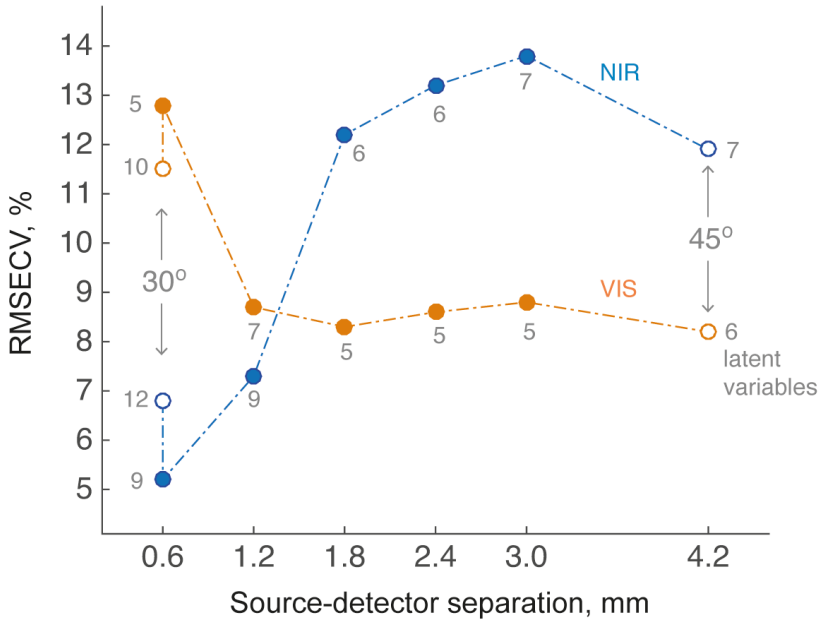

Figure 6.9 Averaged RMSECV obtained by the rdCV method on models developed with NIR (blue circles) and Vis (orange circles) spectra, taken from six polymerisation reactions by using a SARDR probe with different SD distances, from 0.6 to 4.2 mm.

By analysing the results from models built with spectra from individual SD distances, it is clear that the predictive capability becomes poorer as the distance between the source and the NIR detector increases. Models developed with SD distances beyond 1.2 mm did not provide good results, reaching prediction errors higher than 10% and coefficient of determinations below 0.90. Meanwhile, Vis-

based models achieved similar predictive capability for most of the SD distances, with the exception of the closest fibre (SD distance = 0.6 mm), which generated errors higher than 11 % for both normal and angularly incident fibres. The full summary of statistics for the models built for individual fibres are presented in Appendix 6.1.

6.6.1 Integrating spectra taken at multiple source-detector separations

Table 6.3 presents the summary of results obtained from the rdCV and external validation PLS models developed by combining normal incident fibres by using the multi-block and co-adding approaches. For the rdCV models, the table presents the root mean square error of cross validation (RMSECV) and prediction (RMSEP) and their respective confidence interval based on the standard errors of prediction. The coefficient of determination (R^2) and the number of latent variables are also shown. The statistics of the external validation models are highlighted in light blue columns for both multi-block and co-adding approaches. They contain the calibration errors (RMSEC) obtained by the updated calibration set and the RMSEP values achieved by validating the external reactions R3 and R5. The number of blocks (nb) used to build each model is also shown.

Table 6.3 Summary of results from SARDR-NIRS PLS-based models built by using the multi-block and co-adding approaches for the o determination of monomer conversion in suspension polymerisation reactions. For the rdCV calibration and validation results, the RMSECV and RMSEP values are shown with their confidence interval for every SD distance, together with their respective coefficient of determination, R^2, and the optimum number of latent variables. For the external validation, the RMSEC and RMSEP values are also shown for each SD distance. **nb** = number of blocks

Multi-block									Co-adding								
		PLS Repeated Double CV				External validation					PLS Repeated Double CV				External validation		
SD distance combination	**nb**	**RMSECV** % ±CI95%	**RMSEP** % ±CI95%	**R^2**	**lv**	**RMSEC** %	**RMSEP** %		**SD distance combination**	**nb**	**RMSECV** % ±CI95%	**RMSEP** % ±CI95%	**R^2**	**lv**	**RMSEC** %	**RMSEP** %	
							R3	**R5**								**R3**	**R5**
0.6	**1**	5.2 ± 0.02	5.1 ± 0.04	0.97	9	4.8	7.6	5.6	**0.6**	**1**	5.2 ± 0.02	5.1 ± 0.04	0.97	9	4.8	7.6	5.6
fused to 1.2	**2**	5.5 ± 0.03	5.4 ± 0.06	0.97	12	4.4	7.8	4.0	**co-added to 1.2**	**1**	5.8 ± 0.03	5.7 ± 0.06	0.96	9	5.0	6.6	5.0
and 1.8	**3**	6.0 ± 0.03	5.8 ± 0.07	0.96	12	4.7	6.9	6.3	**and 1.8**	**1**	7.6 ± 0.04	7.6 ± 0.07	0.92	8	6.6	7.0	8.1
and 2.4	**4**	6.4 ± 0.04	6.3 ± 0.07	0.96	12	5.2	6.3	6.1	**and 2.4**	**1**	8.4 ± 0.04	8.3 ± 0.08	0.91	7	7.6	8.1	17.3
and 3.0	**5**	6.8 ± 0.04	6.6 ± 0.07	0.96	13	5.1	5.9	6.9	**and 3.0**	**1**	9.1 ± 0.05	9.0 ± 0.09	0.89	7	7.9	11.7	18.7

Overall, for most of the fibre combinations, the rdCV calibration models estimated conversion with sufficient accuracy, and were successfully validated upon test samples. However, the predictive power of these fibre combinations decreased as farther fibres were included to the models, no matter which data arrangement approach was employed.

Fibre combinations containing fibres with SD distances larger than 1.8 mm provided less accurate models, especially for the models built by using the co-adding approach. In fact, by fusing four extra fibres (SD distance = 1.2, 1.8, 2.4 and 3.0 mm) to the 0.6 mm fibre, the RMSEP values increased from 5.2 to 9.1 % in the co-adding model, and to 6.8% in the multi-block one. For all fibre combinations, models developed by the co-adding approach provided poorer estimations of monomer conversion than the ones developed by the multi-block approach.

The fibre combination comprising the two shortest SD distances (SD distance = 0.6 and 1.2 mm) provided the best results when validation was performed on the external reactions, as shown in the grey background columns from Table 6.3. Despite the increasing errors in the rdCV calibration, in the external validation, the calibration errors (RMSEC) decreased from 4.8 to 4.4 % by including one extra SD distance to the starting model. This combination provided a good predictive capability by estimating monomer conversions with errors around 7.8 % for the reaction R3 and 4.4%, for the reaction R5, while for the co-adding approach an error of 6.6% and 5.0 % were achieved, respectively.

Although providing slightly higher errors, co-adding models were much simpler than the multi-block ones, requiring three less latent variables to describe data variance. One possible explanation for this better parsimony is the lower number of **X** variables in the co-added models, because their input block has the same number of variables independent of the number of fibres co-added, i.e., 335 variables in this

case. In multiblock PLS, although the blocks are initially treated separately, their scores vectors are concatenated forming a super scores matrix, which has multiples of 335 variables according to the number of blocks, hence the number of **X** variables ranges from 670 in the two-block model to 1675 variables, in the five-block model. As a consequence, PLS modelling requires additional latent variables to describe the information contained in these extra **X** variables. However, this argument alone does not explain why the latent variables numbers decrease by adding further fibres in the co-adding approach.

Because the absorbances are summed up in the co-adding approach, models are more sensible to problematic samples or measurements with poorer signal-to-noise ratio, which happens when increasing SD distance in the NIR range. This information may explain why the co-adding approach presented lower performance than the multi-block one, especially in the longer SD distances.

In a similar fashion adopted for the NIRS models, PLS models based on repeated double cross validation (rdCV) were developed between Vis spectra taken from samples of six polymerisation reactions and their respective monomer conversions. As explained earlier, Vis spectroscopy is not a natural method of choice for this application, especially due to the absence of absorption bands related to styrene and polystyrene in the visible range. Despite its limitations, models based on the visible range were successfully used for monitoring particle concentration in Chapters 4 and 5. Such models, though, were based in a simple system, consisting only in polystyrene and water. In actual reactions, where the level of complexity is much higher, the validity of the predictions from these models needs to be further analysed. Table 6.4 presents the results obtained by the PLS-based models developed by using the visible spectral range.

Table 6.4 Summary of results from SARDR-Vis PLS-based models built by using the multi-block and co-adding approaches for the determination of monomer conversion in suspension polymerisation reactions. For the rdCV calibration and validation results, the RMSECV and RMSEP values are shown with their confidence interval for every SD distance, together with their respective coefficient of determination, R^2, and the optimum number of latent variables. For the external validation, the RMSEC and RMSEP values are also shown for each SD distance. nb = number of blocks

Multi-block									Co-adding								
		PLS Repeated Double CV				External validation					PLS Repeated Double CV				External validation		
SD distance combination	**nb**	**RMSECV** % ±CI95%	**RMSEP** % ±CI95%	**R^2**	**lv**	**RMSEC** %	**RMSEP** %		**SD distance combination**	**nb**	**RMSECV** % ±CI95%	**RMSEP** % ±CI95%	**R^2**	**lv**	**RMSEC** %	**RMSEP** %	
							R3	**R5**								**R3**	**R5**
3.0	1	8.8 ± 0.05	8.8 ± 0.09	0.89	5	8.4	28.1	11.9	**3.0**	1	8.8 ± 0.05	8.8 ± 0.09	0.89	5	84	28.1	11.9
fused to 2.4	2	7.5 ± 0.04	7.4 ± 0.06	0.92	8	6.6	12.3	12.6	**co-added to 2.4**	1	7.4 ± 0.04	7.3 ± 0.06	0.92	6	6.9	19.1	13.5
and 1.8	3	7.4 ± 0.04	7.2 ± 0.07	0.93	10	6.0	14.4	13.6	**and 1.8**	1	7.0 ± 0.04	6.9 ± 0.06	0.93	7	6.1	11.1	9.0
and 1.2	4	7.6 ± 0.04	7.4 ± 0.08	0.93	11	6.4	7.3	18.9	**and 1.2**	1	7.1 ± 0.03	7.0 ± 0.06	0.93	6	6.5	6.5	15.8
and 0.6	5	7.3 ± 0.04	7.1 ± 0.09	0.94	13	5.7	29.1	12.6	**and 0.6**	1	7.3 ± 0.03	7.3 ± 0.06	0.92	6	6.7	22.1	17.5

As expected, the errors obtained by the VIS models were in general higher than the ones obtained by NIRS. However, unlike NIRS models, for both co-adding and multi-block approaches, the accuracy improved when including information from extra source-detector distances. In the multi-block models, RMSECV values decreased from 8.8 % to 7.3% by including four extra SD distances, whereas for the co-adding approach an improvement from 8.8 to 7.0 % was possible by adding two SD distances. However, despite being successfully validated upon rdCV test sets, all Vis-based models were unable to estimate the monomer conversion on the external reactions with sufficient accuracy, resulting in RMSEP values beyond 10%. This may be explained by the fact that Vis models are not based on chemical information contained in the spectra, but indirect one related to the scattering of beads. The scattering baseline characteristic from the external reactions may be different from that one modelled in the rdCV calibration phase, resulting in poor estimations of monomer conversion.

6.6.2 Can we improve model performance by adding angularly resolved measurements?

In the previous section, diffuse reflectance spectra collected from a normal incident source-fibre were used as the starting block for integrating measurements taken at different source-detector distances. To this spectral block, the output from all subsequent farther (for the NIRS) or closer (for the VIS) normal incident fibres was added one-at-the time, and PLS-based regression models were built and validated by using two different data fusion approaches. The aim was to evaluate the potential of spatially resolved measurements for the *in situ* monitoring of suspension polymerisation reactions.

In this section, to investigate whether the addition of angularly incident measurements could improve the predictive capability of such models, diffuse reflectance spectra collected from a 30° angularly incident fibre (SD distance = 0.6 mm for NIR and 4.2 mm for Vis) were added to all normal incident fibre combinations discussed in the previous section. Likewise, PLS-based models were developed for each combination in the same fashion done for the normal incident fibres.

Figure 6.10 shows the impact on the RMSECV caused by adding an angularly orientated fibre to the combinations of normal incident fibres, for both multi-block and co-adding approaches. The RMSEPs are not shown in this figure, but are similar or better than the cross-validation errors. The complete summary of results can be found in Appendix 6.2.

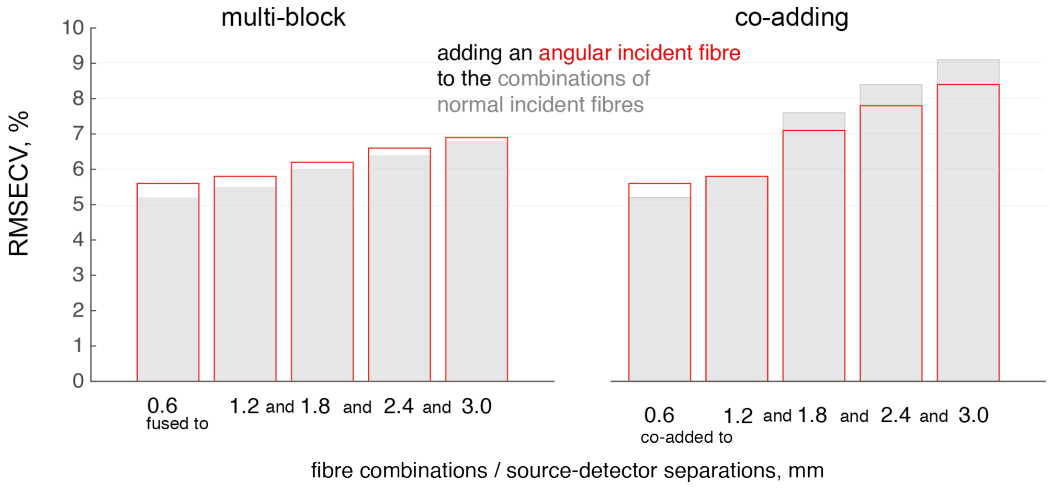

Figure 6.10 Impact on the RMSECV by adding an angular incident fibre to the combinations of normal incident measurements. RMSECV obtained from (a) multi-block and (b) co-adding PLS models based on NIR spectra. The grey closed bars represent the combinations of normal incident fibres of increasing SD distance, while the red open bars represent the same combinations with the fused angular orientated fibre.

The rdCV results for the multi-block models show no significant improvement in the estimation of monomer conversion when including angularly incident measurements to the normal incident combinations. The RMSECVs obtained by the models containing spectra from the angularly orientated fibre (grey bars) were slightly higher than the ones built with only normal incident fibres, and, in general, required more latent variables to describe the variance in the data. However, as more fibres were included to the model, the difference in performance decreased, becoming similar in terms of error and parsimony when all five normal incident fibres were included to the model.

By contrast, when the co-adding approach was used to build the models, the addition of an angular orientated fibre provided similar or improved accuracy. The complexity of the models did not increase as more blocks were added, and the errors were slightly lower than when using only normal incident fibres. However, by adding information from longer SD distances, co-adding models provided poorer results than the multi-block ones. Moreover, despite the improvement by the addition of the angular fibre, the calibration and prediction errors are still appreciably higher than the ones obtained by using shorter SD distances.

The benefit of including an angularly orientated fibre becomes evident when the extended calibration set was used to validate the two external reactions. Figure 6.11 shows the RMSEPs obtained for the external reactions R3 and R5, by both multi-block and co-adding approaches. The 30^{o} angularly incident fibre was added to each combination of normal incident fibres (grey circles), starting with the individual 0.6 mm distant fibre.

Figure 6.11 clearly shows that the inclusion of the 30^{o} angular incident fibre improved the estimation of monomer conversion on both reactions for most of the fibre combinations, regardless of the data fusion approach used.

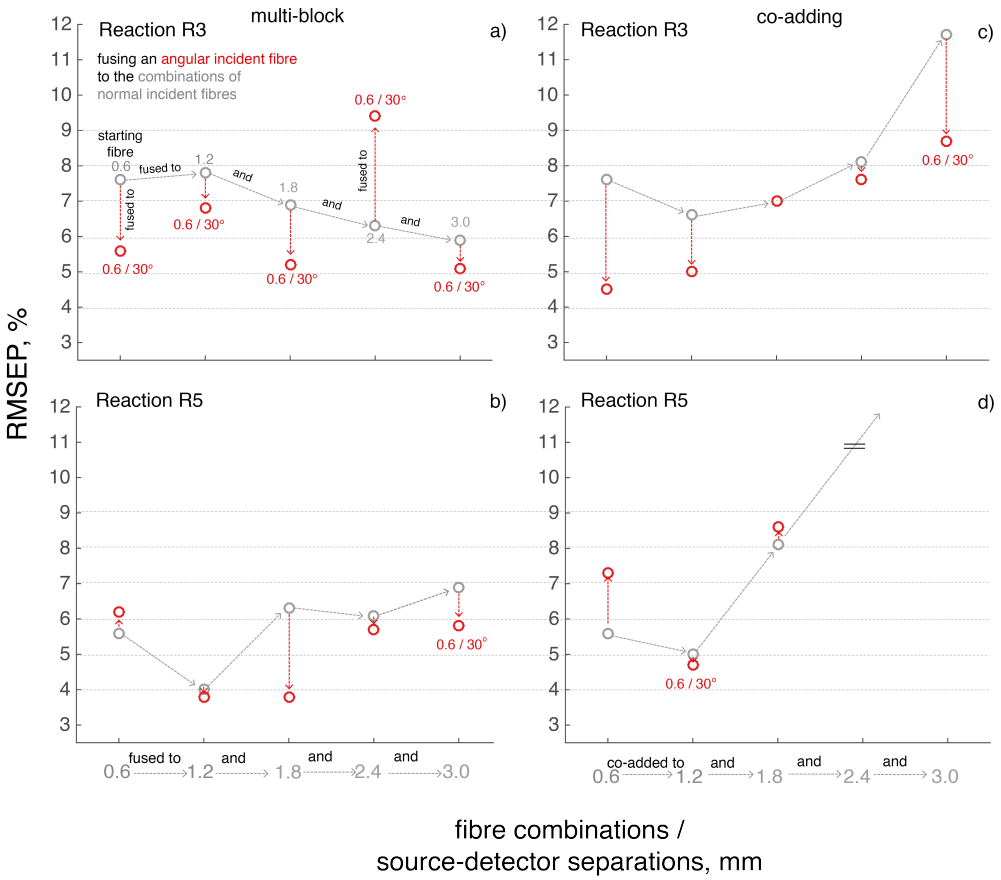

Figure 6.11 External validation. RMSEP for the two external reactions, R3 and R5, for both multiblock and co-adding approaches based on NIR spectra. Gray circles represent the models developed by using only normal incident fibres, while red ones represent the addition of the 30° angular incident fibre.

The multi-block approach provided consistent results for most of the models. The extended calibration model was able to estimate both reactions with satisfactory results for most of the fibre combinations, especially when the angularly incident fibre was included to the models. In particular, the model developed by combining the three shortest SD distances (0.6, 1.2 and 1.8 mm) to the angularly orientated fibre provided the best results. This combination was able to accurately estimate the monomer conversion for both reactions with errors of 5.2 and 3.8%, respectively.

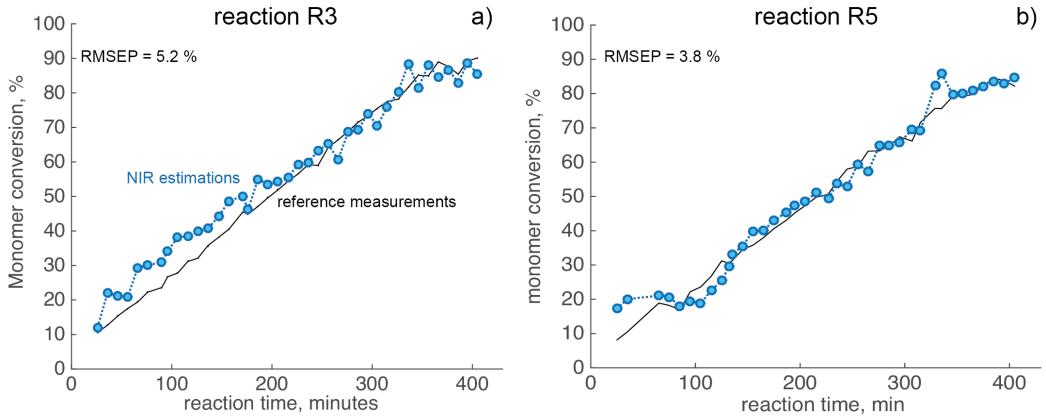

Figure 6.12 NIR predicted monomer conversion (blue circles) versus reference values (black curves) throughout two suspension polymerisation reactions, a) R3 and b) R5, not used in calibration development. Predictions based on the multi-block fused diffuse reflectance spectra obtained by adding an angularly incident fibre (30^{o} – 0.6mm) to three normal incident fibres (SD distance = 0.6, 1.2, 1.8 mm).

Figure 6.12 shows that monomer conversions were better estimated along the reaction R5. It presents the monomer conversion measured by gravimetric analysis (black curves) through out the reaction, together with the estimated values by NIR spectroscopy (blue dotted circles connected by blue curves). Apart from samples at the very beginning of the reaction, and other two samples around 320 minutes, the agreement between the reference values and the NIR estimations was excellent. When the reaction reached 120 minutes, the stabiliser SDBS was added to the reaction vessel, altering the reaction media and consequently, affecting reference and spectroscopic measurements. Note that the NIR measurements were able to detect these changes, but overestimated the monomer for these initial samples. The samples at the end of the reaction, which are very important for end point detection, were estimated with excellent accuracy. For the reaction R3, the NIR predictions overestimated the monomer conversion, especially for the first 200 minutes, but stabilised after that obtaining reliable estimates.

Important to highlight that without the inclusion of the angular measurements, the RMSEP for this fibre combination would be 6.9 and 6.3 %, respectively. Moreover, if only the closer fibre (SD distance = 0.06mm) was used, RMSEP of 7.6 and 5.6 %, respectively, would be obtained.

Despite the poorer results obtained in the calibration phase, the co-adding approach estimates were particularly good, reaching optimum results when combining the two shortest normal incident fibres to the angular incident one. This combination allowed an estimation of conversion with errors of 5.0 and 4.7%, for the reactions R3 and R5, respectively. Models containing fibres with SD distances higher than 1.2 mm provided poor estimates. For the reaction R5, the RMSEP obtained for a combination containing SD distance farther than 1.8 was higher than 15%, and so it was not shown here.

Considering the better parsimony of the co-adding models, we are tempted to conclude that its best performing model is superior than the one obtained by the multi-block approach. The co-adding models using only three fibres was able to estimate conversion well, and needed only 9 latent variables, compared to the 13 needed for the 4-block multiblock model.

For the visible spectroscopy modelling, the improvement caused by adding angularly resolved measurements was more evident in both rdCV and external validation phases. Figure 6.13 presents the summary of statistics for all model combinations containing the 30^{o} angular incident fibre, for both multi-block and co-adding approaches.

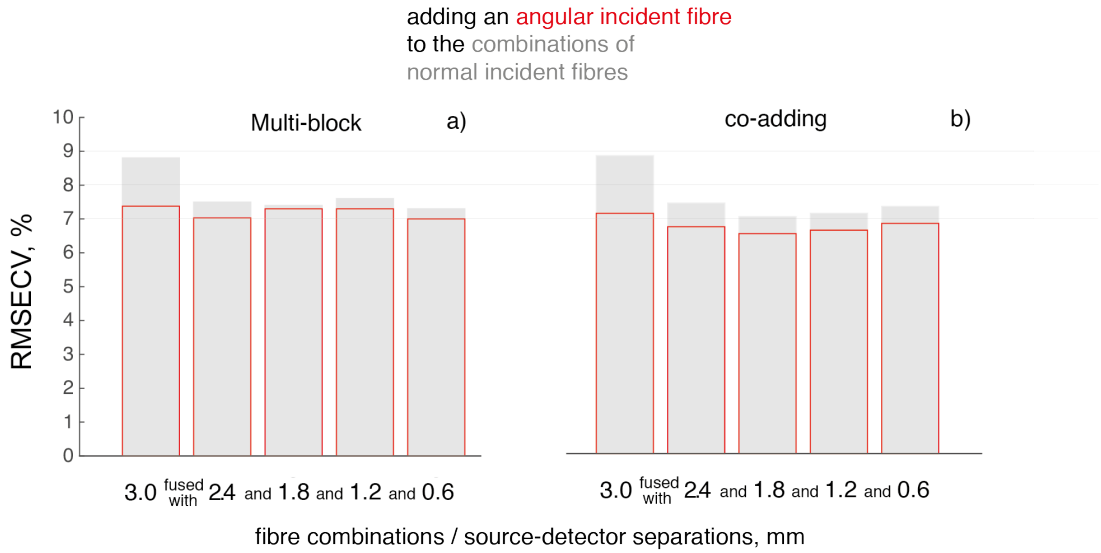

Figure 6.13 Impact on the RMSECV by adding an angular incident fibre to the combinations of normal incident measurements. RMSECV obtained from (a) multi-block and (b) co-adding PLS models based on Vis spectra. Grey closed bars represent the combinations of normal incident fibres of increasing SD distance, while the red open bars represent the same combinations with the fused angular orientated fibre.

Unlike NIRS models, by adding angular incident measurements, the rdCV cross-validation and prediction errors decreased for both multi-block and co-adding approaches, as can be seen by the smaller red bars in Figure 6.13 a and b. The fusion of one extra fibre to the 3.0 mm distant starting fibre improved considerably the accuracy of the models. However, as more fibres were added to the model, the performance of visible models were not significantly affected, since the RMSECV values obtained held constant around 7 %.

Despite the similar performance in the rdCV calibration and validation phases, multi-block and co-adding approaches provided contrasting results when subjected to external validation. Figure 6.14 presents the validation of the extended calibration set upon the two external reactions, R3 and R5, for both data fusion approaches.

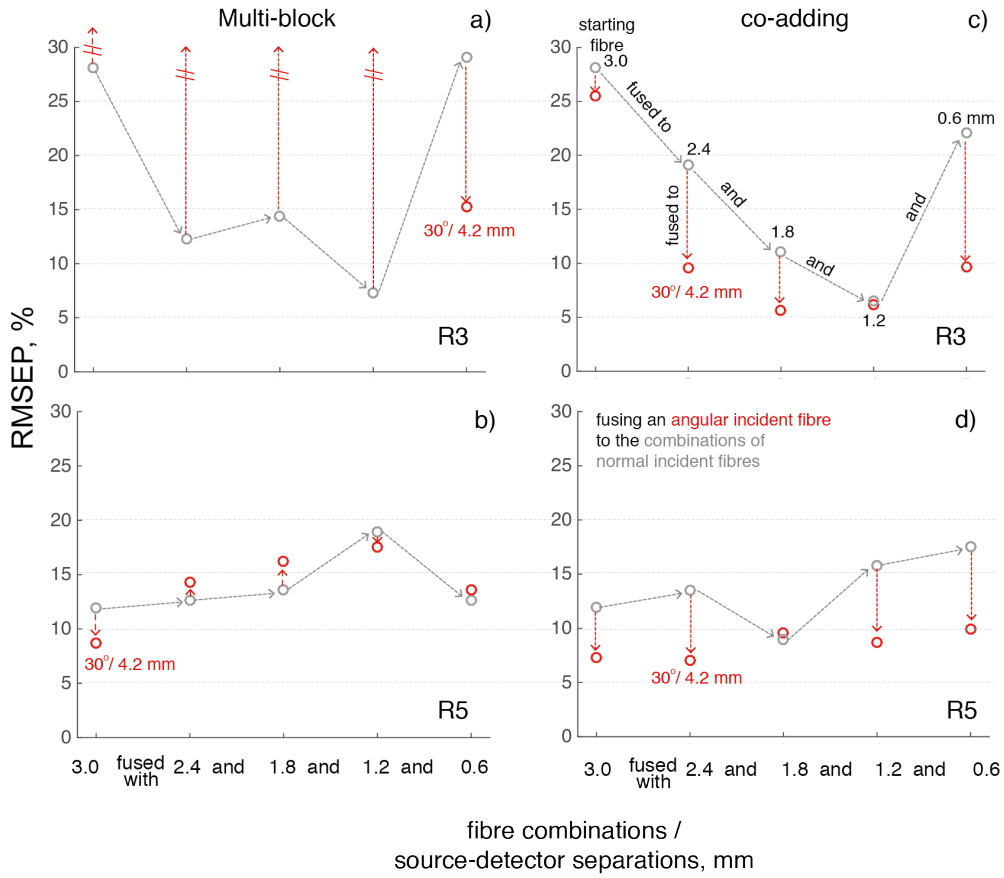

Figure 6.14 External validation. RMSEP for the two external reactions, R3 and R5, for both multiblock and co-adding approaches based on Vis spectra. Gray circles represent the models developed by using only normal incident fibres, while red ones represent the addition of the 30° angular incident fibre.

The multi-block PLS models were unable to provide reliable results for any combination of fibres, whether or not the angularly incident fibre was added. Monomer conversion was poorly estimated in both reactions, but estimations for reaction R3 were particularly poorer.

Co-adding models also provided poor predictive capability for both reactions, especially for the combinations of normal incident fibres. However, the addition of the angularly incident fibre dramatically improved the RMSEP values in both reactions. Considering the joint results for both reactions, the combination of the

three farther SD distances (1.8, 2.4 and 3.0 mm) to the angular incident fibre provided the better predictive capability, presenting RMSEP values of 5.6 and 9.6 %. Important to remember that for the reaction R3, the RMSEP values obtained by the individual 0.6 mm distant fibre and by the combination of the three normal incident fibres were 28.1 and 11.1 %, respectively, and for the reaction R5, 11.9 and 9.0 %, respectively. Thus, a considerable improvement was achieved by adding angularly incident measurements to the combinations of normal incident fibres. Furthermore, the calibration models needed only six latent variables to describe the variability in the data. Although better results were achieved by NIR based models, these results for visible spectroscopy deserve attention.

6.6.3 Impact of scattering correction

In this chapter, a large quantity of models was developed for the comparison of spectroscopy techniques, fibre combinations and modelling strategies. For didactic reasons, the results discussed so far were based only on diffuse reflectance spectra linearised to absorbance (log 1/R). However, although models based on the absorbance spectra provided satisfactory predictive capability, the results might be improved by a further pre-processing step focused on light scattering effects.

Light scattering manifests in the diffuse reflectance spectra as non-linear distortions like baseline fluctuations, additive and multiplicative scattering. These uncertain non-linear variations are caused mainly by differences in the shape and size of beads in the medium. In turbid media like polymerisation reactions, these uncertain non-linear variations are even more complicated, because the size distribution of droplets and polystyrene beads changes along the polymerisation, together with the viscosity of the reaction media. Such variation in the spectra deteriorate the predictive efficiency of PLS models for estimating monomer conversion along the reaction.

To evaluate the impact of pre-processing on the predictive capability of the VIS and NIR models, each dataset was pre-processed by using the SNV, SNV-Detrending, MSC and EMSC techniques, and PLS-based models were developed for all fibre combinations discussed in the previous sections. For multi-block data sets, the pre-processing techniques were applied to each block separately to preserve the inherent information captured by each of them, whereas for the co-added data sets, the pre-processing was applied before the co-adding step. To simplify the discussions, only the best results for each pre-processing technique are shown and compared to the one obtained by the absorbance spectral dataset. Table 6.5 presents the statistic summary of results obtained for NIR based models developed for both data fusion approaches.

Table 6.5 Impact of pre-processing to the predictive capability of SARDR-NIR PLS-based models. Summary of the best results obtained from pre-processing the absorbance spectra by using the SNV, SNV-DT, MSC and EMSC techniques. For the rdCV calibration and validation results, the RMSECV and RMSEP values are shown with their confidence interval for every SD distance, together with their respective coefficient of determination, R^2, and the optimum number of latent variables. For the external validation, the RMSEC and RMSEP values are also shown for each SD distance. **nb** = number of blocks

Multi-block									Co-adding								
PLS Repeated Double CV						External validation			PLS Repeated Double CV						External validation		
SD distance	**nb**	**Pre-processing**	**RMSECV** % ±CI95%	**R^2**	**lv**	**RMSEC** % w/w	**RMSEP** % w/w		**SD distance**	**nb**	**Pre-processing**	**RMSECV** % ±CI95%	**R^2**	**lv**	**RMSEC** % w/w	**RMSEP** % w/w	
							R3	**R5**								**R3**	**R5**
0.6A-1.8	4	none	6.1 ± 0.03	0.96	13	5.1	5.2	3.8	**0.6A-1.2**	1	none	5.8 ± 0.03	0.96	9	5.2	5.0	4.7
0.6A-1.8	4	SNV	5.2 ± 0.03	0.97	11	3.8	9.9	4.2	**0.6A-0.6**	1	SNV	5.6 ± 0.03	0.96	13	4.3	5.1	5.8
0.6A-0.6	2	SNV-DT	4.9 ± 0.03	0.97	10	4.1	5.6	4.3	**0.6A-1.8**	1	SNV-DT	6.3 ± 0.03	0.96	7	5.6	3.9	7.4
0.6A-1.8	4	MSC	5.5 ± 0.03	0.97	11	4.0	8.6	4.2	**0.6A-1.8**	1	MSC	6.3 ± 0.03	0.95	7	5.8	6.4	7.8
0.6A-1.8	4	EMSC	5.9 ± 0.04	0.96	7	4.9	6.4	6.4	**0.6**	1	EMSC	5.4 ± 0.04	0.90	8	4.8	9.2	5.6

The four-block combination resulted in the best predictive capability for all pre-processing techniques, excepting the SNV-DT, which best result was achieved by the two-block combination of angular and normal incident 0.6 mm fibres. The models developed on the spectra pre-processed by using both SNV variants provided the best performance on the rdCV calibration, obtaining RMSECV values of 5.2 % and 4.9 %, for the SNV and SNV-DT. However, only the SNV-DT was able to provide good estimation for both reactions in the external validation.

Although the model developed with the unprocessed absorbance spectra obtained the poorest performance in the rdCV phase, it worked well in the external validation. This model based on absorbance spectra presented errors of 5.2 and 3.8 %, slightly better than the ones obtained by the SNV-DT technique, which errors were 5.6 % and 4.3 %. Figure 6.15 presents the RMSECV curves of the best performer models built by the different pre-processing techniques.

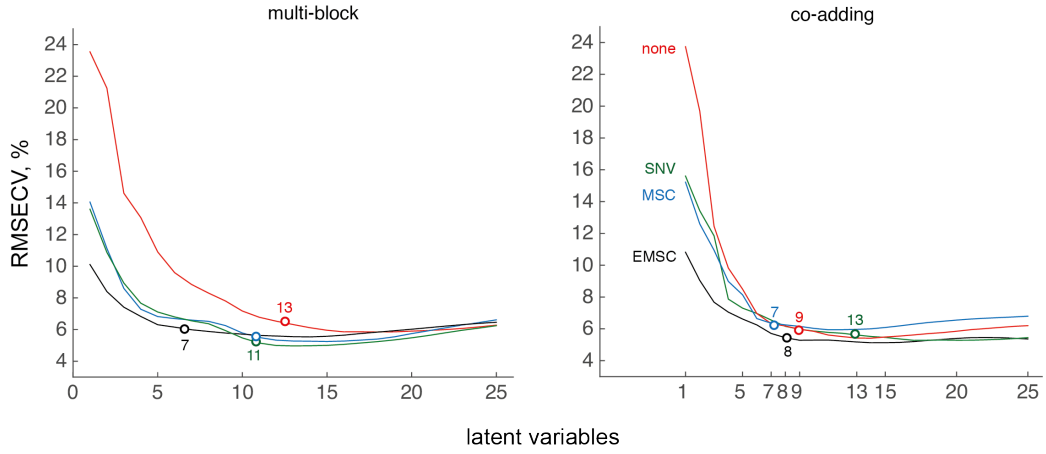

Figure 6.15 rdCV RMSECV curves for estimating monomer conversion by using the (a) multi-block and (b) co-adding approaches on data processed by different pre-processing techniques.

The scree plot profiles for both approaches present similarities. First, they show that the application of pre-processing is usually beneficial compared to using only the

absorbance spectra. For the first latent variables, the models based on absorbance data resulted in the highest errors, meaning that for the absorbance data, these first latent variables are not directly related to the styrene-polystyrene conversion. Second, MSC and SNV presented similar results for both approaches, which is normally expected to happen[42,43]. EMSC initially provided the lowest errors, however, as the number of latent variables increased, both approaches have shown differences. In the multi-block approach, after nine latent variables, the errors obtained for the EMSC stabilised at around 6%, whereas the SNV and MSC they continued to decrease. In the co-adding approach, the EMSC maintained the lowest errors for all latent variables, and the unprocessed spectra reached similar accuracy as the SNV and MSC-treated spectra at around 7 latent variables.

Models based on SNV-DT pre-processed spectra were the simplest ones for both multi-block and co-adding approach. In only required 10 and 7 latent variables, respectively, to describe the variance in the data. In the multi-block models, although SNV-DT needed three extra latent variables than the EMSC, it only required two blocks of spectra to achieve superior performance.

In general, the predictive power and parsimony of the rdCV models were improved by pre-processing. A few of these models were able to predict the monomer conversion from the external reaction with satisfactory accuracy, obtaining RMSECV errors among the better ones. However, most of the pre-processed models had difficulties in the external validation, especially when estimating the monomer conversion for the reaction R3. Considering that this reaction was performed before probe maintenance, as discussed earlier, this may indicate that the pre-processing tools were not robust enough to discern or correct for the undesired effects in the spectra on the face of two different levels of baseline effects.

Although the absorbance spectra did not provide the best results in terms of RMSECV, a comprehensive analysis of the results indicate they obtained the most consistent performance on all scenarios and combinations, obtaining errors in the external validation similar to the rdCV errors. Hence, it provided a simplified model with results as accurate as the pre-processing ones.

6.7 Conclusions and Outlook

The goal of this chapter was to evaluate the potential of Visible and NIR spatially resolved diffuse reflectance measurements for the *in situ* monitoring of suspension polymerisation reactions. Instead of the usual approach employed in Spatially Resolved Spectroscopy (SRS) data, in which the absorption and scattering spectra are extracted from multiple measurements by invoking the computationally intensive radiative transfer theory, we proposed an empirical approach for manipulating the spatially and angular resolved information. Particular interest was placed on studying whether introducing new variation in measurement configurations could improve the predictive capability of multivariate calibration models. For this, several issues regarding the manipulation of the multidimensional data, its pre-processing and modelling parameters were investigated here.

For this purpose, PLS based models were developed on a range of combinations of SD distances and angles of incidence. All models were subjected to a robust calibration and validation procedure based on the repeated double cross validation. Furthermore, the capability of the models to estimate monomer conversion from future samples was evaluated and validated by using an external validation data set.

Selecting the best model complexity I investigated different approaches for selecting the optimum model complexity in the scope of the repeated double cross validation. Different parsimony levels were studied in the one-standard deviation rule, which were compared to the Haaland Thomas procedure. The results indicated that in the one-standard deviation rule with a one standard deviation, as proposed by Filzmoser et al [36], the choice of the optimum number of latent variables was too conservative, leading to underfitting. The results improved by relaxing the parsimony factor to 0.25 SD. However, by using the zero SD parsimony factor, which is in fact the minimum MSE method, the method overfitted the models, and required a large

number of latent variables. The Haaland Thomas criterion obtained similar fit in both calibration and external validation, selecting an acceptable number of latent variables.

Data fusion approaches I investigated whether the manner in which the SRS data are combined might impact the predictive capability and parsimony of the models. For this, we compared the performance of the multi-block and the co-adding approaches. Although the co-adding approach outperformed the multi-block approach in a few cases, in general, multi-block based models provided the best accuracy. In terms of parsimony, the co-adding approach provided models needing less latent variables.

Impact of pre-processing The advantage of pre-processing was clear in the rdCV results. However, most of the pre-processed models were not able to estimate monomer conversion from the external reaction, especially for the reactions performed before the major probe maintenance. While their performance was good, this may indicate that the pre-processing tools were not robust enough to discern or correct for the undesired effects in the spectra, especially considering that the calibration and external validation sets may present different levels of baseline effects.

Adding SD distances

For the NIRS based models, the higher errors obtained by adding information from longer source-detector distances in both multi-block and co-adding approaches could be a consequence of their low signal-to-noise ratio, since the intensity of collected light becomes weaker as the SD distance is increased. Indeed, in Chapter 4 it was shown that the farthest fibres provided poor quality near infrared spectra. Another contribution factor is the increase in the path length travelled by the photons, which increases the number of scattering events.

Adding angular information

When the monomer conversion from external reactions were estimated by the models, improved predictive capability was obtained by adding angular incidence measurements to the combinations of normal incident fibres, regardless the adopted data fusion approach, spectroscopy method and pre-processing technique.

Outlook

The results show that the Visible and NIR Spatially and Angularly resolved spectroscopies are able to provide reliable online estimates of monomer conversion from suspension polymerisation reactions. The technique is cost-efficient, non-invasive and allows in-situ measurements.

Although this chapter focused on the determination of monomer conversion, the techniques studied here have great potential for particle size characterisation. In special, due to the its higher tendency for light scattering, Visible spectroscopy has a strong potential for providing reliable estimations. Unfortunately this application will not be explored in this thesis due to the lack of reliable particle size measurements in our experiment.

In the next chapter, a new probe configuration will be tested in order to evaluate if better signal-to-noise ratio could improve model predictive capability. In this probe, besides the stronger light source, instead of individual illuminating fibres, two rings of fibres at two different distances from the detecting fibre will be used to improve signal acquisition. This probe will be tested on the two-component system studied in the Chapters 4 and 5. This experiment may suggest possible alterations to be done in a next version of the SARDR probe.

6.8 Bibliographic references

1. Kotoulas, C. & Kiparissides, C. in *Polymer Reaction Engineering* 209–232 (Blackwell Publishing Ltd, 2007). doi:10.1002/9780470692134.ch5

2. Kiparissides, C., Daskalakis, G., Achilias, D. S. & Sidiropoulou, E. Dynamic Simulation of Industrial Poly(vinyl chloride) Batch Suspension Polymerization Reactors. *Ind. Eng. Chem. Res.* **36,** 1253–1267 (1997).

3. Holmes, R. L., Burford, R. P. & Bertram, C. D. Preparation and characterization of poly(methyl methacrylate) beads. *J. Appl. Polym. Sci.* **109,** 1814–1822 (2008).

4. Dowding, P. J. & Vincent, B. Suspension polymerisation to form polymer beads. *Colloids and Surfaces A: Physicochemical and Engineering Aspects* **161,** 259–269 (2000).

5. Kotoulas, C. & Kiparissides, C. A generalized population balance model for the prediction of particle size distribution in suspension polymerization reactors. *Chemical Engineering Science* **61,** 332–346 (2006).

6. Gonçalves, O. H., Nogueira, A. L., Araújo, P. H. H. & Machado, R. A. F. Effects of Operational Parameters on Particle Size Distributions in Methyl Methacrylate Suspension Polymerization. *Ind. Eng. Chem. Res.* **50,** 9116–9122 (2011).

7. Brooks, B. Suspension Polymerization Processes. *Chemical Engineering & Technology* **33,** 1737–1744 (2010).

8. Dowding, P. J., Goodwin, J. W. & Vincent, B. The characterization of porous styrene–glycidyl methacrylate copolymer beads prepared by suspension polymerization. *Colloids and Surfaces A: Physicochemical and Engineering Aspects* **145,** 263–270 (1998).

9. Lu, M. J. in *Column Handbook for Size Exclusion Chromatography* 3–26 (Elsevier, 1999). doi:10.1016/B978-012765555-0/50002-6

10. Araújo, P. H. H., Sayer, C., Giudici, R. & Poço, J. G. R. Techniques for reducing residual monomer content in polymers: A review. *Polym Eng Sci* **42,** 1442–1468 (2004).

11. Leiza, J. R. & Pinto, J. C. in *Polymer Reaction Engineering* 315–362 (Blackwell Publishing Ltd, 2007). doi:10.1002/9780470692134.ch8

12. Kiparissides, C. & Morris, J. Intelligent manufacturing of polymers. *Computers & Chemical Engineering* **20,** S1113–S1118 (1996)

13. Celis, M.-T. & Rubio, L. H. G. Spectroscopy as a Tool for Characterization of Monomer Emulsions. *Journal of Dispersion Science and Technology* **28,** 271–278 (2007).

14. Cardenas-Valencia, A. M. & García-Rubio, L. H. Reflectance Spectroscopy Probes for Continuous Process Monitoring of Particle Size, Concentration and Chemical Composition. in 14th International Congress of Chemical and Process Engineering.

15. Florenzano, F. H., Enohnyaket, P., Fleming, V. & Reed, W. F. Coupling of near infrared spectroscopy to automatic continuous online monitoring of polymerization reactions. *European Polymer Journal* **41,** 535–545 (2005).

16. Pasquale, A. J. & Long, T. E. Real-Time Monitoring of the Stable Free Radical Polymerization of Styrene via in-Situ Mid-Infrared Spectroscopy. *Macromolecules* **32,** 7954–7957 (1999).

17. Elizalde, O., Leiza, J. R. & Asua, J. M. On-line monitoring of all-acrylic emulsion polymerization reactors by Raman spectroscopy. *Macromol. Symp.* **206,** 135–148 (2004).

18. Brearley, A. M. & Foulk, S. J. in *Process Analytical Technology* **5,** 493–520 (John Wiley & Sons, Ltd, 2010).

19. Dehanov, N. & Thennadil, S. Monitoring of conversion during emulsion polymerisation using near infrared spectroscopy. *NIR news* **25,** 9 (2014).

20. Reis, M. M., Araújo, P. H. H., Sayer, C. & Giudici, R. In Situ Near-Infrared Spectroscopy for Simultaneous Monitoring of Multiple Process Variables in Emulsion Copolymerization. *Ind. Eng. Chem. Res.* **43,** 7243–7250 (2004).

21. Aldridge, P. K., Kelly, J. J., Callis, J. B. & Burns, D. H. Noninvasive monitoring of bulk polymerization using short-wavelength near-infrared spectroscopy. *Anal. Chem.* **65,** 3581–3585 (1993).

22. Cherfi, A. & Févotte, G. On-line conversion monitoring of the solution polymerization of methyl methacrylate using near-infrared spectroscopy. *Macromolecular Chemistry and Physics* **203,** 1188–1193 (2002).

23. Nogueira, E. S., Borges, C. P. & Pinto, J. C. In-Line Monitoring and Control of Conversion and Weight-Average Molecular Weight of Polyurethanes in Solution Step-Growth Polymerization Based on Near Infrared Spectroscopy and Torquemetry. *Macromolecular Materials and Engineering* **290,** 272–282 (2005).

24. Kappler, B. *et al.* Real-time monitoring of ethene/1-hexene copolymerizations: determination of catalyst activity, copolymer composition and copolymerization parameters. *Polymer* **44,** 6179–6186 (2003).

25. Barnes, R. J., Dhanoa, M. S. & Lister, S. J. Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra. *Appl Spectrosc* **43,** 772–777 (1989).

26. Chen, Y.-C. & Thennadil, S. N. Insights into information contained in multiplicative scatter correction parameters and the potential for estimating particle size from these parameters. *Analytica Chimica Acta* **746,** 37–46 (2012).

27. Martens, H. & Stark, E. Extended multiplicative signal correction and spectral interference subtraction: new preprocessing methods for near infrared spectroscopy. *J Pharm Biomed Anal* **9,** 625–635 (1991).

28. Thennadil, S. N., Martens, H. & Kohler, A. Physics-based multiplicative scatter correction approaches for improving the performance of calibration models. *Appl Spectrosc* **60,** 315–321 (2006).

29. Steponavicius, R. & Thennadil, S. N. Extraction of chemical information of suspensions using radiative transfer theory to remove multiple scattering effects: application to a model multicomponent system. *Anal. Chem.* **83,** 1931–1937 (2011).

30. Farrell, T. J., Wilson, B. C. & Patterson, M. S. The use of a neural network to determine tissue optical properties from spatially resolved diffuse reflectance measurements. *Phys. Med. Biol.* **37,** 2281–2286 (1992).

31. Hjalmarsson, P. & Thennadil, S. N. Spatially resolved in vivo measurement system for estimating the optical properties of tissue in the wavelength range 1000-1700 nm. *Diagnostic Optical Spectroscopy in Biomedicine IV* **6628,** 662805–662805–9 (2007).

32. Kourti, T. Process Analytical Technology Beyond Real-Time Analyzers: The Role of Multivariate Analysis. *Critical Reviews in Analytical Chemistry* **36,** 257–278 (2007).

33. Boiret, M. & Chauchard, F. Use of near-infrared spectroscopy and multipoint measurements for quality control of pharmaceutical drug products. *Anal Bioanal Chem* **409,** 1–9 (2016).

34. Igne, B., Talwar, S., Feng, H., Drennen, J. K. & Anderson, C. A. Near-Infrared Spatially Resolved Spectroscopy for Tablet Quality Determination. *Journal of Pharmaceutical Sciences* **104,** 4074–4081 (2015).

35. Kohonen, J. *et al.* Multi‐block methods in multivariate process control. *J. Chemometrics* **22,** 281–287 (2008).

36. Filzmoser, P., Liebmann, B. & Varmuza, K. Repeated double cross validation. *J. Chemometrics* **23,** 160–171 (2009).

37. Haaland, D. M. & Thomas, E. V. Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. *Anal. Chem.* **60,** 1193–1202 (1988).

38. Chang, S. Y. & Wang, N. S. in *Multidimensional Spectroscopy of Polymers* **598,** 147–165 (American Chemical Society, 2009).

39. Burns, D. A. & Ciurczak, E. W. *Handbook of near-infrared analysis*. (2007).

40. Rasheed, T. M. A., Moosad, K. P. B., Nampoori, V. P. N. & Sathianandan, K. Overtone spectra of styrene and polystyrene in the visible and near infrared regions. 1–5 (2011).

41. Workman, J. & Weyer, L. *Practical Guide and Spectral Atlas for Interpretive Near*. (Infrared Spectroscopy, 2012).

42. Gemperline, P. in *Practical Guide To Chemometrics, Second Edition* 69–104 (CRC Press, 2009). doi:10.1201/9781420018301.ch4

43. Helland, I. Related versions of the multiplicative scatter correction method for preprocessing spectroscopic data. *Chemometrics and Intelligent Laboratory Systems* **29,** 233–241 (1995).

Chapter 7

7 Dual-ring diffuse reflectance probe

In this chapter, I wish to investigate an alternative optical geometry, simpler in design and smaller in outer diameter, for improving the amount of light that is collected by the detector. In this proposed design the illumination is carried out by rings of fibres, instead of individual illuminating fibres. Based on the results from previous chapters, it is envisaged that the consequent increase in the intensity of the signal may improve the accuracy of models, especially for NIR. Thus, the prototype fibre-optic multi-sensor probe for simultaneous measurements of polymer properties is evaluated by using the two-component polystyrene-water experiment used in Chapters 4 and 5, and the results are compared with the ones obtained by the SARDR probe.

7.1 Introduction

Process Analytical Technology research is moving towards analytics that allow monitoring a process through multiple simultaneous measurements[1] collected at different sampling locations or by different measuring principles. The rationale behind this trend is that by gathering complementary information about a process, we can better understand it and, consequently, be better prepared to monitor it[2,3].

In previous chapters, I evaluated whether spectroscopic measurements collected at different spatial distances and angles of incidence could be used as a PAT tool for *in-situ* monitoring of suspension polymerisation reactions. The results pointed out to models with improved predictive capability when incorporating such information on modelling. Nevertheless, further development and optimisation of Vis and NIR sensors are still required, especially concerning the quality of the measurements. In fact, the signal-to-noise ratio offered by the prototype SARDR probe limited the accuracy of the predictive models, especially for longer source-detector distances in the near-infrared range. The results suggested that a modification in the optical geometry of the probe would be required for optimised results. In this chapter, I wish to investigate an alternative optical geometry, simpler in design and smaller in outer diameter, for improving the amount of light that is collected by the detector. In this proposed design the illumination is carried out by rings of fibres, instead of individual illuminating fibres. It is envisaged that the consequent increase in the intensity of the signal has the potential to improve the accuracy of the models. For this purpose, the proposed prototype will be evaluated on the two-component system studied in Chapters 4 and 5, for the estimation of composition and size of polymer beads. As the probe is still an early prototype, it was not possible to use it on polymerisation reactions, owing to incompatibility between the reactions constituents and the adhesive used to attach the window to the probe head.

7.1.1 Dual-ring diffuse reflectance probe

The prototype probe used in this chapter was specifically developed for the OPTICO FP7 European Project. It was designed and built by Fibre Photonics Ltd., under specifications by the Measurements and Analytics Team at University of Strathclyde. The probe head is illustrated in Figure 7.1, which also shows the fibre optical geometry and sizes.

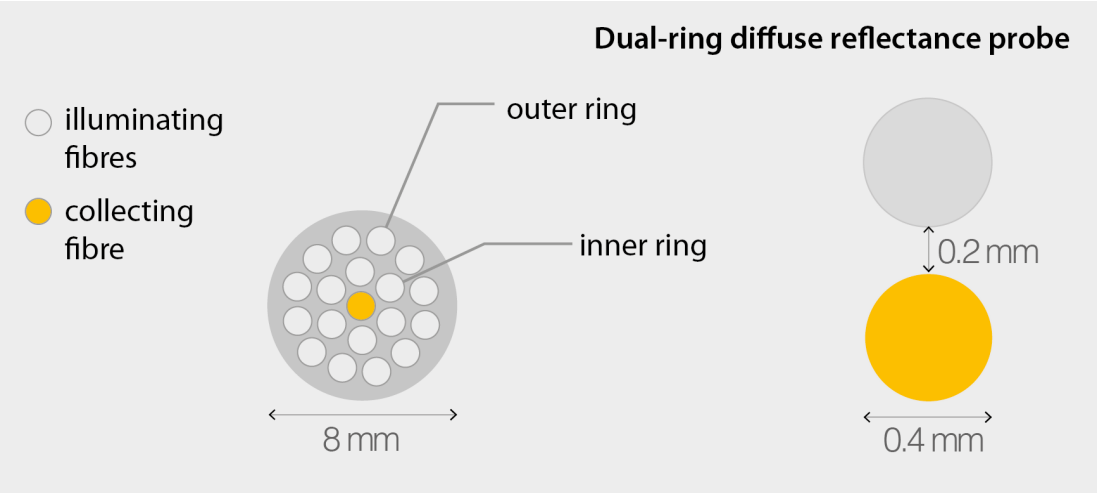

Figure 7.1 Schematic illustration of the dual-ring diffuse reflectance probe

Similarly to the SARDR probe, the 400 μm core diameter was maintained for all fibres. To improve signal-to-noise ratio, the probe geometry consists of a central fibre surrounded by two concentric illumination rings in a diameter ratio close to 2:1. The inner illumination ring is formed from six fibres bundled together, whereas the outer ring uses twelve. Both bundles are separately connected to the light source. The fibres are held together and bonded in to a ceramic ferrule. The complete fibre assembly is 4 m in length and trifurcates after 3 metres into a signal fibre and two separate illumination bundles. The probe head is 8 mm in diameter, which is much smaller compared to the 30 mm in diameter SARDR probe head.

7.2 Experiments and methods

The two-component system studied in Chapters 4 and 5 is investigated here. A total of 56 sample runs were performed, spanning a mean bead diameter range of 0.400 to 1.540 mm and a concentration range of 15 to 50%, at 5% intervals (Table 3.1). The experimental runs were randomly performed over several different days. Further details of the experiment can be found in section 3.1.1 of Chapter 3.

Table 7.1 Mean bead sizes and their span calculated for each diameter range

Range	Diameter range, mm	D50, mm
1	0.300:0.500	0.405 ± -
2	0.500: 0.630	0.565 ± 0.190
3	0.630: 0.800	0.699 ± 0.222
4	0.800: 1.000	0.890 ± 0.201
5	1.000: 1.250	1.108 ± 0.187
6	1.250: 1.400	1.314 ± 0.115
7	1.400: 1.600	1.540 ± 0.264

7.2.1 Spectral acquisition and pre-processing

Measurements in the near infrared region were obtained by a NIR spectrometer (NIRQuest-512, Ocean Optics) whereas spectra in the Visible region were registered by a visible range spectrometer (USB-4000, Ocean Optics). Each of the fibre bundles was connected to the port of a specific spectrometer.

Vis and NIR spectra were acquired from each one of two illuminating rings. For this test phase, the spectrometers were switched manually by removing and inserting the probe bundle from the spectrometer port. Reference and background spectra were also recorded from each illuminating ring. Further details of the spectral acquisition procedure can be found in Chapter 3, section 3.3.1.

After acquisition, all spectra were subjected to correction, normalisation with the integrating sphere spectrum, denoising by wavelet transform {Galvao:2004ie}, conversion from reflectance to logarithmic units, and variable scaling.

7.2.2 Model development and validation

Similarly to the previous chapters, repeated double cross-validation (rdCV)[4] was employed for PLS regression, in both model building and validation steps. The validation in the rdCV approach is performed on test set samples not used in the calibration phase, ensuring the estimations are not biased. A total of 100 repetitions of the double cross validation were performed, based on which the optimum number of latent variables was selected.

7.3 Integrating diffuse reflectance spectra from different fibre geometries and spectroscopy modes

As the collecting fibre can detect light from each of the two rings of fibres separately, it is possible to obtain four different combinations that make use of information acquired from both rings. The probe can be used as a single-sensor probe, exclusively collecting either (a) NIR or (b) Vis spectra, or as a multi-sensor probe, by collecting both spectroscopy modes, each of them in a separate ring (c,d).

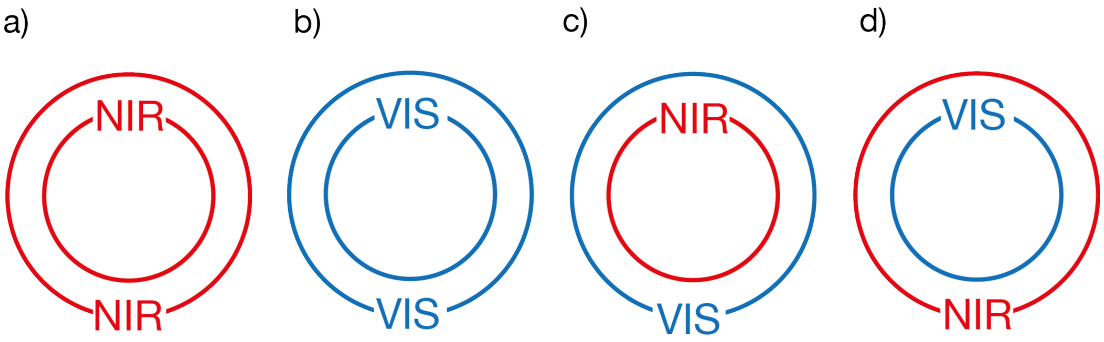

Figure 7.2 Different uses of the dual-ring diffuse reflectance probe. Using the probe for acquiring VIS or NIR spectra exclusively (a and b), or by using each ring for either NIR or Visible illumination (c and d).

Based on the combinations depicted in Figure 7.2, MB-PLS models were built by using the rdCV method for estimating bead concentration and size. The predictive capability of these models was compared to the ones obtained from standard PLS models developed on each spectroscopy and individual rings, to evaluate the impact of integrating different measurement geometries and spectroscopic modes. For all models, every spectral dataset was mean centred and divided by their standard deviation. Further normalisation was not needed since all blocks have the same number of wavelengths and share similar intensities. The results obtained by the dual-ring diffuse reflectance probe were also compared to the best models obtained by the SARDR probe, as discussed in Chapters 4 and 5.

7.3.1 Determination of EPS bead concentration

Before discussing the fusion of different spectroscopic sensors in the dual-ring diffuse reflectance probe, I will first present the results obtained when detecting the light from each illuminating ring independently by Vis and NIR spectroscopy. Table 7.2 summarises the results for the determination of bead concentration. It shows the root mean square error of cross validation (RMSECV) and prediction (RMSEP), their confidence interval based on the standard errors of prediction, together with their coefficient of determination (R^2) and the optimum number of latent variables chosen by the rdCV method.

Table 7.2 Summary of results of rdCV PLS regression models for the determination of EPS bead concentration in water by using the dual-ring diffuse reflectance probe in the Visible and NIR ranges. Mean RMSECV and RMSEP values calculated from 100 repetitions of the double cross validation are shown, together to their respective standard error at 95% confidence interval.

models	PLS Repeated Double CV			
	RMSECV % w/w (± CI95%)	**RMSEP** % w/w (± CI95%)	R^2	LV
Visible				
inner	2.2 ± 0.01	2.2 ± 0.03	0.97	4
outer	1.9 ± 0.03	1.7 ± 0.05	0.99	8
Near-infrared				
inner	2.0 ± 0.03	1.8 ± 0.05	0.99	9
outer	2.0 ± 0.02	1.9 ± 0.03	0.98	7

Bead concentration range: 15 – 50 % w/w

By considering solely the predictive power of the models, it can be noted that Vis and NIR models were not significantly affected by the two different illumination regimes, having obtained errors around 2% by weight. The main difference was in

the parsimony of the visible models. Despite the slightly smaller errors obtained by the outer ring, the model required twice the number of latent variables to describe the variance in the data. In fact, by comparing the trajectories in the RMSECV scree plot of Figure 7.3 a, where the RMSECV values are plotted against the number of latent variables selected for modelling, it can be seen that errors obtained by the outer ring between the latent variables three and eight do not follow a steady decrease as would be preferable.

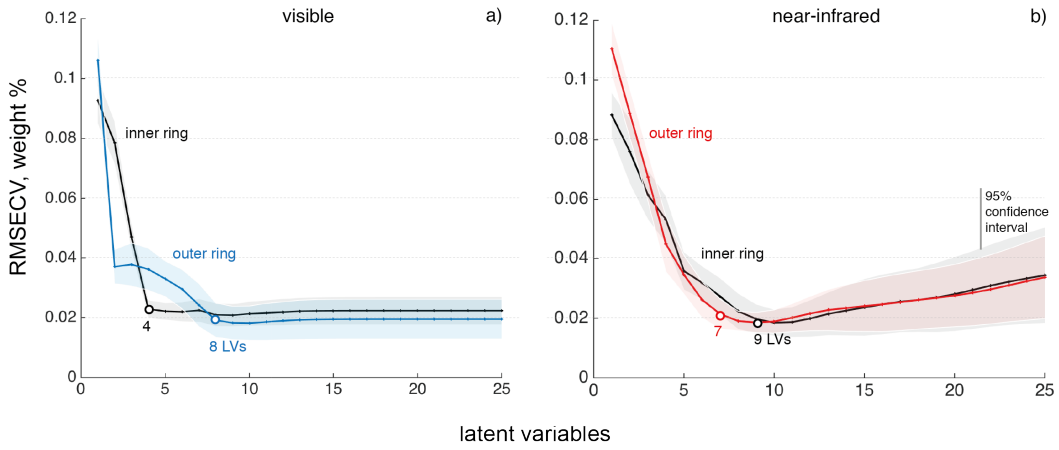

Figure 7.3 Performance of the a) Vis and b) NIR sensors on each illumination ring of the dual-ring reflectance probe. RMSECV values obtained by the first 25 latent variables from rdCV PLS models for estimating bead concentration.

Because the experiment under study in this chapter consists of water and polystyrene only, we would expect that no more than two latent variables would be needed to explain the variance in the data. However, since the polystyrene beads absorb and scatter light, it requires extra latent variables due to the nonlinear effects of multiple light scattering. Moreover, as the size and number density of beads - or a combination of both - influence differently the concentration of particulate species, more latent variables may be needed to account for this complex interplay. As visible spectroscopy can be more affected by scattering, and this being a result of the change in physical traits, this may explain why the Visible inner ring was more efficient in

concentrating on the first latent variables the direct or indirect information on the concentration of polystyrene in the suspension.

Another important reason for requiring extra latent variables may be the influence that water absorption has on the models. By inspecting the loadings plot for the first four latent variables for both NIR inner and outer rings (Figure 7.4), we can see that the main water absorption band around 1450 nm is present in the loading spectra for all latent variables.

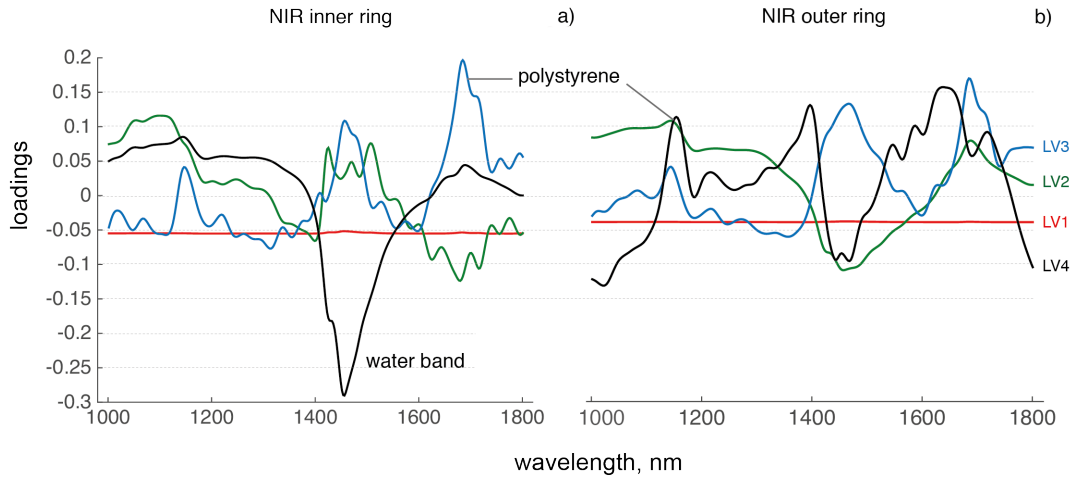

Figure 7.4 Loading curves for the four first latent variables obtained by NIR models developed on the a) inner and b) outer rings.

The first latent is similar for both rings, presenting little influence on both water and polystyrene absorption. Please note that although resembling a flat line, they present features related to water and polystyrene, but smaller when compared to the further latent variables. The outer ring seems to capture more information on the polystyrene concentration, which can be observed by the positive loadings for the first latent variables around the 1680 nm region, due to the C-H stretching vibrations of polystyrene. In the inner ring, the LV 3 captures more information about polystyrene, whereas the latent variable 4 has strong negative loadings in the water region.

The predictive capability of NIR models was equivalent, although the outer ring needed a lesser number of components and provided a slightly smoother RMSECV curve (Figure 7.3 b). Since the intensity of light shone through the outer ring is twice of that which leaves the inner ring, we would expect measurements from the outer ring to have better signal-to-noise ratio, and consequently, to provide better models. However, because the outer ring is located at a farther distance to the collecting fibre, more scattering and absorption events may take place, influencing the amount of light that reaches the detector. These two effects may explain why both rings provided similar results in terms of accuracy.

As shown in Figure 7.2, Vis and NIR spectra collected from both rings can be combined together in four different ways. In the Table 7.3, we present the results obtained by combining these ring combinations by using the multi-block method. When fusing the same spectroscopic data on both rings, the co-adding approach was also used and the results were compared to the multi-block ones. The results obtained by the dual-ring diffuse reflectance probe are compared to the best models developed by using the SARDR probe, as discussed in Chapter 5. These results are highlighted in light grey coloured cells.

Table 7.3 Summary of results of PLS models based on the MB-PLS and co-adding approaches, for the determination of EPS bead concentration in water by using the dual reflectance probe in the visible and NIR ranges. The best results obtained by the SARDR probe are also presented (gray cells) for comparison.

Model	PLS Repeated Double CV					
	Fusion approach	**nb**	**RMSECV** % w/w ± CI95	**RMSEP** % w/w ± CI95	**R^2**	**LV**
NIR inner/outer	multi-block	2	1.6 ± 0.02	1.5 ± 0.04	0.99	9
NIR inner/outer	co-adding	1	1.4 ± 0.01	1.3 ± 0.03	0.99	7
Vis inner/outer	multi-block	2	1.3 ± 0.02	1.2 ± 0.03	0.99	10
Vis inner/outer	co-adding	1	2.0 ± 0.01	2.0 ± 0.03	0.95	5
NIR-inner / Vis-outer	multi-block	2	1.9 ± 0.03	1.7 ± 0.05	0.99	9
Vis-inner / NIR-outer	multi-block	2	1.7 ± 0.02	1.6 ± 0.04	0.99	10
Vis-0.6/3.0/3.0A	multi-block	3	2.3 ± 0.05	2.3 ± 0.05	0.96	10
0.6A/0.6 NIR	multi-block	2	4.0 ± 0.03	4.1 ± 0.08	0.94	6

Bead concentration range: 15 – 50 % w/w

The first conclusion we can draw from Table 7.3 is that the dual reflectance probe presented significantly better predictive capability than the SARDR probe. The prediction and calibration errors obtained by the SARDR best models for Vis and NIR (including combinations) were significantly higher than the ones obtained by any ring combination of the dual reflectance probe.

The second main finding is that the fusion models built by using the co-adding approach were simpler than the ones built by multi-block. The number of latent variables needed to describe the data decreased from 9 to 7 in the NIR models, while decreasing by half in the Vis based models, as it is also shown in the RMSECV scree plots of Figure 7.5.

In the NIR fusion model, the co-added model obtained estimations for bead concentration with better accuracy and precision, obtaining an average cross validation error of 1.4 % + 0.01, while the multi-block model estimates had an average error of 1.6 % + 0.02. However, for the fused Vis based model, the co-adding models had errors much higher than the ones obtained by the multi-block approach, and also higher than the models based on individual rings.

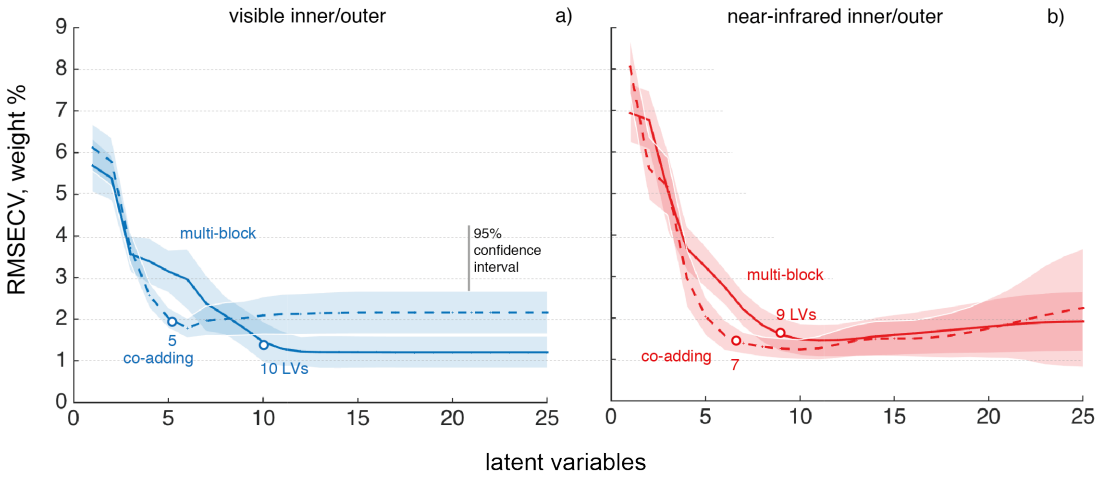

Figure 7.5 Performance of the dual-ring reflectance probe being used as a single-sensor probe for a) Vis and b) NIR. RMSECV values obtained by the first 25 latent variables from rdCV MB-PLS and co-adding PLS models for estimating bead concentration. Solid lines represent the multi-block RMSECV curves whereas dotted lines the co-adding approach.

The results in Figure 7.5 shows that the co-adding approach was better in capturing most of the relevant information in the first latent variables than the multi-block approach. However, for Vis based models, the use of more latent variables was justified for a better accuracy on multi-block models.

Figure 7.6 presents the RMSECV scree plot obtained by all models developed based on the dual-ring fibre. In Figure 7.6 a, we describe models developed when the probe has multi-spectroscopies, whilst in Figure 7.6 b, only one spectroscopy is used in both rings. For this case, both co-adding and multi-block curves are shown. For

comparing the performance of fusion models against the individual ones, the RMSECV curves obtained by the individual rings are also shown.

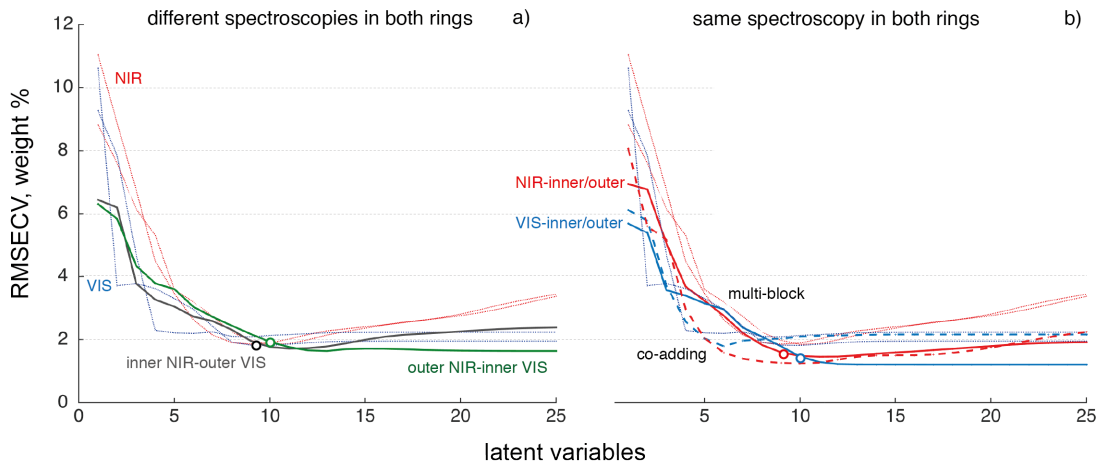

Figure 7.6 Performance of the dual-ring reflectance probe being used as a a) multi-sensor probe and as a b) single-sensor probe, being compared to individual ring models (light dotted line). RMSECV values obtained by the first 25 latent variables from rdCV MB-PLS and co-adding PLS models for estimating bead concentration. Solid lines represent the multi-block RMSECV curves whereas dotted bold lines the co-adding approach.

By analysing Figure 7.6 together with Table 7.3, it can be noted that the integration of spectra from both rings improved the performance of the models. The errors in the first latent variables decreased considerably, meaning that captured more useful information than the single ring models. For the multi-block models, although the errors obtained by using the same spectroscopy were lower than using different spectroscopies in both rings, the overall trend of the RMSECV was not remarkably different, especially in the first latent variables, which concentrates most of the useful information. However, for the co-adding models, the first latent variables captured more direct or indirect information about the polystyrene concentration.

7.3.2 Determination of EPS bead size

Similarly to the previous section, the results obtained by detecting light by Vis and NIR spectroscopy from each ring independently are presented. Table 7.4 summarises the results for the determination of mean bead size.

Table 7.4 Summary of results of PLS regression models for the determination of EPS bead size by using the dual reflectance probe in the visible and NIR ranges, and the best results from the SARDR probe.

Model	PLS Repeated Double CV					
	fusion approach	**nb**	**RMSECV** mm ± CI 95%	**RMSEP** mm ± CI 95%	**R^2**	**LV**
Vis inner	-	1	0.091 ± 0.001	0.086 ± 0.002	0.97	7
Vis outer	-	1	0.054 ± 0.001	0.053 ± 0.001	0.99	6
NIR inner	-	1	0.056 ± 0.001	0.056 ± 0.001	0.99	7
NIR outer	-	1	0.046 ± 0.001	0.045 ± 0.001	0.99	7
NIR inner/outer	multiblock	2	0.033 ± 0.001	0.030 ± 0.001	0.99	10
NIR inner/outer	co-adding	1	0.042 ± 0.001	0.042 ± 0.001	0.99	7
Vis inner/outer	multiblock	2	0.032 ± 0.001	0.031 ± 0.001	0.99	10
Vis inner/outer	co-adding	1	0.065 ± 0.001	0.064 ± 0.001	0.99	8
NIR-inner / Vis-outer	multiblock	2	0.039 ± 0.001	0.039 ± 0.001	0.99	9
Vis-inner / NIR-outer	multiblock	2	0.025 ± 0.001	0.023 ± 0.001	0.99	11
Vis-3.0/1.8/0.6	multiblock	3	0.050 ± 0.001	0.048 ± 0.001	0.99	11
NIR-0.6A/0.6/1.2	multiblock	1	0.081 ± 0.001	0.076 ± 0.002	0.99	10

Bead size range: 0.405 to 1.540 mm

Again, the accuracy from the models developed by using the dual-ring reflectance probe was superior to the one obtained by the SARDR probe. The outer ring provided the best accuracy among the dual reflectance probe based models, obtaining

errors around 0.050 mm, and needing 6 and 7 latent variables to describe the data, for VIS and NIR, respectively. The performance of the models are better illustrated in Figure 7.7, which presents the RMSECV scree plot obtained for all individual ring models.

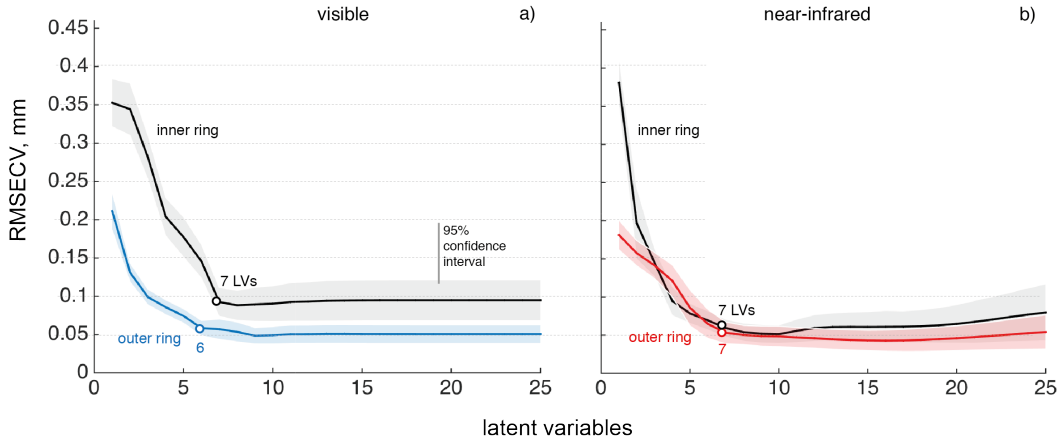

Figure 7.7 Performance of individual sensors on each illumination ring. RMSECV values along 25 latent variables obtained by the rdCV models for estimating bead size by using individual illuminating rings for (a) visible and (b) near-infrared spectroscopy based on the dual reflectance probe.

It is clear from Figure 7.7 a and b that the inner ring provided poorer models, especially for the visible spectroscopy. For the NIR models, the first latent variables had a much higher error, but equalised the outer ring when the model selected 7 latent variables. Again, for all models, a relatively high number of latent variables were needed to describe the data variance.

The results improved when spectra collected from both rings are combined, either by using the same spectroscopy mode in each ring, or alternating them in both rings. Whereas the RMSECV for the best model using one-ring-one-spectroscopy was 0.046 mm, as obtained by the NIR outer rings, when we combine them, the errors dropped to 0.025 mm for the best case, reaching 0.039 mm in the worst case.

When we compare the performance of multi-block and co-adding one-probe-one-spectroscopy models, it is clear that the multi-block provided better accuracy for the Visible models. For the NIR models, the overall RMSECV curve for the co-adding method was smoother, despite its first latent variable accounting for less information on mean bead size. In fact, equalising the number of latent variables to the same chosen for multi-block, i.e., 10 latent variables, the performance would be similar. Overall, the Visible model built by multi-block gave similar results than the co-adding NIR model.

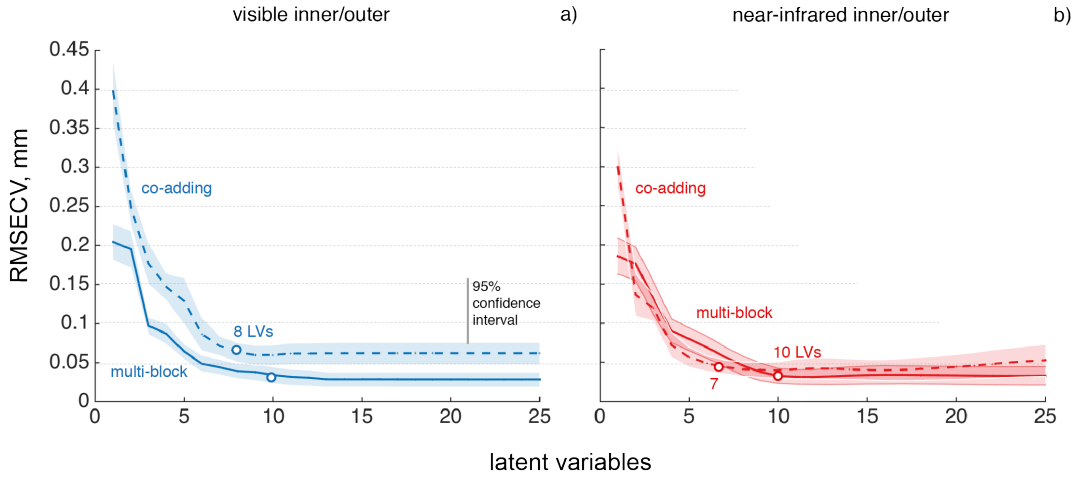

Figure 7.8 Performance of the dual-ring reflectance probe being used as a single-sensor probe for a) Vis and b) NIR. RMSECV values obtained by the first 25 latent variables from rdCV MB-PLS and co-adding PLS models for estimating mean bead size. Solid lines represent the multi-block RMSECV curves whereas dotted lines the co-adding approach.

The RMSECV scree plots including the fusion models exemplifies better the improvement, as shown in Figure 7.9 a and b for different spectroscopies in both rings, and same spectroscopy in both rings, respectively.

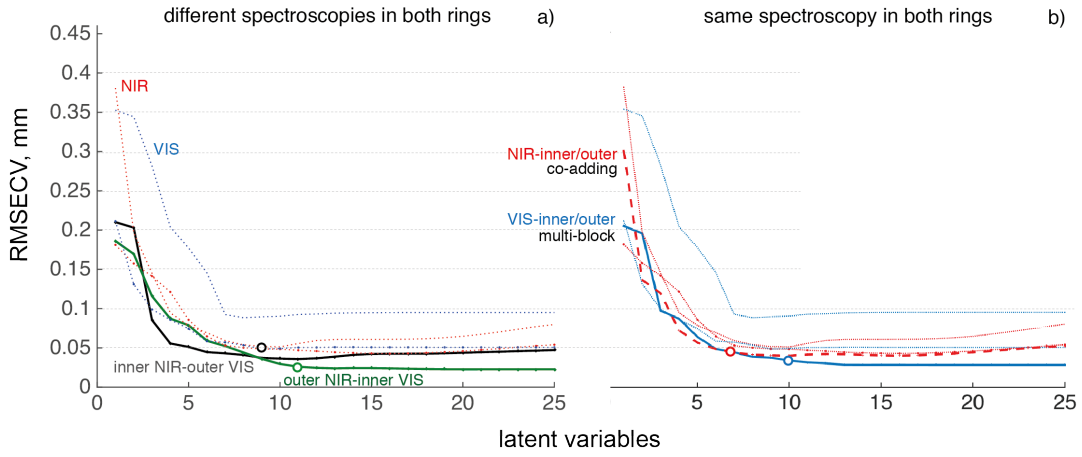

Figure 7.9 Performance of the dual-ring reflectance probe being used as a a) multi-sensor probe and as a b) single-sensor probe, being compared to individual ring models (light dotted line). RMSECV values obtained by the first 25 latent variables from rdCV MB-PLS and co-adding PLS models for estimating mean bead size. Solid lines represent the multi-block RMSECV curves whereas dotted bold lines the co-adding approach.

Interestingly, all RMSECV curves from fused multi-sensor probe mode crossed each other on the ninth latent variable, but followed different trajectories. The combination made by NIR in the inner ring and Vis in the outer ring provided the lowest errors until the ninth latent variable, but maintained the same error level after it, while the errors continued to decrease for other combinations. Although the latent variable selection criterion has chosen 9 LV for this particularly combination, 6 latent variables would suffice. In fact, almost 24 % of the 100 repetitions of the double cross validation models selected 6 latent variables.

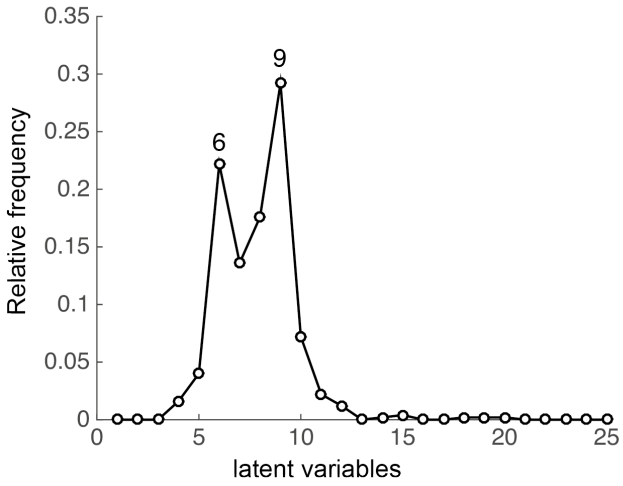

Figure 7.10 Relative frequency of the optimum number of latent variables obtained from 100 repetitions of the double cross validation.

When the ring order was inverted, i.e., NIR spectra were collected from the outer ring whereas Vis spectra from the inner ring, the errors dropped to 0.025 mm. However, for this model, eleven latent variables were required. Figure 7.11 a and b presents the measured versus estimated bead size plots for these two combinations.

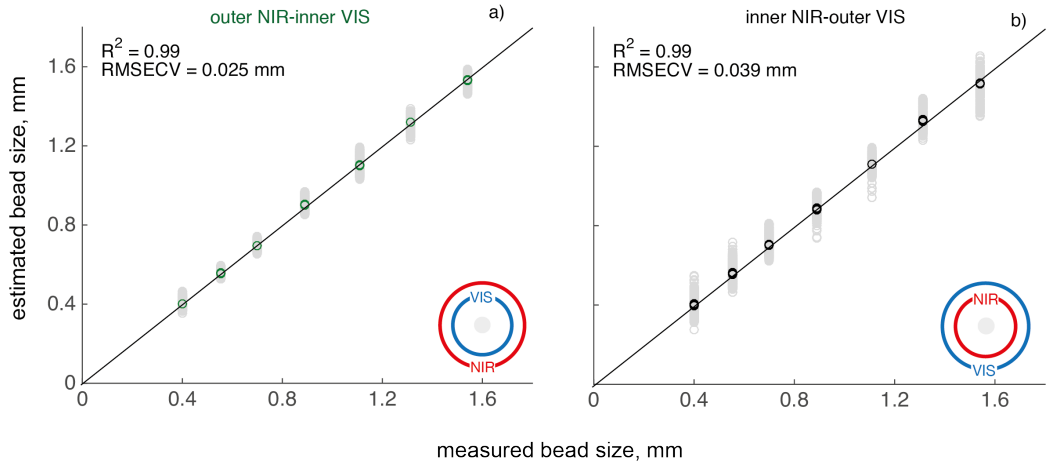

Figure 7.11 Measured versus estimated bead sizes obtained from fused models between (a) outer NIR spectra and inner Visible spectra and (b) inner NIR spectra and outer Visible spectra. Grey coloured circles are all the individual measurements, whereas the coloured circles are the mean values obtained by 100 models.

Although better fit was obtained by the combination of outer NIR and inner Visible spectra is clear, it required more latent variables to describe the variance in the data. Since the difference in the errors may be not significant, and based on the RMSECV curves from Figure 7.9, we conclude that the combination of inner NIR/Outer Visible was a preferable choice for bead size determination.

7.3.3 Best results

For the determination of bead concentration, the co-added NIR spectra obtained from inner and outer rings results in the best model, considering the better parsimony and low errors. It provided estimations with an average error of 1.4 %, needing only 7 latent variables to describe the variance contained in the data.

For mean bead size determination, all models provided excellent accuracy. By evaluating their parsimony, and analysing the RMSECV scree plot, the inner NIR/Outer Visible was the preferable combination. However, by using the probe as a single-sensor Vis probe, comparable results were also obtained. Due to fast measurements, the probe can be used to acquire NIR and Visible spectra in both rings for optimum results on both bead size and concentration.

7.4 Conclusions and Outlook

The goal of this chapter was to evaluate whether improving light throughput would enhance the accuracy of multivariate calibration models based on Vis and NIR spectroscopy. For this, an alternative optical probe design was presented for the estimation of polymer properties. PLS based models were developed on the spectra collected from the light shone through each of the illuminating rings, first independently, and later, fused as a single-sensor or multi-sensor probe.

When compared to the SARDR probe, the results obtained in this chapter show that the predictive capability of models can be largely improved by enhancing the quality of the measurements. The findings here confirm that the poor signal-to-noise ratio of the NIR SARDR measurements compromised their predictive capability, limiting their potential on estimating bead concentration, a property for which NIR is normally a *par excellence* choice. When stronger light throughput is available, which is the case of the multi-fibre rings of the dual-ring reflectance probe, the NIR measurements resulted in models with excellent predictive power, surpassing the Visible models. Overall, the performance of the models was influenced by different combinations of sensors and rings used, as well as by the data fusion approach, especially concerned to the parsimony.

Combinations of rings/sensors Both single-sensor or multi-sensor use of the dual-ring diffuse reflectance probe provided reliable estimates of bead size and concentration. The accuracy of the models increases by combining information from both illumination ring and or both sensors.

Data fusion approaches Data fusion by using the co-adding approach outperformed the multi-block approach only in the NIR models. For Visible models, the multi-block provided improved results. In terms of parsimony, the co-adding approach provided models needing less latent variables for both Visible and NIR models.

As the probe is still in early days of development, it was not possible to use it on polymerisation reactions, owing to incompatibility between the reactions constituents and the adhesive used to attach the window to the probe head. However, since the models developed by using SARDR measurements were up scaled very well from the simple two-component system to the complex and heterogeneous full polymerisation reactions, it is expected that the dual-ring reflectance probe would perform even better on the polymerisation system.

7.5 Bibliographic references

1. Mahony, N. O., Murphy, T., Panduru, K., Riordan, D. & Walsh, J. Smart sensors for process analytical technology. in 1005–1010 (IEEE, 2016). doi:10.1109/AIM.2016.7576901

2. Bakeev, K. A. & Menezes, J. C. in *Process Analytical Technology* **29,** 521–543 (John Wiley & Sons, Ltd, 2010).

3. Rantanen, J. & Khinast, J. The Future of Pharmaceutical Manufacturing Sciences. *Journal of Pharmaceutical Sciences* **104,** 3612–3638 (2015).

4. Filzmoser, P., Liebmann, B. & Varmuza, K. Repeated double cross validation. *J. Chemometrics* **23,** 160–171 (2009).

Chapter 8

8 Conclusions and Future Work

This thesis proposed for the first time the use of spatially and angularly resolved diffuse reflectance spectroscopy as a PAT tool for *in-situ* monitoring of suspension polymerisation reactions. For this, multivariate calibration and data fusion modelling strategies were also proposed for optimum extraction, interpretation, visualisation, fusing and modelling of such multi-dimensional spectroscopic information

The proposed SARDR probe, measurement system and modelling strategies were experimentally evaluated by using a two-component system composed of polystyrene beads suspended in water at varying concentrations and mean bead sizes (Chapter 4 and 5), and by full batch styrene suspension polymerisation reactions (Chapter 6). In Chapter 7, a prototype multimode probe based on a dual ring geometry was also proposed and evaluated on the two-component system, and its performance compared to the one obtained by the SARDR probe.

Throughout this thesis, by using these two experimental systems, I aimed to investigate to what extent the predictive capability of multivariate calibration models is affected by different wavelength regions, source-to-detector distances, angles of incidence, rings of fibres, and probe geometries. But mainly, I was interested in whether combining such complementary information could improve the estimation of chemical and physical properties from polymer products.

8.1 Individual measurements and their integration

The benefits of integrating complementary spectroscopic information were clearly demonstrated. Compared to the best predictive models developed based on individual measurements, data fusion models have shown improved predictive capability, either by integrating different (i) source-detectors distances; (ii) angular incidences; (iii) spectroscopic regions; or (iv) rings of fibres.

i) Source-detector separations

Source-detector separation largely influences the quality and amount of information captured by diffuse reflectance measurements, especially in the NIR wavelength range. Although longer SD distances tend to capture richer information coming from deeper layers in the sample, the best NIR models were obtained by using shorter SD distances. While these models may have been aided by any inherent information contained in such shorter pathlength, the main reason for their improved performance may reside in the quality of the measurements. Although 400 μm fibres were employed in the SARDR probe, NIR light throughput from farther illuminating fibres was not strong enough to reach the detecting fibre with sufficient signal-to-noise ratio, affecting negatively the sensitivities of the acquired spectra readings, as demonstrated in Chapter 4. This finding is also corroborated by experiments from Chapter 7, where considerably lower errors of prediction were obtained by using stronger illumination.

It is important to remember from the experimental description in Chapter 3, that the integrating time needed to record a SARDR NIR spectrum was 80 to 100 times higher than the one for a visible spectrum. Consequently, the NIR signal obtained before normalisation was very noisy. The background spectrum acquired by using the integrating sphere was even poorer, which contributed to the low quality of the reflectance signals obtained after normalisation. Hence, in the visible range the

impact of SD distance was much weaker, although longer SD separations tended to give simpler models and lower prediction errors.

Overall, the predictive capability of Vis and NIR models was improved upon fusing spectra from more than one normal incident fibre. The main effect was achieved by combining spectra from two or three SD distances, with Visible spectroscopy being the one that benefited the most. By fusing the spectra from three fibres (SDD = 0.6, 1.8 and 3.0 mm), errors dropped for both mean bead size and concentration determination. For NIR, bead size estimation was improved by using the two closest fibres (SDD = 0.6 and 1.2 mm), but no improvement was obtained for bead concentration. When monomer conversion was estimated along full suspension polymerisation reactions, Visible calibration models benefited by fusing information from different SD distances, although they were not able to successfully estimate monomer conversion from external reactions. Unlike Visible, NIR fused models were able to provide improved estimations of conversion from external reactions, although not presenting advantages in the rdCV calibration and prediction phases, when compared to the individual SD distance.

ii) Individual angles of incidence

Angularly incident fibres, when modelled individually, did not surpass the predictive capability of normal incident fibres. Relative to their respective SD separation, the results follow the ones obtained by using normal incident fibres, with NIR performing better at shorter SD distances (0.6 mm), whereas Vis at longer distances (SDD=4.2 mm). In this case, the fibre with the incidence 30° to the normal provided the best results for both Visible and NIR.

The inclusion of the 30° angularly incident fibre to the combinations of normal incident ones largely improved the predictive capability for most of the models in both two-component system and polymerisation reactions, regardless the adopted

data fusion approach, spectroscopy method and pre-processing technique. The inclusion of the 45° angularly incident fibre did not bring any improvements, especially due to its wider SD distance (4.2mm) for the NIR.

iii) Spectroscopic regions

Predictive models based on Visible spectra were better when the only factor changing was the proportion of beads to water. For most of the SD distances, models based on visible spectra were able to yield estimates with considerably lower errors, tighter residual distributions, and a lower number of latent variables to describe the variability in the data. However, when such models were developed and evaluated in a highly heterogeneous system containing, with more chemical and physical variations coming into play, NIR models were able to better capture information from such variations related to monomer conversion.

In the dual reflectance probe, which was only evaluated in the two-component system, NIR and Visible models had more comparable performance, with the Visible ones providing similar or slightly better predictive capability than NIR models.

iv) Rings of fibres (dual-reflectance probe)

Both single-sensor and multi-sensor use of the dual-ring diffuse reflectance probe provides reliable estimates of bead size and concentration. The accuracy of the models increases by combining information from illumination rings or spectroscopic modes.

When modelled independently, Visible spectra acquired by using the outer ring yielded better estimations than the inner ring. For NIR, both rings obtained similar results.

8.2 Data fusion approaches

The ways in which SRS data can be combined were investigated in their impact on the predictive capability and parsimony of the models. For this, the performance of the multi-block and the co-adding approaches were compared in Chapters 6 and 7. Although the co-adding approach outperformed the multi-block approach in a few cases, in general, multi-block based models provided better accuracy. In terms of parsimony, however, the co-adding models required less latent variables.

The multi-block approach was employed here instead of the data augmentation approach since it allows a better model interpretation. However, in cases where the only purpose is predictive modelling, data augmentation can be used as a simpler alternative since it gives similar results than the multi-block approach.

8.3 Selecting the best model complexity

An investigation was performed to evaluate the different approaches for selecting the optimum model complexity in the scope of the repeated double cross validation. Different parsimony levels were studied in the one-standard deviation rule, which were compared to the Haaland Thomas procedure. It was found that in the one-standard deviation rule, as proposed by Filzmoser et al, the choice of the optimum number of latent variables was too conservative, leading to underfitting. The results were improved by relaxing the parsimony factor to 0.25 SD, but required careful choice, since lowering it too much may cause overfitting and a selection of a higher number of latent variables. In fact, this is the case when using the minimum MSE. The Haaland Thomas criterion obtained similar fit in both calibration and external validations, selecting an acceptable number of latent variables.

8.4 Impact of empirical scatter correction

The advantage of pre-processing was clear in the rdCV results. However, most of the pre-processed models were not able to estimate monomer conversion from the external reaction, especially for the reactions performed before the major probe maintenance. While their performance was good, this may indicate that the pre-processing tools were not robust enough to discern or correct for the undesired effects in the spectra in the presence of different levels of baseline effects.

8.5 Recommendations for Future Work

In view of the conclusion that the signal-to-noise ratio of NIR measurements was a limiting factor of the predictive capability of the NIR models, further research could be done in the following aspects:

- **Optimisation of the probe geometry:** as it was demonstrated that the addition of an angularly incident fibre can improve the predictive power of PLS models, a new probe geometry could be developed by adding several angularly incident fibres at the same SD distance, as well as normal incident fibres, increasing the intensity of light that reaches a central collecting fibre. The fibres could be placed in a cross geometry, allowing four illuminating fibres for each SD distance, as the example presented in Figure 8.1.

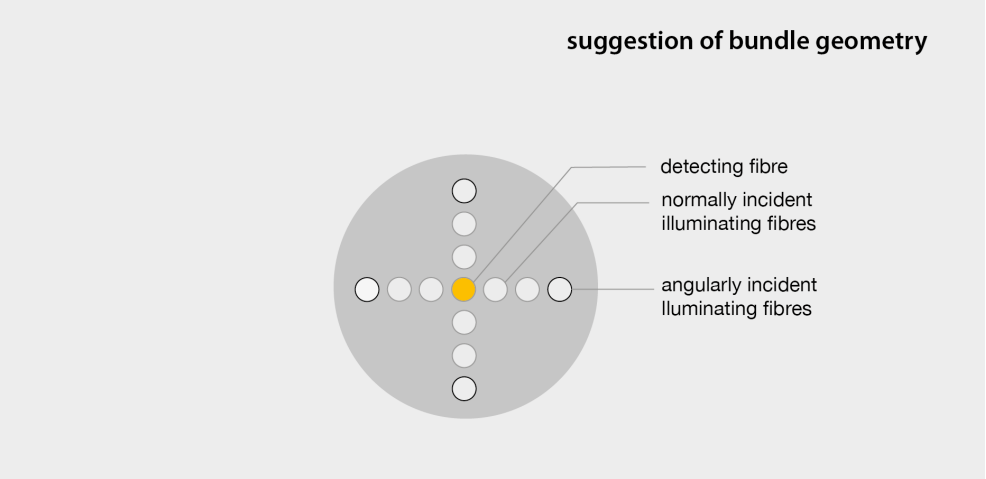

Figure 8.1 Schematic illustration of a potential geometry for the optical fibre bundle.

- **Optimisation of the optimum integrating time selection:** In this thesis, integrating times for spectral collection were defined at the beginning of the experiment. As the spectral intensities change considerably throughout the reaction, the selected integrating time can be too low or too high at certain samples, decreasing the signal-to-noise ratio or causing saturation. In the present experiment, the instances of saturation were manually corrected by repeating the measurement at a lower integrating time. An automatic procedure for updating the integrating time is important to obtain the best possible measurement during the whole reaction.

Bead size monitoring. Considering the good results obtained in the two-component system for the determination of mean bead size, there is a large potential for monitoring mean bead size in suspension polymerisation reactions.

How general is the approach? The experiments for this thesis were carried out at a laboratory scale. Further investigation on the scale-up for spectroscopic models could be performed to assess their performance in a pilot plant.

The proposed technique can be extended and applied to other polymerisation reactions. For example, it would be interesting to see how the spectroscopic system and related modelling strategies will work in the nano scale, which is the size scale of the particles produced by emulsion polymerisation reactions. Applications can be developed for other industries, as an example of the food and pharmaceutical particulate systems.

Extraction of absorption and scattering coefficients. As the efficacy of the scatter correction methods was limited in the results obtained in this thesis, it would be interesting to evaluate how models based on first principles would cope with the polymerisation spectral data. The extraction of absorption and scattering coefficients, and their use for predictive modelling of monomer conversion and mean bead size, respectively, would be an important addition to the research.

Appendix

Table A1 Prediction performance of MB-PLS models for determination of mean EPS bead diameter from fused NIR-SARDR spectra collected at different normal incident source-detector distances and from a 0.6 mm angularly incident fibre. RMSECV and RMSEP were obtained from rdCV models.

PLS Repeated Double CV					
SD Distance mm	**nb****	**RMSECV** % w/w CI95%	**RMSEP** % w/w CI95%	**R^2**	**LV**
0.6-30° fused to 0.6	2	0.091 ±0.001	0.100 ±0.003	0.99	8
and 1.2	3	0.081 ±0.001	0.076 ±0.002	0.99	10
and 1.8	4	0.092 ±0.001	0.088 ±0.003	0.99	13
and 2.4	5	0.092 ±0.001	0.091 ±0.002	0.99	11
and 3.0	6	0.105 ±0.001	0.106 ±0.003	0.98	9

Bead concentration range: 15 – 50 % w/w
*SD distance = 0.6 mm at 30° to the normal
**nb= number of blocks

Table A2 Prediction performance of MB-PLS models for determination of mean bead size from fused Visible-SARDR spectra collected at different normal incident source-detector distances and from a 4.2 mm 30° angularly incident fibre. RMSECV and RMSEP were obtained from rdCV models.

PLS Repeated Double CV

SD Distance mm	**nb**	**RMSECV** % w/w ± CI95%	**RMSEP** % w/w ± CI95%	**R^2**	**LV**
3.0	1	0.077 ±0.001	0.076 ±0.002	0.97	4
4.2 30°	1	0.151 ±0.001	0.155 ±0.003	0.96	7
4.2 30° / 3.0	2	0.077 ±0.001	0.076 ±0.002	0.98	7
4.2 30° / 3.0 / 2.4	3	0.071 ±0.001	0.070 ±0.002	0.99	8
4.2 30° / 3.0 / 2.4 / 1.8	4	0.077 ±0.001	0.075 ±0.002	0.99	10
4.2 30° / 3.0 / 2.4 / 1.8 / 1.2	5	0.087 ±0.001	0.084 ±0.003	0.99	13
4.2 30° / 3.0 / 2.4 / 1.8 / 1.2 / 0.6	6	0.069 ±0.001	0.063 ±0.001	0.99	15
4.2 30° / 3.0 / 1.8 / 0.6	4	0.063 ±0.001	0.061 ±0.001	0.99	11
4.2 30° / 3.0 / 0.6	3	0.057 ±0.001	0.057 ±0.001	0.99	9

Mean bead diameter range: 0.405 to 1.540 mm