FLOW INJECTION AND MULTIVARIATE CALIBRATION TECHNIQUES FOR PROCESS ANALYSIS

being a thesis submitted for the degree of

DOCTOR OF PHILOSOPHY

Department of Environmental Sciences
University of Plymouth

in collaboration with
ICI Chemicals & Polymers

by

PAUL MACLAURIN

August 1993

ABSTRACT

FLOW INJECTION AND MULTIVARIATE CALIBRATION TECHNIQUES FOR PROCESS ANALYSIS

PAUL MACLAURIN

The role of process analytical chemistry is summarised in chapter one with particular emphasis on a multidisciplinary approach and the instrumental requirements for on-plant analysis. These concepts are extended to process FIA, highlighting its potential for simultaneous multicomponent determinations.

The development of an automated FIA monitor for the on-line determination of sulphite in potassium chloride brine is covered in the second chapter. Reaction stability is demonstrated and the results of on-plant validation and on-line trials are presented.

The next chapter deals with the concepts of multivariate calibration. Direct multicomponent analysis, principal components regression and partial least squares regression are critically examined in practical spectroscopic terms and statistical terms. The relative predictive abilities of these techniques are compared in chapter four for the resolution of a multicomponent UV-visible spectrophotometric data set.

Chapter five describes the development of an automated FIA-diode array system for the simultaneous determination of phosphate and chlorine. The implications of combining reaction chemistries and the influence of a number of calibration parameters are considered in detail.

Finally, the jackknife is presented as a means of dimensionality estimation and bias correction in PLS modelling. Data sets from the literature are analysed and the results compared with those obtaining using commercial software.

ACKNOWLEDGEMENTS

I would like to thank Paul Worsfold for his guidance, foresight and patience throughout the course of this work. His invaluable ability to distil the worthwhile from the mediocre is attributable to his faithful support of AFC Bournemouth.

My thanks also to Alan Townshend for his advice at the beginning of this project.

I am grateful to many people in various parts of ICI for their extensive support. I would particularly like to thank the infamous trio of Phil Norman, Mike Crane and Neil Barnett (the legend lives on).

I gratefully acknowledge financial support from the SERC, ICI Chemicals & Polymers and Mike Crane's cost centre.

I have been helped & hindered; aided & abetted; cajoled & corrupted; and inspired & intoxicated by a super crowd during my time at both Hull and Plymouth. I warmly thank you all, especially Aardvark, Ant & Ant & Ant, Beno, Blox, DBA, Dr Nick, Eugene, Fishcake, Gland, Jimbo, Kev'n'Trev, Lancs, Lottie, Moo, Rev, Tiny, Wavey, Whiplash, Wilko and Yan.

To Mum & Dad

AUTHOR'S DECLARATION

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award.

This study was financed with the aid of a studentship from the Science and Education Research Council, and carried out in collaboration with ICI Chemicals & Polymers under the CASE scheme.

A programme of advanced study was undertaken, which included plant and mammalian physiology, cytogenetics and ecology.

Relevant scientific seminars and conferences were regularly attended at which work was often presented; external institutions were visited for consultation purposes, and several papers were prepared for publication.

Publications:

Analytica Chimica Acta, 1990, **238**, 171.

Analyst, 1991, **116**, 701.

Analytical Proceedings, 1992, **29**, 65.

Michrochemical Journal, 1992, **45**, 178.

Analytical Proceedings, 1993, **30**, 143.

Analyst, 1993, **118**, 617.

Presentations:

Anatech '90, Noordwijkerhout, The Netherlands.

Flow Analysis V, Kumamoto, Japan.

R & D Topics, Birmingham, UK.

SAC '92, Reading, UK.

Euroanalysis VIII, Edinburgh, UK

Signed.....

Date.....20-9-93.....

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter One

Introduction

1.1 PROCESS ANALYTICAL CHEMISTRY

Interest in process analytical chemistry (PAC) has grown considerably in recent years, and with developments in instrumentation and procedures for on-line analysis, the number of reported process applications has risen accordingly. The desire to acquire information of a chemical nature about a manufacturing process can be attributed to economic and environmental reasons. The economic reasons are related to product quality and the optimal use of raw materials, labour, energy and time. The environmental reasons include concern for occupational hygiene, emission control and the wider environment [1,2,3]. The development of PAC is also demonstrated by the diversity of applications, the numerous measurement techniques that are now being used [4,5] and the launch of a journal devoted to the area [6].

Process control, the domain of chemical and process engineers, has traditionally relied on the measurement of physical parameters such as pressure, temperature and viscosity supplemented by the occasional chemical measurement; pH for example. More complicated chemical analysis would be carried out in centralised laboratory facilities by teams of highly trained technicians using expensive multi-tasking equipment. This would cover raw material testing, final product certification and intermediate stage analysis for non-continuous processes. Within such a framework, samples are logged and stored until a sufficient number has accumulated to warrant carrying out a particular analytical procedure. The delay between sampling and the communication of a result can therefore run to a number of days and often represents a "post-mortem" rather than an interactive approach.

In order to achieve the level of control needed to meet the increasingly high standards required in today's chemical manufacturing industry, much closer attention to chemical composition has to be maintained. To enable the collection of chemical information about dynamic chemical processes

fundamental shifts in the philosophy underlying analytical procedures have resulted in PAC. PAC has developed as a sub-discipline of analytical chemistry but in practice it requires a multi-disciplinary problem-orientated approach [7].

The development of PAC has addressed the following issues;

1. location of analysis,
2. analysis time,
3. dependability,
4. cost.

One of the first steps towards reducing the delay between sampling and the generation of analytical information is to transfer analysis from the laboratory to the plant. The so-called "at-line" approach involves the installation of a laboratory instrument close to the process sampling point. Such instrumentation is generally less sophisticated and therefore less expensive than the instrumentation of a centralised facility and more amenable to operation by the process personnel. Information could then be accumulated quickly and with a greater frequency, thus providing the process control staff with a better picture of system performance.

Laboratory based analysis has benefited enormously from the automation of instrumentation and procedures [8], and it is the automation of process sampling and analysis that distinguishes the "on-line" approach. The advantages of laboratory automation such as increased sample throughput and improved precision apply equally well to process monitoring and with the reduced analysis time associated with the process location, near real time chemical analysis is feasible.

The on-line monitoring of process streams requires more than moving the laboratory equipment to the plant however [9]. If an analytical procedure is to provide information for process control then the entire system, from sampling

to communication of results, needs to be dependable. One of the most difficult aspects of developing an on-line monitor is the provision of an automated sampling system that will provide representative samples for extended periods of time. This tends to remain the responsibility of the process engineer but no less trivial is the development of rugged analysers suitable for long-term unattended operation. The analytical performance characteristics such as selectivity, sensitivity, accuracy and precision need to be addressed with the application in mind but the corrosivity of the atmosphere and other matrix effects of the sample also need to be considered.

The cost of developing, installing, operating and maintaining such systems needs to be kept in the perspective of the value of collecting the process information. Savings can be made in terms of raw materials and labour but primarily non-financial justifications such as safety must also be considered. In addition, on-line analysers can be deployed in remote or hazardous locations and can operate on a 24-hour basis.

Various spectroscopic approaches have been taken for on-line analysis. After the necessary modifications have been made to the instrumentation and a suitable sampling system installed, ultraviolet-visible (UV-vis), near-infrared (NIR), mid-IR and FT-IR instruments can be deployed for continuous monitoring; in fact, commercial process NIR systems are already available [10]. Moreover, apparently inappropriate techniques such as mass spectrometry, nuclear magnetic resonance spectroscopy and X-ray fluorescence spectroscopy have been developed for use in particularly problematic applications. It is often the case, however, that some kind of physico-chemical or mathematical selectivity enhancement must be incorporated into the procedure to ensure reliable results. High performance liquid chromatography (HPLC), for example, for UV-vis detection and multivariate calibration routines for NIR analysis.

Continuous monitoring systems [11] permit a derivatisation stage to be included prior to detection without disturbing the continuous output. This is in contrast to chromatographic techniques which immediately preclude continuous analysis. The frequency of the intermittent output is dependent on the speed of the chromatography; some HPLC separations can be quite time consuming whereas gas chromatography can, for some applications, produce a very rapid response. Continuous flow analysis (CFA) [11] and flow injection analysis (FIA) [12,13] are further examples of intermittent techniques, neither of which rely on a chromatographic separation but present some form of the process stream to the detector.

The concept of on-line analysis can be extended still further to "in-line" and "non-invasive" analysis as defined by Callis et al. [2]. Chemical sensors which can be placed directly inside process pipework remove the need for sampling and a system that requires no direct contact with the process stream represents the ultimate process monitor. The most obvious in-line sensor is the pH electrode; rugged versions of which are used in chemical processing. Many other types of sensor are available but have yet to be used routinely due to their poor long-term reliability. Spectroscopic techniques have also been developed for in-line analysis, whereby some form of optrode may be placed in-situ and connected to a remote spectrometer by fibre-optics. Furthermore, multiplexing allows numerous optrodes to be monitored by a single spectrometer. This can be taken one stage further, whereby optical windows are incorporated in process pipework, allowing non-invasive spectroscopic analysis in the NIR region for example. Other examples of non-invasive analysis include IR emission, X-ray absorption and acoustic emission analysis.

The final approach taken to solving a process analysis problem should be carefully considered. Ideally, a working party consisting of an analytical chemist, a process chemist and an electrical/electronic engineer should consider the following issues:

1. Analysis objectives;
    i. purpose,
    ii. analyte/s,
    iii. frequency,
    iv. delay.
2. Economic justification.
3. Analytical feasibility;
    i. accuracy, precision and sensitivity,
    ii. selectivity and matrix interference,
    iii. instrumental reliability.

Potential analysis procedures should be subjected to rigorous laboratory and plant validation trials and the instrumentation provided with ample technical support after installation. Routine maintenance schedules should be implemented to ensure minimum down-time and enable effective control.

1.2 PROCESS FIA

FIA is now widely accepted as a laboratory tool for routine analysis and research [14,15,16]. It is taught as an integral part of courses in analytical chemistry, has been the subject of five major international conferences [17,18,19,20,21], has four monographs devoted to its theory and applications [10,11,22,23] and a dedicated periodical; the Journal of FIA [24]. FIA is an unsegmented flow technique. This is in contrast to CFA (such as the Technicon AutoAnalyzer™ system) which relies on air bubble segmentation as a means of keeping successive samples apart. The essence of FIA is controlled dispersion, enabling reproducible mixing of sample and reagents without excessive dilution, to produce measurable transient signals proportional to analyte concentration. Dispersion in CFA is minimised by air bubbles which create a series of isolated reaction chambers where homogeneous mixing produces steady state signals. Due to non-segmentation and

heterogeneous mixing FIA offers a number of distinct advantages over CFA, notably;

1. short response time,
2. improved reproducibility,
3. greater versatility,
4. less complex instrumentation.

These advantages are particularly pertinent within the context of PAC, where the near real time pseudo continuous output achievable with FIA systems is of particular importance. The suitability of FIA for process monitoring and control was first reported in 1982 [25] and several papers discussing its potential followed [26,27,28,29,30,31,32,33]. The merits of process FIA are summarised below:

1. Fast response,
2. High sample frequency,
3. Rugged and dependable hardware,
4. Ease of automation and self-calibration,
5. On-line sample treatment capabilities,
6. Low reagent consumption,
7. Compatibility with liquid process streams,
8. Stay clean properties,
9. Wide range of established laboratory methods,
10. Low operational and maintenance costs.

Microcomputer control of FIA components is shown in Fig. 1.1 and illustrates the collection and manipulation of data and the communication of information to a central process computer. Long-term accuracy is maintained by the incorporation of self-calibration procedures and diagnostic routines can monitor precision and instrument performance. Due to the low reagent consumption capability of FIA large vessels of potentially hazardous chemicals do not have to be accommodated and with stable reagents then long-term unattended

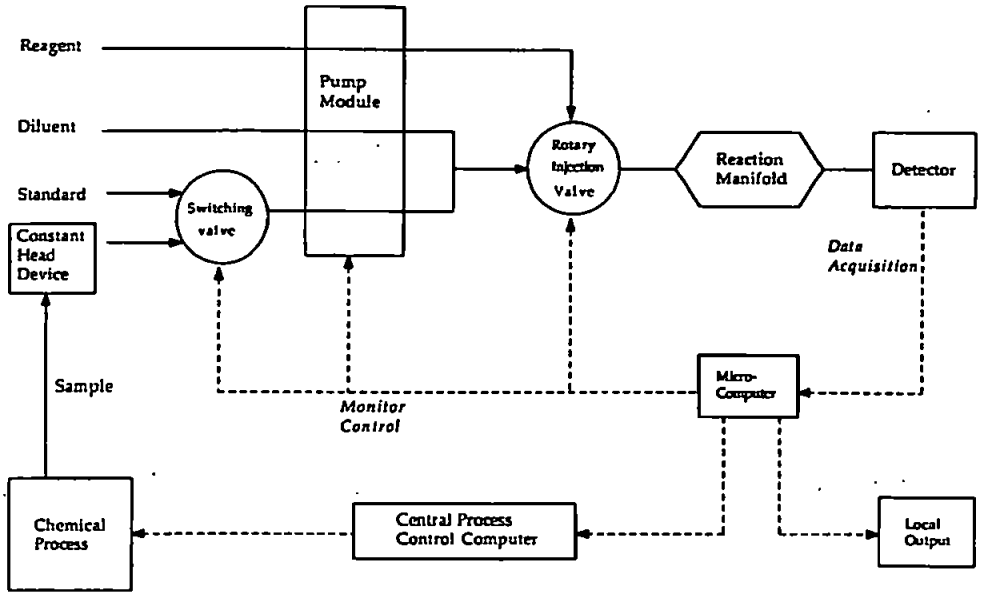


Figure 1.1 Schematic diagram of a process FIA system

operation is possible. The use of "reagent injection" manifolds [34,35] can reduce consumption still further; this is particularly useful when expensive reagents are required. Manifold simplicity is a prerequisite for process FIA and Fig. 1.2 illustrates the reagent injection approach within this framework along with the concept of continuous monitoring systems [11] or completely continuous flow analysis [36].

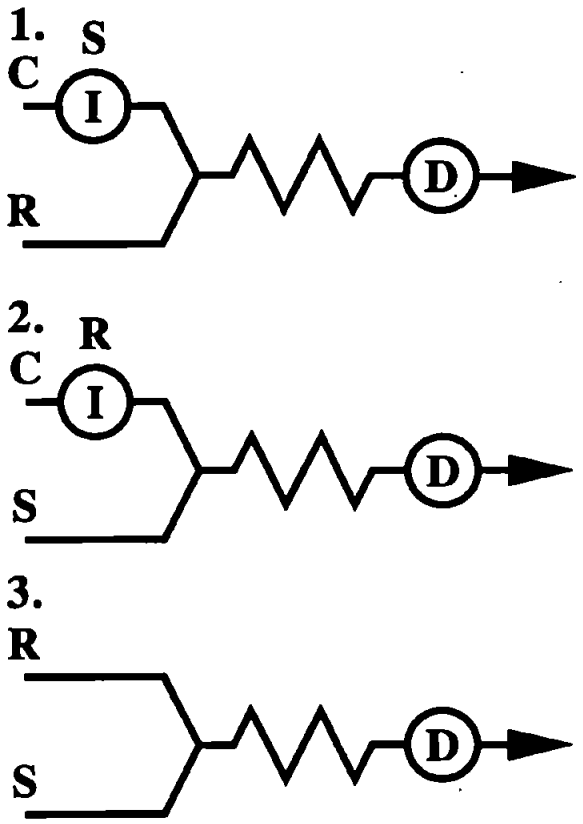


Figure 1.2 Process FIA manifolds:
1. Sample injection.
2. Reagent injection.
3. Continuous flow.
S, sample; C, carrier; R, reagent; I, injector; D, detector

Despite the obvious potential of process FIA the number of practical on-line applications discussed in the literature remains

Table 1.1 Applications of Process FIA

Field	Analysis	References
Chemical production	Sulphide in DIPA solution	28
	Azo compounds	30,37
	Sulphates and phosphates in effluents	38
Water quality monitoring	Phosphate	38
	Nitrate	39,40,41,42
	Ammonia	43
	Fluoride	44
	Aluminium	45
Biotechnology	FDH and L-LeuDH	46
	L-phenylalanine	47
	Glucose, lactic acid and protein	48
	Protein	49
	Cellulase activity	50
	Glucose, ammonium and protein	51

very small. This can be attributed, in part at least, to problems of industrial confidentiality [52]. The applications that have been reported can be divided into three distinct areas as listed in Table 1.1. Although their numbers are few, these examples demonstrate some of the salient features of process FIA; particularly its flexibility for monitoring diverse analytes. The ability of process FIA to deal with harsh sample matrices has been proven by the analysis of dye production liquors [30,38] and fermentation broths [47-52], and long term application has been demonstrated with a nitrate monitor that has been operating

continuously in a remote site for several years [40]. Possibly the greatest interest in process FIA has been shown in the field of biotechnology; this is demonstrated by the proceedings of the Anabiotec meetings [53,54], a special issue of the Journal of Biotechnology [55] and, most recently, an extensive review article [56].

A number of developments in laboratory FIA practice are directly relevant to process analysis. These include further use of membrane separation techniques [28,44,57] for gas analysis [58,59] and preconcentration [60], novel approaches to calibration [61,62,63], and the combination of sequential injection [64] with sinusoidal flow [65,66]. With the sinusoidal flow pump the concept of constant flow is replaced by variable but reproducible nonlinear flow created by a cam-driven, computer controlled piston. Among the advantages for process applications are the simplicity of construction, absence of pump-tubing and check valves, pulseless flow, and the capability of handling aggressive liquids. Sequential injection for zone penetration is achieved by using a simple selector valve, which, in combination with the sinusoidal flow pump, offers a single line manifold suitable for a number of analyses without the need for physical reconfiguration.

Another area to be exploited for laboratory analysis is that of simultaneous multicomponent determinations by FIA, whereby several analytes in the same sample are measured from a single injection [67]. Although this is possibly the best demonstration of the capacity and versatility of FIA [33], it is yet to be applied in process analysis.

1.3 SIMULTANEOUS MULTICOMPONENT FIA

Simultaneous determinations in FIA were reviewed in 1984 by Luque de Castro and Valcarcel [68] and the framework for classification described therein is still generally applied. FIA techniques for speciation were reviewed in 1986

[69] and the term simultaneous, as opposed to sequential, was clarified in terms of multidetection and multideterminations [67].

The methodology for multideterminations by FIA was divided into two groups; conventional FIA and those methods based on differential kinetics. In spite of some innovative procedures being available [70,71,72], the very nature of kinetic determinations renders them unsuitable for continuous monitoring, particularly in a process environment, and are not considered here. Of the conventional FIA methods for multicomponent analysis those most suited to process applications utilise simple manifolds, stable chemistries, and, ideally, only one detection system and injector.

An elegant approach has been exploited by Townshend and co-workers for the speciation of iron [73] and cerium [74], and the simultaneous determination of anions [75,76]. This type of manifold, shown in Fig. 1.3, allows the splitting of a single sample injection for different treatments followed by remerging and detection. The problem of irreproducibility of splitting was circumvented by placing the pumps after the splitter. The speciation studies gave the concentration of the lower oxidation state from the untreated stream and the utilisation of a Jones reductor column in the other stream gave total analyte concentration. Binary mixtures of nitrate/chloride and nitrate/sulphate were resolved using a suppressor column in one stream. Detection was achieved via the iron(III) / thiocyanate complex after displacement of thiocyanate from an anion-exchange microcolumn. The concept of sample splitting has also been applied to binary [77,78,79,80,81,82,83] and ternary multideterminations [84] using immobilised enzyme reactors.

Various injection techniques have been used to produce doublet peaks for multideterminations, including internally coupled valves [85,86,87,88], sandwich techniques [89,90], and reversed injector loading [91]. Whitman et al. [92] used minimal dispersion and the inherent absorbance of

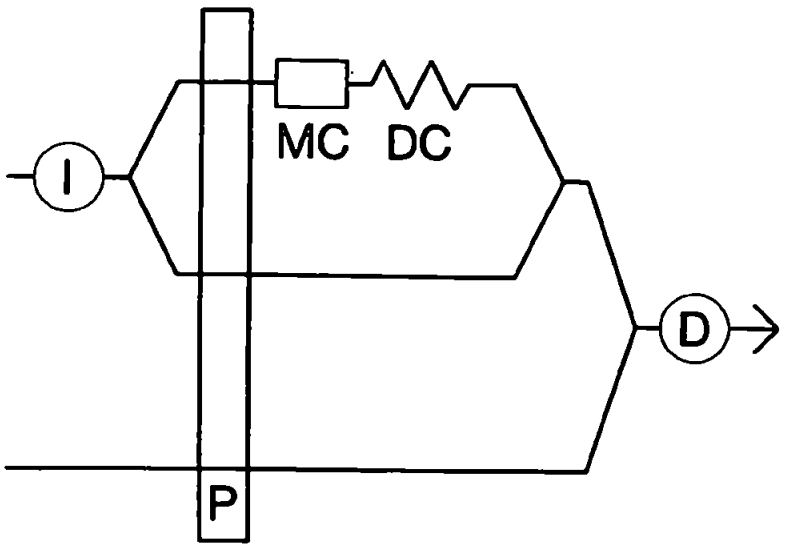


Figure 1.3 General manifold design for stream splitting; I, injection valve; P, pump; MC, micro column; DC, delay coil; D, detector.

aqueous nickel(II) for its simultaneous determination with iron(II) (detected via the thiocyanate complex), thus eliminating the need for stream splitting or multi-injection. Trojanowicz and Spunzar-Lobinska [93] recently developed a low-cost multi-light emitting diode (LED) detector to determine aluminium and zinc.

Electroanalytical systems have been less widely applied to multideterminations. Three interesting applications, however, are the voltammetric determination of phenolic compounds [94], an investigation into Kalman filtering for improved resolution [95], and the use of amperometry for the simultaneous enzymatic determination of glucose and ascorbic acid [96].

Multidetection systems greatly enhance the capabilities of FIA to perform multideterminations and offer a number of advantages for process analysis. A multidetection system is a single device capable of recording a number of analytical signals simultaneously, examples of which include; electrochemical sensor arrays, multi-LED devices (as discussed above) and photodiode-array spectrophotometers (PDA).

PDAs have been commercially available since 1979 and are a product of the

revolution in microprocessor technology [97]. The principle of "reverse-optics" is illustrated in Fig. 1.4, where it can be seen that after passing through the sample polychromatic light is dispersed onto a diode-array. This is in direct contrast to conventional spectrophotometry in which monochromatic light passes through the sample to a photomultiplier. The advantages of the PDA arrangement over conventional scanning spectrophotometers can be summarised as:

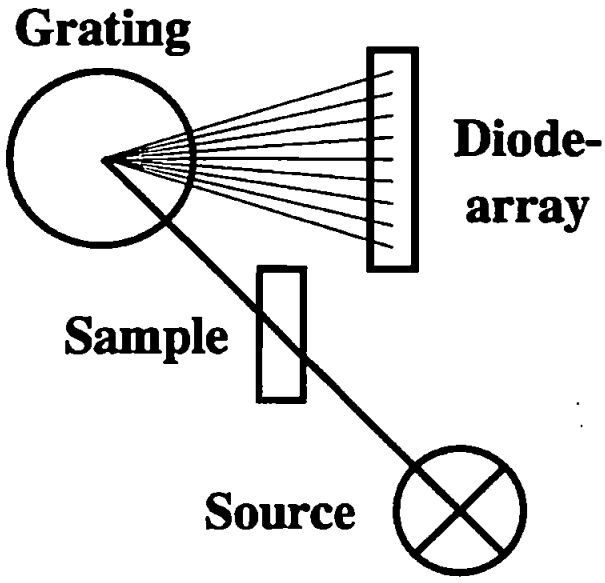


Figure 1.4 Schematic diagram of PDA optical arrangement.

1. Rapid acquisition of complete UV/visible spectra,
2. Mechanical simplicity,
3. Wavelength resettability,
4. Measurement statistics,

The linear diode-array is made up of a number of photodiodes positioned in series on a silicon crystal. Light impinging on a diode causes the capacitor to which it is connected to discharge and the extent to which it needs recharging is proportional to the light intensity. The diodes are multiplexed to allow simultaneous measurement and a holographic grating ensures that small portions of the spectrum selectively impinge on each diode. This arrangement allows the collection of an entire UV/visible spectrum in as little as one tenth of a second. The elimination of the need for monochromatic light renders the PDA mechanically very simple and reliability is improved due to the minimal use of moving parts. This has the added attraction of improving confidence in the resettability of wavelength and because a number of measurements can be taken quickly then a statistical measure of data quality can be made at each

wavelength.

The speed with which PDAs can acquire and store multiwavelength data is particularly important for the monitoring of dynamic systems such as FIA and HPLC [98,99]. They are widely used in routine HPLC analysis where multivariate analysis of the column eluent allows peak purity checks to be made. Furthermore, commercial software is available from some instrument manufacturers for statistical selectivity enhancement, aimed particularly at the pharmaceutical industry. PDAs have been less widely used as detectors for FIA although their full-spectrum capabilities are complimentary to the rapid and highly reproducible sample treatment and delivery features of automated FIA. Their combination offers a great deal in terms of versatility and simplicity for simultaneous multicomponent determinations and the removal of matrix interferences. A schematic representation of a typical FIA-PDA arrangement is shown in Fig. 1.5.

The first applications of the FIA-PDA combination were reported in 1986 for chemical equilibrium studies [100], and for the simultaneous determinations of copper(II) and iron(II) using a 1:10 phenanthroline/neocuproine mixed

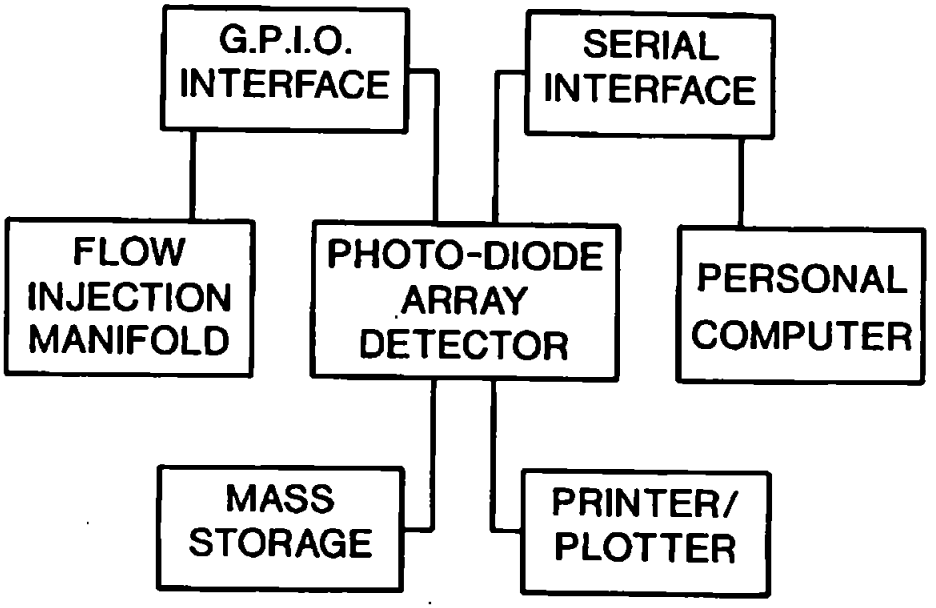


Figure 1.5 Schematic diagram of a typical FIA-PDA system.

reagent [101] and the enzymatic determination of ethanol and acetaldehyde [102]. For the simultaneous determinations, the wavelengths of maximum absorbance of the reaction products were monitored and the results were calculated with due consideration of synergistic effects. The monitoring of wavelengths away from lambda max and a series of wavelengths around lambda max were shown to aid dilution and amplification methods for the determination of nitrate [102]. These techniques were further studied for the formaldehyde/pararosaniline/sulphite system [103]. Mixed reagents and absorbance maxima were also used for the simultaneous determination of iron and copper in blood serum [104]. Examples of simultaneous determinations using a single indicator reaction have been reported for aromatic isomers after diazo-coupling [105], iron(II) and copper(II) with PAN-7S [106], iron(III) and aluminium(III) as oxinates [107], and nickel(II) and zinc(II) with PAN [108].

1.4 RESEARCH OBJECTIVES

The general aim of this research was to investigate the potential of FIA for the on-line analysis of chemical parameters in process streams.

The particular aims were as follows:

1. To develop a single analyte FIA procedure to plant specifications and prove the system reliability with on-line trials.
2. To evaluate the most appropriate FIA approaches to simultaneous multideterminations for process analysis.
3. To investigate the potential of recent developments in quantitative chemometrics.
4. To develop a simultaneous multi-analyte FIA procedure suitable for on-line process monitoring.

Chapter Two

On-line FIA determination of sulphite

2.1 INTRODUCTION

Potassium hydroxide is an important intermediate in the manufacture of potassium salts in the chemical and agricultural industries. It has traditionally been produced by the electrolysis of potassium chloride brine using the mercury cell process but recent developments in cell technology have led to the commissioning of a membrane electrolysis facility. This has advantages in terms of power consumption and environmental impact.

Imperial Chemical Industries have been producing KOH at the Castner Kellner Works on Merseyside since the 1950s. This new plant has been built to replace existing facilities and meet the increased demand for KOH liquor both in the UK and overseas. The plant, opened in 1989, is designed to produce 75,000 tonnes of 50% KOH and 24,000 tonnes of chlorine per annum. It operates as a single unit for KOH production; comprising KCl resaturation/purification, electrolysis and caustic evaporation. The chlorine and hydrogen produced by the plant are moved to other areas of the works for treatment and distribution.

Potassium chloride (muriate of potash) is the basic raw material and major cost for the process. It is transported by road from the main sources of supply in Cleveland (UK), France and Germany. Because of the high cost of the raw material, a resaturation process is employed whereby weakened brine is strengthened by redissolving KCl. A schematic diagram of the KOH plant is shown in Fig. 2.1.

The process can be divided into three main areas:

1. Brine Purification.

   Depleted chlorinated KCl brine is adjusted to pH 2 and fed into the tops of towers where air is drawn upwards to remove most of the chlorine. Any remaining chlorine is removed by adjusting the pH to 10.5 and the addition of potassium sulphite. The dechlorinated weak brine is passed

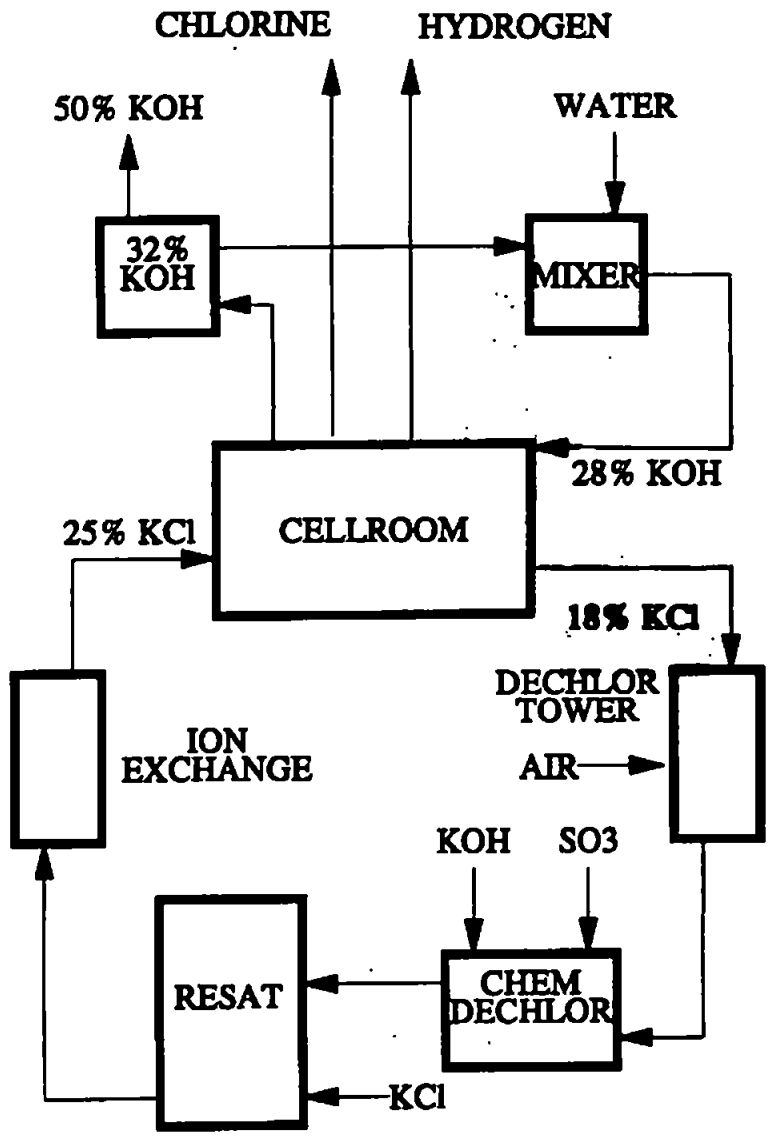


Figure 2.1 Schematic diagram of the KOH plant.

up through a 'bed' of potassium chloride where it is resaturated and then through filters to remove insolubles. The brine is then fed through ion exchange columns to remove soluble contaminants, particularly the group II metals.

2. The Cellroom.

   25.5% m/v high purity KCl brine and 28% m/v KOH, each at approximately 70°C, are fed into the respective sides of each cell. A high current is passed through the liquors producing 32% m/v KOH, chlorine, hydrogen and 18% m/v KCl brine.

3. Caustic Evaporation.

   The 32% m/v KOH is passed through heat exchangers and evaporators

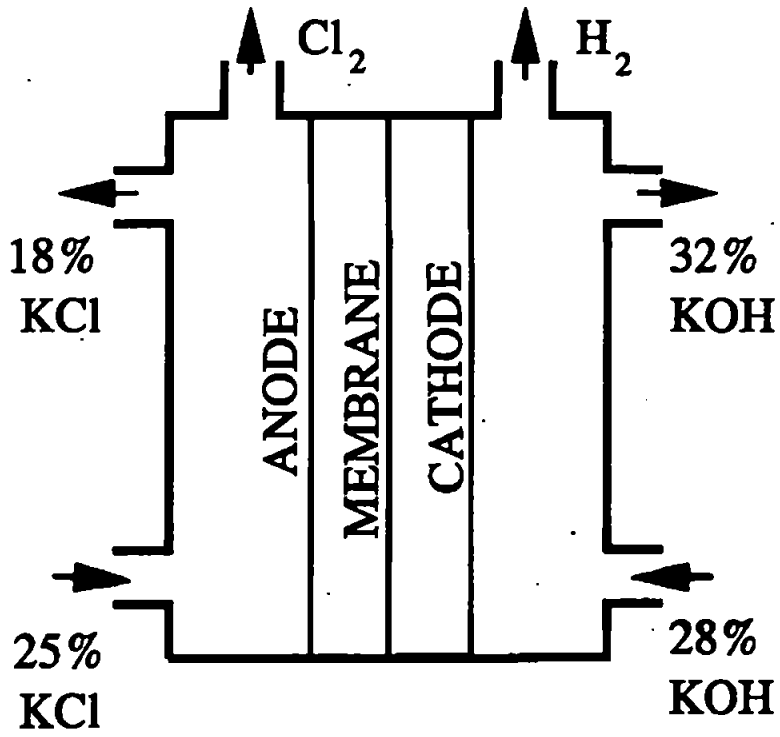


Figure 2.2 Schematic diagram of the membrane electrolysis cell.

to strengthen the liquor to 50% m/v KOH before pumping away to storage.

A schematic representation of the membrane cell is given in Fig. 2.2. KCl brine is fed into the anode compartments of the cell and KOH solution into the cathode compartments. Under the influence of the current which passes between the electrodes through the liquors and the membrane, chlorine gas is liberated at the surface of the anode. Potassium ions are transported through the chemically inert and selective membrane to the cathode. As hydrogen is liberated at the cathode the resultant hydroxyl ions balance the flux of potassium ions, leaving the cell as strengthened KOH. The membrane serves as a physical separator, preventing the mixing of chlorine and hydrogen gases, and the brine and potassium hydroxide. Some back migration of caustic into the anolyte compartment does occur however, leading to a small loss in current efficiency.

The cell room consists of 30 FM21 SP cells arranged in two rows of 15. Each cell consists of 60 titanium anodes and 60 nickel cathodes arranged alternately

with membranes sandwiched between. Each membrane has an active electrolytic area of 0.21 m^2. The membrane, manufactured by Du Pont, is Nafion 430, which is a perfluorosulphonic acid polymer reinforced with a PTFE mesh.

During brine purification, potassium sulphite is added to the recirculating brine as a chlorine scavenger. This in turn leads to an increase in the sulphate concentration according to the equation:

$$SO_3^{2-}(aq) + H_2O(l) + Cl_2(g) \rightarrow 2H^+(aq) + 2Cl^-(aq) + SO_4^{2-}(aq)$$

Residual chlorine is removed to maintain the efficiency of the ion exchange resin but the increased sulphate level necessitates a continual brine purge. As discussed earlier, potassium chloride is very expensive and the purge needs to be kept to an absolute minimum. Any increase in sulphate concentration is directly proportional to the rate of potassium sulphite addition, therefore continuous monitoring of the sulphite concentration would allow closer control and hence a lower purge rate. A system capable of measuring sulphite in the process liquors on a near-real time basis is therefore required with the following specifications.

Plant specification:

Analyte;	sulphite
Dynamic range;	1-20 mg l^{-1}
Matrix;	18 % m/m potassium chloride
Temperature;	70-80 °C
pH;	11-12
Response time;	15 min
Accuracy;	± 10 %
Precision;	± 5 %
Maintenance;	<1 h per week

In addition to the above application, sulphite is widely used as an antioxidant in the pharmaceutical and food industries, and as an oxygen scavenger in water for steam generation and in paper pulping. There are several reported methods for the determination of sulphite, principally by spectrophotometry [109,110,111,112,113,114], amperometry[115,116], potentiometry [117], chromatography [118,119], enzymatic analysis [120] and chemiluminescence [121]. Several of the above laboratory based methods use flow injection analysis (FIA) techniques for sample treatment and presentation to the detector [113-116,121].

As discussed in section 1.2, spectrophotometric FIA is particularly suitable for process monitoring and this chapter describes the development and validation of an FIA monitor to meet the specification listed above.

2.2 EXPERIMENTAL

Reagents

All solutions were prepared in distilled, de-ionised water and all reagents were of AnalaR grade (Merck) unless otherwise indicated. A 1000 mg l^{-1} sulphite (as SO_3^{2-}) stock solution was prepared by dissolving 1.5743 g of sodium sulphite (dried for 2 h at 105°C) in 1 l of 1 x 10^{-3} mol dm^{-3} ethylenediaminetetraacetic acid (EDTA) (0.3722 g of EDTA sodium salt dissolved in 1 l of water). Appropriate dilutions of this stock solution were made in water, 20 % m/v standard potassium chloride and potassium chloride brine for the respective calibrations. Sulphite is readily oxidised during the preparation of aqueous solutions and hence, prior to use, the concentration of stock solutions was determined iodimetrically. A solution containing an excess of iodine was acidified with hydrochloric acid, the sulphite solution was added carefully with stirring. The remaining iodine was titrated with sodium thiosulphate [122].

Solutions of 2,2'-dinitro-5-5'-dithiodibenzoic acid (DTNB) (Aldrich) were

prepared by dissolving an appropriate amount in ethanol (5 ml l^{-1} of solution) and diluting with pH 6.9 buffer. The pH 6.9 buffer was prepared by dissolving 3.55 g of disodium hydrogen orthophosphate and 3.41 g sodium dihydrogen orthophosphate in 1 l of water. The pH 9.9 buffer was prepared by dissolving 19.07 g of disodium tetraborate decahydrate (borax) and 2.0 g of sodium hydroxide in 1 l of water. The pH 11.7 buffer was prepared by dissolving 3.80 g of trisodium phosphate (BDH; general purpose reagent) in 1 l of water. All pH adjustments were made using hydrochloric acid and sodium hydroxide of various concentrations.

Instrumentation

A schematic diagram of an automated FIA monitor is shown in Fig. 1.1. The FIA manifolds were made with 0.8 mm i.d. polytetrafluoroethylene (PTFE) tubing (Anachem) with PTFE T-pieces for stream merging. The absorbance was monitored by a spectrophotometer (LKB Ultrospec II) fitted with an 18 µl silica flow cell with a path length of 1 cm (Hellma), and the analogue output relayed to strip chart recorder (Chessel BD 40 04). Injections (20 µl) were made *via* a 12 V solenoid-activated injection valve (Chemlab Instruments) and standard/sample selection controlled by a 2-way 12 V solenoid valve (Lee). Sample, reagent and carrier streams were propelled by two peristaltic pumps (Ismatec Mini S-820) with poly(vinyl chloride) (PVC) pump tubing (Labsystems).

The system was controlled by single board microcomputer (Control Universal) as described by Clinch [123] and Benson [124]. Data acquisition and data output was achieved by incorporating additional cards into the system. The individual cards are described below:

1. *Control and data processing*: EuroBEEB with 6502 8-bit 2 MHz microprocessor and 8 Kb RAM or 16 Kb EPROM.
2. *Data Storage*: CU-MEM Selecta with 32 Kb RAM for the storage of

raw data.

3. *Signal capture*: CUBAN-12A 16 channel analogue to digital converter with 13 bit accuracy and 1 mV resolution.
4. *Output*: JOBBER interface enabling data output to VIEWLINE, a 24 character by 2 row liquid crystal display, and RACKPRINT, a 24 character per line miniature impact printer.

Control software was written in MosB4, an extended version of BBC BASIC language and details of the software protocols are given in the procedures section.

Spectral measurements and kinetic studies were carried out using a diode array spectrophotometer (Hewlett Packard 8451A) fitted with a 1 cm pathlength silica cuvette. A thermostatically controlled, heated water bath (Grant W14) was used for the temperature effect study.

Sample presentation to the on-line monitor was facilitated by a 1 l constant head device. This was plumbed into the process stream *via* a length of polypropylene pipe, thus ensuring a continuously replenished real-time supply of the process liquor.

Procedures

Batch experiments

In all experiments the response corresponds to the addition of a 3.0 ml aliquot of sample solution (0.0 or 16.7 mg l^{-1} of SO_3^{2-}) to 0.3 ml of DTNB solution (4.0 g l^{-1}; 24-fold concentration excess). The sample solution was added directly to the cuvette containing the DTNB reagent positioned in the spectrophotometer sample holder and measurement started immediately at 412 nm.

Automated laboratory analysis

The FIA manifold for the laboratory analysis (shown in Fig 2.3) used a pH 9.9

borax-sodium hydroxide buffer. The sensitivity of this system to sample pH was investigated by analysing standard solutions of 10 mg l^{-1} of sulphite in water and in 20 % m/v potassium chloride solution adjusted to various pH values. The effect of sample temperature was studied by analysing a standard solution of 10 mg l^{-1} of sulphite in 20 % m/v potassium chloride solution at pH 11.0 maintained at various temperatures in a thermostated water-bath. Sensitivity to potassium chloride concentration was determined by analysing standard solutions of 10 mg l^{-1} of sulphite prepared in potassium chloride solutions covering the range 0-20 % m/v at pH 11.0.

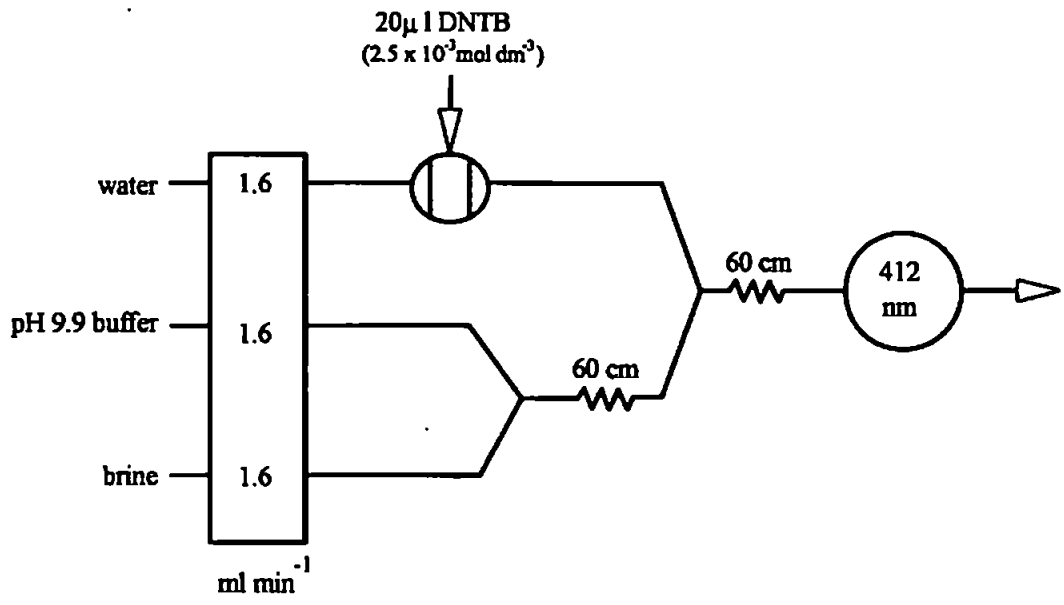


Figure 2.3 Flow injection manifold for the determination of sulphite in high ionic strength potassium chloride brine.

Development software was written for control and data processing in a form that facilitated operator interaction. This ensured that modifications to the sampling frequency, data acquisition mode and data treatment could be implemented from a general program.

The modified 4-line manifold used for on-line analysis is shown in Fig. 2.4.

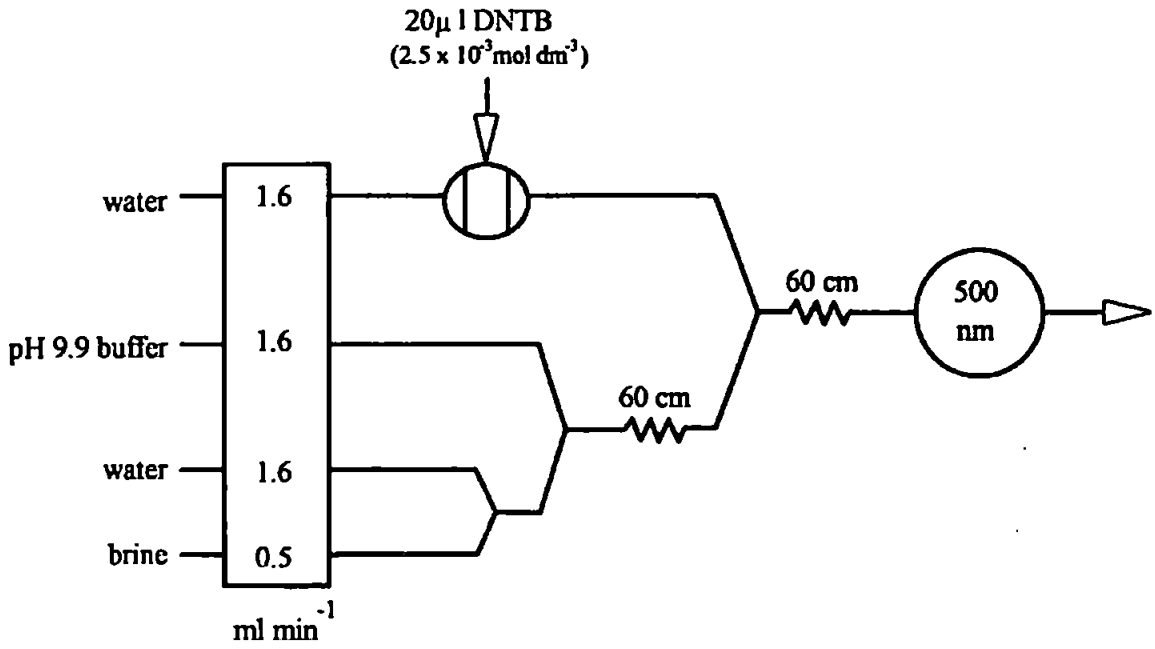


Figure 2.4 Modified flow injection manifold for the on-line determination of sulphite in high ionic strength potassium chloride brine.

The control software was developed for unattended operation. This incorporated a triplicate standard analysis every 60 min and a triplicate sample analysis every 15 min. Each sulphite concentration was calculated by ratioing each mean sample response to the preceding mean standard response. This ensured that every result was automatically calibrated to the standard response measured at the most 60 min beforehand, thus compensating for any signal drift.

All hardware control was achieved *via* a series of commands communicated through the serial output of the EuroBEEB card. The events were timed to maintain the reproducibility of the system protocol allowing sufficient time for sample and standard flushing to reduce any memory effects.

The CUBAN-12A A/D card was configured to capture the signal generated by the spectrophotometer and after conversion the digital data was stored in the

CUMEM RAM. Upon completion of data collection and storage a peak find algorithm was activated. This processed the raw data, isolated the baseline and peak maximum absorbances and computed the difference between them as the peak height. In an attempt to minimise the gathering of spurious data, the algorithm was designed to use a rate of change of absorbance function to locate the peak maximum. This enabled differentiation of the analytical signal from spikes due to entrained air. In addition, the precision of the triplicate analysis was monitored and a relative standard deviation of >5 % led to the result being discarded and the analysis cycle repeated.

Upon completion of every 15 min analysis cycle, the time, sulphite concentration and relative standard deviation were down-loaded to the local printer and visual display *via* the JOBBER card.

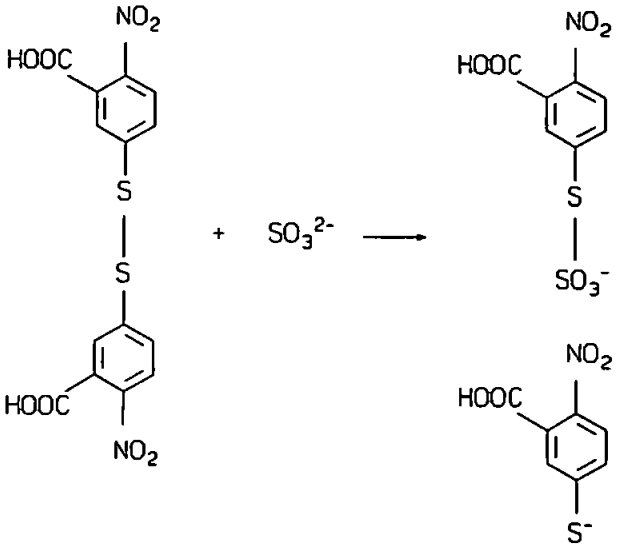
2.3 RESULTS & DISCUSSION

Reaction chemistry

Of the available spectrophotometric procedures for the analysis of sulphite, the methods based on *p*-rosaniline [108] and 1,10-phenanthroline [109] are particularly sensitive to sample pH and were considered inappropriate for development. The methodology described by Humphrey *et al* [110] offered greater tolerance to pH however and, furthermore, had been successfully developed into a laboratory-based FIA procedure [112]. The reagent used was an organic disulphide; 2,2'-dinitro-5-5'-dithiodibenzoic acid (DTNB). Sulphite reacts quantitatively with DTNB to produce a chromophoric thiolate species, 2-nitro-5-mercaptobenzoic acid, as shown overleaf.

Batch experiments

While the DTNB reaction had been shown to be more tolerant to pH than the *p*-rosaline and 1,10-phenanthroline reactions, initial studies indicated some variance at higher pH. For the analysis of aqueous sulphite standards a pH 6.9



buffer was found to be adequate but the process liquor is in the range pH 11-12 and a buffer of higher pH was considered more appropriate. Experiments conducted using a pH 11.7 buffer indicated that precise pH control in alkaline media is much more important than in neutral solutions.

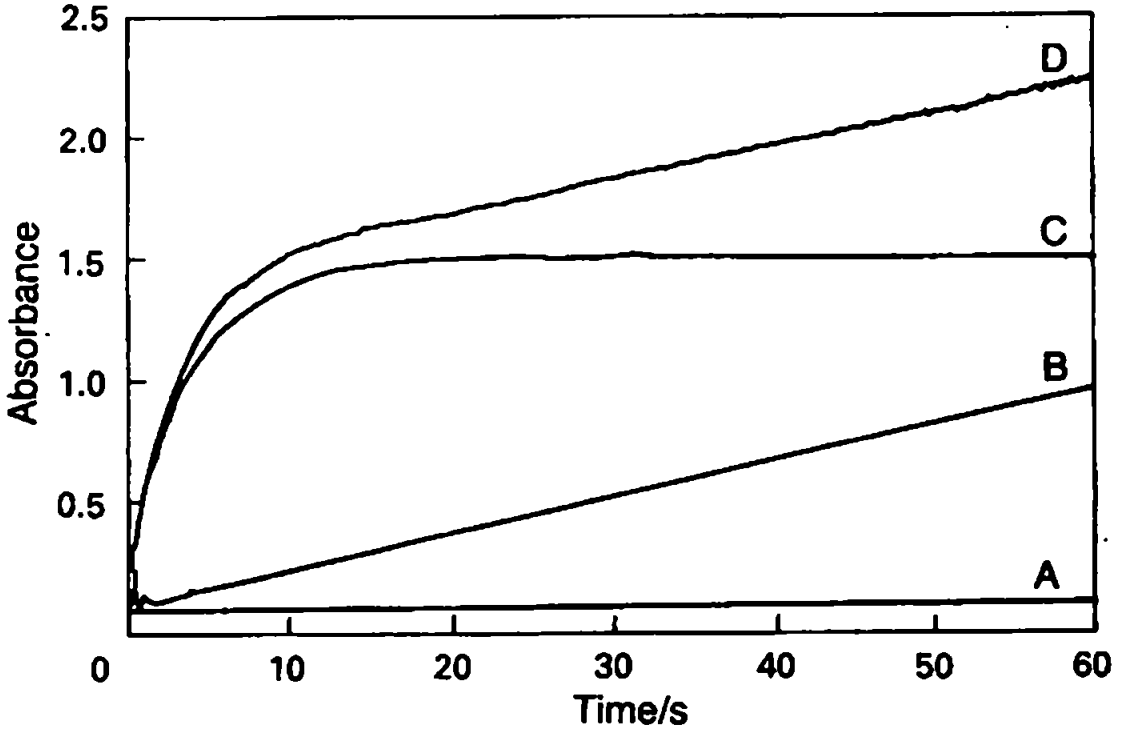


Figure 2.5 Response profiles for the reaction of DTNB: A, pH 6.9 buffer; B, pH 11.7 buffer; C, pH 6.9 buffer + 10 mg l^{-1} sulphite; and D, pH 11.7 buffer + 10 mg l^{-1} sulphite.

As can be seen in Fig. 2.5, the reaction profile is significantly affected by

increasing the pH of the reaction medium from pH 6.9 to pH 11.7. The ultra-violet spectrum (200-800 nm) of DTNB in pH 11.7 buffer is very similar to that of the thiolate anion (the monitorand of the DTNB-sulphite reaction), suggesting cleavage of the DTNB sulphur-sulphur bond at high pH in the absence of sulphite. It is known that aromatic disulphides, particularly nitro-substituted aromatic disulphides, are susceptible to cleavage in alkaline conditions [125, 126], yielding the corresponding thiolate anion and sulphinic acid. The molar extinction coefficient calculated in terms of the thiolate anion was found to be in agreement (ε=13500 l mol^{-1} cm^{-1}) with that previously reported, thus confirming cleavage of the DTNB sulphur-sulphur bond to yield the thiolate anion in a 1:1 ratio.

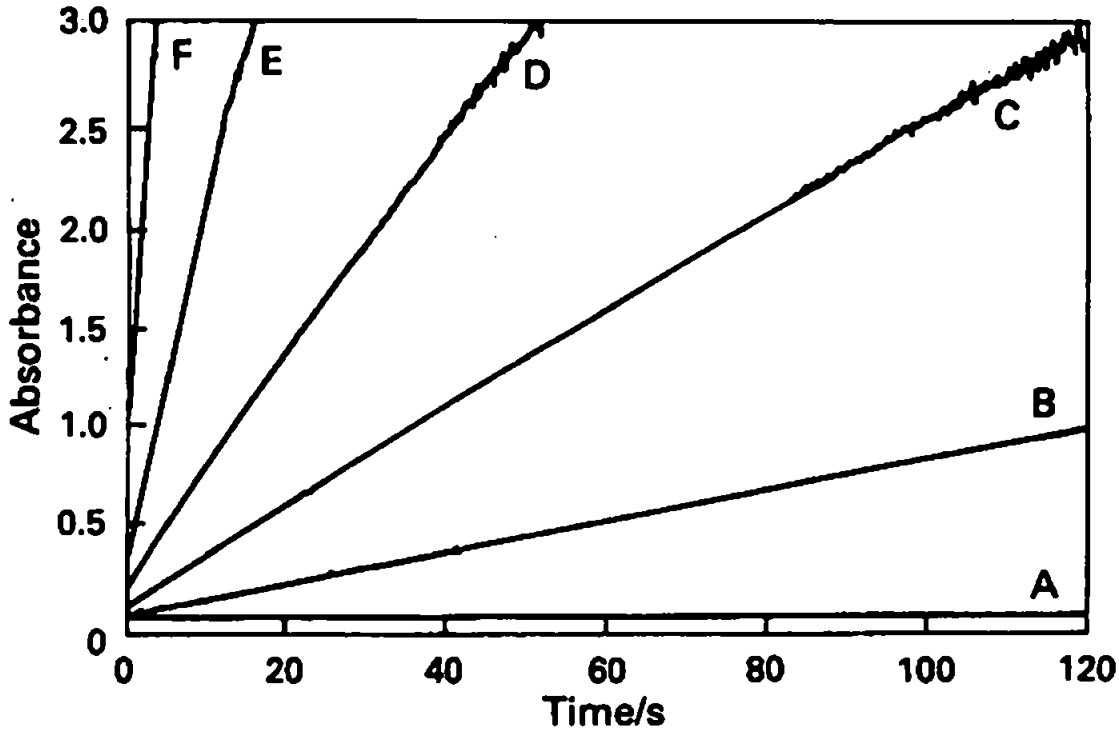


Figure 2.6 Response profiles for the cleavage of the DTNB sulphur-sulphur bond: A, pH 10.9; B, pH 11.5; C, pH 11.8; D, pH 12.1; E, pH 12.5; and F, pH 12.8.

The effect of pH on cleavage of the DTNB sulphur-sulphur bond in the absence of sulphite was further investigated over the pH range 10.9-12.8. Fig. 2.6 shows that at pH 10.9 there is no significant increase in absorbance with time but at a higher pH the rate of cleavage is significantly affected by very small

pH changes. For on-line analysis it is therefore necessary to buffer the sample stream to pH 10.9 or below in order to minimise this cleavage.

Automated laboratory analysis

Calibration

Calibration data are presented in Table 2.1 for sulphite in water, 20 % m/v potassium chloride standard solution and potassium chloride brine (obtained from ICI Chemicals & Polymers). A sample injection frequency of 60 samples h^{-1} was used throughout and ten replicate analyses of each solution were made. The results indicate good correlation in the range 0.1-20 mg l^{-1} of sulphite.

Table 2.1 Calibration data for sulphite in aqueous media.

Sulphite Conc. mg l^{-1}	Matrix (n=10)					
	Water		KCl (20 % m/v)		KCl brine	
	AU	RSD %	AU	RSD %	AU	RSD %
0	0.050	1.0	0.059	3.3	0.056	4.2
5	0.228	0.4	0.251	0.8	0.237	0.7
10	0.409	0.6	0.442	0.9	0.417	0.6
15	0.555	0.5	0.616	0.3	0.584	0.7
20	0.716	0.3	0.777	0.4	0.743	0.4

Linear regression data.

(n=5)	Water	KCl (20 % m/v)	KCl brine
Slope	0.033	0.036	0.034
Intercept	0.060	0.069	0.063
Correlation coefficient	0.9992	0.9993	0.9996

pH stability

Owing to the sensitivity of the reaction to pH, the effect of sample pH on the response of the system was investigated for standards in water and in 20 % m/v potassium chloride. The results given in Table 2.2 reveal that below pH 12.0 there is no significant variability in response and that above pH 12.0 only a slight increase in signal is observed. The manifold is therefore suitable for process applications because the pH of the sample stream rarely exceeds 12.0.

Table 2.2 Effect of sample pH on system response for 10 mg l^{-1} standard sulphite solutions prepared in water and 20 % m/v KCl (n=5).

Matrix	Sample pH	Mean signal (AU)	RSD (%)
Water	6.6	0.297	0.6
	9.5	0.289	0.4
	10.4	0.293	1.0
	11.2	0.298	0.3
	11.9	0.299	0.7
	12.2	0.311	0.4
KCl (20 % m/v)	5.6	0.423	0.5
	10.4	0.422	0.4
	10.8	0.423	0.5
	11.4	0.427	0.9
	12.1	0.435	0.9
	12.5	0.436	0.5

Temperature stability

The temperature of the KCl brine on plant is maintained in the range 70-80 °C and with its transfer and holding in the constant head device, the temperature of the abstracted sample may vary considerably. This could have a significant effect on the rate of reaction but the results given in Table 2.3 show that any temperature effect is eliminated by sample dilution in the FIA manifold.

Table 2.3 Effect of sample temperature on system response for a 10 mg l^{-1} standard sulphite solution in 20 % m/v KCl at pH 11.0 (n=5).

Sample temp. (°C)	Mean signal (AU)	RSD (%)
25	0.463	0.4
35	0.464	0.3
45	0.463	0.6
55	0.466	0.1
70	0.470	0.2
90	0.462	0.5
95	0.468	0.7

Effect of potassium chloride concentration

Comparison of the calibration data for water, 20 % m/v potassium chloride standard solution and potassium chloride brine, suggests that the potassium chloride concentration has an effect on sensitivity. This is confirmed in Fig. 2.7, which reveals the increased response for a 10 mg l^{-1} standard sulphite solution with increasing potassium chloride concentration. This is thought to be due to an increased rate of reaction with increased ionic strength. However, the process stream contains 18 % m/m potassium chloride, which corresponds to the region exhibiting the least variation in response, and small changes in process ionic strength will therefore have minimal effect.

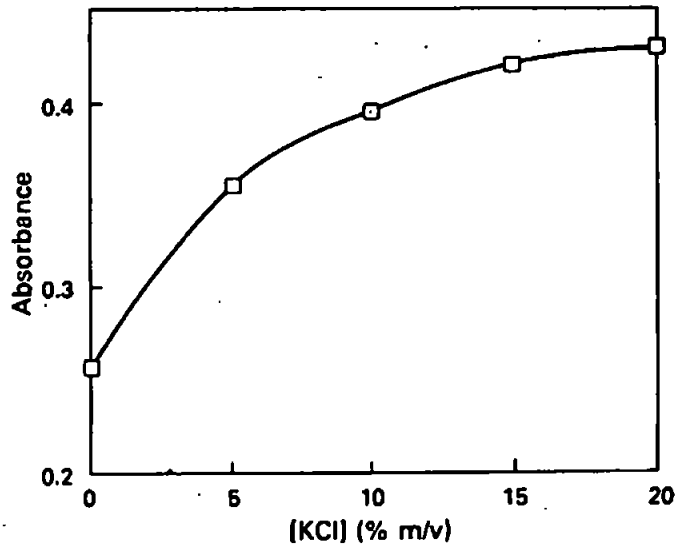


Figure 2.7 Effect of potassium chloride concentration on the response.

On-line analysis

Calibration

A 16 h continuous trial of the manifold and software using a 10 mg l^{-1} sulphite standard solution produced a mean response of 0.452 A.U. with a relative standard deviation (RSD) of 0.46% (n=16). Analysis of a simulated sample solution gave a mean concentration of 15.85 mg l^{-1} with an RSD of 0.51% (n=64). This concentration was subsequently confirmed by iodimetric analysis.

Preliminary on-line trials revealed that the sulphite concentration in the process liquor was outside the linear range of the proposed method (0.1-20 mg l^{-1}). In order to extend the linear range, the absorbance of the thiolate anion was measured at increasing wavelengths from 412 nm (the wavelength of maximum absorbance). Measurement at 500 nm extended the linear range from 0.5 to 40 mg l^{-1} with a subsequent reduction in sensitivity to 0.139 A.U. for a 40 mg l^{-1} standard. The linear range could not be extended further in this manner owing to the poor signal-to-noise ratio observed at higher wavelengths. Furthermore, it was not possible to increase the DTNB concentration because of its limited solubility. It was therefore necessary to dilute the sample further prior to analysis. This was achieved on-line by modifying the FIA manifold as shown in Fig. 2.4. Calibration data for the modified manifold covering the range 3-200 mg l^{-1} are presented in Table 2.4. Sulphite can be determined over the range 3-200 mg l^{-1} and the response is linear for the concentration range 3-100 mg l^{-1} (n=5, r=0.9999).

Validation of the on-line method

Results from the on-line method were validated against the standard iodimetric procedure over an 8 h period. A portion of the process liquor was abstracted from the constant head device every 15 min to coincide with the analysis cycle of the monitor. Three hours into the trial, the addition of potassium sulphite solution to the process stream was increased, slowly and in a step-wise manner, over a period of 2.5 h. After the trial had been in progress for 5.5 h, the rate

Table 2.4 Calibration data for the on-line determination of sulphite in 20 % m/v KCl at pH 11.0 (n=10).

Concentration (mg l^{-1})	Mean signal (AU)	RSD (%)
0	0.015	2.9
25	0.038	2.6
50	0.063	-
75	0.088	0.9
100	0.112	0.8
150	0.150	0.6
200	0.177	0.7

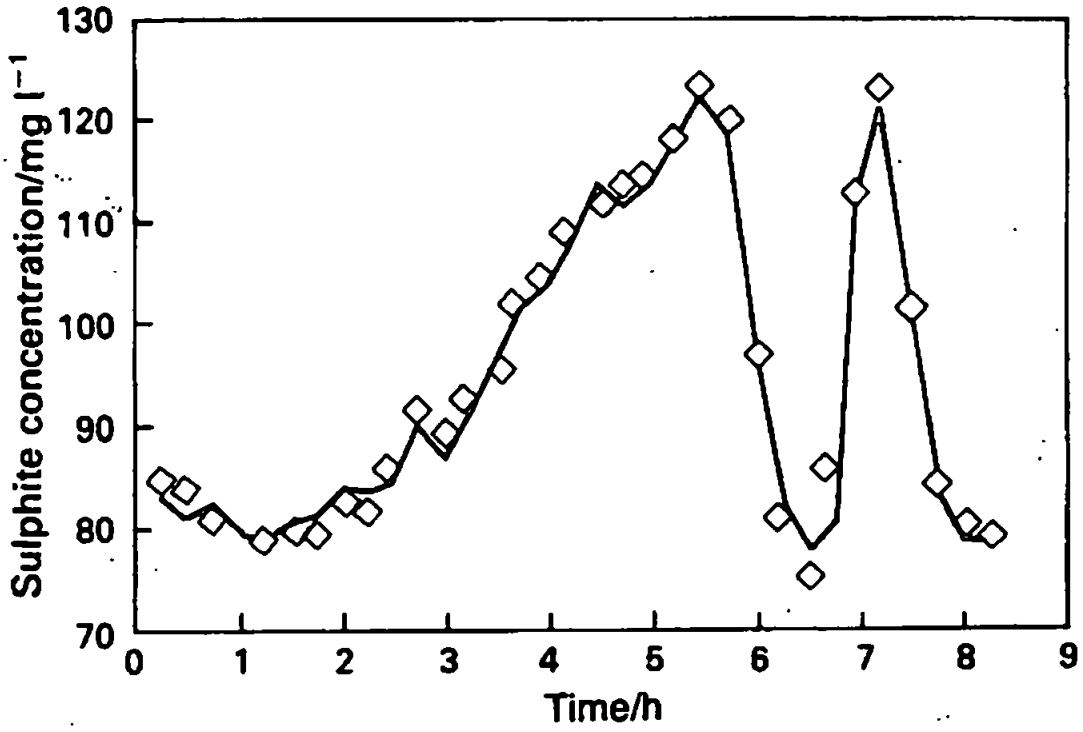


Figure 2.8 Method comparison study: solid line, monitor response; and squares, off-line iodimetric results.

of sulphite addition was reduced to its original level and then rapidly increased and reduced again over a 1 h period. It can be seen from Fig. 2.8 that the corresponding changes in sulphite concentration are closely followed by the monitor and by the iodimetric procedure. Fig. 2.9 shows the regression of the monitor response on the results of the standard iodimetric analysis. The results in Table 2.5 reveal no observable systematic error.

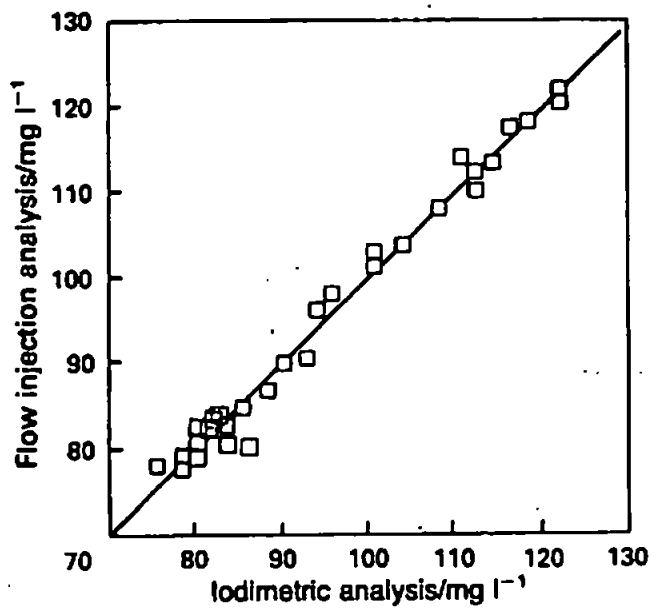


Figure 2.9 Regression of the flow injection results on the off-line iodimetric results.

Table 2.5 Regression data from Fig. 2.9

Slope	Intercept	Correlation coefficient
0.990 ± 0.031	0.64 ± 2.94	0.9964

Extended on-line trial

Fresh reagents were prepared weekly and details of reagent consumption are given in Table 2.6. The response to a standard sulphite solution (62.3 mg l^{-1}) over a 1 week operating period (168 triplicate determinations) was 0.084 A.U. with an RSD of 2.1%. Peristaltic pump tubing was replaced after 14 days, and in 21 days of continuous on-line use only one failure was reported (owing to blockage of the injection valve).

Table 2.6 Reagent consumption over a 7 day period (672 analyses).

Reagent	Consumption (l)
De-ionised water	10.8
pH 9.9 buffer	5.4
DTNB reagent	1.4
Sulphite standard	0.9

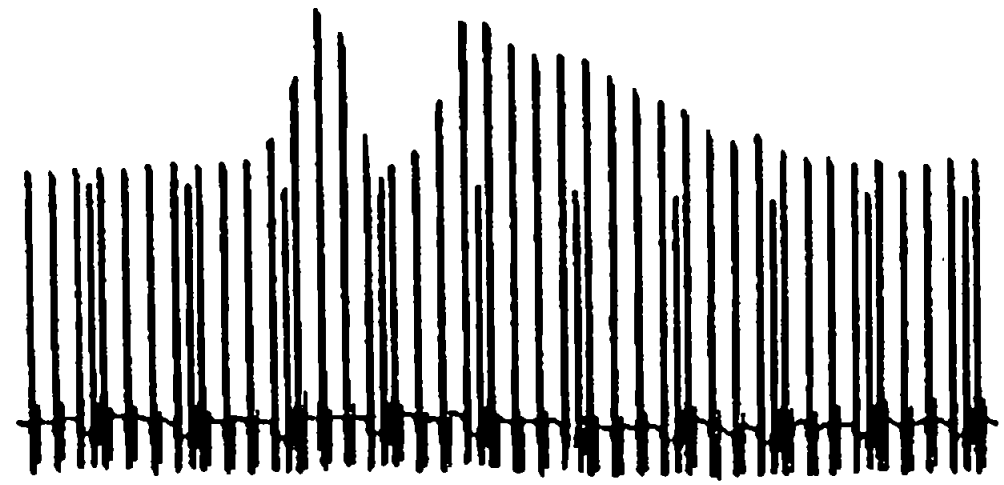


Figure 2.10 Analogue output of 10 h on-line monitoring of the process liquor.

A plot of the analogue output from the monitor is shown in Fig. 2.10 and emphasises changes in the process sulphite concentration relative to the constant response due to the standard. Fig. 2.11 shows the monitor output over the period of the 21 d on-line trial. The sharp increase in concentration at approximately 400 h corresponds to a temporary plant shut-down, during which time sulphite addition was maintained.

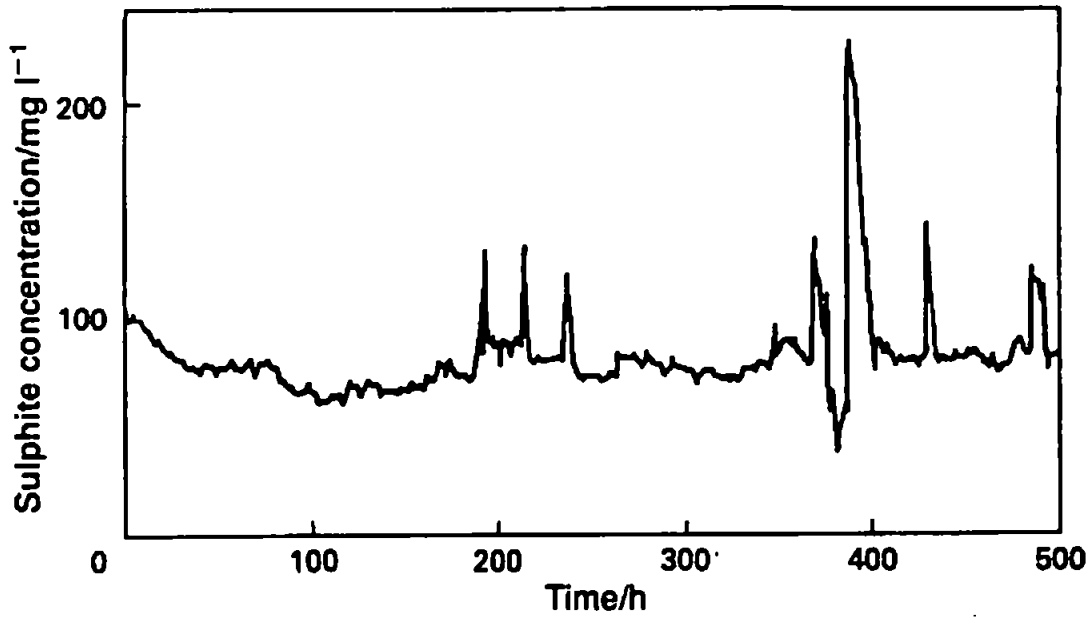


Figure 2.11 Results of a 21 day on-line trial.

The overall performance characteristics of the proposed monitor are summarised in Table 2.7. It can be seen that the system meets all of the criteria set out in the original plant specification and boasts modest purchase and operational costs.

Table 2.7 Performance characteristics and specifications.

Parameter	Plant specification	Proposed monitor
Over-all accuracy	± 10 %	± 3 %
Precision	± 5 %	± 1 %
Response time	15 min.	< 5 min.
Dynamic range	1-20 mg l^{-1}	0.1-100 mg l^{-1}
Maintenance	< 1 hr. $week^{-1}$	30 min. $week^{-1}$
Running costs	-	< £1.00 day^{-1}
System costs	-	£6,500

2.4 CONCLUSIONS

The proposed FIA monitor has been shown to meet plant specifications for the on-line determination of sulphite in a real chemical processing environment.

The reaction chemistry has been shown to be sufficiently stable in terms of the pH, temperature and potassium chloride strength of the process liquor. Moreover, the on-line performance of the system has been validated with an off-line standard procedure and the instrumentation reliability has been demonstrated with a 21 day trial.

Chapter Three

Multivariate calibration techniques

3.1 INTRODUCTION

In a period of less than twenty years, chemometrics has grown from merely a collection of statistical techniques into a dynamic area of research with wide reaching implications. Chemometrics has been defined by Massart et al. [127] as,

> "...the chemical discipline that uses mathematical, statistical, and other methods employing formal logic (a) to design or select optimal measurement procedures and experiments, and (b) to provide maximum relevant chemical information by analysing chemical data."

This was more succinctly stated by Malinowski [128] as,

> "...the use of mathematical and statistical techniques for handling, interpreting, and predicting chemical data."

According to Svante Wold [129] chemometrics started in 1920 with Gosset's "Student's t-test" but it was not until 1974 that he and Kowalski founded the International Chemometrics Society. In the interim chemists from different research areas had been applying well-established mathematical and statistical techniques to chemical problems in isolation from each other. The advent of chemometrics was therefore perceived by its critics to be "more of the same" [130] but by the mid 1980's the discipline was firmly established and now boasts two dedicated journals [131,132] and a number of specialised monographs [127,133,134,135]. The development of chemometrics can be split into three phases:

1. Academic research into algorithms and associated software and its application to selected data sets.
2. Commercialisation of user-friendly software resulting in much wider application across the analytical community.
3. Marketing of analytical instrumentation with dedicated chemometric software rendering routine analysis possible.

The historical development of chemometrics has been discussed in greater detail

by Vandeginste [136] and Wold [129].

The revolution in laboratory computing and computerised analytical instrumentation coupled with the development of chemometric methods has transformed many areas of analytical chemistry into an information science [137]. Furthermore, the use of appropriate mathematical routines can in some cases reduce the complexity and cost of chemical analysis [138] especially as analytical equipment increases in price and computer hardware becomes cheaper [139].

Some of the most dramatic advances in chemometrics have been made in the area of calibration [140]. A number of different approaches to calibration are possible and it is the aim of this chapter to present these techniques, firstly in practical spectroscopic terms and secondly in a statistical/algorithmic sense.

3.2 PRACTICAL CALIBRATION

Calibration is the determination of a mathematical function that can be used to predict quantitative information from measured data. In practice, this means taking transmittance or absorbance values in spectroscopy or peak area or peak height values in chromatography, and finding their relationship to known analyte concentrations in order to predict the concentration of analytes in unknown samples.

Univariate Calibration

Calibration in chemistry has traditionally relied on the measurement of a single variable to predict one analyte concentration. For example, manual spectrophotometric analysis for a single analyte would involve the preparation and single wavelength analysis of a set of standard solutions. The measured values would then be plotted against the corresponding concentrations and a straight line or curve fitted. This graph would subsequently be used to predict

the unknown analyte concentration in samples by interpolation. Statistical software packages are now widely used to produce calibration graphs but, as Miller points out [141,142], care should be taken to visually inspect the resulting fit. The use of such software provides the analyst with the least squares regression curve, its associated errors, and interpolated results with their confidence intervals.

The univariate approach is excellent in cases where the analytical response is entirely selective for the analyte under observation. Much of the available spectrophotometric methodology includes steps to minimise the effect of potential interferents by techniques such as solvent extraction or derivatisation; an obvious selectivity enhancement tool is chromatography. It is possible however, that the analytical procedure in question has failed to account for all potential interferents, and the measured response reflects a positive or negative contribution due to something other than the analyte. This interference may be in the form of a species absorbing at the same wavelength, or something that effects the derivatisation procedure, eg. pH or competitive complexation. This will lead to erroneous predictions, which may go unnoticed by the analyst [143].

In routine analysis, more than one assay is often required on each sample and rather than carrying out numerous procedures, some form of multidetermination could prove to be more efficient. Unfortunately, univariate techniques preclude the collection of simultaneous multi-analyte information. This, and problems of selectivity can, in many cases, be solved by collecting multivariate data and utilising one of the mathematical selectivity enhancement tools that are the product of quantitative chemometrics.

In spectroscopy, multivariate calibration relates a set of signals from a multichannel instrument to the concentration of one or more analytes in a sample [144,145] and is most often applied to the quantitative

interpretation of non-selective chemical data [146,147]. Some multivariate calibration techniques such as direct multicomponent analysis have been in use for quite some time, but others, partial least squares regression for example, are relatively recent products of chemometric research.

The Beer-Lambert Model

The Beer-Lambert or linear mixture model states that the absorbance at a particular wavelength is a linear function of the concentrations of the absorbing species present in the solution under examination. Therefore, in the simplest case, binary mixtures can be analysed by the construction and solution of two simultaneous equations using the absorbance data from two appropriate wavelengths. This can be extended to the determination of J components from J equations at J wavelengths providing that the response at each wavelength is sufficiently different for each component. This approach is seldom useful except under ideal circumstances due to the effect of random noise and the selectivity problems discussed earlier.

By including the absorbance data from more wavelengths than J (over-determined systems) the effect of noise can be reduced by ordinary least squares fitting [144], and reduced still further by using weighted least squares. The Beer-Lambert model is most often applied to spectrophotometric data as an over-determined or full-spectrum technique known as direct multicomponent analysis (DMA) or direct unmixing. It is direct in the sense that the spectrum of every absorbing species needs to be known in advance. Interferences not explicitly modelled or inter-analyte interactions that influence the spectral data can yield erroneous information on prediction. If the spectral interference is sufficiently different from any linear combination of the pure spectra, then the least squares residual will be higher than expected, suggesting an outlier. However, if the converse is true, then it is unlikely that the "rogue" sample will be spotted. DMA routines are often provided as on-board software by spectrophotometer manufacturers. This makes them particularly easy to

implement as no transfer of data is required between the instrument and a secondary computer.

It is also possible to use Beer-Lambert models for indirect calibration; if pure component spectra cannot be measured then a statistical estimate of them can be made and used to make predictions.

Multiple Linear Regression

Indirect calibration methods are more generally applicable because they do not require pure component spectra to be known in advance or to be calculated. Multiple linear regression (MLR) is conceptually the most simple of the indirect calibration methods. It can be viewed as a multivariate extension to univariate linear regression. MLR assumes that concentration is a linear function of instrumental response (cf. the Beer-Lambert models). To ensure successful prediction with MLR, the wavelengths used in calibration need to be carefully selected. This is due to the phenomenon known as "multicollinearity", whereby variables approximate to linear combinations of other variables; a problem often encountered in spectrophotometry. Selection of the wavelengths to be used in MLR may be achieved statistically by stepwise MLR; a method that chooses the "best" subset of variables according to some predefined criterion. Alternatively, selection may be made by judicious choice according to the analyst's knowledge of the samples and their spectra. Whilst MLR, carefully executed, has the advantage over DMA of not requiring interferences to be known before-hand, the interferences do have to be incorporated into the calibration. Also, due to the wavelength selection requirements of MLR, the full-spectrum advantages of DMA are lost.

Most general statistical software packages, such as Statgraphics™, can handle MLR quite satisfactorily, and Minitab™ includes a stepwise MLR procedure.

Factor Analysis

According to Martens and Naes [140] flexible calibration methods are needed that can simultaneously overcome the problems of:

1. selectivity,
2. collinearity,
3. lack of prior knowledge.

The Beer-Lambert methods can deal with problems of selectivity and collinearity but require pure component spectra and are unable to account for analyte interactions. MLR, on the other hand, suffers from data redundancy and leaving variables out of the calibration reduces the effectiveness of outlier detection. Stepwise MLR provides a means of data compression or "rank reduction" that utilises selected wavelengths which attempt to represent all the relevant information in the spectra.

A different approach to data compression which utilises all of the spectral variables is known as factor analysis. Here, the spectral information is concentrated onto a few factors, which can be used as variables in an MLR regression equation. A factor is a linear combination of the original variables. As with MLR, factor analysis techniques assume that concentration is a function of instrumental response, but due to rank reduction the problem of collinearity can be overcome. The term bilinear refers to the way that the spectral data is expressed as the product of two linear parameters, known as the scores and loadings.

The estimation of these parameters is very useful for qualitative data analysis as well as in calibration. Consider a series of spectra of solutions containing two analytes, without any physical or chemical interferences. Factor analysis of the data should reveal two factors which describe the original data. The structure of this data can be expressed by plotting the scores of factor 1 against the scores of factor 2. Samples with similar profiles would tend to be grouped together in such a plot and the presence of an isolated sample could be

indicative of unusual characteristics. Furthermore, plots of the factor loadings can reveal the identity of the wavelengths that are influencing the factors to the greatest extent. This can be useful for the interpretation of the physical phenomena influencing particular spectral regions.

Most importantly, inspection of the scores and loadings plots can reveal situations where the spectra are more complicated than anticipated. For example, a third and unexpected phenomenon could be influencing the spectra of the two analyte solutions. Models based on the linear mixture model would seldom identify a third constituent, whereas factor analysis will reveal the need for a third factor to successfully describe the original spectra. In practice, the determination of the number of factors required to describe the data (the dimensionality of the data) is rarely straightforward. Factor analysis incorporates measurement noise and non-linearities into the model, often leading to dimensionalities far in excess of the number of analytes. Here again, visual inspection of the model parameters is vital to safeguard against "overfitting" which can lead to unstable models. The importance of dimensionality estimation is dealt with in more detail in Chapter 6.

As with all indirect calibration methods, factor analysis techniques require that there is a relationship between the concentration of the analyte and the spectra that is sufficiently unique to allow quantitation. In addition, the spectra used to build the calibration model must have known analyte concentrations and span all anticipated analyte and interference levels independently. Once a model has been built that satisfies these criteria, then the analyte concentration of unknown samples can be estimated regardless of interferences. If the data in question does not fit the model, then the sample may be identified as an outlier. This may be due to an unexpected interference which was not included in the calibration set, or simply that the level of analyte is outside those spanned in calibration.

Factor analysis, a product of research in the behavioral sciences, has appeared in the chemical literature under a number of different guises and although the terminology differs, the methodology is often very similar. Two methods to which factor analysis is fundamental are discussed below.

Principal Components Regression

The decomposition of the data matrix into its most dominant factors has been referred to, amongst other things, as principal components analysis [148,149] (PCA), principal factor analysis (PFA) and singular value decomposition (SVD). All of these methods are equivalent, although the means of achieving the decomposition may vary according to the author. PCA lends its name to principal components regression (PCR) which uses the most dominant factors for calibration.

The first principal component is a linear combination of the original variables that best describes the measured spectra. It is calculated in the least squares sense to yield the lowest residuals. Subsequent principal components are successively calculated in the same way to explain the remaining variance. Projection of the original data onto this reduced dimension space gives the factor scores and regression of the concentration data onto the score matrix gives the calibration coefficients. For prediction of an unknown sample, the scores of the new spectra are calculated and the concentration determined via the regression equation.

By utilising these dominant factors PCR can provide a much more flexible and robust calibration than full-spectrum MLR and has come to replace MLR and stepwise MLR in many NIR calibrations [150,151,152]. Its weakness, however, lies in the data compression stage. It is conceivable that the most dominant factors are not those that best describe the analyte concentration; factor 1 may be largely describing measurement noise for example.

Partial Least Squares Regression

Just as PCR utilises the most dominant factors in the spectral data, partial least squares regression (PLSR) attempts to define the factors which are most relevant to the concentration of the analyte in question [140,145,153]. This is achieved by simultaneously estimating the factors in both the spectral and the concentration data, and actively using the concentration data in the bilinear decomposition of the spectral data. In this way PLSR can reduce the influence of dominant but irrelevant factors and in some cases yield models of lower dimensionality which are subsequently easier to interpret. PLSR also has the advantage of being able to model a number of analytes simultaneously.

Partial least squares has developed from the early work by Hermann Wold between 1960 and 1980 [154] and is being increasingly used as a calibration technique in chemistry today. PLSR has been applied to data from various analytical techniques including liquid chromatography with UV detection [155] and the following spectroscopies; NIR [145], molecular fluorescence [156,157], X-ray diffraction [158], FTIR [159,160,161], FT Raman [162] and UV-visible [163].

A number of specialised chemometric software packages are commercially available with which to perform these bilinear modelling techniques. Among the most popular is the Unscrambler II™ program which has PCA, PCR and PLSR facilities.

Unscrambler II™

Unscrambler [164], an interactive program for multivariate calibration and prediction [165], is the product of research at the Norwegian Food Research Institute by Harald Martens and co-workers. The program is well structured and with its menu driven interface is relatively easy to use.

The development of a multivariate calibration model and subsequent predictions

using Unscrambler can be broken down into a number of distinct stages.

1. *Problem definition and experimental design.*

Before commencing any chemical analysis it is important to clearly define one's objectives. This is no less true when dealing with multivariate calibration. Consider an analysis involving the spectrophotometric determination of three inherently absorbing analytes in an aqueous matrix. One must first address the following issues:

i) Can artificial calibration standards be prepared?

ii) Are the expected levels of all analytes and interferences known?

iii) What are the general analytical requirements in terms of accuracy, precision, sensitivity and detection limits?

If artificial standards can be prepared then the analyst has much greater control over the experimental design. This would generally be true in spectrophotometry, but in other cases (whole grain analysis of wheat by NIR reflectance for example) the analyte levels may have to be determined on real samples by a reference method. Furthermore, if the analyte and interference levels are well characterised then a structured experimental design spanning all anticipated events can be used for the training set.

The choice of experimental design is a non-trivial matter [143,166,167,168], indeed a detailed discussion is beyond the scope of this chapter. It is important, however, to avoid making unnecessary replications and whilst ensuring that all expected phenomena are spanned, care should be taken to avoid very large designs. This can be particularly problematic when factorial designs are used. Consider a 2-level 3-factor design which requires $2^3 = 8$ samples; if this is increased to a 3-level 4 factor design then $3^4 = 81$ samples are required. Nevertheless, the preparation, analysis and data processing of such a large experiment can be avoided by using fractional factorial designs. This approach effectively ignores higher order interactions, which are often negligible, and can reduce the experiment size considerably.

For example, the 3^4 design could be reduced to only 27 samples by using a full 3^3 design with the fourth factor assigned according to the sum of the first three factor effects. This is designated as a 3^{4-1} design. Another approach to design which necessitates fewer experiments is known as the central composite design [166,167].

Once the calibration has been designed, then the analytical variables such as wavelength range and integration time must be considered and the data must be stored in a manner such that it will be easily retrievable for processing.

2. *Data transfer.*

The transfer of data files is often one of the most problematic stages of carrying out a multivariate calibration. Fortunately, Unscrambler has a number of routines designed to aid the import and export of data files and, most importantly, can accept ASCII files in a number of formats. Spectrophotometers that are not controlled by a personal computer, have the added complication of transferring data electronically from instrument to computer. This can be achieved by using one of a number of communication software packages, eg. Kermit™.

3. *Preprocessing.*

With the raw data converted to a suitable format, a number of processing or preprocessing routines can be conducted using the Unscrambler software. It should be noted, however, that great care needs to be taken when manipulating raw data as this can have dramatic effects on the final results. Many of the transformations that can be carried out are linearisation procedures, but curvature is not generally a problem in absorbance mode spectrophotometry. A smoothing function is available which can be used to remove measurement noise. This is a simple averaging function which reduces the number of variables; a box-car moving average function would be a welcome addition here. Spectral derivatives, which can enhance resolution, can also be calculated

but each derivatisation leads to a depreciation of the signal-to-noise ratio.

Another form of preprocessing is the weighting of variables. With bilinear modelling techniques, the data sets are always mean-centred prior to decomposition. Mean-centring simply involves the subtraction of the variable mean from each individual variable. Normalisation is a scaling technique that sets all variable values to unit length across individual spectra. This is particularly useful when variables of different units are used in the same calibration, but is seldom required in spectrophotometry. Autoscaling is used to set all variables to equal variance after mean-centring and is carried out by dividing the individual variables by the standard deviation of that variable across all objects. This can cause problems due to unimportant variables making a significant contribution to the model. As the analyst becomes more familiar with particular data sets then weights may be attached to individual variables according to their relative importance but as a general rule, for new data sets and when in doubt, avoid any form of preprocessing.

4. *Calibration method.*

The choice of calibration method was discussed earlier but there is often little difference between the predictive ability of PCR and PLSR. For some data sets PLSR can yield less complex models than PCR and has the ability to estimate more than one analyte simultaneously. Obviously, if the analyst is only interested in studying the underlying spectral information rather than quantitation then PCA can be employed.

5. *Model validation.*

Selection of the optimum number of factors or dimensions to be used for future predictions is arguably the most important stage in reduced dimension multivariate calibration. If too many factors are included there is the risk of over-fitting the data; the calibration set may be well modelled but subsequent predictions will be unreliable due to the incorporation of noise. Conversely,

using too few factors can lead to under-fitting, leaving important interactions and interferences unmodelled, therefore yielding similarly unreliable predictions.

In order to compare the predictive ability of the model at different dimensionalities some kind of validation needs to be carried out. In practice this means the direct comparison of actual and predicted values for a given set of objects. The objects may be those used in the calibration stage or a subset of them, so called internal validation; or a new and independent set, known as external validation.

When large data sets are available, and a representative subset can be defined, then external validation using this subset is possible. However, this approach is wasteful of data and rarely used in routine work. Internal validation uses the calibration data for measuring predictive ability. Calibration fitting estimates can be used, but this is not validation in the predictive sense and is of little value due to its tendency to underestimate prediction errors. The method of choice is cross validation (CV) which uses independent validation subsets without wasting data. Full CV (leave one out) successively divides the data set (n objects) into a modelling subset (n-1) and a validation subsample until all possible divisions have been made. The predictive ability is calculated at each dimension for each object left out, and hence the optimal model for prediction can be estimated. Full CV in Unscrambler is achieved by setting the number of CV segments equal to the number of objects in the calibration set.

Unscrambler selects the optimum which has the first local minimum of prediction error as the dimensionality is increased one factor at a time. This generally provides a good compromise between over and underfitting of data, whilst remaining computationally simple. Depending on the data in question, especially considering large data sets, the prediction error may never reach a local minimum within the number of factors being considered. In this case the first local minimum is in fact the global minimum. Visual inspection of the

loadings vectors can also be helpful in assessing the extent to which noise is being modelled.

6. *Prediction.*

The concentration of unknown samples can be determined by using the validated model. The program will ask for the number of model factors to be used for prediction and will output estimates of the analyte concentrations and their deviations. In addition, the presence of objects which do not fit the model will be flagged by the program as outliers.

3.3 MULTIVARIATE CALIBRATION ALGORITHMS

A study of the multivariate calibration literature will reveal not only that the same techniques have a number of names but also that the presentation of the algorithms varies widely, both in style and the variables used. The aim of this section is to present the four multivariate calibration techniques which have been applied in this work, in a consistent algorithmic manner.

The following notation is used throughout the thesis:

bold uppercase letters	→	matrix
bold lowercase letters	→	vector
plain lowercase letters	→	scalar

X	→	matrix of spectral variables (independent)
Y	→	matrix of analyte concentrations (dependent)

ie.

$$\mathbf{X}=\begin{vmatrix} x_{11} & \cdot & \cdot & \cdot & x_{1K} \\ \cdot & & & & \cdot \\ x_{I1} & \cdot & \cdot & \cdot & x_{IK} \end{vmatrix}$$

where,

$$\mathbf{x}_i=\begin{vmatrix} x_{i1} & x_{i2} & \cdot & \cdot & \cdot & x_{iK-1} & x_{iK} \end{vmatrix}$$

and,

$\mathbf{x}_i$	→	spectrum i from wavelength $1 \rightarrow K$
I	→	sample or object index
J	→	analyte index
K	→	wavelength index
c	→	concentration
ε	→	molar absorptivity
K	→	matrix of sensitivity coefficients
E	→	matrix of spectral residuals
F	→	matrix of concentration residuals
ß	→	matrix of regression coefficients
'	→	transpose of a matrix or vector
^	→	estimated value (hat)

Matrix inversion

The solution of multivariate expressions in the development of calibration models often involves matrix inversion. Inversion of a matrix is the multivariate equivalent of division [140]. The product of a matrix and its inverse is equal to the identity matrix,

$$\mathbf{X}\mathbf{X}^{-1}=\mathbf{X}^{-1}\mathbf{X}=\mathbf{I}$$

where I is the identity matrix and is the matrix equivalent of 1. A matrix is only invertible if it satisfies the following criteria [128,146];

1. **X** is a square matrix, ie. $I = K$
2. **X** is non-singular, ie. neither the rows or columns are linearly dependent.

In spectroscopy, the first criterion is rarely fulfilled by either the concentration or the spectral matrix, but is overcome by implementing the generalised inverse method. This yields what is referred to as either the generalised inverse, the pseudoinverse or the Moore-Penrose inverse.

If **X** is nonsingular and square then the model

$$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}$$

is solved by

$$\boldsymbol{\beta}=\mathbf{X}^{-1}\mathbf{y}$$

however, when **X** is not square, then a least squares estimate has to be made minimising the squared residuals of **y** such that

$$\boldsymbol{\beta}=[\mathbf{X}'\mathbf{X}]^{-1}\ \mathbf{X}'\mathbf{y}$$

where,

$$[\mathbf{X}'\mathbf{X}]^{-1}\ \mathbf{X}'=\mathbf{X}^{+}$$

which is the pseudoinverse of **X**. Note that the determination of the pseudoinverse includes the inversion of **X'X**, which by definition is a square matrix. There is, however, no guarantee that **X'X** is non-singular. If the variables of a matrix approximate to linear combinations of other variables it is said to exhibit singularity or collinearity. Spectrophotometric data sets contain a great deal of linear dependence and are said to be multicollinear. The generalised inverse of a non-singular matrix is likely to be unstable and can yield models with poor predictive ability.

Direct multicomponent analysis

DMA is a technique based on the Beer-Lambert model which states that absorbance at wavelength k can be expressed as the sum of the component concentrations multiplied by their molar absorptivity at wavelength k. Assuming a cell of fixed path length then:

$$\mathbf{A}=\varepsilon\mathbf{c}$$

therefore, at wavelength k,

$$A_k=\varepsilon_{k1}c_{k1}+\varepsilon_{k2}c_{k2}+.....+\varepsilon_{kj-1}c_{kj-1}+\varepsilon_{kj}c_{kj}$$

In the literature, DMA has been referred to as classical least squares, reverse least squares and K-matrix calibration. The K-matrix refers to the matrix of molar absorptivities or sensitivity coefficients. The model can be expressed in matrix terms as:

$$\mathbf{X}=\mathbf{Y}\mathbf{K}'+\mathbf{E}$$

where $\mathbf{X}$ is an I sample by K wavelength matrix of spectra, $\mathbf{Y}$ is an I sample by J analyte matrix of concentrations, $\mathbf{K}$ is a J analyte by K wavelength matrix of normalised pure component spectra and $\mathbf{E}$ is an I sample by J wavelength matrix of spectral residuals. The calibration model in DMA is usually built from the spectra of pure individual component spectra and therefore no estimation of $\mathbf{K}$ is required. The prediction of a new sample is estimated by:

$$\hat{\mathbf{y}}_i'=\mathbf{x}_i'\mathbf{K}[\mathbf{K}'\mathbf{K}]^{-1}$$

If $\mathbf{K}$ is unknown then the least squares estimate minimising the squares of the spectral residuals is given by:

$$\hat{\mathbf{K}}=\mathbf{X}'\mathbf{Y}[\mathbf{Y}'\mathbf{Y}]^{-1}$$

again using the generalised inverse.

Future predictions are calculated by:

$$\hat{\mathbf{y}}_i' = \mathbf{x}_i'\hat{\mathbf{K}}[\hat{\mathbf{K}}'\hat{\mathbf{K}}]^{-1}$$

Every component that has an absorbance in the region of the spectrum under analysis must be considered in the calibration because the spectra are defined as a function of the individual component absorbances and their concentrations. Omission of a component would yield a large residual error upon prediction.

Multiple linear regression

The mathematical inverse of DMA and related techniques is known as MLR; concentration is defined as a function of the absorbance data. This is illustrated by comparing the model from the last section with that for MLR:

$$c_{ij} = \sum_{k=1}^{K} A_{ik}\beta_{jk}$$

ie. the concentration of the analyte j in sample i is equal to the absorbance at k wavelengths multiplied by the regression coefficients for analyte j at the k wavelengths. MLR is also known as inverse least squares, indirect calibration and forward calibration. In matrix terms, MLR can be represented as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{F}$$

where $\mathbf{Y}$ is an I sample by J analyte matrix of concentrations, $\mathbf{X}$ is an I sample by K wavelength matrix of spectra, $\boldsymbol{\beta}$ is a K wavelength by J analyte matrix of regression coefficients and $\mathbf{F}$ is an I sample by J analyte matrix of concentration residuals.

The regression coefficients are determined in the least squares sense to minimise the squares of the concentration residuals according to:

$$\hat{\boldsymbol{\beta}} = [\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Y}$$

minimising,

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij}-\hat{y}_{ij})^2=\sum_{i=1}^{I}\sum_{j=1}^{J}f_{ij}^2$$

Inspection of this procedure reveals that all the variation in **X** is being used to model **Y** in accordance with the least squares principle. This will include any noise and irrelevant information in addition to the information pertinent to the concentrations. From the regression coefficients predictions can be made using:

$$\hat{\mathbf{y}}_i'=\mathbf{x}_i'\hat{\boldsymbol{\beta}}$$

If noise has been incorporated into the regression coefficients then the subsequent predictions are likely to be inaccurate.

Collinearity is a major problem in full-spectrum MLR calibrations. The generalised inverse is utilised for the calculation of the regression coefficients but **X** is unlikely to be nonsingular. This is the reason for the popularity of selective wavelength routines particularly in NIR spectroscopy, SMLR for example. This type of data reduction renders the inverse stable, but leads to a compromise in both the signal-to-noise ratio and outlier detection.

Principal components regression

Principal components analysis decomposes the spectral matrix into its most dominant factors, where the first principal component describes the greatest variance and subsequent principal components describe the remaining variance. In matrix algebra terms, **X** is approximated by two smaller matrices, **T** and **P**, which describe the patterns in **X**. Thus PCA can be represented as follows:

$$\mathbf{X}=\hat{\mathbf{T}}\hat{\mathbf{P}}'+\mathbf{E}$$

The Columns of **T**, known as principal component scores, are orthogonal and describe the concentration patterns of the objects in **X**. Similarly, the rows of **P**, known as the principal component loadings, are also orthogonal and describe the spectral patterns of the variables in **X**. PCA can be geometrically interpreted as the projection of **X** on a reduced dimension subspace by the

projection matrix **P**. The coordinates of the objects on this hyperspace are the score vectors, **T**.

Decomposition of the **X** matrix can be achieved by a number of numerical algorithms [169]. Of these methods, singular value decomposition (SVD) is recognised as the superior method when all of the principal components are required [170]. However, the non-iterative partial least squares (NIPALS) algorithm is computationally much faster, and in calibration when only the first few factors are required, then NIPALS is the algorithm of choice [146].

The NIPALS algorithm can be summarised as follows:
For each dimension, *a*:

1. Select the score vector $\mathbf{t}_a$ which corresponds to the column of $\mathbf{X}_{(a-1)}$ with the largest remaining variance
2. Calculate the loading vector **p**

$$\hat{\mathbf{p}}'_a = (\hat{\mathbf{t}}'_a \hat{\mathbf{t}}_a)^{-1} \hat{\mathbf{t}}'_a \mathbf{X}_{a-1}$$

3. Scale the length of **p**

$$\hat{\mathbf{p}}_a = \frac{\hat{\mathbf{p}}_a}{\sqrt{(\hat{\mathbf{p}}'_a \hat{\mathbf{p}}_a)}}$$

4. Calculate a new score vector **t**

$$\hat{\mathbf{t}}_a = \mathbf{X}_{a-1} \hat{\mathbf{p}}_a (\hat{\mathbf{p}}'_a \hat{\mathbf{p}}_a)^{-1}$$

5. Check for convergence: If **t** in No. 4. is different to **t** in No. 1. then return No. 2
6. Calculate the residual

$$\hat{\mathbf{E}} = \mathbf{X}_{a-1} - \hat{\mathbf{t}}_a \hat{\mathbf{p}}'_a$$

7. The data matrix is then reassigned as the residual for the next dimension

$$\mathbf{X}_a = \hat{\mathbf{E}}$$

Once the optimal number of dimensions has been determined (see Chapter 6) the principal components regression is obtained by regressing y onto the score vectors

$$\mathbf{y} = \hat{\mathbf{T}}\boldsymbol{\beta} + \mathbf{f}$$

The regression coefficients are determined by least squares estimation minimising the residuals in **f**

$$\hat{\boldsymbol{\beta}} = [\hat{\mathbf{T}}'\hat{\mathbf{T}}]^{-1}\hat{\mathbf{T}}'\mathbf{y}$$

It should be noted that in cases where the number of factors A equals the number of wavelength variables K then the regression coefficients **ß** are equivalent to those in MLR.

Prediction of an unknown sample first requires the calculation of its scores vector

$$\hat{\mathbf{t}}_i = \hat{\mathbf{p}}'\mathbf{x}_i$$

The score vector can then be multiplied by the regression coefficients

$$\hat{\mathbf{y}}'_i = \hat{\mathbf{t}}'_i\hat{\boldsymbol{\beta}}$$

to yield the analyte concentrations.

In the discussion of the NIPALS algorithm, the concentration data has been represented as the vector **y**. In the real analysis of data sets, this could in fact be a matrix of analyte concentration data **Y**. However, the number of analytes under inspection, and their covariance, has no effect on the principal component analysis. The principal components are computed to describe the **X** variance only and therefore, all J analytes use the same data prior to regression. This is

the fundamental difference between PCR and the other biased regression technique used in this work, partial least squares regression.

Partial least squares regression

PLSR is conceptually very similar to PCR. It is an indirect, full spectrum, biased method of regression which uses bilinear modelling. It differs in its approach to the calculation of scores and loadings. Principal components describe the variance in the spectral data, whereas PLS factors are calculated with regard to the concentration variance. According to Martens and Naes [140],

> "...the intention of partial least squares in regression is to optimise parsimony: Produce bilinear calibration models with as few dimensions as possible and in such a way that these dimensions are as relevant as possible."

The choice of factors that describe the spectral variance correlated to the analyte concentrations ensures this relevance.

Similar again to PCA, the PLS principle has its roots in econometrics and the social sciences and various different algorithms are available for matrix decomposition. In chemometrics, PLS models the relationship between two matrices, **X** and **Y**, by a sequence of simple, partial models fitted by least squares. As a consequence, PLS algorithms tend to be more complex than those used in PCR, with more variables requiring computation.

The algorithm may be represented as follows:

For each dimension *a*:

1. Select the score vector $\mathbf{u}_a$ that corresponds to the column of $\mathbf{Y}_{a-1}$ with the largest remaining variance
2. Calculate the loading weight vector **w** by regressing $\mathbf{X}_{a-1}$ onto the concentration scores according to the local model

$$\mathbf{X}_{a-1} = \hat{\mathbf{u}}_a \mathbf{w}'_a + \mathbf{E}$$

The weights are found by least squares estimation minimising the residuals in **E**

$$\hat{\mathbf{w}}_a = \mathbf{X}'_{a-1}\, \hat{\mathbf{u}}_a [\hat{\mathbf{u}}'_a \hat{\mathbf{u}}_a]^{-1}$$

3. Scale the length of **w**

$$\hat{\mathbf{w}}_a = \frac{\hat{\mathbf{w}}_a}{\sqrt{(\hat{\mathbf{w}}'_a \hat{\mathbf{w}}_a)}}$$

4. Calculate a new score vector **t**

$$\hat{\mathbf{t}}_a = \mathbf{X}_{a-1} \hat{\mathbf{w}}_a [\hat{\mathbf{w}}'_a \hat{\mathbf{w}}_a]^{-1}$$

5. Calculate the loading weights **c** by regressing $\mathbf{Y}_{a-1}$ onto the spectral scores according to the local model

$$\mathbf{Y}_{a-1} = \hat{\mathbf{t}}_a \mathbf{c}'_a + \mathbf{F}$$

The weights are found by least squares estimation minimising the residuals in **F**

$$\hat{\mathbf{c}}_a = \mathbf{Y}'_{a-1}\, \hat{\mathbf{t}}_a [\hat{\mathbf{t}}'_a \hat{\mathbf{t}}_a]^{-1}$$

6. Scale the length of **c**

$$\hat{\mathbf{c}}_a = \frac{\hat{\mathbf{c}}_a}{\sqrt{(\hat{\mathbf{c}}'_a \hat{\mathbf{c}}_a)}}$$

7. Calculate a new score vector **u**

$$\hat{\mathbf{u}}_a = \mathbf{Y}_{a-1} \hat{\mathbf{w}}_a [\hat{\mathbf{w}}'_a \hat{\mathbf{w}}_a]^{-1}$$

8. Check for convergence: If **u** in No.7 is different to **u** in No. 1 then return to No.2

9. Calculate the **X** and **Y** loading vectors, **p** and **q**

$$\hat{\mathbf{p}}'_a = (\hat{\mathbf{t}}'_a \hat{\mathbf{t}}_a)^{-1} \hat{\mathbf{t}}'_a \mathbf{X}_{a-1}$$

$$\hat{\mathbf{q}}'_a = (\hat{\mathbf{u}}'_a \hat{\mathbf{u}}_a)^{-1} \hat{\mathbf{u}}'_a \mathbf{Y}_{a-1}$$

10. Calculate the residuals

$$\hat{\mathbf{E}} = \mathbf{X}_{a-1} - \hat{\mathbf{t}}_a \hat{\mathbf{p}}'_a$$

$$\hat{\mathbf{F}} = \mathbf{Y}_{a-1} - \hat{\mathbf{t}}_a \hat{\mathbf{q}}'_a$$

11. The data matrices are then reassigned as the residuals for the next dimension

$$\mathbf{X}_a = \hat{\mathbf{E}} \qquad \mathbf{Y}_a = \hat{\mathbf{F}}$$

Once the optimal number of dimensions has been established (see Chapter 6), the scores vectors for unknown samples can be calculated according to the following sequence.

For each dimension, *a*:

1. Calculate a new scores vector according to the model

$$\mathbf{x}_{a-1} = \mathbf{t}_a \hat{\mathbf{w}}'_a + \mathbf{E}$$

minimising the residual by least squares estimation

$$\hat{t}_{i,a} = \mathbf{x}'_{i,a-1} \hat{\mathbf{w}}_a$$

2. Calculate the residual and reassign **x**

$$\mathbf{x}_{i,a} = \mathbf{E} = \mathbf{x}_{i,a-1} - \hat{t}_{i,a} \hat{\mathbf{p}}'_{i,a}$$

Using these data the analyte concentrations can be estimated according to

$$\hat{y}_i = \bar{y} + \sum_{a=1}^{A} \hat{t}_{i,a} \hat{\mathbf{q}}_a$$

Chapter Four

Multicomponent analysis of a model spectrophotometric data set

4.1 INTRODUCTION

Relative to its application in vibrational spectroscopy, multivariate calibration has been little used in UV-visible spectrophotometry. This can largely be attributed to the nature of the procedures applied using these techniques. In very generalised terms, IR spectroscopists tend to quantify the spectroscopic data derived from virgin or matrix isolated samples. In the mid-IR region, it is sometimes possible to find a fundamental bend or stretch that is directly attributable to the analyte, or analytes, of interest. In such cases, providing that a reproducible base-line can be established then univariate procedures can be more than adequate. It is more likely however, that due to known and unknown interferences, single frequency procedures will be inadequate and the multivariate calibration routines discussed in Chapter 3 will be more suitable. The case is accentuated as one moves into the near-IR region where the assignment of combination and overtone bands can be ambiguous. Here multivariate techniques are essential to the success of real sample analyses.

This is in contrast to the manner in which UV-visible spectrophotometry is applied. Traditionally, quantitative spectrophotometric measurements are carried out in the latter stages of a derivatisation procedure. Such procedures are designed to produce a highly absorbing chromophore, the absorbance of which is directly proportional to the analyte of interest. Implicit in this procedure is not only the enhancement of analytical sensitivity but also minimisation of the effect of potential interferents, ie. analytical selectivity. Accordingly, UV-visible spectrophotometry has tended to involve measurement at a single wavelength corresponding to the wavelength of maximum absorbance, followed by a univariate calibration. In the simplest terms, visible spectroscopists have generally utilised physico-chemical approaches for selectivity enhancement, whereas a more mathematical approach has been used in quantitative IR studies.

Multivariate analysis has, however, been applied to visible spectrophotometric

data for the resolution of multianalyte systems. Thus, in moving from single to multicomponent analysis, the other multivariate advantages such as interference removal and noise reduction are adopted.

The aim of this chapter is to investigate the relative predictive abilities of three of the multivariate calibration techniques discussed in the previous chapter. An extension of the transition metal 'model' system described by Wolf [171] has been selected as a means of evaluation. Rather than monitoring derivatisation products the model system relies on the inherent absorbance of the species under investigation, thus reducing the number of random error sources. In total, four visible spectrophotometric data sets were collected and subjected to direct multicomponent analysis, principal components regression and partial least squares regression. The data sets were designed to provide varying spectroscopic complexity and a range of physical and chemical interferences.

4.2 EXPERIMENTAL

Reagents

All solutions were prepared in Milli R-O water (Millipore) and all reagents were AnalaR grade (Merck). Solutions (0.1 mol dm^{-3}) of chromium (III) potassium sulphate 12-hydrate, iron (II) sulphate 7-hydrate, cobalt (II) sulphate 7- hydrate, nickel (II) sulphate 7-hydrate and copper (II) sulphate 5-hydrate were prepared in 1% v/v sulphuric acid. Barium sulphate was added, where indicated in the text, as the solid.

Apparatus

Absorbance and derivative spectra were measured using a Hewlett-Packard 8451A PDA fitted with a 1 cm path-length silica cuvette, and the data were stored using a HP 9121 disk drive. Data from the PDA were downloaded in ASCII format to a personal computer via a HP 82939A serial interface card

using 'Kermit' serial communication software.

Software

DMA was carried out using the weighted least squares on-board software of the PDA. PCR and PLSR were carried out using the Unscrambler v. 3.2 multivariate data analysis package (Camo A/S, Norway) which incorporates matrix handling routines thus allowing manipulation of ASCII files.

Procedures

The spectra of all solutions were measured in triplicate, against a 1% v/v sulphuric acid blank, with an integration time of 25 s. The spectra were averaged and their means were stored for use in calibration/prediction. A wavelength range of 302-800 nm with a 2 nm interval was used throughout yielding 250 data points per spectrum.

A three component system was developed by dilution of the Co (II), Ni (II) and Cu (II) solutions; 0.025M solutions of the metal sulphates were used in the DMA and the calibration set used for PCR and PLSR is shown in Table 4.1. The predictive ability of each method was determined using the test set also described in Table 4.1. Incorporation of Cr (III) gave a four component system, a 0.025M sulphate solution again being used for DMA. The calibration set for PCR and PLSR and the test set is given in Table 4.2. After measurement of the spectra, various amounts of barium chloride were added to the training and test set solutions in a non-quantitative manner, thus creating the effect of a physical interference due to the scatter and absorbance caused by the barium sulphate precipitate. DMA (with and without the barium sulphate standard spectrum), PCR and PLSR were repeated using these solutions which simulate suspended solids. Finally the Fe (II) solution was incorporated to give a five component system; the calibration and test sets are shown in Table 4.3.

Table 4.1 Concentration data of the training set and test set for the three-component system (mol dm^{-1})

Training set	Co (II)	Ni (II)	Cu (II)
A	0.025	0.025	0
B	0.025	0	0.025
C	0	0.025	0.025
D	0.010	0.010	0.010
E	0.010	0.005	0.020
F	0.005	0.010	0.020
G	0.010	0.020	0.005
H	0.005	0.020	0.010
J	0.020	0.010	0.005
K	0.020	0.005	0.010
Test set			
1	0.010	0.010	0.020
2	0.010	0.020	0.010
3	0.020	0.010	0.010
4	0.020	0.020	0.010
5	0.020	0.010	0.020
6	0.010	0.020	0.020
7	0.025	0.005	0
8	0.025	0	0.005
9	0.005	0.025	0
10	0	0.025	0.005
11	0.005	0	0.025
12	0	0.005	0.025
13	0.015	0	0
14	0	0.015	0
15	0	0	0.015

Table 4.2 Concentration data of the training set and test set for the four-component system (mol dm^{-1})

Training set	Cr (III)	Co (II)	Ni (II)	Cu (II)
L	0	0.025	0.025	0.025
M	0.025	0.025	0.025	0
N	0.025	0.025	0	0.025
O	0.025	0	0.025	0.025
P	0.010	0.010	0.010	0.010
Q	0.020	0.005	0.010	0.015
R	0.015	0.020	0.005	0.015
S	0.010	0.015	0.020	0
T	0.005	0.010	0.015	0
Test Set				
19	0.025	0	0.005	0.010
20	0.010	0.025	0	0.005
21	0.005	0.010	0.025	0
22	0	0.005	0.010	
23	0.020	0.010	0.010	0.025
24	0.010	0.010	0.010	0.010
25	0.010	0.010	0.020	0.020
26	0.010	0.020	0.010	0.010
27	0	0.025	0.020	0.010
28	0	0	0.025	0
29	0.020	0	0	0.020
30	0.025	0.020	0	0.025
31	0.015	0	0.015	0
32	0	0.015	0	0
33	0.015	0.015	0	0.015
34	0	0	0.015	0
35	0.015	0.015	0.015	0.015
				0.015

Table 4.3a Concentration data of the training set for the five-component system (mol dm^{-1})

Training set	Cr (III)	Fe (II)	Co (II)	Ni (II)	Cu (II)
A	0.050	0.050	0	0	0
B	0.050	0	0.05	0	0
C	0.050	0	0	0.050	0
D	0.050	0	0	0	0.050
E	0	0.050	0.050	0	0
F	0	0.050	0	0.050	0
G	0	0.050	0	0	0.050
H	0	0	0.050	0.050	0
I	0	0	0.050	0	0.050
J	0	0	0	0.050	0.050
K	0.033	0.033	0.033	0	0
L	0.033	0	0.033	0	0.033
M	0.033	0.033	0	0.033	0
N	0.033	0.033	0	0.033	0
O	0.033	0.033	0	0	0.033
P	0.033	0	0	0.033	0.033
Q	0	0.033	0.033	0.033	0
R	0	0.033	0.033	0	0.033
S	0	0.033	0	0.033	0.033
T	0	0	0.033	0.033	0.033
U	0.025	0.025	0.025	0.025	0
V	0.025	0.025	0.025	0	0.025
W	0.025	0.025	0	0.025	0.025
X	0.025	0	0.025	0.025	0.025
Y	0	0.025	0.025	0.025	0.025
Z	0.020	0.020	0.020	0.020	0.020

Table 4.3b Concentration data of the Test set for the five-component system (mol dm^{-1})

Test set	Cr (III)	Fe (II)	Co (II)	Ni (II)	Cu (II)
1	0	0.010	0.020	0.030	0.040
2	0.010	0.050	0.040	0	0
3	0.020	0.010	0.010	0.020	0.040
4	0.030	0.020	0	0.050	0
5	0.040	0	0.050	0	0.010
6	0.050	0.020	0	0	0.030
7	0.020	0	0.030	0.040	0.010
8	0.040	0.010	0	0	0.050
9	0	0.020	0.020	0.040	0.010
10	0	0.030	0.050	0	0.020
11	0.050	0.040	0	0.010	0
12	0	0.048	0	0.030	0.020
13	0.030	0.040	0	0.010	0.020
14	0	0	0.010	0.050	0.040
15	0.020	0.040	0.020	0.010	0.010
16	0.050	0	0.030	0.020	0
17	0	0.010	0.040	0	0.050
18	0	0.030	0.050	0.020	0
19	0.040	0.010	0.020	0	0.0320
20	0	0.050	0.040	0.010	0
21	0.040	0.010	0.010	0.020	0.020
22	0	0.020	0	0.030	0.050
23	0.010	0	0.050	0.040	0
24	0.030	0.020	0	0.050	0
25	0.010	0.020	0.030	0.040	0
26	0.050	0.040	0	0	0.010
27	0.020	0.010	0.010	0.020	0.040
28	0.020	0	0.050	0	0.030
29	0	0.050	0	0.010	0.040
30	0.020	0	0	0.030	0.050

All PCR and PLSR models were developed from mean-centred data and the optimal dimensionality was defined as the first local minimum of the PRESS (prediction error sum of squares) relative to the number of factors used. The PRESS is defined as:

$$\text{PRESS} = \sum_{i=1}^{I} (y_i - \hat{y}_i)^2$$

where y_i is the concentration of object i, $\hat{y}_i$ is the predicted concentration of object i and I is the total number of objects used in the calibration. The standard error of prediction (SEP) is calculated from the PRESS and has units the same as the original concentration data.

$$\text{SEP} = \sqrt{\frac{\text{PRESS}}{I}}$$

The SEP is also referred to as the root mean square error (RMSE), which can be expressed in relative terms (similar to the relative standard deviation) as the relative RMSE.

$$\text{RRMSE} = \frac{100}{\bar{y}} . \text{SEP}$$

where $\bar{y}$ is the mean analyte concentration. The RRMSE is used for comparisons of both the cross validation models (RRMSECV) and the prediction of an independent test set (RRMSEP). In both cases no degrees of freedom are lost. In this work a hybrid of the RRMSEP has been used for comparison of the predictive ability. Defined as the relative error of prediction (REP (%)), it represents the cumulative RRMSEP for all analytes predicted by the model

$$\text{REP}\,(\%) = \frac{100}{\bar{y}_{ij}} \cdot \sqrt{\frac{\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij}-\hat{y}_{ij})^2}{N}}$$

where $\bar{y}_{ij}$ is the mean concentration of all the analytes in the prediction set, $\hat{y}_{ij}$ is the predicted concentration of analyte j in sample i, y_{ij} is the true concentration of analyte j in sample i and N is the total number of predictions ($I.J$).

All PLSR models were developed in the PLS-2 mode.

4.3 RESULTS & DISCUSSION

The absorbance and first derivative spectra of the five metal ion standard solutions are shown in Figs. 4.1 & 4.2. The relative prediction errors for the three calibration methods used for the three component system using absorbance, first derivative and second derivative data are given in Table 4.4. The use of first derivative data with DMA has led to significantly better predictions than DMA with absorbance data and the PCR and PLSR methods for absorbance and first derivative data. There was no significant difference in the REP from the absorbance and first derivative data with PCR and PLSR. With all three calibration methods the second derivative data yielded significantly less accurate predictions.

Analysis of the results for the four component system revealed no significant difference in the REP values between the three procedures for absorbance or first derivative data. The second derivative data again yielded much less accurate predictions. The relative prediction errors are given in Table 4.5. Fig. 4.3a shows the absorbance spectra for the calibration set for comparison with

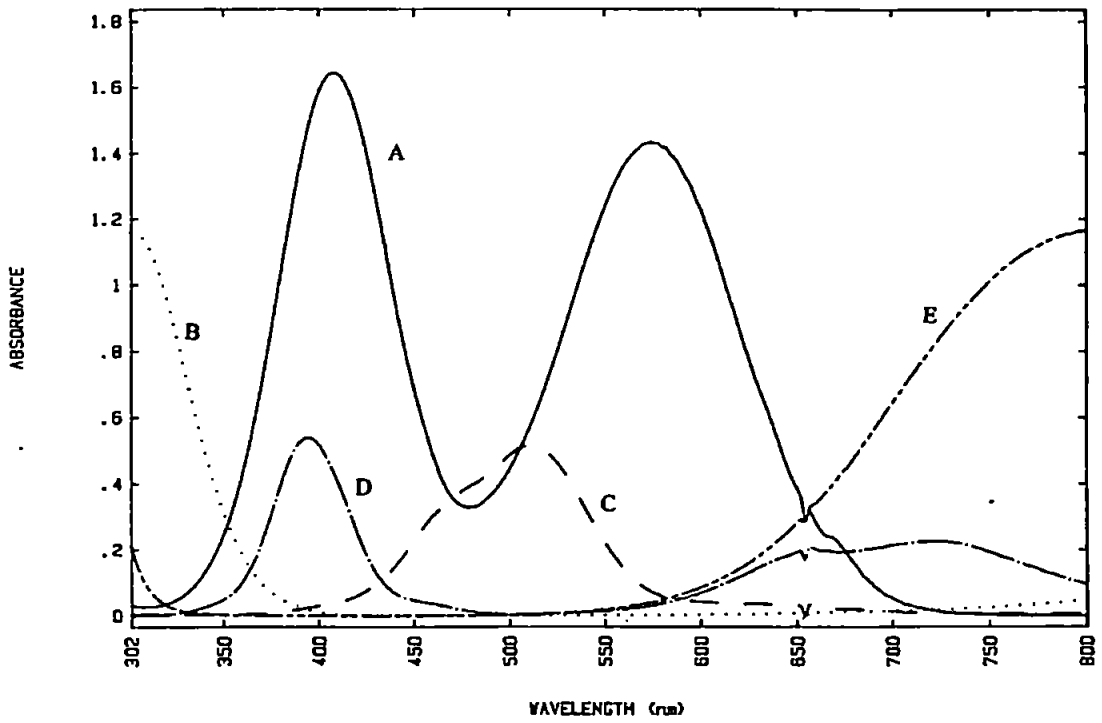


Figure 4.1 Absorbance spectra of the five standard metal sulphate solutions: A, Cr; B, Fe; C, Co; D, Ni; and E, Cu.

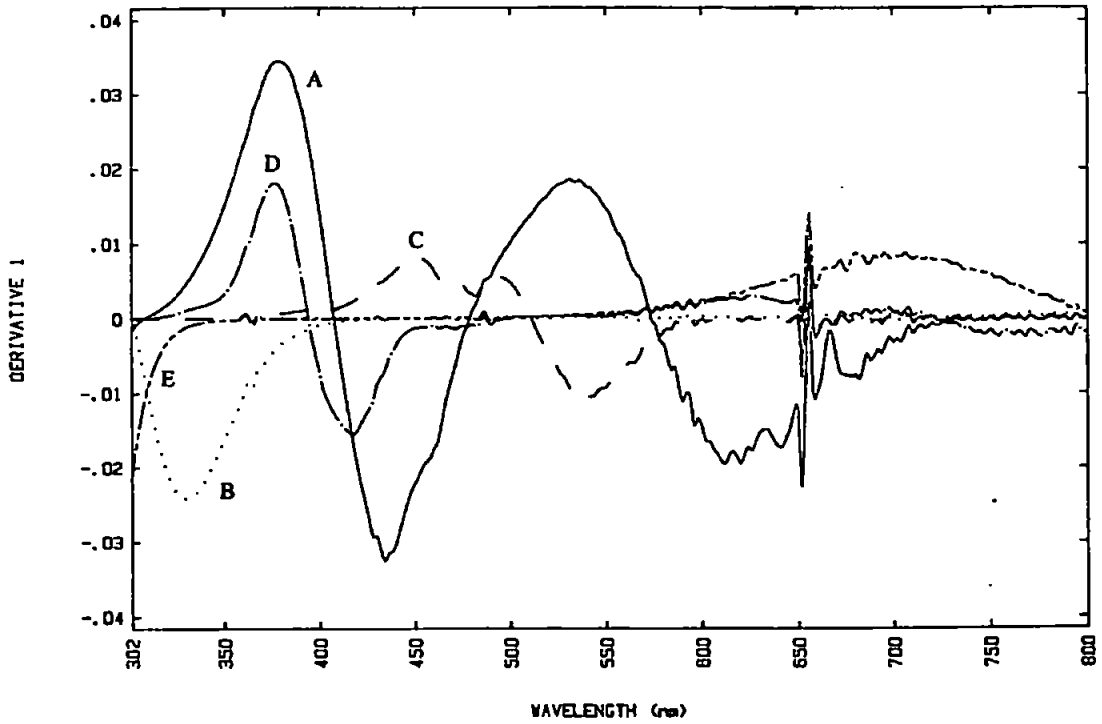


Figure 4.2 First-derivative spectra of the five standard metal sulphate solutions: A, Cr; B, Fe; C, Co; D, Ni; and E, Cu.

procedures with the barium sulphate interference are summarised in Table 4.6. As expected, the prediction ability of DMA is very poor when the barium sulphate interference is not included as a calibration standard. When the interference is included the predicted values with DMA are still significantly less accurate than those obtained with PCR and PLSR using both absorbance and first derivative data. There is no significant difference between the REP values for the absorbance and first derivative data in either PCR or PLSR. The second derivative data once again yields results markedly less accurate for all three methods.

Table 4.4 Relative error of prediction values for the three-component system.

	Absorbance		First derivative		Second derivative	
	REP(%)	Dim.	REP(%)	Dim.	REP(%)	Dim.
DBA	3.91	-	1.20	-	107.53	-
PCR	3.62	3	2.70	3	13.66	5
PLSR	3.61	3	2.69	3	11.88	5

Table 4.5 Relative error of prediction values for the four-component system.

	Absorbance		First derivative		Second derivative	
	REP(%)	Dim.	REP(%)	Dim.	REP(%)	Dim.
DMA	1.86	-	2.08	-	27.73	-
PCR	2.47	4	2.15	4	10.23	5
PLSR	2.47	4	2.15	4	10.21	5

The five component system incorporating Fe (II) presents the most challenging problem for calibration and prediction due to partial oxidation of the ferrous ion in the presence of copper, which markedly effects the visible spectrum.

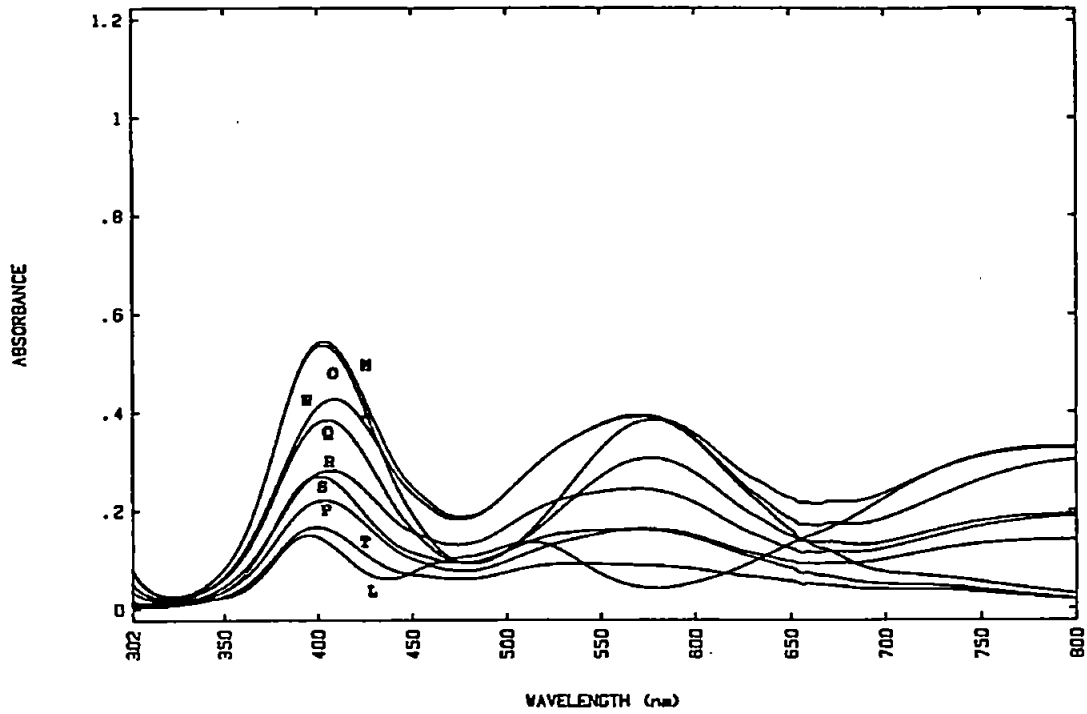


Figure 4.3a Absorbance spectra of the four-component test set (refer to Table 4.2 for concentration data).

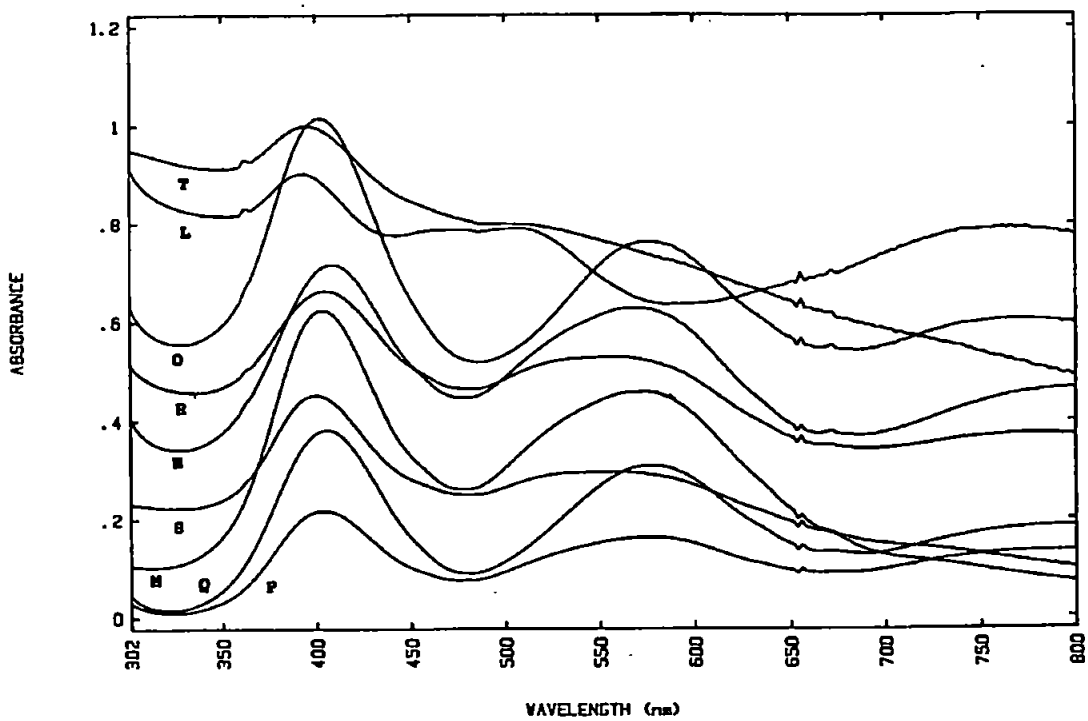


Figure 4.3b Absorbance spectra of the four-component test set after the non-quantitative addition of barium chloride.

Analysis of the prediction residuals revealed that soft modelling techniques offer a much more robust calibration procedure in this particular case. There was no significant difference between the REP values obtained by the PCR and PLSR methods for both absorbance and first derivative data. The relative error of prediction values for each procedure are presented in Table 4.7. The linear regression data for the regression of the predicted values on the true values and the relative error of prediction for each analyte from the PLSR model with absorbance data are presented in Table 4.8. It can be seen that the analytes best modelled are those with the most distinct spectra; Fe (II) having the poorest relative error of prediction due to the interferences explained above.

Table 4.6 Relative error of prediction values for the four-component system with barium sulphate interference.

	Absorbance		First derivative		Second derivative	
	REP(%)	Dim.	REP(%)	Dim.	REP(%)	Dim.
DMA 1	484.92	-	33.55	-	122.09	-
DMA 2	21.52	-	9.52	-	126.49	-
PCR	2.28	6	4.61	5	20.51	5
PLSR	2.28	6	4.58	5	19.9	5

Table 4.7 Relative error of prediction values for the five-component system.

	Absorbance		First derivative	
	REP(%)	Dim.	REP(%)	Dim.
DMA	110.77	-	100.87	-
PCR	6.40	5	8.25	6
PLSR	6.40	5	8.24	6

Table 4.8 Linear regression data for the regression of the predicted values on the true values for individual analytes from the PLSR calibration on the absorbance data.

Analyte	Slope	Intercept	Correlation coefficient	REP (%)
Cr	1.021 ± 0.005	0.000 ± 0.000	1.0000	3.44
Fe	1.015 ± 0.030	-0.002 ± 0.001	0.9984	11.89
Co	1.013 ± 0.009	0.000 ± 0.000	0.9999	3.55
Ni	1.016 ± 0.021	0.000 ± 0.001	0.9993	4.57
Cu	1.021 ± 0.010	0.000 ± 0.000	0.9998	3.83

4.4 CONCLUSIONS

The DMA procedure provides an accurate means of prediction in the well behaved three and four component systems. However, under less well behaved circumstances, the PCR and PLSR routines offer more robust models by implicitly accounting for interferences in the calibration stage.

The use of second derivative data consistently led to significantly less accurate predictions due to the much poorer signal-to-noise ratio.

No significant difference was observed in the predictive ability of the PCR and PLSR routines in any of the experiments.

Chapter Five

Partial least squares resolution of multianalyte FIA data

5.1 INTRODUCTION

In the preceding chapters, the concepts of process FIA and multivariate calibration of UV-visible spectrophotometric data have been investigated. The aim of this chapter is to draw these two threads together and present multianalyte FIA suitable for the process environment.

Multivariate calibration has been applied to data from a number of FIA determinations as shown in Table 5.1. Blanco et al. [171] applied DMA to the absorbance and derivative spectra of 2, 3 and 4 component mixtures of etafedrine, phenylephrine, doxylamine and theophylline using FIA as a sample presentation technique for the PDA. The same authors also compared univariate, DMA and MLR methods for the FIA-PDA speciation of iron [172]. MLR has also been applied to the resolution of ternary mixtures of aromatic amines after retention on a C_{18} bonded silica support in a PDA flow-cell [179].

The first application of PLSR to FIA data was published in 1988 by Lukkari and Lindberg [174]. They exploited a gradient system for the simultaneous FIA-titration of up to five organic acids by utilising firstly the signal shape at a single wavelength and secondly the absorbance vs. time matrix from a PDA. The second-order data for each sample was unfolded to a vector (and bunched) before calibration. Gerritsen *et al* used PLSR to quantify teniposide in blood plasma [181] thus demonstrating the resolving power of this technique in the presence of an interfering matrix. MLR, PCR and PLSR have recently been applied to first-order data for the determination of nickel and iron by FIA utilising a double-injection zone penetration technique [184], and PLSR has been used for multicomponent analysis of FIA-FTIR data [185]. There is as yet no reported use of an on-line FIA-multidetection system, with or without multivariate calibration, for process analysis.

Table 5.1 Examples of the application of multivariate calibration techniques to FIA data.

Analytes	Calibration Method	Reference
Pharmaceutical compounds	DMA	172
Fe(II)/Fe(III)	DMA,MLR	173
Organic acids	PLSR	174
Lanthanoids	MLR	175
Cu/Fe	MLR	176
pH indicators	SMCR	177
Ca/Mg	DMA	178
Th(IV)/La(III)	MLR	179
2,4-DNPH/2-NPH/4-NPH	MLR	180
Teniposide	PLSR	181
Fe/free acid	PLSR	182
Rare earth metals	MLR	183
Ni(II)/Fe(II)	MLR,PCR,PLSR	184
Acetone/ethanol/THF	PLSR	185

The conclusions of Chapter 4 indicate that in ideal circumstances DMA performs no worse than PCR and PLSR for the multicomponent resolution of visible spectrophotometric data. However, when physical and chemical interferences were incorporated into the experiments, the bilinear modelling techniques consistently produced significantly better predictions. The same experiments revealed no significant difference between the predictions made by PCR and PLSR. These findings agree with those of other workers [150,151,152,153] but if one considers its theoretical advantages and optimal performance over a wide range of conditions [154], PLSR can be considered the general method of choice.

This chapter describes the development of a combined reaction FIA system with

photodiode array detection and data treatment with PLSR. The primary objective is to demonstrate the feasibility of this integrated approach for simultaneous multianalyte determinations in a process environment. With this in mind, the emphasis is on the investigation of a number of calibration criteria using a physically simple manifold. Nonetheless, the model system considered here is a real one. Zinc phosphate and chlorine are added to industrial cooling water systems as a corrosion inhibitor and biocide respectively, and on-line information is desirable for control purposes. The practical implications of combining established spectrophotometric methods for analytes of a diverse nature are considered and the influences of a number of calibration parameters are considered in detail.

5.1 EXPERIMENTAL

Reagents

All solutions were prepared in Milli-Q water (Millipore) and all reagents were of AnalaR grade (Merck) unless otherwise indicated. A stock phosphate solution containing 1000 mg l^{-1} phosphorus (PO_4-P) was prepared by dissolving 4.390 g potassium dihydrogen orthophosphate (dried for 2 h at 105 °C) in 1 l of water. A stock hypochlorite solution containing 1000 mg l^{-1} free chlorine as chlorine was prepared by dilution of an appropriate volume of iodometrically standardised sodium hypochlorite solution (Merck; general purpose reagent). Calibration and test set solutions were prepared by serial dilution of these stock solutions and are subsequently referred to as phosphate and chlorine solutions. The DPD reagent was prepared by dissolving 1.5 g N,N-diethyl-p-phenylene diamine sulphate (Aldrich) (4-N,N-diethylaminoaniline sulphate) in 1 l of water. The acid/molybdate reagent was prepared by dissolving 10 g of ammonium heptamolybdate in 1 l of 0.4 M nitric acid and the ascorbic acid solution was prepared by dissolving 80 g in 1 l of water. A solution of *o*-tolidine was prepared by dissolving 0.86 g of *o*-tolidine dihydrochloride (Fluka; purum) in 2 M hydrochloric acid.

Caution: *o*-tolidine is highly toxic and should be handled with extreme care.

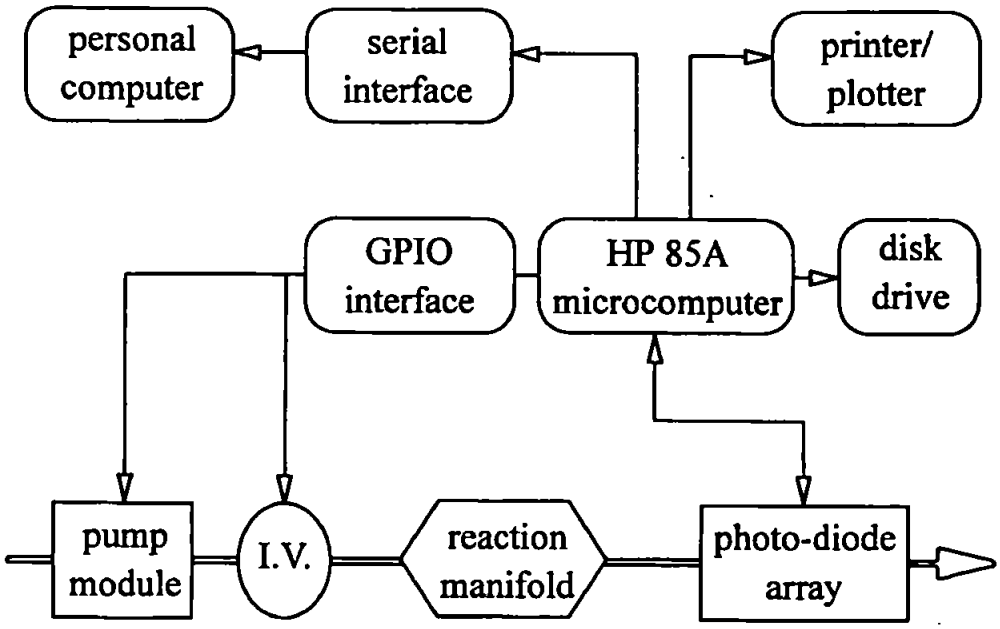


Figure 5.1 Schematic diagram of the automated FIA-PDA system.

Instrumentation

A schematic diagram of the automated FIA-PDA arrangement is shown in Fig. 5.1. The FIA manifold was constructed from 0.8 mm i.d. polytetrafluoro-ethylene (PTFE) tubing and in-house PTFE T pieces. Absorbance spectra were measured using a Hewlett-Packard HP 8451A PDA spectrophotometer fitted with an 18 µl glass flow cell with a path length of 1 cm (Hellma). Raw data were stored using an HP 9121 disk drive and output in ASCII format using an HP 82939A serial interface to a Viglen 386 DX personal computer with 8 Mb of RAM. All subsequent data processing was carried out using this computer. Sample injections (150 µl) were made using a pneumatic valve control unit (P.S. Analytical) and all solutions were propelled by two peristaltic pumps (Ismatec Mini S-820) with PVC pump tubing (Labsystems). Control of the valves and pumps was maintained via an HP 82940A GPIO interface.

Software

A general purpose program was written in HP basic to control the FIA components, measure and record spectra, and carry out some basic data processing. A further program was used to transmit spectral data to the personal computer via the serial interface. Kermit serial interface software version 3.01 was used to collect and store data in ASCII format on the personal computer. All multivariate data analysis was carried out using Unscrambler II Extended version 4.00 (Camo A/S, Norway) which incorporates matrix handling routines allowing manipulation of the ASCII files.

Procedures

Batch Experiments

In order to evaluate the compatibility of the phosphate and chlorine reaction chemistries, a number of preliminary experiments were carried out. The visible spectra of combinations of the molybdate, DPD and *o*-tolidine reagents were recorded after addition of combinations of water and solutions of 10 mg l^{-1} phosphate and 10 mg l^{-1} chlorine.

Flow Injection Experiments

The FIA manifold used in all experiments is shown in Fig. 5.2. The absorbance was measured every 2 nm over the wavelength range of 352-550 nm yielding 100 data points per spectrum and one spectrum was recorded every second between 1 and 60 s after injection. This resulted in a total of 6000 data points for each injection. All spectra were measured against a reagent blank. The control software was designed to calculate and store to disk the mean spectrum of the three spectra nearest to the peak maximum for each injection. All solutions were measured in triplicate and the overall mean spectrum of the three injections was also stored to disk. This overall mean spectrum was used for all subsequent data processing unless otherwise stated.

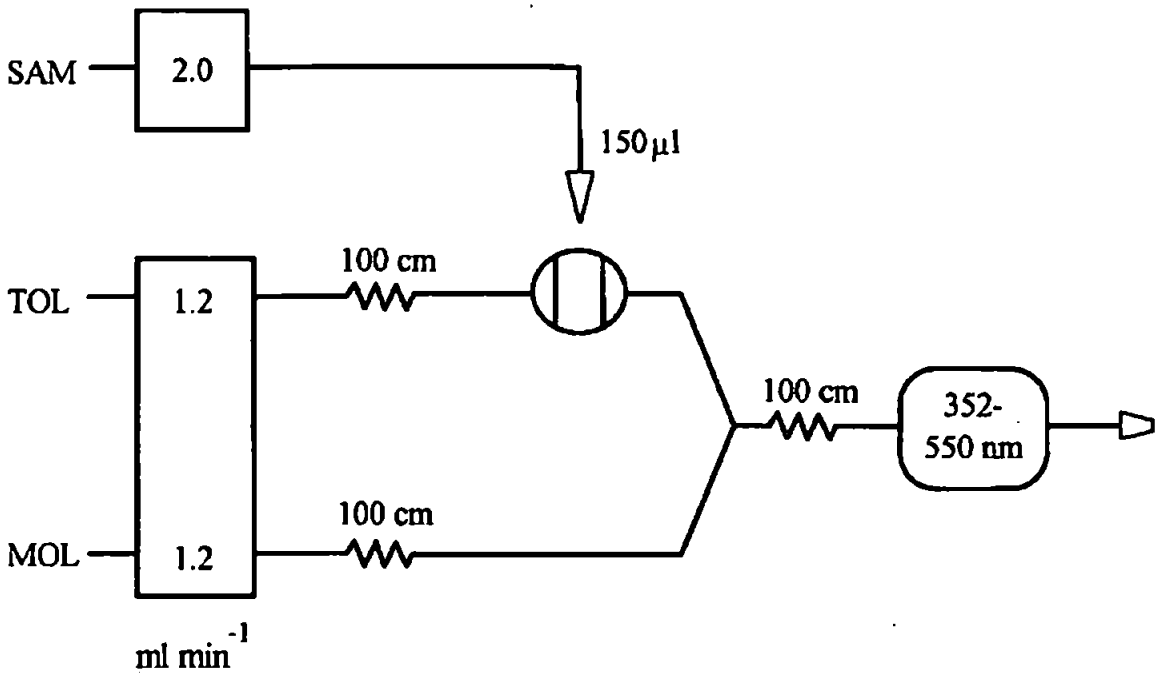


Figure 5.2 Flow-injection manifold for the simultaneous determination of phosphate and chlorine: SAM, sample; TOL, *o*-tolidine; and MOL, acid-molybdate.

Calibration set solutions (training set) were prepared to cover the range 2-10 mg l^{-1} phosphate and 1-5 mg l^{-1} chlorine in a 5-level factorial design. A further 20 samples were prepared and analysed 48 hours later as an independent test set. The concentration details are given in Table 5.2. Both the calibration and test sets were analysed in random order to reduce any drift effects.

All PLSR models were developed in PLS-2 mode and the optimal dimensionality was defined as the first local minimum of PRESS relative to the number of factors included.

Table 5.2 Concentration data of the calibration set and test set (mg l^{-1})

Sample Number	Calibration Set		Test Set	
	PO_4	Cl	PO_4	Cl
1	2	1	3	1
2	2	2	3	2
3	2	3	3	3
4	2	4	3	4
5	2	5	3	5
6	4	1	5	1
7	4	2	5	2
8	4	3	5	3
9	4	4	5	4
10	4	5	5	5
11	6	1	7	1
12	6	2	7	2
13	6	3	7	3
14	6	4	7	4
15	6	5	7	5
16	8	1	9	1
17	8	2	9	2
18	8	3	9	3
19	8	4	9	4
20	8	5	9	5
21	10	1	-	-
22	10	2	-	-
23	10	3	-	-
24	10	4	-	-
25	10	5	-	-

5.3 RESULTS & DISCUSSION

Batch Experiments

One reason for the widespread use of FIA techniques is the breadth of established spectrophotometric procedures that can be implemented. The analytes under investigation in this study are routinely determined by spectrophotometric procedures; the molybdenum blue method for phosphate and the DPD method for chlorine [186]. Furthermore, both reaction chemistries have been successfully used in FIA methods for phosphate [39,187,188] and chlorine [189,190]. Initial experiments were conducted to combine these reaction chemistries to enable simultaneous determinations. However, batch experiments revealed that the two procedures were incompatible due to differing pH requirements; acid media for the molybdate reaction and pH 6.2-6.5 for the DPD reaction [186]. Another method for the spectrophotometric determination of chlorine uses *o*-tolidine and can be carried out over a wide pH range [191]. This reaction has also been used in an FIA method [192]. However, when the *o*-tolidine reaction was combined with the molybdenum blue reaction the chlorine response was lost completely. This was found to occur instantaneously upon addition of the ascorbic acid solution. Ascorbic acid is added in the determination of phosphate to reduce phosphomolybdic acid to the molybdenum blue complex. The monitorand for chlorine, in contrast, is a holoquinone; the product of chlorine oxidising the *o*-tolidine. Stannous chloride was found to have the same effect on the chlorine reaction suggesting that the holoquinone is being reduced by the ascorbic acid. Nevertheless, the yellow phosphomolybdic acid can also be monitored spectrophotometrically, thus eliminating the need for ascorbic acid reduction. This approach, while less sensitive than the molybdenum blue approach, has also been used in FIA [193].

Flow Injection Experiments

The successful implementation of process analytical methods requires the

fulfilment of a number of important criteria, one of the most important of which concerns instrument reliability. Any instrumentation which is to be installed in a manufacturing environment needs to be robust and the overall procedure must be dependable, especially if the information is going to be used for process control. In FIA terms, the manifold design must be kept as physically simple as is permissible with the analytical requirements. This would be a single line manifold in ideal situations. In this work, a single injection, two line manifold with one detector and a second pump for sample loop filling was used throughout (Fig. 5.2). This configuration was required for two reasons. Firstly, injection of sample into a molybdate stream caused a large negative response due to reagent dilution and secondly, a mixed molybdate/*o*-tolidine reagent was found to be unstable.

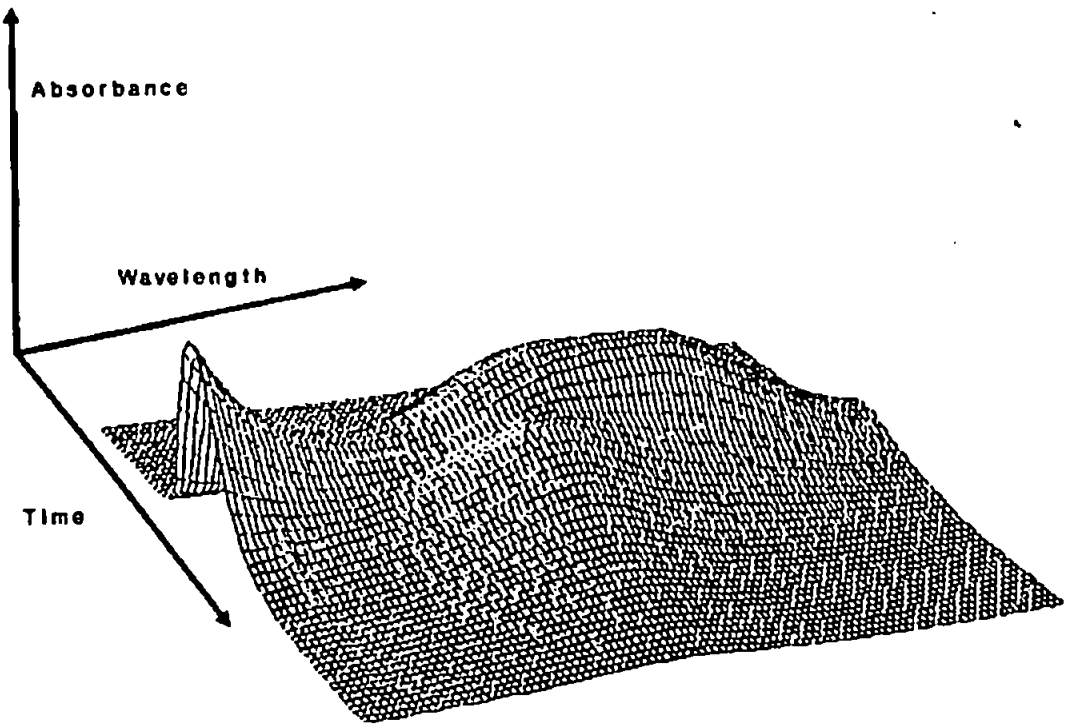


Figure 5.3 3-D FIA response profile for a solution containing chlorine at 5 mg l^{-1} and phosphate at 10 mg l^{-1}.

A typical 3 dimensional FIA response profile obtained from this manifold is shown in Fig. 5.3., representing 60 spectra measured at 2 nm intervals over the 352-550 nm range. Fig. 5.4 shows the mean spectra recorded at the FIA peak

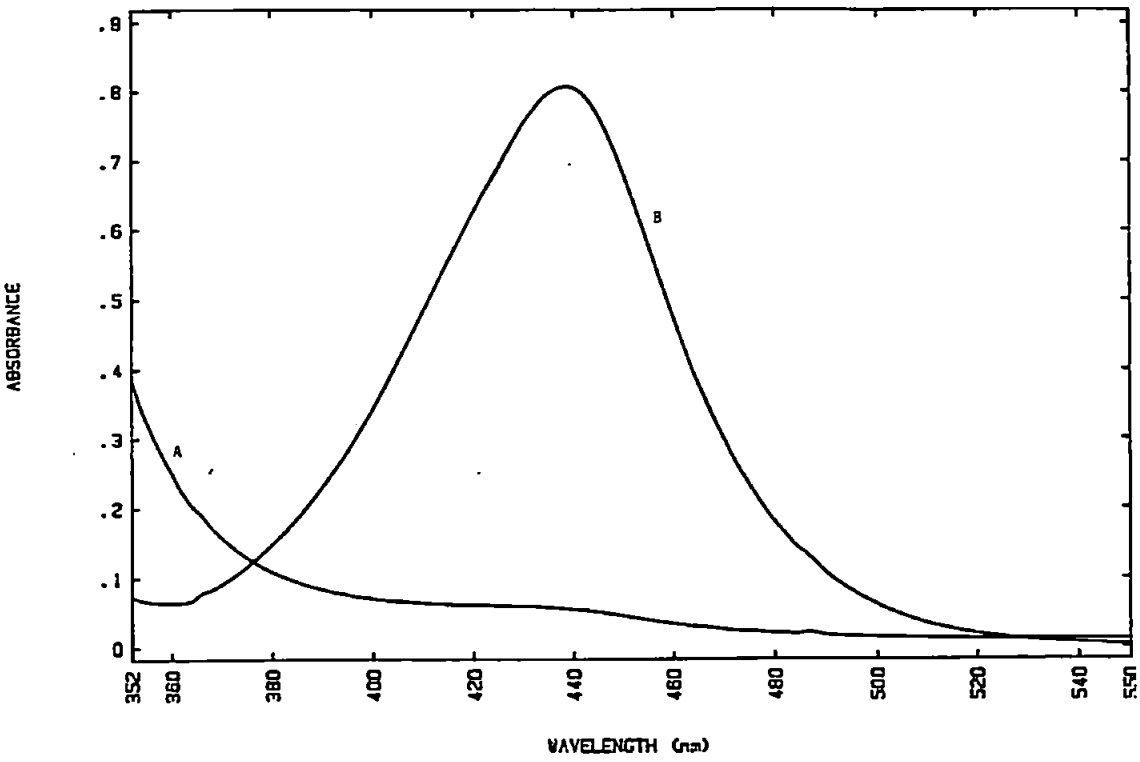


Figure 5.4 Mean spectra recorded at the FIA peak maximum for solutions containing: A, 10 mg l^{-1} phosphate; and B, 5 mg l^{-1} chlorine.

maximum for phosphate only and chlorine only standards. The univariate procedures on which this work was based used 362 nm for phosphate [193] and 438 nm for chlorine [192] and it can be seen that the same spectral regions are active after combination of the reaction chemistries. It is obvious from Fig. 5.3, however, that when phosphate and chlorine are present in the same solution a more complex picture arises. Most noticeable is the emergence of a shoulder at wavelengths greater than 460 nm in the chlorine active region of the spectrum. This is more distinct in Fig. 5.5 which shows the mean spectra recorded at the FIA peak maximum for each of the 25 calibration standards of the 5^2 experimental design. For reasons of clarity the spectra have not been labelled. Variance in the chlorine active region of the spectra is particularly evident and grouping of equal chlorine concentration samples is noticeable, especially at lower concentrations.

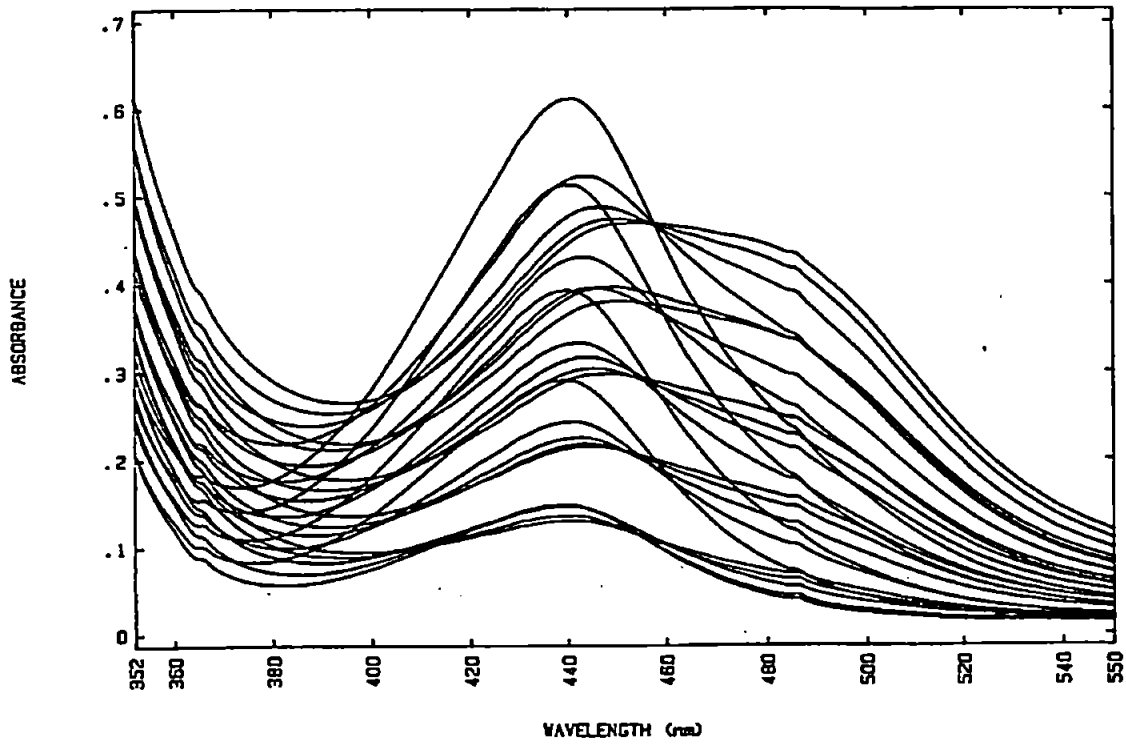


Figure 5.5 Mean spectra recorded at FIA peak maximum for each of the 25 solutions of the 5^2 experimental design.

Calibration

The first three PLSR loading vectors for the 5^2 calibration model are shown in Fig. 5.6. In the process of PLSR modelling, the covariance between the spectral scores and a single analyte is maximised. This often leads to the loadings of the first PLSR factor approximating to the pure component spectrum of the analyte under examination. PLS-2 however, maximises the covariance between the spectral scores and a linear combination of a number of variables (2 in this case). The physical significance of the loadings therefore becomes less clear. Inspection of the plot of the scores of the first PLSR factor versus the second factor reveals a very interesting structure (Fig. 5.7). The samples are aligned, as expected, in the order of the 5^2 experimental design but not in an equidistant fashion. This is particularly noticeable between the samples containing 2 and 4 mg l^{-1} phosphate, where the distance between pairs of samples of equal chlorine concentration increases with chlorine concentration. This would suggest some kind of non-linear relationship caused by the combination of the

phosphate and chlorine reaction chemistries [194,195]. Although PLSR is a linear method, it can handle non-linearities by the inclusion of additional factors [140] and this could explain the need for three factors to describe a two component system.

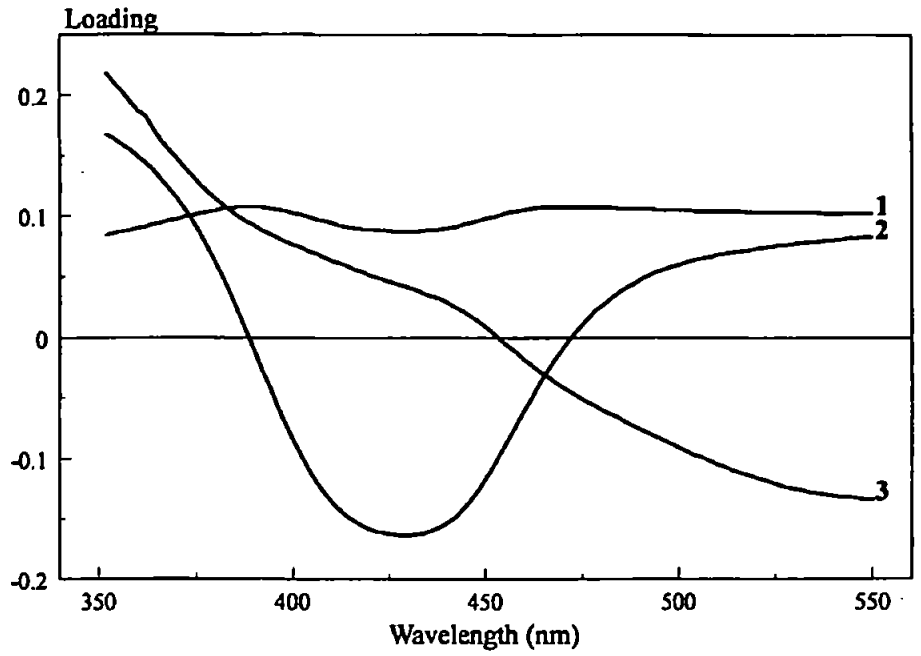


Figure 5.6 Overlay of the loading vectors of the first three PLS-2 factors as a function of wavelength.

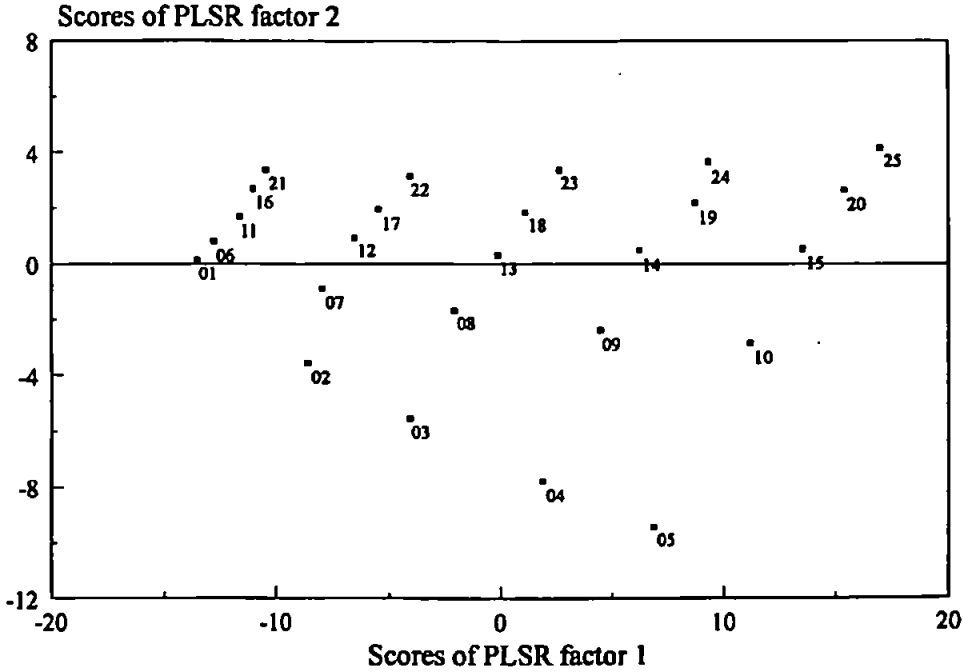


Figure 5.7 PLS-2 scores of factor 1 *versus* factor 2.

Preprocessing

The effect of a number of preprocessing techniques on the RRMSECV of the 5^2 experimental design are shown in Table 5.3. Mean-centring [196] is traditionally applied in PCR and PLSR and, as the name suggests, involves the subtraction of the variable mean from the individual variable values. Whilst the model dimensionality has not been reduced by mean-centring in this case, the phosphate predictions are significantly improved. Setting all variables to equal variance by dividing the mean-centred values by their standard deviation is known as autoscaling and it can be seen that autoscaling has had a small but beneficial effect on this data set. Normalisation on the other hand, which sets all spectra to unit length, has had a grossly detrimental effect.

Table 5.3 Effect of a number of preprocessing techniques on the relative prediction errors of PLSR and PCR models

Preprocessing Techniques	PLSR			PCR		
	No. Factor	RRMSECV		No. Factor	RRMSECV	
		PO_4	Cl		PO_4	Cl
None	3	11.9	1.9	3	12.0	1.9
Mean-centring (MC)	3	5.4	2.0	3	5.6	1.9
MC & Autoscaling (AS)	3	4.0	2.4	3	4.0	2.4
MC AS & normalisation	6	12.8	14.7	5	15.1	15.1
MC AS & 1st deriv.	3	6.5	1.8	3	6.5	1.8
MC AS & 2nd deriv..	4	11.8	4.3	4	14.9	4.5

Spectral derivatives which can enhance resolution generally lead to a depreciation in the signal to noise ratio with each derivatisation. Both the first and second derivatisations had an overall detrimental effect on the RRMSECV of this data set. PCR models were also built using the preprocessed data and,

as expected, resulted in dimensionality and RRMSECV values very similar to those for PLSR.

Wavelength selection and averaging

The effect of the size of the spectral data matrix on the prediction ability was studied in two ways. Firstly, wavelength variables were simply selected from the original data set and used to build PLS-2 models after mean-centring and autoscaling. Selection was made by taking every second variable to reduce the number from 100 to 50 and the same approach was taken for the selection of the 25 and 10 point data sets. The 5 point data set was selected according to the perceived importance of the variables; 360, 400, 440, 470 and 510 nm were used. The RRMSECV values for the five models are given in Table 5.4 together with the RRMSEP values for the independent test set. The prediction error for chlorine is very stable with decreasing data set size but that for phosphate increases. In the second case, the data set was reduced by averaging the spectral variables before mean-centring and autoscaling. Inspection of Table 5.5 reveals that both the phosphate and chlorine predictions are stable to the data set averaging.

Table 5.4 Effect of wavelength selection on the relative prediction errors

Number of wavelengths	RRMSECV		RRMSEP	
	PO_4	Cl	PO_4	Cl
100	4.0	2.4	4.0	2.9
50	4.1	2.4	4.2	2.9
25	4.3	2.4	4.9	2.8
10	4.3	2.5	7.2	3.3
5	5.2	2.2	7.0	2.3

Table 5.5 Effect of wavelength averaging on the relative prediction errors.

Number of wavelengths	RRMSECV		RRMSEP	
	PO_4	Cl	PO_4	Cl
0 (100)*	4.0	2.4	4.0	2.9
2 (50)	4.0	2.4	4.0	2.9
4 (25)	4.0	2.4	4.0	2.9
10 (10)	4.0	2.4	4.1	2.9
20 (5)	4.1	2.3	4.2	2.9

* Values in parenthesis indicate the number of data points used.

In averaging the spectral variables the original data is largely retained, albeit in a modified form, whereas information is lost in wavelength selection. This could explain the small increase in RRMSEP for phosphate using the selected variable data sets. The practical implications of these findings are that full spectra should be collected and stored at the measurement stage and that some wavelength averaging could be carried out before model building. However, the only advantage of wavelength averaging is a reduction in the time taken for model building, which for data sets of this size is not problematic and, given the loss of qualitative information associated with reducing the data set size, it would be provident to use the full spectra.

Calibration design

The effect of the size of the calibration set on the RRMSEP of the independent test set was determined by reducing the number of levels of the experimental design. The four level design includes the samples at 2, 4, 8 and 10 mg l^{-1} phosphate and 1, 2, 4 and 5 mg l^{-1} chlorine, and the three level design was constructed from the 2, 6 and 10 mg l^{-1} phosphate and 1, 3 and 5 mg l^{-1} chlorine samples. The samples containing 2 and 10 mg l^{-1} phosphate and 1 and 5 mg l^{-1} chlorine made up the two level design and finally the 6 mg l^{-1} phosphate and 3

mg l^{-1} chlorine sample was included to give a calibration set of 5 samples. The results given in Table 5.6, show a general increase in the RRMSEP as the number of calibration samples is reduced. Nevertheless, this increase is not dramatic, and in a situation where analysis time is an important consideration, the use of a 9 sample calibration set requires only a small compromise in prediction error.

Table 5.6 Effect of reducing the size of the calibration set on the prediction errors of an independent test set.

Calibration design	Size of calibration set	RRMSEP	
		PO_4	Cl
5 level	25	4.0	2.4
4 level	16	4.5	3.5
3 level	9	4.7	3.3
2 level	4	6.8	3.7
2 level +1	5	6.3	3.2

Predictions

Finally the predicted values of the independent test set are given in Table 5.7. Predictions were made using the model built from the 5 level experimental design after mean-centring and autoscaling the data. The RSD of three replicate injections and the percentage difference between the added and calculated concentrations of phosphate and chlorine are listed. The absolute errors and the precision of these predictions would fulfil the process specifications for the on-line monitoring of phosphate and chlorine in industrial cooling waters.

Table 5.7 Predictions of the independent test set

Sample number	Phosphate			Chlorine		
	Added mg l^{-1}	Found mg l^{-1}	Diff. %	Added mg l^{-1}	Found mg l^{-1}	Diff. %
13	7.0	7.1	+1.4	3.0	2.8	-6.7
8	5.0	5.1	-2.0	3.0	2.8	-6.7
17	9.0	9.4	+4.4	2.0	1.9	-5.0
18	9.0	9.4	+4.4	3.0	3.0	-
6	5.0	5.2	+4.0	1.0	1.0	-
20	9.0	9.1	+1.1	5.0	5.1	+2.0
19	9.0	9.2	+2.2	4.0	3.9	-2.5
5	3.0	2.6	-13.0	5.0	4.9	-2.0
15	7.0	6.7	-4.3	5.0	5.0	-
2	3.0	2.9	-3.3	2.0	1.9	-5.0
9	5.0	4.8	-4.0	4.0	3.9	-2.5
7	5.0	5.1	+2.0	2.0	1.9	-5.0
1	3.0	3.0	-	1.0	1.0	-
11	7.0	7.0	-	1.0	1.0	-
3	3.0	2.8	-6.7	3.0	2.9	-3.3
10	5.0	4.5	-10.0	5.0	5.1	+2.0
4	3.0	2.7	-10.0	4.0	3.9	-2.5
16	9.0	9.0	0	1.0	1.0	-
12	7.0	7.0	-	2.0	1.9	-5.0
14	7.0	6.7	-4.3	4.0	4.0	-

5.4 Conclusions

A physically simple, combined reaction FIA system with PDA detection integrated with PLSR of the data has been shown to be a feasible approach to simultaneous multianalyte determinations.

The combination of established spectrophotometric methods is a non-trivial matter and judicious choice of reaction chemistries is required to avoid gross interference. Visual inspection of the scores and loadings of the multivariate calibration model has been shown to reveal some of the underlying effects of the reaction combination.

Mean-centring and autoscaling of the data sets were found to be profitable, whilst selection and averaging of the spectral variables had no beneficial effect. Reducing the number of calibration standards used in modelling increased the error of prediction, but not prohibitively so.

A procedure has been developed for the simultaneous determination of phosphate and chlorine and the prediction of analyte concentrations for an independent test set, prepared and analysed 48 h after calibration, yielded RRMSEP values of 4.0% for phosphate and 2.4% for chlorine.

Chapter Six

Jackknife estimation of PLS models

6.1 INTRODUCTION

Two of the most important stages in the development and implementation of multivariate calibration models are the estimation of optimal dimensionality and the estimation of the errors in prediction. Arguably, however, both of these fundamental aspects of the procedure have been somewhat neglected. It is the aim of this chapter to present some of the currently practised methods, highlight some of their shortcomings and investigate the potential of a different approach.

Model validation

Selection of the optimum number of factors or dimensions to be used for future predictions is critical to the success of reduced dimension multivariate calibration models. As discussed in Chapter 3, the inclusion of too many factors leads to overfitting and the incorporation of noise, whereas underfitting leaves important interactions unmodelled. The consequence, in either case, is poor prediction of future samples. Because the overall objective of the calibration procedure is the accurate prediction of future samples, then an estimate of predictive ability provides a good means of comparing the different dimensionalities.

As part of a validation exercise, the models are usually compared in terms of the predictive error sum of squares (PRESS). PRESS is calculated as follows

$$\text{PRESS} = \sum_{i=1}^{I} (y_i - \hat{y}_i)^2$$

and therefore gives a direct comparison of the actual analyte values and the values predicted by the model; the smaller the value of PRESS the closer the model fits the true values. The objects included in the calculation of the PRESS are governed by the type of validation implemented.

The most rigorous form of validation uses a completely new and independent

test set. This external validation approach was adopted for the transition metal model system discussed in Chapter 4. It is particularly suited to large, well understood data sets, from which a representative subset can be selected or synthetic samples can be easily prepared and analysed. Accordingly, it tends not to be used in routine studies because the very data being used for validation can enhance the quality of the model by being accounted for at the calibration stage.

Extension of this concept leads to a number of routines known as internal validation. These routines actively use the calibration data for measuring predictive ability. One form of internal validation is calibration fitting. Here the PRESS is calculated using all *n* objects from the calibration data set and as such does not consider forward prediction. However internal validation can be carried out in the predictive direction by subdividing the calibration set through cross validation . Full cross validation (leave one out) successively divides the data set (*n* objects) into a modelling subset (*n*-1) and a validation subsample until all possible divisions have been made. The value of PRESS is calculated at each dimension for each object left out, and hence the optimal model for prediction can be estimated.

The criterion for the choice of optimal dimensionality can be;

i) the global minimum in PRESS,

ii) the first local minimum in PRESS,

iii) related to the significance of incremental changes in PRESS,

iv) the visual inspection of the loadings vectors,

v) a combination or combinations of i) to iv).

Selection based on the absolute minimum in PRESS has been shown to have poor statistical properties [197] and is not considered further. The first local minimum, however, generally makes a good compromise between over and underfitting of data, while remaining computationally simple. Depending on the data under consideration, the PRESS may never reach a local minimum within

the number of factors being considered. This is often encountered for large data sets. In this case the first local minimum is in fact the global minimum. This, of course, may well be the optimal model, but in situations where the incremental difference in PRESS is very small, then some form of significance test could avoid overfitting. With any of these criteria, a visual inspection of the loadings vectors can be helpful in assessing the extent to which noise is being modelled. Assuming that the chosen criterion has been satisfied, then cross validation will have provided an estimated optimal model for prediction and an estimated measure of error associated with any future predictions. An apparently similar approach might be to implement jackknife theory in validation. As with cross validation, the jackknife is based on the leave one out principle and it can be used to estimate optimal dimensionality and prediction ability.

Jackknife theory

The jackknife is a general nonparametric method for reducing the bias in an estimator and for obtaining a measure of the estimator variance by sample reuse [198]. The statistical similarity between the jackknife and CV runs no deeper than this resampling of data [199]. The estimator was introduced by Quenouille [200] for bias reduction and this version was subsequently utilised by Tukey [201] to develop a general method for obtaining approximate confidence intervals. This was referred to as the "jackknife". The jackknife can be used to calculate estimators of the bias and variance of PLSR coefficients and, by implementing a "double-jackknife", an estimation of dimensionality can be made.

Consider the regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\varepsilon}$ is a matrix of random errors with mean 0 and variance σ^2 . PLSR

is used to calculate the regression coefficients $\boldsymbol{\beta}$ from the data **y** and **X**.

Estimation of the RMSEP of a future prediction, is possible under the assumption that the model holds for future observations.

The mean square error of prediction (MSEP) can be decomposed as follows:

$$\begin{aligned} \mathrm{MSEP}(\hat{\mathbf{y}}_0) &= \mathrm{var}(\hat{\mathbf{y}}_0) + \| E\hat{\mathbf{y}}_0 - \mathbf{x}_0'\boldsymbol{\beta} \|^2 \\ &= \sigma^2 + \mathbf{x}_0'\mathrm{var}(\hat{\boldsymbol{\beta}})\mathbf{x}_0 + \| \mathbf{x}_0' E\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \|^2 \end{aligned}$$

$$\text{where, variance of } \hat{\boldsymbol{\beta}} = \mathbf{x}_0'\mathrm{var}(\hat{\boldsymbol{\beta}})\mathbf{x}_0$$

$$\text{bias of } \hat{\boldsymbol{\beta}} = \| \mathbf{x}_0' E\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \|^2$$

The estimators of the bias and variance of $\hat{\boldsymbol{\beta}}$ are determined by jackknifing:

1. Leave out the *i*-th object.
2. Perform the dimension estimation on the remaining *I*-1 objects.
3. Calculate the regression coefficient $\hat{\boldsymbol{\beta}}_{-i}$
4. Calculate the prediction $\hat{\mathbf{y}}_{-i} = \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{-i}$
5. Repeat steps 1-3 for i=1,...,*I*.
6. From the *I* values of $\mathbf{y}_i$ and $\hat{\mathbf{y}}_i$,calculate the PRESS.
7. From the *I* values of $\hat{\boldsymbol{\beta}}_{-i}$,calculate the jackknife estimators of the bias and variance of $\hat{\boldsymbol{\beta}}$ according to Efron's formulae [198]:

$$\text{bias} = (I-1)\,(\hat{\boldsymbol{\beta}}_{-i} - \hat{\boldsymbol{\beta}})$$

$$\text{variance} = \frac{I-1}{I}\sum_{I=1}^{I}(\mathbf{P}_{*i} - \bar{\mathbf{P}}_{*})^2$$

where, $\hat{\boldsymbol{\beta}}_{-i}$ = jackknife PLSR estimators

$\hat{\boldsymbol{\beta}}$ = PLSR coefficients

$$\text{pseudo-value, } \mathbf{P}_{*i} = \hat{\boldsymbol{\beta}}_{-i} + I(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-i})$$

$$\bar{\mathbf{P}}_{*} = \frac{(\mathbf{P}_{*1} + \mathbf{P}_{*2} \ldots\ldots + \mathbf{P}_{*I})}{I}$$

8. From the values of PRESS and the bias and variance of $\hat{\boldsymbol{\beta}}$ calculate the value of σ

$$\text{MSEP}\,(\hat{\mathbf{y}}_{-i}) = \frac{\text{PRESS}}{I}$$

$$\text{hence, } \sigma^2 = \frac{\text{PRESS}}{I} - \text{variance} - \text{bias}$$

To account for the estimation of the dimension a double jackknife must be implemented; whereby a second jackknife is nested within the first.

6.2 PROCEDURES

Software

MATLAB™ software was used for the development of a program to perform the jackknife and related procedures. MATLAB (Matrix Laboratory) is a high performance interactive software package, designed for scientific and engineering calculations, which combines numerical analysis, matrix computation and signal processing in one environment [202]. Command

sequences are logical and usually require few statements, although the graphical output of MATLAB is somewhat limited.

In overview, the program operates as follows:

1. Initial file sorting.
2. Optional output of scores information.
3. Calculation of CV model and estimation of PRESS minimum.
4. Calculation of jackknife model, estimation of PRESS minimum and the model, predictions and prediction error at the minimum.
5. Calculation of the mean jackknife estimated model.
6. Output of the regression and dimensionality information.
7. Calculation of the bias, variance and bias corrected model.
8. Calculation of the MLR model.
9. Output of the regression information and data storage.

The double-jackknife is carried out according to the nested loops as follows:

```
for i=1:n                                    % outer loop
ind_i=[1:i-1,i+1:n];
      for k=1:n-1                            % inner loop
      if k<i, j=k,: else j=k+1; end;
      ind_ij=ind_i(:,[1:k-1,k+1:n-1]);
      [Xbar,ybar,B2]=pls(X(ind_ij,:), y(ind_ij), maxA);
      E2(k,:)=y(j)-(ybar+(X(j,:)-Xbar)*[zeros(B2(:,1))B2]);
      end
.....
end
```

An object is removed by the outer loop and the remaining objects are successively left out according to the inner loop. Within the inner loop PLSR

models are calculated according to the MATLAB function file "pls.m". The absolute error of prediction for the object left out is calculated and presented as the PRESS by dimension for that object. By choosing the optimum dimensionality at the first local minimum of PRESS, the optimum model can be determined according to I-1 objects. This process is repeated for each of the I objects.

A program was developed to make predictions of new and independent objects. This routine produces a hard copy of the predictions with their confidence interval and compares the predicted values to those which were obtained by the reference method. This was specifically incorporated to provide a means of comparing the jackknife model predictions with those produced by the commercially available Unscrambler™ PLSR software. The Unscrambler package produces a confidence interval that has no theoretical foundation, rather it is "an empirically found relationship that has given satisfactory indications on the uncertainty in predictions for a large range of applications". A similar program was written for comparison with MLR predictions.

A number of data sets from the literature were used to assess the potential of the double-jackknife:

Wold [203]

Data set consisting of the observed ß-receptor agonist activity for 15 structurally similar phenethylamines, and 8 independent variables relating to the morphological and physico-chemical properties of these compounds. All 8 objects were used in calibration and predictions were made using the same data.

Naes [204]

Fat concentration (%) of 45 fish samples (rainbow trout) and independent variables of the absorbance at 9 wavelengths measured after sample homogenisation. Calibration was undertaken on the whole data set (45x9) and

after arbitrarily splitting the data set; #1-24 for calibration and #25-45 as an independent test set.

Fearn [205]

Data set consisting of the measured protein content (%) (Kjeldahl) of 50 ground wheat samples and the log reciprocal reflectance at 6 NIR wavelengths. Calibration was again carried out using the entire data set and after splitting the data #1-24 for calibration and #25-50 as an independent test set, as used in the original publication.

6.3 RESULTS & DISCUSSION

Wold

Figure 6.1 reveals that cross validation of the complete data set produces a typical relationship between the PRESS and increasing dimensionality. After rapidly dropping to a local minimum (also the global minimum) at a dimensionality of 2, the value of PRESS gently increases with each extra factor. A 2-factor model would therefore be selected as the optimum model for prediction on the basis of cross validation. Similarly, examination of the score plots, such as the scores of factor 1 versus factor 2 shown in Figure 6.2, revealed no strong grouping of the objects. However, if the PRESS versus dimensionality plot is studied for the double-jackknife model (Figure 6.3), it can be seen that one object is having a particularly strong influence; object number 13. When #13 is left out of the inner jackknife, instead of the PRESS reaching a minimum after 2 factors and beginning to rise once again, it remains very low with the inclusion of each factor. This suggests that by including #13 in the calibration set an increasing amount of noise is being modelled with each added factor. Inspection of the regression coefficients (Figure 6.4) reveals dramatically different responses for each variable at #13. The score plots, however, do not indicate #13 as an outlier.

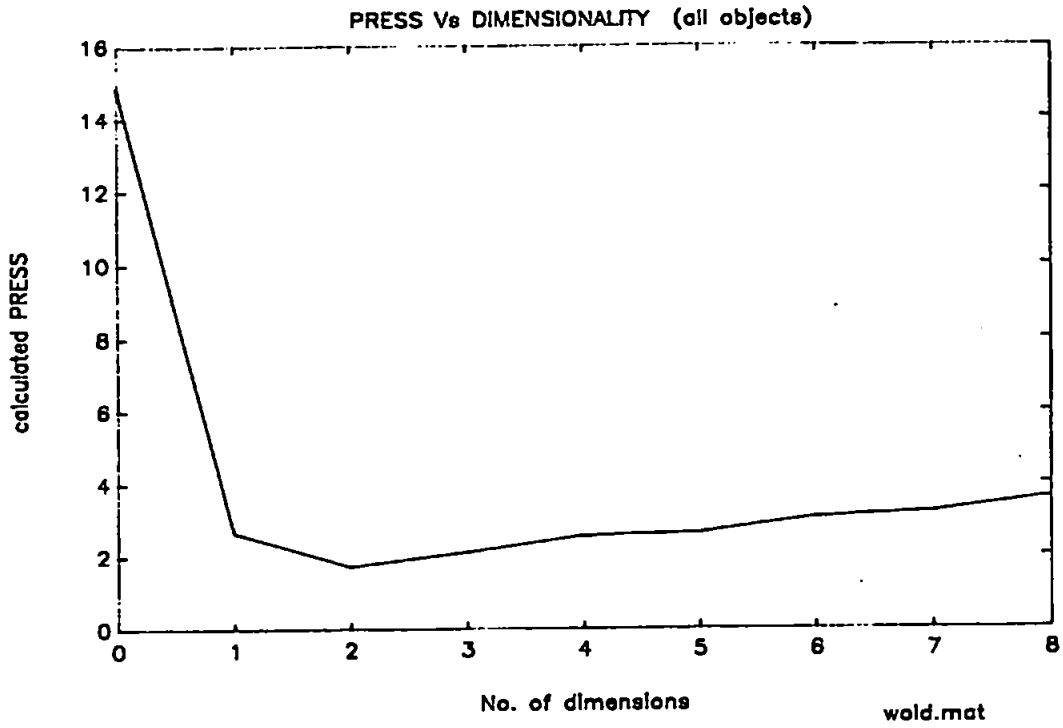


Figure 6.1 PRESS vs dimensionality curve for cross validation

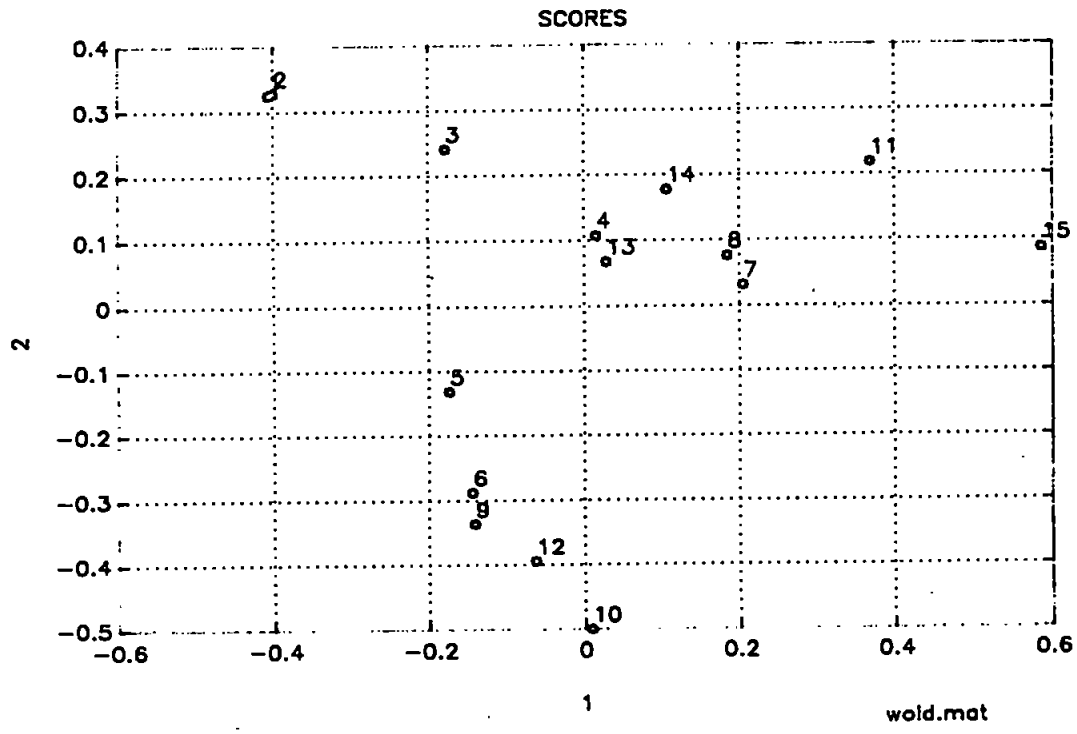


Figure 6.2 PLSR factor scores plot

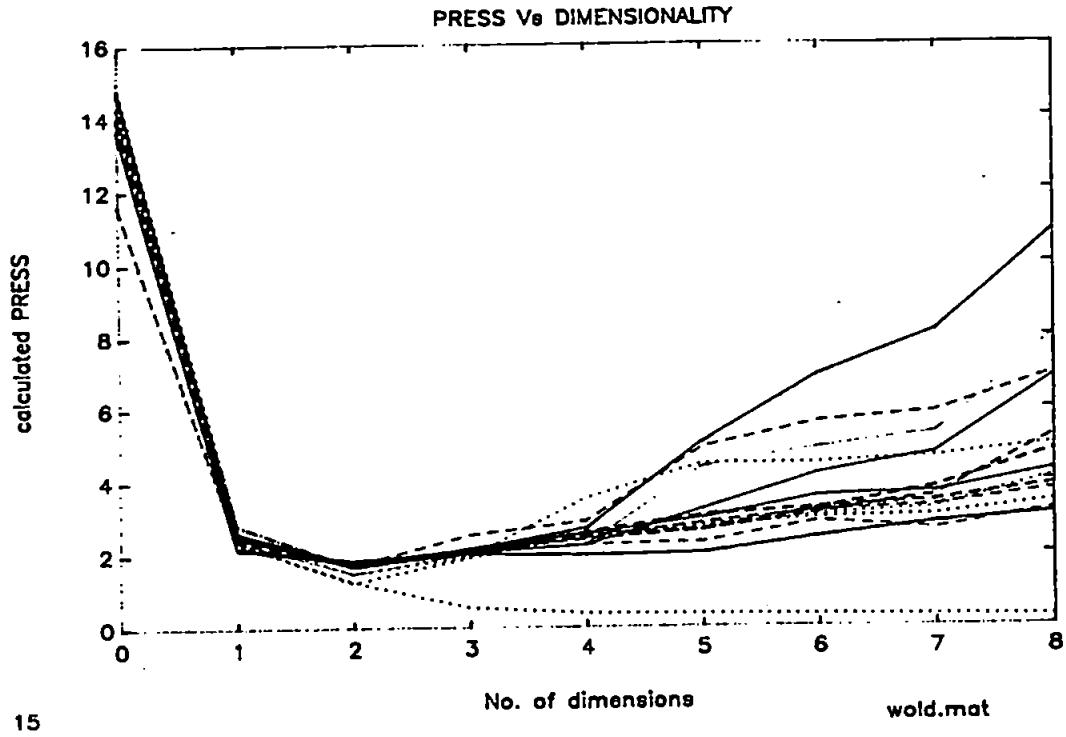


Figure 6.3 PRESS vs dimensionality curves for jackknife
Each curve represents the object left out of the outer jackknife.
The identifiers have been removed for clarity

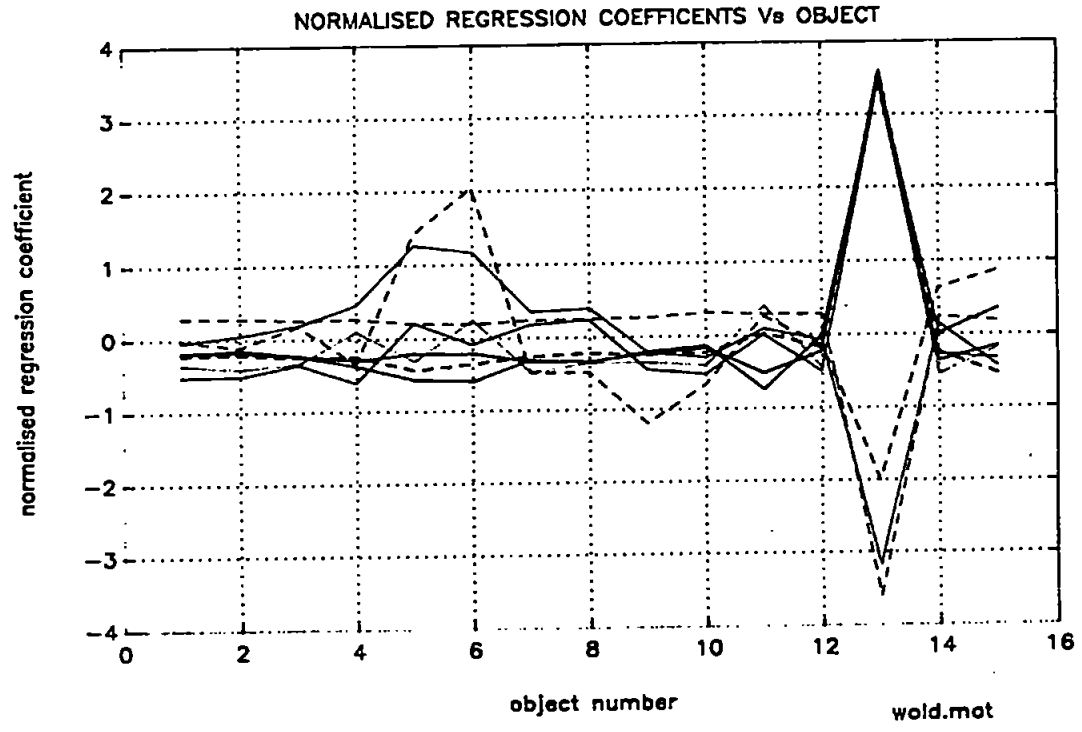


Figure 6.4 Regression coefficients for jackknife model

Predictions of the calibration objects were made using the jackknife model and the Unscrambler package. The UNSC model was fully cross validated and has an optimum dimensionality of 2, as expected. The results from the UNSC model, shown in Figure 6.5 reveal a tight confidence interval around the predicted values, however in 6 out of 15 cases (40%) the actual value lies outside this interval. Predictions from the jackknife model (Figure 6.6) yield a much less optimistic confidence interval and accordingly only one of the actual values lies outside the confidence interval. Conversely, the UNSC model yields predicted values closer to the actual values than the jackknife model; with calculated PRESS values of 1.0 and 2.9 respectively.

During the model building stage the jackknife has selected an optimum dimensionality of 8 when #13 was left out. Inspection of the PRESS curve indicates that 3 factors would have been a more realistic choice. The incorporation of this element into the final bias corrected model could conceivably lead to inaccurate predictions. For this reason and the indications from the regression coefficients, #13 was removed from the data set and the model recalculated (wold-1).

The PRESS versus dimensionality plot for the CV model, shown in Figure 6.7, is very similar to that for the complete data set model. However, the jackknife model dimensionality estimates again reveal a different data structure (Figure 6.8). Here it can be seen that by leaving out either #14 or #15 has a similar effect to that seen when #13 was left out of the complete dataset model, (i.e. after 2/3 dimensions the PRESS value remains very low with each additional dimension). With #14 removed the first local minimum is reached after 3 factors, and after 8 factors with #15 left out. The regression coefficients also reveal markedly different responses for these two objects.

Predictions from the UNSC model (Figure 6.9) reveal a very tight confidence interval with the actual values of three objects (21%) lying outside. The

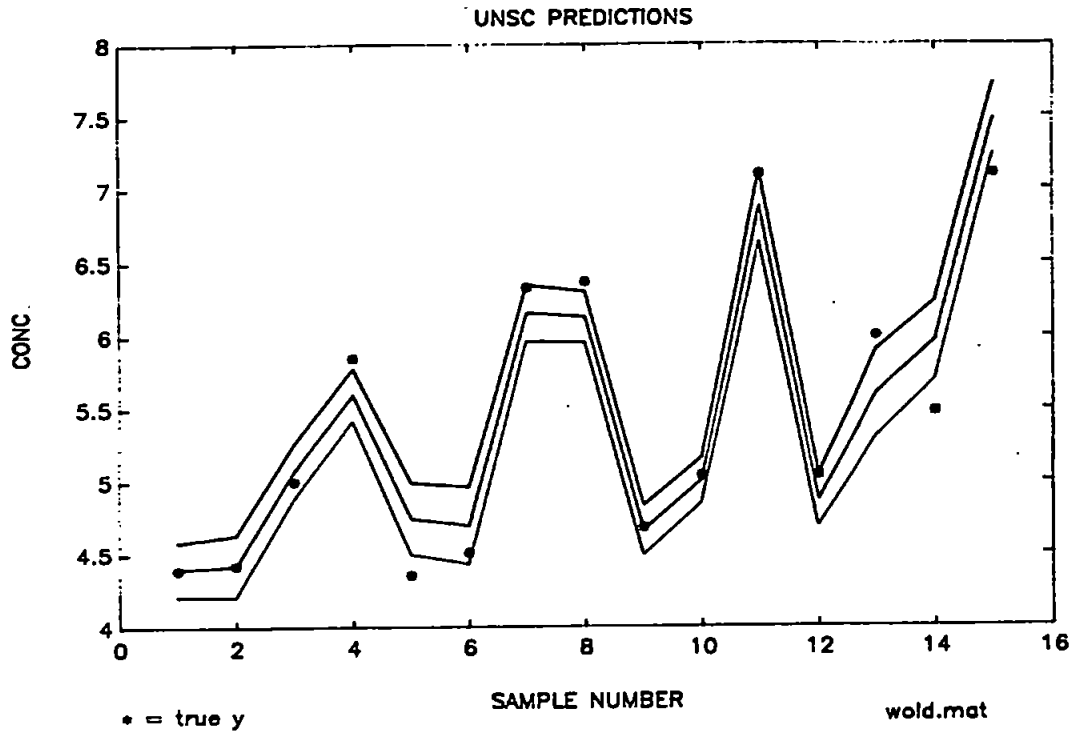


Figure 6.5 Predictions and confidence interval from the Unscrambler package, actual y_i again shown as *

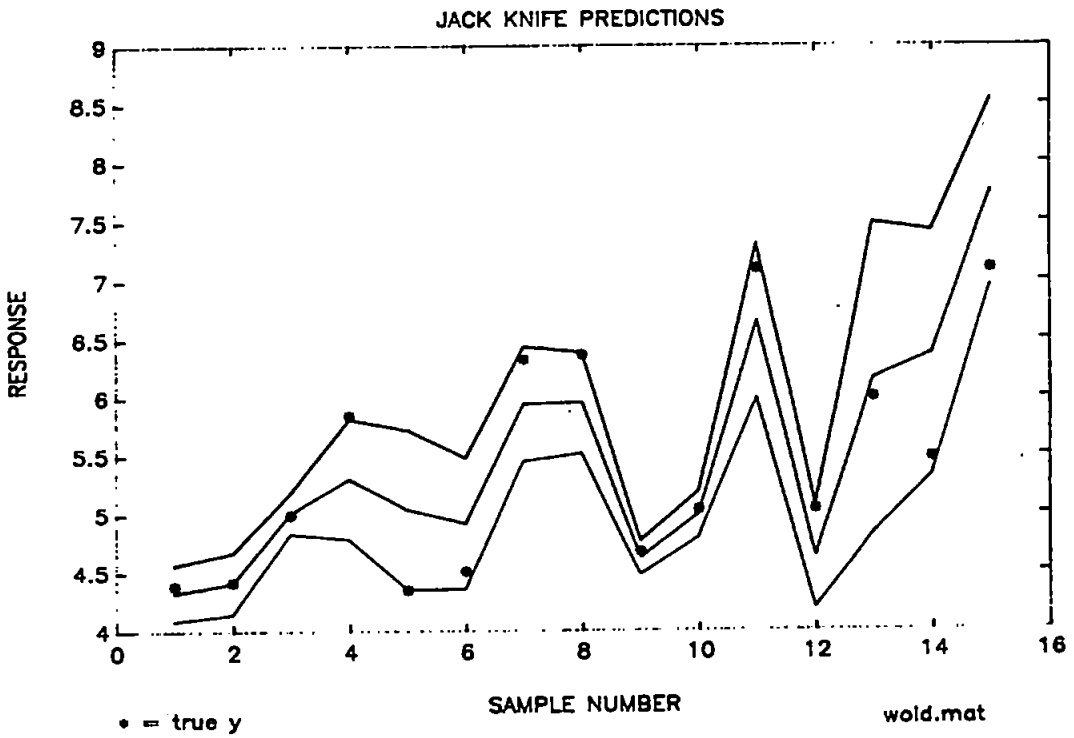


Figure 6.6 Jackknife predictions and confidence interval with the actual values of y_i shown as *.

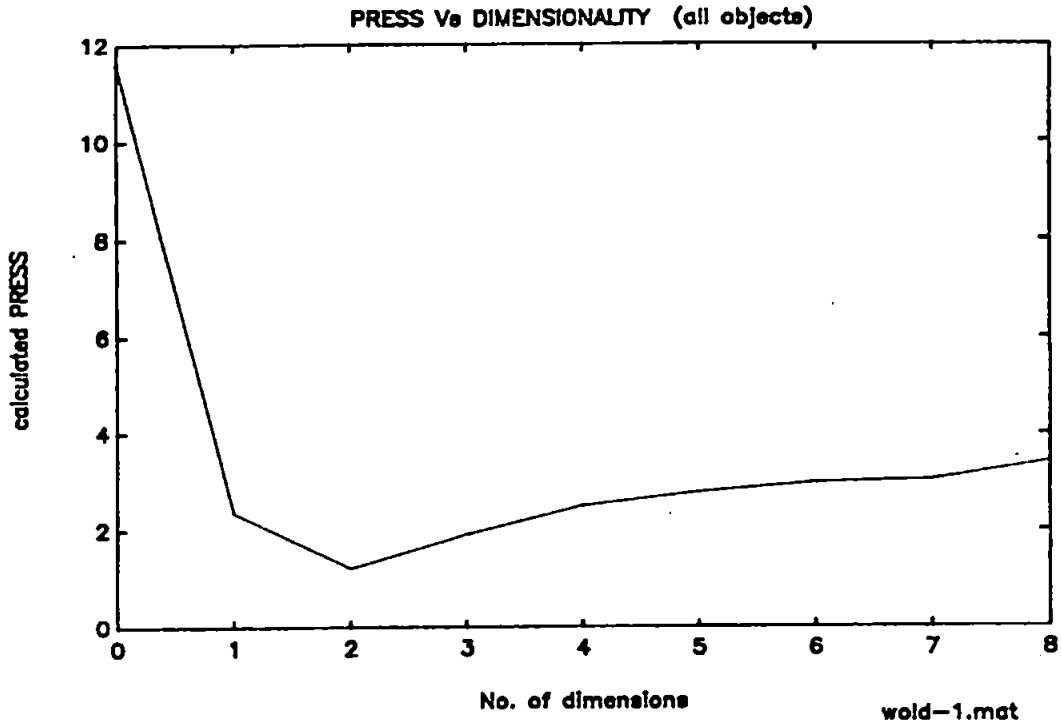


Figure 6.7 Press vs dimensionality curve for cross validation

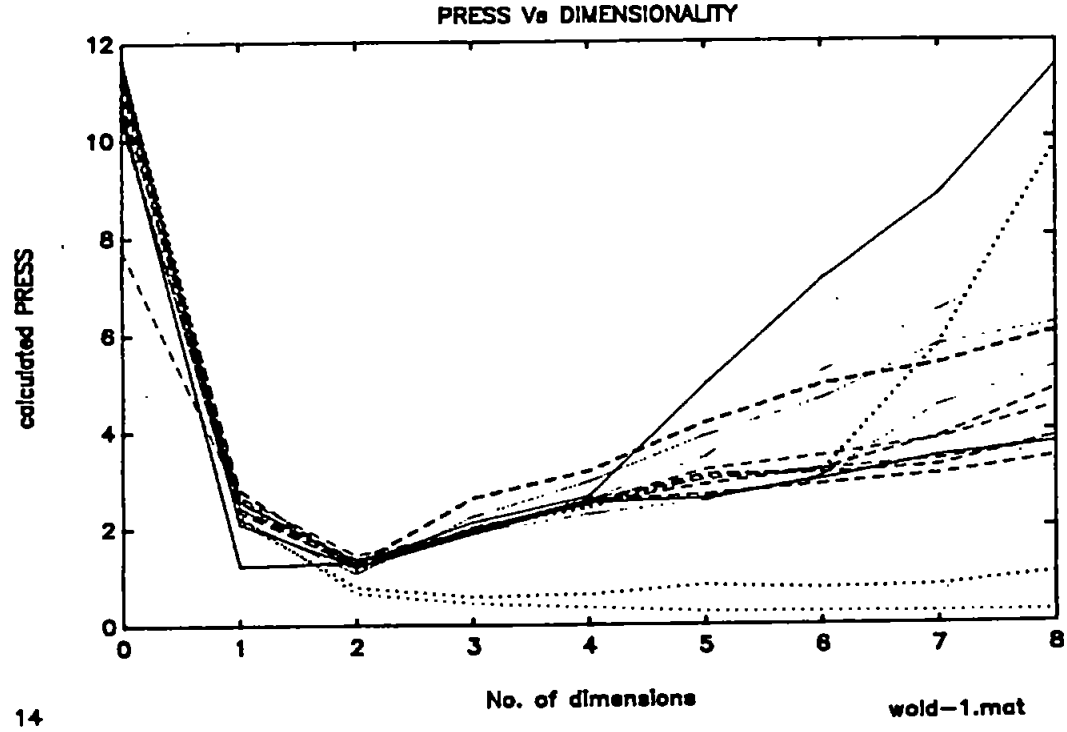


Figure 6.8 Press vs dimensionality curve for jackknife model

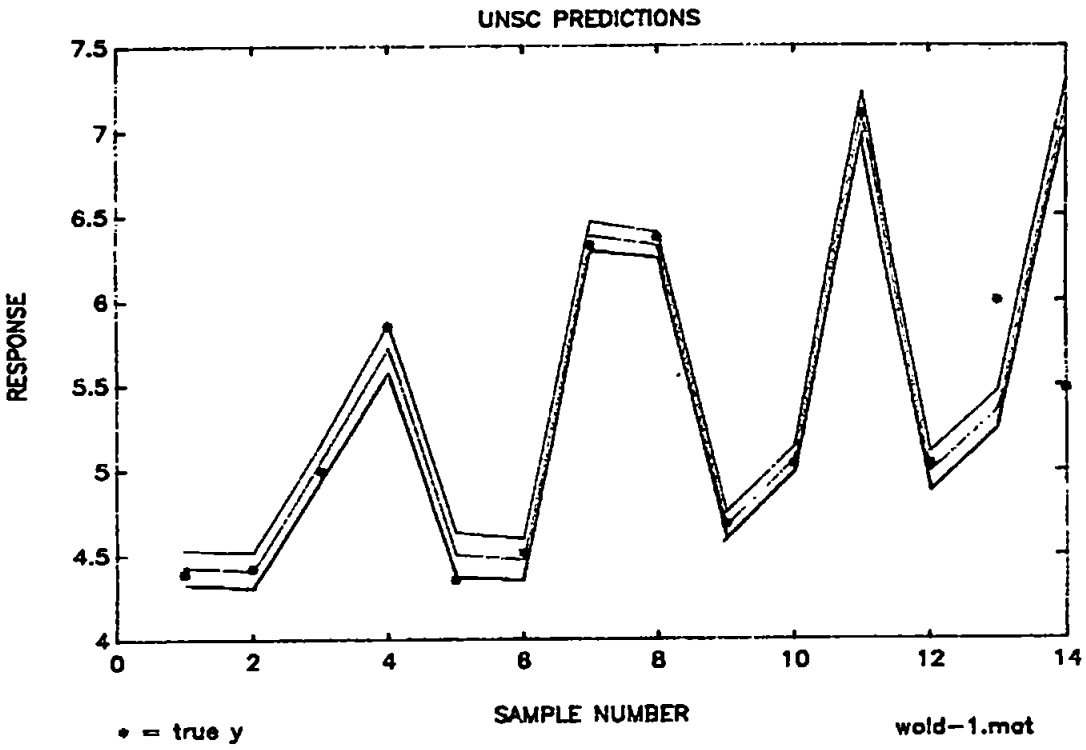


Figure 6.9 Predictions and confidence interval from Unscrambler

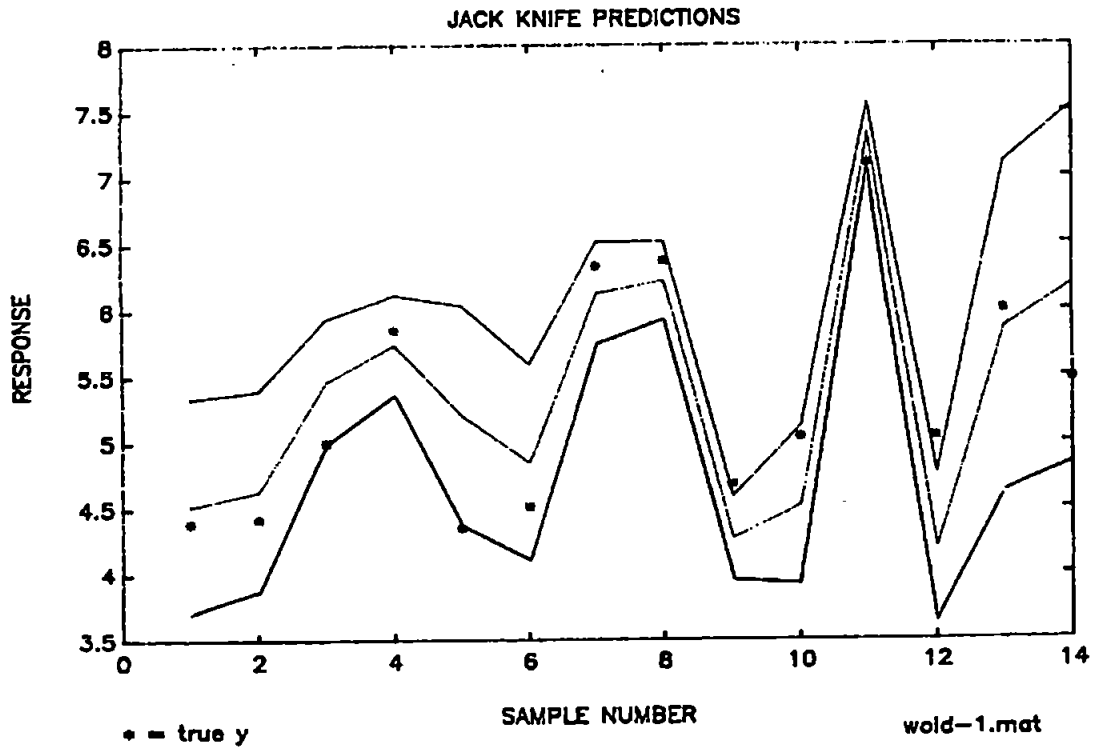


Figure 6.10 Jackknife predictions and confidence interval

jackknife predictions (Figure 6.10) again have a much less optimistic confidence interval, but with the actual values of 3 objects falling outside the interval, as with the UNSC model. The PRESS values are also very similar, 3.3 and 2.9 for the UNSC and jackknife predictions respectively.

#14 & #15 have been removed and the models recalculated with 12 objects (wold-3). Figure 6.11 reveals that the CV model has a first local minimum in PRESS at a dimensionality of 6, although there appears to be no significant change after 3 dimensions. This pattern is followed by the jackknife model (Figure 6.12) with the first local minimum in PRESS being found after 3,4,5 or 6 dimensions. In all cases it is unlikely that any useful information is being modelled after 3 factors are incorporated. The predicted values from the UNSC (Figure 6.13) model are very good (PRESS<0.1) although the actual values for two of the predictions still lie well outside the confidence interval. For the jackknife predictions (Figure 6.14) the actual value lies outside the confidence interval in 5 cases (42%) and the PRESS is 2.0.

For each of the models described above, the MLR predictions are very close to the actual values and mostly lie within the confidence interval. This is to be expected with MLR when predictions are being made using the same data as that used in calibration; MLR is fitting the data without accounting for the variance in the independent variables.

For this particular data set, jackknife estimation has identified a number of objects that appear to be outlying which CV has failed to recognise. In terms of the predictive ability of the jackknife it is difficult to assess in this case because predictions were only carried out on the data used in the calibration stage. The initial model did, nevertheless, illustrate how the jackknife estimations of bias and variance led to predictions boasting a more realistic confidence interval, which encompassed all but one of the actual values.

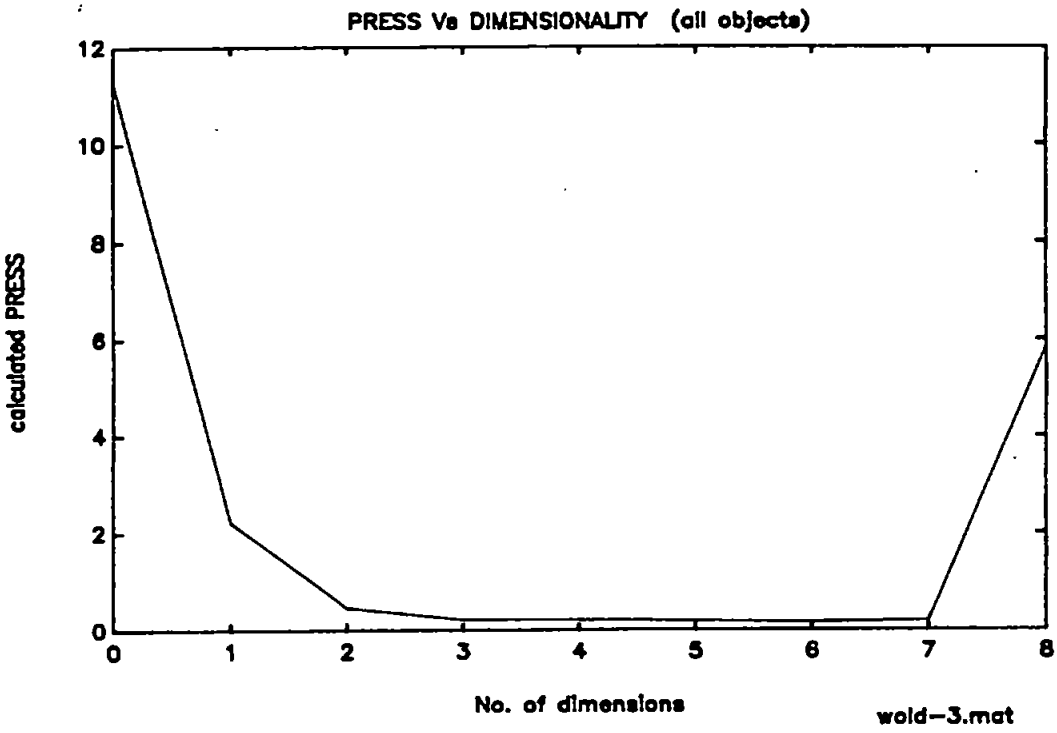


Figure 6.11 PRESS vs dimensionality curve for cross validation

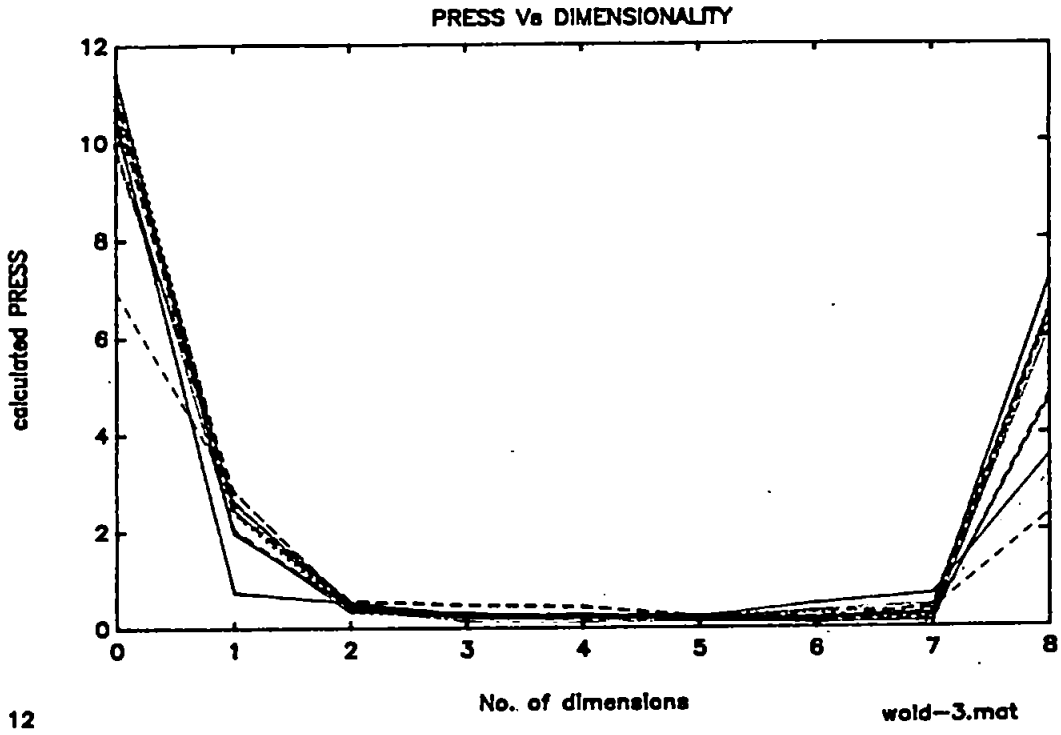


Figure 6.12 PRESS vs dimensionality curve for jackknife model

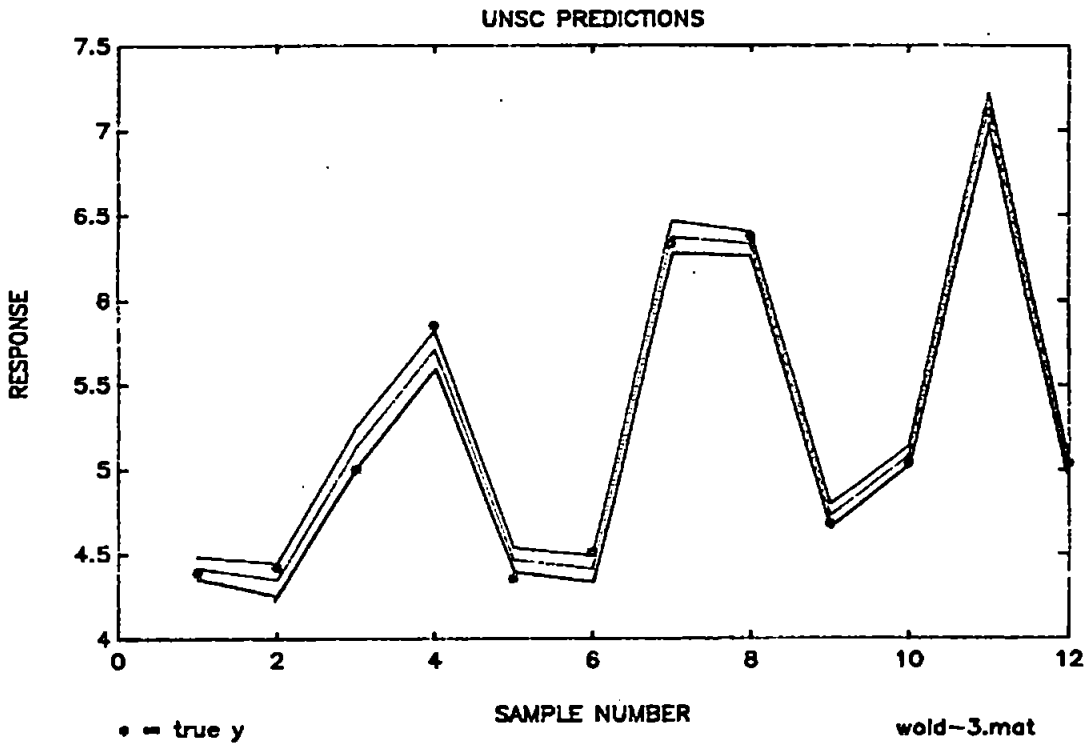


Figure 6.13 Unscrambler predictions and confidence interval

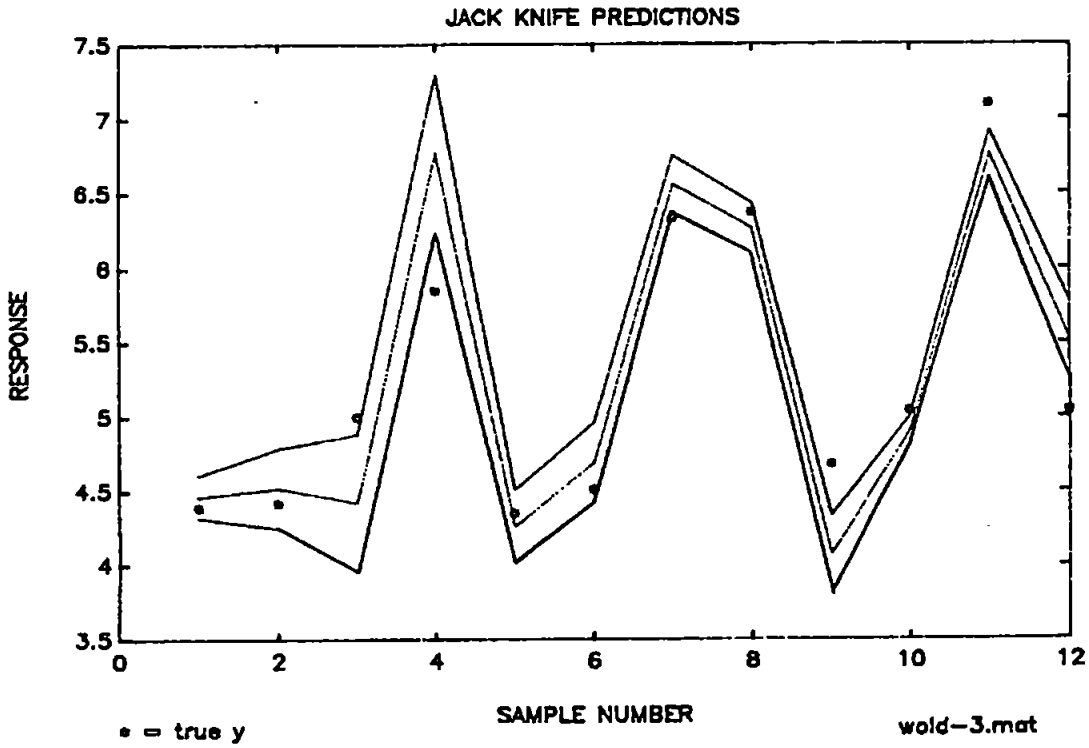


Figure 6.14 Jackknife predictions and confidence interval

Naes

Jackknife modelling of the whole data set (Figure 6.15) revealed an optimum dimensionality of 4 in all but three cases; CV likewise selected an optimum of 4. It does, however, seem likely that 3 factors would have been sufficient. #43-45 appear to be outlying from the scores and the regression coefficients as well as when left out during the jackknife procedure. This has been recognised by the large confidence interval attached to these objects by the jackknife model as can be seen in Figure 6.16. The UNSC model (Figure 6.17) in this case yields 11 actual values lying outside the attached confidence interval, whereas the jackknife model yields only 4. The calculated values of PRESS for the predictions are 68 and 147 for the UNSC and jackknife models respectively.

After splitting the data set and remodelling, the optimum dimensionality has been reduced to 3 for the CV model and in all but one case for the Jackknife model (Figure 6.18). The prediction results for the independent objects are very similar; a PRESS of 89 and 109 for the UNSC and jackknife models (Figures 6.19 & 6.20). The actual values lying outside the calculated confidence interval were shown to be 4 and 3. The estimated size of the confidence interval is very similar for both models, showing a marked widening for the outlying objects.

The jackknife model has again recognised potentially outlying objects and attached a realistic confidence interval to the predictions for the whole data set. With the split set there is little difference in the predictions made by the two models. The selected dimensionality is uniform for the jackknife model (with one exception) and equal to the that estimated by CV. The resulting models are therefore very similar, with subtle differences due to a combination of bias correction and influence of the single 5-factor optimum.

Fearn

The PRESS versus dimensionality plot for the CV model revealed a first local minimum at a dimensionality of 0. This was ignored by the incorporation

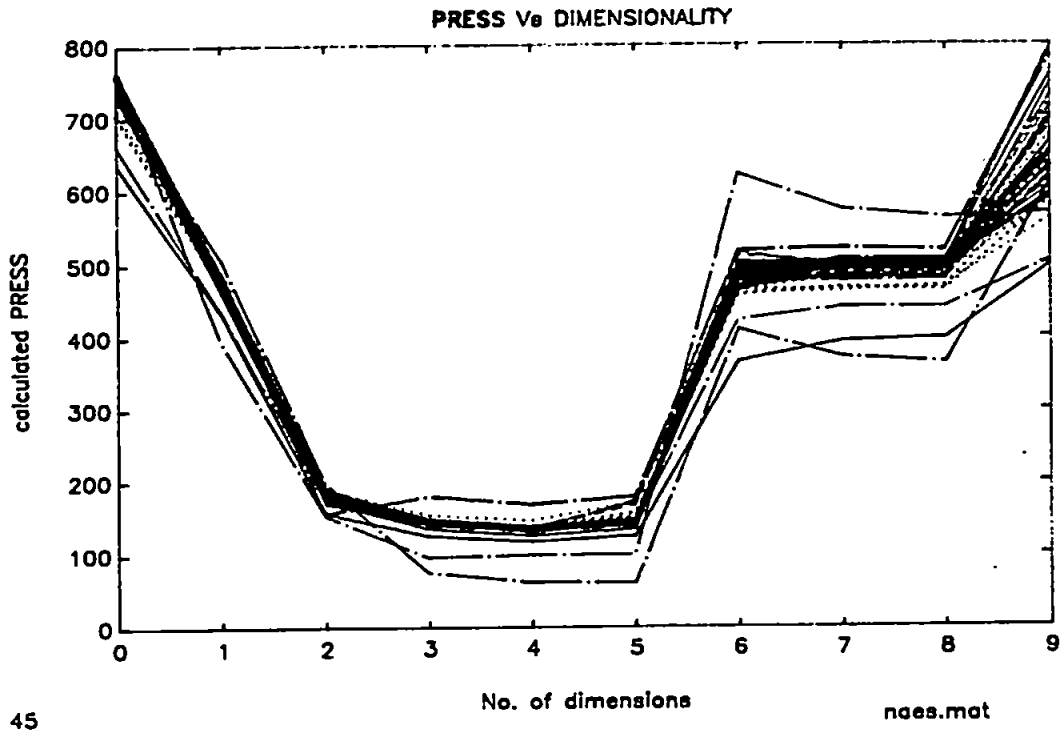


Figure 6.15 PRESS vs dimensionality curve for jackknife model

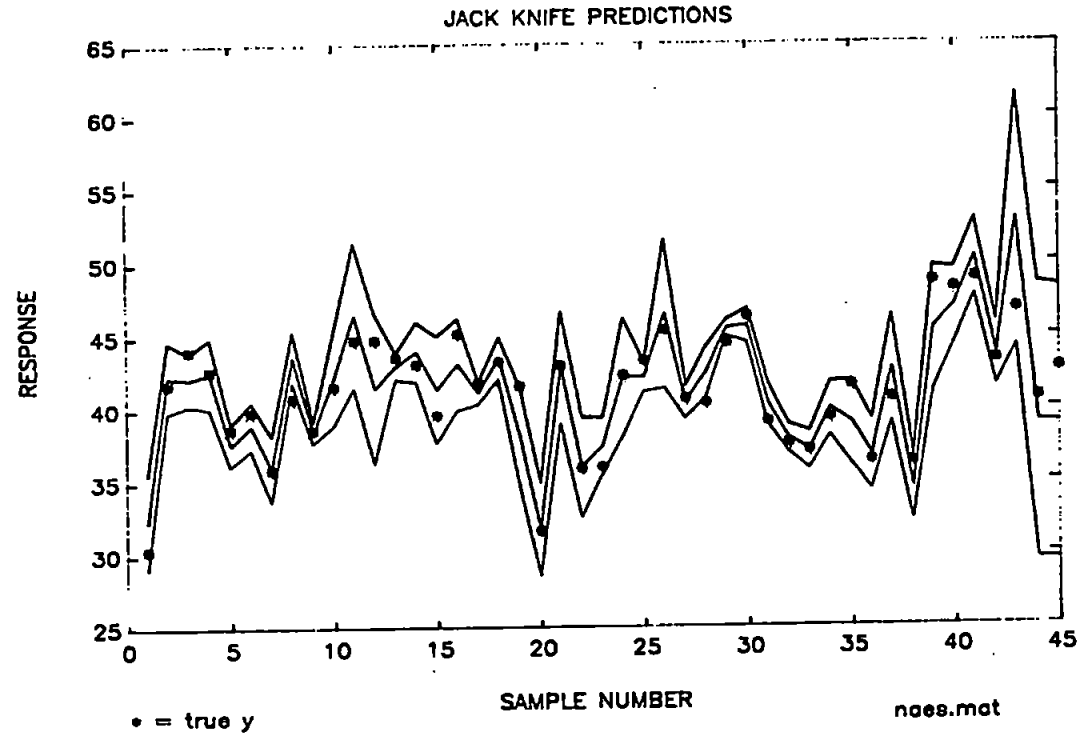


Figure 6.16 Jackknife predictions and confidence interval

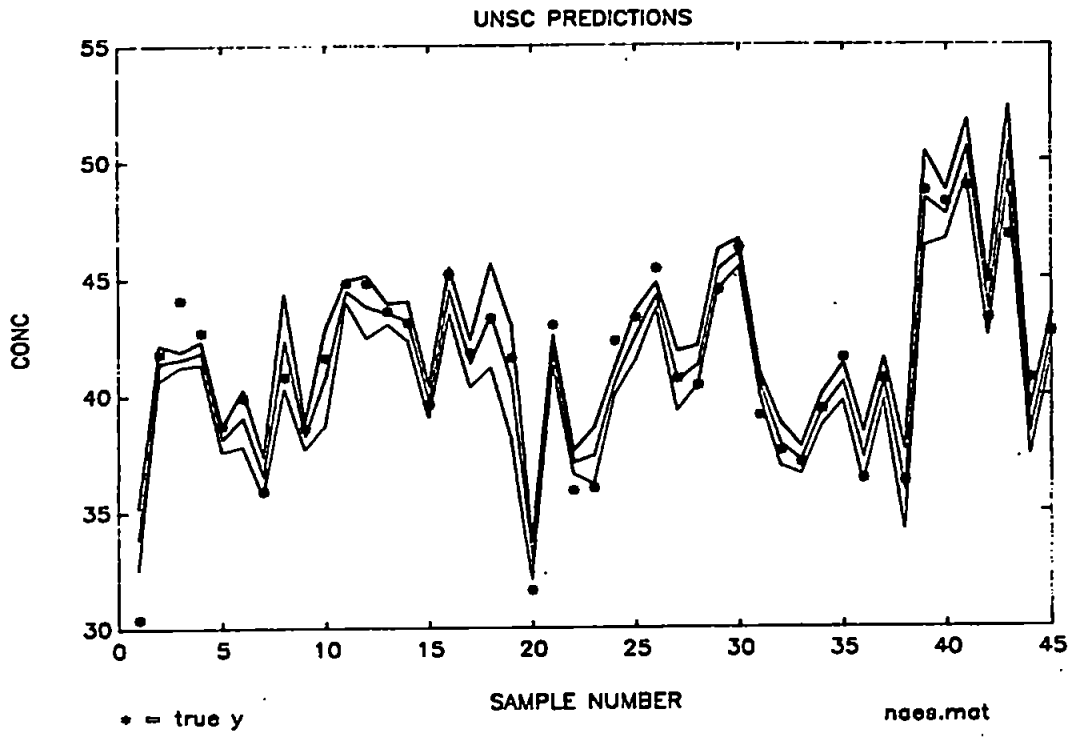


Figure 6.17 Unscrambler predictions and confidence interval

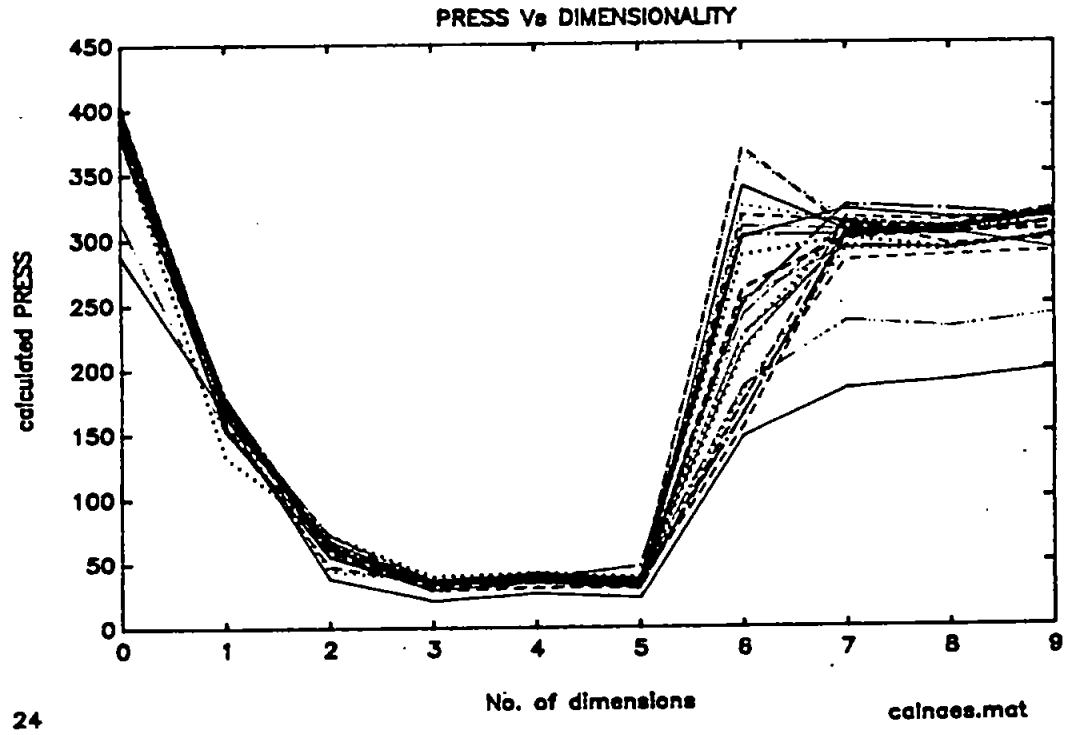


Figure 6.18 PRESS vs dimensionality for jackknife model

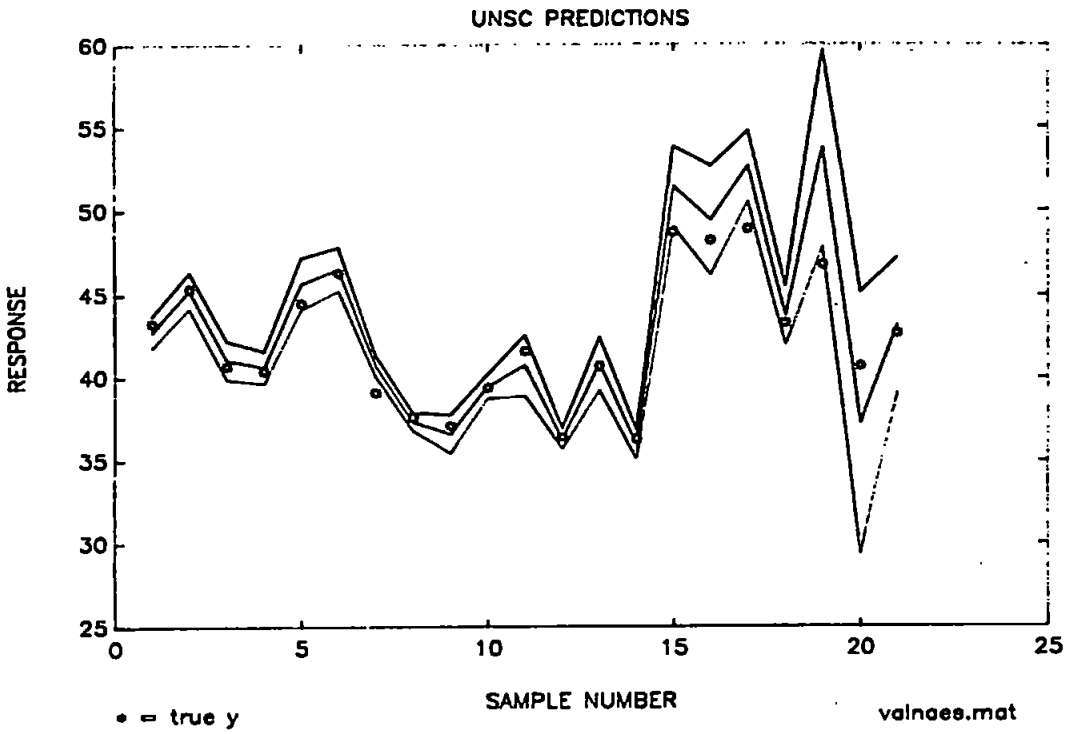


Figure 6.19 Unscrambler predictions and confidence interval

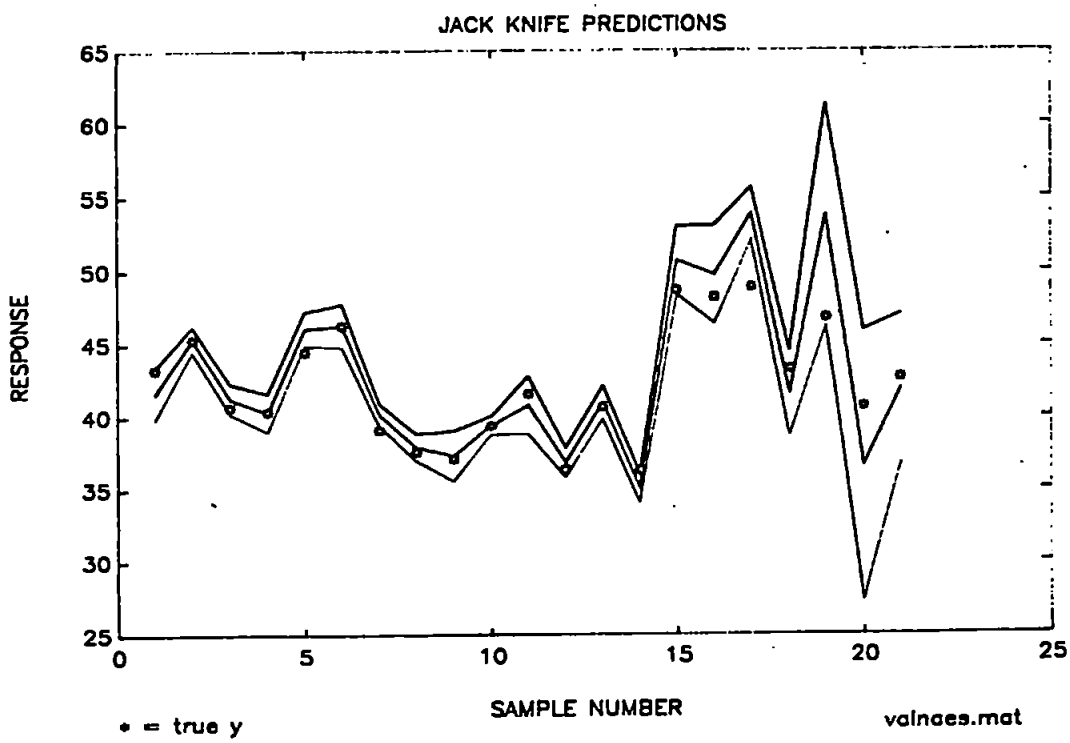


Figure 6.20 Jackknife predictions and confidence interval

of a loop within the program, and hence a 3-factor model was selected by CV. The jackknife model, shown in Figure 6.21, also yields optima at a dimensionality of 3 in 80% of cases with the remainder being split between 5 and 6 factors. No outliers are obvious from the PRESS curves or the regression coefficients, although #25 appears to be outlying from the scores plots. Predictions on the whole data set produced a PRESS of 2.9 for the UNSC model (Figure 6.22) with 18 actual values lying outside the confidence interval. The jackknife model (Figure 6.23) predictions proved disappointing with a PRESS of 140 and 35 actual values lying outside the estimated confidence interval.

After splitting the data set, both models select optima at a dimensionality of 4, with one jackknife the exception as shown in Figure 6.24. In all cases there appears to be little difference between the PRESS at dimensionalities of 3,4 and 5; with 3 factors probably the most appropriate. In both cases the predictors are good, with PRESS values of 2.9 and 6.0 and the number of actual values lying outside the confidence interval, 9 and 7 for the UNSC and jackknife models respectively (Figures 6.25 & 6.26). As noted for the Naes data set, the differences in the models are presumably related to the bias correction.

The poor prediction performance of the jackknife model for the whole data set can be attributed to the overestimation of dimensionality recorded for 20% of the jackknife estimators. The jackknife model has, nevertheless, provided reliable predictions from an independent test set, although there were a considerable number of actual values lying outside the estimated confidence interval. When making this consideration it should be noted that the assumption that the independent data fit the calibration model may not necessarily hold.

For the Naes and Fearn data sets the MLR predictions appear to be as good as, if not better than those predicted by the PLSR models, for the whole and the sub-divided data sets.

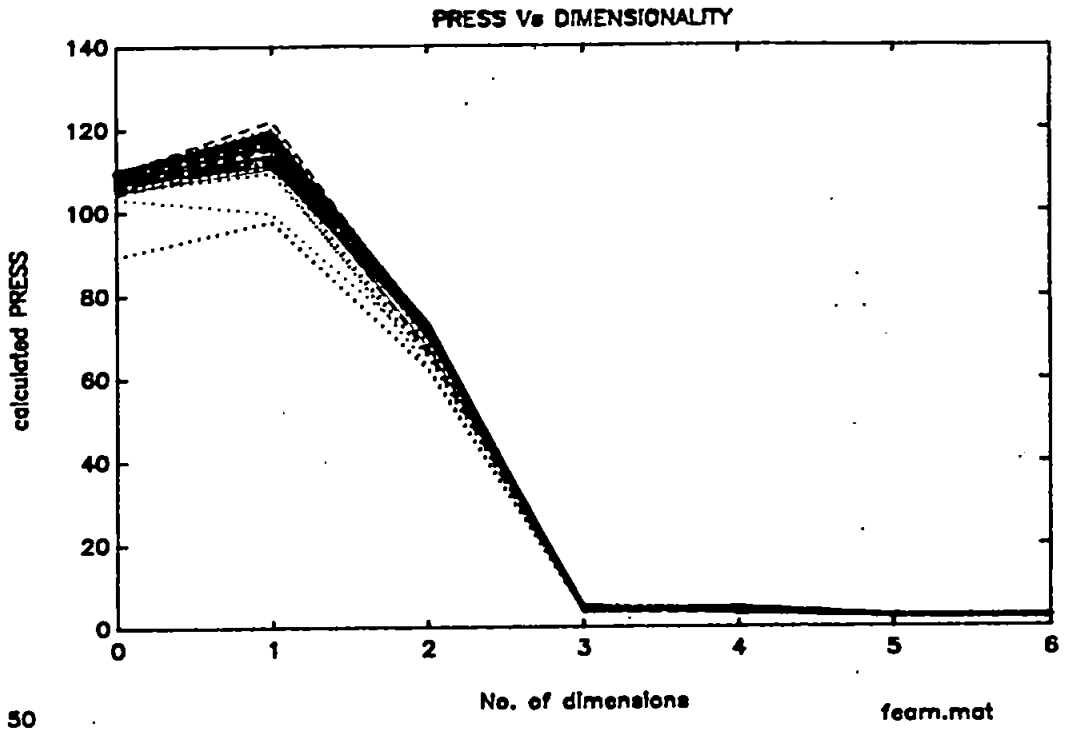


Figure 6.21 PRESS vs dimensionality curve for the jackknife model

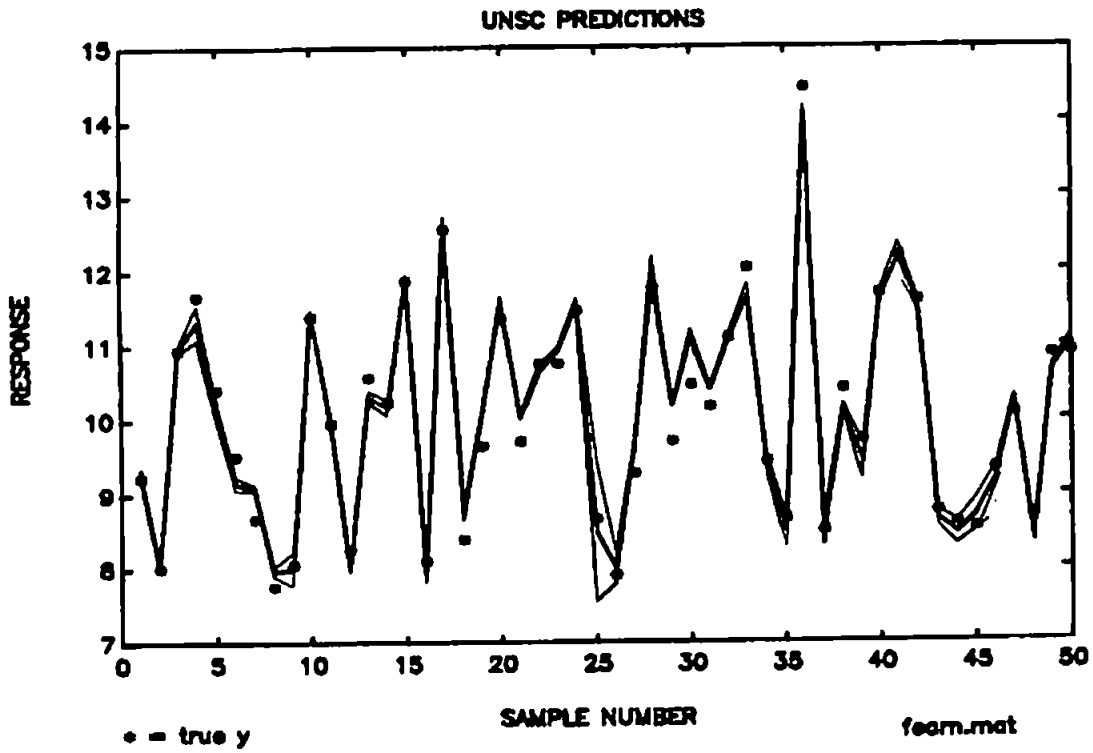


Figure 6.22 Unscrambler predictions and confidence interval

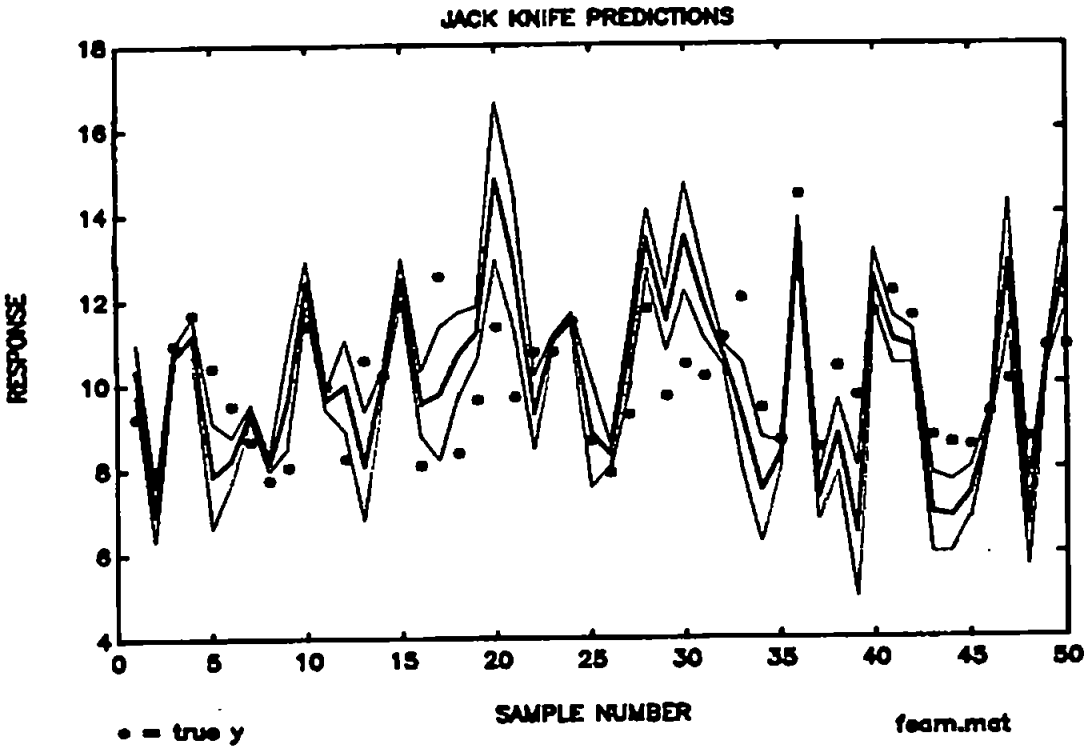


Figure 6.23 Jackknife predictions and confidence interval

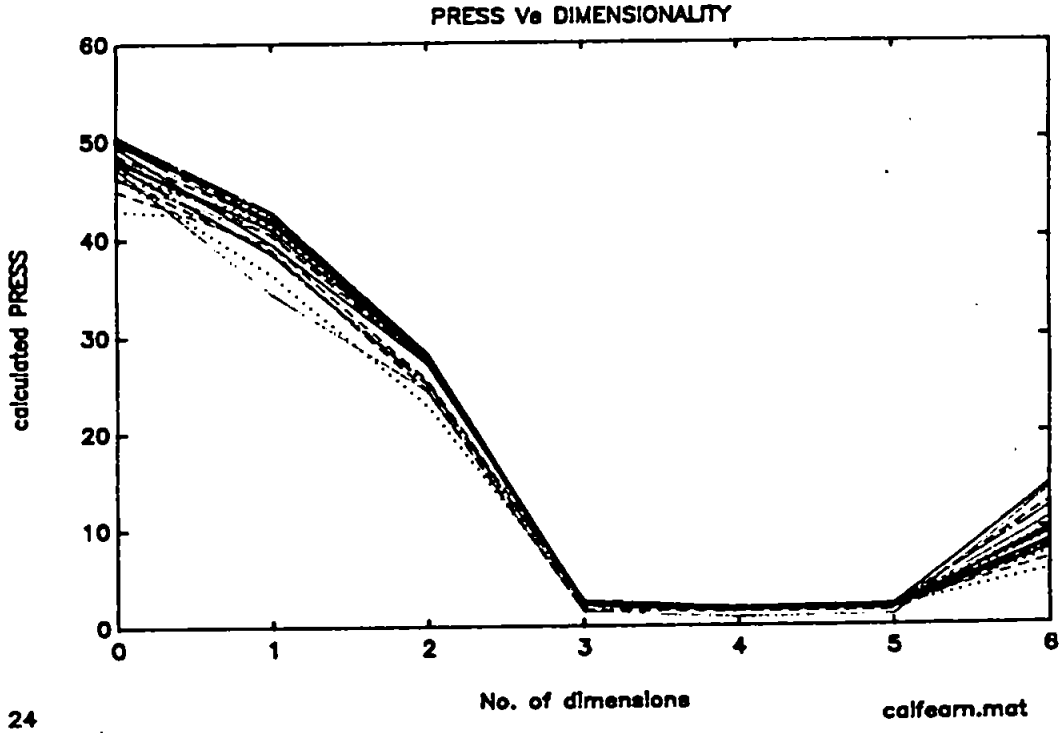


Figure 6.24 PRESS vs dimensionality curve for the jackknife model

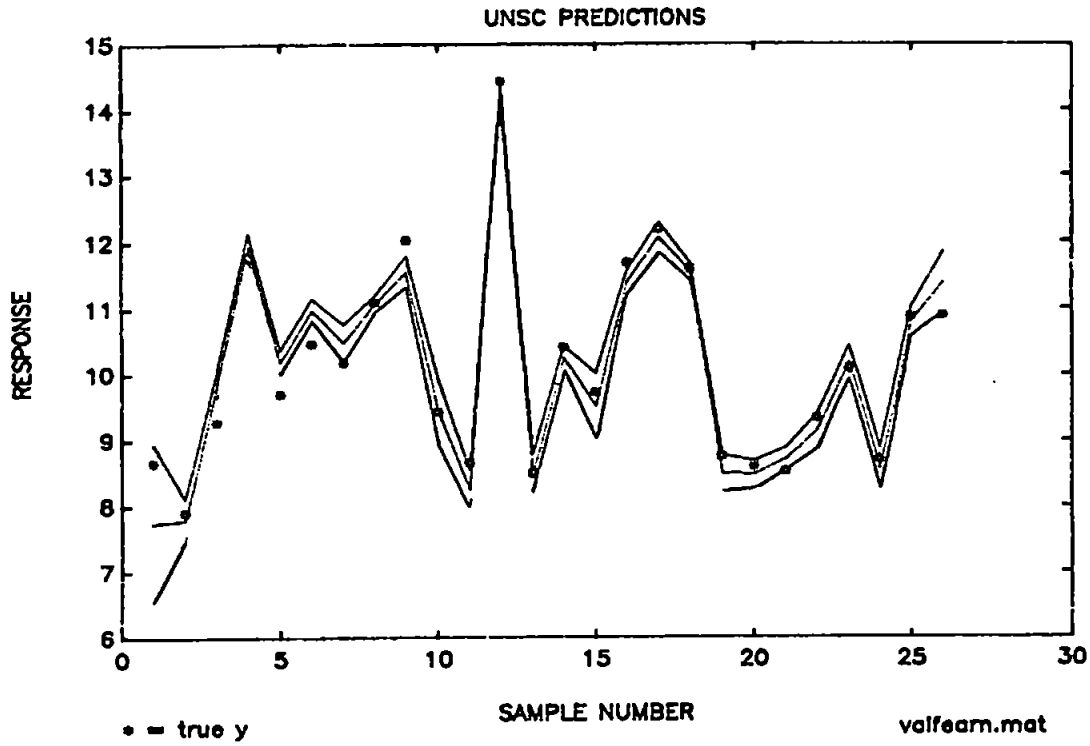


Figure 6.25 Unscrambler predictions and confidence interval

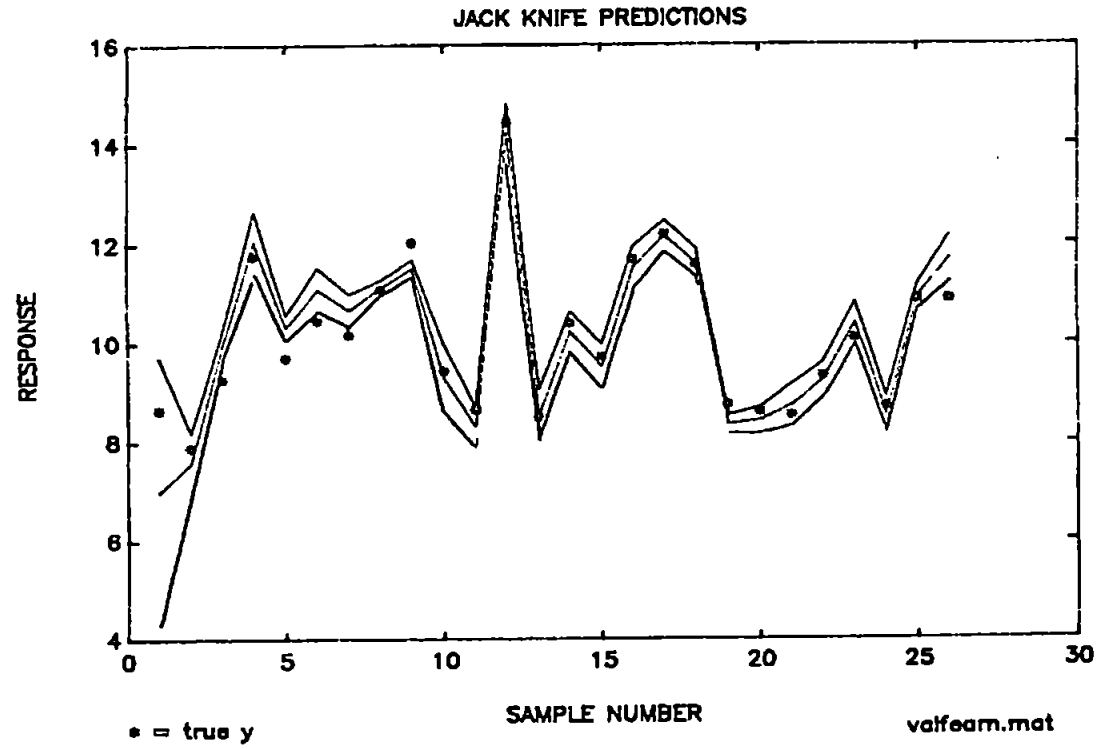


Figure 6.26 Jackknife predictions and confidence interval

6.4 CONCLUSIONS

1. The jackknife has been shown to reveal outlying objects which have not been detected by inspection of scores plots.
2. Implementation of the double-jackknife enables a different approach to dimensionality estimation.
3. Computation of the jackknife coefficients allows the model bias to be estimated and hence bias correction to be undertaken.
4. Computation of model error from a theoretically sound basis allows the allocation of realistic confidence intervals to future predictions.

However, the ultimate goal of any calibration is effective prediction of future samples and for a number of data sets the jackknife procedure has produced models with poor predictive ability. This has been due to overestimation of the optimum dimensionality to be used for prediction. A method of dimensionality selection based on the significance of improved prediction with added factors may prove more successful than the first local minimum approach used in this work.

Chapter Seven

Conclusions & future work

7.1 FINAL CONCLUSIONS

From the preceding chapters, the following general conclusions can be drawn:

1. Automated flow injection analysis is suitable for the on-line single analyte monitoring of chemical processes.

   This is demonstrated by the installation of a monitor for the on-line determination of sulphite in 20 % m/v potassium chloride brine. The accuracy and precision of the system are ± 3 % and ± 1 % respectively, with a response time of <5 min and a dynamic range of 0.1-100 mg l^{-1}. The procedure is valid relative to a standard iodimetric method and the monitor is reliable over 21 days on-line analysis.

2. Multivariate calibration enables multianalyte resolution of UV-visible spectrophotometric data.

   A model data set consisting of mixtures of transition metal sulphate solutions demonstrates the application of direct multicomponent analysis, principal components regression and partial least squares regression. Spectra, collected using a photo-diode array detector, can be resolved using commercial software.

3. Partial least squares regression offers good calibration performance over a wide range of physical and chemical conditions.

   Comparison of the relative prediction abilities of the three multivariate calibration techniques under varying degrees of calibration complexity and interferences reveals that:
   Direct multicomponent analysis is reliable in the absence of gross interferences but yields large errors when chemical interactions or

physical interferences are incorporated.

Both principal components regression and partial least squares regression are capable of accurate predictions under such circumstances.

4. The combination of automated flow injection analysis, photo-diode array detection and partial least squares regression offers a physically simple means of simultaneous multianalyte determinations suitable for the process environment.

   The simultaneous determination of phosphate and chlorine is possible utilising a two-line flow injection manifold, single injector and single detector. Partial least squares modelling can reveal the non-linear effect of combining the established spectrophotometric reactions in a single procedure. RRMSEP values of a new and independent test set of 20 samples prepared and analysed 48 hours after calibration are 4.0 % for phosphate and 2.4 % for chlorine.

5. The jackknife offers a means of dimensionality estimation, bias correction and outlier detection in partial least squares modelling.

   The jackknife estimates of dimensionality curves help reveal potential outlying objects and the jackknife estimation of model error allows realistic confidence intervals to be attached to future predictions. A more robust form of dimensionality selection is required to improve predictive ability.

7.2 SUGGESTIONS FOR FUTURE WORK

Significant developments could be made in three distinct areas, within which short and long term aims can be defined:

Flow injection analysis

Short term Further examination of the suitability of flow injection analysis for process monitoring by the development, implementation, validation and extended on-line use of single analyte monitoring systems. Extension of the flow injection analysis, photo-diode array, multivariate calibration approach to a three (or more) component multianalyte system.

Long term Development and on-plant implementation of a cam-driven piston pump system for the handling of aggressive materials.

Detection

Short term Investigation of vibrational spectroscopic techniques for single and multianalyte process flow injection analysis.

Long term Development of a low cost, process worthy photo-diode array spectrophotometer and its use for on-line multideterminations.

Chemometrics

Short term Implementation of a significance based dimensionality selector for jackknife estimation of partial least squares models.

Long term Examination of the applicability of quantitative multivariate curve resolution techniques for spectrophotometric flow injection data.

References

1. Illman, D.L., *Trends Anal. Chem.*, 1986, **5**, 164.

2. Callis, J.B., Illman, D.L. and Kowalski, B.R., *Anal. Chem.*, 1987, **59**, 624A.

3. Riebe, M.T. and Eustace, D.J., *Anal. Chem.*, 1987, **62**, 65A.

4. Proceedings of Anatech 1986, *Anal. Chim Acta*, 1986, **190**, 1-288.

5. Proceedings of Anatech 1990, *Anal. Chim. Acta*, 1990, **238**, 1-262.

6. *Process Control and Quality.*

7. Jacobs, S.M. and Mehta, S.M., *Amer. Lab.*, 1987, Dec., 15.

8. Valcarcel, M. and Luque de Castro M.D., *Automatic Methods of Analysis*, Elsevier, Amsterdam, 1988.

9. Blaser, W.W., Ruhl, H.D. and Bredeweg, *Amer. Lab.*, 1989, Jan., 69.

10. Bickel, A., *Amer. Lab.*, 1990, Oct., 94.

11. van der Linden, W.E., Classification and definition of analytical methods based on flowing media, *IUPAC note*, 1991.

12. Ruzicka, J. and Hansen, E.H., *Flow Injection Analysis*, 2nd Edn., Wiley, New York, 1988.

13. Valcarcel, M. and Luque de Castro, M.D., *Flow Injection Analysis: Principles and Applications*, Ellis Horwood, Chichester, 1987.

14. Ruzicka, J., *Frezenius Z. Anal. Chem.*, 1986, **324**, 745.

15. Stockwell, P.B., *J. Auto. Chem.*, 1990, **12**, 92.

16. Whitaker, M.J., *Amer. Lab.*, 1983, Mar., 154.

17. Proceedings of Flow Analysis I, *Anal. Chim. Acta*, 1980, **114**, 1-338.

18. Proceedings of Flow Analysis II, *Anal. Chim. Acta*, 1983, **145**, 1-226.

19. Proceedings of Flow Analysis III, *Anal. Chim. Acta*, 1986, **179**, 1-518.

20. Proceedings of Flow Analysis IV, *Anal. Chim. Acta*, 1988, **214**, 1-486.

21. Proceedings of Flow Analysis V, *Anal. Chim. Acta*, 1992, **261**, 1-582.

22. Karlberg, B. and Pacey, G.E., *Flow Injection Analysis: a Practical Guide*, Elsevier, Amsterdam, 1989.

23. Burguero, J.L., *Flow Injection Atomic Spectroscopy*, Dekker, New York, 1989.

24. *Journal of Flow Injection Analysis*, The Japanese Association for Flow Injection Analysis.

25. Ranger, C.B., *Flow Injection Analysis: A new approach to near real time monitoring*. In *Automated Stream Analysis for Process Control*, (Manka, D.P., Ed.) pp. 39-67, Academic Press, New York, 1982.

26. Mowery, R.A., *Intech*, 1984, **31**, 51.

27. Mowery, R.A., *ISA Trans*, 1985, **24**, 1.

28. van der Linden, W.E., *Anal. Chim. Acta*, 1986, **179**, 91.

29. Ruzicka, J., *Anal. Chim Acta*, 1986, **190**, 155.

30. Gisin, M. and Thommen, C., *Anal. Chim. Acta*, 1986, **179**, 165.

31. Lazaro, F., Luque de Castro, M.D. and Valcarcel, M., *J. Pharm. Biomed. Anal.*, 1988, **6**, 585.

32. Christian. G.D. and Ruzicka, J., *Chem. Eng. (NY)*, 1988, **95**, 57.

33. Luque de Castro, M.D., *Talanta*, 1989, **36**, 561.

34. Bergamin, H.F., Zagatto, E.A.G., Krug, F.J. and Reis, B.F., *Anal. Chim. Acta*, 1978, **101**, 17.

35. Frenzel, W., *Ferenius Z. Anal. Chem.*, 1988, **392**, 668.

36. Goto, M., *Trends Anal. Chem.*, 1983, **2**, 92.

37. Gisin, M. and Thommen, C., *Trends Anal. Chem.*, 1989, **8**, 62.

38. Worsfold, P.J., Clinch, J.R. and Casey, H., *Anal. Chim. Acta*, 1987, **197**, 43.

39. Clinch, J.R., Worsfold, P.J. and Casey, H., *Anal. Chim. Acta*, 1987, **200**, 523.

40. Clinch, J.R., Worsfold, P.J., Casey, H. and Smith, S.M., *Anal. Proc*, 1988, **25**, 71.

41. Casey, H., Clarke, R.E., Smith, S.M., Clinch, J.R. and Worsfold, P.J., *Anal. Chim. Acta*, 1989, **227**, 379.

42. Benson, R.L., Worsfold, P.J. and Sweeting, F., *Anal. Proc.*, 1989, **26**, 385.

43. Clinch, J.R., Worsfold, P.J. and Casey, H., *Anal. Chim. Acta*, 1988, **214**, 401.

44. Chen, D., Luque de Castro, M.D. and Valcarcel, M., *Anal. Chim. Acta*, 1990, **230**, 137.

45. Benson, R.L., Worsfold, P.J., and Sweeting, F., *Anal. Chim. Acta*, 1990, **238**, 177.

46. Recktenwald, A., Kroner, K.-H. and Kula, M.-R., *Enzyme Microb. Technol.*, 1985, **7**, 607.

47. Nalbach, U., Schiemenz, H., Stamm, W.W., Hummel, W. and Kula, M.-R., *Anal. Chim. Acta*, 1988, **213**, 55.

48. Nikilajsen, K., Nielsen, J. and Villadsen, J., *Anal. Chim. Acta*, 1988, **214**, 137.

49. Stamm, W.W., Pommerening, G., Wandrey, C. and Kula, M.-R., *Enzyme Microb. Technol.*, 1989, **11**, 96.

50. Worsfold, P.J., Whiteside, I.R.C., Pfeiffer, H.F. and Waldhoff, H., *J. Biotechnol.*, 1990, **14**, 81.

51. Chung, S., Wen, X., Vilholm, K., de Bang, M., Christian, G.D. and Ruzicka, J., *Anal. Chim. Acta*, 1991, **249**, 77.

52. van der Linden, W.E., Mulder, R.J. and Overman, L.J., *Anal. Chim. Acta*, 1986, **190**, 1.

53. Proceedings of Anabiotec 1988, *Anal. Chim. Acta*, 1988, **213**, 1-282.

54. Proceedings of Anabiotec 1991, *Anal. Chim. Acta*, 1991, **249**, 1-302.

55. *J. Biotechnol.*, 1990, 14.

56. Bradley, J., *Process Control Qual.*, 1991, **1**, 157.

57. Frenzel, W., *Fresenius Z. Anal. Chem.*, 1990, **336**, 21.

58. Frenzel, W., *Fresenius Z. Anal. Chem.*, 1992, **342**, 817.

59. Kuban, V. and Dasgupta, P.K., *Anal. Chem.*, 1992, **64**, 1106.

60. Gonzalo, E.R., Pavon, J.L.P., Ruzicka, J., Christian, G.D. and Olsen, D.C., *Anal. Chim. Acta*, 1992, **259**, 37.

61. Alonso, J., Bartroli, J., Del Valle, M., Escalada, M. and Barber, R., *Anal. Chim. Acta*, 1987, **199**, 191.

62. Whitman, D.A. and Christian, G.D., *Talanta*, 1989, **36**, 205.

63. Rios, A., Luque de Castro, M.D. and Valcarcel, M., *Talanta*, 1989, **36**, 612.

64. Ruzicka, J., *Anal. Chim. Acta*, 1990, **237**, 329.

65. Ruzicka, J., Marshall, G.D. and Christian, G.D., *Anal. Chem.*, 1990, **62**, 1861.

66. Gubeli, T., Christian, G.D. and Ruzicka, J., *Anal. Chem.*, 1991, **63**, 2407.

67. Luque de Castro, M.D. and Valcarcel, M., *Trends Anal. Chem.*, 1986, **5**, 71.

68. Luque de Castro, M.D. and Valcarcel, M., *Analyst*, 1984, **109**, 413.

69. Luque de Castro, M.D., *Talanta*, 1986, **33**, 45.

70. Koruda, R., Nara, T. and Oguma, K., *Analyst*, 1988, **113**, 1557.

71. Muller, H., Muller, V. and Hansen, E.H., *Anal. Chim. Acta*, 1990, **230**, 113.

72. Romero-Saldana, M., Luque de Castro, M.D. and Valcarcel, M., *Talanta*, 1991, **38**, 291.

73. Faizullah, A.T. and Townshend, A., *Anal. Chim. Acta*, 1985, **167**, 225.

74. Al-Sowdani, K.H. and Townshend, A., *Anal. Chim. Acta*, 1986, **179**, 469.

75. Faizullah, A.T. and Townshend, A., *Anal. Chim. Acta*, 1986, **179**, 233.

76. Devi, S. and Townshend, A., *Anal. Chim. Acta*, 1989, **225**, 331.

77. Masoom, M. and Townshend, A., *Anal. Chim. Acta*, 1985, **171**, 185.

78. Yao, T. and Wasa, T., *Anal. Chim. Acta*, 1985, **175**, 301.

79. Lazaro, F., Luque de Castro, M.D. and Valcarcel, M., *Anal. Chem.*, 1987, **59**, 1859.

80. Yao, T. and Wasa, T., *Anal. Chim. Acta*, 1988, **207**, 319.

81. Massoom, M., *Anal. Chim. Acta*, 1988, **214**, 173.

82. Matsumoto, K., Kamikado, H., Matsubara, H. Otsajima, Y., *Anal. Chem.*, 1988, **60**, 147.

83. Cosana, J.S., Calle, J.L., Pinillos, J.L., Linares, P. and Luque de Castro, M.D., *Anal. Chim. Acta*, 1989, **221**, 173.

84. Morishita, F., Nishikawa, Y. and Kojima, T., *Anal. Sci.*, 1986, **2**, 411.

85. Rios, A., Luque de Castro, M.D. and Valcarcel, M., *Anal. Chem.*, 1986, **58**, 663.

86. Rios, A., Luque de Castro, M.D. and Valcarcel, M., *Anal. Chim. Acta*, 1986, **187**, 139.

87. Bermudez, B., Rios, A., Luque de Castro, M.D. and Valcarcel, M., *Talanta*, 1988, **35**, 810.

88. Pavon, J.L.P., Cordero, B.M., Maendez, J.H. and Agudo, R.M.I., *Anal. Chem.*, 1989, **61**, 1789.

89. Alonso, J., Bartroli, J., Del Valle, M. and Barber, R., *Anal. Chim. Acta*, 1989, **219**, 345.

90. Arauja, A.N., Lima, J.L.F.C., Rangel, A.O.O.S., Alonso, J., Bartroli, J. and Barber, R., *Analyst*, 1989, **114**, 1465.

91. Pavon, J.L.P., Pinto, G.C., Cordero, B.M. and Mendez, J.H., *Anal. Chem.*, 1990, **62**, 2405.

92. Whitman, D.A., Christian, G.D. and Ruzicka, J., *Anal. Chim. Acta*, 1988, **214**, 197.

93. Trojanowicz, M. and Spunzar-Lobinska, J., *Anal. Chim. Acta*, 1990, **230**, 125.

94. Canete, F., Rios, A., Luque de Castro, M.D. and Valcarcel, M., *Anal. Chim. Acta*, 1988, **214**, 375.

95. Scolari, C.A. and Brown S.D., *Anal. Chim. Acta*, 1985, **178**, 239.

96. Matuszewski, W., Trojanowicz, M. and Ilcheva, L., *Electroanalysis*, 1990, **2**, 147.

97. Owen, A.J., *The Diode-array Advantage in UV/visible Spectroscopy*, Hewlett Packard, 1988.

98. Wolf, K. and Worsfold, P.J., *Anal. Proc.*, 1986, **23**, 365.

99. Lazaro, F., Rios, A., Luque de Castro, M.D. and Valcarcel, M., *Analusis*, 1986, **14**, 378.

100. Vithanage, R.S. and Dasgupte, P.K., *Anal. Chem.*, 1986, **58**, 326.

101. Lazaro, F., Rios, A., Luque de Castro, M.D. and Valcarcel, M., *Anal. Chim. Acta*, 1986, **179**, 279.

102. Lazaro, F., Luque de Castro, M.D. and Valcarcel, M., *Anal. Chim. Acta*, 1986, **185**, 57.

103. Rios, A., Lazaro, F., Luque de Castro, M.D. and Valcarcel, M., *Anal. Chim. Acta*, 1987, **199**, 279.

104. Kuban, V., Gladilovich, D.B. and Sommer, L., *Talanta*, 1989, **36**, 463.

105. Bermudez, B., Lazaro, F., Luque de Castro, M.D. and Valcarcel, M., *Analyst*, 1987, **112**, 535.

106. Wada, H., Murakawa, T. and Nakagawa, G., *Anal. Chim. Acta*, 1987, **200**, 515.

107. Leon, L., Rios, A., Luque de Castro, M.D. and Valcarcel, M., *Lab. Rob. Autom.*, 1989, **1**, 295.

108. Malgarejo, A.G., Pavon, J.M.C. and Castro, A.R. *Anal. Chim. Acta*, 1990, **241**, 153.

109. West, P.W. and Gaeke, G.K., *Anal. Chem.*, 1956, **28**, 1816.

110. Stephens, B.G. and Lindstrom, F., *Anal. Chem.*, 1964, **36**, 1309.

111. Humphrey, R.E., Ward, M.H. and Hinze, W., *Anal. Chem.*, 1970, **42**, 698.

112. Lazaro, F., Luque de Castro, M.D. and Valcarcel, M., *Analusis*, 1987, **15**, 183.

113. Brown, D.S. and Jenke, D.R., *Analyst*, 1987, **112**, 899.

114. Fogg, A.G., Wang, X. and Tyson, J.F., *Analyst*, 1990, **115**, 305.

115. Grandos, M., Maspoch, S. and Blanco, M., *Anal. Chim. Acta*, 1987, **192**, 445.

116. Fogg, A.G., Guta, C.W. and Chamsi, A.Y., *Analyst*, 1987, **112**, 253.

117. Koukli, I.I. and Calokerinos, A.C., *Anal. Chim Acta*, 1987, **192**, 333.

118. Jenke, D.R. and Raghaven, N., *J. Chromatogr. Sci.*, 1985, **23**, 75.

119. Jenke, D.R., *J. Chromatogr. Sci.*, 1986, **24**, 352.

120. Masoom, M. and Towhshend, A., *Anal. Chim. Acta*, 1986, **179**, 399.

121. Al Tamrah, S.A., Townshend, A. and Wheatley, A.R., *Anal. Chim. Acta*, 1987, **112**, 883.

122. Vogel, A.I., *Quantitative Inorganic Analysis*, Longman, London, 1961.

123. Clinch, J.R., *PhD Thesis*, University of Hull, 1988

124. Benson, R.L., *PhD Thesis*, University of Hull, 1991

125. Jocelyn, P.C., *Biochemistry of the SH Group*, Academic Press, London, 1972.

126. Parker, A.J. and Kharasch, N., *Chem. Rev.*, 1959, **59**, 583.

127. Massart, D.L., Vandeginste, B.G.M., Deming, S.N., Michotte, Y. and Kaufman, L., *Chemometrics: a Textbook*, Elsevier, Amsterdam, 1988.

128. Malinowski, E.R., *Factor Analysis in Chemistry*, 2nd. Edn., Wiley, Chichester, 1991.

129. Wold,S. in Geladi, P. and Esbensen, K., *J. Chemom.*, 1990, **4**, 337.

130. de Jong, S., *Mikrochim. Acta*, 1991, **2**, 93.

131. *Chemometrics and Intelligent Laboratory Systems*

132. *Journal of Chemometrics*

133. Sharaf, M.A., Illman, D.L. and Kowalski, B.R., *Chemometrics*, Wiley, Chichester, 1990.

134. Brereton, R.G., *Chemometrics, Applications of Mathematics and Statistics to laboratory systems*, Ellis Horwood, Chichester, 1990.

135. Haswell, S.J., *Practical Guide to Chemometrics*, Marcel Dekker, New York, 1992.

136. Vandeginste, B.G.M., *Top. Curr. Chem.*, 1987, **141**, 1.

137. Kowalski, B.R., *Trends Anal. Chem.*, 1981, **1**, 71.

138. Borman, S.A., *Anal. Chem.*, 1982, **54**, 1379A.

139. Petersen, K., Lopez, J.L. and Dasgupta, P.K., *J. Chemom.*, 1989, **3**, 601.

140. Martens, H. and Naes, T., *Multivariate Calibration*, Wiley, Chichester, 1989.

141. Miller, J.C. and Miller J.N., *Statistics for Analytical Chemistry*, 2nd. Edn., Ellis Horwood, Chichester, 1988.

142. Miller, J.N., *Analyst*, 1991, **116**, 3.

143. Martens, H. and Naes, T., *Trends Anal. Chem.*, 1984, **3**, 204.

144. Beebe, K.R. and Kowalski, B.R., *Anal. Chem.*, 1987, **59**, 1007A.

145. Sanchez, E. and Kowalski, B.R., *J. Chemom.*, 1988, **2**, 247.

146. Martens, H., *PhD. Thesis*, Technical University of Trondheim, 1985.

147. Naes, T. and Martens, H., *Trends Anal. Chem.*, 1984, **3**, 266.

148. Wold, S., Esbensen, K. and Geladi, P., *Chemom. Intel. Lab. Sys.*, 1987, **2**, 37.

149. Gemperline, P.J., Miller, K.H., West, T.L., Weinstein, J.E., Hamilton, J.C. and Bray, J.T., *Anal. Chem.*, 1992, *64*, 523A.

150. Naes, T. and Martens, H., *J. Chemom.*, 1988, **2**, 155.

151. Naes, T. and Isaksson, T., *App. Spec.*, 1989, **43**, 328.

152. Downey, G., Robert, P., Bertyrand, D. and Devaux, M.F., *J. Chemom.*, 1989, **3**, 397.

153. Geladi, P. and Kowalski, B.R., *Anal. Chim. Acta*, 1986, **185**, 1.

154. Geladi, P. *J. Chemom.*, 1988, **2**, 231.

155. Lindberg, W., Ohman, J., Wold, S. and Martens, H., *Anal. Chim. Acta*, 1985, **174**, 41.

156. Sjostrom, M., Wold, S., Lindberg, W., Persson, J. and Martens, H., *Anal. Chim. Acta*, 1983, **150**, 61.

157. Lindberg, W., Persson, J. and Martens, H., *Anal. Chem.*, 1983, **55**, 643.

158. Karstang, T.V. and Eastgate, R.J., *Chemom. Intel. Lab. Sys.*, 1987, **2**, 209.

159. Haaland, D.M. and Thomas, E.V., *Anal. Chem.*, 1988, **60**, 1202.

160. Haaland, D.M., *Anal. Chem.*, 1988, **60**, 1208.

161. Guzman, M., de Bang, M., Ruzicka, J. and Christian, G.D., *Process Cont. Qual.*, 1992, **2**, 113.

162. Haaland, D.M., Higgins, K.L. and Tallant, D.R., *Vib. Spec.*, 1990, **1**, 35.

163. Carey, W.P. and Wangen, L.E., *Chemom. Int. Lab. Sys.*, 1991, **10**, 245.

164. Unscrambler II User's Guide, CAMO A/S, Trondheim, Norway, 1992.

165. Tysso, V., Esbensen, K. and Martens, H., *Chemom. Intel. Lab. Sys.*, 1987, **2**, 239.

166. Davies, O.L. (Ed.), *The Design and Analysis of Industrial Experiments*, Oliver & Boyd, London, 1956.

167. Morgan, E., Burton, K.W. and Church, P.A., *Chemom. Intel. Lab. Sys.*, 1989, **5**, 283.

168. Morgan, E., *Chemometrics: Experimental Design*, Wiley, Chichester, 1991.

169. Press, W.H., Flannery, B.P., Teukolsky, S. and Vetterling, W.T., *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press, Cambridge, 1986.

170. Wold, S., Esbensen, K. and Geladi, P., *Chemom. Intel. Lab. Sys.*, 1987, **2**, 37.

171. Wolf, K., *PhD. Thesis*, University of Hull, 1988.

172. Blanco, M., Gene, J., Iturriaga, H. and Maspoch, S., *Analyst*, 1987**112**, 619.

173. Blanco, M., Gene, J., Iturriaga, H., Maspoch and Riba, J., *Talanta*, 1987, **34**, 987.

174. Lukkari, I. and Lindberg, W., *Anal. Chim. Acta*, 1988, **211**, 1.

175. Kuban, V. and Dolezel, P., *Collect. Czech. Chem. Commun.*, 1988, **53**, 543.

176. Kuban, V. and Gladilovich, D.B., *Collect. Czech. Chem. Commun.*, 1989, **218**, 303.

177. Erickson, B.C., Ruzicka, J. and Kowalski, B.R., *Anal. Chim. Acta*, 1989, **218**, 303.

178. Blanco, M., Coello, J., Gene, J., Iturriaga, H. and Maspoch, S., *Anal. Chim. Acta*, 1989, **224**, 23.

179. Blanco, M., Coello, J., Gene, J., Iturriaga, H. and Maspoch, S., *Quim. Anal.*, 1989, **8**, 223.

/ H

180. Fernandez-Band, B., Lazaro, F., Luque de Castro, M.D. and Valcarcel, M., *Anal. Chim. Acta*, 1990, **229**, 177.

181. Gerritsen, M.J.P., Kateman, G., van Opstal, M.A.J., van Bennekom, W.P. and Vandeginste, B.G.M., *Anal. Chim. Acta*, 1990, **241**, 23.

182. Lindberg, W., Clark, G.D., Hanna, C.P., Whitman, D.A., Christian, G.D. and Ruzicka, J., *Anal. Chem.*, 1990, **62**, 849.

183. Blanco, M., Coello, J., Gene, J., Iturriaga, H. and Maspoch, S., *Fresenius' J. Anal. Chem.*, 1990, **338**, 831.

184. Whitman, D.A., Seasholtz, M.B., Christian, G.D., Ruzicka, J., Kowalski, B.R., *Anal. Chem.*, 1991, **63**, 775.

185. Guzman, M., Christian, G.D., Ruzicka, J., and Shelley, P., *Vib. Spec.*, 1991, **2**, 1.

186. *Standard Methods for the Examination of Water and Wastewater*, American Public Health Association, American Water Works Association, Water Pollution Control Federation, Washington DC, 17th Edn., 1989.

187. Johnson, K.S. and Petty, R.L., *Anal. Chem.*, 1982, **54**, 1185.

188. Lacy, N., Christian, G.D. and Ruzicka, J., *Quim. Anal.*, 1989, **8**, 201.

189. Leggett, D.J., Chem, N.H. and Mahadevappa, D.S., *Fresenius' Anal. Chem.*, 1983, **315**, 47.

190. Gordon, G., Sweetin, D.L., Smith, K. and Pacey, G.E., *Talanta*, 1991, **107**, 145.

191. Johnson, J.D. and Overby, R., *Anal. Chem.*, 1969, **41**, 1744.

192. Leggett, D.J., Chen, N.H. and Mahadevappa, D.S., *Analyst*, 1982, **107**, 433.

193. Ruzicka, J. and Hansen, E.H., *Anal. Chim. Acta*, 1975, **78**, 145.

194. Isaksson, T. and Naes, T., *Appl. Spectrosc.*, 1988, **42**, 1273.

195. Kowalski, B.R. and Seasholtz, M.B., *J. Chemom.*, 1991, **5**, 129.

196. Seasholtz, M.B. and Kowalski, B.R., *J. Chemom.*, 1992, **6**, 103.

197. Osten, D.W., *J. Chemom.*, 1988, **2**, 39.

198. Gray, H.L. and Schucany, W.R., *The Generalised Jackknife Statistic*, Dekker, New York, 1972.

199. Efron, B., *The Jackknife, the Bootstrap and other Resampling Plans*, Society for Industrial and Applied Mathematics, Philadelphia, 1982.

200. Quenouille, M., *Biometrika*, 1956, **43**, 353.

201. Tukey, J.W., *Ann. Math. Stat.*, 1958, **29**, 614.

202. *MATLAB 386 User's Guide*, The Mathworks Inc., South Natick, 1989.

203. Wold S., Ruhe, A., Wold, H. and Dunn III, W.J., *SIAM J. Sci. Stat. Comput.*, 1984, **5**, 735.

204. Naes, T., *Technomet.*, 1985, **27**, 301.

205. Fearn, T., *Appl. Statist.*, 1983, **32**, 73.