MULTIVARIATE CALIBRATION FOR ICP-AES

by

MICHAEL LEE GRIFFITHS

A thesis submitted to the University of Plymouth

in partial fulfilment for the degree of

DOCTOR OF PHILOSOPHY

Department of Environmental Sciences

Faculty of Science

In collaboration with

The Analytical Innovation Programme

Johnson Matthey Ltd.

Department of Trade and Industry

March 2001

ABSTRACT

MULTIVARIATE CALIBRATION FOR ICP-AES

MICHAEL LEE GRIFFITHS BSc. (Hons)

The analysis of metals is now a major application area for ICP-AES, however, the technique suffers from both spectral and non-spectral interferences. This thesis details the application of univariate and multivariate calibration methods for the prediction of Pt, Pd, and Rh in acid-digested and of Au, Ag and Pd in fusion-digested autocatalyst samples.

Of all the univariate calibration methods investigated matrix matching proved the most accurate method with relative root mean square errors (RRMSEs) for Pt, Pd and Rh of 2.4, 3.7, and 2.4 % for a series of synthetic test solutions, and 12.0, 2.4, and 8.0 % for autocatalyst samples. In comparison, the multivariate calibration method (PLS1) yielded average relative errors for Pt, Pd, and Rh of 5.8, 3.0, and 3.5 % in the test solutions, and 32.0, 7.5, and 75.0 % in the autocatalyst samples.

A variable selection procedure has been developed enabling multivariate models to be built using large parts of the atomic emission spectrum. The first stage identified and removed wavelengths whose PLS regression coefficients were equal to zero. The second stage ranked the remaining wavelengths according to their PLS regression coefficient and estimated standard error ratio. The algorithms were applied to the emission spectra for the determination of Pt, Pd and Rh in a synthetic matrix. For independent test samples variable selection gave RRMSEs of 5.3, 2.5 and 1.7 % for Pt, Pd and Rh respectively compared with 8.3, 7.0 and 3.1 % when using integrated atomic emission lines. Variable selection was then applied for the prediction of Au, Ag and Pd in independent test fusion digests. This resulted in RRMSEs of 74.2, 8.8 and 12.2 % for Au, Ag and Pd respectively which were comparable to those obtained using a more traditional univariate calibration approach.

A preliminary study has shown that calibration drift can be corrected using Piecewise Direct Standardisation (PDS). The application of PDS to synthetic test samples analysed 10 days apart resulted in RRMSEs of 4.14, 3.03 and 1.88%, compared to 73.04, 44.39 and 28.06 % without correction, for Pt, Pd, and Rh respectively.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

A	
Autoscaled (As)	Form of pre-processing that adjusts all variables to have the same influence in a multivariate mathematical regression model
B	
Background correction points (BCPs)	Used to correct for baselines changes in ICP-AES
C	
Certified reference material (CRM)	Material with known elemental composition
Confidence interval (CI)	Limits within which a point estimation lies with some degree of confidence, e.g. 95 %
Cumulative sum (CS)	Sum of x, x + y, x + (x + y) +.....e.t.c.
Classical least squares (CLS)	Mathematical form relating r = c x S
D	
Degrees of freedom (DOF)	Number of independent units of information relevant to estimating some parameter
Direct standardisation (DS)	Multivariate technique used to standardise instrumental output using the full spectrum
G	
Graphite furnace atomic absorption spectroscopy (GFAAS)	Atomic absorption spectroscopy using a graphite furnace as the excitation source
H	
High training range (HTR)	Calibration data using high concentrations
I	
Independent & identically distributed (i.i.d.)	Bearing no correlation to one another
Interelement correction factor (IEC)	Used to correct for spectral interference's in ICP-AES
Inductively coupled plasma atomic emission spectroscopy (ICP-AES)	Spectroscopy using atomic and ionic electromagnetic emission in a quantifiable manner
Inductively coupled plasma mass spectroscopy (ICP-MS)	Mass spectroscopy using charge/mass ratio in a quantifiable manner

Informative variable degradation (IVD)	Multivariate technique used to find informative variables
Individual wavelengths (IW)	As opposed to the segmented full spectrum
J	
Jackknife (JK)	Technique used to bypass statistical theory and find the standard error values of complicated statistical estimators
L	
Low training range (LTR)	Calibration data using high concentrations
Leave one out (LOO)	The iterative procedure of calculating a parameter of interest, usually an error value, with all calibration/test samples present bar one. This is performed with all samples iteratively left out, and the parameter calculated as an overall average.
Limit of detection (LOD)	Lowest quantifiable concentration
M	
Mean centered (Mc)	Form of preprocessing, making the mean equal to zero, thereby making differences in the sample spectra more discernible
Multiple linear regression (MLR)	Regression using multiple variables
N	
Not-given (n/g)	Not available
Non-detectable (n/d)	Below limit of detection
Non-linear iterative partial least squares (NIPALS)	Algorithm which attempts to solve the eigenvalue problem
O	
Orthogonal array (OA)	A vector space in which the vectors are at right angles to one another
P	
Parts per billion (PPB)	1 part in 1 x 10^9 parts (ng ml^{-1})
Parts per million (PPM)	1 part in 1 x 10^6 parts (µg ml^{-1})
Partial least squares 1 (PLS1)	Multivariate techniques used to compress data and also perform multiple regression
Piecewise direct standardisation (PDS)	Multivariate technique used to standardise instrumental output using a partial spectrum

Principal component (PC)	A projection of points in multidimensional space onto a line in two dimensional space
Principal components analysis (PCA)	Multivariate technique used to compress complex data into a set of PCs
Principal components regression (PCR)	See PLS (note, there exist major differences in these two regression techniques)
R	
Relative standard error (RSE %)	Error relative to the mean as a percentage
Relative root mean square error (RRMSE %)	Used in multivariate regression as an assessment of overall error using all predicted values and expressed as a percentage of the true mean value
Relative root mean square of cross validation (RRMSECV %)	Used in multivariate regression as an assessment of overall error by using a LOO strategy and expressed as a percentage of the true mean value
Root mean square of cross validation (RMSECV %)	Used in multivariate regression as an assessment of overall error by using all samples and expressed as a percentage
Relative root mean square error of calibration (RRMSEC %)	As in RRMSE but for the calibration
Relative root mean square error of prediction (RRMSEP %)	As in RRMSE but for the prediction
S	
Segmented full spectrum (SFS)	Spectral output from the Perkin-Elmer Optima 3000 is in the form of disjointed regions of the spectrum, i.e. 194-195 nm, 196-197 nm, e.t.c The full spectrum is covered in terms of limits (190-782 nm) but portions are missing, i.e. it is segmented
Standard error (Se)	The standard deviation of a statistical parameter, e.g. σ
Stepwise multiple linear regression (SMLR)	As in MLR, but with variables, but using only those variables that match some specific criterion
U	
Uninformative variable elimination by partial least squares (UVE-PLS)	Multivariate technique used to remove uninformative variables

V

Type I error variables ($Var_{Type\ I}$)	A variable which is the result of a type I error
UVE-PLS datasets (Var_{sel})	The dataset selected by UVE-PLS
Informative variable degradation ratio (Var_j^{ivd})	The ratio of a variables regression coefficient to the standard error of the regression coefficient

X

X-Ray fluorescence (XRF)	A branch of spectroscopy

ACKNOWLEDGMENTS

Many people have helped me during the course of this piece of work, none more so than my supervisors Hywel Evans and Paul Worsfold without whos guidance over the past three years this manuscript would not be here in front of you now. Their help and guidance has been invaluable to me.

I am very grateful to the financial support that I have received from the University of Plymouth, the Department of Trade and Industry, the Analytical Innovation Programme and Johnson Matthey Ltd. I would particularly like to thank my industrial sponsors Johnson Matthey Ltd, in particular Les Guy for his genius in getting data out of the Optima 3000 and Peter Ash for all his help at Johnson Matthew's research facility.

I would also like to thank John Kalivas and his family for a wonderful stay in America last summer and for the opportunity of studying at Idaho State University. Particularly I would like to thank Lorie for her wonderful cooking and John for his unending enthusiasm and his help is solving some tough problems.

My thanks must also go to all the staff on the fifth floor of the Davey building for their patience when all the glassware went missing, and for their express delivery of many kinds of weird and wonderful chemicals. I would also like to express my heart-felt gratitude to Rob and Anita for keeping the Optima on its tracks when everyone else would have given up.

Last, but certainly not least, I would like to thank Helen my wife for her unending support in every way possible and for her gentle nudging when writing this thesis, and it is to Helen that I dedicate this work.

AUTHOR'S DECLARATION

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

This study was financed with the aid of studentship from the University of Plymouth and with joint CASE equivalent support from the Department of Trade and Industry, the Analytical Innovation Programme and Johnson Matthey Ltd.

A program of advanced study was undertaken, which included a five-week study visit to Idaho State University (Idaho, USA) under the tutelage of Professor John Kalivas, and short courses on Atomic Spectroscopy. Relevant scientific seminars and conferences were regularly attended at which work was presented and four papers were prepared for publication.

Publications

1 Comparison of traditional and multivariate calibration techniques applied to complex matrices using inductively coupled plasma atomic emission spectroscopy, Mike L. Griffiths, Daniel Svozil, Paul J. Worsfold, Sue Denham and E. Hywel Evans, Journal of Analytical Atomic Spectroscopy, 2000, 15, 967-972.

2 Uninformative Variable Elimination Partial Least Squares (UVE-PLS) Procedure for Quantitative Analysis using Atomic Emission Spectrometry
Mike L. Griffiths, Daniel Svozil, Paul J. Worsfold, Sue Denham and E. Hywel Evans, Submitted to the Journal of Chemometrics, March 2001.

3 Uninformative Variable Elimination Partial Least Squares (UVE-PLS) Procedure for the Quantitative Analysis of Complex Industrial Samples using Atomic Emission Spectrometry, Mike L. Griffiths, Daniel Svozil, Paul J. Worsfold, Sue Denham and E. Hywel Evans, Submitted to the Journal of Analytical Atomic Spectroscopy, March 2001.

4 The Application of Piecewise Direct Standardisation (PDS) in conjunction with an Uninformative Variable Elimination Partial Least Squares (UVE-PLS) Procedure for Quantitative Analysis using Atomic Emission Spectrometry.

Conferences and Presentations:

1 The Ninth Biennial National Atomic Spectroscopy Symposium (BNASS), Bath 8-10 July 1998, *conference & poster presentation.*

2 Short Courses in Analytical Chemistry (University of Plymouth in conjunction with the LGC. 29 June – 9th July 1998, *lectures.*

3 One day course in Multivariate Calibration, Cambridge (Royal Society of Chemistry), Cambridge, 10th August 1998, *lectures.*

4 'Multivariate Calibration and ICP-AES', European Winter Conference, Pau, France 10th-15th January 1999, conference & poster presentation.

5 'Optimisation of the Perkin-Elmer Optima 3000' Elemental Assay Seminar (LGC and the DTI) March 15th 1999, oral presentation.

6 Analytical Science Network – Young Scientists Meeting, Wednesday, 17th November 1999, University of Plymouth, *conference & poster.*

7 Matlab for Environmental Science Seminar, 1st March 2000, *oral presentation.*

8 Analytical Division of the RSC, meeting for the "Young Researchers' Meeting and Specialist Symposia, combined with 'The Age of the Molecule' Annual Conference 2000," 16th to 20th April, 2000, University of Manchester (UMIST), Manchester, U.K, *conference & poster presentation.*

9 Five week stay at Idaho State University, Idaho, U.S.A. to visit Professor J. H. Kalivas July-August 2000, *lectures and individual tuition.*

10 Departmental and Research Colloquia, weekly meetings, March 1998 to March 2001.

11 Quarterly meetings with Johnson Matthey Ltd. (industrial sponsor) and the Laboratory of the Government Chemist (Analytical Innovation programme / Department of Trade and Industry).

External Contacts:

Dr. Peter Ash, Johnson Matthey Ltd.

Dr. Peter Lyne, Department of Trade & Industry.

Signed .....

Date .....

For Helen

CHAPTER 1 - INTRODUCTION

1.1 The Determination of Metal Concentrations in Industrial Samples by Inductively Coupled Plasma Atomic Emission Spectrometry (ICP-AES).

1.1.1 Inductively Coupled Plasma Atomic Emission Spectroscopy (ICP-AES)

Of the various plasma sources used for elemental analysis during the past 25 years, the inductively coupled plasma (ICP) has had the most significant impact on the field of atomic spectroscopy (1). Other plasma sources, such as the direct-current plasma (DCP) and the microwave induced plasma (MIP) have been found useful for spectrochemical analysis, however, the ICP continues to be the primary source for assay and trace level metals analysis. In 1980, Stanley Greenfield, one of the prominent pioneers of ICP atomic emission spectrometry (ICP-AES) presented a plenary lecture at the Fifth International Conference of the Society for Analytical Chemistry, in Lancaster, England. In his talk, "Plasma Spectroscopy Comes of Age"(1), Greenfield reflected on the growth of ICP-AES and stated that the technique had achieved respectability and had indeed "come of age".

In practice, all but a few elements (e.g. noble gases, some halogens e.t.c.) can be determined by commercially available equipment. The analysis of metals is now a major application area for ICP-AES (2), and the technique offers several important advantages for the analysis of these materials. Speed of analysis, relatively free of chemical interferences, high sensitivity, the range of determinable elements and the ability to measure simultaneously trace, minor and major constituents by virtue of the plasma's high excitation temperature compared with the flame, are significant factors in its popularity. There are various situations where ICP-AES is especially well suited.

i) Segregated samples which require a dissolution procedure to render them homogenous.

ii) The relatively low cost of the equipment compared with, for example, ICP-MS systems.

iii) Elements which posses very few isotopes (e.g. Rh) and which cannot be analysed using ICP-MS due to matrix complexity.

The analysis of metals in industrial products is a necessary step for manufacturers to verify and control the quality of products and services. The range of metals, and consequently the range of matrices, is vast. These include such materials as aluminium and aluminium alloys, tableted powder products for aluminium and copper processing, copper based alloys, ferro-alloys, boron-containing alloys, hardmetals, chromium metal, magnet alloys, refractory products, precious metals, steels and ferrous metals.

There are many examples of the need for determining trace level (<5 $\mu g\ ml^{-1}$(3)) concentrations of elements within industrial samples. An example is the recovery of Platinum Group Metals (PGM's), such as Pt, Pd, Rh, Au, Ag, Ir, and Ru from materials like spent catalytic converters (used in vehicles with internal combustion engines to convert toxic gases into less harmful gases), electronic and jewellery scrap. Johnson Matthey Plc (Industrial partners to this project), a UK-based precious metals group, processed approximately £400 million worth of precious metal in 1994/95.

The most commonly used techniques for the determination of trace elements are based on atomic spectrometry (GFAAS, ICP-AES, ICP-MS and XRF), which involve the absorption or emission of electromagnetic radiation (GFAAS, ICP-AES, XRF) or the determination of mass/charge (m/z) ratio (ICP-MS) so that meaningful quantitative and qualitative information about a sample can be obtained.

1.2 Inductively Coupled Plasma - Atomic Emission Spectroscopy (ICP-AES)

Inductively Coupled Plasma – Atomic Emission spectroscopy (ICP-AES) is well established as a powerful technique for multielemental analysis (4). The high temperature of the inductively coupled plasma (ICP) minimises matrix effects and produces adequate sensitivity for most metals and some non-metals (e.g. Si and C), in concentrations ranging from percent to ppb.

Reviews of ICP-AES give credit to Babat (5, 6) as the individual who first succeeded in sustaining induction heated plasmas at atmospheric pressure. The stabilisation of an inductively heated plasma operated at atmospheric pressure in gases flowing though an open-ended tube wasn't achieved until 1961/62 by Reed(7-9). Fassel (10, 11) and Greenfield (12) recount their independent analytical studies of ICP's, in 1962, the first results of which were communicated in 1964 (13) and 1965 (14). Major efforts by the groups of Fassel (14-17) and Greenfield (18-21) during the 1960's established the viability of the ICP as a spectrochemical source. A paper published in 1969 by Dickinson and Fassel (15) heralded a new era in ICP-AES by reconfiguring the torch geometry. These authors succeeded in reducing the detection limits to the 0.1-10 ng ml^{-1} range for many elements, which meant an improvement by two or more orders of magnitude compared to the results achieved previously. This report stands as a major landmark in ICP progress (22).

1.2.1 Atomic Emission

When ground state atoms absorb energy, this can result in either an increase in kinetic energy or excitation to a higher energy state causing an electron to become promoted from its ground state to one with a higher energy. This makes the atom less

stable and it will decay back to the ground state by emitting a photon, the wavelength of which is dependent upon the energy difference between the two energy levels and is characteristic of the element. If the energy absorbed by the atom is sufficiently high, an electron may become dissociated from the atom, leaving a positively charged ion. The energy required for this process, known as the ionisation potential, is different for each element. Ions also possess ground and excited states through which they can absorb and emit energy by the same excitation and decay processes as neutral atoms.

When a system is in thermodynamic equilibrium the level population, i.e. the number of atoms N_j in the excited state, is given by the Boltzman distribution law:

$$N_j = N_o \frac{g_j}{g_0} exp[-(E_j - E_0)/kT] \qquad \text{Eqn. 1.1}$$

where N_0 is the number of atoms in the ground state with relative energy $E_0 = 0$, E_j is the energy of the excited state, g_j and g_0 are the statistical weights of the jth (excited) and ground states, respectively. (where $g = 2J + 1$, J is the third quantum number), k is the Boltzman constant and T is the temperature.

The difference in energy between the ground and excited state of a radiative transition defines the wavelength of the radiation that is involved in that transition. Figure 1.1 shows the excitation, ionisation and emission processes schematically. The horizontal lines of this highly simplified Grotian diagram represent the energy levels of an atom. The vertical arrows represent energy transitions.

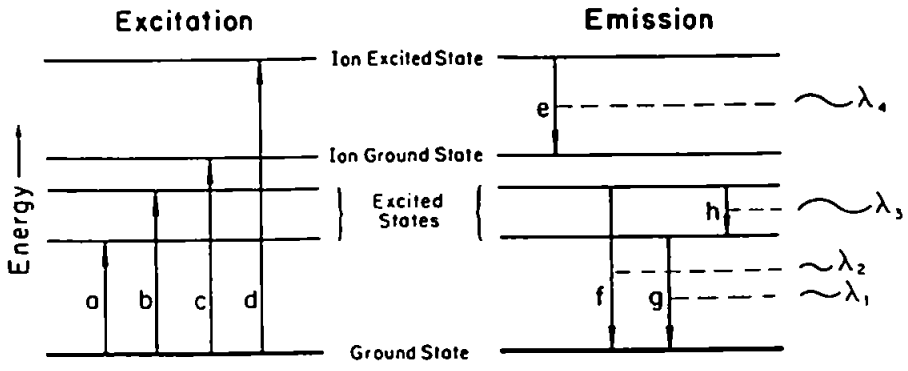


Figure 1.1 Highly simplified Grotian diagram showing energy transitions where a and b represent excitation, c is ionisation, d is excitation, e is ion emission, and f, g and h are atom emission

The relationship between the energy difference and wavelength can be derived through Planck's equation, Eq 1.2

$$\Delta E = h \qquad \text{Eq 1.2}$$

Where E is the energy difference between the two levels, h is Planck's constant, and ν is the frequency of the radiation. Substituting c/λ for ν, where c is the speed of light and λ is the wavelength, we arrive at

$$\Delta E = \frac{hc}{\lambda} \qquad \text{Eqn. 1.3}$$

Each element has its own characteristic set of energy levels and thus its own unique set of emission wavelengths. It is this property that makes atomic spectrometry useful for element-specific analytical techniques.

1.2.2 ICP-AES Instrumentation

1.2.2.1 Sample Introduction

Most samples are introduced into the plasma as a liquid in the form of an aerosol generated using a pneumatic nebuliser. Two commonly used pneumatic nebulisers are the concentric and cross-flow nebuliser. However, these nebulisers can only tolerate solutions containing less than 0.1-1% dissolved solids (23), as such precautions must be made to prevent 'clogging' of the tip. A pneumatic meinhard concentric glass type nebuliser is shown in Fig. 1.2. The analyte solution is fed through the central channel of the nebuliser and the nebuliser gas flows around it. The rapidly flowing nebuliser gas creates a low-pressure region at the end of the capillary (the venturi effect), which results in the liquid being drawn and 'fragmenting' into droplets. The aerosol is then swept through a spray chamber, the primary functions of which are to remove the larger aerosol droplets and smooth out pulses that occur during nebulisation, and then into the injector which directs the sample into the ICP.

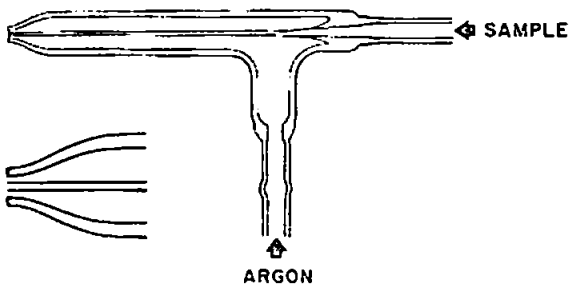


Figure 1.2 Meinhard concentric glass nebuliser

In general, spray chambers for the ICP are designed to allow the droplets with diameters of around 10 μm or smaller into the plasma. With typical nebulisers, this droplet range constitutes about 1-5% of the sample that is introduced to the nebuliser, the remaining 95-99% going to waste (24).

1.2.2.2 Plasma Generation

Figure 1.3 is a schematic drawing of an ICP torch, which consists of an assembly of three concentric tubes, most frequently made of silica or quartz. The torch is placed within a water-cooled radio frequency (RF) generator coil. When RF power (typically 700-1500 W) is applied to the load coil, an alternating current oscillates within the coil, at a rate corresponding to the frequency of the supplying generator, typically 27 or 40 MHz. An argon gas flow is introduced tangentially into the intermediate and outer tubes and the application of a spark causes electron collision and partial ionisation of the argon. The RF oscillation of the current in the coil generates oscillating magnetic fields with lines of force axially oriented inside the coil. These induced oscillating magnetic fields generate in turn high frequency, annular oscillating electric currents in the conductor, which accelerate the

stripped electrons which then heat the gas conductor as a result of its Ohmic resistance thus forming an inductively coupled plasma.

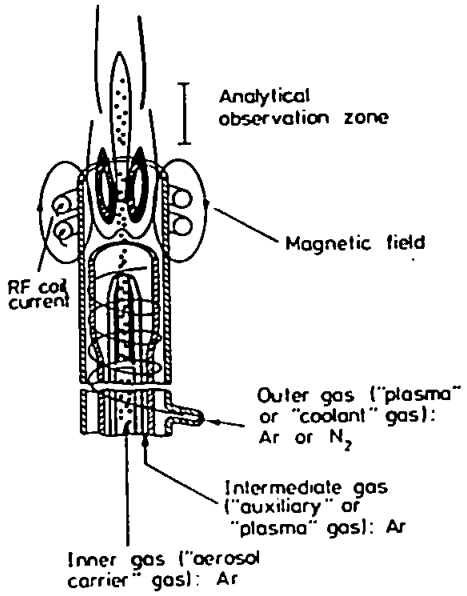


Figure 1.3 Schematic of an ICP Torch(1)

The torch consists of three concentric tubes designated the 'outer tube' (plasma gas), 'intermediate tube' (auxiliary gas), and ' inner tube' (nebuliser gas). Thermal isolation of the plasma is achieved by using a tangentially introduced argon gas flow, which also has the added benefit of stabilising the position of the plasma. The gas flowing through the inner tube, the carrier gas, is used to carry the nebulised sample to the plasma, whereas the auxiliary gas maintains the plasma at the appropriate height above the injector, preventing it from melting the delicate injector tip.

A major advantage of the ICP compared to other emission sources is its ability to efficiently vaporise, atomise, excite, and ionise a wide range of elements present in many different sample types. One of the important reasons for the superiority of the ICP over flames and furnaces is its high temperature in the region of 6800 K, compared to 3000 K

for a flame. The high temperature also eliminates many of the chemical interferences found in flames and furnaces, such as analyte depression caused by the formation of less volatile compounds which are difficult to dissociate or analyte depression due to occlusion into refractory compounds.

An important advantage of the ICP as an emission source arises from the torch design and the so-called skin effect. RF coupling occurs mostly in the outer part of the discharge, so sample aerosol can be directed into the centre of the plasma without significantly altering the energy-coupling process and hence the temperature. Thus, the plasma has a higher temperature in its outer region compared to its inner region, so the radiation emitted from analyte atoms in the centre does not pass through zones of lower temperature and is subjected to considerably less self-absorption effects compared with flame and furnace techniques. As a result of this ICP-AES is characterised by a large dynamic range, typically five orders of magnitude. Detection limits are generally very low for the majority of elements, falling within the range 1 – 100 ng ml^{-1} (23) depending on the analyte and the sample matrix.

When a sample droplet is introduced into the plasma the processes depicted in Fig. 1.4 take place. First, the aerosol is desolvated usually leaving the sample as microscopic salt particles. Next, decomposition of the salt particles into a gas of individual molecules (vaporisation) occurs, which are then dissociated into atoms (atomisation). These processes occur predominantly in the preheating zone (PHZ). The atoms are then excited and/or ionised in the plasma predominantly in the initial radiation zone (IRZ) and the normal analytical zone (NAZ). The NAZ is the region of the plasma from which analyte emission is typically measured.

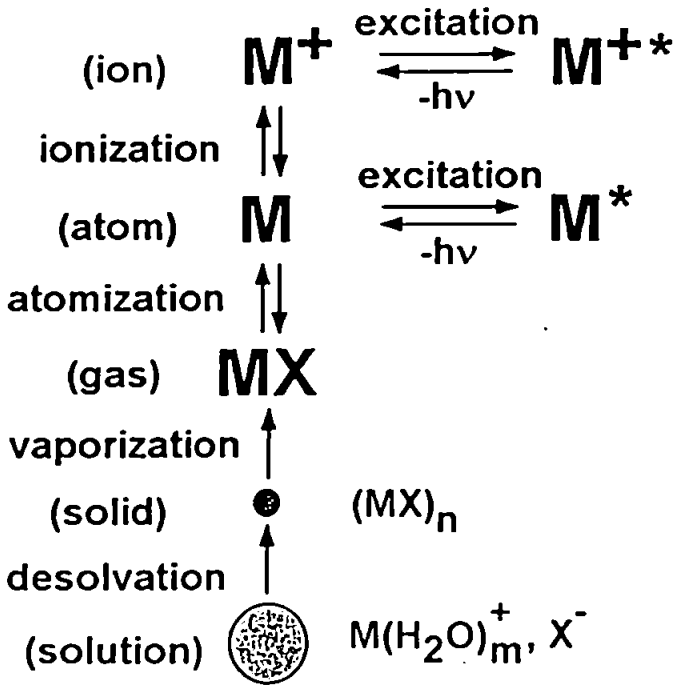


Figure 1.4 Processes that take place when a sample droplet is introduced into an ICP discharge (24)

1.2.2.3 Spectrometer

The physical separation of the emission radiation into its constituent wavelengths is carried out by the spectrometer. The monochromator is an instrument that can isolate a narrow range bandwidth (e.g. 0.01-0.001 nm)(1) anywhere in the spectrum (for atomic emission spectrometry typically 190-750 nm). The commonest configuration for an ICP-AES is for the monochromator and detector to view the plasma-side on as shown in Fig. 1.5. There is, therefore, an optimum viewing height in the plasma (dependent on the vertical plasma temperature profile) which gives the maximum signal intensity, lowest background and least interferences.

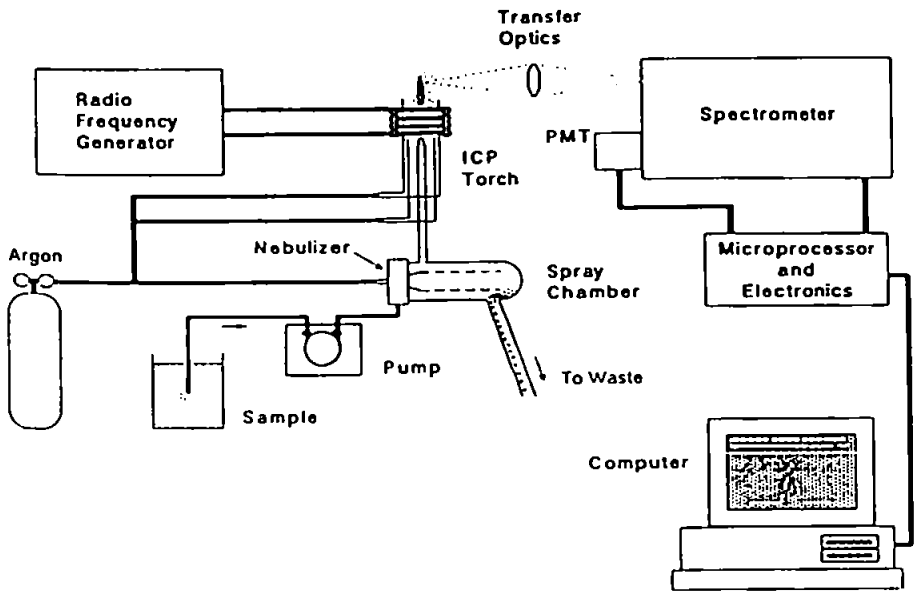


Fig. 1.5 Typical configuration for ICP-AES instruments: side on viewing of the ICP.

The dispersion of the different wavelengths by a diffraction grating is by far the most common method. Since, up to the early 1990's most ICP-AES analyses were performed using diffraction grating-based devices, the following discussion will first focus on the more conventional instruments based on their use. However, there is an additional wavelength dispersive device, called an echelle grating, that has become an important component in ICP echelle-based instruments of the 1990's. Such instruments use combinations of 'conventional' and echelle gratings, and will therefore be discussed separately.

Atomic emission from the plasma can be focused onto the entrance slit of a monochromater using a combination of convex or plano-convex lenses or a concave mirror. Two common lens arrangements are the Czerny-Turner or Ebert mountings (Fig.1.6). Most modern instruments use a diffraction grating with between 600 to 4200 lines per millimetre etched

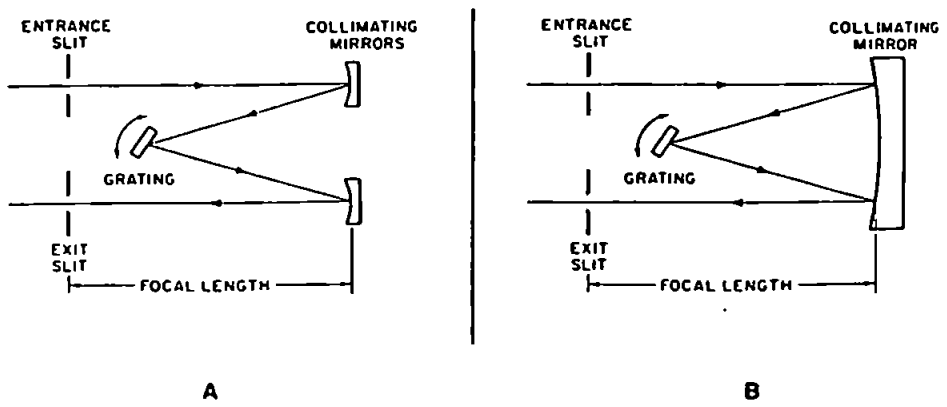


Figure 1.6 Czerny-Turner (a) and Ebert (b) monochromator mounts.

into it. Light striking the surface of the grating will be diffracted at an angle that is dependent on the wavelength of the light and the line density of the grating hence a spatial separation of wavelengths is achieved. In general, the longer the wavelength and the higher the line density, the higher the angle of diffraction will be. Figure 1.7 shows schematically the paths that light rays of two different wavelengths would take when diffracted from a grating.

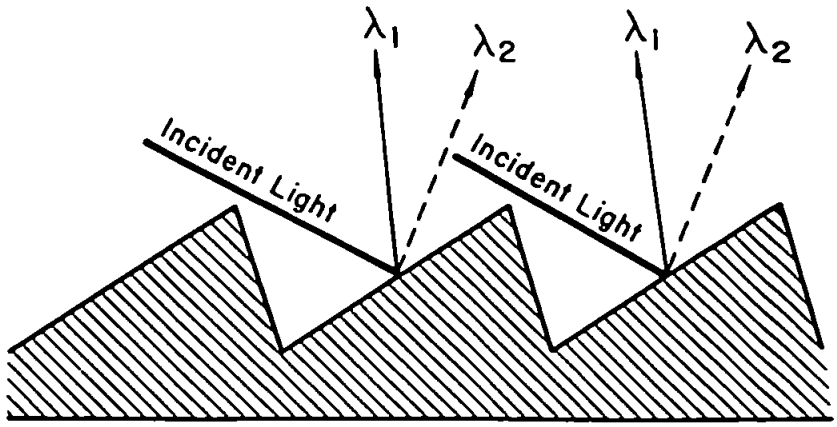


Figure 1.7 Diffraction grating separating two wavelengths of light.

To separate polychromatic light, the grating is incorporated in a spectrometer. The function of the spectrometer is to form the light into a well-defined beam, disperse it according to wavelength with a grating, and focus the dispersed light onto an exit plane or circle. One or more exit slits on the exit plane or circle are then used to allow certain wavelengths to pass to the detector while blocking out others. The diffracted light is collimated and focused, using a mirror, or combination of mirrors (Fig. 1.6) onto an exit slit, and the diffraction grating can be rotated so that different wavelengths are focused on to the exit slit in turn, giving rapid sequential multielement analysis. A prism can also be used to disperse polychromatic radiation into its characteristic wavelengths.

In the 1970's it was shown that certain advantages might be obtained by combining the characteristics of two dispersing systems such as the diffraction grating and the prism or two diffraction gratings. The two optical components are positioned perpendicular to each other. One of the dispersive devices is in general, a coarsely-lined echelle grating which separates the polychromatic radiation by wavelength and produces multiple,

overlapping spectral orders. The second dispersing device, either a grating with a ruling density of greater than 350 lines/mm or a prism, separates, or cross disperses, the overlapping orders into a two dimensional pattern called an echellogram. A typical optical configuration for this echelle-type of spectrometer is illustrated in Fig. 1.8. A two dimensional display results in the exit plane, wavelength in one direction and spectral order in the other as shown in Fig. 1.9.

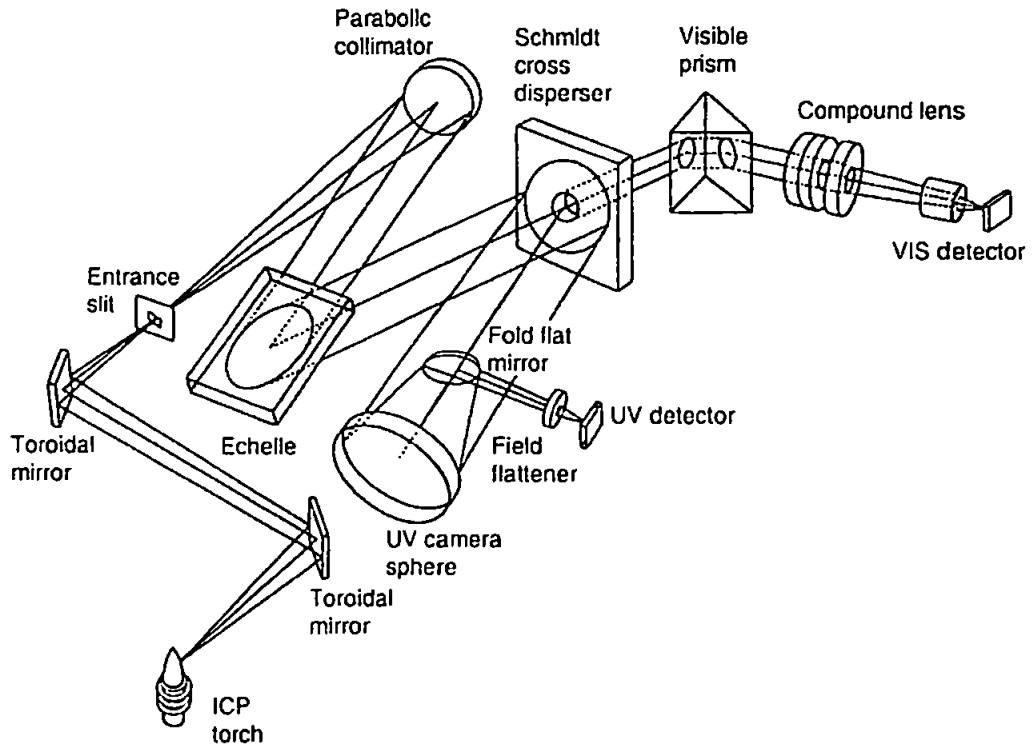


Figure 1.8 An echelle optical mount (24)

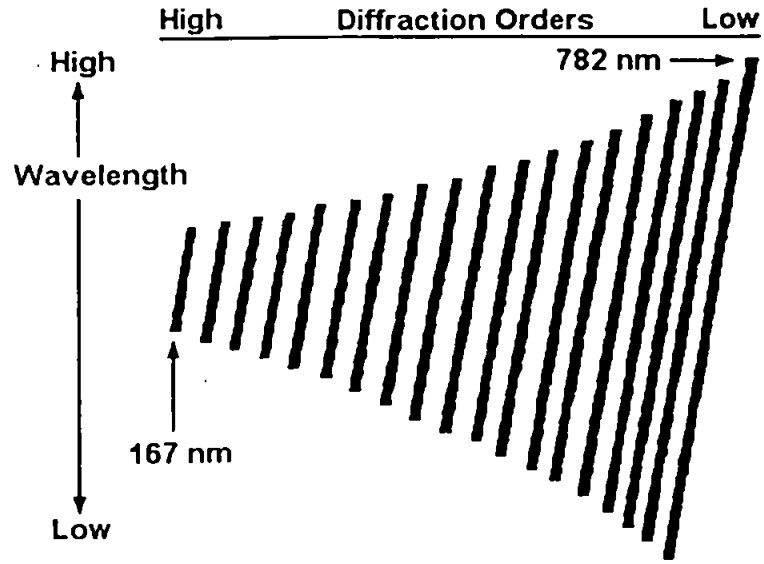


Figure 1.9 Exit plane illustrating the two-dimensional array produced by the echelle mount (24).

Echelle grating-based spectrometers offer some advantages over conventional spectrometers using non-echelle dispersive systems. First, the optics result in high efficiency in each of the spectral orders, whereas conventional diffraction gratings are generally optimised at a particular wavelength, called the blaze wavelength, and for a particular order, usually the first order. Second, the system has excellent resolution since it is generally used in the higher spectral orders (resolution enhancements are exhibited with increasing order). The two-dimensional echelle pattern resulting from the optical design of the echelle spectrometer lends itself well to the use of solid state detectors.

Three types of advanced solid-state detectors, with high sensitivity and resolution, have been developed for spectroscopic applications namely, the photodiode array (PDA), the charge injection device (CID) and the charge-coupled device (CCD). The CID and CCD belong to the broad class of silicon-based devices called charge transfer devices

(CTD). CTDs are two-dimensional arrays of photosensitive metal oxide semiconductor capacitors arranged within a single, solid-state integrated circuit, and are capable of collecting and quantifying photo-generated electrical charge (25).

In the case of the CID, photons striking the surface of each detector element (pixel) generate a proportional degree of electrical charge, which is shifted between two electrodes within the pixel (*i.e.* intra-cell transfer). Voltage fluctuations resulting from this transfer are detected and are proportional to the intensity of light striking the pixel. However, the CID system has a higher dark current than the CCD system, and requires cooling to liquid nitrogen temperatures to decrease the noise. In a CCD detector, photo-generated charge accumulated within each pixel is transferred to a serial register, and then to a charge-sensing output amplifier (*i.e.* inter-cell transfer) (25, 26). CIDs have been used for wide dynamic range imaging applications (*e.g.* atomic spectroscopy), whereas CCDs are preferred for low intensity spectroscopic (e.g. Raman and fluorescence spectroscopy) imaging owing to their superior signal-noise ratio. A pixel is typically ~ 22 x 22 μm in size, while CCD arrays are generally arranged in a two-dimensional silicon wafer configuration of 578 x 385, 512 x 512, 1152 x 298, or 4096 x 4096 pixels.

Recently, a new kind of detector has been introduced, a Segmented-array-Charge-Coupled device detector (SCD) (27) for an echelle ICP instrument. Instead of using a large CCD with hundreds of thousands of continuous pixels, the SCD has been designed with subarrays located at specific *x-y* locations corresponding to the 236 or more of the most important ICP spectral lines of the 70 elements observed in ICP spectrometry. Each subarray contains between 20 to 80 pixels. Each pixel is 12.5 μm wide and 80 to 170 μm in height, the height depends on the wavelength and spectral order of the line. Each subarray has its own interface logic which permits individual subarrays to be addressed and read-out. Most commercial CIDs and CCDs have poor sensitivity below 350 nm because of photon absorption by electrodes on the surface of the device. Because the pixels of the

individual subarrays of the SCD have no embedded electrodes, the SCD has excellent photosensitivity from 160 to 782 nm (24).

1.2.2.4 Performance Characteristics of ICP-AES

Many wavelengths of varying sensitivity are available for each element, making ICP-AES suitable for analysis over a wide concentration range, from trace levels to major components. Although precision is concentration dependent, relative standard deviations of 0.5 – 2% are common. When coupled to a suitable detection device, such as the Segmented Charge Coupled Detector (SCCD), the simultaneous determination of a large number of elements can be accomplished.

1.2.3 Types of Interferences, Conventional Correction and their Limitations

Interferences can be broadly divided into two groups: spectroscopic and non-spectroscopic.

1.2.3.1 Spectroscopic Interferences

Spectroscopic interferences can be caused by: (i) line broadening, (ii) shifting baseline and line coincidences (iii) structured molecular band emission (iv) stray light, (v) plasma continuum, and (vi) radiative recombination continuum. Line coincidences, stray light, line broadening, and recombination continuum emission are the most troublesome and, as they arise from elements in the sample, their type and severity depend on the sample composition (28).

An example of line broadening interference is shown in Fig 1.10 for Ca II 393.4 and 396.8 nm (1000μg ml^{-1}) on a solution of 1 μg ml^{-1} Al. A substantial background enhancement is observed which seriously interferes with the determination of low concentrations of Al in the presence of Ca using the most prominent Al line.

Interference caused by a shifting baseline or line coincidences can be categorised as: simple ("flat") background; sloping background; direct line overlap or complex line overlap, all of which are shown schematically in Fig 1.11a. The complexity of the interference increasing from (1) to (4).

Interferences caused by structured molecular band emission vary with the ICP operating conditions particularly observation height (1). Scans such as reproduced in Fig. 1.12 give an excellent overall impression of the wavelength regions where molecular bands (e.g. OH, C_2 Mulliken, CN violet and C_2 Swan system) dominate the spectrum and where interferences can be expected. The elements present in the sample may also contribute molecular bands. For example, some elements may form stable diatomic oxide radicals with the oxygen originating from the dissociation of water or entrained from the surrounding air, examples of which are: YO, AlO, BaO and MgO.

Fig. 1.13 illustrates the background continuum spectrum for (i) a solution of deionised water and (ii) a 5000μg ml^{-1} solution of Mg. The rise and levelling off of the continuum spectrum is clearly visible at approximately 300 nm and contributes significantly to the Mg signal. Each class of interference presents problems in calibration, the extent of which depends on the calibration method used.

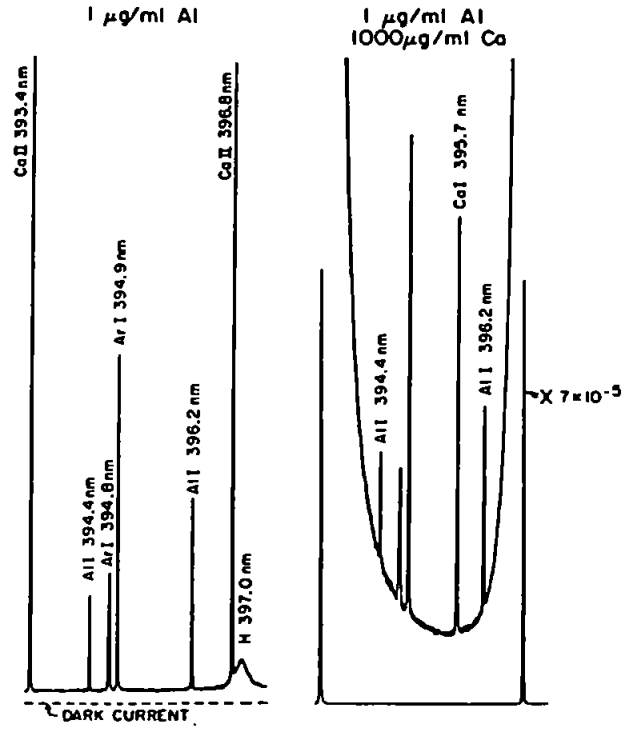


Figure 1.10 Background enhancement produced by broadening of Ca II 393.4 and 396.8 nm lines (29).

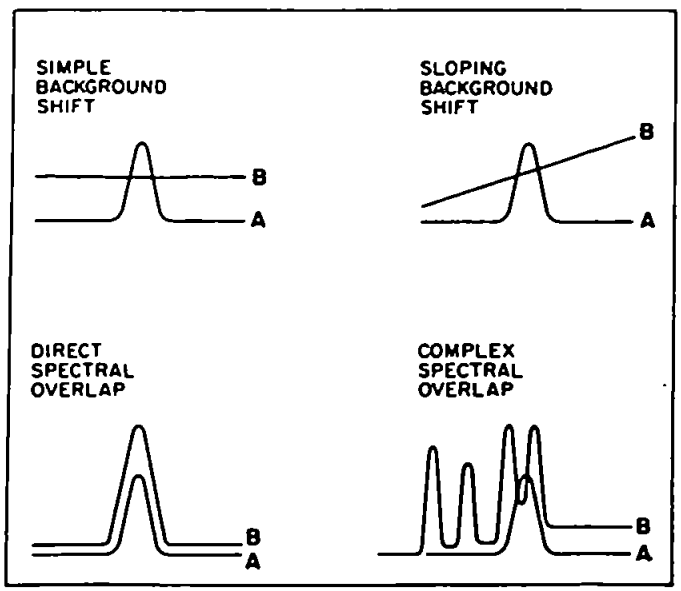


Figure 1.11a Schematic representation of the basic types of background Interference (1)

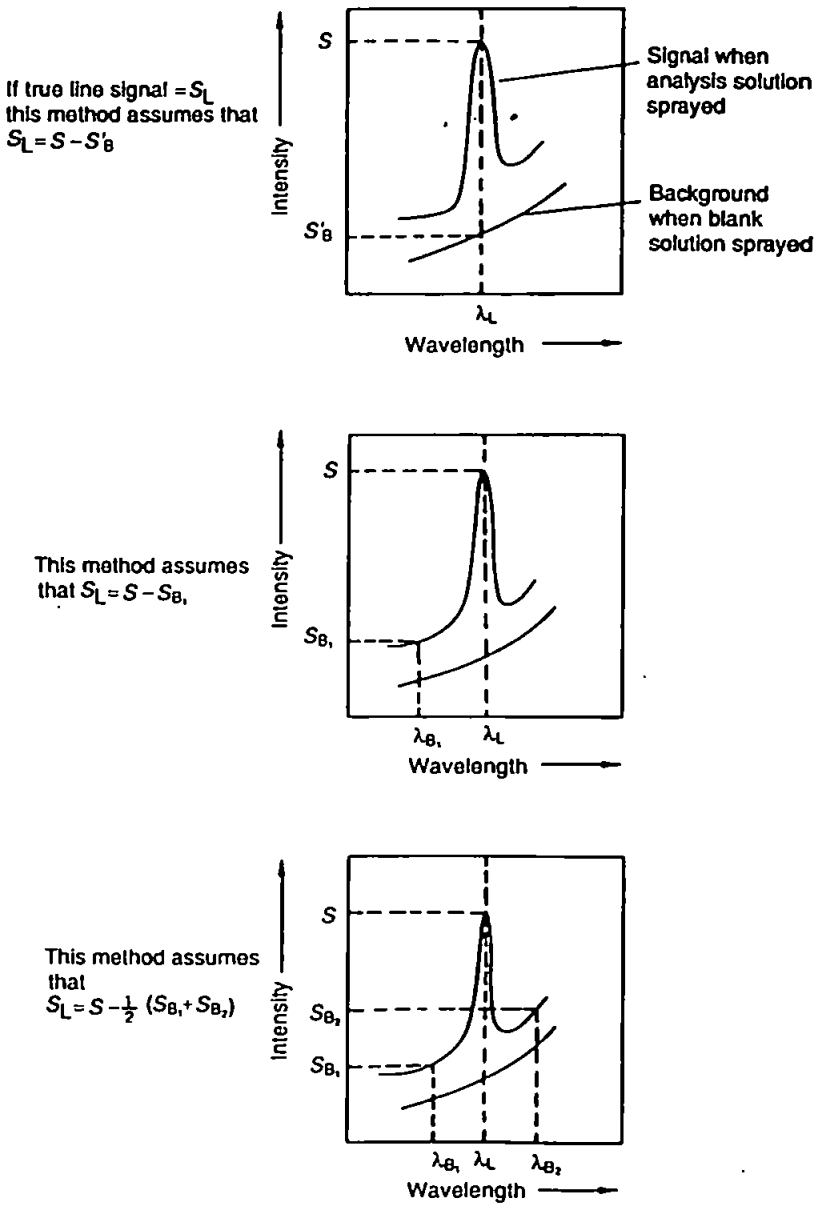


Figure 1.11b Background correction methods used in atomic emission spectrometry (23)

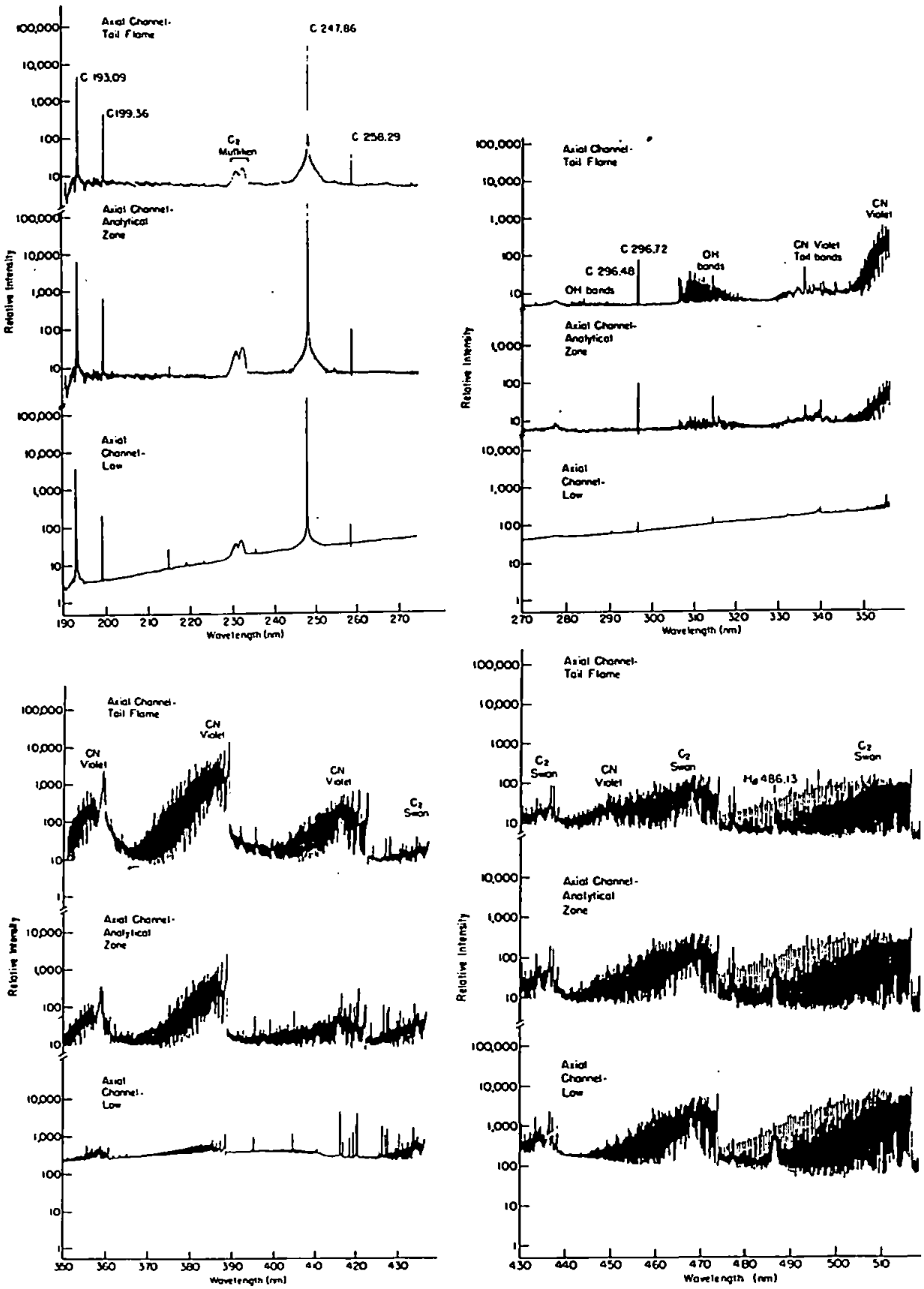


Figure 1.12 190 –520 nm (oil and MIBK (1:10 m/v)) for three observation heights in the centre of an argon axial ICP channel (30)

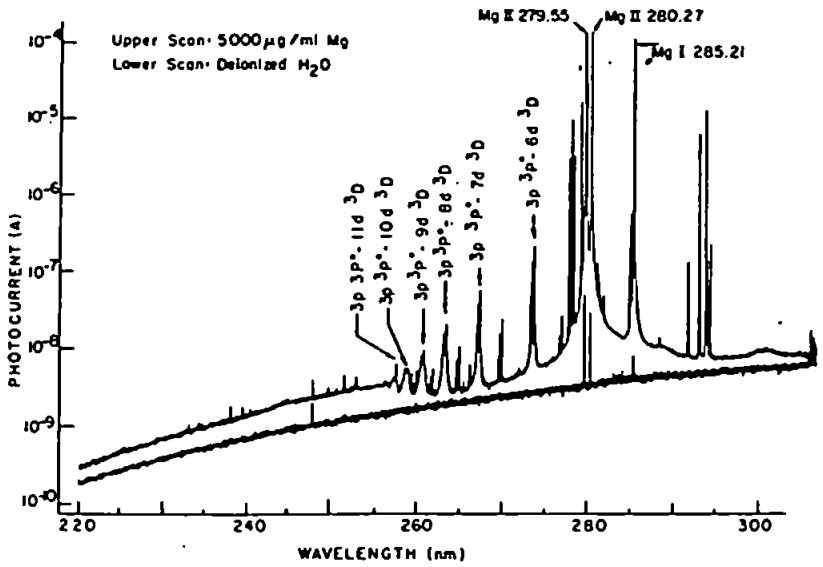


Figure 1.13 Wavelength scan for a Mg solution aspirated into the ICP(31)

Many attempts have been made to circumvent interferences in ICP-AES. The most common the optimisation of line selection with ICP wavelength tables (32). Tables complied by the Massachusetts Institute of Technology (MIT)(33, 34) and those by Meggers, Corliss and Scribner (35) are two of the most commonly used reference sources, but they cannot predict all possible spectral interferences in ICP-AES since the properties of the inductively coupled plasma are quite different from those of the DC arc and spark for which they were compiled. Wohlers (36) has compiled a set of tables for ICP-AES, but they are not comprehensive enough to predict all interferences. Detailed tables of spectral interferences in geological analysis were presented by Church (37). The most widely used spectral interference tables used however are: *An Atlas of Spectral Interferences in ICP Spectroscopy*, by Parsons, Foster and Anderson (38) and *Bouman's Line Coincidence Tables for ICP-AES*, by Boumans (39). The use of tables is limited in that satisfactory appraisal of interferences requires comprehensive ICP tables. Even if such tables were

available the usefulness of the information would be dependent upon the experimental conditions such as the relative concentrations of the analyte to the interferents and the instrumental parameters. In addition, selection of the 'best' analyte lines requires an *a priori* knowledge of the sample composition (especially all possible interferents), which is often impossible or impractical.

Traditionally, interference correction methods for spectroscopic interference have taken the form of background correction or the use of interelement correction factors (40). The use of background correction is limited, however, to relatively simple situations, such as an increase in the continuum background emission which can be easily compensated for by subtraction of the background adjacent to the analytical line. For a sloping background measurements must be made on both sides of the line and usually the mean value is subtracted. These options are summarised in Fig. 1.11b. When complete or complex spectral overlap occurs (Fig. 1.11) background correction is not possible.

1.2.3.1.1 Interelement Correction Factors

The determination of empirical spectral interelement correction factors (IECs) for a given ICP-AES instrument is, in principle, straightforward. A calibration graph for the analyte of interest is first constructed in the usual way, and solutions of the suspected interferents (1000 $\mu g\ ml^{-1}$) are aspirated while monitoring the analyte line(s) of interest. The apparent analyte concentration is then determined at the line of interest, and an IEC factor calculated by dividing the apparent analyte concentration at that particular wavelength by the interferent concentration. Eqn. 1.1 is then applied to obtain the corrected analyte concentration.

$$C_c = C_u - \sum_{i=1}^{i} (F_i x I_{ic}) \qquad \text{Eqn. 1.1}$$

Where C_u, C_c, F_i and I_{ic} are uncorrected analyte concentration, corrected analyte concentration, IEC factor, and interferent concentration respectively for each interfering element *i*. There are, however, many practical difficulties to overcome in order to obtain reliable correction factors.

In practise truly pure single element solutions of interfering elements are impossible to obtain, especially at the high concentrations likely to cause interference. Trace quantities of elements, other than the element required, are often present in the source material or in the solvent used to dissolve the material. If these impurities are not identified then correction factors become inaccurate. In addition the procedure requires proper continuum background correction in the form of background correction points. Hence, complex variable matrices make this method very difficult to apply successfully.

Eqn. 1.1 implies that the concentration of the interferent must be known; to achieve this the concentration of all the interferents must be determined separately. The assumption made here is that the signal output from the interferent element in question is itself not suffering from spectral or other interferences. In complex mixtures this assumption may not be true, even when alternative lines are used. To overcome this one may consult spectral tables, but because of the problems already mentioned, this is not a satisfactory approach to the problem.

1.2.3.1.2 Matrix Matching

An alternative to using an internal standard is to matrix match calibration standards to the samples being analysed. This is straightforward if all the elements present in the

sample are known. However, if there are unidentified elements present this method may give poor predictive results, depending on the severity of the interference.

1.2.3.1.3 Standard Additions

The method of standard additions is commonplace in spectroscopy, however, there are major drawbacks to this method. The method requires at least three, preferably five, aliquots for each sample being analysed (if large numbers of samples are to be analysed this method can easily become burdensome and expensive in terms of instrumental time). The additions themselves must be small in comparison to the sample volume to prevent dilution effects which would render the standard addition process invalid. The technique of backward linear extrapolation itself leads to errors and sample aliquots are required for the determination of each analyte (i.e. the method is univariate). The method cannot correct for spectral interferences and background changes (emission continuum) which gives it a limited value when analysing complex samples.

1.2.3.2 Non-Spectroscopic Interferences and Internal Standardisation

Non-spectroscopic interferences give rise to enhancement or suppression of the analyte signal as a result of perturbations in sample nebulisation efficiency and sample transport (resulting from changes in the sample matrix), by physical effects such as alterations in the electron density in the plasma itself often referred to as matrix-induced suppression or enhancement, or by chemical effects such as the formation of compounds containing the analyte which possess much higher ionisation energies.

The use of internal standards (41, 42) can compensate for such effects, however, the internal standard must be chosen with care so that it matches the analyte chemically and

spectroscopically (i.e. atomic lines with similar excitation energies, partition functions and transition probabilities must be used). Myers and Tracy (42) have demonstrated that a judicious application of internal standards in ICP-AES can improve the % RSD by a factor of 10. These significant improvements in precision were only obtained by using those internal standards with similar excitation characteristics as the analyte lines of interest. For internal standards and analyte lines with much different excitation characteristics such improvements were not found. The use of internal standardisation is, however, reliant upon interferent-free lines being available. It is highly probable that for complex samples such lines will be unavailable.

Ramsey and Thompson (43) devised the parameter-related internal standard method (PRISM) to overcome the limitations of the internal standard method in ICP-AES by obviating the requirement for the matching of analyte and internal standard emission lines. However, successful use of their method still requires the availability of interferent-free internal standard lines.

1.3 Univariate Calibration

Calibration is the process of determining a mathematical model to relate measured instrumental response (dependent variable, e.g. emission) to a known parameter (independent variable, e.g. concentration) of a sample analyte, and using this model to predict the same parameter in unknown samples. Typically, response measurements are obtained for a series of samples, usually greater than five, in which accurate analyte concentration values have been determined independently. These calibration standards are measured by the analytical instrument under the same conditions as those subsequently used for the unknown samples. Once the calibration curve (mathematical model) has been

established the analyte concentrations in any unknown sample can be obtained by interpolation (44). The simplest form of calibration model is the linear univariate calibration model, in which a single instrumental measurement is used to determine the concentration of a single analyte. However, with the development of instrumentation capable of rapidly obtaining multiple response data (e.g. full spectrum emission measurements), it has become desirable to adopt calibration techniques which can fully utilise the available multivariate data. The relative merits of univariate and multivariate calibration are discussed below.

1.3.1 Univariate Classical Least Squares

One of the most commonly applied univariate calibration procedures in analytical chemistry is the 'classical' probabilistic model (Fig. 1.14) which assumes a linear relationship between instrumental measurement and concentration. The simplest from of this model is the first-order (straight-line model) (45).

$$y_i = \beta_i x_i + e_i \qquad \text{Eqn. 1.2}$$

where y_i and x_i are the instrumental response (dependent variable) and analyte concentration (independent variable) respectively, β_i is the calibration coefficient determined by least-squares regression of instrumental response on analyte concentration and e_i is the random error component.

The principal advantages of univariate calibration techniques are their simplicity of application and ability to produce accurate calibration models using a relatively small number of calibration samples. However, in order to obtain accurate predictions with the

univariate approach instrumental measurements must be highly selective (no interferences affecting instrumental response) with respect to the analyte of interest. This can only be achieved if the sample matrix is of low complexity. If these requirements are not met, then predictions of analyte concentrations in unknown samples are likely to be inaccurate.

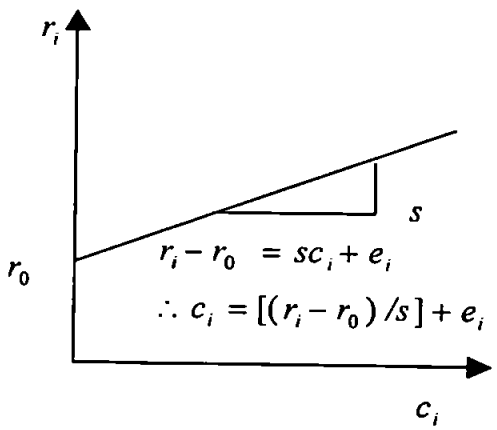


Figure 1.14 The classical straight line model.

The removal of such interferences by chemical or physical means may alleviate this problem, but this is not always possible. A further limitation is that univariate calibration has no fault-detection capabilities, as illustrated in Fig.1.15 where the true relationship between the instrument response and concentration is represented by the solid line. It follows that a sample concentration of c_{true} corresponds to an instrument response of r_{true}. Errors occur if an unknown interferent is present and the instrument has a significant sensitivity to it. In Fig. 1.15 assuming $r_{true+interference}$ is the response of the instrument to a sample with concentration c_{true} plus some additional response due to an interferent. Using the calibration curve yields a predicted concentration, $\hat{c}$, for the analyte. The presence of the interferent has resulted in the overestimation of the concentration of the analyte. In

particular, industrial processes can require the analysis of extremely complicated samples and it is often impossible to obtain highly selective measurements or to separate the analyte of interest from all potential interferences. Univariate calibration is generally inappropriate under these circumstances, and for this reason multivariate calibration methods are becoming increasingly widespread.

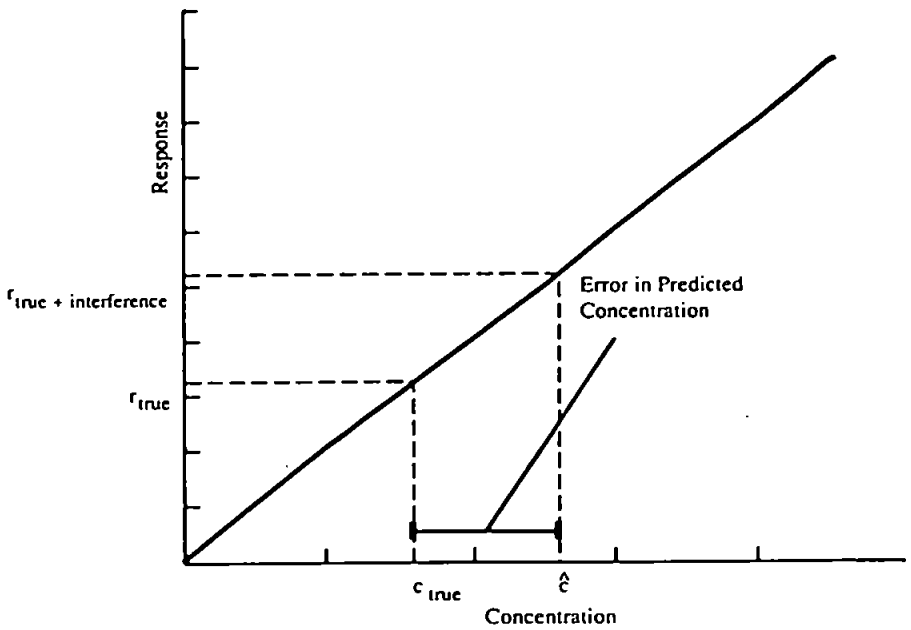


Figure 1.15 Illustration of the inability of univariate methods to detect the presence of interferents

1.4 Multivariate Calibration

In all cases, the following format will be used for algorithmic expression:

MATRICES written in bold uppercase;

Vectors written in bold lowercase (all vectors are *column* vectors, and all transposed vectors are *row* vectors);

scalars written in italics (lowercase)

The algebraic notation used throughout this section is defined in Table 1.1.

Multivariate and univariate calibration are similar insofar as they both involve the construction of a mathematical calibration model relating instrumental response to analyte concentration for a set of known calibration standards, and the use of this model to predict analyte concentrations in unknown samples. However, multivariate calibration incorporates multiple instrumental measurements of each sample (e.g. the spectra obtained by multiwavelength spectrometers) into the calibration model.

Multivariate calibration makes it possible to determine multiple components simultaneously. This is not possible with univariate analysis and can reduce the amount of time spent on method development. Multiple measurements can also provide improved prediction precision. Statistics show that repeating a measurement n times and calculating a mean value will give a factor of reduction of $\sqrt{n}$ in the standard deviation of the mean, commonly termed signal averaging.

It is also possible to account for sources of system variation without isolating and characterising their source using multivariate calibration. Hence there is no need for time consuming separation processes to remove interferences, the only requirement being that sufficient variation is incorporated in the calibration phase. This capability for implicit modelling gives this multivariate methods powerful advantages over the univariate approach. It is also possible to apply diagnostics to assess model confidence and the reliability of the predicted values. These diagnostics can also be used to investigate the nature of the problem when a particular sample is "flagged" as being unusual (46). The least complex, most widely available, and therefore most frequently applied multivariate techniques are those that assume a linear relationship between response signal and analyte concentration. These include classical least squares (CLS), direct and indirect, principal components regression (PCR) and partial least squares (PLS). The last two of which are collectively known as inverse least squares modelling techniques.

Table 1.1 Definition of notation used for multivariate calibration algorithms.

C or Y	Component concentrations matrix (*i.e.* analyte concentrations for all samples in the calibration data set
R or X	Instrumental response matrix(*i.e.* emission spectra for all samples in the calibration set)
$\mathbf{c}_i$	Component concentrations vector for sample *i*
$\mathbf{r}_i$	Instrumental response vector (i.e. spectrum) for sample *i*
$\mathbf{T}(t_i)$	Principal components scores matrix (vector)
$\mathbf{P}(p_i)$	Principal components loadings matrix (vector)
i	Total number of samples
j	Total number of variables / wavelengths
h	Total number of principal components
β_j or (b_j)	Regression coefficient (relating concentration to instrumental response)
Q	Matrix of regression coefficients (relating **T** to **C**)
w_h	Partial least squares weight vector for PC *h*
$\mathbf{E}_i$	Spectral residuals matrix
^	Estimated parameter
S	Sensitivity matrix ($\mathbf{R}/\mathbf{C}$)
T	Transpose of a matrix or vector

1.4.1 Classical Least Squares

Both direct and indirect CLS assume linear additivity of the chemical components at each wavelength. This holds in situations where there are no chemical or physical interactions, so can only be used for simple chemical systems. For wavelength *j*, linear additivity is expressed mathematically by Eqn. 1.3

$$r_{ABj} = r_{Aj} + r_{Bj} = c_A s_{Aj} + c_B s_{Bj} \qquad \text{Eqn. 1.3}$$

where r_{ABj} is the instrument response to a mixture of analytes A and B, r_{Aj} and r_{Bj} are the spectral responses of the instrument to analyte A and B respectively for *j* wavelengths, c_A and c_B are the concentration of analyte A and B respectively and s_{Aj} and s_{Bj} are the sensitivities of analyte A and B respectively for *j* wavelengths. This equation can be written as Eqn. 1.4

$$\left[r_{AB1}, r_{AB2} \ldots\ldots r_{ABj}\right] = \left[c_A, c_B\right] \times \begin{bmatrix} s_{A1}, s_{A2} \ldots\ldots s_{Aj} \\ s_{B1}, s_{B2} \ldots\ldots s_{Bj} \end{bmatrix} \qquad \text{Eqn. 1.4}$$

and in matrix notation as $\mathbf{r} = \mathbf{c} \times \mathbf{S}$.

To construct a CLS model, pure spectra for all the analytes are obtained to form the **S** matrix. Two CLS methods are available, direct, where pure spectra are measured directly, and indirect, where pure spectra are computed from spectra of mixtures with known composition.

To perform prediction, an unknown sample spectrum is measured (**r**). Given **r** and **S**, it is possible to solve Eqn. 1.4 for **c**:

$$\mathbf{r} = \mathbf{cS} \qquad \text{Eqn. 1.5}$$

$$\mathbf{rS}^T = \mathbf{cSS}^T \qquad \text{Eqn. 1.6}$$

$$\mathbf{rS}^T(\mathbf{SS}^T)^{-1} = \mathbf{c}(\mathbf{SS}^T)(\mathbf{SS}^T)^{-1} \qquad \text{Eqn. 1.7}$$

$$\mathbf{rS}^T(\mathbf{SS}^T)^{-1} = \hat{\mathbf{c}} \qquad \text{Eqn. 1.8}$$

By letting $\mathbf{S}^+ = \mathbf{S}^T(\mathbf{SS}^T)^{-1}$, the pseudo-inverse of $\boldsymbol{S}$:

$$\hat{\mathbf{c}} = \mathbf{rS}^+ \qquad \text{Eqn. 1.9}$$

where $\hat{c}$ contains the predicted concentrations.

These classical methods have the advantage of being full spectrum calibration techniques, which offer greater precision than models limited to a smaller number of response variables because of their signal averaging capabilities (47) However, for classical methods to work well pure component spectra of all the analytes and interferents in unknown samples must be obtained and explicitly incorporated into the $\boldsymbol{S}$ matrix, and there must be a linear relationship with concentration (i.e. linear additivity of pure component spectra must be assumed)

This is a significant limitation of the technique, since it is seldom possible to provide the model with information for all the elements within a complex sample matrix, and any unmodelled spectral interferences will produce large residual errors in future sample prediction. These errors can be minimised by selecting those spectral regions in which unknown components do not significantly interfere with the response of the analytes of interest, although in many cases the entire spectral range can be subject to interference effects.

1.4.2 Inverse Least Squares (ILS)

The common feature of inverse methods is how the relationship between response and concentration is modelled. The concentrations are treated as a function of the response, as shown in Eqn. 1.10

$$\mathbf{c} = \mathbf{R}\hat{\mathbf{b}} \qquad \text{Eqn 1.10}$$

where $\mathbf{c}$ contains the concentrations, $\mathbf{R}$ is the matrix of measurements (emission), and the vector $\hat{\mathbf{b}}$ contains the model regression coefficients. From a comparison of the inverse model (Eqn. 1.10) to the model for the classical method ($\mathbf{c} = \mathbf{rS}$) it is not obvious that the approaches are significantly different. To illustrate the difference, the matrix algebra for Eqn. 1.10 is given for one sample (Eqn. 1.11):

$$[c_1] = [r_{1,1}, r_{1,2}, r_{1,3}, r_{1,4}] \times \begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \\ \hat{b}_3 \\ \hat{b}_4 \end{bmatrix} \qquad \text{Eqn. 1.11}$$

Equation. 1.11 shows that the concentration value (c_1) is equal to the weighted sum of the responses for different variables (i.e. $r_{1,1} \times \hat{b}_1 + r_{1,2} \times \hat{b}_2 + r_{1,3} \times \hat{b}_3 + r_{1,4} \times \hat{b}_4$). However, the classical approach fits a combination of linear pure spectra to an unknown spectrum (Eqn's. 1.5-1.9).

This major difference in the modelling approaches gives the inverse methods significant advantages over classical methods. Classical approaches require the explicit inclusion of sources of variation. With the ILS methods, it is possible to predict unknown sample analyte concentration even when additional chemical and physical sources of

variation are present provided that all significant sources of variation are present when the calibration models are estimated. Additional requirements are that the analytical instrument adequately differentiates between the component of interest from other sources of variation and that the instrumental response is sufficiently linear with concentration.

Although it is important to ensure that all significant sources of variation are present when the calibration models are estimated, these other sources of variance are not included as extra variables, as is the case with CLS methods, but are implicitly modelled. Any source of variation that does not change during the calibration phase will not implicitly be included in the model. One approach, therefore, is to manipulate all the sources of variation using an experimental design. If designed experiments are not possible, for example when complex matrices are involved, another approach is the use of historical data.

The prediction of future unknown sample concentrations is achieved by obtaining a new response matrix, $\mathbf{r}_{unk}$. Using this response matrix and the known concentrations of only one of the components in $\mathbf{c}$, the regression coefficients in Eqn. 1.10 can be estimated as in Eqn. 1.11:

$$\hat{\mathbf{b}} = (\mathbf{R}^T\mathbf{R})^{-1}\mathbf{R}^T\mathbf{c} \qquad \text{Eqn. 1.11}$$

where $\hat{\mathbf{b}}$ contains the estimated regression coefficients, and where $(\mathbf{R}^T\mathbf{R})^{-1}\mathbf{R}^T$ is known as the pseudo-inverse of $\mathbf{R}$. Given the spectrum of a single unknown sample ($\mathbf{r}_{unk}$) it is now possible to use the estimated regression vector $\hat{\mathbf{b}}$ to predict the concentration of the components of interest according to Eqn. 1.12:

$$\hat{\mathbf{c}} = \mathbf{r}_{unk}\hat{\mathbf{b}} \qquad \text{Eqn. 1.12}$$

In Eqn. 1.11, the key to the model-building step is the inversion of the matrix ($\mathbf{R}^T\mathbf{R}$). This is a square matrix with row (calibration samples) and column (variables (nvars)) numbers equal. From theory, a number of independent samples in the calibration set greater than or equal to nvars is needed to invert this matrix. For most analytical systems, nvars is much greater than the number of samples, and therefore $\mathbf{R}^T\mathbf{R}$ cannot be inverted. However, by manipulating the variables, calculation of the pseudo-inverse is possible. How this manipulation is accomplished distinguishes between the different inverse modelling methods.

1.4.2.1 Multiple Linear Regression

Multiple linear regression (MLR) with variable selection makes the matrix inversion by selecting a subset of the original variables. Multiple linear regression is a relatively simple multivariate technique, and is therefore limited to simple systems (i.e. small numbers of components and other sources of variation) (46). It is important that the variables in the $\mathbf{R}$ matrix are informative, i.e. they are correlated to the analyte of interest, since MLR will attempt to model all the data present in $\mathbf{R}$, including any irrelevant information.

Collinearity in the response data can pose a problem for MLR, particularly in the calibration of multiwavelength spectroscopic data. A data set is collinear if at least one variable is an exact or approximate linear combination of the others (i.e. a linear or near-linear relationship exists between data points). Such collinearity will produce instability (large variance in some elements of $\hat{\mathbf{b}}$) in the $\hat{\mathbf{b}}$ vector during the inversion of $\mathbf{R}$. The collinearity and noise problem can be overcome by selecting a suitable subset of the response data. Such variable subsets can be statistically determined by a number of techniques, an example of which is stepwise multiple linear regression (SMLR) (48).

SMLR can be performed as either forward selection, which begins with one variable and progressively incorporates more variables into the model until a certain criterion is met, or backwards elimination, which starts with the full spectrum and deletes one variable from the model at each step until the predefined criterion is achieved. The stopping criterion is typically an *F*- statistic, which tests the significance of the regression coefficients for each variable by examining model error. In forward selection, the variable with the most significant coefficient at each step is retained, and this continues until no added variable is significant. In backwards elimination, the variable with the lowest *F*-statistic at each step is removed until the point is reached when all the remaining variables are significant. In this way SMLR can circumvent the problem of collinear data, though, the signal-averaging capabilities of full-spectrum techniques are reduced.

A major limitation of SMLR is the constraint on the number of permitted variables. As an illustration, if a data set has, for example, n calibration samples and p variables, the maximum number of variables (v) allowed in either the forward of backward SMLR model must be $v \leq p$. Therefore, although a variable may have a significant *F*–statistic and may reduce predictive error, if the maximum number of variables has been reached it cannot be included as it may cause unstable coefficients in the $\hat{\mathbf{b}}$ vector during the inversion of $\mathbf{R}$. Spectroscopic instrumentation typically produces thousands of variables so if the number of calibration samples is, for example, 50 then $v \leq 50$. Such a reduction in the number of variables may not give optimum predictive results. The minimum number of variables must also be greater or equal to the number of chemical components in the system (46). For complex samples the number of variables required for an adequate predictive model may exceed v. The selection of a variable subset that produces optimum predictive error is therefore not theoretically possible within such constraints. Additionally the calibration sample to variable ratio must be of the order of 40 to 1(49), which is clearly impracticable for modern instruments capable of delivering thousands of variables.

1.4.2.2 Principal Components Analysis and Principal Components Regression

The most popular data compression method used to describe variation in large data sets in chemistry is principal components analysis (PCA) (50, 51).

Principal components analysis (PCA) is a method of data compression derived from factor analysis, a technique first developed to describe patterns in large data sets in terms of a much smaller number of underlying factors (*i.e.* to reduce the dimensionality of the data set) (52).

The first PC is that which best describes the variability, based on changes in the response data and not their absolute values, within the matrix, while the second PC and subsequent PCs successively describe the remaining variance, under the constraint that each PC is orthogonal (i.e. perpendicular) to the previous one. This is illustrated in Fig. 1.16a, which represents the **R** matrix as ten points when plotted in two-dimensional column space, and in Fig. 1.16b where the principal components are plotted also. Figure. 1.16b illustrates PCA in terms of only two dimensions, though it is important to realise that this technique is equally applicable to large matrices for which *n*PCs are required to describe the variability in *n* dimensions.

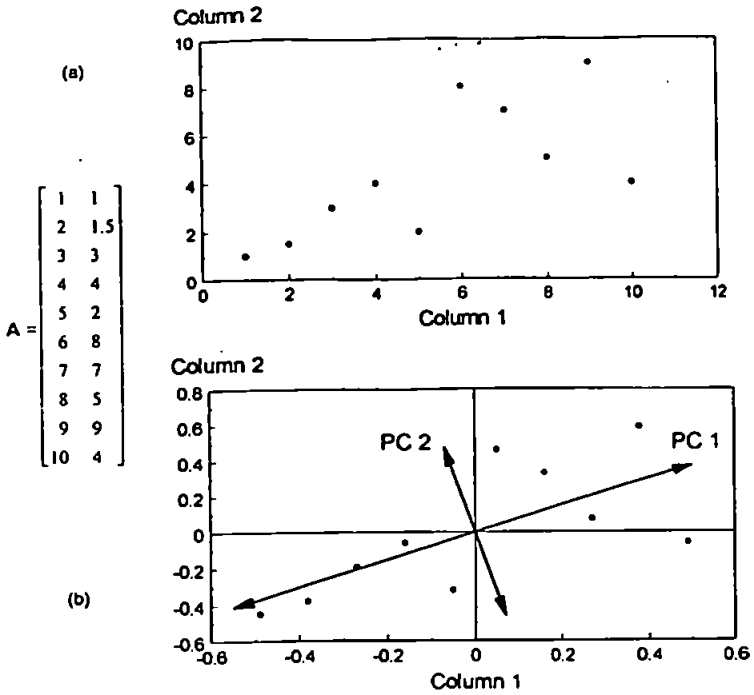


Figure 1.16 (a) The matrix A plotted in column space; (b) first and second principal components (PC 1 and PC 2) for A following mean-centering and variance scaling of the columns

Mathematically, PCA relies upon an eigenvector decomposition of the covariance or correlation matrix of the **R** matrix. The process is one of elimination such that by iterative eliminations of each independent variation from **R** in series it is possible to create a set of PCs that capture the variation in **R**. For a given data matrix **X** with m rows and n columns (each row a sample and each column a variable), the covariance matrix of **X** is defined as Eqn. 1.13.

$$cov(\mathbf{X}) = \frac{\mathbf{X}^T\mathbf{X}}{m-1} \qquad \text{Eqn. 1.13}$$

provided that the columns of **X** have been "mean centered" (*i.e.* adjusted to have zero mean). If the columns have been "autoscaled" (i.e. adjusted to zero mean and unit variance), Eqn. 1.13 gives the correlation matrix of **X**. In order to reduce unwanted

sources of variation and to highlight any relative differences within a dataset mean centering or autoscaling is often performed prior to PCA. PCA decomposes the data matrix $\mathbf{X}$ as a sum of the outer product of vectors $\hat{\mathbf{t}}_i$ and $\hat{\mathbf{p}}_i$ plus a residual matrix $\hat{\mathbf{E}}$ as in Eqn 1.14:

$$\mathbf{X} = \hat{\mathbf{t}}_1\hat{\mathbf{p}}_1^T + \hat{\mathbf{t}}_2\hat{\mathbf{p}}_2^T + ... + \hat{\mathbf{t}}_n\hat{\mathbf{p}}_n^T + \hat{\mathbf{E}} \qquad \text{Eqn. 1.14}$$

where k must be less than or equal to the smallest dimension of $\mathbf{X}$, *i.e.* $k \leq min(m,n)$. To illustrate what the $\hat{\mathbf{t}}_i$ and $\hat{\mathbf{p}}_i$ mean, an example for two variables, is shown in Fig. 1.17.

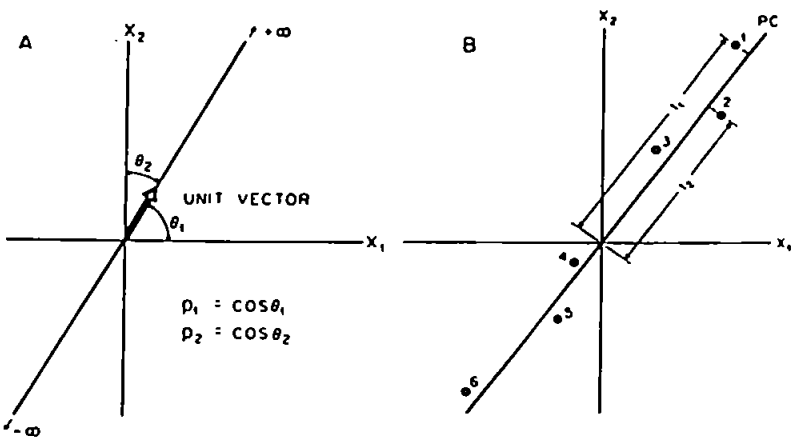


Figure 1.17 A principal component in the case of two variables x_1 and x_2: (A) loadings are the angle cosines of the direction vector; (B) are the projections of the sample points (1-6) on the principal component direction. (Note that the data are mean centered)

The $\hat{\mathbf{t}}_i$ vectors (scores) contain information on how the samples relate to each other, the closer in value the scores, the more related the samples and vice-versa. The $\hat{\mathbf{p}}_i$ (loadings) contain information on how the variables relate to one another. The extent to which a measurement variable contributes to a PC (its loading value) depends on the

relative orientation in space of the PC and variable axis. If a PC points exactly in the same direction as an individual variable, the angle between them is 0, and the cosine is 1, indicating that the PC describes all of the variation in that variable axis. Similarly, if a PC is perpendicular to an individual variable axis, the cosine is 0, indicating that none of the variation is contained in the PC. It should be noted that the length of $\hat{\mathbf{p}}_i$ is 1 because $[cos(\theta_1)^2 + cos(\theta_2)^2 = 1]$, hence the relative importance of each variable can be ascertained. Similar rules exist for more than two dimensions.

For the $\hat{\mathbf{p}}_i$ vectors of the covariance matrix Eqn. 1.15 holds

$$cov(\mathbf{X}) = \lambda_i \hat{\mathbf{p}}_i \qquad \text{Eqn. 1.15}$$

where λ_i is the eigenvalue (λ_i is a measure of the amount of variance described by the $\hat{\mathbf{t}}_i, \hat{\mathbf{p}}_i$ pair) associated with the eigenvector $\hat{\mathbf{p}}_i$. The $\hat{\mathbf{t}}_i$ form an orthogonal set $(\hat{\mathbf{t}}_i^T \hat{\mathbf{t}}_j = 0, i \neq j)$, while the $\hat{\mathbf{p}}_i$ are orthonormal $(\hat{\mathbf{p}}_i^T \hat{\mathbf{p}}_j = 0, i \neq j, \hat{\mathbf{p}}_i^T \hat{\mathbf{p}}_j = 1)$. For $\mathbf{X}$ and any $\hat{\mathbf{T}}_i, \hat{\mathbf{P}}_i$ pair Eqn. 1.16 holds

$$\hat{\mathbf{T}}_i = \hat{\mathbf{X}}\hat{\mathbf{P}}_i \qquad \text{Eqn. 1.16}$$

i.e. the score vector $\hat{\mathbf{t}}_i$ is the linear combination of the original X data defined by $\hat{\mathbf{p}}_i$ Hence Eqn. 1.17, which shows the elements of the first PC, is true also:

$$\hat{\mathbf{t}}_{1,1} = \mathbf{x}_{1,1}\hat{\mathbf{p}}_{1,1} + \mathbf{x}_{1,2}\hat{\mathbf{p}}_{2,1} + ... + \mathbf{x}_{1,j}\hat{\mathbf{p}}_{j,1} \qquad \text{Eqn. 1.18}$$

$$\hat{\mathbf{t}}_{2,1} = \mathbf{x}_{2,1}\hat{\mathbf{p}}_{1,1} + \mathbf{x}_{2,2}\hat{\mathbf{p}}_{2,1} + ... + \mathbf{x}_{2,j}\hat{\mathbf{p}}_{j,1}$$

.

.

$$\hat{\mathbf{t}}_{n,1} = \mathbf{x}_{n,1}\hat{\mathbf{p}}_{n,1} + \mathbf{x}_{n,2}\hat{\mathbf{p}}_{2,1} + ... + \mathbf{x}_{n,j}\hat{\mathbf{p}}_{j,1}$$

The variables of $\mathbf{X}$ are now replaced by new ones, $\hat{\mathbf{t}}_i$ (now linear combinations of the original variables), that have better properties (orthogonality) and also span the multidimensional space of $\mathbf{X}$. From of Eqn. 1.16 we have:

$$\mathbf{X}_i = \hat{\mathbf{T}}\hat{\mathbf{P}}_i^T \qquad \text{Eqn. 1.19}$$

which can be expanded to Eqn. 1.14 ($\mathbf{X}_{n,1} = \hat{\mathbf{t}}_1\hat{\mathbf{p}}_1^T + \hat{\mathbf{t}}_2\hat{\mathbf{p}}_2^T + ... + \hat{\mathbf{t}}_n\hat{\mathbf{p}}_n^T$).

The $\hat{\mathbf{t}}_i, \hat{\mathbf{p}}_i$ pairs are arranged in descending order according to the associated λ_i. Because the $\hat{\mathbf{t}}_i, \hat{\mathbf{p}}_i$ pairs are arranged in descending order of λ_i, the first pair capture the largest amount of variance of any pair in the decomposition. It is possible to show that the first $\hat{\mathbf{t}}_1, \hat{\mathbf{p}}_1$ pair capture the greatest amount of variation in the data that is it possible to capture with a linear factor, and that each subsequent pair captures the greatest possible amount of variance remaining after subtracting $\hat{\mathbf{t}}_n\hat{\mathbf{p}}_n^T$ from $\mathbf{X}$. Hence each successive PC captures less and less variance.

A number of algorithms can be used to perform the decomposition of the response matrix into successive PCs. Non-iterative partial least squares (NIPALS) is one of the more frequently applied methods owing to its simplicity and speed of computation (48). For each successive PC $\hat{\mathbf{t}}_i$ and $\hat{\mathbf{p}}_i$ are calculated from $\mathbf{R}$ as follows:

(a) The initial score vector $\hat{\mathbf{t}}_i$ is selected as the column of $\mathbf{R}$ with the largest variance;

(b) A new loading vector is estimated for this PC by projecting $\mathbf{R}$ onto $\hat{\mathbf{t}}_i$:

$$\hat{\mathbf{p}}_i^T = (\hat{\mathbf{t}}_i^T \hat{\mathbf{t}}_i)^{-1} \hat{\mathbf{t}}^T \mathbf{R}$$

(c) The length of p_i^T is scaled to 1.0 to avoid scaling ambiguity:

$$\hat{\mathbf{p}}_i^T = (\hat{\mathbf{p}}_i^T \hat{\mathbf{p}}_i)^{1/2} \hat{\mathbf{p}}_i^T$$

(d) A new score vector is estimated by projecting $\mathbf{R}$ onto $\hat{\mathbf{p}}_i^T$;

$$\hat{\mathbf{t}}_i = (\hat{\mathbf{p}}_i^T \hat{\mathbf{p}}_i)^{-1} \hat{\mathbf{p}}_i^T \mathbf{R}$$

(e) If the difference between the newly estimated t and the previous estimated is less than a pre-defined criterion, then the method has achieved convergence with respect to this factor. If not, then repeat the process from step (b);

(f) Subtract the effect of this from $\mathbf{R}$:

$$\mathbf{R}_{i-1} = \mathbf{R} - \hat{\mathbf{t}}_i \hat{\mathbf{p}}^T$$

If quantitative information is required, then principal components regression (PCR) can be used. PCR is conceptually similar to MLR, but the calibration model is constructed using the matrix of PC scores, $\mathbf{T}$, in place of the original response matrix $\mathbf{R}$. By reducing the number of variables in this manner the problems of collinearity can be overcome and the inversion of $\mathbf{R}^T\mathbf{R}$ is then possible. A major drawback to the PCR approach is that the PCs which best describe the variance in the response matrix may not be the best description of the variance in the analyte concentrations matrix (*e.g.* instrumental noise may be responsible for the largest component of the $\mathbf{R}$ variance). If this is the case, then the

resulting calibration model will produce poor predictive results for any unknown samples (53).

1.4.2.3 Partial Least Squares

Partial least squares (PLS) regression is conceptually similar to PCR, however PCR and PLS use different approaches for choosing the linear combinations of variables that make up the score vectors, $\hat{\mathbf{t}}_i$. Specifically PCR only uses the $\mathbf{R}$ matrix to determine the linear combinations of variables, but with PLS the variance of the concentration matrix, $\mathbf{c}$, is used in addition to the variance in $\mathbf{R}$ to generate the linear combinations. The major advantage of this method is that variation in $\mathbf{R}$ that is not correlated with the concentrations of interest is not used to construct $\hat{\mathbf{T}}$, as it is in PCR. In this way, PLS is able to determine which PCs in the response matrix are most relevant to the variance in the concentration matrix, thereby reducing the influence of the irrelevant PCs upon the calibration model (53-55)].

The technique of PLS was first introduced in 1977, following development work carried out largely by Herman Wold during the 1960s and 1970s. In analytical chemistry it is increasingly being applied to the calibration of multicomponent spectroscopic data, particularly UV-visible (56, 57), NIR (58, 59) and FT-IR (60).

The form of PLS employed in chemometrics is actually a modification of the NIPALS algorithm used in PCA, as described above. The modifications enable PLS to calculate loading vectors which contain the maximum amount of predictive information in the earlier vectors. This is achieved by using the information in the concentration matrix when performing the decomposition of the response matrix, so that the loading vectors are concentration dependent (53). There are several ways of presenting the NIPALS PLS

algorithm, e.g. based on the work of Hoskuldsson (55) and Frank (61) or Wold *et. al.* (62) and indeed several ways in which to calculate PLS model parameters, e.g. SIMPLS devised by de Jong (63). The PLS version described here is based upon non-linear iterative partial least squares (NIPALS) and is the version used in the PLS_Toolbox 2.00 (Mathworks Inc.).

NIPALS calculates scores, $\hat{\mathbf{t}}$, and loadings, $\hat{\mathbf{p}}$, and an additional set of vectors known as weights, $\hat{\mathbf{w}}$. The addition of weights is required to maintain orthogonal scores. A vector of "inner-relationship" coefficients, $\hat{\mathbf{b}}$, which relate the $\mathbf{X}$ and $\mathbf{Y}$ block scores, must also be calculated. To use the concentration information during the decomposition of the response matrix, scores, $\hat{\mathbf{u}}$, and loadings, $\hat{\mathbf{q}}$, are also calculated for $\mathbf{Y}$. Using NIPALS the scores, weights, loadings and inner-coefficients are calculated sequentially for PCs $1 = h$ as shown:

(a) The PLS decomposition is started by selecting one column of $\mathbf{Y}$ as the starting estimate for $\hat{\mathbf{u}}_h$ (usually the $\mathbf{Y}$ column with the greatest variance is chosen). In the situation where only 1 analyte is being determined, $\hat{\mathbf{u}}_h = \mathbf{y}$.

(b) Starting in the $\mathbf{X}$ data block the weights $\hat{\mathbf{w}}_h$ and scores $\hat{\mathbf{t}}_h$ are calculated

$$\hat{\mathbf{w}}_h = \frac{\mathbf{X}^T \hat{\mathbf{u}}_h}{\| \mathbf{X}^T \hat{\mathbf{u}}_h \|}$$

$$\hat{\mathbf{t}}_h = \mathbf{X}\hat{\mathbf{w}}_h$$

(c) In the y data, the loadings and scores are calculated

$$\hat{\mathbf{q}}_h = \frac{\hat{\mathbf{u}}_h{}^T \hat{\mathbf{t}}_h}{\| \hat{\mathbf{u}}^T{}_h \hat{\mathbf{t}}_h \|}$$

$$\hat{\mathbf{u}}_h = \mathbf{Y}\hat{\mathbf{q}}_h$$

(d) Check for convergence by comparing $\hat{\mathbf{t}}_h$ in step (b) with the value from the previous iteration. If they are equal within rounding error, proceed to step (e). If they are not, return to step (b) and calculate $\hat{\mathbf{w}}_h$ using $\hat{\mathbf{u}}_h$ from step (c). If the *Y* block is univariate step (c) can be omitted, $\hat{\mathbf{q}}_h = 1$, and no iteration is required.

(e) Calculate the **X** data loadings and rescale the scores and weights accordingly:

$$\hat{\mathbf{p}}_h = \frac{\mathbf{X}^T \hat{\mathbf{t}}_h}{\| \hat{\mathbf{t}}_h{}^T \hat{\mathbf{t}}_h \|}$$

$$\hat{\mathbf{p}}_{hnew} = \frac{\hat{\mathbf{p}}_{hold}}{\| \hat{\mathbf{p}}_{hold} \|}$$

$$\hat{\mathbf{t}}_{hnew} = \hat{\mathbf{t}}_{hold} \| \hat{\mathbf{p}}_{hold} \|$$

$$\hat{\mathbf{w}}_{hnew} = \hat{\mathbf{w}}_{hold} \| \hat{\mathbf{p}}_{hold} \|$$

(f) Find the regression coefficients $\hat{\mathbf{b}}$ for the inner relation:

$$\hat{\mathbf{b}}_h = \frac{\hat{\mathbf{u}}_h{}^T \hat{\mathbf{t}}_h}{\hat{\mathbf{t}}_h{}^T \hat{\mathbf{t}}_h}$$

(g) After the scores and loadings have been calculated for the first PC, the **X** and **Y** residuals are calculated as follows:

$$\hat{\mathbf{E}}_h = \mathbf{X} - \hat{\mathbf{t}}_h \hat{\mathbf{p}}_h$$

$$\hat{\mathbf{F}}_h = \mathbf{Y} - \hat{\mathbf{b}}_h \hat{\mathbf{t}}_h \hat{\mathbf{q}}_h{}^T$$

(h) The entire procedure is now repeated for the next PC starting from step (b). Both $\mathbf{X}$ and $\mathbf{Y}$ are replaced with their residuals $\hat{\mathbf{E}}_h$ and $\hat{\mathbf{F}}_h$ respectively.

Prediction can then be performed using the following procedure:

(a) Calculate the PLS regression coefficients as

$$\hat{\boldsymbol{\beta}}_{PLS} = \hat{\mathbf{W}}(\hat{\mathbf{P}}^T\hat{\mathbf{W}})^{-1} diag(\hat{\mathbf{b}})\hat{\mathbf{Q}}^T \text{ for multivariate } \mathbf{Y}$$

$$\hat{\boldsymbol{\beta}}_{PLS} = \hat{\mathbf{W}}(\hat{\mathbf{P}}^T\hat{\mathbf{W}})^{-1} diag(\hat{\mathbf{b}}) \text{ for univariate } \mathbf{Y}$$

(b) New sample concentrations are then calculated as

$$\hat{\mathbf{Y}}_{new} = \mathbf{R}_{new}\hat{\boldsymbol{\beta}}_{PLS}$$

1.5 Application of Univariate and Multivariate Calibration in Spectroscopy

The application of univariate calibration in atomic spectroscopy can be divided into the following techniques: univariate calibration (UC); uc incorporating matrix matching; standard additions; and interelement correction (IEC).

In general spectroscopic fields the use of univariate calibration by itself is limited to chemical systems where the analyte line under investigation is interference free or where the interferences are slight. Univariate regression was used by Peralta-Zamora *et. al.* for

the quantitative determination of a set of 20 different mixtures of Ce, Pr and Nd and Sm concentrations in a commercial rare-earth product using FAAS. However, significantly better precision was obtained using the multivariate calibration method partial least squares 1 (PLS1) (64). Another example of the use of univariate calibration was given by Luis *et. al.*, in which the determination of chlorthalidone and spironolactone by HPLC gave poor results because of the scarcity of interference free wavelengths (65). In those instances where the interferences are significant then univariate calibration gives poor predictive results. An effective method of correcting for both spectroscopic and non-spectroscopic interferences is matrix matching, where the calibration standards used to construct a univariate calibration curve are matched, chemically and physically to the samples. In theory, any analyte line now used to determine analyte response in the calibration samples will incorporate any interferent contribution present in the sample. Signal enhancement or suppression effects will also be incorporated in the calibration standards. This technique is widespread in atomic spectroscopy for a variety of instrumental techniques including ICP-AES (66), HPLC-TLS (high performance liquid chromatography-thermal lens spectrometry) (67), ICP-MS (68) and GFAAS(69). However, this technique does have limitations. For accurate analyte determination the calibration standards must be matched with the sample in terms of chemical composition. Failure to do so will give poor predictive results (40). This means that the composition of the sample must be known, *a priori* which is not always possible.

The method of standard additions, an extension of univariate calibration is again widespread throughout atomic spectrometry (70-72), but has several limitations which have been discussed earlier.

One of the most widespread uses of univariate calibration is in conjunction with interelement correction (73-75). Griffiths *et. al.* (40) showed that the predictive errors for the determination of Pt, Pd and Rh in a complex matrix using interelement correction were

acceptable and were of the same order when matrix matching was used. The successful use of this method, is dependent upon *a priori* knowledge of the sample being available and it does not correct for non-spectroscopic interferences. If significant interferents are not accounted for (in the form of IEC factors), any subsequent predictions will be poor. The use of interelement correction is therefore restricted to samples with known complexity.

Another approach to minimise the effects of interferences is the use of mathematical and statistical procedures, collectively referred to as multivariate calibration. Multivariate calibration techniques can be divided into two main groups: linear and non-linear. Although in most real situations the chemical additivity principle is not strictly fulfilled, linear multivariate methods can provide good approximations of many types of non-linearities (76, 77).

Multivariate techniques are of great interest in chemical analysis and have been applied to various spectroscopic methods of analysis, e.g. infrared (58, 60, 78), UV/VIS spectroscopy (57, 79, 80), NMR spectroscopy (81), mass spectroscopy (82) and emission spectroscopy (4, 40, 83-85).

The multivariate techniques themselves fall broadly into four common methods, i.e. partial least squares (PLS) (47, 86, 87), principal components regression (PCR) (80, 88, 89), stepwise multiple linear regression (SMLR) (83, 90, 91) and principal components analysis (PCA) (82, 92, 93) and have become common tools for the analytical chemist.

Other, less widely used techniques include: numerical derivatives (94, 95); curve resolution (96, 97); orthogonal polynomials (98) and Kalman filtering (99-101).

Multivariate calibration methods have been applied to various analytical problems using a wide variety of instrumental techniques. Mc Shane *et. al.* (102) successfully determined the concentrations of lactate, ammonia, glutamate and glutamine from NIR spectroscopy using partial least squares coupled with a variable selection algorithm. In the

field of mass spectrometry both partial least squares and principal components regression have been successfully applied to the correction of spectral and non-spectral interferences in samples containing a mixture of Fe, Ni, Cu, Zn and Pb (103).

The growth in the use of array detectors has resulted in a commensurate increase in the application of chemometric procedures to make the best use of this type of spectral data acquisition. The predominant use of multivariate methods in ICP-AES is in the area of calibration and the removal of interferences. Methods to correct for spectral interference have included peak purity assessment by matrix projection (104)and non-linear deconvolution to test different line shapes (105). The rare earth elements (REEs) produce some the most line rich spectra and two studies have used multivariate methods to improve the determination of trace REEs in rare earth element matrices. The first used multi-component analysis (106) and the other a spectral fitting approach based on least squares minimisation (107). Other multivariate techniques include stepwise multiple linear regression (83), numerical derivatives (95), PCR (108), PLS (40), the generalised standard additions method (GSAM) (90, 109, 110) and Kalman filtering (99-101).

The work of Sadler (108) successfully determined the concentrations of Cr, Mn and V (0.5 $\mu g\ ml^{-1}$) in a matrix containing Ce (100 $\mu g\ ml^{-1}$), La (40$\mu g\ ml^{-1}$), Dy (40$\mu g\ ml^{-1}$), Fe (20$\mu g\ ml^{-1}$), Mo(20 $\mu g\ ml^{-1}$) and Co (20 $\mu g\ ml^{-1}$) and 5% v/v HNO_3 using PCR. The predicted concentrations for Cr, Mn and V were 0.51 ± 0.01, 0.49 ± 0.01 and 0.51 ± 0.01 $\mu g\ ml^{-1}$ respectively. However, in more complex systems it would be expected, that PLS would generally give better predictive results, because of its ability to obtain more correlated information in the first few PCs. Similarly, Pimental *et. al.*(4) successfully used PLS and PCR to predict the concentrations of Mn (0.4-2.0%), Mo (0.6-4.0%), Cr (8.0-28.0%), Ni (4.0-28.0%) and Fe (40.0-88.0%) in natural waters. Instead of using analytical lines, important regions of the full spectrum were located which possessed, both intense analyte and matrix lines with a minimum of spectral overlap. Although the predictions

were acceptable, such a methodology does not necessarily locate the most important modelling regions from the viewpoint of the PLS algorithm. Areas of major spectral overlap will only be considered by PLS as important depending on how well correlated it is with the analyte in question.

With two exceptions, Veen *et. al.* (85) and Pimental *et. al.*(4), these techniques were not applied to the full atomic emission spectrum. Most of the instrumental calibration methods applied to ICP-AES have utilised data in the form of integrated line intensities, which has traditionally been the form in which the data has been presented for multivariate calibration. However, the use of data in this form presents a number of problems as follows:

(i) Suitable analytical lines must be selected beforehand, using subjective criteria (e.g. intensity, analyte line, matrix line), so that useful information may be missed.

(ii) Each spectral line must be integrated to yield integrated line intensities, during which substantial errors can be introduced.

(iii) If net intensities are required then a suitable background correction method must be employed, which can often yield erroneous results for low intensity lines or those with potential spectral interferences.

Hence, it would be desirable for as much of the ICP-AES spectrum as possible to be used, rather than selected lines, a technique known as full spectrum modelling.

1.6 Research Objectives

The objectives of this research were to investigate the potential of applying suitable multivariate calibration techniques to the full segmented spectrum obtained from samples introduced into an inductively coupled plasma atomic emission spectrometer by developing a suitable wavelength variable selection strategy.

The specific objectives were:

1 To compare 'traditional' calibration techniques, such as univariate calibration, matrix matching e.t.c. with 'traditional' multivariate calibration techniques, e.g. multivariate calibration using individual analyte and matrix lines.

2 To develop a methodology for utilising a full segmented ICP-AES spectrum using multivariate techniques coupled to variable selection and to transfer this methodology to complex industrial samples.

3 To assess multivariate model-transferability over time by using multivariate signal transformation techniques.

4 To develop a technique capable of estimating test sample confidence intervals using parameters from a multivariate model.

CHAPTER 2 - INVESTIGATION OF INTERFERENCES IN ICP-AES USING INTERELEMENT CORRECTION

2.1 Introduction

Interferences caused by partial or direct spectral overlap can result in incorrect estimates of concentration. These are primarily the result of interferences from other chemical elements in the sample or structured molecular interference formed by plasma-sample interactions. The magnitude of these interferences is an important consideration when any method of calibration is performed. Hence, the nature and extent of expected spectral interferences within a synthetic autocatalyst sample, and their influence on the use of interelement correction factors was investigated in order to provide a baseline study and rationale for the subsequent investigation of univariate and multivariate calibration.

2.2 Experimental

2.2.1 Instrumentation and Reagents

All data were collected using a simultaneous echelle-based inductively coupled plasma atomic emission spectrometer (Perkin-Elmer Optima 3000 ICP, Norwalk, USA) equipped with a segmented-charge-coupled array detection (SCCD) system. Instrumental operating conditions were optimised using simplex optimisation and are given in Table 2.1. Simplex optimisation (44) was used in this work to optimise plasma gas, auxiliary gas, viewing height and power. The optimisation procedure was performed for a 1 $\mu g\ ml^{-1}$ Pt standard, measuring the emission at 214.423 nm, the figure of merit for the optimisation being the limit of detection (LOD). The simplex routine used was programmed in-house and

was of variable-simplex construction, thus allowing the limit of detection (LOD^{-1}) (simplex programmed to find response surface maximum) to be found more quickly and accurately compared to the non-variable simplex design. Because of the potentially high solid content of the acid-digest and fusion samples, a wide bore injector tube (2.5 mm) was used, resulting in an optimised nebuliser gas flow rate of 0.93 l min^{-1}.

Single and multielement solutions were prepared by serial dilution of ultra-pure stock standards (10,000 and 1000 μg ml^{-1}, Johnson Matthey plc, Royston, Hertfordshire). Water was double deionised, (18 MΩ quality) and acids were of Aristar grade (Merk-BDH, Poole, Dorset). All glassware was acid washed in 10 % v/v nitric acid for 24 hours then rinsed thoroughly with 18 MΩ water. All plasticware was metal-free high-density polypropylene (Anachem, Luton, Bedfordshire). Calibration and test solutions containing varying concentrations of Pt, Pd, Rh, Al, Mg, Ce, Zr and Ba, were prepared from the stock solutions and stored in high-density polypropylene tubes.

Table 2.1 'Simplex' Optimised instrumental (Optima 3000) parameters used for the collection of all data.

Nebuliser gas flow (l min^{-1})	0.93
Auxiliary gas flow (l min^{-1})	0.5
Plasma gas flow (l min^{-1})	16
Viewing height above the load coil (mm)	9 (IRZ/NAZ)
Power (W)	1286
Spray chamber	Ryton, double-pass
Nebuliser	Seaspray, glass concentric
Injector diameter (mm)	2.5
Resolution Mode (nm $pixel^{-1}$)	High (0.01)
Read time/ integration time (s)	3/0.2
Sample uptake (ml min^{-1})	1.75

2.2.2 Procedure

The calculation of interelement correction factors can be summarised as follows:

i) The instrument is calibrated for a specific analyte at a specific wavelength (e.g. Pt 214.423 nm with two-point background correction using a zero and a top standard equal to the expected maximum in the unknown samples).

ii) A 1000 μg ml^{-1} solution of the interferent element (e.g. Al) is aspirated while monitoring the peak wavelength and background signals for the analyte element (e.g. Pt 214.423 nm)

iii) The *apparent* concentration of the analyte element (Pt) is calculated using the calibration curve and the *apparent* background corrected Pt signal due to the Al interference.

iv) A blank solution (matrix of the Pt standard minus Pt, e.g. double deionised 18 MΩ water) is analysed and the *apparent* Pt concentration is calculated in the same way as iii above. If the interferent solution and future standards are matrix matched then the *apparent* blank Pt concentration is not needed.

v) The correction factor (F) is calculated by dividing the *apparent* analyte concentration by the concentration of the interfering element (1000 μg ml^{-1} in this case).

vi) When a sample is analysed the concentration of the interfering element (Al_{conc}) is determined as part of the analysis in the usual way, and this value is multiplied by the correction factor (F) to obtain the corrected value which must be subtracted from the uncorrected concentration of the analyte (Pt_{uncorr}) in order to obtain the corrected value (Pt_{corr}). Mathematically this process can be described by Eqn. 2.1:

$$Pt_{corr} = Pt_{uncorr} - (F \times Al_{conc}) \qquad \text{Eqn. 2.1}$$

Equation 2.1 can be changed (Eqn. 2.2) so as to accommodate multiple interferents.

$$Pt_{corr} = Pt_{uncorr} - \sum_{i=1}^{i} (F_i \times I_{ic}) \qquad \text{Eqn. 2.2}$$

where F_i and I_{ic} are the *ith* correction factor and *ith* interferent contributing to the apparent Pt signal. Correction factors calculated in this way will be specific to a particular instrument and subject to change over time. The above procedure (steps i – vi) was performed on those analyte lines and interferents listed in Table 2.2.

Table 2.2 Analytes and interferents used in IEC study

Analyte line (nm)	Interferents
Pt 265.945	Mg, Ce, Zr
Pt 193.700	Mg, Ce, Zr
Pt 204.937	Mg, Ce, Zr
Pd 248.892	Mg, Ce, Zr
Pd 363.470	Mg, Ce, Zr
Rh 343.489	Mg, Ce, Zr

These particular analyte lines and interferents were chosen to represent a broad range of interference types including direct and partial spectral overlap and background shift.

2.3 Results and Discussion

2.3.1 Positioning of Background Correction Points

Before interelement correction factors can be calculated it is first necessary to ensure correct integration of the analytical lines. This is only possible with the correct positioning

of background correction points as shown in Figure 2.1. Incorrect positioning of the BCPs (BC1-BC3, dashed line) would omit integration of the hatched area under the peak thus resulting in an incorrect peak area. The presence of interferents can further complicate the positioning of the BCPs as can be seen in Fig. 2.2 where the interferent has caused the shifting of the BCP to a new position (original position obtained from pure analyte solution spectra). The new BCP position will be dependent on both the position of the interferent peak and also the peak intensity (i.e. interferent concentration). Incorrect positioning of the BCPs (Fig. 2.3, BCPs 3 and 4) can also result in negative net intensities. Hence, the presence of neighbouring interferents must be taken into account when positioning the BCPs which is only possible when the sample composition is well characterised.

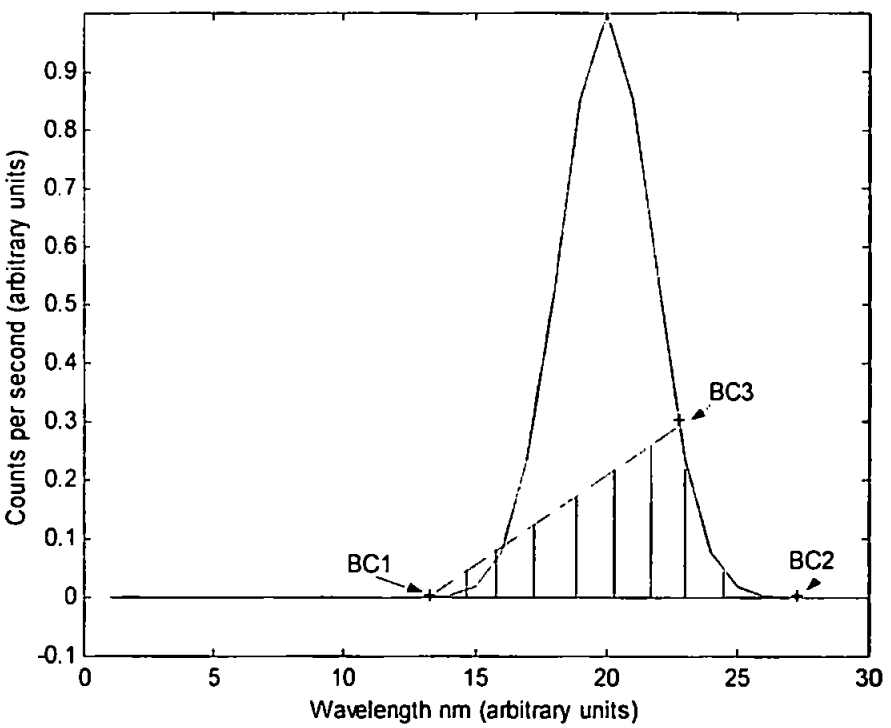


Figure 2.1 Schematic of correct background correction point (BCP) positioning (BC1-BC2) and incorrect BCP positioning (BC1-BCP3)

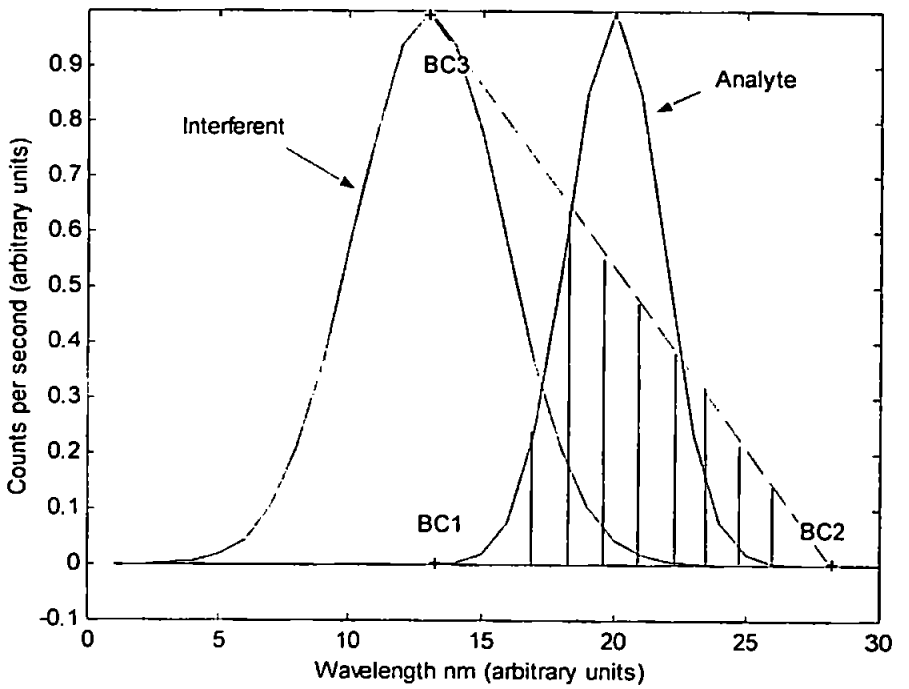


Figure 2.2 Schematic of correct background correction point (BCP) positioning and incorrect BCP positioning caused by a neighbouring interferent.

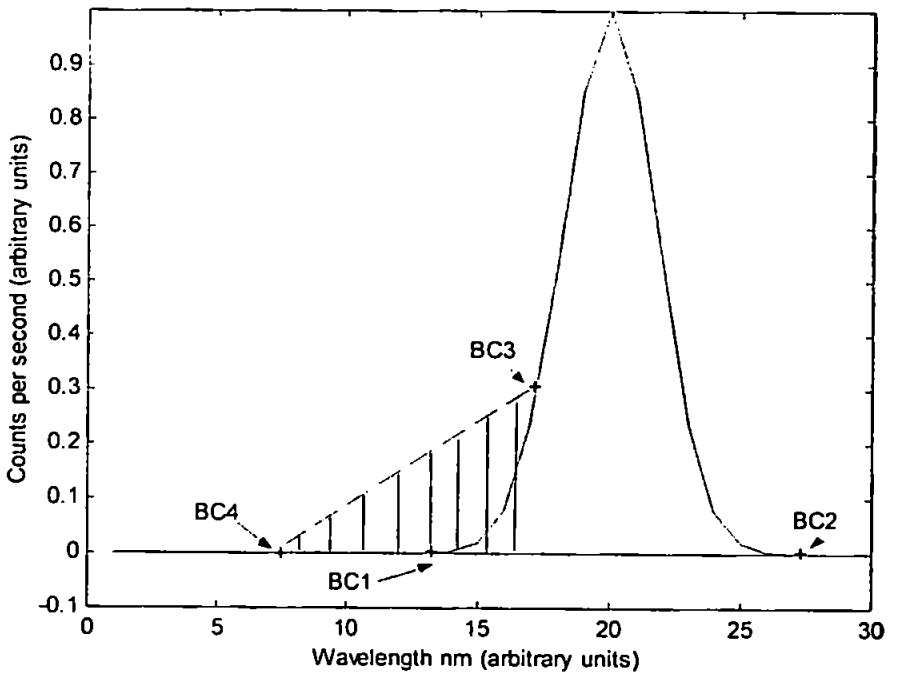


Figure 2.3 Schematic of incorrect background correction point (BCP) positioning (BCPs 3 and 4) causing a negative net intensity.

Background effects are illustrated by observing the analyte lines Pt (20 μg ml^{-1}) at 265.945 nm (Fig.2.4) and Pt at 193.700 nm (Fig. 2.5) with the interferents Mg, Ce and Zr (500 μg ml^{-1}). The interference experienced by Pt at 265.945 nm is a simple background shift which is easily corrected for by examining a mixed solution of Pt and Mg and positioning two BCPs accordingly (Fig. 2.4). In terms of calibration this is an adequate correction provided that future samples contain no other interferents that show significant signal intensity at this wavelength.

A more complicated background shift and spectral interference is experienced by Pt at 193.700 nm with the addition of Mg, Ce and Zr. Here, the background shift is combined with partial spectral overlap from Ce on the left and Zr on the right (Fig. 2.5). The addition of two BCPs will provide adequate correction for these interferences, but as in the last case, any additional interferents not compensated for by the inclusion of the BCPs will lead to incorrect calibration results. As can be seen from Fig. 2.5a, if the Ce peak were positioned directly under the Pt line at 193.680 nm then it would be impossible to assign BCP's, because the Pt signal would be the sum of Pt + Ce (i.e. direct spectral overlap).

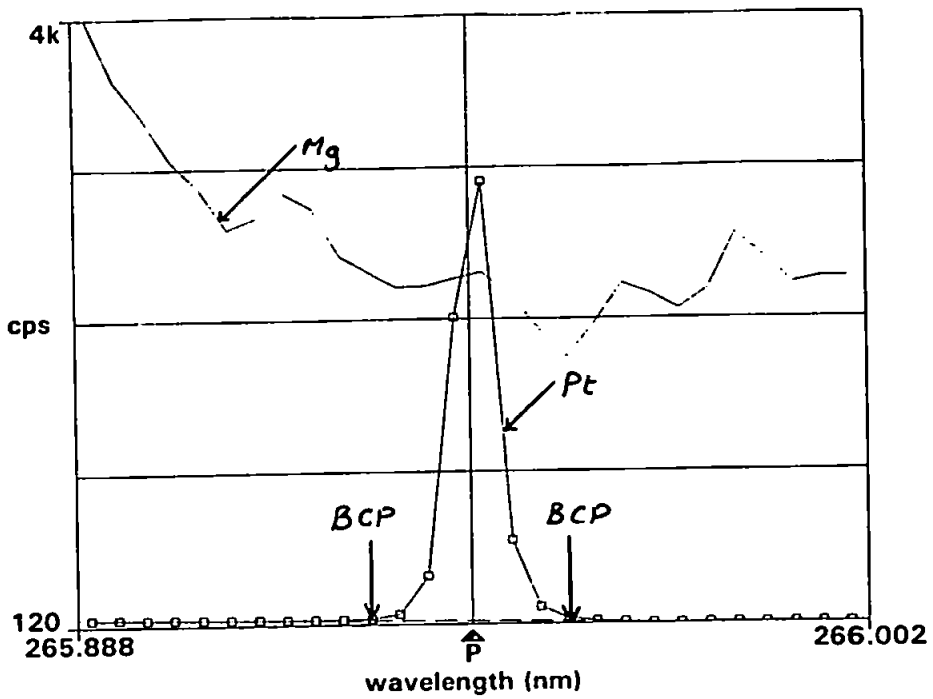


Figure 2.4a Spectra of Pt (20 μg/ml) at 265.945 nm with the interferent Mg (500 μg ml^{-1}) overlaid showing BCPs positioned using a pure Pt solution.

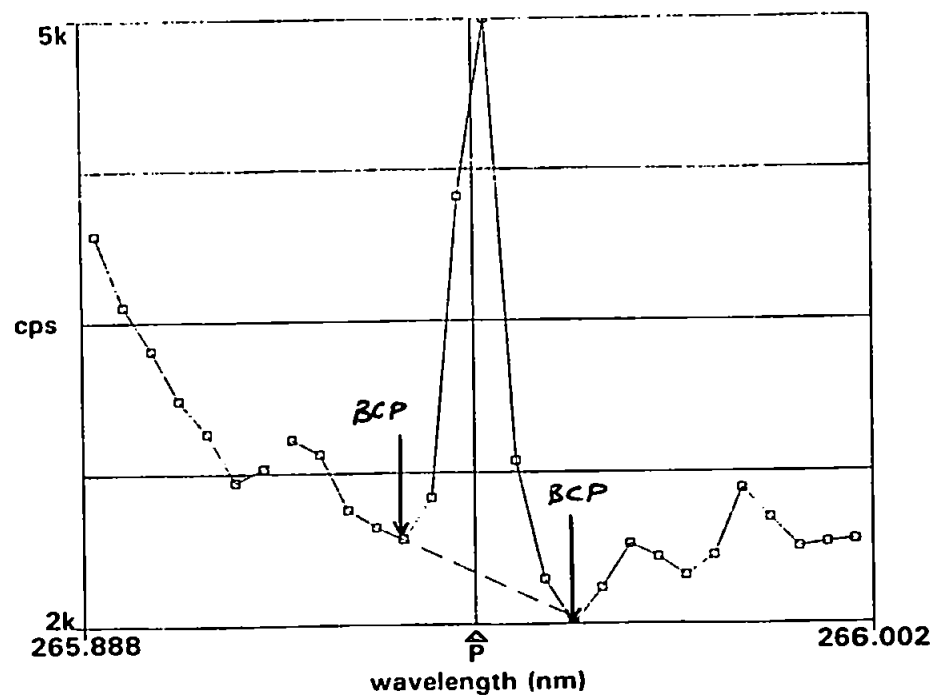


Figure 2.4b Spectra of Pt (20 μg/ml) at 265.945 nm with the interferent Mg (500 μg ml^{-1}) showing BCPs positioned using a mixed Pt + Mg solution.

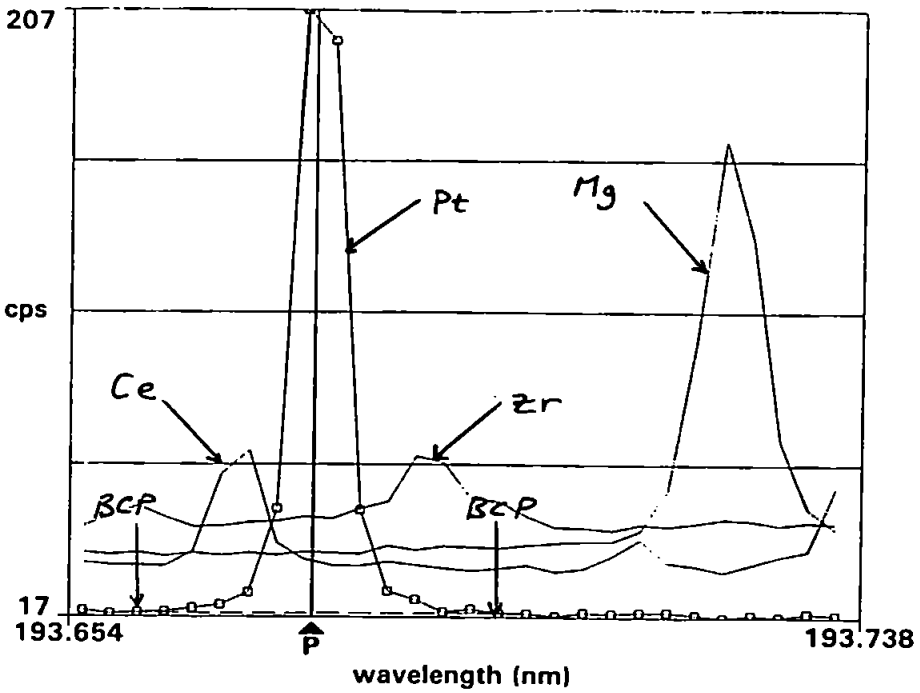


Figure 2.5a Spectra of Pt (20 µg/ml) at 193.700 nm with the interferents Mg, Ce and Zr (500 µg ml^{-1}) overlaid showing BCPs positioned using a pure Pt solution.

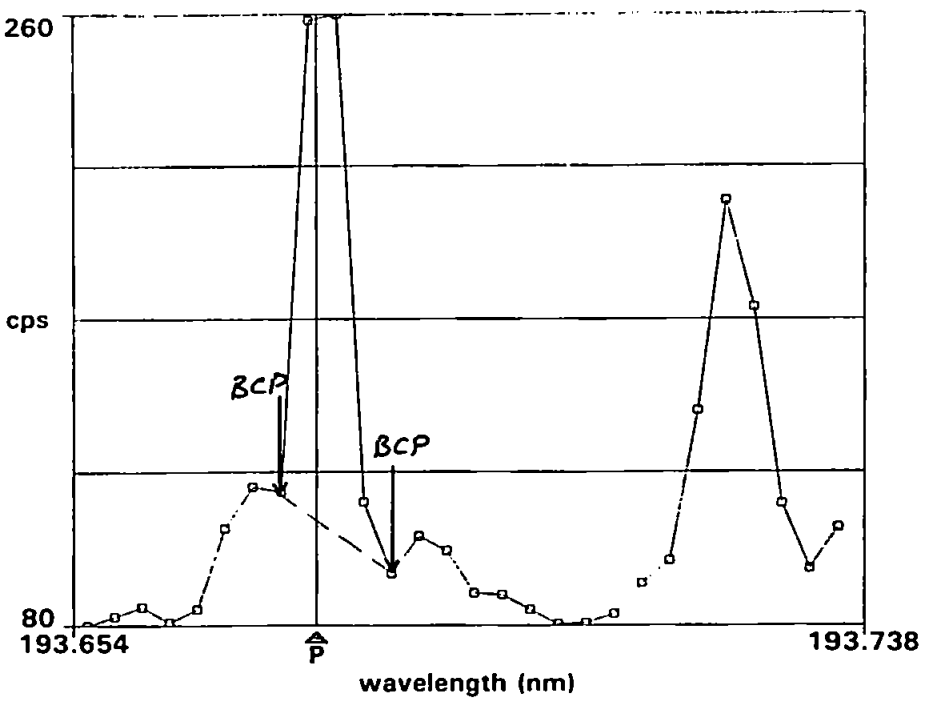


Figure 2.5b Spectra of Pt (20 µg/ml) at 193.700 nm with the interferent Mg, Ce and Zr (500 µg ml^{-1}) showing BCPs positioned using a mixed Pt + Mg + Ce + Zr solution.

2.3.2 Effect of Interferent Correction on IEC Factors

To study the effect of increasing interferent contribution on the magnitude of the IEC factors, the factors were calculated for a number of analyte elements in the presence of varying concentrations of several different matrix elements. Results shown in Fig 2.6 and Table 2.3, summarising the effect of Mg on the Pt 265.945 nm line, are typical of those obtained. As can be seen, the IEC factor increased (i.e. became less negative) as the concentration of Mg increased from 500 to 2500 μg ml^{-1}. Examination of the equation used to calculate the IEC factor (Eqn. 2.1) would lead one to expect that it would remain constant regardless of the concentration of Mg. This would be so providing that the interference caused by Mg on the Pt 265.945 nm line was linearly additive in nature, however, this was evidently not the case as can be seen in Fig 2.6.

Table 2.3 Apparent analyte concentration (Pt 265.945 nm) and IEC factor calculation as a function of interferent concentration (Mg). BCP's positioned using Pt solution.

Interferent concentration (μg/ml)	Apparant analyte concentration (μg/ml)	Calculation of IEC factor	IEC factor
500	-2.91	-2.9/500	-0.0058
1000	-3.43	-3.4/1000	-0.0034
1500	-3.71	-3.7/1500	-0.0025
2000	-3.92	-3.9/2000	-0.0020

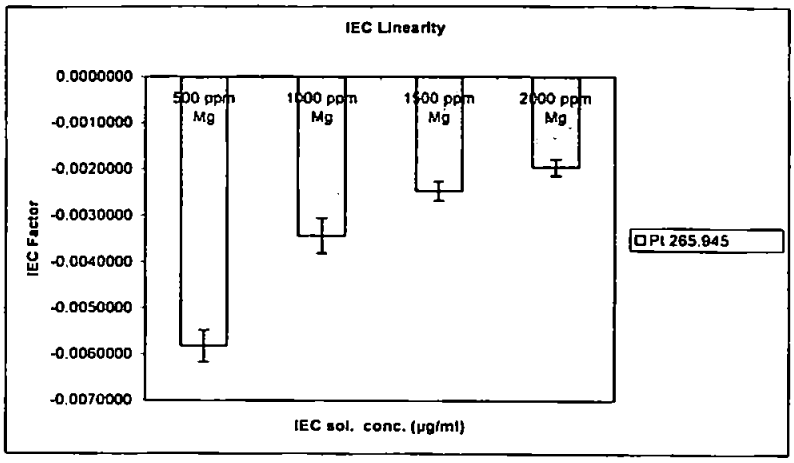


Figure 2.6 The effect of increasing Mg concentration on the determination of the IEC factor at Pt 265.945 nm (errors bars indicate 3x σ of signal intensity)

The positioning of the BCPs shown in Fig. 2.7 was effected by aspirating a pure Pt solution (20µg ml^{-1}), however on aspiration of the interferent Mg solution (Fig. 2.8) it becomes obvious that the positioning of the BCPs leads to the erroneous results shown in Fig 2.2 and Table 2.3 providing one takes into account how the instrument software integrates the peak areas (Figs 2.1 – 2.3). As can be seen in Fig. 2.8, the original positions of the BCPs (position 1) resulted in an integrated area indicated by the single hatched area which was calculated as a negative value resulting in negative values for the IEC factors. As the concentration of Mg was increased the negative apparent concentration increased, but at a lower rate than the increase in Mg, thereby leading to an apparent increase in the magnitude of the IEC factor (Table 2.3). However, the actual interference caused by Mg on Pt 265.945 is shown by the double-hatched area in Fig. 2.8, which was correctly integrated when the BCPs were altered to position 2, which resulted in the IEC factors shown in Fig 2.10 and Table 2.3. The correct positioning of BCPs can, therefore, only be accomplished when all of the interferents are known, as shown in Fig 2.9b. Figure 2.9a shows how the

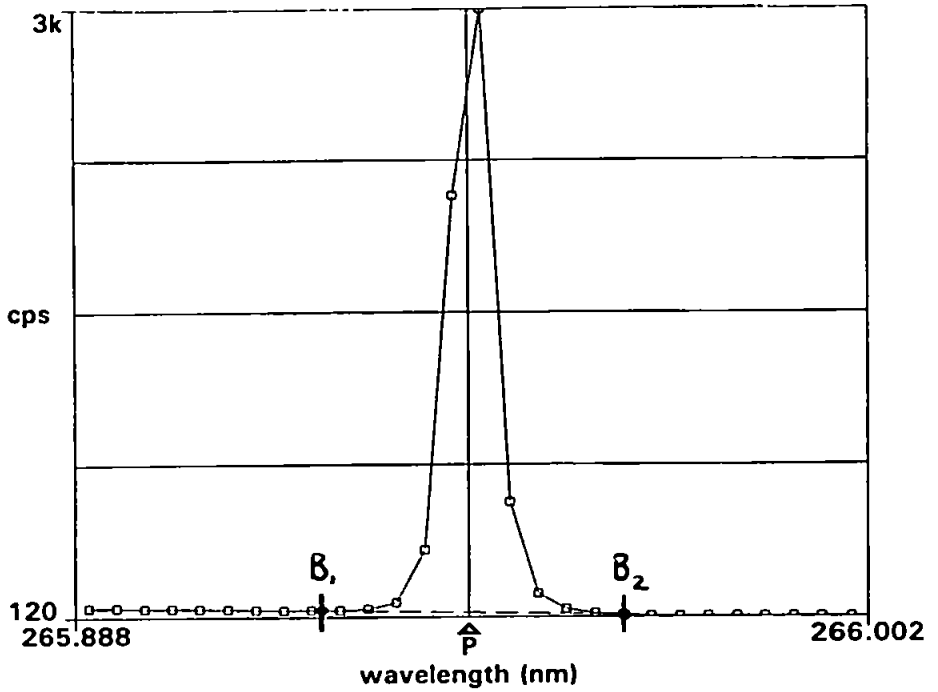


Figure 2.7 Spectrum of Pt 265.945 nm (20 μg ml^{-1}), showing positions of background correction points B1 and B2.

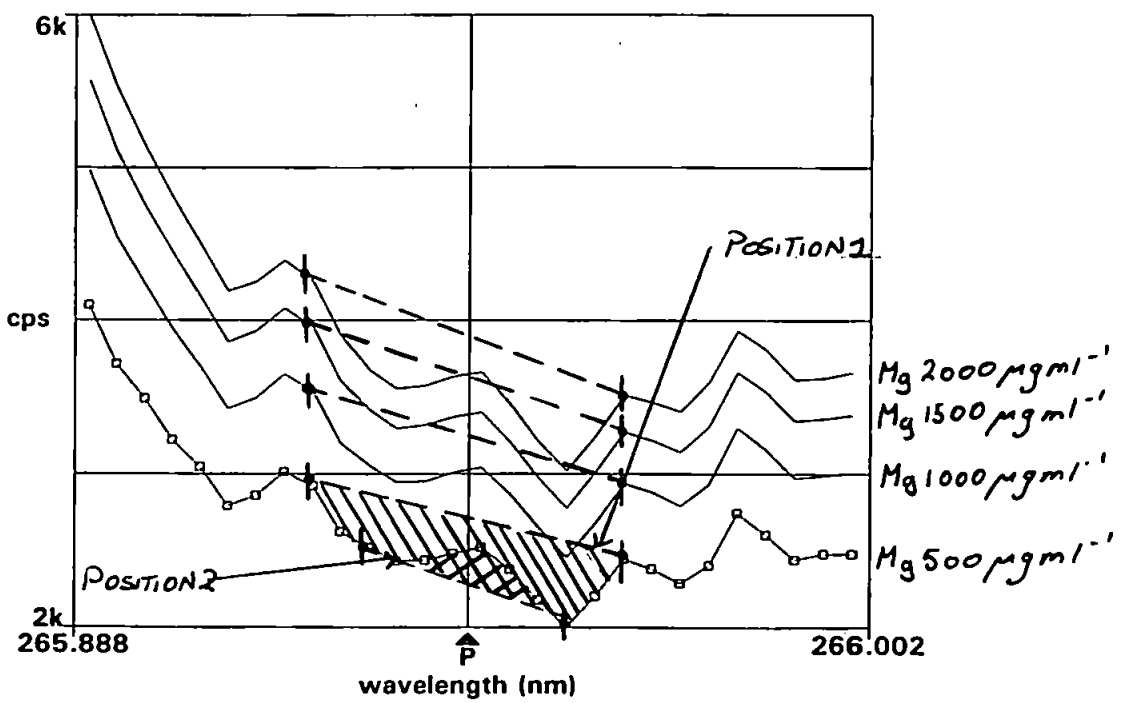


Figure 2.8 Spectra of Mg (500 μg ml^{-1}) at 265.945 nm (BCPs positioned using a Mg solution).

BCPs are positioned using only a pure Pt (20 μg ml^{-1}) solution, whereas Fig. 2.9b shows the BCP positions using a mixed solution of Pt (μg ml^{-1}) and the suspected interferent Mg (500 μg ml^{-1}). Only using Pt, when positioning the BCPs, has resulted in the integration of two areas either side of the Pt 265.945 nm peak (hatched) which are integrated as negative intensities (Fig. 2.9a). These are subsequently subtracted form the hatched area, which represents the majority of the Pt 265.945 nm peak. However, the use of a mixed Pt and Mg solution has given BCPs which give the correct integrated intensity (Fig. 2.9b).

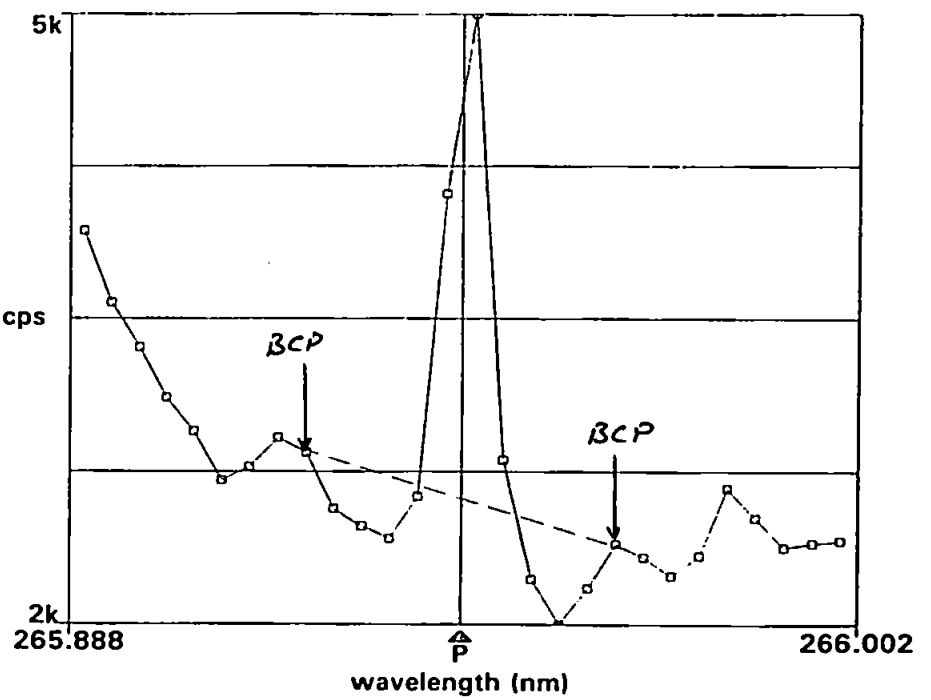


Figure 2.9a Spectrum of Pt (20 μg ml^{-1}) + Mg (500 μg/ml) at 265.945 nm (BCPs positioned according using a Pt solution).

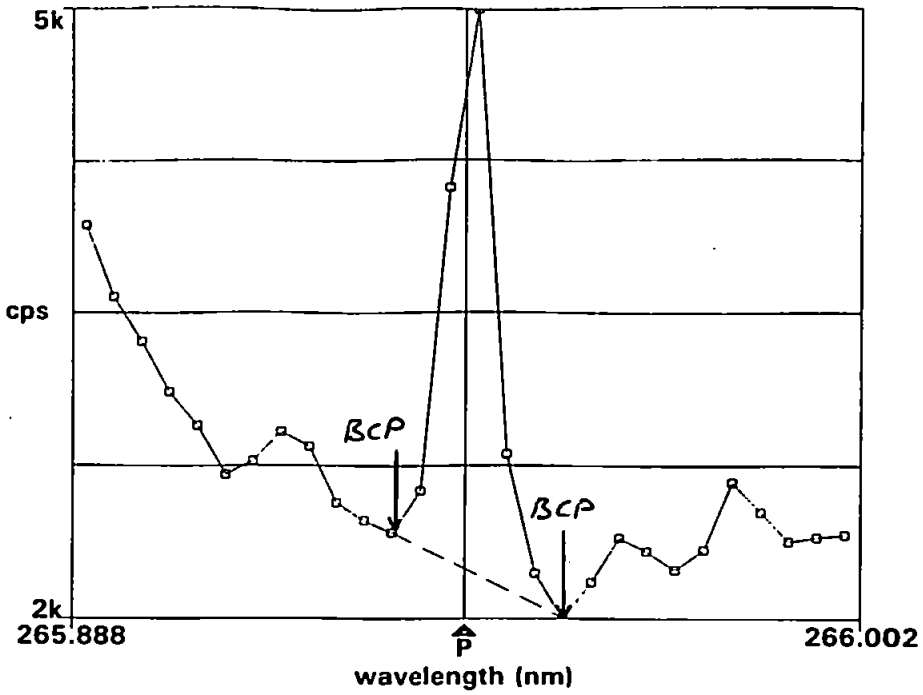


Figure 2.9b Spectrum of Pt (20 μg ml^{-1}) + Mg (500 μg/ml) at 265.945 nm (BCPs positioned using a mixed Pt + Mg solution).

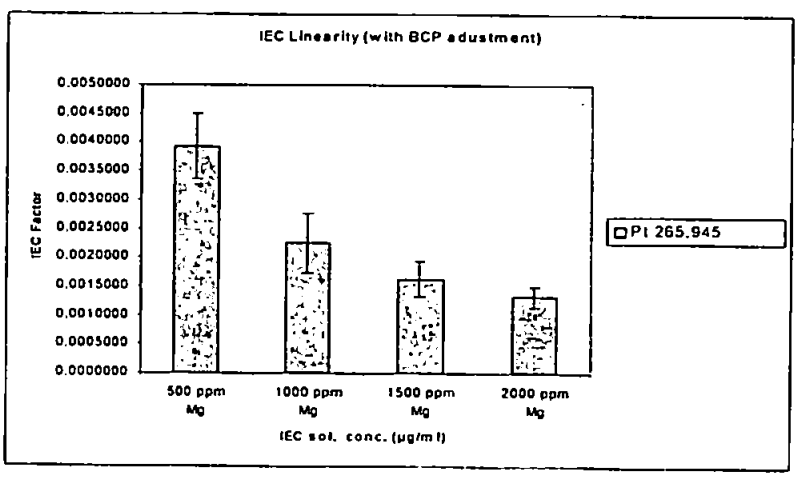


Figure 2.10 The effect of increasing Mg concentration on the determination of the IEC factor at Pt 265.945 nm (errors bars indicate 3x σ of signal intensity)

Even with the correct positioning of the BCPs the IEC factors still varied with concentration of Mg, however, in this case they were all positive due to the BCP positioning taking into consideration the presence of the interfering Mg line, thus resulting in positive apparent analyte concentrations at Pt 265.945 nm (Table 2.4). The non-linear

additive behaviour of the interfering signal is probably a combination of several processes including changes in nebulisation efficiency, ionisation suppression and energy transfer processes due to the introduction of increasing amounts of Mg.

Table 2.4 Apparent analyte concentration (Pt 265.945 nm) and IEC factor calculation as a function of interferent concentration (Mg). BCPs positioned using Pt + Mg solution.

Interferent concentration (µg/ml)	Apparant analyte concentration (µg/ml)	Calculation of IEC factor	IEC factor
500	2.01	2.01/500	0.0039
1000	2.30	2.30/1000	0.0022
1500	2.50	2.50/1500	0.0016
2000	2.69	2.69/2000	0.0013

2.3.3 Additivity of IEC Factor

This discussion will look at the additivity of IEC factors where the BCPs have been positioned using pure analyte solutions as this is the most common method of positioning. In general, IEC factors are calculated separately for each interfering element then subsequently summed in order to derive an overall IEC factor (Eqn. 2.2) for the interfering elements present in a particular matrix. However, this assumes that IEC factors, which are determined independently are additive in nature. As has been shown in the previous section the IEC factors can change depending on the concentration of the interferent element that was used to determine it, so it is possible that the assumption of additivity is not valid.

In order to test this hypothesis IEC factors were calculated by summing individual factors for different interferents and also by determining the IEC factors using a mixed solution of the interfering elements in question. Results are shown in Tables 2.5 and 2.6 and Figs. 2.11 and 2.12, from which two main conclusions can be drawn.

The first concerns the precision of the two types of IEC factors. For example, at Pd 363.470 nm (Pd) with interferents Mg and Zr, the IEC factor for the mixed interferent solution was –0.0036 and the sum of the individual IEC factors for these interferents was –0.0037 (Table 2.5 and Fig. 2.11). The errors (1 x σ) were $\pm$ 0.0003 and $\pm$ 0.0007 for the mixed interferent solution and summed IEC factor solutions respectively. However, some combinations yielded much worse results in terms of precision. Pt 265.945 nm with a mixed interferent solution of Ce and Zr gave an IEC factor of, -0.0014, and –0.0015 by summing the individual IEC factors but with a corresponding error of $\pm$ 0.0024 compared to $\pm$ 0.0001 (Table 2.5 and Fig. 2.11). This would be expected if one of the interferents had a particularly low signal at the analyte line of interest, and in this particular case the lower signal was due to Ce (240 cps) compared to Zr and Pt (360 and 3000 cps respectively). In comparison, for the mixed interferent solution, the low intensity signal was combined with the other interferent signals resulting in a higher overall signal which was less prone to error. Hence, it may be beneficial in cases where there are multiple interferents to calculate IEC factors with mixed solutions to compensate for this.

Table 2.5 IEC additivity and the effect on IEC factor precision.

Element Line (nm)	Mixed Sol'n (Mg + Ce 500 µg/ml) IEC Factor	SD	Mg IEC Factor	Ce IEC Factor	Summed IEC Factor	SD
Pt 193.700	0.0010	0.0005	-0.0002	0.0010	0.0008	0.0010
Pt 204.937	0.0005	0.0006	-0.0006	0.0009	0.0003	0.0008
Pt 265.945	-0.0070	0.0001	-0.0058	-0.0015	-0.0073	0.0003
Pd 248.892	-0.0007	0.0002	-0.0005	-0.0004	-0.0010	0.0005
Pd 363.470	0.0003	0.0000	-0.0001	0.0004	0.0003	0.0000
Rh 343.489	-0.0122	0.0001	0.0000	-0.0129	-0.0129	0.0037

Element Line (nm)	Mixed Sol'n (Ce + Zr 500 µg/ml) IEC Factor	SD	Ce IEC Factor	Zr IEC Factor	Summed IEC Factor	SD
Pt 193.700	0.0000	0.0005	0.0010	-0.0011	-0.0001	0.0000
Pt 204.937	0.0006	0.0009	0.0009	-0.0006	0.0003	0.0005
Pt 265.945	-0.0014	0.0001	-0.0015	0.0000	-0.0015	0.0024
Pd 248.892	-0.0004	0.0001	-0.0004	-0.0003	-0.0007	0.0005
Pd 363.470	-0.0032	0.0009	0.0004	-0.0036	-0.0032	0.0006
Rh 343.489	-0.0126	0.0001	-0.0129	0.0000	-0.0129	0.0131

Element Line (nm)	Mixed Sol'n (Mg + Zr 500 µg/ml) IEC Factor	SD	Mg IEC Factor	Zr IEC Factor	Summed IEC Factor	SD
Pt 193.700	-0.0013	0.0004	-0.0002	-0.0011	-0.0013	0.0015
Pt 204.937	0.0005	0.0007	-0.0006	-0.0006	-0.0012	0.0029
Pt 265.945	-0.0055	0.0001	-0.0058	0.0000	-0.0058	0.0094
Pd 248.892	-0.0005	0.0003	-0.0005	-0.0003	-0.0008	0.0007
Pd 363.470	-0.0036	0.0003	-0.0001	-0.0036	-0.0037	0.0007
Rh 343.489	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001

Element Line (nm)	Mixed Sol'n (Mg + Zr + Ce 500 µg/ml) IEC Factor	SD	Mg IEC Factor	Zr IEC Factor	Ce IEC Factor	Summed IEC Factor	SD
Pt 193.700	0.00001	0.00048	-0.00022	-0.00111	0.00103	-0.00030	0.00125
Pt 204.937	-0.00032	0.00227	-0.00056	-0.00061	0.00086	-0.00030	0.00244
Pt 265.945	-0.00668	0.00015	-0.00582	0.00004	-0.00149	-0.00727	0.00243
Pd 248.892	-0.00064	0.00022	-0.00052	-0.00029	-0.00045	-0.00127	0.00039
Pd 363.470	-0.00317	0.00203	-0.00013	-0.00361	0.00040	-0.00334	0.00007
Rh 343.489	-0.01205	0.00011	0.00004	0.00002	-0.01292	-0.01286	0.01362

SD Standard deviation

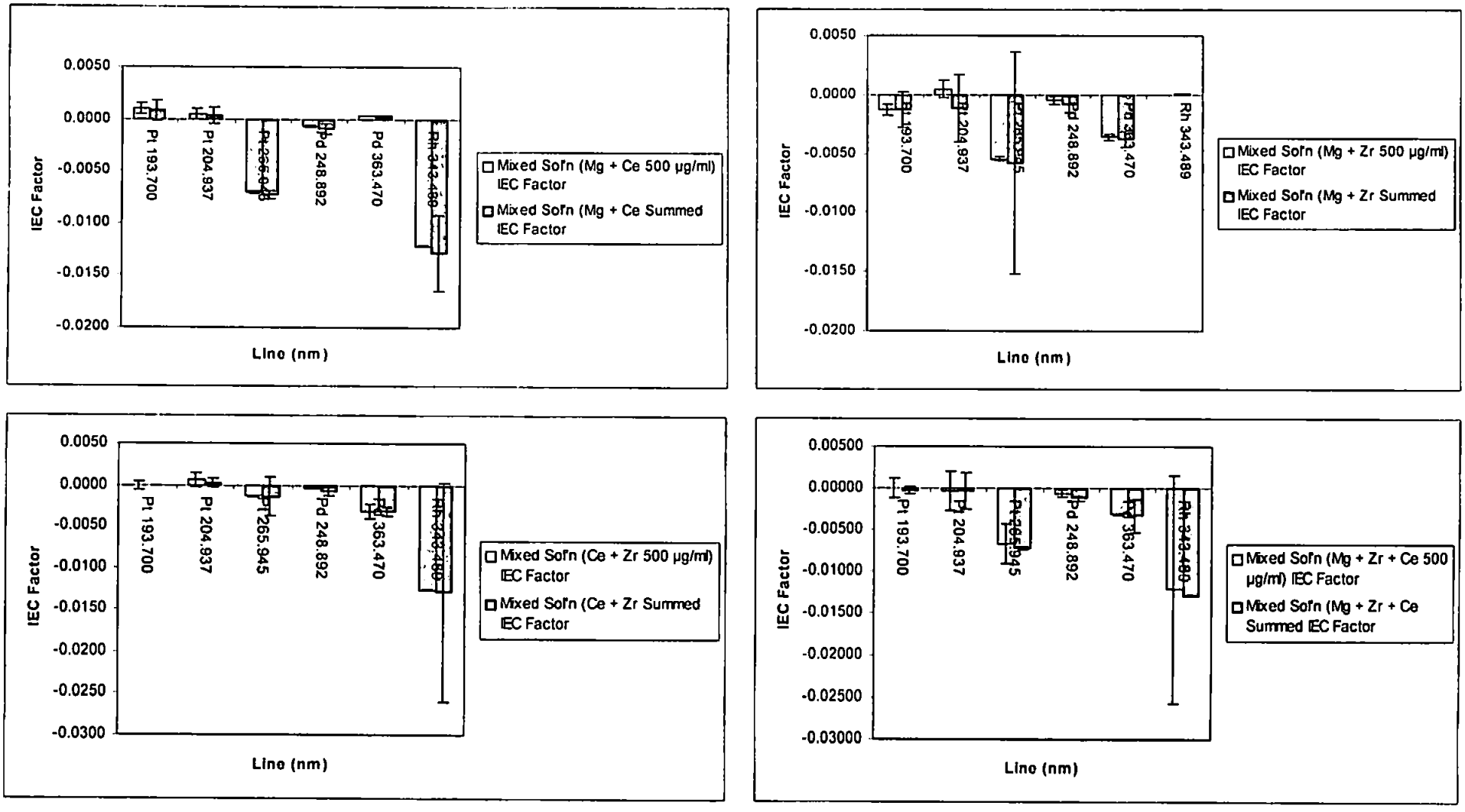


Figure 2.11 IEC additivity and precision

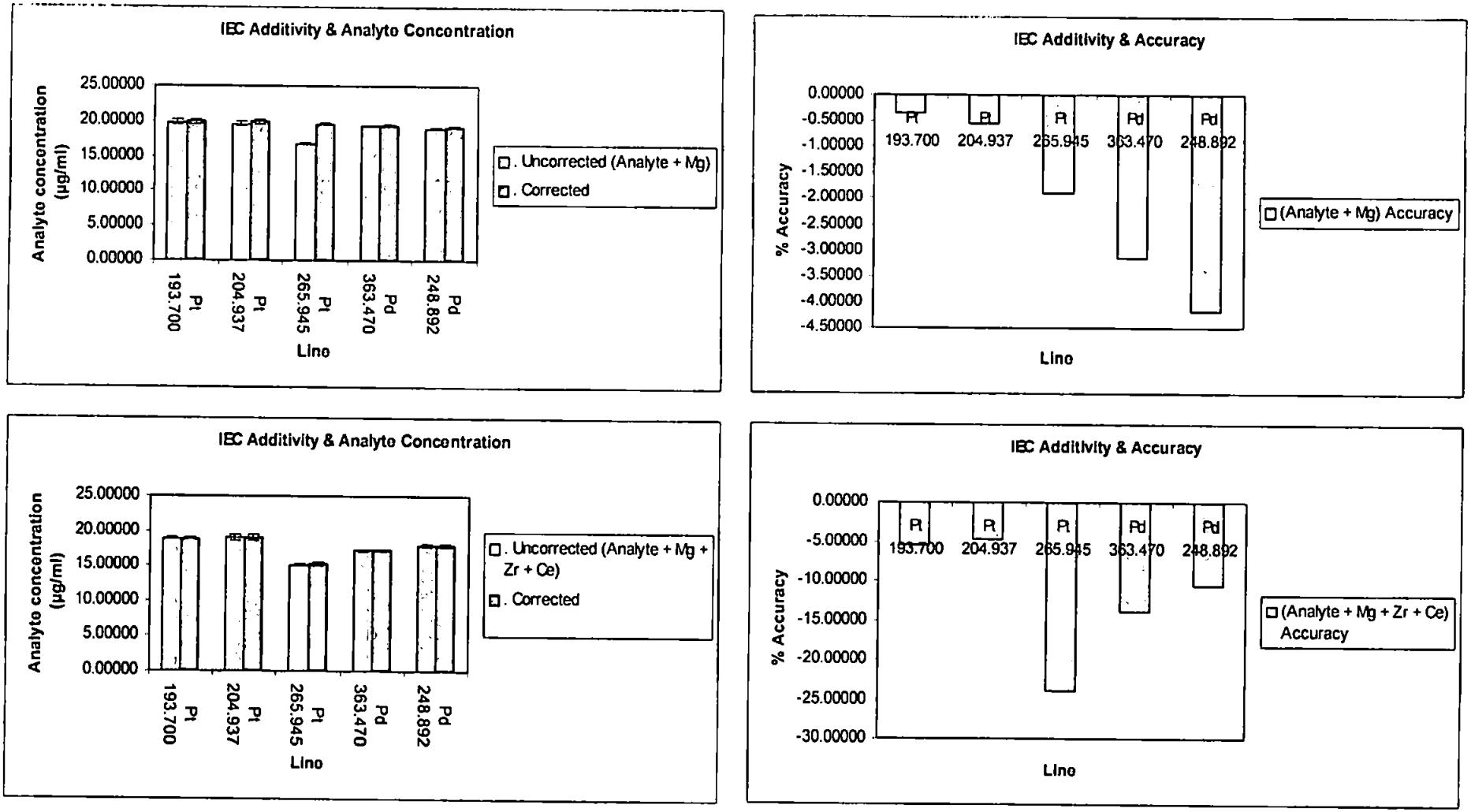


Figure 2.12 IEC additivity and analyte prediction accuracy

Table 2.6 IEC additivity and analyte prediction accuracy

	Pt + Mg				
Analyte lines	Actual concentration	Uncorrected concentration	Corrected concentration	Uncorrected % Accuracy	Corrected % Accuracy
Pt 193.700	20	19.82	19.93	-0.91	-0.36
Pt 204.937	20	19.61	19.89	-1.97	-0.56
Pt 265.945	20	16.71	19.62	-16.45	-1.89
Pd 363.470	20	19.30	19.37	-3.48	-3.16
Pd 248.892	20	18.90	19.17	-5.48	-4.17
	Pt + Mg + Zr +Ce				
	Actual concentration	Uncorrected concentration	Corrected concentration	Uncorrected % Accuracy	Corrected % Accuracy
Pt 193.700	20	18.91	18.91	-5.43	-5.43
Pt 204.937	20	19.05	19.06	-4.74	-4.71
Pt 265.945	20	15.12	15.23	-24.38	-23.87
Pd 363.470	20	17.16	17.22	-14.19	-13.92
Pd 248.892	20	17.87	17.88	-10.66	-10.61

The affects of IEC additivity, using mixed solution IEC factors, on the accuracy of predicted analyte concentrations are shown in Table 2.6 and Fig. 2.12. An inverse trend was observed where the % accuracy became worse with an increase in solution complexity. This was observed for all the analyte lines studied. For Pd at 248.892 nm with Mg as the only interferent the corrected predicted concentration % accuracy was – 4.2%, whereas in a solution with Mg, Ce, and Zr, the % accuracy fell to 10.66%. The largest decrease in accuracy was for Pt 265.945 nm from –1.89 % with only Mg as the only interferent to –23.87 % when Mg, Ce and Zr were the interferents. From Table 2.6 it would appear that as the number of interferents increases the ability of the method to eliminate the interferences decreases, i.e. because the BCPs were assigned using pure analyte solutions the decrease in accuracy may be due to the increased number of interferents which have altered the BCPs position. However, this effect may also be the result of matrix suppression which the use of IEC factors cannot compensate for, or a combination of both of the above.

2.4 Conclusions

For the synthetic samples investigated the interferences ranged from relatively simple background shifts which were easily corrected for, to more complex spectral overlaps which cannot be corrected for using IEC factors. Generally, the number and nature of interferents has been shown to affect the accuracy and precision of IEC factors, with accurate prediction only possible with the correct positioning of the BCPs, which requires *a priori* sample knowledge, or when matrix effects are minimal. The contribution to the analyte signal by a relatively small interference can lead to a degradation in precision of the IEC factor compared to the use of a mixed solution containing the same

interferents. The major disadvantage of this method of background correction, the positioning of the BCPs, has served to illustrate the complexity of PGM emission spectra and also the large range of intensities from several hundred to several thousand cps.

CHAPTER 3 - UNIVARIATE CALIBRATION AND PRELIMINARY COMPARISON WITH PARTIAL LEAST SQUARES

3.1 Introduction

Inductively Coupled Plasma Atomic Emission Spectrometry (ICP-AES) is now well established as a powerful technique for multielement analysis(4), but can suffer from both spectral and non-spectral interferences which limit the accuracy, repeatability and reproducibility of the information obtained. The nature of the interference is often complex and it is not always possible to apply the required corrections in order to achieve accurate and precise analysis. In contrast, the information obtained by molecular spectroscopic techniques has been greatly enhanced by the application of data handling tools by the removal of redundant information and the concurrent decrease in model bias (58-60, 78, 86, 87, 111-114).

Traditionally, for quantitative analysis in atomic spectroscopy, a single spectral line is chosen, based upon the criteria of line sensitivity and freedom from spectral interferences. Many attempts have been made to correct for spectral interferences in ICP-AES, including standard additions, matrix matching, inter-element correction, and optimisation of line selection. However, these methods suffer from serious limitations when a sample with a complex matrix is presented for analysis. For example, the use of interelement correction factors requires interferent lines that can be used in a univariate fashion to determine the concentration of the interferents themselves, which is not always possible with complex samples.

The use of chemometric approaches to correct for interferences in ICP-AES has emerged as an attractive alternative and various multivariate calibration techniques have

been proposed. In the mid 1960s, computerised instrumental methods of chemical analysis began to generate very large amounts of data. This gave rise to the so-called "data explosion" in analytical chemistry. Previous to this, chemists had based their decisions on only a few, hard-won, expensive pieces of data. Now they had to base their decisions on vast amounts of easily obtained, inexpensive data. In an attempt to make sense of all of this data, and extract only useful, relevant information, chemists began to use increasingly sophisticated mathematical and statistical techniques borrowed from other disciplines.

Chemometrics can be defined in the following manner:

Chemometrics is the chemical discipline that uses mathematical and statistical methods,

- *to design or select optimal measurement procedures and experiments, and*
- *to provide maximum chemical information by analysing chemical data* (115).

The use of chemometrics in mainstream analytical chemistry first became apparent in the early 1970s with the introduction of a number of new mathematical techniques, such as pattern recognition and multivariate statistics (116). The term *Chemometrics* was proposed by the Swedish physical organic chemist Svante Wold and the American analytical chemist Bruce R. Kowalski, as a generic name for the discipline of chemistry in which mathematical and statistical techniques are used for the purposes of optimising experimental design and maximising the information obtainable from analytical data (117, 118). Two years later Wold formed the Chemometrics society in association with Bruce R. Kowalski, in order to provide an international forum for chemists applying formal logic to chemical analysis. Since this time, chemometrics has expanded into a very prominent area of chemical research with two specialist journals (*Journal of Chemometrics*, Elsevier;

Chemometrics and Intelligent Laboratory Systems, Wiley) dedicated to the progression of this field.

One of the most important applications in the field of analytical chemistry is multivariate calibration (48, 119), which can be applied to the quantification of single or multiple analytes when more than one data point is collected for each variable (i.e. multivariate data).This is particularly appropriate in the case of multiwavelength spectroscopic techniques, such as ICP-AES.

In this study multivariate calibration is compared to traditional univariate calibration with interelement correction and matrix matching for the determination of platinum group metals in autocatalyst digests. The multivariate models were built using intensity data for 248 lines which included the most intense Pt, Pd and Rh lines and many of the most intense matrix lines as opposed to using the full spectrum which is dealt with in a later section.

One requirement of multivariate methods is that the factor space defined by the multi-element standards used for model calibration encompasses all likely constituents (analytes and interferents) and concentrations of the real sample matrices. To ensure that this requirement is met it is necessary to acquire data for the multivariate calibration model using an appropriate experimental design, such as orthogonal arrays, or historical data for which accurate analyte concentrations are available.

3.1.1 Orthogonal Arrays and Fractional Factorial Experimental Designs

In mixture experiments it is essential that the elements are uniformly distributed over the 'mixture space', thereby giving a balanced element concentration distribution (composition), a condition which is often overlooked resulting in poor predictions and

unbalanced designs. This is why features such as orthogonality (element vectors at right angles to each other, Fig. 3.1) are extremely important if a good model is to be achieved. When two factors (relating to the concentration of two compounds) have a correlation coefficient of 0 this is equivalent to stating that both span each other's mixture space evenly (Fig. 3.1), and hence that the design is balanced and predictions will depend only on the distance from the centre of the design (120).

The problem of constructing such a multilevel orthogonal design involves the theory of orthogonal arrays (120). According to this theory, if l is the number of levels, in order to produce an orthogonal design (e.g. $l = 5$) the minimum number of experiments (N) is equal to $N=l^2$ (i.e. 25) and the maximum number of factors (elements) permitted is N-1 (i.e. 24), but generally only (N-1)/(l-1) (i.e.6) of the N calibration experiments are mutually orthogonal. Adherence to the mathematical constraints of the orthogonal array design consisting of 49 experiments, 7 levels, and 8 factors used later in this section are shown in Table 3.1. Although this is true for two to four-level designs, special circumstances allow up to 12 factors to be mutually orthogonal for five-level designs. It is therefore possible to construct a 25x12 design matrix in which each pair of columns has a correlation coefficient equal to 0.

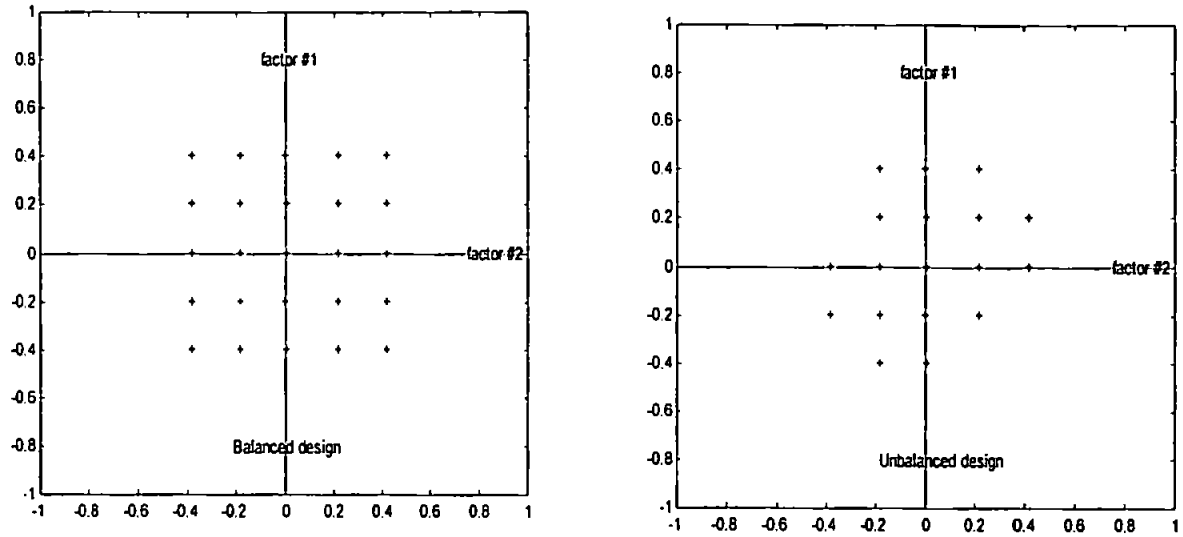


Figure 3.1 Example of a 5 level (-0.4 ,-0.2, 0, 0.2, 0.4) balanced and unbalanced design. The factors in the balanced design are mutually orthogonal $\Rightarrow$ perpendicular $\Rightarrow$ their correlation coefficient is zero equivalent to stating that the factors span each others mixture space evenly. Note that in the case of the unbalanced design the correlation coefficient $\neq 0$, a pattern in the data exists.

Table 3.1 Number of mutually orthogonal factors possible with 25 and 49 experiments

	No. of experiments ($N = l^2$)	
	25 experiments	49 experiments
No. of mutually orthogonal factors	$25 = 5^2$	$49 = 7^2$
1.$(N - 1)/(l - 1)$ up to $l = 4$	25 -1 / 5 – 1 = 6 (12 for $l > 4$)	49 - 1 / 7 - 1 = 8
2. (N-1) not all are mutually orthogonal (1. Plackett-Burman methodology) (2. Methodology according to Brereton (120))	24	49 - 1= 48 (16 factors mutually orthogonal)

In order to construct an orthogonal array several steps are required. A generalised treatment is given below. The five levels will be represented by -0.4 ,-0.2, 0, 0.2 and 0.4 respectively, representing five equally spaced concentrations, the central one coded by the figure 0. Twenty five experiments are necessary and up to 24 possible factors can be included in the design.

The first step is to select the so called repeater level, this is preferably the central level which allows more than six mutually orthogonal factors. This is the level at which the first calibration experiment is performed for all the factors (elements). The next step is to produce a cyclic generator for the remaining four levels. One possible generator can be of the form $-2 \rightarrow -1 \rightarrow 2 \rightarrow 1 \rightarrow -2$, if the repeater level is 0 as is illustrated in Figure 3.2. Thus if we denote the levels *a*, *b*, *c*, *d* the cyclic generator converts *a* to *b*, *b* to *c*, *c* to *d* and *d* to *a*. The key is to generate the first column of the design matrix, after which all other columns can be generated. The 25 experiments in the first column can be divided into four sets of five experiments and five unique experiments (at the level of the repeater) as illustrated in Fig. 3.2.

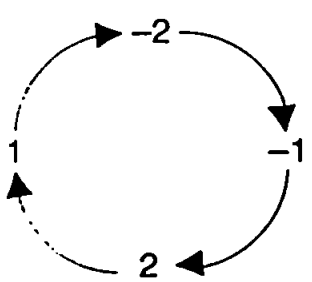


Fig. 3 Possible cyclic generator for a five-level design.

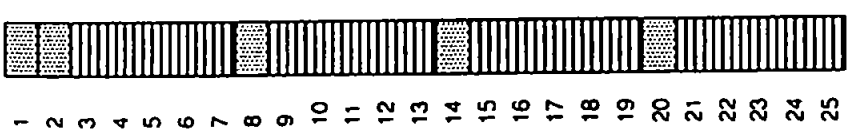


Figure 3.2 Arrangement of experiments in the first column of a five-level design.

The four (l-1) blocks of five experiments are determined by the first block of five experiments as follows. The levels in each block are cyclically shifted from the previous block by one (clockwise), using the cyclic generator (e.g. level –0.4 is shifted to level –0.2). This cycling ensures that each level is represented five times over the 25 experiments. So that each individual element is measured at each of the five concentration levels five times, an important prerequisite for an orthogonal design. Thus it is possible to determine the overall design. The next, and most difficult aspect, is to establish which combinations of levels are possible for the first block of five experiments (starting vector) in column 1 and how to choose a feasible cyclic generator, both of which are described in detail by Brereton (120). As a result, there are only two possible cyclic generators that are capable of generating twelve uncorrected factors: $-2 \rightarrow -1 \rightarrow 2 \rightarrow 1 \rightarrow 2$ and $-2 \rightarrow 1 \rightarrow 2 \rightarrow -1 \rightarrow -2$ and three possible starting vectors: [-2, -2, 2, -1, 2]; [-2,-1, -2, 2, 2] and [-2,2, 2, 1, 2].

The major advantage of the orthogonal array is cost efficiency. In the above example only 25 calibration experiments are required to study 12 factors at 5 levels. A full factorial design studying the same numbers of factors and levels would require 244×10^6 experiments, (experiments = l^f). Generally, full factorial designs are only useful where the number of factors is relatively limited (121). Such arrays, therefore, allow the mixture space to be covered in an orthogonal manner, with a minimum number of runs in the experiment. Latin square, 2**(k-p) fractional factorial, Plackett-Burman (in particular), and Box-Behnken designs are also aimed at accomplishing this goal. In fact, many of the standard orthogonal arrays tabulated by Taguchi (122) are identical to many of the more established designs, such as the fractional two-level factorial and Plackett-Burman designs. However, many of these well established designs are available at only two levels and in multivariate calibration especially several concentration levels must be used.

3.2 Experimental

3.2.1 Instrumentation and Reagents

All data were collected using a simultaneous echelle inductively coupled plasma atomic emission spectrometer (Perkin-Elmer Optima 3000 ICP, Norwalk, USA) equipped with a segmented-charge-coupled array detection system (123, 124). Instrumental operating conditions were optimised using simplex optimisation and are given in Table 2.1 Chapter 2., section 2.3.1.

Single and multielement solutions were prepared by serial dilution of ultra-pure stock standards (10,000 and 1000 μg ml^{-1}, Johnson Matthey plc, Royston, Hertfordshire). Water was double deionised (18 MΩ quality) and acids were of Aristar grade (Merk-BDH, Poole, Dorset). All glassware was acid washed in 10 % v/v nitric acid for 24 hours then rinsed thoroughly with 18 MΩ water. All plasticware was metal-free high-density polypropylene (Anachem, Luton, Bedfordshire). Calibration and test solutions containing varying concentrations of Pt, Pd, Rh, Al, Mg, Ce, Zr and Ba, plus the internal standards In, Sc and Y were prepared from the stock solutions and stored in high-density polypropylene tubes. Digests of autocatalyst samples (Johnson Matthey plc, Royston, Hertfordshire) which had been validated by comparison with a NiS fire assay, were used for method validation.

3.2.2 Procedure

3.2.2.1 Univariate Calibration

Univariate calibration was performed by two point calibration using a zero standard containing 10% v/v aqua regia (3:1 $HCl:HNO_3$) and a single element calibration solution for each analyte at the concentrations expected in the samples (i.e. Pt 50 μg ml^{-1}, Pd 50 μg ml^{-1}, Rh 10 μg ml^{-1}. A calibration graph was prepared using the net integrated peak areas (obtained using 2 point background correction) for the analyte lines (Pt 214.423 nm, Pd 248.892 nm and Rh 343.489 nm) and the concentrations of the analytes in the samples were determined by interpolation.

3.2.2.2 Interelement Correction (IEC)

Interelement correction factors for each of the suspected interfering elements were determined. A calibration graph was first constructed in the usual way, and solutions of the suspected interferents, (1000 μg ml^{-1}), were aspirated while monitoring the analyte line(s) of interest. The apparent analyte concentration was then determined at the line of interest, and an IEC factor calculated by dividing the apparent analyte concentration at that particular wavelength by the interferent concentration. Eqn. 3.1 was then applied to obtain the corrected analyte concentration.

$$C_c = C_u - \sum_{i=1}^{i} (F_i \bullet I_{ic}) \qquad \text{Eqn. 3.1}$$

Where C_u, C_c, F_i and I_{ic} are the uncorrected analyte concentration, corrected analyte concentration, IEC factor, and interferent concentration respectively for each interfering element, *i*.

3.2.2.3 Matrix Matched Calibration

Calibration was carried out using a zero standard containing 10% v/v aqua regia (3:1 $HCl:HNO_3$) and In, Sc and Y at concentrations of 50, 25 and 25 μg ml^{-1} and with Al and Mg at 1000 and 500 μg ml^{-1} respectively. A multi-element calibration solution containing each element at the concentrations expected in the samples (i.e. Pt 40 μg ml^{-1}, Pd 40 μg ml^{-1}, Rh 10 μg ml^{-1}, Al 1000 μg ml^{-1}, Mg 500 μg ml^{-1}, Ce 300 μg ml^{-1}, Zr 100 μg ml^{-1}, Ba 25 μg ml^{-1}, and the internal standards In 50 μg ml^{-1}, Sc 30 μg ml^{-1} and Y 30 μg ml^{-1}) was also used. The internal standards were included in the solutions but were subsequently not used in any of the calibrations due interferences form other elements. The calibration standard contained both the analyte and matrix elements at the highest concentrations expected in the autocatalyst matrix. The composition of these matrix-matched standards is given in Table 3.2.

3.2.2.4 Multivariate Calibration

Prior to ICP-AES analysis all calibration and independent test samples were randomised to prevent any instrumental bias from affecting subsequent models. Several multivariate-modelling algorithms were studied, including Principal Components Regression, Partial Least Squares 1 and 2, and Multiple Linear Regression Analysis, using Matlab Software Version 5.0, and the PLS_Toolbox 2.0 (Mathworks Inc.). Of these four techniques PLS1 gave the best predictive results and only those are reported here. All data were mean centred, and full-set random cross validation was used throughout this work to minimise systematic error in forming the models.

The root mean square of cross-validation (RRMSECV, the equation for which is given in Eqn. 3.2) is a measure of the models ability to predict new samples and consists of

leaving out one calibration sample from the calibration set, building the model on the remaining calibration samples, then predicting the value for the left out calibration sample and computing the prediction residuals. The process is repeated until all calibration samples have been left out once; then all prediction residuals are combined to compute the root mean square of cross-validation. To assess the model fit to the calibration data the root mean square error of calibration (RMSEC) value was used. The format is the same as that used in Eqn. 3.2 except that $\hat{y}$ are the values of the predicted variables when all calibration samples are left included in the model.

$$RMSECV = \sqrt{\frac{\Sigma(\hat{y}_i - y_i)^2}{N}} \qquad \text{Eqn. 3.2}$$

Table 3.2 Concentrations (μg ml^{-1}) of the standards for the matrix matched calibration.

Mixed calibration standard	Zero standard	Calibration standard
Pt[a]	0	40
Pd[a]	0	40
Rh[a]	0	10
Ba[b]	0	25
Ce[b]	0	300
Zr[b]	0	100
In[c]	50	50
Sc[c]	25	25
Y[c]	25	25
Al[b]	1000	1000
Mg[b]	500	500

[a]analyte, [b]matrix element, [c]internal standard

3.2.2.5 Estimation of Errors

The errors of prediction relative to the known values, for individual test solutions and autocatalyst samples, were compared using relative standard error (RSE) values, calculated as shown in Eqn. 3.3:

$$RSE(\%) = 100 \times \frac{(\hat{y}_i - y_i)}{y_i} \qquad \text{Eqn. 3.3}$$

The overall efficacy of the different calibration methods was compared using the relative root mean square error (RRMSE), defined in Eqn. 3.4, which gives a general idea of the error of prediction for a range of concentrations:

$$RRMSE(\%) = 100 \times \frac{1}{mean(y)} \sqrt{\frac{\sum (\hat{y}_i - y_i)^2}{N}} \qquad \text{Eqn. 3.4}$$

where y is the known concentration, $\hat{y}$ is the predicted concentration, and N is the number of experiments.

3.2.2.6 Experimental Design

In a working laboratory it is desirable to maximise the time spent analysing samples compared to the calibration step. Traditionally, multivariate calibration datasets have been acquired using experimental designs based on a factorial or partial factorial approach. However, for the 8-factor problem studied here, where 8 elements must be uniformly distributed over an 8-dimensional mixture space (Fig.1 for 2-dimensional mixture space), such approaches would result in an impracticably large number of experiments if more than a few levels were used (e.g. 65,536 for a design with 4 concentration levels and 8 factors (4^8)). Hence, the calibration set for multivariate analysis in this work was prepared using a Taguchi orthogonal array design (122, 125, 126) in order to cover the required factor space with the minimum number of experiments. The concentration ranges of the elements were determined from historical data on the composition of autocatalyst digest samples. The orthogonal array contained 8 factors at 7 levels with a total of 49 experiments, represented as $OA_{49}(7^8)$ [cf. 5,764,801 experiments for a full factorial design]. The levels and factors in the design are shown in Table 3.3.

Table 3.3 Concentrations levels (μg ml^{-1}) and element factors in the orthogonal array design

	Level						
Factor	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Pt	0	5	10	20	30	40	50
Pd	0	5	10	20	30	40	50
Rh	0	1	2	4	6	8	10
Ba	0	1	5	10	50	100	200
Ce	0	1	10	50	100	300	500
Zr	0	1	10	50	100	300	500
Mg	0	1	10	50	100	300	500
Al	0	1	10	100	200	500	1000

3.3 Results and Discussion

Individual RSE values for the test solutions are shown in Fig. 3.3 a-c, and RRMSE values are shown in Table 3.4. Individual RSE values for the autocatalyst digests are shown in Fig, 3.4 a-c.

3.3.1 Univariate Calibration

When univariate calibration was used the RRMSEs of prediction for the test solutions for Pt, Pd and Rh were 19.22, 15.23, and 54.09 % respectively (Table 3.4). The high errors for Rh were probably due to matrix induced suppression or a combination of matrix-induced suppression and spectral interference. There were no direct spectral overlap interferences on the Rh 343.489 nm line which would cause this high error but there was a baseline shift due to the presence of Zr (Fig. 3.5c). Net intensity data were calculated by baseline subtraction, using background correction points (BCPs) positioned either side of the line, so the baseline shift should be accounted for, however, the positioning of the

BCPs for Rh 343.489 was hampered by the proximity of the Ce 343.521 nm line (Fig. 3.5c), resulting in failure of the background correction routine. In addition, the average Rh concentration in the test samples was only 3.8 $\mu g\ ml^{-1}$ compared with 15.6 and 15.2 $\mu g\ ml^{-1}$ for Pt and Pd respectively (Table 3.5) giving a lower spectral response. Coupled with the incorrect positioning of the BCPs present at 343.489 nm, the relatively low Rh signal has resulted in a much larger predictive error (Fig. 3.3) compared with Pt and Pd.

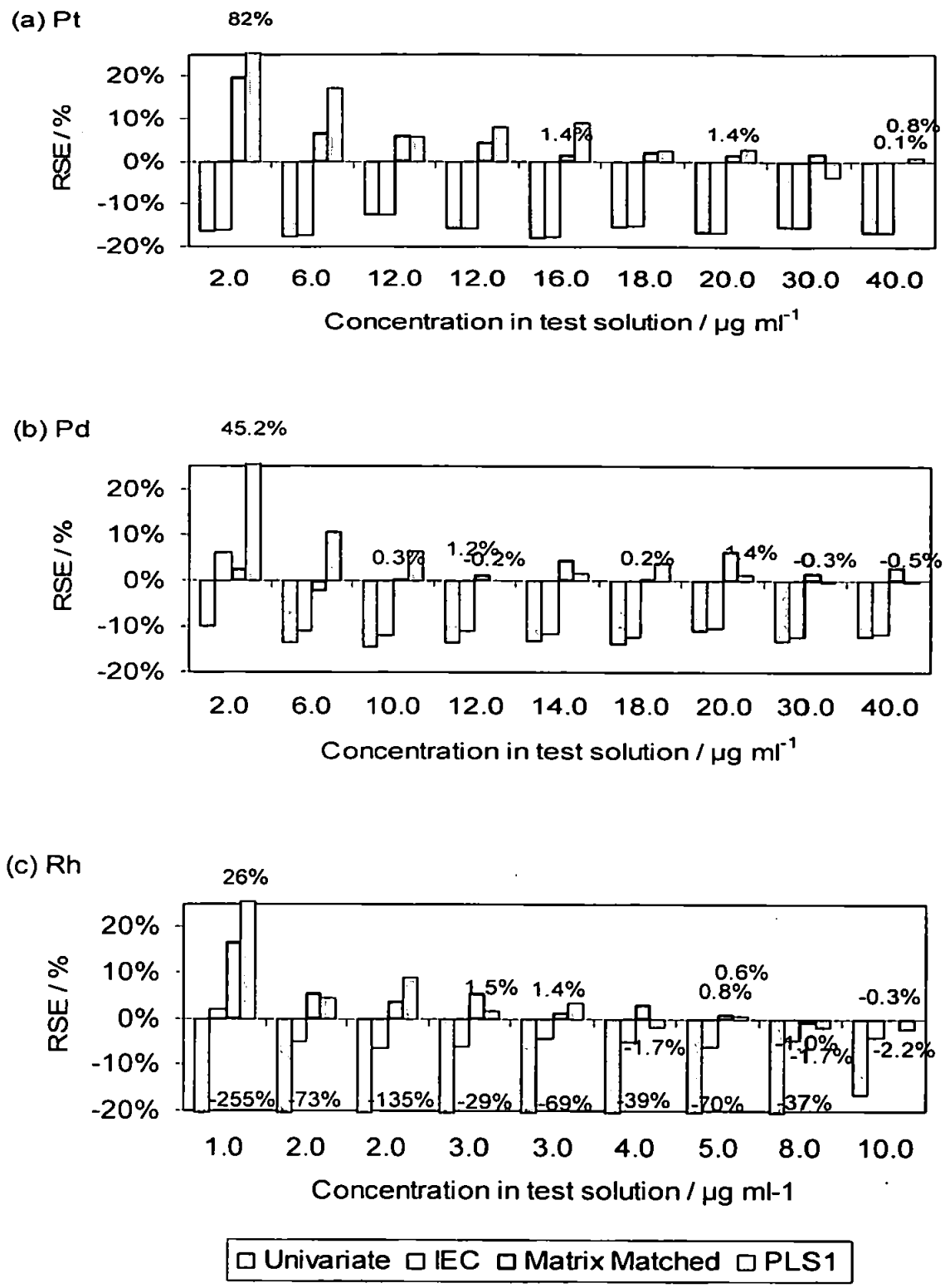


Figure 3.3 Comparison of RSE of prediction obtained using univariate, univariate with IEC, matrix matched and PLS1 for: (a) Pt; (b) Pd; (c) Rh in synthetic test samples

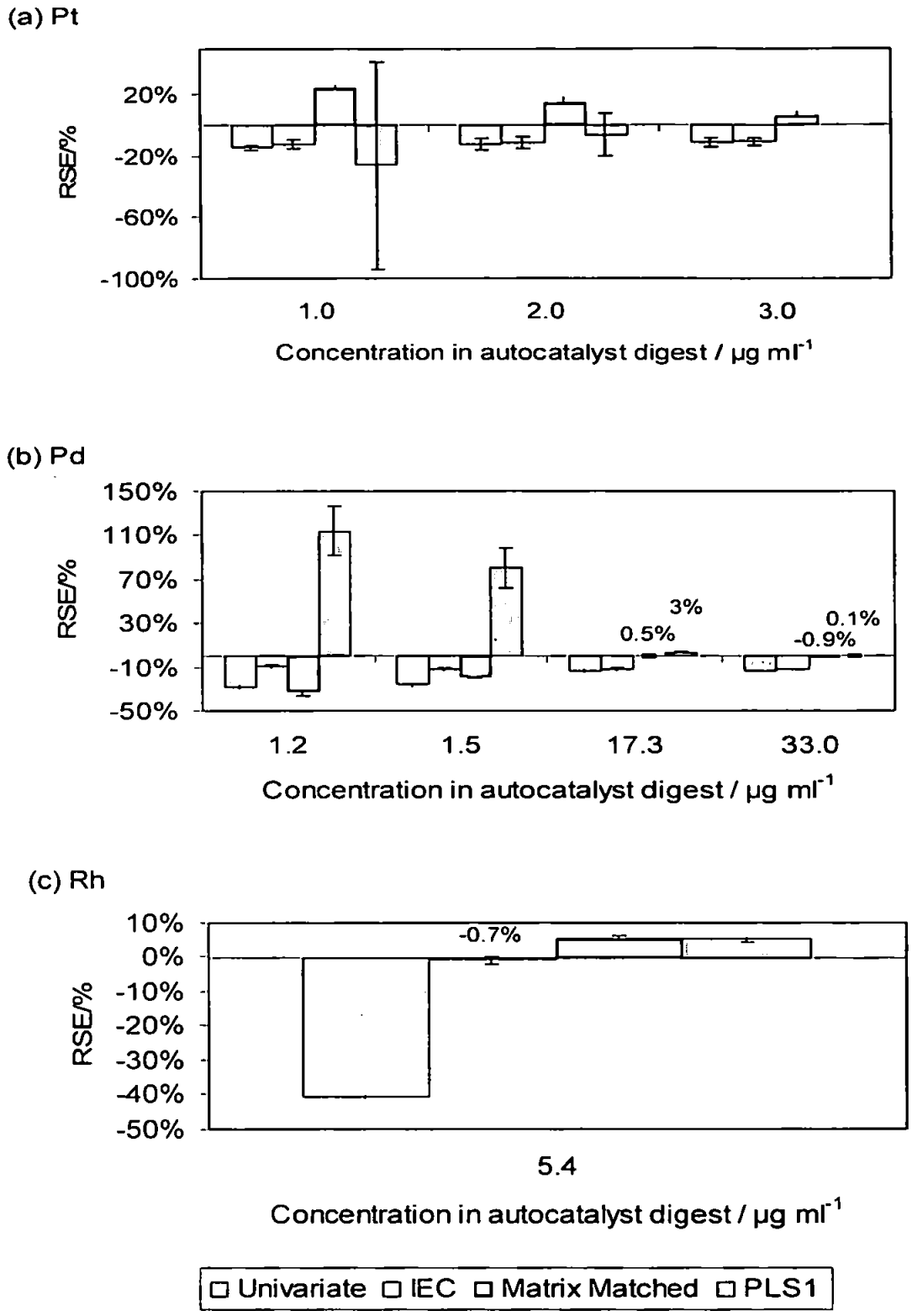


Figure 3.4 Comparison of RSE prediction obtained using univariate, univariate with IEC, matrix matched and PLS1 calibration for: (a) Pt; (b) Pd; (c) Rh in autocatalyst samples. Each result is the mean of three replicate analyses, and the error bars represent ± σ

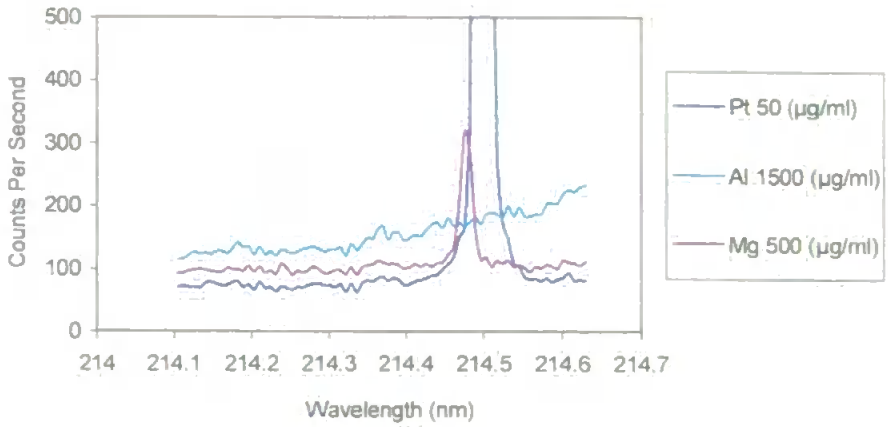
Table 3.4 RRMSE of prediction for the concentration of Pt, Pd and Rh in the synthetic test samples and autocatalyst digests.

	Test solutions			Autocatalyst digests		
Calibration method	Pt	Pd	Rh	Pt	Pd	Rh
Univariate	19.22	15.23	54.09	13.47	18.54	87.69
Univariate with IEC	19.17	13.97	5.75	12.48	16.95	2.08
Matrix matched	2.4	3.67	2.44	12.03	2.39	8.03
PLS1	5.77	2.96	3.46	32.38	7.44	75.33

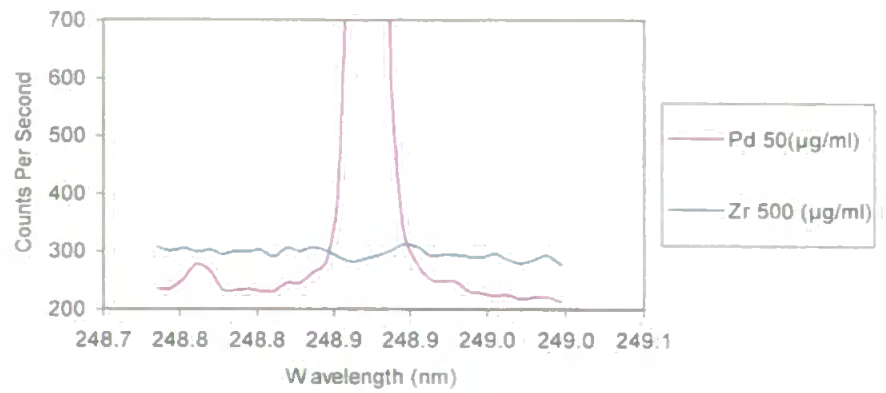
Table 3.5 known and predicted concentrations ($\mu g\ ml^{-1}$) of the test solutions after the application of univariate calibration on net signal intensity obtained using 2-point background correction.

	Known			**Predicted**		
Test solution	Pt 214.423	Pd 248.892	Rh 343.489	Pt 214.423	Pd 248.892	Rh 343.489
Te1	12	20	3	10.48	17.82	2.14
Te2	16	12	5	13.13	10.38	1.52
Te3	20	18	2	16.66	15.49	-0.69
Te4	12	14	4	10.1	12.16	2.45
Te5	18	10	3	15.24	8.54	0.93
Te6	6	30	1	4.94	26.04	-1.55
Te7	2	6	2	1.67	5.18	0.55
Te8	40	2	8	33.32	1.8	5.03
Te9	30	40	10	25.41	35.1	8.34
Te10	0	0	0	-0.01	-0.19	-1.33

(a)



(b)



(c)

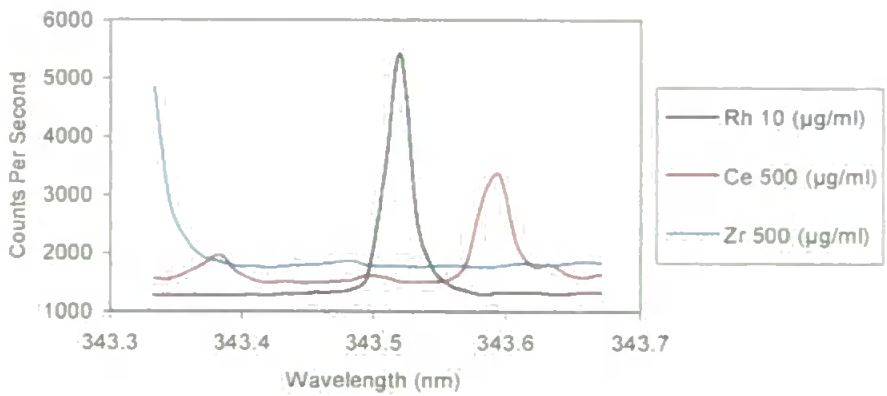


Figure 3.5 Interferences on the: (a) Pt 214.423 nm; (b) Pd 248.892 nm; (c) Rh 343.489 nm lines

Predicted concentrations for Pt and Pd were much closer to the known values and there was no trend in the relative error with concentration (Fig. 3.3a-b), despite the presence of some small spectral interference and baseline shift (Figs 3.5a and 3.5b). For example, even at a relatively high Mg concentration of 500 μg ml^{-1} the spectral interference was small relative to the peak for Pt (Fig. 3.5a), and the baseline shift caused by 500 μg ml^{-1} of Zr was relatively minor at the Pd 248.892 line (Fig. 3.5b).

The RSEs for the prediction of Pt, Pd and Rh in the autocatalyst samples are shown in Fig. 3.4a-c. Each point represents the average of three replicate sample digests corrected for mass of sample. Concentrations of Pt, Pd and Rh were predicted to be lower than their actual values. The RRMSEs for Pt, Pd and Rh were 13.47, 18.54 and 87.69 % respectively, again confirming that Rh was not predicted as well as the other analytes.

3.3.2 Interelement Correction (IEC)

The magnitude of the IEC factors varied considerably from -5.5×10^{-6} to 9.9×10^{-3} (Table 3.6). For the prediction of analyte concentrations in the synthetic test solutions the application of IEC generally resulted in an improvement in the accuracy of prediction. The greatest improvement was observed for Rh, with the RRMSE falling from 54.09 to 5.75 % when IEC was applied (Table 3.4). The RRMSE values for Pt and Pd did not change significantly with values of 19.17 % and 13.97 % respectively after the application of IEC factors (Table 3.4), which is also reflected in the magnitude of the RSE values for the individual solutions (Fig 3.3a-b). Results for the autocatalyst samples followed the same trend where the RRMSEs for Pt, Pd and Rh were 12.48, 16.95 and 2.08 % respectively (Table 3.4). Hence, IEC correction had a significant effect for the correction of spectral interferences on Rh at low concentrations in this instance.

Table 3.6 Magnitude of the IEC factors used to correct for spectroscopic interferences

Interfering element	IEC factor at analyte line (nm)		
	Pt 214.423	Pd 248.892	Rh 343.489
Pt	-	9.86×10^{-3}	-2.23×10^{-6}
Pd	3.67×10^{-4}	-	5.45×10^{-5}
Rh	-7.15×10^{-5}	-6.52×10^{-4}	-
Al	-1.47×10^{-5}	-8.32×10^{-5}	3.18×10^{-5}
Mg	6.38×10^{-6}	-3.66×10^{-4}	2.59×10^{-5}
Ce	1.08×10^{-5}	-2.3×10^{-4}	-6.68×10^{-3}
Zr	-4.16×10^{-5}	-8.75×10^{-5}	8.03×10^{-6}
Ba	-3.26×10^{-5}	-1.37×10^{-4}	1.67×10^{-5}
In	-1.16×10^{-5}	-7.23×10^{-5}	-5.53×10^{-6}
Sc	-6.78×10^{-5}	-5.64×10^{-4}	4.75×10^{-4}
Y	-4.22×10^{-6}	-5.06×10^{-5}	-1.33×10^{-3}
Sr	-2.95×10^{-4}	-1.08×10^{-4}	1.02×10^{-5}

3.3.3 Matrix Matched Calibration

Results obtained using a matrix matched standard are shown Figs 3.3a-c and Fig. 3.4a-c for the test solutions and autocatalyst samples respectively. Overall, this was the most accurate of the univariate calibration methods for the prediction of Pt and Pd in the test solutions, and was comparable to, or better than, IEC correction for Rh for all but the lowest concentration (Fig. 3.3). This overall improvement in accuracy of prediction for the test solutions is reflected in RRMSE values of 2.40, 3.67 and 2.44 % for Pt, Pd and Rh respectively (Table 3.4).

For the autocatalyst samples the RRMSEs for Pt, Pd and Rh were 12.03, 2.39 and 8.03 % respectively (Table 3.4). Evidently, the accuracy of prediction for Pt and Pd in the autocatalyst samples was not as good as for the test solutions, which was probably due to the concentrations of the matrix matched elements not being exactly the same as their concentrations in the autocatalyst samples. Matrix-induced suppression of the analyte signal will only be effectively corrected for by matrix matching the standards to the

samples. The relatively low RRMSEs for all three analytes, achieved when matrix matching was used, suggests that matrix-induced suppression had a greater effect on the accuracy of the results than spectral interferences at the lines chosen for analysis. The main disadvantage of matrix matching is that the matrix matched standards contain matrix elements at fixed concentrations, whereas the concentrations in the samples may vary considerably. This is demonstrated in Fig. 3.6, which shows the RSEs obtained for Rh in the individual test solutions when the concentrations of Ce and Zr were reduced by a factor of ten, to 30 and 10 μg ml^{-1} respectively. As can be seen, the RSEs increased considerably at all concentrations when incorrect matrix matching was employed, and a comparison of Figs. 3.6 and 3.3c reveals that the results were very similar to those obtained when non-matrix matched univariate calibration was used.

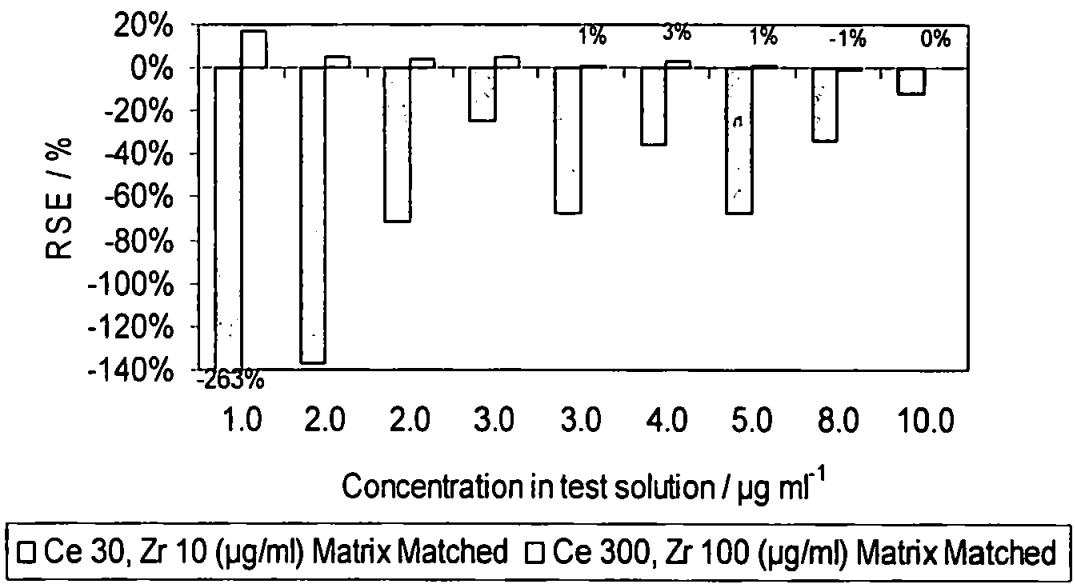


Figure 3.6 Effect on the RSE for Rh in the test samples when matrix matching is applied with different concentrations of the main interferents (Ce and Zr) on the line chosen (Rh 343.489 nm)

3.3.4 Multivariate Calibration: Partial Least Squares 1 (PLS1)

Multivariate calibration was initially performed using net signal intensities. In order to obtain the net signal from the instrument a 2-point, or 1-point where appropriate, background correction was performed by the instrument software, however, the selection of optimum background correction points for all 248 lines proved problematical due to the presence of adjacent spectroscopic interferences. Hence, it was decided to model the data using gross integrated as well as net integrated signal intensities. The definition of gross integrated intensity being that area under the peak profile curve, defined by five points, without the use of BCPs. This, therefore, includes any area between the peak profile (defined by the five points) and the baseline due to any sample/standard matrix differences. All data were translated along the co-ordinate origin by mean-centering each variable. Autoscaling was also tried but gave slightly worse results, probably due to the noise being scaled equally with the informative data. Results obtained using net integrated signal intensity data were worse than those obtained when using gross integrated signal data most probably due to the incorrect assignment of BCPs. It appears therefore, that any non-linear 'shifts' in the plasma continuum (gross signal intensity) giving a non-linear overestimation of signal intensity had less effect than incorrect positioning of the BCPs. As such only results for gross integrated line intensity are presented here. The number of optimal PCs was chosen using cross-validation (Figs. 3.7-3.9 for Pt, Pd and Rh respectively).

Results for the model validation are shown in Table 3.7, indicating that the model for Pd had the lowest error (i.e. lowest RMSECV and RMSEC values). The RRMSE values for Pt, Pd and Rh in the test solutions were 5.77, 2.96 and 3.46 % respectively (Table 3.4), indicating that, overall, this calibration strategy was as good as matrix matching.

Table 3.7 RMSEC and RMSECV values for the PLS1-model calibration data set constructed using gross mean centred data

Analyte (PCs)	RMSEC	RMSECV
Pt (9)	0.77	1.05
Pd (6)	0.40	0.48
Rh (6)	0.12	0.24

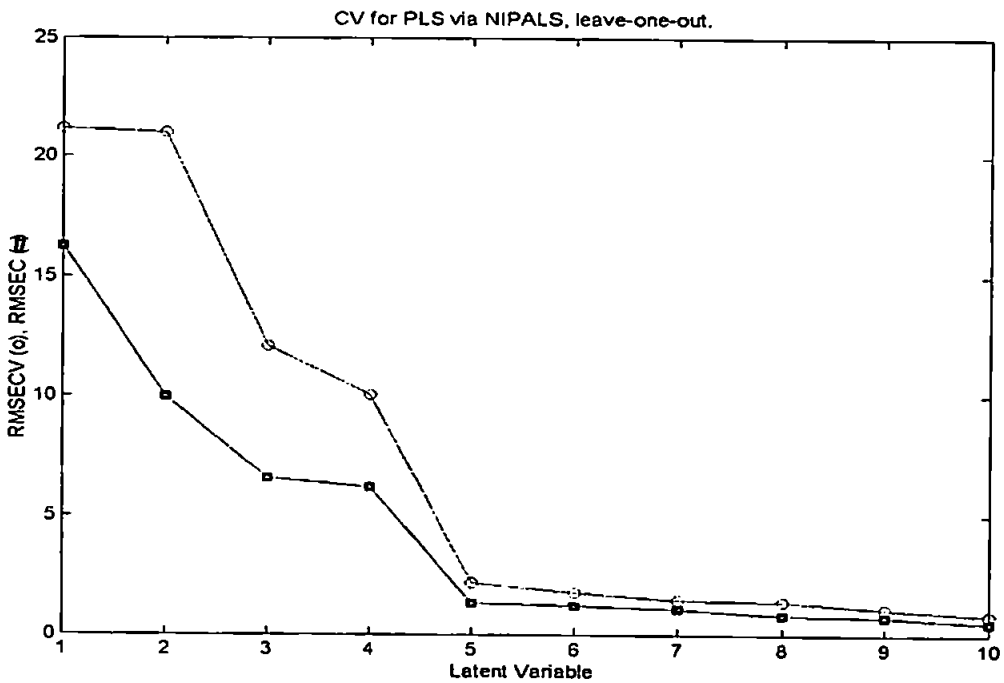


Figure 3.7 Cross validation plot for Pt (RMSEC included also)

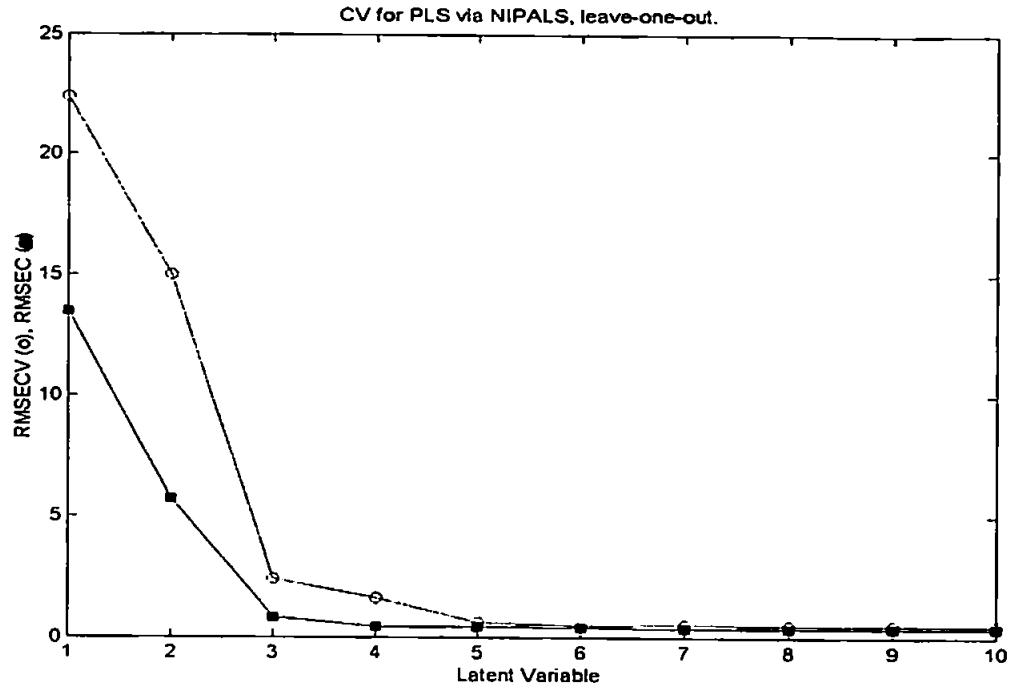


Figure 3.8 Cross validation plot for Pd (RMSEC included also)

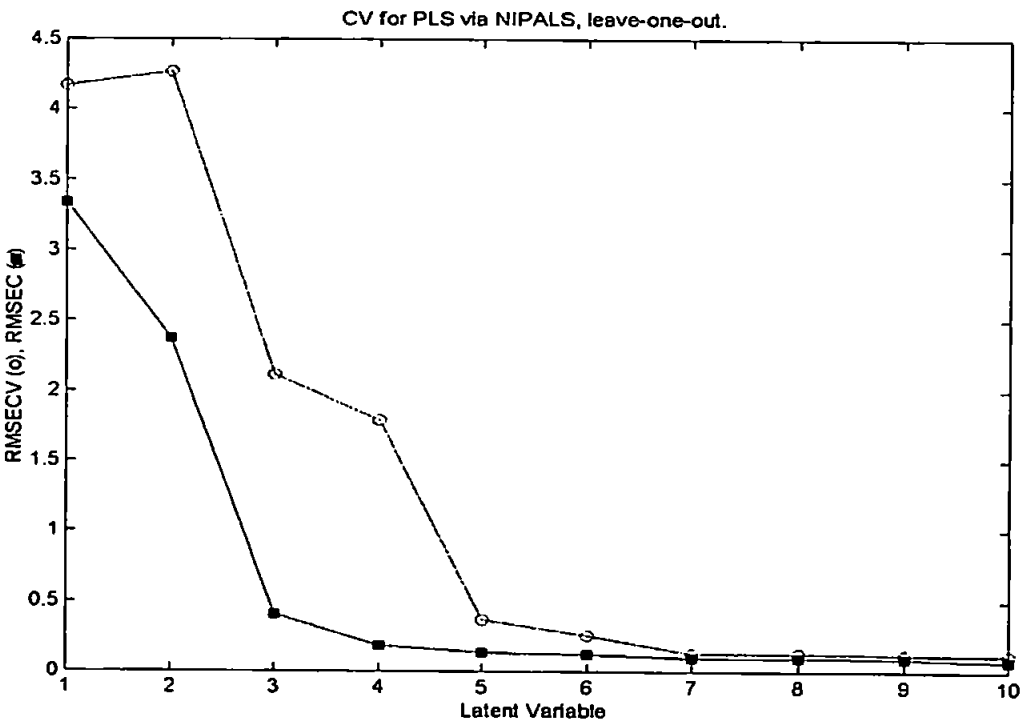


Figure 3.9 Cross validation plot for Rh (RMSEC included also)

The individual results for the synthetic test solutions are shown in Table 3.8 and Fig. 3.3 respectively. In the majority of cases the RSEs for the analytes after multivariate treatment were lower than the corresponding values when univariate and interelement correction was applied. It is evident from Fig. 3.3 that the predictive accuracy of the PLS1 model was highly dependent on analyte concentration. For example, the RSE for Pd in the test solutions decreased from 45 % for 2 $\mu g\ ml^{-1}$ to 11 % for the 6 $\mu g\ ml^{-1}$ solution, and only -0.5 % for 40 $\mu g\ ml^{-1}$. This pattern was repeated for the autocatalyst samples, with the RSE for Pd changing from 114 % at ca. 1.2 $\mu g\ ml^{-1}$ to 0.1 % at ca. 33 $\mu g\ ml^{-1}$ (Fig. 3.4).

Table 3.8 Known and predicted concentrations (μg ml^{-1}) of Pt, Pd and Rh in the synthetic test solutions (Te) and autocatalyst samples, predicted using PLS1

	Pt		Pd		Rh	
Sample	Known	Predicted	Known	Predicted	Known	Predicted
Te1	12	12.71	20	20.27	3	3.04
Te2	16	17.45	12	11.98	5	5.03
Te3	20	20.55	18	18.68	2	2.18
Te4	12	12.95	14	14.21	4	3.93
Te5	18	18.43	10	10.64	3	3.1
Te6	6	7.04	30	29.9	1	1.26
Te7	2	3.64	6	6.64	2	2.09
Te8	40	40.33	2	2.9	8	7.86
Te9	30	28.93	40	39.81	10	9.78
Te10	n/d	2.2	n/d	0.44	n/k	0.36
S1 R 1	1.01	1.26	1.22	2.46	n/k	-0.11
S1 R 2	1.02	-0.03	1.19	2.85	n/k	0.14
S1 R 3	0.99	0.98	1.22	2.45	n/k	0.1
S2 R 1	1.99	1.55	1.51	2.77	n/k	0.3
S2 R 2	1.88	1.97	1.48	2.36	n/k	0.1
S2 R 3	1.99	1.97	1.49	2.93	n/k	-0.34
S4 R 1	n/d	1.06	17.62	18.19	5.36	5.63
S4 R 2	n/d	1.11	17.86	17.92	5.36	5.62
S4 R 3	n/d	1.36	17.36	17.59	5.33	5.65
S5 R 1	n/d	1.47	33.7	32.63	0.05	2.83
S5 R 2	n/d	0.25	34.13	33.84	0.03	2.93
S5 R 3	n/d	0.63	33.73	32.7	0.03	2.95

n/d: not detectable
n/k: not known

For the autocatalyst samples the RRMSEs for Pt, Pd and Rh were 32.38, 7.44 and 75.33 % (Table 3.4). The relative failure of the model at low concentrations can be partly explained by the fact that the lowest concentrations used in the multivariate calibration data set were 5, 5 and 1μg ml^{-1} for Pt, Pd and Rh respectively, so the lower end of the concentration range was not modelled sufficiently well to enable accurate prediction of analyte concentration below these concentrations. This hypothesis is lent credibility by the

fact that nearly all of the synthetic test solutions were predicted with lower RSE error values than the autocatalyst samples, indicating also that interfering elements may have been present in the autocatalyst samples which were not included in the calibration model. Another possible cause may be due to the use of gross integrated intensities, which would not have corrected for any changes in the plasma continuum due to different element concentrations. Any non-linear fluctuation in the continuum would have a significant effect at low rather than higher analyte concentrations. One possible solution to this would be to model the lower concentration range, where noise is likely to have a greater influence, separately from the higher concentration range.

3.4 Conclusions

A number of calibration methods have been compared for the simultaneous determination of Pt, Pd and Rh in test solutions containing a synthetic matrix and autocatalyst samples containing varying concentrations of these analytes. Traditional calibration showed that for those elements in the test samples the predominant interferences on the Pt 214.423 nm, Pd 248.892 nm and Rh 343.489 nm lines were caused by matrix induced suppression and spectroscopic interference, a combination of both or poor background correction procedures.. Several calibration methods were compared, with the best being matrix matching and multivariate calibration using PLS1. Matrix matching failed when the standards were not matched correctly, which will often be the case with variable matrices, and PLS1 yielded good results at high concentrations but was less effective at low concentrations due to the noise contribution.

The PLS1 models were built using intensity data for 248 lines which included all of the most intense Pt, Pd and Rh lines and many of the most intense matrix lines. Thus, the assumption was made that these lines would be the optimum set for

modelling, however this may not be the case. It may be far better to allow the PLS1 model to select 'informative' regions of the spectrum based upon both PLS1 principles coupled with some selection criterion. This is investigated in the next section.

CHAPTER 4 - PLS AND VARIABLE SELECTION USING THE FULL AVAILABLE SPECTRUM USING COMPLEX SYNTHETIC SOLUTIONS AND INDUSTRIAL AUTOCATALYST SAMPLES

4.1 Introduction

In Chapter 3, multivariate calibration routines were applied to gross integrated line intensities (no BCPs) acquired using ICP-AES. The rationale for using integrated line intensities being that they represent those regions of the spectrum which contain the most useful information with respect to the analyte and matrix elements under study. However, there are several disadvantages to this approach:

1). The integration must be able to correctly integrate the peaks in circumstances where the optimal assignment of background correction points may not always be possible (see Chapter 2).

2). Only a limited number of lines can be included in the analytical method, hence an *a priori* choice must be made as to which analytical lines are included, thereby excluding potentially useful data.

3). Data processing time and complexity is increased.

A simpler and more elegant method would be to use the raw spectral data (i.e. signal intensity at each pixel in the spectral array). The raw spectral data has up to 5684 data points for each sample acquisition, much of which will be redundant or noise. Hence, it is necessary to reduce the data to a manageable size while retaining as much of the useful information as possible.

Partial least squares regression is now a popular multivariate calibration tool for the quantitative analysis of spectral data because of its potential to overcome common problems such as collinearity, spectral overlaps and matrix effects(48, 53). However, the results of any calibration technique are only as good as the data the technique is given to work with. It has been noted experimentally that including data with poor, or uninformative, information regarding the parameter of interest results in less than optimal calibrations (127, 128).

Selection procedures have been coupled with a variety of different modelling techniques, with most work concentrated in the area of conventional multiple linear regression (MLR), e.g. forward and backward stepwise multiple linear regression. Factor based techniques, such as PLS and PCR have been considered with wavelength selection algorithms less often, although improvements in performance have been reported in several studies (129-132). Regardless of the regression method employed, the thrust behind wavelength selection is the identification of a subset of spectral wavelengths that will produce the smallest possible errors when used to predict chemical concentration. In obtaining this optimal subset of wavelengths the full available spectrum is used. This adds a major advantage to this modelling methodology in that it is possible to build predictive models without the need for expert knowledge of the chemical system and the selection of appropriate wavelengths.

4.2 Preliminary Studies of Variable Selection

4.2.1 Introduction

Preliminary work on variable selection focused on the theory used by McShane *et. al.* (133). The approach was based on providing PLS with those regions with greatest spectral variance. Because PLS looks for correlated variation within a data set, those spectral regions with large variation may contain important information for modelling.

4.2.2 Theory of Variable Reduction

4.2.2.1 Mathematical benefits to Partial Least Squares

The nature of any data set can be expressed as Eqn. 4.1:

$$\mathbf{Y}_1 = \mathbf{x}\boldsymbol{\alpha}' + \mathbf{E}_1, \mathbf{Y}_2 = \mathbf{E}_2 \qquad \text{Eqn. 4.1}$$

Assuming the classical model: $\mathbf{Y}_1(N \times p_1)$ and $\mathbf{Y}_2$ $(N \times p_1)$ are the spectral data for N samples at p_1 and p_2 wavelengths, $\mathbf{x}$ is the vector of N sample concentrations, $\boldsymbol{\alpha}$ is a vector of unknown parameters for each of P total wavelengths ($P = p_1 + p_2$), and $\mathbf{E}_1$ and $\mathbf{E}_2$ are the random error matrices of mean zero and variance σ_1^2 and σ_2^2.

We now define the predictor of future concentration $\mathbf{X}^*$ based on the inverse calibration model, Eqn. 4.2

$$\tilde{\mathbf{X}}^* = \mathbf{Y}_1^*\tilde{\boldsymbol{\beta}}'_1 + \mathbf{Y}_2^*\tilde{\boldsymbol{\beta}}'_2 \qquad \text{Eqn. 4.2}$$

However, assuming $\mathbf{Y}_2$ is a set of "noise" or useless variables, this equation clearly adds bias to the estimator. Therefore we wish to define a predictor for $\mathbf{X}^*$ based on the inverse calibration model without the noise component, Eqn. 4.3

$$\tilde{\mathbf{X}}^* = \mathbf{Y}_1^* \hat{\boldsymbol{\beta}}_1' \qquad \text{Eqn. 4.3}$$

where only informative variables are considered. It can be proven that even for a small number of samples, $\hat{\boldsymbol{\beta}}_j(\mathbf{Y}_1)$ produces a smaller mean square error than $\hat{\boldsymbol{\beta}}_j(\mathbf{Y}_1 + \mathbf{Y}_2)$ and $\hat{\mathbf{X}}^*$ has a smaller mean square error of prediction than $\tilde{\mathbf{X}}^*$ (133), and in fact, this result has been experimentally observed (127-129).

4.2.3 Experimental

4.2.3.1 Instrumentation and Reagents

See Chapter 3, section 3.2

4.2.3.2 Experimental Design

The calibration set for multivariate analysis was prepared using a Taguchi orthogonal array (section 3.2.2.6) in order to cover the required factor space with the minimum number of experiments. The concentration ranges of the elements were determined from historical data on the composition of autocatalyst digest samples. The orthogonal array contained 8 factors at seven levels with a total of 49 experiments, represented as $OA_{49}(7^8)$. The levels and factors for the design are shown in Table 4.1.

Table 4.1 Experimental design analyte and matrix concentrations (μg ml^{-1})

	High concentration range (μg/ml)						
Pt	0	5	10	20	30	40	50
Pd	0	5	10	20	30	40	50
Rh	0	1	2	4	6	8	10
Ba	0	1	5	10	50	100	200
Ce	0	1	10	50	100	300	500
Zr	0	1	10	50	100	300	500
Mg	0	1	10	50	100	300	500
Al	0	1	10	100	200	500	1000

4.2.3.3 Data and Data Preprocessing

In order to ascertain the effectiveness of the variable selection procedure, models were prepared using three different datasets and compared. The range of the analytes and matrix elements used in the experimental design are shown in Table 4.1. The first dataset comprised the reduced variables only; the second, the unreduced spectrum (i.e. all 5684 wavelength data points from all 201 subarrays on the SCCD); and a third was prepared using the more traditional method of choosing 166 individual spectral lines representing the most intense analyte and matrix lines in the spectrum from which gross integrated line intensities were then modelled.

4.2.3.4 Algorithm Validation

All computer algorithms were developed within the Matlab environment (Matlab Software Version 5.0, and PLS_Toolbox 2.0 (Mathworks Inc)). Matlab was chosen because of its ability to quickly perform complex matrix operations and the availability of open access software, e.g. PLS_Toolbox 2.0.

The basis of all the in-house programs was the PLS routine contained in PLS_Toolbox 2.0. All programming was followed by rigorous validation which was composed of three stages:

1) Program design.

2) Programming / program alteration and visual inspection.

3) Visual inspection of complete program.

4) Testing of program with:
   i) In-house data (data with known parameters)
   ii) Referenced data with known parameter values (where available).

4.2.3.5 Procedure

4.2.3.5.1 Variable Selection using the Standard Deviation of Wavelength Intensity

Selection begins with the classification of all points in the spectrum as noise. Variables are placed into the model and if instrumental response is sufficiently sensitive, the standard deviation will be large at variables corresponding to points of interest. The number of variables placed into the model at any one time is governed by the total number of variables, here 5684. The addition of one variable at a time, although preferential for finding those variables that are important, greatly increases computational time. Therefore,

the variables were added in 'bits' consisting of twenty five variables at a time. Model error was assessed using relative root mean square error (RRMSE %, Eqn. 3.4).

A flow-chart outlining the selection algorithm, used for one analyte at a time, is outlined in Fig. 4.1 and was applied as follows:

i) All data was left unprocessed and classified as noise.

ii) The variables were ranked in descending order of their signal standard deviation.

iii) The size of the 'bit' to be placed into PLS was chosen (e.g. 25).

iv) The number of latent variables was selected and the data autoscaled (see later).

v) The RRMSE % value for the first 'bit' was calculated using the first PC, then using the first and second and so on until the RRMSE % value had been calculated using all PCs.

vi) The next 'bit' was then placed into the PLS routine and step v) repeated until all the data had been modelled.

vii) A plot of the RRMSE % value for each PC was then produced and the number of variables giving the lowest RRMSE % value chosen as the optimal data-set.

In step vi) the data was autoscaled as this gave the lowest predictive errors. It is emphasised that this was carried out *after* the data was ranked according to the standard deviation of the intensity.

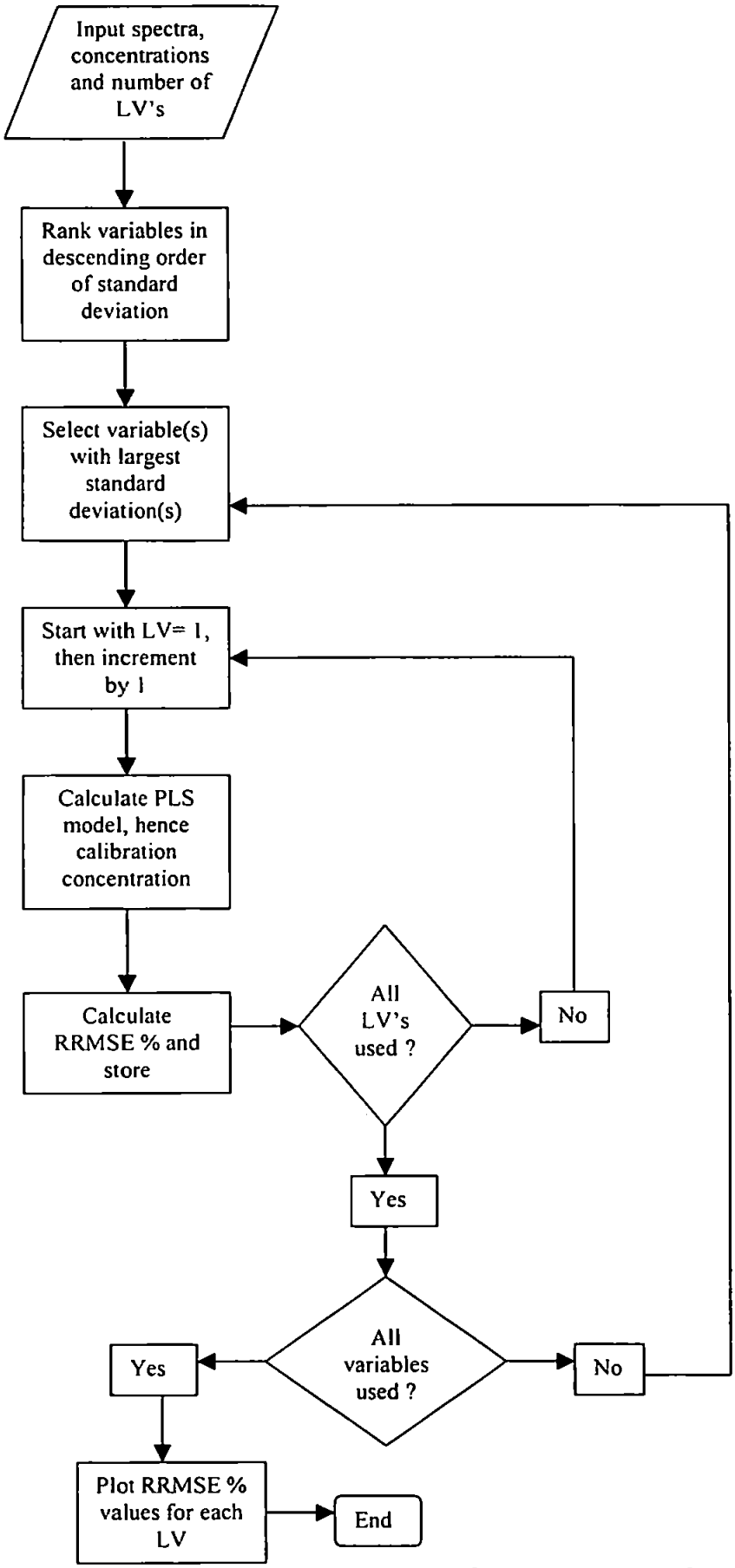


Figure 4.1 Flowchart for the execution of the variable ranking algorithm

In step vi) the data was autoscaled as this gave the lowest predictive errors. It is emphasised that this was carried out *after* the data was ranked according to the standard deviation of the intensity.

4.2.3.5.2 Preprocessing after Variable Selection

For the data set consisting of 166 individual analyte and matrix lines both mean centering and autoscaling were performed, however, in this particular case, autoscaling the data gave better predictive results. For the data set obtained after the application of the variable selection routine, again both methods of preprocessing were performed and autoscaling was chosen as it gave better models in terms of predictive accuracy.

4.2.3.5.3 Test Sample Confidence Interval calculation using the 'Jackknife' estimation of Standard Error

Chapter 4, section 4.4.2.2 describes the procedure for calculating test sample confidence intervals using the 'jackknife' method.

4.3 Results and Discussion

4.3.1 Variable Reduction

The effect of the variable reduction procedure is shown in Figs. 4.2a/b, 4.3 and 4.4. The minimum RRMSE % values were approximately 2, 5 and 2 % for Pd, Rh and Pt respectively, and were arrived at using a relatively small number of variables. For example

the RRMSE % values for Pd fell to 2% with only 300 variables, the Rh RRMSE % value fell to approximately 5% using only 600 variables, and for Pt the RRMSE % error value was reduced to only 2 % with 800 variables. Although all variables with high variances were not guaranteed to be correlated to any analyte in particular, sufficient numbers of variables were evidently present to ensure the relatively successful prediction of the calibration samples.

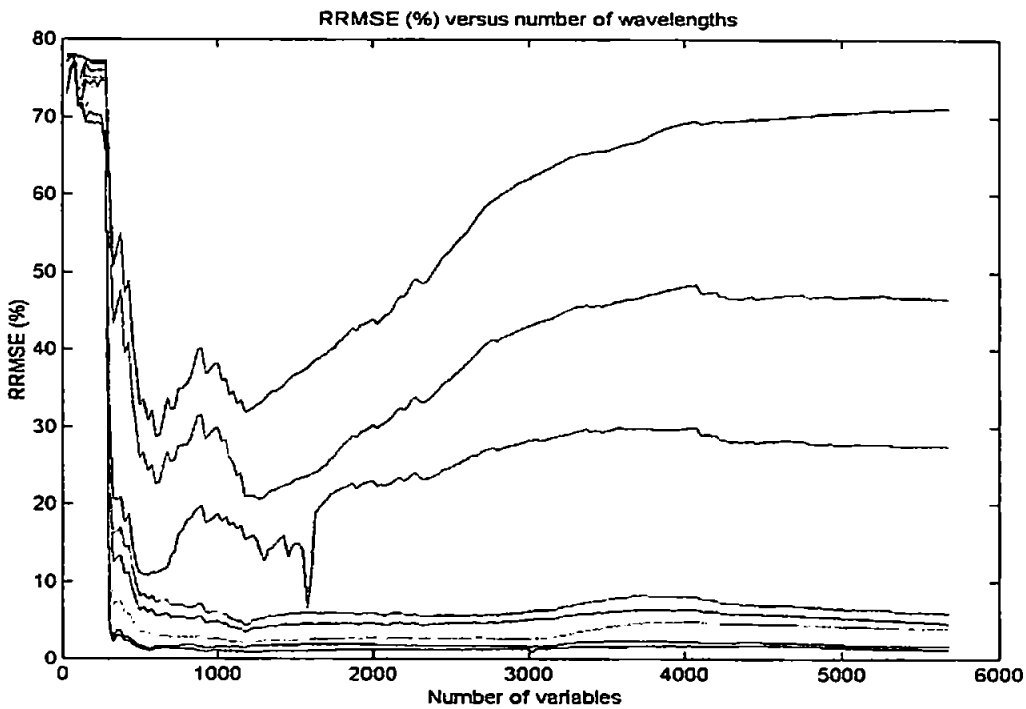


Figure 4.2a Pd RRMSE % as a function of variables in descending order of σ (8 PCs).

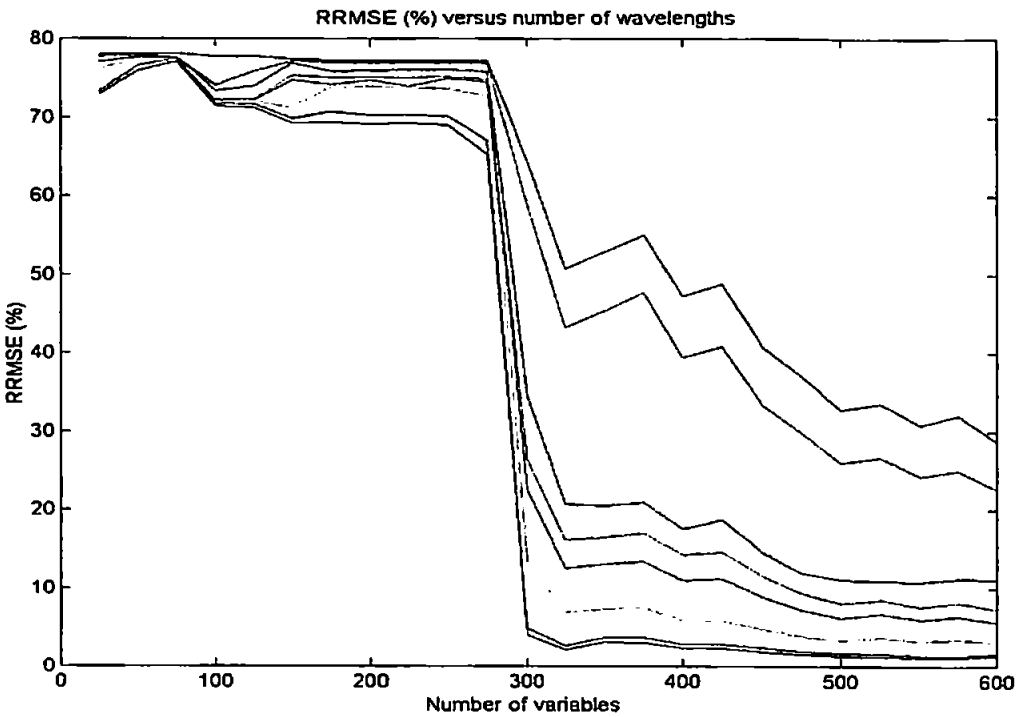


Figure 4.2b Pd RRMSE % from variables 1 – 600 (8PCs).

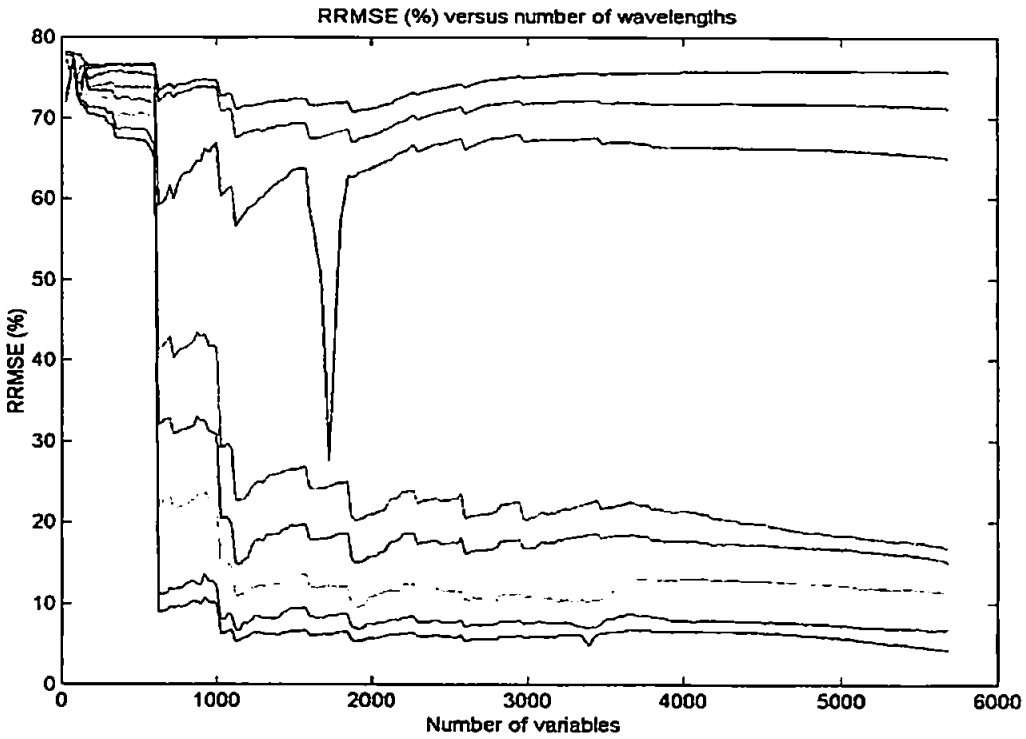


Figure 4.3 Rh RRMSE % as a function of variables in descending order of σ (8 PCs).

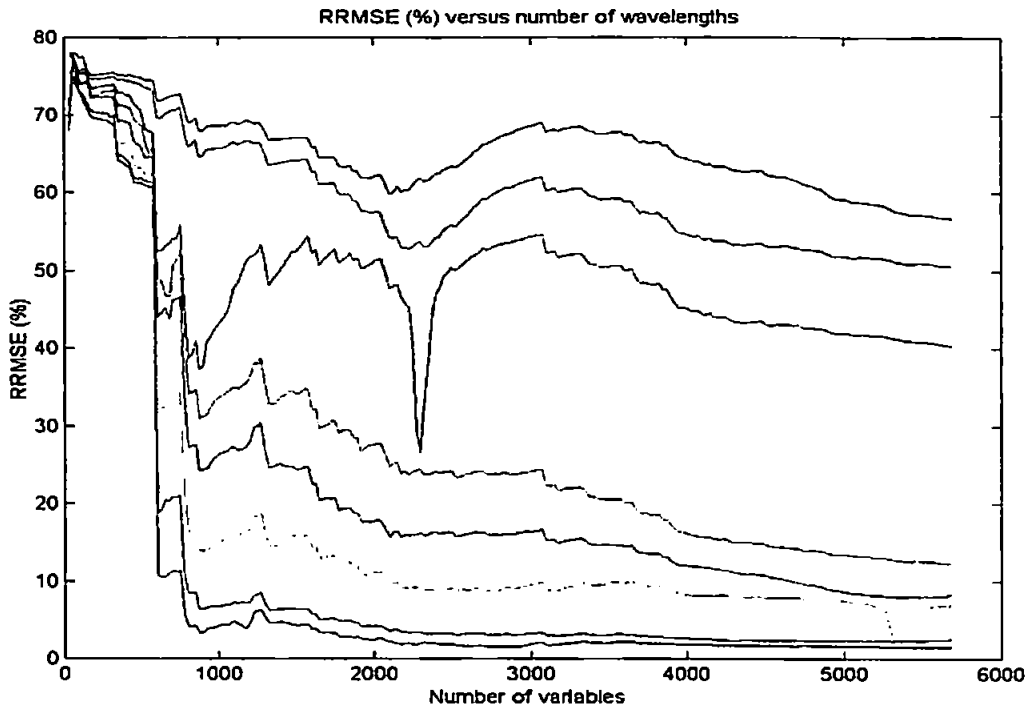


Figure 4.4 Pt RRMSE % as a function of variables in descending order of σ (8 PCs).

One worrying feature was the appearance of steps in the graphs (e.g. Fig. 4.3 for Rh). This was caused by variables with high a SD, but which were less correlated with the analyte of interest than the previous 'block' of variables. If the variable reduction procedure were working correctly then a smooth initial decrease in the RRMSE value to a global minimum would be expected. This would indicate that the ranking procedure had been successful in ranking the most important variables first.

4.3.2 Multivariate Calibration and Quantitative Prediction

Predictive results using variable selection for Pd, Rh and Pt are given in Tables 4.2 – 4.4. The RRMSE % values for the Pd, Rh and Pt test solutions were 4.39, 13.23 and 10.46 % respectively. These relatively low error values are seen graphically in Figs. 4.5 –4.7 where the majority of the samples fall on the 45° line. For the autocat samples only Pd was

predicted with any degree of accuracy, with an RRMSE % value of 18.44 %. (Table 4.2) In comparison, extremely poor accuracy was obtained for Rh and Pt (Table 4.3 and 4.4) which, had RRMSE % values of 141 and 6.5 x 10^3 respectively. Evidently those variables with high variance were well correlated to Pd and did not experience, to the same extent, interference from the unmodelled interferents, as did the Pt and Rh high variance lines. The RRMSE % errors obtained using the variable selection method compared with the use of all 166 individual lines and the full spectrum are given in Table 4.5. The lowest RRMSE % value obtained was for Pd (4.39 %) using the variable selection and autoscaling as the preprocessing step. By autoscaling, all variables are given the same weight in the PLS algorithm. Evidently for the prediction of Pd, there were very few uninformative wavelengths present after variable selection. For Rh and Pt the RRMSE % values were 13.32 and 10.46 % respectively, whereas using the individual line dataset (using mean centering) the RRMSE % values were 3.18 and 8.38 %. The highest predictive errors were obtained using the full spectrum., which contained 5684 wavelength data points with a substantial amount of noise. However, autoscaling gave the lowest predictive errors, this would appear to contradict the previous statement. However, if the intensity of the most informative predictors were relatively low, then autoscaling would have given these predictors an equal chance of participating in the model. Predictive errors for Rh and Pt in the autocat solutions ranged from 141.34 % for Rh to 6.5 x 10^3 % for Pt using variable selection, indicating the presence of unmodelled interferents in the autocat samples.

Table 4.2 Actual and predicted concentration for Pd (μg ml^{-1}) using variable selection.

Pd							
	Test samples oncentration (μg/ml)		95%		Autocat samples Concentration (μg/ml)		95%
Sample	Actual	Predicted	C. I.	Sample	Actual	Predicted	C. I.
Te1	20	19.48	2.14	Au1R1	1.07	0.97	>50
Te2	12	11.69	9.31	Au1R2	0.96	0.74	>50
Te3	18	17.92	1.70	Au1R3	0.90	0.79	>50
Te4	14	13.91	1.77	Au2R1	1.22	0.66	>50
Te5	10	10.32	1.76	Au2R2	1.19	1.05	>50
Te6	30	29.68	2.05	Au2R3	1.22	1.14	>50
Te7	6	5.68	2.06	Au31R	17.62	16.70	>50
Te8	2	2.64	2.03	Au3R2	17.86	16.74	>50
Te9	40	38.60	2.24	Au3R3	17.35	17.15	>50
Te10	0	1.18	1.81	Au41R	33.70	28.84	>50
				Au4R2	34.12	29.51	>50
				Au4R3	33.73	28.63	>50
RRMSEP %		4.39				18.44	

Table 4.3 Actual and predicted concentrations for Rh ($\mu g\ ml^{-1}$) using variable selection.

Rh						
Test samples oncentration (µg/ml)			95%	Autocat samples Concentration (µg/ml)		
Sample	Actual	Predicted	C. I.	Sample	Actual	Predicted
Te1	3	3.58	1.53	Au1R1	n/d	0.00
Te2	5	4.90	4.01	Au1R2	n/d	0.25
Te3	2	2.15	1.71	Au1R3	n/d	0.55
Te4	4	4.55	1.56	Au2R1	3.31	-1.53
Te5	3	3.28	1.56	Au2R2	3.32	0.65
Te6	1	1.53	1.98	Au2R3	3.21	0.46
Te7	2	2.74	1.93	Au31R	5.36	10.45
Te8	8	7.77	1.85	Au3R2	5.36	9.00
Te9	10	9.55	1.39	Au3R3	5.33	9.75
Te10	0	0.83	2.24	Au41R	0.05	-4.42
				Au4R2	0.03	-3.07
				Au4R3	0.03	-4.85
RRMSEP %		13.23				141.34

n/d Not detectable

Table 4.4 Actual and predicted concentrations for Pt (μg ml^{-1}).

Pt							
Test samples				Autocat samples			
	Concentration (μg/ml)		95%		Concentration (μg/ml)		95%
Sample	Actual	Predicted	C. I.	Sample	Actual	Predicted	C. I.
Te1	12	12.77	6.80	Au1R1	n/d	64.07	
Te2	16	17.25	21.25	Au1R2	n/d	63.86	
Te3	20	19.51	6.52	Au1R3	n/d	63.48	
Te4	12	11.69	6.27	Au2R1	1.01	73.31	>50
Te5	18	17.38	6.00	Au2R2	1.02	63.88	>50
Te6	6	7.26	7.90	Au2R3	0.99	64.10	>50
Te7	2	2.72	6.43	Au31R	n/d	57.30	
Te8	40	36.67	8.11	Au3R2	n/d	61.85	
Te9	30	26.87	7.78	Au3R3	n/d	59.60	
Te10	0	0.85	7.81	Au41R	n/d	134.94	
				Au4R2	n/d	129.89	
				Au4R3	n/d	135.53	
RRMSEP %		10.46				6,570	

Table 4.5 RRMSE % values for the synthetic test and autocat samples using variable reduction (VR), individual wavelengths and the full spectrum. The number of PCs are shown in parenthesis.

	Synthetic test solutions RRMSE %			Autocat samples RRMSE %		
	Pt	Pd	Rh	Pt	Pd	Rh
VR [As]	10.46 (10)	4.39 (10)	13.23 (10)	6570.0	18.44	141.34
Number of variables	1000	650	1250	1000	650	1250
Individual wavelengths (166) [Au]	8.38 (5)	7.06 (5)	3.18 (7)	n/a	3.38	n/a
Full spectrum [Au]	12.64 (8)	8.31 (8)	27.15 (8)	7480 (8)	622 (8)	1734 (8)

As Autoscaled
n/a Data not available
( ) Number of PCs used in the final PLS model

The variable selection method (Fig. 4.1 flow diagram of the selection procedure) makes the assumption that those spectral regions with large variance are correlated to the analyte of interest. In addition to this, this method assumes that there are no interactions between variables or groups of variables and hence that there is no important predictive information contained within such interactions which is not a safe assumption. A continuation of this methodology is to examine the correlation of individual wavelengths with the analyte of interest by using all information present within a data set. This can be accomplished through the use of multivariate calibration techniques, such as PLS. Here the individual regression coefficients are not estimated from the data directly, wavelength by wavelength, but are estimated using all the relevant information (via principal components), giving regression coefficients that are much lower in noise and which can be selected not only on the basis of their correlation with a particular analyte, but whilst taking into consideration the error associated with that particular correlation.

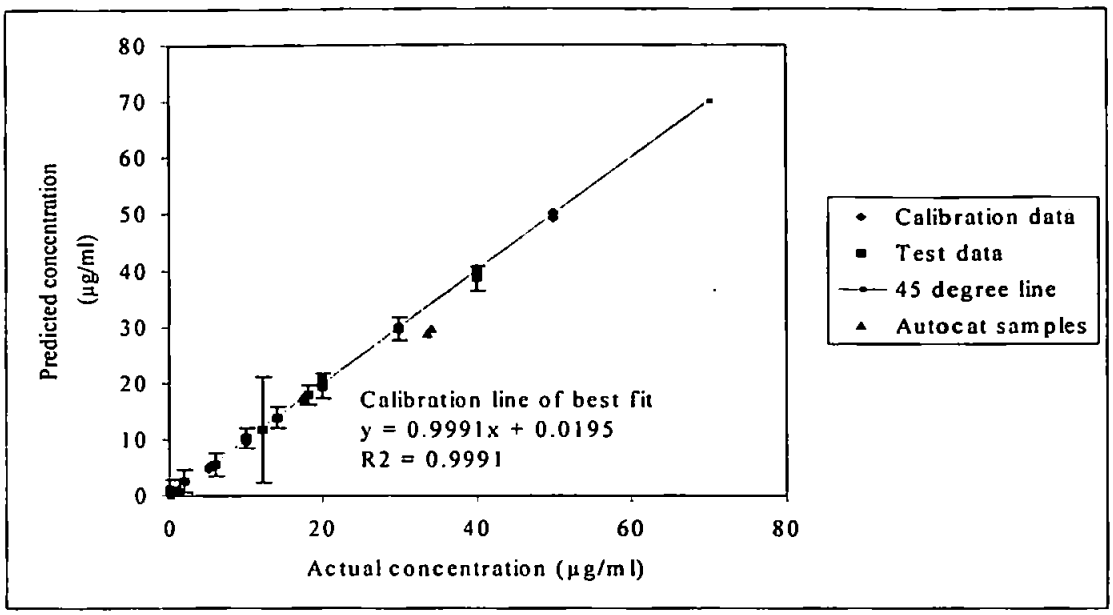


Figure 4.5 Actual and predicted concentrations for Pd calibration, test and autocatalyst samples (µg ml^{-1}).

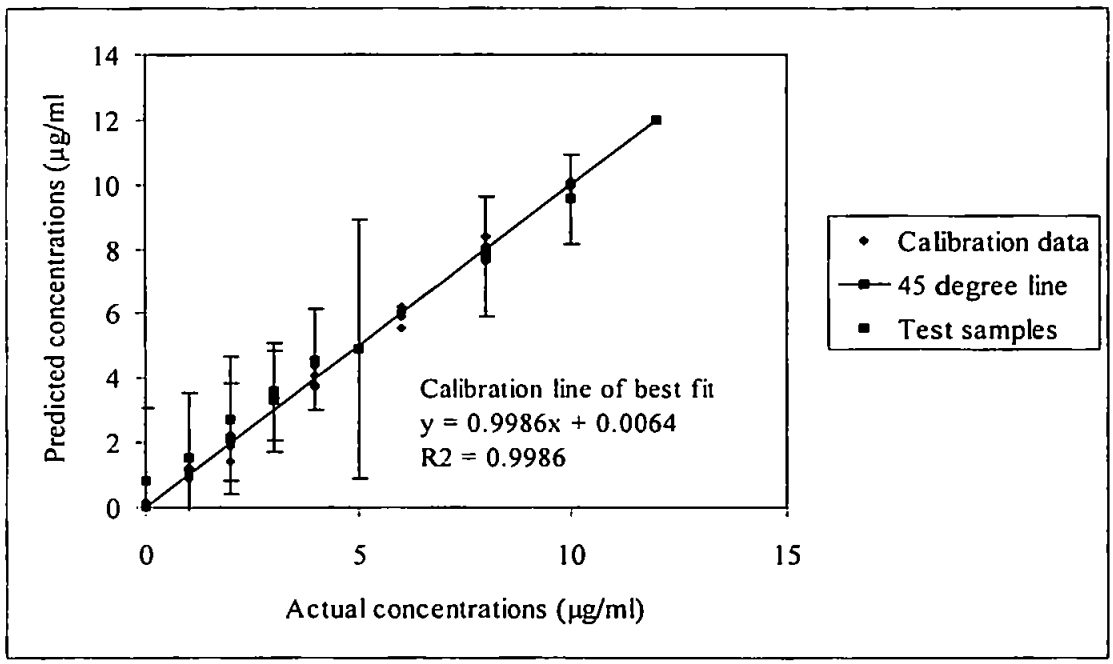


Figure 4.6 Actual and predicted concentrations for Rh calibration, test and autocatalyst samples (µg ml^{-1}).

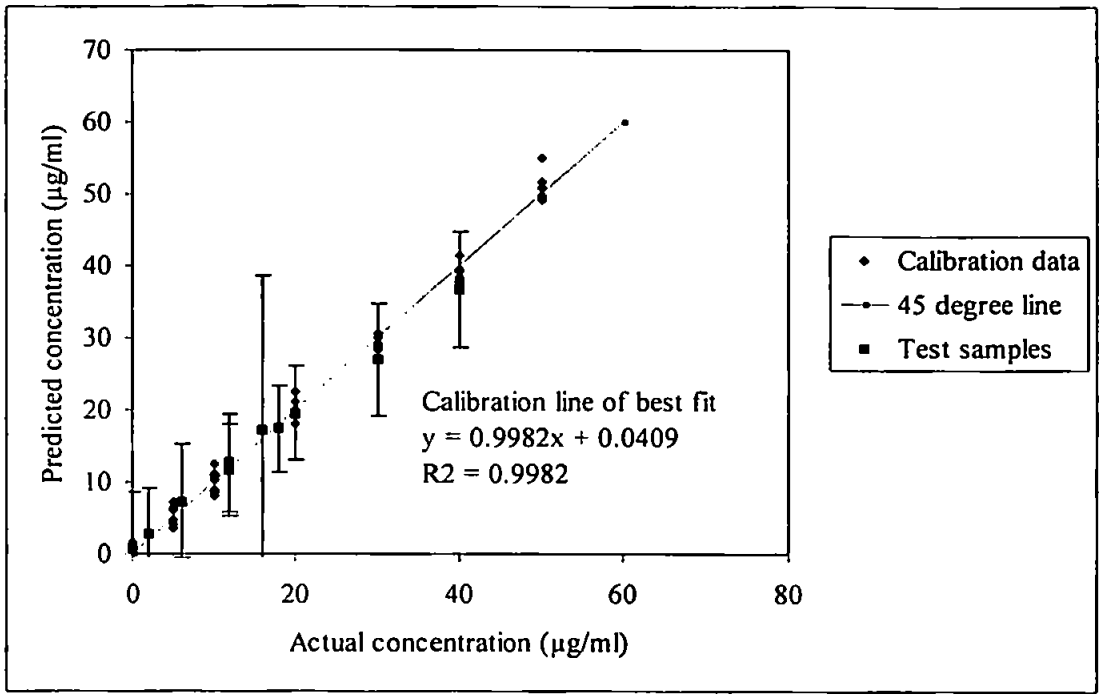


Figure 4.7 Actual and predicted concentrations for Pt calibration, test and autocatalyst samples (µg ml^{-1}).

4.3.3 Conclusions

Selection using variable standard deviation, has shown that it is able to select variables useful for prediction, but that large numbers of variables may be necessary. The main disadvantage of this technique, however, is that it is not analyte specific; variables are ranked without taking into consideration their correlation with the analyte of interest which makes many variables redundant, increasing the predictive error. The predictive ability of this technique for the autocat samples showed that it was only possible to predict Pd with any reasonable degree of accuracy. This may be due to the presence of interferents, the absence of informative variables, or a combination of both. Additionally, using standard deviation as the ranking criterion, well correlated variables with small intensity are overshadowed (ranked after) by less correlated variables with larger intensity, giving them less importance than they deserve within the dataset.

The use of variable relative standard deviation as a variable selection criterion would prevent this overshadowing effect, however, this may also give noisy baseline regions the same importance as important variables and the method would still be non-selective. Because of the problems associated with non-selectivity this ranking criterion was not investigated any further.

4.4 Variable Selection using Uninformative Variable Elimination by Partial Least Squares (UVE-PLS) and Informative Variable Degradation by Partial Least Squares (IVD-PLS)

4.4.1 Introduction

The variable selection method used in the previous section had the disadvantage that it was not analyte specific. Because the variables were ranked in descending order of their standard deviation there was no guarantee that variables with the largest variance were correlated with the analyte of interest. The variables were also considered in a univariate fashion, i.e. there was no consideration of how one variable influenced another. By using multivariate techniques it is possible to reduce a large data set into a number of PCs which progressively describe less and less of the correlated variance in that data set, until all that is left is noise. In this way all of the variables and their effects on one another, are considered simultaneously. Each PC will have associated with it a number of regression coefficients equal in number to the number of original variables, which give information on how important those variables were in forming that particular PC. By taking the cumulative sum of the regression coefficient for a particular number of PCs, the importance of each original variable can be determined. Because the noise component of

the model is modelled in the later PCs the noise component of the regression coefficients can be reduced by selecting the optimum number of PCs.

4.4.2 Statistical Theory

4.4.2.1 Traditional Theory and Complex Estimators

For some data x, an estimate $\hat{\theta} = t(x)$ for a particular parameter of interest (θ) is calculated. In the most familiar case, x consists of observations $x_1,....,x_n$ independently sampled from an unknown probability distribution F: as an example the parameter of interest θ is the true mean (population mean) of F (i.e. $\mu(F) = \int x\,dF(x)$); and the statistic $t(x)$ is the sample mean ($\bar{x}$).

Having selected an estimator $\hat{\theta}$, e.g. $\hat{\theta} = t(x)$, it is important to assess the accuracy of $\hat{\theta}$ as an estimator of the true value θ. The standard error of $\hat{\theta}$, Eqn. 4.4, is the most common measure of accuracy for estimators $\hat{\theta}$ that are unbiased (134):

$$Se\{\hat{\theta};F\} = [var_F\{t(x)\}]^{1/2} \qquad \text{Eqn. 4.4}$$

The formula for the standard error of the mean ($\theta = \bar{x}$) is given by Eqn. 4.5:

$$\sigma^2(F) = \int_{-\infty}^{\infty}\{x-\mu(F)\}^2\,dF(x) \qquad \text{Eqn. 4.5}$$

where $\sigma^2(F)$ is the variance of F. To relate $Se(\bar{x};F)$ to σ^2 Eqn. 4.6 is used:

$$Se\{\bar{x};F\} = [\sigma^2(F)/n]^{1/2} \qquad \text{Eqn. 4.6}$$

However, from a practical point of view, this is not helpful since $\sigma^2(F)$ is itself a function of the unknown distribution F. In this case though a simple unbiased estimate exists for $\sigma^2(F)$, in the form of Eqn. 4.7:

$$\hat{\sigma}^2(F) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)} \qquad \text{Eqn. 4.7}$$

Substituting 4.7 into 4.6 gives the estimated standard error for $\bar{x}$ in the form of Eqn. 4.8:

$$Se(\bar{x}) = \left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n(n-1)}\right]^{1/2} \qquad \text{Eqn. 4.8}$$

In 1958 John Tukey revolutionised error estimation with the "Jackknife" method built upon Quenouille's older technique for bias estimation. Tukey's method did not use Eqn. 4.6, but went directly to a generalisation of Eqn. 4.8 which omitted much statistical theory and instead relied upon computing power. This was not an argument against theory, but against unnecessary theory. Most common statistical methods were developed in the 1920s and 1930s, when computation was slow and expensive. Now that computation is fast and cheap major changes are occurring in statistical methodology, such as the Jackknife, a branch of statistics commonly known as resampling statistics.

4.4.2.2 The Jackknife Estimator

To illustrate the theory of the Jackknife, let the dataset x consist of n independent and identically distributed (i.d.d.) observations from an unknown distribution F, i.e. Eqn. 4.9:

$$F(i.i.d.) \rightarrow (x_1, x_2, \ldots\ldots, x_n) = x \qquad \text{Eqn. 4.9}$$

Let $x_{(i)}$ be the dataset with the i-th datum removed and let $\hat{\theta}_{(i)}$ equal $t(x_{(i)})$, the statistic $\hat{\theta}$ re-evaluated for the deleted point data set $x_{(i)}$. The jackknife estimate of the standard error is then given by Eqn. 4.10:

$$Se_{jack}\{\hat{\theta}\} = \left[\frac{(n-1)}{n}\sum_{i=1}^{n}(\hat{\theta}_{(i)} - \bar{\theta}_{(.)})^2\right]^{1/2} \qquad \text{Eqn. 4.10}$$

where $\bar{\theta}_{(.)} = \sum_{i=1}^{n}\hat{\theta}_{(i)}/n$. It is easy to verify that Eqn. 4.10 reduces to Eqn. 4.6 (135) when $\hat{\theta} = \bar{x}$ (i.e. when $\hat{\theta}$ has a simple algebraic form).The beauty of Tukey's Jackknife is that it can produce a standard error estimate for even the most complicated estimator (136) provided that the estimator is 'smooth'. The definition 'smooth' defines the way in which the estimator changes with changes in the data. A 'smooth' estimator will display only small changes with small changes in the data, such as the mean or regression coefficient. All that is required is the ability to recompute $\hat{\theta}_i$, n times, once for each deleted-point data set $x_{(i)}$. Because the estimates of PLS regression coefficient standard error do not have exact forms (136), the traditional approach requires the use of the jackknife procedure.

A subsidiary benefit of this Jackknife procedure is that an improved estimate of θ can be derived in those situations where the original estimator of θ is biased. The

estimator ($\tilde{\theta} = n\hat{\theta} - (n-1)\bar{\theta}$) is known as the Jackknife estimator. If $\hat{\theta}$ has bias of the order $1/n$ then the bias of $\tilde{\theta}$ is substantially reduced, to order $1/n^2$.

4.4.3 Uninformative Variable Elimination by Partial Least Squares (UVE-PLS)

The method proposed here is based upon that of Centner *et al.*(78) and modified by both Westad *et al.*(87) and Faber (137). In the original paper by Centner, a reliability criterion, Eq. 4.11, was used in conjunction with the addition of random noise. The reliability criterion c_j was based on an analogy with stepwise MLR. Because, the estimated standard error, $s(\beta_j)$, cannot be computed directly for PLS. Centner *et. al.* proposed to estimate the regression coefficient (β_j) as a mean and its estimated standard error $s(\beta_j)$ for the j-th variable using a 'leave one out' strategy.

$$c_j = \beta_j / s(\beta_j) \qquad \text{Eq. 4.11}$$

Two sets of data were collected: one set comprising ordinary calibration data and a second set comprising artificial random variables (c_{noise}). These random variables were used to define a subjective cut-off level. The problem of determining a subjective cut-off level was achieved using Eq. 4.12.

$$In_{X_j}(c_j) = < |max(c_{noise})| \qquad \text{Eq. 4.12}$$

where $In_{X_j}(c_j)$ is the reliability coefficient (c_j) for the j-th informative variable in the experimental data collected by the instrument and $|\max(c_{noise})|$ is the modulus of the

maximum value for c_j from the random noise. Because the random data represent (artificially added) noise, their c_j values will be indicative of the values that can be reached by uninformative variables and hence a cut-off level is given. However, this method was dependent upon the satisfactory estimation of the magnitude of the noise, it is therefore a subjective method. The modification proposed by both Faber (137) and Westad (87) gave an objective cut-off value for the reliabilities in the form of a jackknife corrected estimated standard error which was used to determine whether $\beta_j \neq 0$, (i.e. an informative variable) as opposed to $\beta_j = 0$, (which was classified as an uninformative variable), using the Student's t-test. It is this modified method which is used as a basis here (as is shown in the flow diagram in Fig. 4.8a). An objective cut-off can easily be found using the Student's t-test by using $t_{variable} = \overline{\beta} / Se(\beta)$ and the tabulated T values for the appropriate degrees of freedom (Fig. 4.8b). The degree of freedom used here was equal to the number of calibration samples used (N).

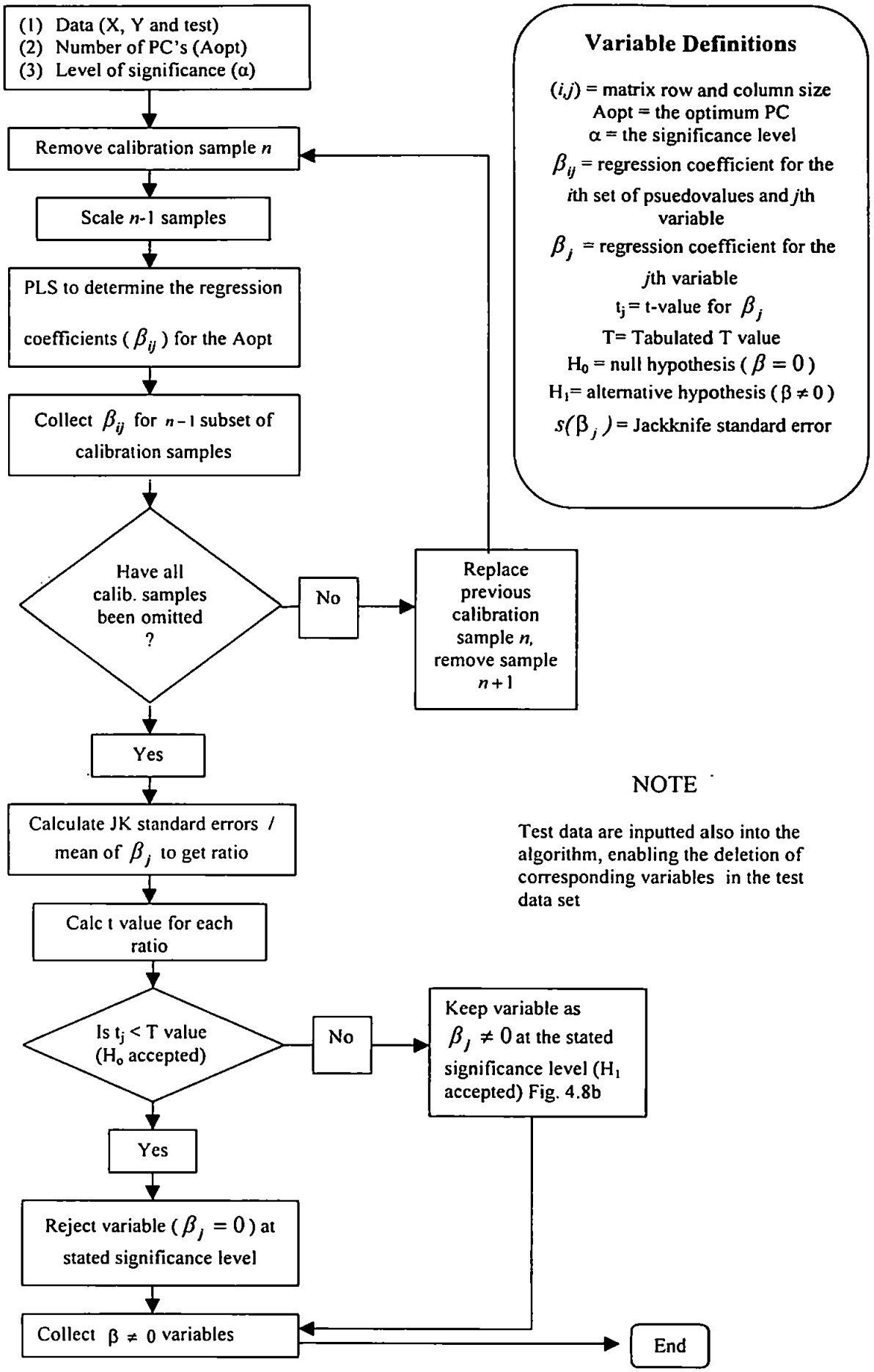


Figure 4.8a Flowchart for the execution of the UVE-PLS algorithm

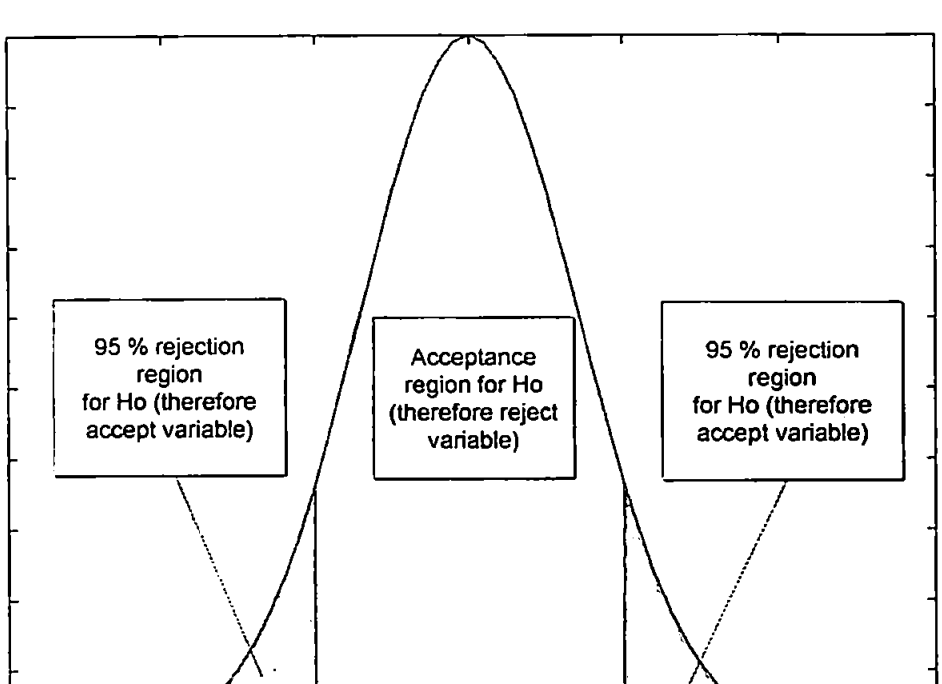


Figure 4.8b T-test illustrating the 95 % confidence interval for $\beta = 0$ (uninformative variable)

4.4.3.1 Limitations of Multiple Statistical Comparisons

If a single statistical test is conducted, a one-sample t-test for example, and $\alpha = 0.05$, then the probability of a Type I error (rejecting Ho when it is true) is under control. If however, 10 independent one-sample t-tests are now performed, each at the 0.025 level, the probability of a false rejection is no longer 0.05. The *overall* α for a set of tests is the probability of at least one rejection when the Ho is true. This is described mathematically by Eqn. 4.13 for independent tests:

$$\alpha = 1 - (1 - \alpha')^k \qquad \text{Eqn. 4.13}$$

which is the probability of at least one type I error. Where α is the significance level and k is the number of tests performed.

The consequences of this are apparent when one considers the above example of 10 multiple t-tests. The probability of a type I error for any single t-test is 0.025, however the combined probability of obtaining at least one type I error using 10 t-tests is given by $1-(1-0.025)^{10}$ which equals approximately 0.22, or 22%. Therefore at least one type I error will occur 22% of the time. If α remains constant, but a large number of t-tests are performed, say 5684 corresponding to an equal number of variables, there is a $1-(1-0.025)^{5684} \sim 100$ % probability of, an average, $0.025 \times 5684 = 142$ type I errors.

For the purposes of variable selection, therefore, not all the UVE-PLS selected variables will be truly informative. Because of this a decision rule was required to find the minimum number of variables that could be accepted given that a proportion were the result of type I errors, denoted $var_{Type\,I}$. The decision rule was based upon several criteria: Type I and Type II errors, test power, the correct estimation of the individual PLS regression coefficients (β) in UVE-PLS and the predictive RRMSE % value. The value of alpha (α) chosen was 0.025, there was therefore a 2.5 % probability of rejecting the null hypothesis (Ho) when it was true for each individual one-sample t-test. Levels of α below this were applied, but proved too stringent resulting in very low numbers of variables after the application of UVE-PLS due to the increase in the Type II error (accepting Ho when it is incorrect) and the concomitant decrease in t-test power.

For the accurate estimation of β_j, a suitable number of PCs were required, too many or too few would have resulted in the addition of bias to the individual elements of $\boldsymbol{\beta}$. Where possible UVE-PLS datasets (Var_{sel}) were chosen, such that the inequality given by Eqn. 4.14 was obeyed:

$$Var_{sel} / \alpha n \approx> 3 \qquad \text{Eqn. 4.14}$$

where α is the accepted probability of a Type I error and n is the number of t-tests performed. Thus the initial probability of choosing a truly informative variable, without replacement, is $\approx> 66\%$, and therefore the probability of a $var_{Type\,I}$ variable is $\approx< 33\ \%$.

4.4.4 Informative Variable Degradation by Partial Least Squares (IVD-PLS)

UVE-PLS can be thought of as a filter which allows through those β-coefficient's that have a high probability of not equalling zero, including $var_{Type\,I}$ variables. A variable may, therefore, have an associated regression coefficient equal to 0.01, whereas another may have a regression coefficient of 10. If regression coefficient standard error is neglected for a moment it is clear that the variable with a regression coefficient equal to 10 is the more important of the two, so in order to select these variables efficiently, a suitable criterion must be found. Centner *et. al.* proposed using a genetic algorithm(78), however one of the dangers of using such algorithms is that of very large numbers of random correlations. Even after the application of the UVE-PLS routine to the full spectrum used in this study the number of samples required to give a satisfactory variable to sample ratio would be in the order of 800-900 , which is clearly impractical (138).

In spectroscopic applications, the β-coefficients cannot be used directly to choose which wavelengths are most important for modelling. Indeed, a large coefficient may indicate a significant variable, but it may also have large variability with little or no correlation to the analyte of interest. This problem can be avoided by autoscaling the data so that a large absolute β-coefficient indicates an important variable (79).

It is essential that the regression coefficients used in the final predictive PLS model are both relatively large and have low standard error. Hence, a ranking scheme is proposed whereby the importance of the mean regression value is assessed by ratioing it to its standard error, with larger ratios being assessed as more important (Eq. 4.15).Informative variables will have both a large multivariate regression coefficient estimate and small estimated standard error, so ranking on the basis of a decreasing IVD ratio should enable the most important variables to be used in the PLS algorithm.

$$var_j^{ivd} = \bar{\beta}_j / s(\beta_j) \qquad \text{Eq. 4.15}$$

where var_j^{ivd} is the IVD ratio for the j-th variable, $\bar{\beta}_j$ is the mean value for the j-th regression coefficient from autoscaled **X** data, and $s(\beta_j)$ is its estimated standard error. Although the presence of $var_{Type\,I}$ variables in the UVE-PLS dataset mean that uninformative variables will be present together with truly informative variables, by choosing UVE-PLS datasets where $Var_{sel} / \alpha n \approx> 3$ the probability of ranking $var_{Type\,I}$ in the first few programme iterations is $\approx< 0.33$, thereby giving weight the truly informative variables.

In order to obtain the correct number of ranked variables, the cumulative sum of var_j^{ivd} is obtained and at specific percentage intervals, e.g. 30 to 100 % stepping 5%, the corresponding root mean square error of cross validation (RMSECV) is obtained for that set of mean centered wavelength data points (as is shown in the flow diagram in Fig. 4.9).

It can be argued that by mean centering preference will be given to variables with a larger variability due to their size whatever their correlation. The alternatives are to autoscale the data, which will give equal weight to variables with less correlation, or to use

non pre-processed data. Mean centering has been chosen in this case, although either pre-processing method has advantages and disadvantages. As will be demonstrated, the whole process results in a minimum being found which then gives the correct number of ranked variables to use. The stepping value can be altered to suit the degree of accuracy required and in this study a value of 5 % was found to be adequate. The final PLS model was built with the corrected number of mean-centered wavelength data points.

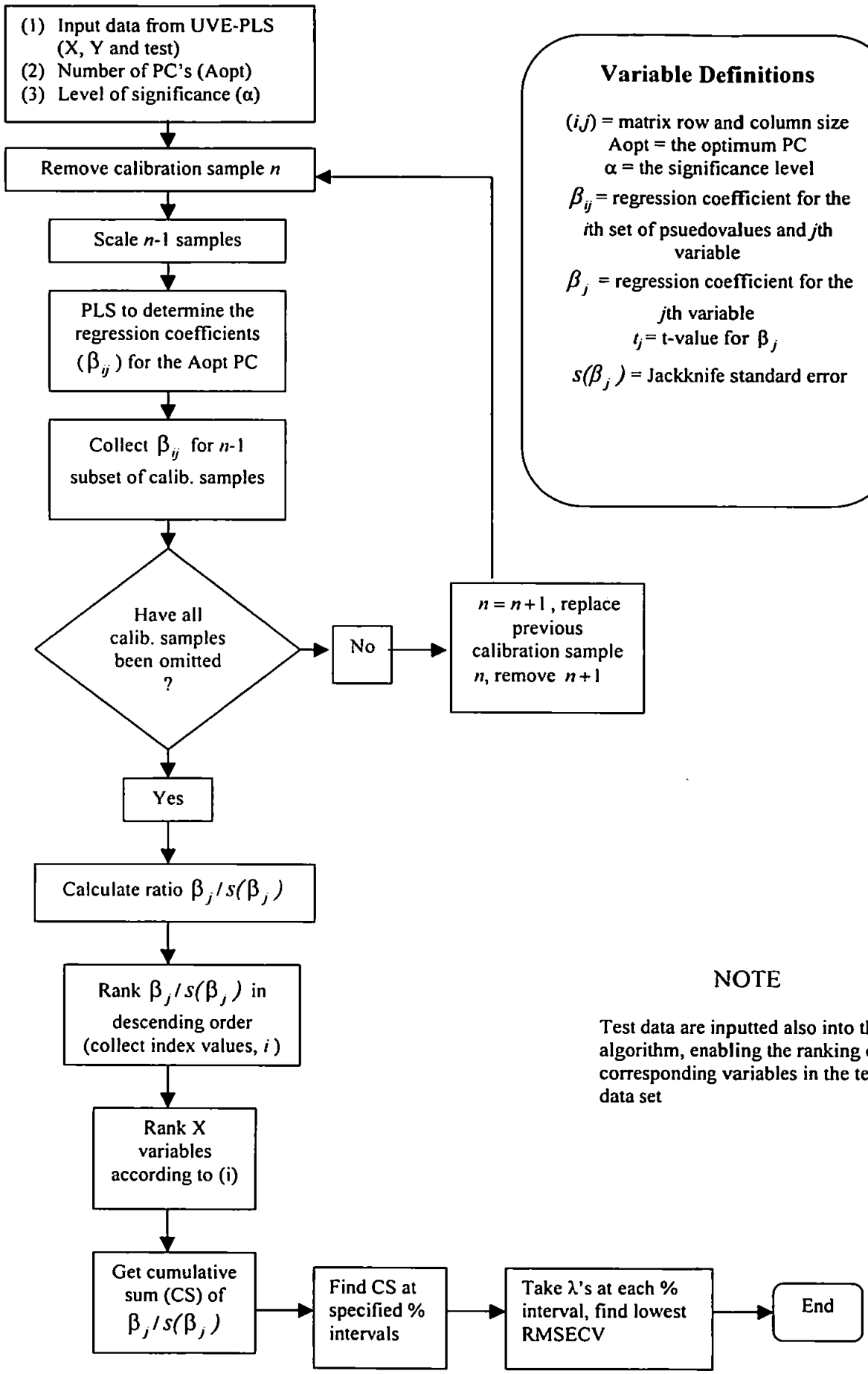


Figure 4.9 Flowchart for execution of IVD-PLS algorithm

Correct estimation of all parameters in the IVD and UVE-PLS routines, is dependent upon the optimum number of PCs being used. The calculation of these parameters is performed prior to the removal of any variables so the estimation of the correct number of PC's by minimising the predictive error will not give a unique solution. This effect can be limited by using a range of PC's in an attempt to reduce predictive errors in the final models.

4.4.5 Confidence Intervals

The multivariate inverse model is of the form (Eqn. 4.16) :

$$y = \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon \qquad \text{Eqn. 4.16}$$

where y is the concentration value; $\beta_1, \beta_2, ..., \beta_k$ are parameters with unknown values (regression coefficients), $x_1, x_2, ...x_k$ are independent information contributing variables that are measured without error and ε is a random error component. It is assumed that the random error has a normal probability distribution with mean equal to 0 (i.e. $E(\varepsilon) = 0$) and variance equal to σ^2 (i.e. $Var(\varepsilon) = \sigma^2$). Further, it assumed that the random errors associated with every pair of y values are probabilistically independent (homoscedastic). That is, the error ε associated with any one y value is independent of the error associated with any other y value.

The assumptions that have been described for an inverse multivariate model imply

that the mean value $E(y)$ for a given set of values $x_1, x_2, \ldots x_k$ is equal to Eqn. 4.17.

$$E(y) = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \qquad \text{Eqn. 4.17}$$

The successful determination of test sample confidence limits depends on the data satisfying these assumptions. Tests to detect deviations from these assumptions are possible and include trend detection in residual plots. The residuals used in this study take the form of concentration residuals ($(c - \hat{c})$ vs. $\hat{c}$), illustrated in Fig. 4.10, where c and $\hat{c}$ are the actual and predicted calibration concentrations respectively. If the assumptions concerning the error term ε are satisfied, concentration residual plots should display random fluctuation about zero with no observable trends and no residuals more than 3 estimated standard deviations (3σ) of ε above or below 0.

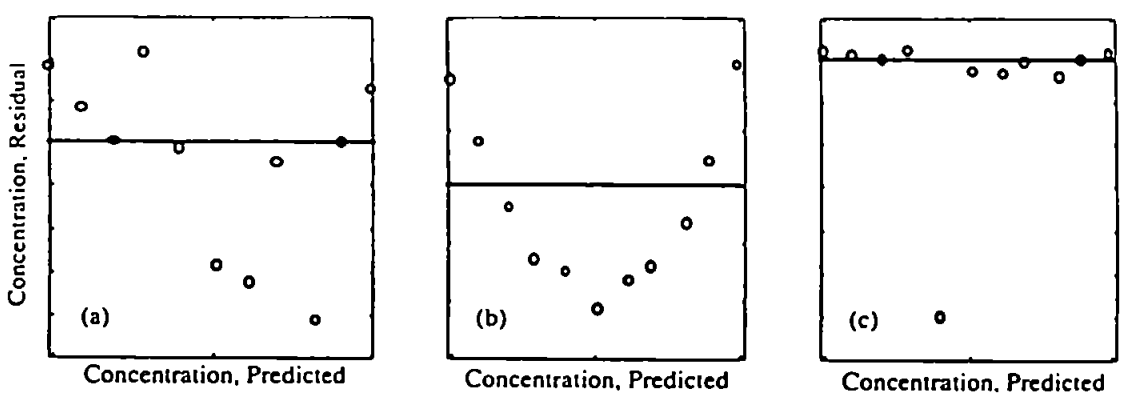


Figure. 4.10 Schematic concentration residuals versus actual concentration plots showing: (a) ideal; (b) non-linear and (c) outlier.

It has already been shown that if the Jackknife method is used it is possible to calculate the standard error of β or any other parameter of interest. For the model given by Eqn. 4.16 it is possible to construct a confidence interval for the β parameters using as follows Eqn. 4.18:

$$\beta \pm t_{\alpha/2} \mathrm{Se}(\beta) \qquad \text{Eqn. 4.18}$$

where $t_{\alpha/2}$ is the tabulated students t-value and Se(β) is the standard error of the regression coefficient using the Jackknife method. An alternative to this is the calculation of a confidence interval for the predicted concentrations themselves. As before each calibration sample in turn is omitted, however instead of calculating the standard error of β, sample concentrations are calculated n times (n = number of calibration samples). These resampled estimates of the sample concentrations are then put into Eqn. 4.10 and the standard error of prediction calculated as Eqn. 4.19. This process is illustrated by Fig. 4.11.

$$\hat{y} = t_{\alpha/2} \mathrm{Se}(\hat{y}) \qquad \text{Eqn. 4.19}$$

where $\hat{y}$ is the predicted value, $t_{\alpha \cdot 2}$ is the tabulated t-statistic for the appropriate degrees of freedom and $\mathrm{Se}(\hat{y})$ the estimated standard error of $\hat{y}$. Because predictions are simply the sum of weighted (β_j = weight for j-th variable) variables, predictions are defined as 'smooth' and are therefore well estimated by the Jackknife method of estimation.

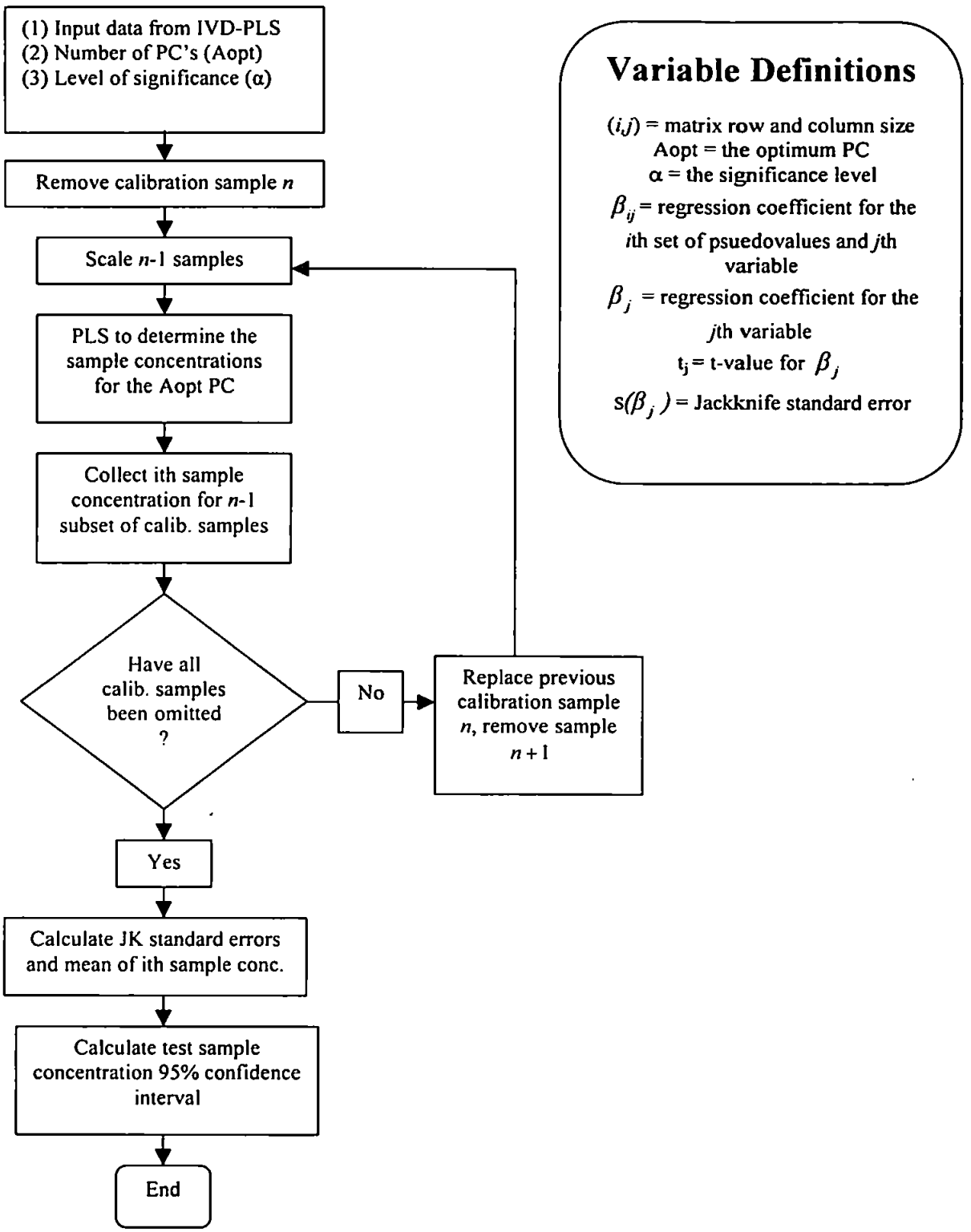


Figure 4.11 Flowchart for execution of confidence interval algorithm

4.5 Experimental

4.5.1 Instrumentation and Reagents

For information on instrumentation and reagents see Chapter 3, Section 3.2.1.

4.5.2 Experimental Design

For information on the experimental design used see Chapter 3, Section 3.2.2.6.

4.5.3 Procedure

4.5.3.1 Data and Data Preprocessing

In order to ascertain the effectiveness of the variable reduction procedure, models prepared using three different datasets were compared. The datasets were constructed spanning both a low analyte training range (LTR) and a high analyte training range (HTR) (Table 4.6). The first dataset comprised the reduced variables only; the second, the unreduced spectrum (i.e. all 5684 wavelengths); and a third was prepared using the more traditional method of choosing individual spectral lines representing the most intense analyte and matrix lines in the spectrum from which gross line integrated intensities were then modelled.

Data was autoscaled prior to the application of UVE-PLS and IVD-PLS algorithms in order to prevent variables with large and / or random variance from dominating the PLS model at the expense of variables with a small variance. This is important for atomic emission spectra in which signal intensities can range from a few hundred units up to

several hundred thousand or more and which can possess very different correlations with the analyte.

4.5.3.2 Partial Least Squares (PLS)

PLS was used because of its ability to reduce the impact of common problems such as collinearity, spectral overlaps, interactions, and matrix affects. Because it has been shown that linear PLS can provide good approximations with many types of non-linearities (76, 139) partial least squares with linear inner relations was used. Excessive non-linear structure in the data may be identified after modelling is performed by examining the concentration residuals.

Table 4.6 High and low training ranges ($\mu g\ ml^{-1}$)

	High concentration range (μg/ml)						
Pt	0	5	10	20	30	40	50
Pd	0	5	10	20	30	40	50
Rh	0	1	2	4	6	8	10
Ba	0	1	5	10	50	100	200
Ce	0	1	10	50	100	300	500
Zr	0	1	10	50	100	300	500
Mg	0	1	10	50	100	300	500
Al	0	1	10	100	200	500	1000
	Low concentration range (μg/ml)						
Pt	**0.05**	**0.1**	**0.2**	**0.4**	**0.8**	**1**	**2**
Pd	**0.05**	**0.1**	**0.2**	**0.4**	**0.8**	**1**	**2**
Rh	**0.05**	**0.1**	**0.2**	**0.4**	**0.8**	**1**	**2**
Ba	0	1	5	10	50	100	200
Ce	0	1	10	50	100	300	500
Zr	0	1	10	50	100	300	500
Mg	0	1	10	50	100	300	500
Al	0	1	10	100	200	500	1000

Bold Low analyte concentration range

4.5.3.3 Uninformative Variable Elimination PLS (UVE-PLS)

A flow-chart outlining the UVE-PLS procedure is shown in Fig. 4.8a (section 4.4.3). The software used was Matlab Software Version 5.0, and the PLS_Toolbox 2.0 (Mathworks Inc). The full spectrum data set was subjected to the UVE-PLS algorithm as follows:

i) The original data array extracted from the ICP-AES instrument was a 49 x 5684 matrix made up of 49 spectra (for the calibration data-set) each containing 5684 data points. One calibration spectrum was initially removed to leave a 48 x 5684 matrix after which the data was autoscaled.

ii) The PLS algorithm was applied and the β_{ij} regression coefficients were extracted for the optimum number of PCs and $\overline{\beta}_j$ calculated for each wavelength.

iii) Jack-knife corrected estimated standard errors were calculated for the β_{ij} regression coefficients, and a two-sided t-test was performed to determine which were equal to zero at the 97.5% confidence level ($\alpha = 0.025$).

iv) Those **X** variables corresponding to $\beta_j = 0$ were rejected from the original 49 x 5684 data matrix before progression onto the next step.

4.5.3.4 Informative Variable Degradation by PLS (IVD-PLS)

A flow-chart outlining the IVD-PLS algorithm is outlined in Fig. 4.9 (section 4.4.4), and was applied as follows:

i) One calibration spectrum was initially removed from the 49 x 5684 data matrix resulting from the application of the UVE-PLS algorithm. The remaining data was then autoscaled.

ii) The PLS algorithm was applied and the β_{ij} regression coefficients extracted for the optimum number of PCs and $\overline{\beta}_j$ calculated for each wavelength.

iii) Jack-knife corrected estimated standard errors, $s(\beta_j)$, and the mean, $\overline{\beta}_j$, were calculated for all regression coefficients.

iv) The $var_j^{ivd} = \overline{\beta}_j / s(\beta_j)$ ratios were calculated (see Eqn. 4) and ranked in descending order.

v) The **X** variables in the data matrix were ranked in accordance with the var_j^{ivd} ratios and the cumulative sum calculated.

vi) A multivariate PLS calibration model was built using mean-centered **X** data contributing to the first 30 % of the cumulative sum of the var_j^{ivd} ratios, and the RMSECV was calculated. The process was repeated using the first 35%, 40% and so on at 5% intervals of the cumulative sum data and the model with the lowest RMSECV value was chosen as optimal.

4.5.3.5 Estimation of Errors

The overall prediction efficacy, and assessment of the models capability to accommodate the calibration data itself, were compared using the relative root mean square error (RRMSE), defined in Eq. 4.20, which gives a general estimate of the error of prediction for concentrations of an element in the range of samples used:

$$RRMSE(\%) = 100 \times \frac{1}{mean(y)} \sqrt{\frac{\sum(\hat{y}_i - y_i)^2}{N}} \qquad \text{Eq. 4.20}$$

where y_i is the known concentration, $\hat{y}$ is the predicted concentration, and N is the number of experiments. However, the ultimate assessment of future prediction is the application of the calibration models to independent test data and the assessment of the confidence interval magnitude.

4.5.3.6 Test Sample Confidence Intervals

Central to the variable reduction routine is the estimation of β coefficient uncertainty. This uncertainty can be projected onto the test samples in order to define symmetric confidence intervals (assuming a normal distribution). This was done using a leave-one-out cross-calibration approach, so that the variation of the calibration models was used to estimate the variation in predictions for an independent sample. This was done n times (n = number of calibration samples) and the usual Jackknife formula used to estimate the standard error of the prediction. Test sample confidence intervals were then based on $TSP \pm (t_{(\alpha=0.05/2, DF=n)} * se(TSP))$ (TSP = test sample prediction), which corresponded to a confidence interval of 95% (i.e. $\sim 2 \times \sigma$).

4.6 Results and Discussion

The UVE and IVD algorithms were applied to the full raw spectral data matrix obtained from the ICP-AES instrument and multivariate calibration models built for the prediction of Pt, Pd and Rh in the synthetic test and autocatalyst samples. To avoid repetition the efficacy of the variable reduction algorithms are discussed using Pd as an example. A comparison of the results obtained for the multivariate calibration and prediction of Pt, Pd and Rh using the reduced data, the unreduced data, and integrated line intensities are then discussed.

4.6.1 Application of UVE-PLS and IVD-PLS Algorithms to the High Training Range (HTR) Dataset

Application of the variable reduction algorithms resulted in the deletion of the majority of the original 5684 spectral data points (Table 4.7). It is evident that the UVE-PLS algorithm had the largest effect, eliminating between 79-84% of the spectral data.

Table 4.7 Effect of applying the UVE-PLS and IVD-PLS algorithms to the original 5684 variables in the data matrix for Pd, Rh and Pt calibration

	Pt	Pd	Rh	Pt	Pd	Rh
No of PCs	Number of variables remaining after application of algorithms					
	UVE-PLS			IVD-PLS		
6	375	933	168	108	99	35
8	906	1186	501	73	103	165
10	1000	1174	537	119	245	140
	142	142	142	*	*	*
αn	142	142	142	*	*	*
	142	142	142	*	*	*
$\alpha n / nvars$	2.64	6.57	1.18	*	*	*
	6.38	8.35	3.52	*	*	*
	7.02	8.26	3.78	*	*	*

αn Number of Type I errors.
nvars Number of variables remaining after UVE-PLS.
Shaded Number of variables finally used in the PLS1.
* No t-test used in IVD-PLS

The IVD-PLS algorithm then reduced the remaining data by between 79-91%, depending on the number of PCs used. Because it was known that there were 8 elements present in the calibration solutions the number of PCs used was 6, 8 and 10 to prevent model under or overfitting.

The $\alpha n / nvars$ ratio, in conjunction with Table 4.8, shows that the lowest RRMSE % values were not generally obtained, as would be expected, with the highest ratio. In fact the variability of the RRMSE % value (Table 4.8) with the $\alpha n / nvars$ ratio was insignificant. This would seem to indicate that the data used, by the IVD-PLS algorithm, was not 'too diluted' by the presence of the uninformative, i.e. Type I variables. However,

for Rh the RRMSE value was 1.62 % with a corresponding $\alpha n / n var s$ ratio of only 1.18, suggesting that those variables ranked first, and having the largest var_j^{ivd} ratio, were truly informative as opposed to variables which were the result of Type I error.

Table 4.8 Synthetic test RRMSE % values for Pt, Pd and Rh using PLS1 with variable selection (6, 8 and 10 PCs), full spectra modelling and the data set containing 166 gross analyte and matrix lines.

	Pt			Pd			Rh		
Variable Selection [Mc]									
RRMSE %	5.27 (6)	5.18 (8)	5.58 (10)	2.51 (6)	2.33 (8)	2.56 (10)	1.66 (6)	1.52 (8)	1.62 (10)
Variables after t-test	375	906	1000	933	1186	1174	168	501	537
Variables selected	108	73	119	99*	103*	245*	35	165	140
Full Spectrum (5684 wavelength points) [As]									
RRMSE %	-	12.64(8)	-	-	8.31(8)	-	-	27.15	-
Individual Wavelengths (166 analyte & matrix lines) [As]									
RRMSE %	-	8.38(5)	-	-	7.06(5)	-	-	3.18(7)	-

* Variables were selected using the IVD plot (e.g. Fig 4.12)
[Mc] Data mean centered
[As] Data autoscaled

Inspection of Fig. 4.12 might lead to the assumption that a mixture of Type I error and truly informative variables were incorporated into the PLS model at first followed by only truly informative variables later, indicted by the sudden decrease in the RRMSE with the addition of the 11th variable subset. However, the type I error variables possess incidental chance correlation and as a result would also give reductions the RMSECV value for the calibration data. Because the $\alpha n / n var s$ ratio was low, it can be concluded that the relatively small number of truly informative variables (27) were of sufficient importance,

to limit the effect of the Type I error variables thereby giving a low Rh RRMSE for the predicted synthetic test data (Table 4.8). However, because of the higher probability of including Type I error variables in the final IVD-PLS data set with a low $\alpha n / nvars$ ratio, it is safer modelling practice to choose UVE-PLS data sets with a much higher $\alpha n / nvars$ ratio, thereby lowering the probability of including Type I error variables.

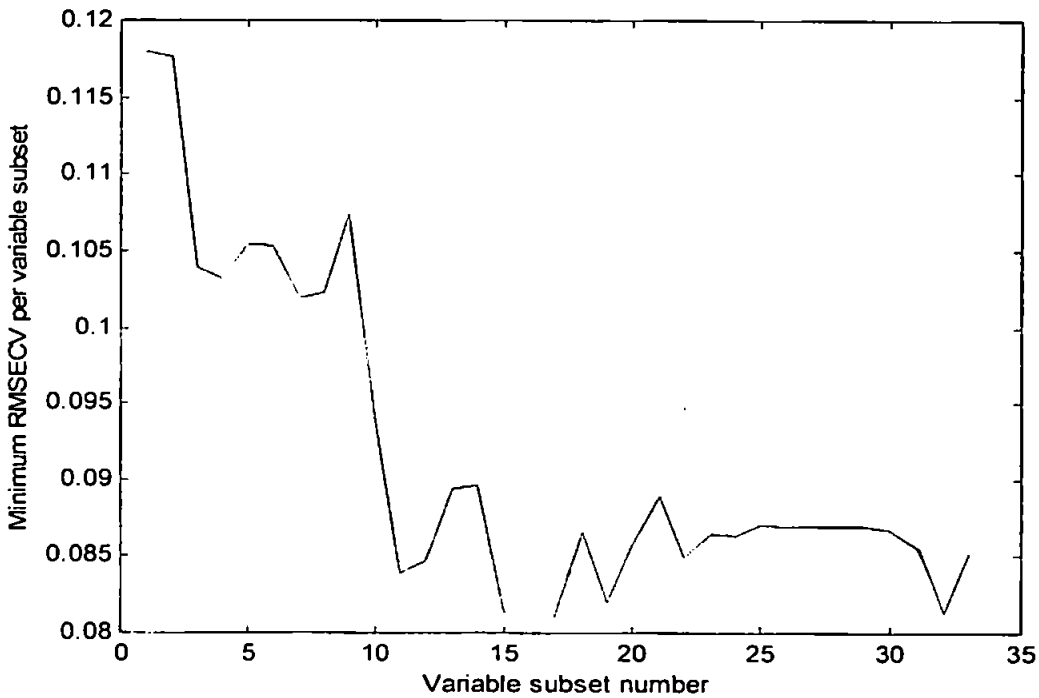


Figure 4.12 Minimum RMSECV for Rh (6 PCs) for each variable subset (step size of 2.5 % starting at 20%.

The overall effectiveness of the IVD-PLS algorithm when the $\alpha n / nvars$ ratio was much larger (6.57, Table 4.7) is illustrated in Fig. 4.13 for Pd. It is evident from Fig. 4.13a that a minimum RMSECV value of 0.515 was obtained when 771 wavelength data points were included in the model. However, a RMSECV value of approximately 0.535 was

obtained by using only 99 variables and it was found that, by inspection of these plots, together with consideration of the $\alpha n / n var s$ ratio, robust models could be constructed using a lower number of variables than the global minimum given by IVD-PLS with very little sacrifice in the RMSECV value.

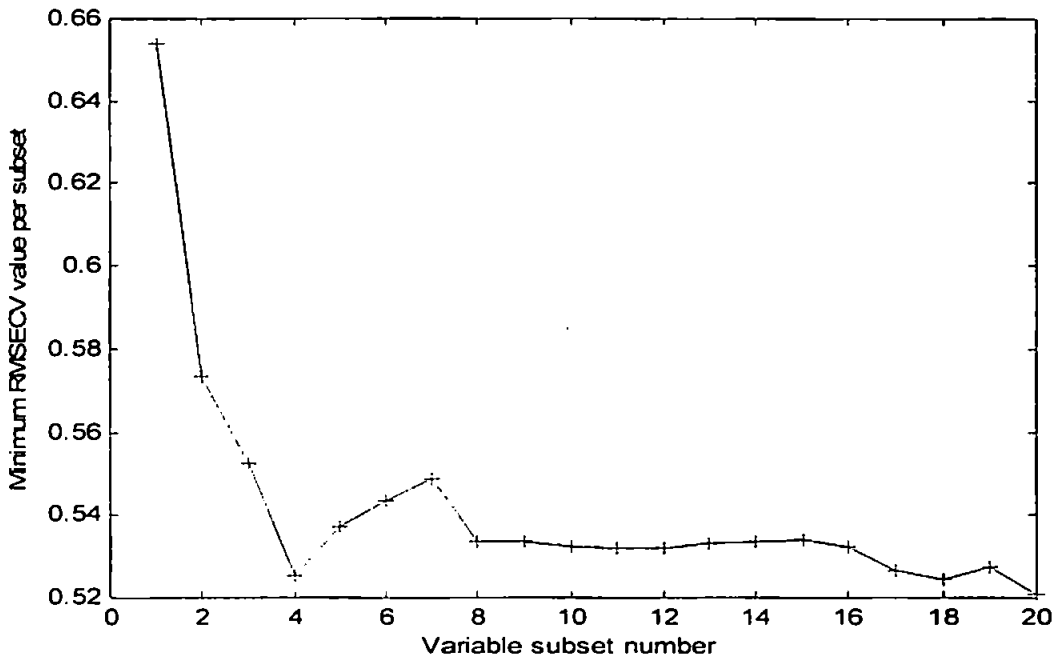


Figure 4.13a Minimum RMSECV for Pd (6 PCs) for each variable subset

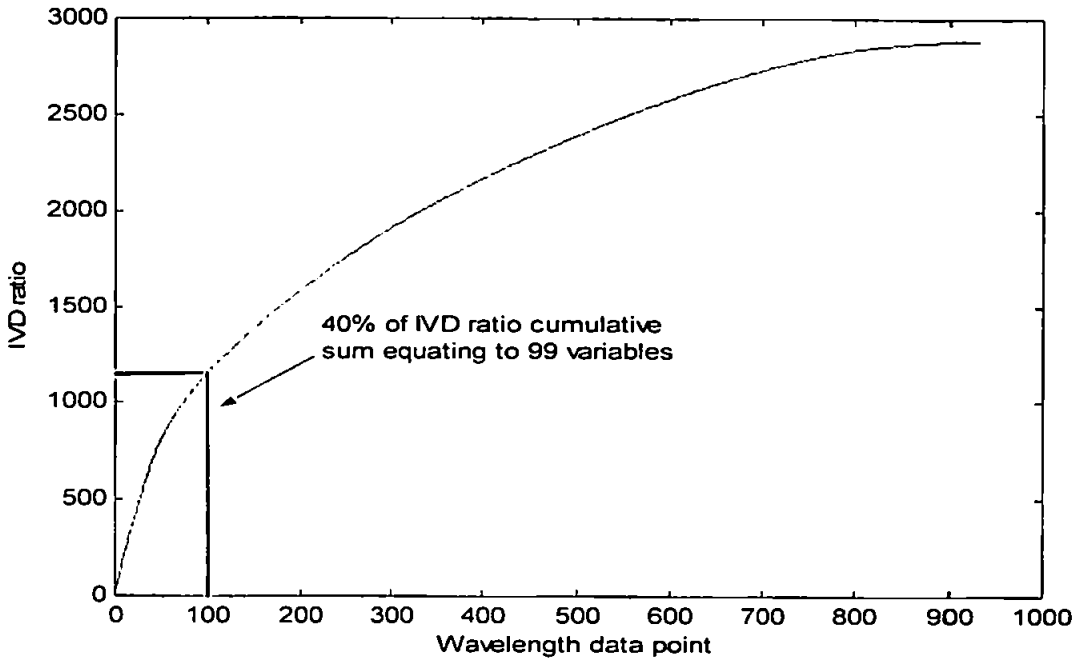


Figure 4.13b Cumulative sum of IVD ratio for Pd (6 PCs)
(Both Fig. 4.13a and 4.13b show the relevant plots for a step size of 5% starting at 5%, usually it is only necessary to begin the IVD-PLS routine at approximately 30% of the maximum IVD value, corresponding to point)

The effectiveness of the variable reduction algorithm in improving the quality of the PLS calibration models built using the data is illustrated by the synthetic test sample RRMSE values obtained for the prediction of Pt, Pd and Rh concentration (Table 4.8). In order to assess the effectiveness of the variable reduction routine, an 8 principal component PLS model was also built using the entire spectrum available with no variable selection. As can be seen from Table 4.8, there was a significant increase in the prediction accuracy of Pd following variable reduction compared to using the unreduced spectral data.

The synthetic test RRMSE values obtained using the reduced spectrum were 5.27, 2.33 and 1.52 % for Pt, Pd and Rh respectively, compared to 12.64, 8.31, and 27.15 % using the unreduced spectral data. The higher RRMSE values obtained when using the unreduced spectral data was probably due to the inclusion of too many uninformative

variables (i.e. noise) in the model, which were deleted using the UVE-PLS and IVD-PLS algorithms.

In contrast to infra-red and UV spectroscopy, where absorption bands are quite broad, atomic emission spectra are comprised of many narrow emission lines, of the order of 0.01 nm width, which can be extremely complex if even a few matrix elements with line-rich spectra are present.

Part of the emission spectrum for Pd is shown in Fig. 4.14 where all the calibration spectra are overlaid.. Also shown are the regions of the spectrum (i.e. the wavelength data points) which were selected by the variable reduction algorithms. It is evident from Fig. 4.14 that the selected parts of the spectrum were often co-incident with analyte lines for Pd (e.g. Pd 324.470 and Pd 340. 458) and also known Pd interferents such as Ce, Pt and Zr. Because PLS looks for linear combinations of variables for which variability is correlated to the analyte of interest it is to be expected that analyte interferents are also selected, as was the case here. On close examination of other parts of the spectrum, it was not obvious why particular spectral regions were selected by the algorithm (e.g. continuum background), however, it is quite possible that parts of the spectrum are correlated with non-spectroscopic matrix effects, such as suppression or enhancement of the emission signal, which can only be identified by the use of multivariate methods. This highlights an extremely desirable aspect of this method of selecting variables from the raw spectral data, namely that it is an objective rather than a subjective method of selecting the most informative variables, so that prejudgements about the usefulness or otherwise of parts of the spectrum are not necessary. Bearing this in mind, it is interesting to compare the RRMSE values obtained

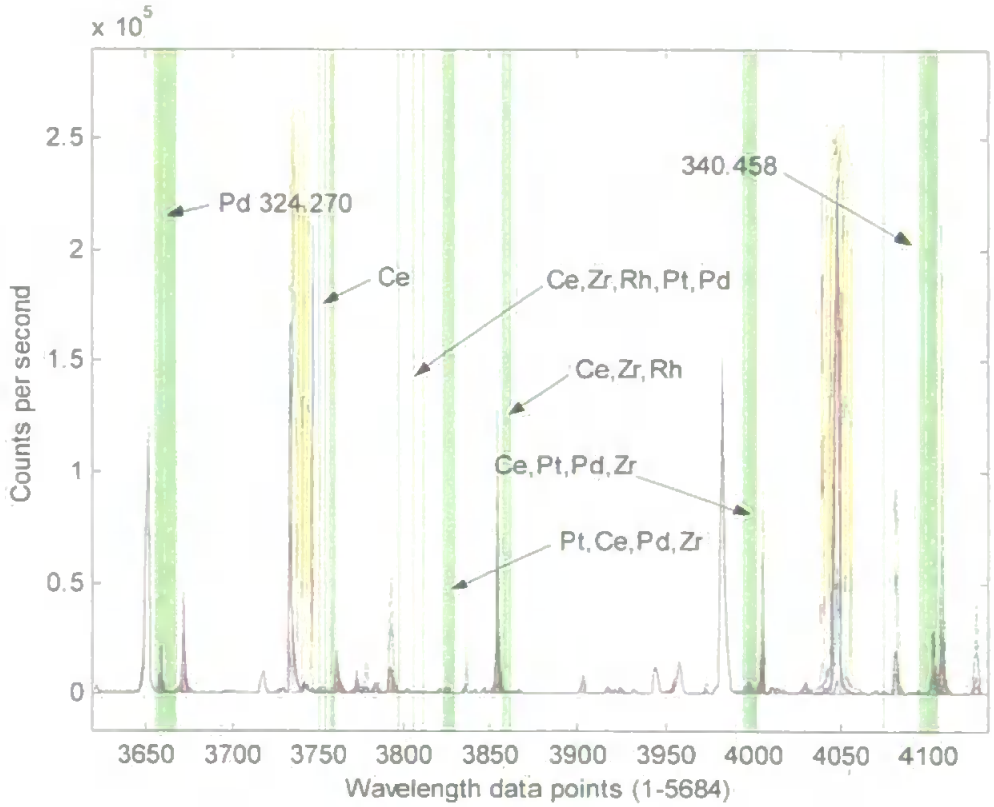


Figure 4.14 Partial spectrum (overlaid calibration samples (49)) showing the selected areas for the Pd model

when the calibration model was constructed using 166 individual spectral lines representing the most intense analyte and matrix lines in the spectrum, the traditional method. The variable reduction method resulted in an improvement compared to this approach, presumably because useful regions of the spectrum were not subjectively omitted from inclusion in the model, as they inevitably must be if preselected lines are used.

In order to avoid model under- or over-fitting the number of PCs was heuristically selected on the basis of the known number of elements in the data set. It was known that 8 elements were present in the samples, so it was assumed that between 6 and 10 PC's would

best model the system. It has been suggested that one should select the optimum number of PCs before variable selection has been carried out. However, this method can lead to the selection of a large number of PCs resulting in a grossly overfitted model, especially when line-rich spectra containing a large amount of noise are used. It is recognised that selection of the optimum number of PCs is not always an easy task especially if the method is to be applied to systems for which the exact chemical composition is unknown. In this case, the number of PCs can be selected based upon the 'bracketing approach', or 'boundary condition' adopted here, starting at some sensible number.

4.6.2 Multivariate Calibration for Quantitative Prediction of the High Training Range Data-Sets

In order to evaluate the usefulness of the variable selection algorithm for multivariate calibration and quantitative analysis, a series of independent test solutions were prepared with randomly chosen concentrations of the elements present in the calibration solutions. These were then analysed and the concentrations of Pt, Pd, and Rh predicted using the multivariate calibration model. In addition autocatalyst samples were also analysed and the same analytes determined.

Results are shown in Tables 4.9 – 4.12 and Figs. 4.15 - 4.17. The 95% confidence limits of prediction, obtained using the method described in the theory section, are given for the test samples in Table 4.9, and are shown as error bars in Figs. 4.15b, 4.16b and 4.17b for Pd, Pt and Rh respectively. In most cases the confidence interval for the predicted concentrations encompassed the known value. For the few exceptions to this (e.g. Pd at a known concentration of 18 $\mu g\ ml^{-1}$), the predicted concentrations were still very close to the known concentrations (Table 4.9).

Table 4.9 Upper and lower confidence intervals for the predicted test set concentrations ($\mu g\ ml^{-1}$) using those variables that gave the lowest RRMSE % test set value

Pt				Pd				Rh			
Concentration		Prediction Confidence Limits		Concentration		Prediction Confidence Limits		Concentration		Prediction Confidence Limits	
Actual	Predicted	L.C.L.	U.C.L.	Actual	Predicted	L.C.L.	U.C.L.	Actual	Predicted	L.C.L.	U.C.L.
12	11.70	11.12	12.28	20	19.98	19.80	20.15	3	2.96	2.83	3.08
16	15.96	15.32	16.60	12	11.81	11.47	12.10	5	4.96	4.80	5.12
20	19.50	18.84	20.16	18	17.46	17.21	17.61	2	2.04	1.93	2.14
12	11.61	11.07	12.15	14	13.76	13.59	13.81	4	4.02	3.87	4.17
18	17.24	16.80	17.69	10	9.86	9.61	9.92	3	3.04	2.85	3.23
6	6.14	5.79	6.49	30	29.68	29.52	29.94	1	1.04	0.91	1.18
2	2.16	1.72	2.60	6	5.95	5.75	6.00	2	1.98	1.83	2.14
40	38.15	37.54	38.76	2	1.86	1.63	2.26	8	8.04	7.90	8.18
30	28.65	27.96	29.35	40	39.16	38.80	39.40	10	9.85	9.70	10.00
0	0.43	-0.15	1.01	0	-0.15	-0.35	0.11	0	0.03	-0.11	0.18

Table 4.10 Actual and predicted concentrations for Pt ($\mu g\ ml^{-1}$)

Pt							
Test samples				Autocat samples			
	Concentration (μg/ml)		95%		oncentration (μg/ml)		95%
Sample	Actual	Predicted	C. I. (±)	Sample	Actual	Predicted	C. I. (±)
Te1	12	11.70	0.58	Au1R1	n/d	-2.05	n/g
Te2	16	15.96	0.64	Au1R2	n/d	-1.79	n/g
Te3	20	19.50	0.66	Au1R3	n/d	-1.89	n/g
Te4	12	11.61	0.54	Au2R1	1.01	-0.70	n/g
Te5	18	17.24	0.45	Au2R2	1.02	-1.32	n/g
Te6	6	6.14	0.35	Au2R3	0.99	-0.66	n/g
Te7	2	2.16	0.44	Au31R	n/d	-0.71	n/g
Te8	40	38.15	0.61	Au3R2	n/d	-0.82	n/g
Te9	30	28.65	0.70	Au3R3	n/d	-1.31	n/g
Te10	0	0.43	0.58	Au41R	n/d	-0.59	n/g
				Au4R2	n/d	-0.47	n/g
				Au4R3	n/d	-0.38	n/g
RRMSEP %		5.18				191.27	

n/g Not given
n/d Not detectable

Table 4.11 Actual and predicted concentrations for Pd (μg ml^{-1})

Pd							
Test samples				Autocat samples			
	Concentration (μg/ml)		95%		oncentration (μg/ml)		95%
Sample	Actual	Predicted	C. I. (±)	Sample	Actual	Predicted	C. I. (±)
Te1	20	19.98	0.17	Au1R1	1.07	2.76	0.4386
Te2	12	11.78	0.31	Au1R2	0.96	2.66	0.4263
Te3	18	17.41	0.20	Au1R3	0.90	2.67	0.4272
Te4	14	13.70	0.11	Au2R1	1.22	2.93	0.4244
Te5	10	9.76	0.15	Au2R2	1.19	2.92	0.4265
Te6	30	29.73	0.21	Au2R3	1.22	2.99	0.4305
Te7	6	5.87	0.12	Au31R	17.62	18.89	0.2977
Te8	2	1.95	0.32	Au3R2	17.86	18.93	0.296
Te9	40	39.10	0.30	Au3R3	17.35	18.98	0.2996
Te10	0	-0.12	0.23	Au41R	33.70	34.00	0.2963
				Au4R2	34.12	34.38	0.3037
				Au4R3	33.73	33.85	0.2989
RRMSEP %		2.51				10.44	

Table 4.12 Actual and predicted concentrations for Rh ($\mu g\ ml^{-1}$)

Rh							
Test samples				Autocat samples			
	Concentration (µg/ml)		95%		oncentration (µg/ml)		95%
Sample	Actual	Predicted	C. I. (±)	Sample	Actual	Predicted	C. I. (±)
Te1	3	2.96	0.13	Au1R1	n/d	0.10	n/g
Te2	5	4.96	0.16	Au1R2	n/d	0.10	n/g
Te3	2	2.04	0.11	Au1R3	n/d	0.11	n/g
Te4	4	4.02	0.15	Au2R1	3.31	0.15	n/g
Te5	3	3.04	0.19	Au2R2	3.32	0.10	n/g
Te6	1	1.04	0.13	Au2R3	3.21	0.12	n/g
Te7	2	1.98	0.15	Au31R	5.36	5.81	n/g
Te8	8	8.04	0.14	Au3R2	5.36	5.72	n/g
Te9	10	9.85	0.15	Au3R3	5.33	5.93	n/g
Te10	0	0.03	0.14	Au41R	0.05	-0.46	n/g
				Au4R2	0.03	-0.49	n/g
				Au4R3	0.03	-0.50	n/g
RRMSEP %		1.52				64.66	

n/g Not given
n/d Not detectable

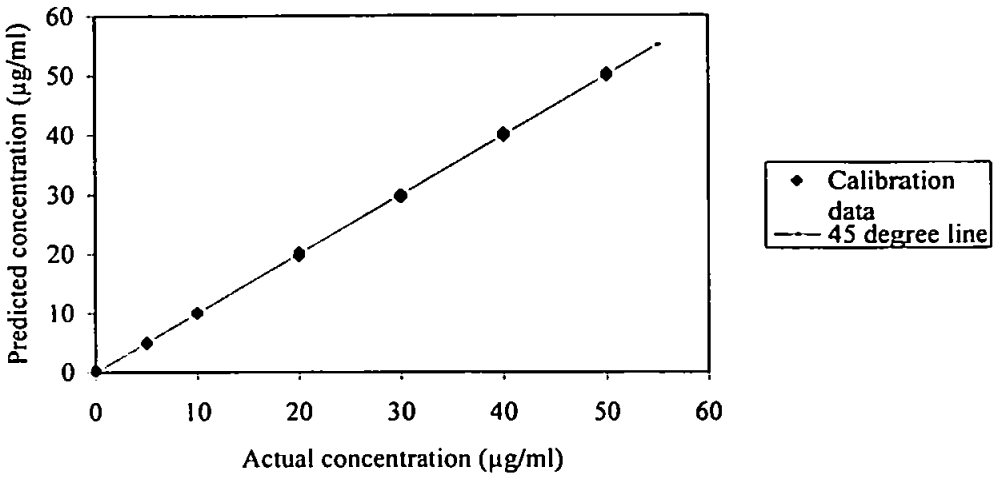
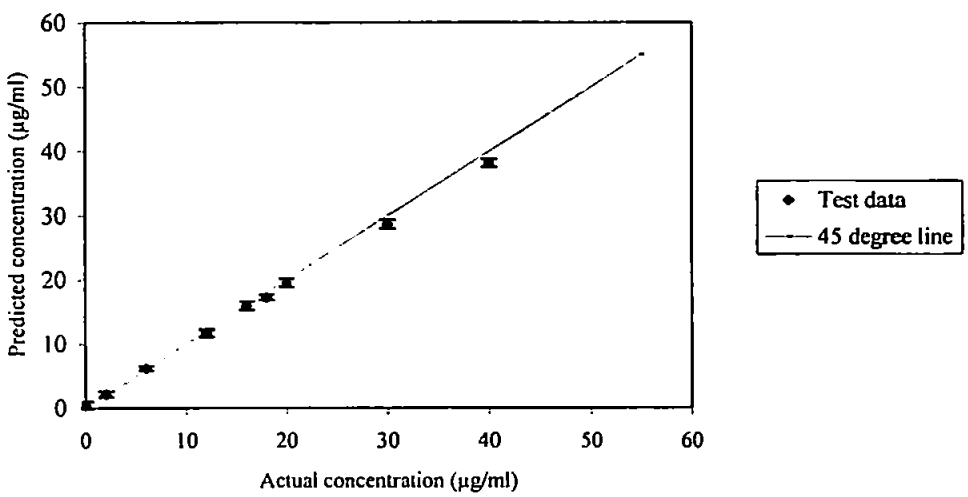


Figure 4.15a Actual vs predicted concentration (µg ml^{-1}) using variable selection for Pt for the calibration data.



	R2	Slope	Intercept
Calibration	0.999	0.999	0.012
Test	0.999	0.984	0.041

Figure 4.15b Actual vs predicted concentration (µg ml^{-1}) using variable selection for Pt for the independent test data with 95 % confidence interval.

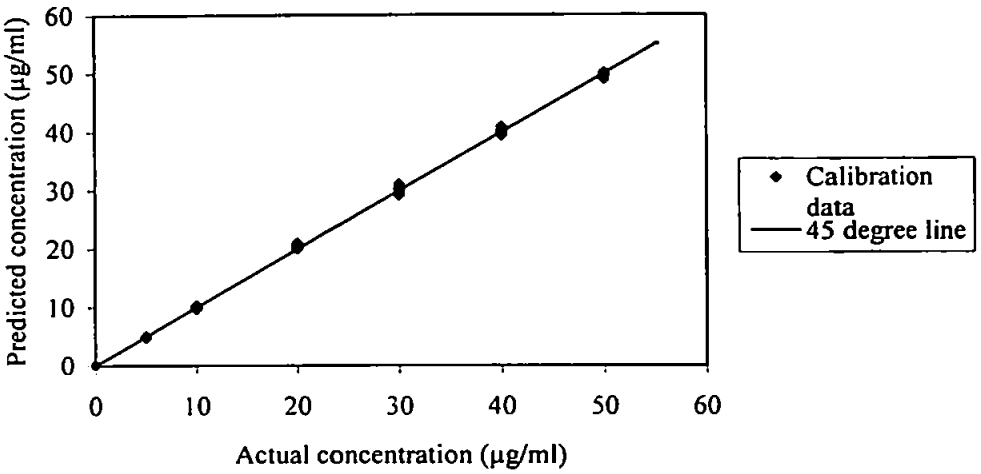
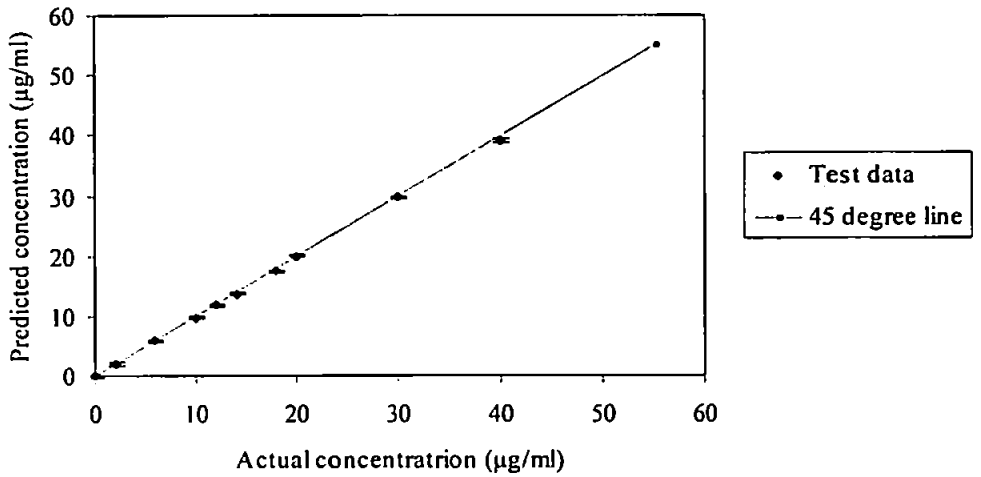


Figure 4.16a Actual vs predicted concentration (µg ml^{-1}) using variable selection for Pd for



the calibration data.

	R2	Slope	Intercept
Calibration	0.999	0.999	0.007
Test	0.999	0.944	0.419

Figure 4.16b Actual vs predicted concentration (µg ml^{-1}) using variable selection for Pd for the independent test data with 95% confidence interval.

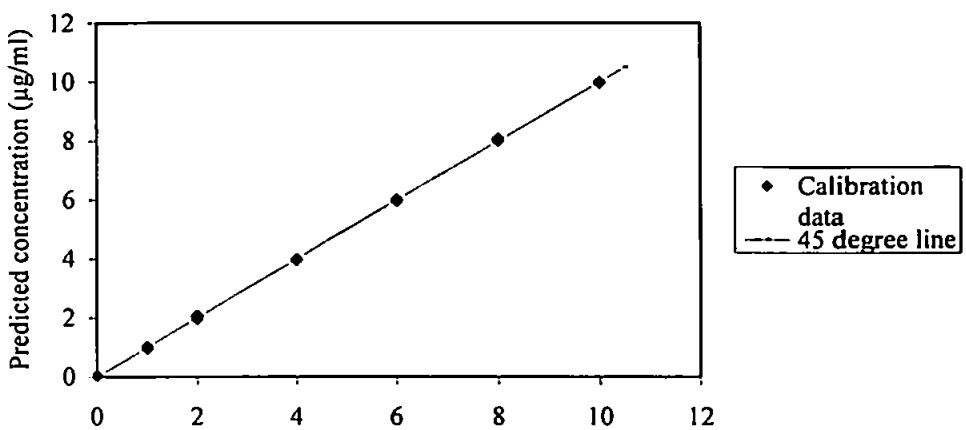
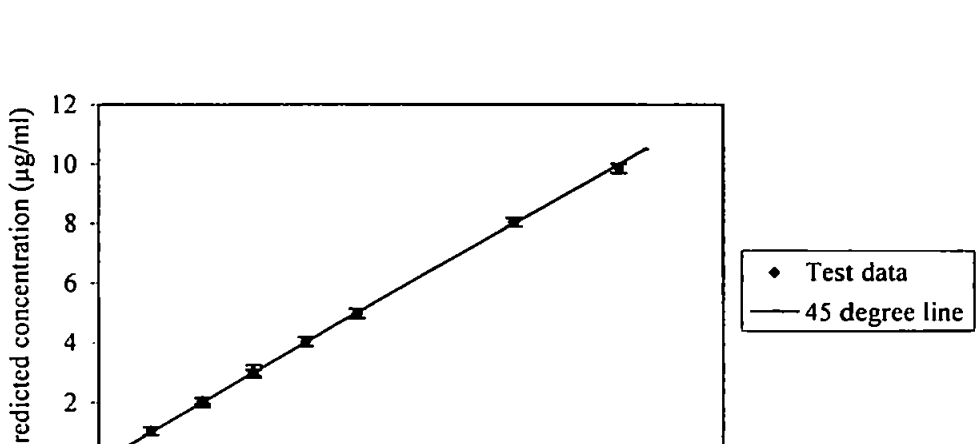


Figure 4.17a Actual vs predicted concentration (µg ml^{-1}) using variable selection for Rh for the calibration data.



	R2	Slope	Intercept
Calibration	0.999	0.999	0.006
Test	0.999	0.9878	0.043

Figure 4.17b Actual vs predicted concentration (µg ml^{-1}) using variable selection for Rh for both the independent test data with 95 % confidence interval.

The RRMSE values, using the individual lines, for the synthetic test samples were 8.38, 7.06 and 3.18 % for Pt, Pd and Rh (Table 4.5, section 4.3.2) respectively. This compares to 5.18, 2.51 and 1.52 % (Tables 4.10-4.12) respectively using the UVE and IVD-PLS algorithms, illustrating their effectiveness at variable selection. These low values, however, are not reflected in the autocatalyst samples where the RRMSE values were 191.27, 10.44 and 64.66 % for Pt, Pd and Rh (Tables 4.10 - 4.12) respectively using the variable selection method. This is most probably due to the presence of unmodelled interferents in the autocatalyst samples. This is also confirmed by the low RRMSE values for the test solutions using the entire spectrum, where the RRMSE values were only 12.64, 8.31 and 27.15 % for Pt, Pd and Rh respectively (Table 4.5, section 4.3.2). If any unmodelled interferents are present they will be accentuated by using the entire spectrum because of the increased amount of unmodelled data compared with the data set consisted of individual lines. Evidently the calibration samples covered the test sample factor space, but not that of the autocatalyst samples. The use of certified reference materials (CRMs), notably standard reference material 2556 (National Institute of Standards and Technology, NIST) - 'Used Autocatalyst' pellets was not analysed owing to the presence of non-certified values in the reference material which may render any experimental design ineffective.

4.6.3 Comparison of Variable Selection techniques

Both of the variable selection techniques used showed that they were capable of selecting suitable variables for the successful prediction of the synthetic test data Table 4.13). Using the method which selected variables based on the standard deviation of the intensity, the RRMSE values were 10.46, 4.39 and 13.23 % for Pt, Pd and Rh respectively.

However, significant improvements in these values were made by using of the UVE and IVD–PLS algorithms. The RRMSE values were reduced by 50 % for Pt (5.18%), 43 % for Pd (2.51 %) and 89 % for Rh (1.51 %). Although the reduction in the autocat RRMSE values followed the same trend, the values remained relatively high for both Pt (191.27 %) and Rh (64.66 %), with Pd at 10.44 % (Table 4.13). Compared with the more traditional method of selecting individual analyte and matrix lines, the UVE and IVD-PLS variable selection techniques gave significantly lower RRMSE values for the synthetic test solutions. The UVE and IVD-PLS algorithms resulted in a significant decrease in the autocatalyst RRMSE values also from 622, 1734 and 7840 % compared with 191.27, 10.44 and 64.66 % for Pt, Pd and Rh respectively. The high RRMSE values again indicate the possible presence of unmodelled interferents in the autocatalyst samples which were not incorporated into the calibration solutions. The best modelling to date, however, has been obtained with the use of 166 individual analyte and matrix lines. This has given test sample RRMSE values of 8.38, 7.06 and 3.18 % for Pt, Pd and Rh respectively, and 3.38 % for Pt in the autocatalyst sample. The relatively low autocatalyst RRMSE value for Pt (3.38 %) using only individual lines compared with the UVE/IVD-PLS RRMSE at over 190 % may indicate the presence of unmodelled interferents in the autocatalyst samples.

Table 4.13 Comparison of Pt, Pd and Rh RRMSE values (test and autocat solutions) for the different variable selection techniques.

	Test			Autocat		
	Pt	Pd	Rh	Pt	Pd	Rh
VR UVE and IVD-PLS [mean centered]	5.18	2.51	1.52	191.27	10.44	64.66
Number of variables	73	99*	165			
Standard deviation [autoscaled]	10.46	4.39	13.23	6750	16.34	141.34
Number of variables	1000	650	1250			
Full Spectrum [autoscaled]	12.64	8.31	27.15	622.66	1734	7480
Individual Wavelengths (166) [autoscaled]	8.38	7.06	3.18	3.38	n/a	n/a

* Variables were selected using the IVD plot (e.g. Fig 4.12a).
n/a Data was not available.

4.7 Application of UVE-PLS and IVD-PLS Algorithms to the Low Training Range Dataset

4.7.1 Results and Discussion

Application of the variable reduction algorithms resulted in the deletion of the majority of the original 5684 spectral data points (Table 4.14). It is evident that the UVE-PLS algorithm had a large effect, eliminating between 96-99% of the spectral data.

Table 4.14 Effect of applying the UVE-PLS and IVD-PLS algorithms to the original 5684 variables in the data matrix for Pd, Rh and Pt calibration

	Pt	Pd	Rh	Pt	Pd	Rh
No of PCs	Number of variables remaining after application of algorithms					
	UVE-PLS			IVD-PLS		
8	20	220	116	*	175	13
12	28	187	205	**	***	41
	142	142	142	**	**	**
αn	142	142	142	**	**	**
$\alpha n / n var s$	0.14	1.54	0.81	**	**	**

* IVD-PLS not carried out because UVE-PLS selected variables are most probably the result Type I errors.
** No t-test used in IVD-PLS algorithm.
*** IVD-PLS not performed because number of UVE-PLS variables decreased

The IVD-PLS algorithm then reduced the remaining data by between 45-87%, depending on the analyte. As the chemical rank of the LTR and HTR data sets were equal it was only necessary to use 8 PCs for the UVE-PLS algorithm (RRMSE % error values in Table 4.8 show very little change with the number of PCs chosen or with changes in the $\alpha n / n var s$ ratio). However, because the number of variables for Pt was less than αn, it was highly probable that all 20 variables were the result of Type I errors, and were therefore present through chance. Because of this UVE-PLS was run using 12 PCs in an attempt to increase the $\alpha n / n var s$ ratio. From Table 4.14 it can be seen that this was unsuccessful, and that the increase was insignificant. Because of this results are presented for Pd and Rh only. The $\alpha n / n var s$ ratio for Rh was < 0 at 0.81, this would seem to indicate that all variables were the result of Type I errors, however, this does not take into

account the variability of α with non-normal distributions. Therefore, because of its proximity to 1 modelling was performed.

The effectiveness of the IVD-PLS algorithm is illustrated in Fig. 4.18a, which shows the cumulative sum plot of the var_j^{ivd} ratios and the RMSECV values for Pd. It is evident from Fig. 4.18a that a minimum RMSECV value of 0.045 was obtained when only 13 wavelength data points were included in the model. Further inclusion of wavelength variables, however, causes the RMSECV value to increase from a minimum of approximately 0.045 to a maximum of approximately 0.0535. This was most probably the result of incorporating less informative variables into the PLS model.

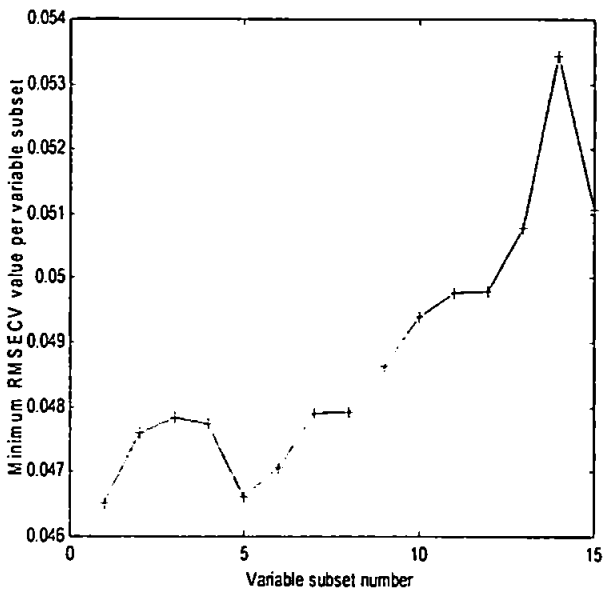


Figure 4.18a Minimum RMSECV for Pd (8 PCs) for each variable subset.

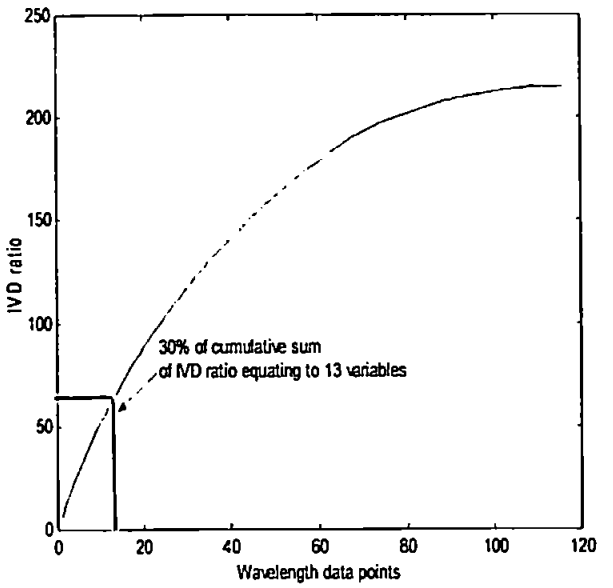


Figure 4.18b Cumulative sum of IVD ratio for Pd (8 PCs).
(Both Fig. 4.18a and 4.18b show the relevant plots for a step size of 5% starting at 30%)

The effectiveness of the variable reduction algorithm in improving the quality of the PLS calibration models is illustrated by the RRMSE values obtained for the external prediction of Pd and Rh in Table 4.15 and Figs. 4.19 and 4.20 for Pd and Rh respectively. In order to assess the effectiveness of the variable reduction routine, an 8 PC PLS model was also built using the entire spectrum available with no variable selection.

Table 4.15 shows that for the synthetic test samples there was a significant increase in the predictive accuracy using variable selection for Pd and Rh with RRMSE reductions of ~ 12 and ~ 43 % respectively compared to the individual wavelengths data. The high RRMSE value (~ 232) obtained for Pt, using the reduced spectral data, indicates that insufficient information was present in the predictive model.

Table 4.15 RRMSE values for Pt, Pd and Rh using PLS1 with variable reduction, full spectra modelling and the data set containing 164 lines.

	Synthetic test solutions RRMSE %			Autocat samples RRMSE %		
	Pt	Pd	Rh	Pt	Pd	Rh
VR [Mc]	232.21 (3)	50.72 (5)	44.64 (5)	79.83 (3)	27.13 (5)	*
Individual wavelengths (164) [As]	282.08 (4)	63.39 (10)	87.68 (8)	272.73 (4)	702.02 (10)	*
Full spectrum [Mc]	210.11 (8)	163.29 (8)	159.43 (8)	93.71 (8)	22.94 (8)	*

* Sample analyte concentration was outside of the calibration range
[Mc] Mean centered
[As] Autoscaled

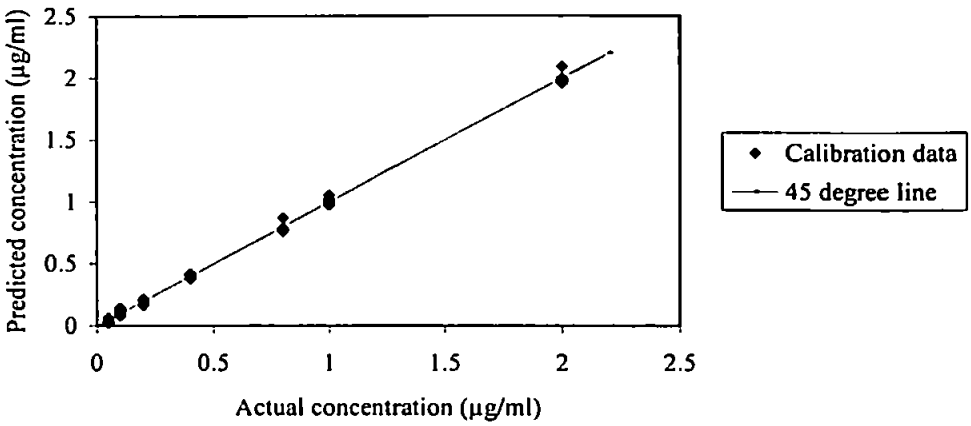
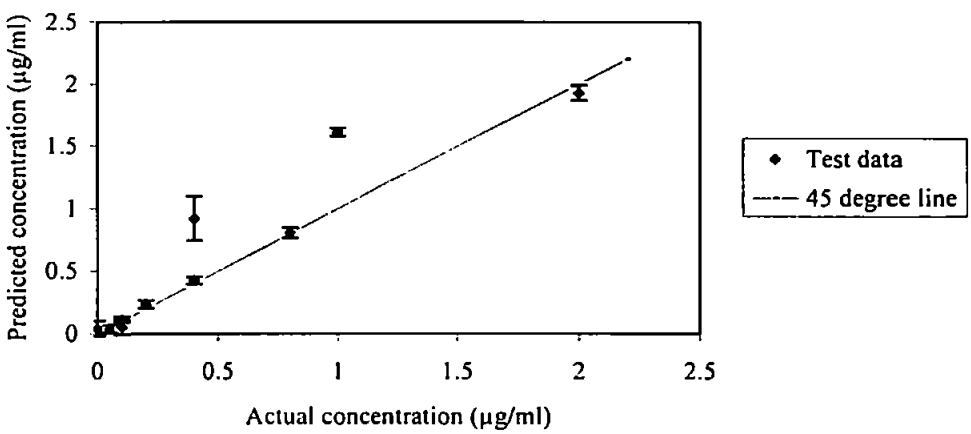


Figure 4.19a Actual vs predicted concentration (µg ml^{-1}) using variable selection for Pd for the calibration data.



	R2	Slope	Intercept
Calibration	0.998	0.998	0.001
Test	0.876	1.037	0.092

Figure 4.19b Actual vs predicted concentration (µg ml^{-1}) using variable selection for Pd for both the independent test data with 95 % confidence interval.

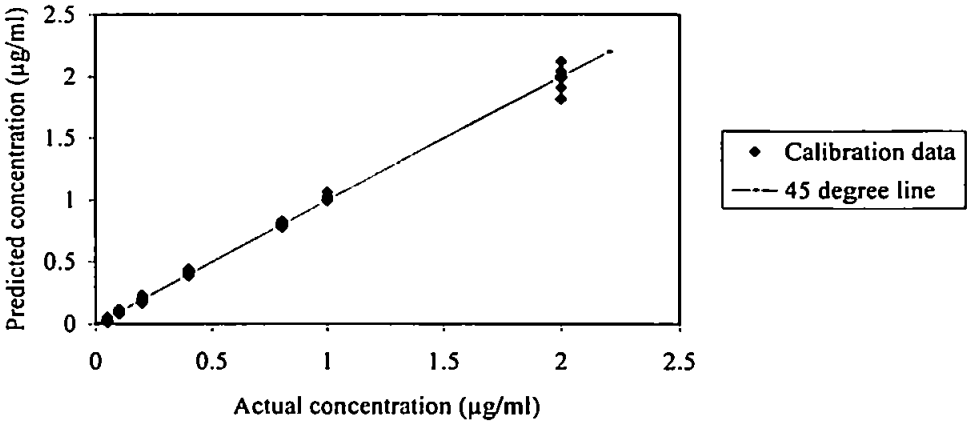
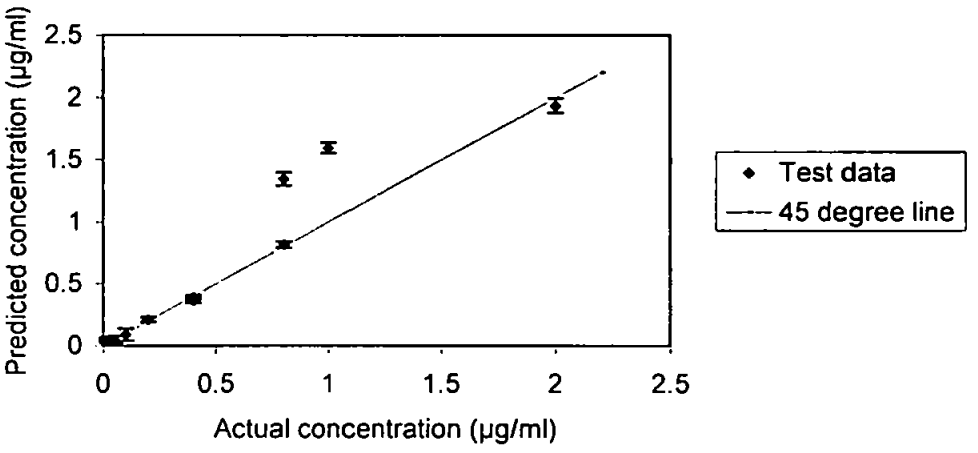


Figure 4.20a Actual vs predicted concentration (µg ml^{-1}) using variable selection for Rh for the calibration data.



	R2	Slope	Intercept
Calibration	0.996	0.996	0.002
Test	0.883	1.082	0.06

Figure 4.20b Actual vs predicted concentration (µg ml^{-1}) using variable selection for Rh for both the independent test data with 95 % confidence interval.

This is lent credibility by Fig. 4.21 which shows the IVD ratio ($var_j^{ivd} = \overline{\beta}_j / s(\beta_j)$) obtained from the IVD-PLS algorithm. Because the data was autoscaled prior to the use of the IVD-PLS algorithm, all variable coefficients were independent of scale, a small regression coefficient, therefore indicates an uninformative variable. From Fig. 4.21 it is evident that Pt had the lowest IVD values of the three analytes, indicating that very little predictive information was available to the PLS model.

The synthetic test RRMSE values obtained using the reduced spectrum were 50.72 and 44.64 % for Pd and Rh respectively (Table 4.15), compared with 2.51 and 1.52 % using the HTR (Table 4.8). The higher RRMSE values using the LTR data set, compared to those obtained for the HTR data set are most probably due to the lower analyte concentrations (Table 4.6) which have experienced the same magnitude of interference. The much larger difference between the analyte and matrix element concentrations may have 'masked' any analyte variation present at important spectral regions, thereby giving the PLS model poor predictive information. The synthetic test RRMSE values for Pd and Rh, using individual wavelengths, were 63.39 and 87.68 % respectively, compared with 163.29 and 159.43 % (Table 4.17 and 4.18) using the full spectrum. Evidently, for these analytes the selected wavelengths were more able to predict than the full spectrum most probably because of excluded noise. For Pt the RRMSE value obtained by using individual wavelengths and the full spectrum was 282.08 and 210.11 % (Table 4.16) respectively, indicating that the wavelengths used had missed the important information that was captured by using the full spectrum.

For the autocat samples the Pd RRMSE value using variable selection was 27.13 % (Table 4.15). When compared with the RRMSE values using the individual wavelengths data, the drop was significant with a reduction of ~ 674 % (Table 4.15). Predictive errors for Rh are not shown because all sample concentrations lay outside of the calibration

range. The RRMSE values for Pd using the full spectrum was 22.94 % (Table 4.15). This value was comparable with that obtained using variable selection. Evidently, by selecting individual wavelengths important information has been omitted which is not the case with either the full spectrum or the variable selection methods. This trend is also followed by Pt, where the RRMSE values are 272.73 and 93.71 % using 164 individual wavelengths (I.W.) and the full spectrum (F.S.) respectively (Tables 4.16 – 4.18 for Pt, Pd and Rh respectively).

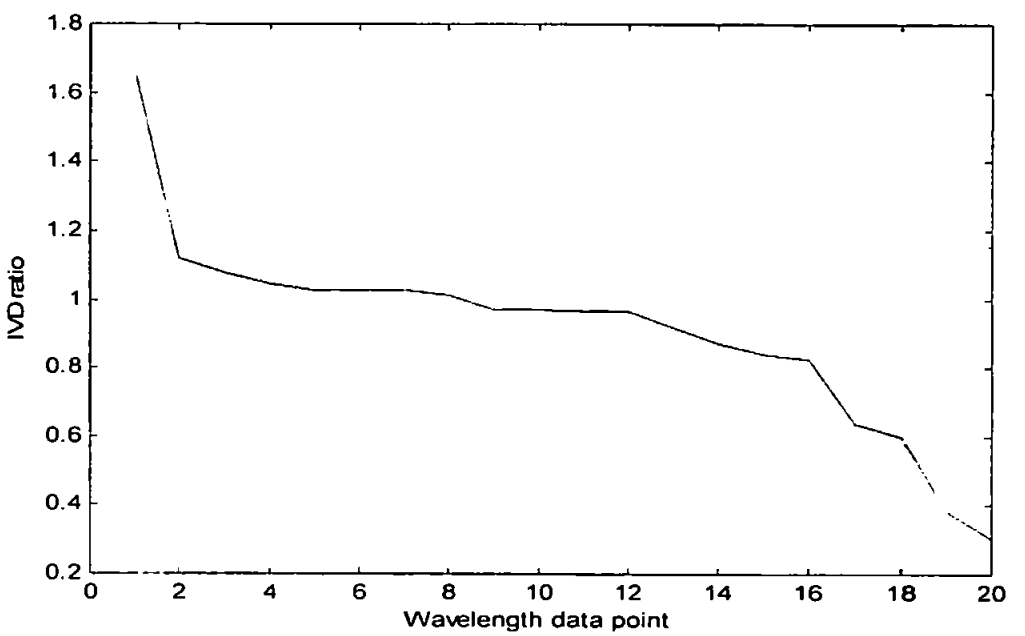


Figure 4.21a IVD ratio versus wavelength data point for Pt

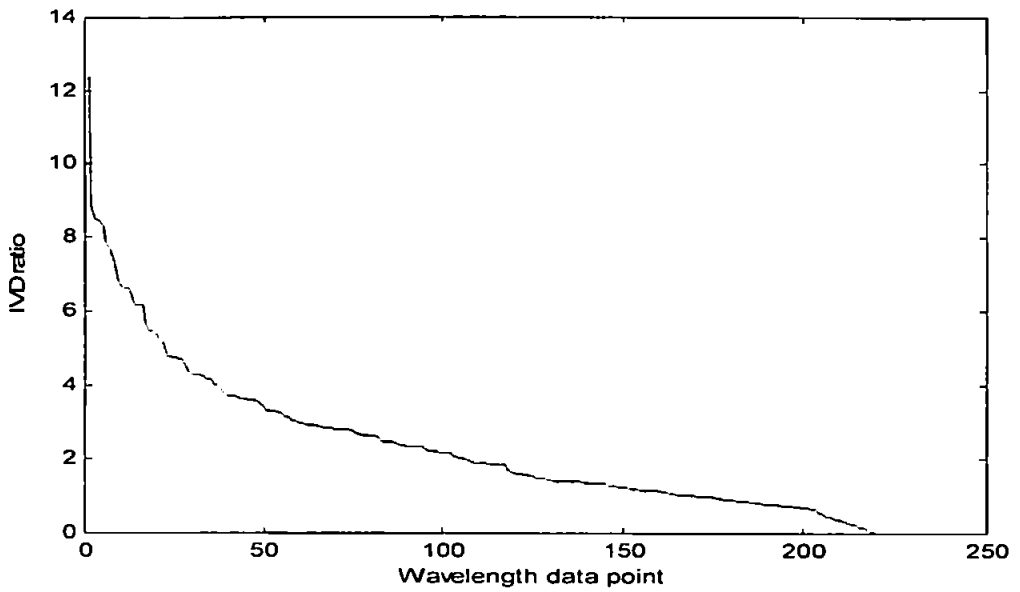


Figure 4.21b IVD ratio versus wavelength data point for Pd

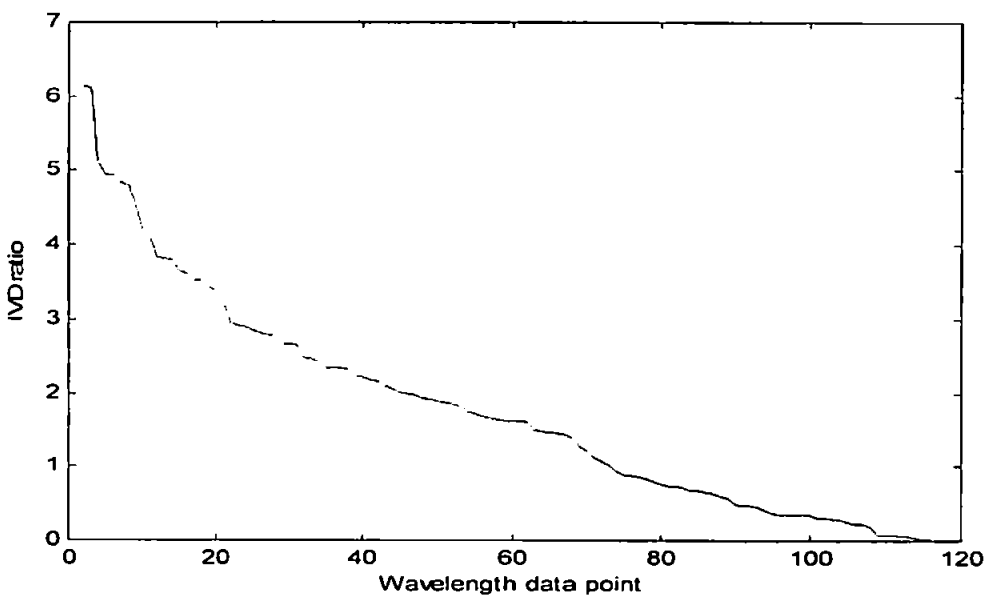


Figure 4.21c IVD ratio versus wavelength data point for Rh

Table 4.16 Actual and predicted concentrations for Pt ($\mu g\ ml^{-1}$)

Pt							
Test samples Concentration (μg/ml)				Autocat samples Concentration (μg/ml)			
Sample	Actual	Predicted (164 I.W.)	Predicted (F. S.)	Sample	Actual	Predicted (164 I.W.)	Predicted (F.S.)
Te1	0.05	1.85	1.73	Au1R1	1.01	3.57	2.742
Te2	0.40	1.71	2.11	Au1R2	1.02	14.43	2.7955
Te3	0.20	1.70	1.46	Au1R3	0.99	3.30	2.7938
Te4	0.40	2.43	1.21	Au2R1	1.99	6.06	2.8023
Te5	1.00	3.21	1.89	Au2R2	1.88	1.15	2.799
Te6	0.80	1.93	1.78	Au2R3	1.99	3.29	2.7997
Te7	0.80	1.89	1.56				
Te8	2.00	1.65	1.78				
Te9	0.05	1.80	1.45				
Te10	0.00	2.00	1.42				
RRMSEP %		282.08	210.11			272.73	93.71

I. W. individual wavelengths
F. S. full spectrum

Table 4.17 Actual and predicted concentrations for Pd (μg ml^{-1})

Pd							
Test samples Concentration (μg/ml)				Autocat samples Concentration (μg/ml)			
Sample	Actual	Predicted (164 I.W.)	Predicted (F. S.)	Sample	Actual	Predicted (164 I.W.)	Predicted (F.S.)
Te1	0.1	0.09	1.43	Au1R1	1.22	7.09	1.15
Te2	0.05	0.01	-0.55	Au1R2	1.19	17.18	0.96
Te3	0.1	0.10	0.29	Au1R3	1.22	8.97	1.01
Te4	0.8	0.84	1.16	Au2R1	1.50	11.25	1.06
Te5	0.4	1.06	0.05	Au2R2	1.48	6.85	1.13
Te6	0.2	0.37	-0.02	Au2R3	1.49	9.55	1.10
Te7	1	1.72	1.33				
Te8	2	2.03	0.12				
Te9	0.4	0.52	1.21				
Te10	0	0.14	0.24				
RRMSEP %		63.39	163.29			702.02	22.94

I. W. individual wavelengths
F. S. full spectrum

Table 4.18 Actual and predicted concentrations for Rh (μg ml^{-1})

Rh							
	Test samples Concentration (μg/ml)				Autocat samples Concentration (μg/ml)		
Sample	Actual	Predicted (164 I.W.)	Predicted (F. S.)	Sample	Actual	Predicted (164 I.W.)	Predicted (F.S.)
Te1	0.40	0.39	0.25	*	*	*	*
Te2	0.10	0.15	2.33				
Te3	0.05	0.19	0.48				
Te4	0.20	0.37	1.01				
Te5	0.80	1.81	1.61				
Te6	0.40	0.62	1.13				
Te7	2.00	2.31	1.21				
Te8	1.00	2.13	1.10				
Te9	0.80	0.98	0.52				
Te10	0.00	0.09	0.83				
RRMSEP %		87.68	159.43				

* samples out of calibration range
I. W. individual wavelengths
F. S. full spectrum

The emission spectrum for all calibration solutions is shown in Fig. 4.22. Also shown are the regions of this spectrum (i.e. the wavelength data points) which were selected by the variable reduction algorithms for the prediction of Rh. Because of the relatively few lines chosen (Fig. 4.22) it was not obvious why particular spectral regions were selected by the algorithm. Because PLS looks for linear combinations of variables for which variability is correlated to the analyte of interest it is to be expected that analyte interferents are selected, e.g. Pt 240.309, Ce 382.000 or Ba 233.527 nm. It is interesting to compare the RRMSE values obtained when the calibration model was constructed using 164 individual spectral lines (Table 4.11) representing the most intense analyte and matrix lines in the spectrum. The variable reduction method resulted in an improvement compared to this approach, presumably because useful regions of the spectrum were not subjectively omitted from inclusion in the model, as they inevitably must be if preselected lines are used.

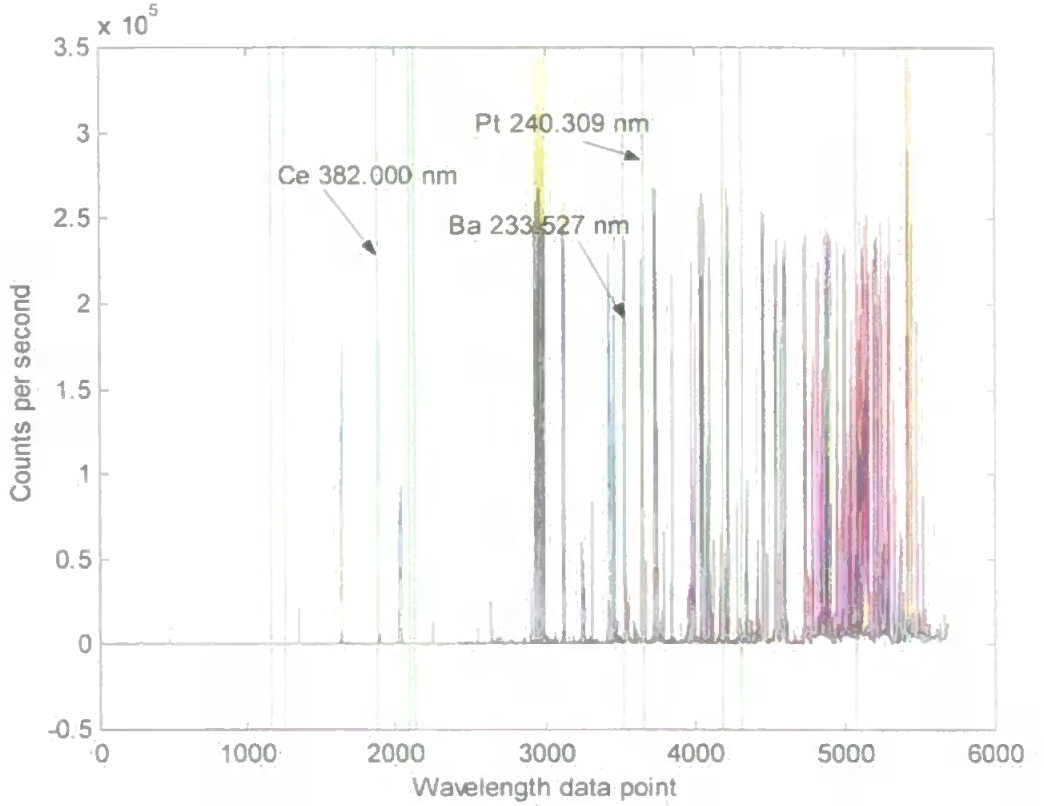


Figure 4.22 Emission spectrum for all calibration samples with those regions selected for the prediction of Rh.

4.8 Conclusions

The application of variable elimination and selection algorithms has shown that it is possible to use the complete available segmented ICP-AES emission spectrum for multivariate modelling without having to resort to line selection or the need to assign background correction points in order to obtain the net analyte signal of individual lines. Indeed one of the benefits of this approach is the selection of parts of the spectrum which appear uninformative, such as continuum background, but which can be highly informative to a bilinear modelling technique such as PLS, which is able to detect useful variation in all parts of the spectrum. The new method has several desirable properties: it is computationally simple, it has significance tests of multivariate model parameters and allows the calculation of independent test data confidence intervals.

The model errors for the independent test data, have shown considerable improvement compared with the errors achieved when using all 5684 wavelengths and marginal improvement compared with the more traditional individual wavelength data set consisting of intense analyte and matrix lines, thus reinforcing the fact that PLS does indeed benefit from selective variable reduction. Of greater interest is the important advantage of being able to utilise the full spectrum because the need for individual line selection, normally based upon expert knowledge of the chemical system is no longer required. In situations where such systems are not fully understood the selection of pertinent individual lines may prove impossible. The fundamental techniques used in this method of variable reduction are applicable to all forms of spectroscopy. The spectra from an ICP-AES instrument, compared to that of a broad spectrum instrument, for example, is relatively complex, as such comparable results should be obtainable for other classes of spectroscopy.

The limitations of the techniques were very different. The use of signal intensity standard deviation was not analyte specific, as such some of the selected variables possessed very little correlation. The use of relative standard deviation was discussed but this would give noisy regions of the spectrum the same weight as informative regions. The use of PLS regression coefficients also had limitations in that the correct number of PCs was required prior to UVE-PLS to estimate the coefficients, thereby leading to a non unique solution. It is recognised that selecting a range of PC's to use is not always an easy task, especially if the method is to be applied to data where the exact chemical composition is unknown. However, by using this heuristic bracketing approach, in conjunction with the variable selection algorithms developed, it has been shown that it is possible to identify those spectral areas within an ICP-AES emission spectrum that can be used in subsequent modelling. The ratio of Type I error variables to truly informative variables proved only to be an important issue when < 0 with these particular datasets. However, this may not be the

case when applied to other chemical systems where the number and strength, in terms of PLS regression coefficient, of informative spectral regions is much less. However, given its limitations this method was analyte specific and was therefore chosen as the most appropriate of the two with which to continue full spectrum investigations.

CHAPTER 5 - PLS AND VARIABLE SELECTION USING THE FULL AVAILABLE SPECTRUMA AND COMPLEX INDUSTRIAL SAMPLES

5.1 Introduction

The theory of variable selection has up to this point been developed using complex synthetic samples containing no more 11 elements (3 analytes, 5 matrix elements and 3 internal standard elements which were subsequently not used). In this section the procedure will be applied to real industrial samples containing combinations of over 40 elements, including Pt, Pd, Rh, Al, Mg, Ce, Zr, Fe, Si, Au, Ag, Ir, As, Cu, Ca, B and Li. The elements Fe, Mg and Al all have extremely line-rich spectra in the 200-450 nm wavelength region thereby posing problems for the determination of elements with emission lines in this region of the spectrum. The complexity of the samples is the result of two factors: first, the samples covered a very broad range of sample types, including electronic scrap, spent catalysts, alumina, silica, and base metal matrices, to name but a few (Table 5.1). The samples were obtained from Johnson Matthey (JM) and were subsamples of material that was to be processed by JM to reclaim the precious PGM metals (e.g. Au, Ag, Pt, Pd and Rh). They had been subdivided by the company into those main categories given in Table 5.1; second, the samples had been digested using a peroxide fusion method, with subsequent digestion in hydrochloric acid, so all the matrix elements were still present.

Because of this complexity it was not possible to use an experimental design to produce a calibration data set covering all the factor space so an alternative method was required. The samples had previously been analysed by JM in two ways. First, they had been subjected to a NiS fire assay (140) procedure to separate the matrix from the precious

metals and the precious metals had then been determined by ICP-AES using a combination of univariate IEC factors and matrix matching. Furthermore the samples had also been determined by independent umpire laboratories and a set of mean values for the precious metals were obtained. Second, the samples which had been subjected to the peroxide fusion digestion, mentioned above, which did not separate the matrix, had their precious and base metal content determined by ICP-AES using suitable wavelengths and a combination of IEC factors and matrix matching. Hence, two sets of data were available with corresponding values for the concentrations of the precious metals in each sample. The objective of the study was to use multivariate calibration to correct for interference in the samples which had been digested by the peroxide fusion (i.e. the samples with large concentrations of matrix elements). To this end, the spectral data from the ICP-AES analysis could be used to build multivariate calibration models and the concentration values obtained for the precious metals obtained using the NiS fire assay method could be used as standard concentration values. The analytes available to this approach were gold (Au), silver (Ag) and palladium (Pd).

Because these samples had been analysed by JM using the Perkin-Elmer optima 3000 ICP-AES specifically for their own work, only 38 % of the segmented spectrum was available, consisting of all spectral points from 76 subarrays (out of a possible 201). Nevertheless, it was considered that the data would be suitable for the variable selection algorithms developed in Chapter 4 and that it would still contain a substantial amount of noise which could be identified and removed.

Table 5.1 Sample classification

Assay Group	Sample Type	Definition (%)	% of Total
1	High silver	$Ag \geqq 10$	3.2
2	Barium Titanate	$Ba \geqq 10$, $Ti \geqq 10$	8.7
4	High Iron	$Fe \geqq 25$, $S \leqq 5$	7.6
5	Calcia	$CaO \geqq 30$	1.4
6	Alumina	$Al_2O_3 \geqq 25$, $SiO_2 \leqq 10$	11.0
7	Silica	$SiO_2 \geqq 25$, $Al_2O_3 \leqq 10$	7.9
8	Matte	$S \geqq 10$, $Cu + Ni + Fe \geqq 10$	5.5
9	Silica/Alumina	$SiO_2 \geqq 10$, $Al_2O_3 \geqq 10$	21.8
10	Carbon base	Sum (base metal) < 10	3.5
11	Mixed base metal	Not as above	20.9
13	High PGM	Any PGM > 20	8.5
	TOTAL		**100**

5.2 Experimental

5.2.1 Instrumentation and Reagents

All peroxide fusion samples were analysed by ICP-AES (for instrument and reagents information see Chapter 3, section 3.2.1) and their base and precious metal content determined. Instrumental conditions were optimised (section 3.2.1) and the wavelengths chosen corresponding to the major analyte and matrix lines. This made available the segmented spectrum from 76 out of a possible 201 subarrays.

5.2.2 Procedures

5.2.2.1 Selection of Calibration and Validation Samples using Principal Components Analysis (PCA)

5.2.2.1.1 Theory

In order to perform any calibration experiment suitable calibration samples must be used. In a situation where structured experimental design can be used this is a matter of deciding which design to choose (e.g. factorial, partial factorial or orthogonal array e.t.c). If the samples under examination are very complex (in this case > 40 elements), or if the elemental content of future samples is known to vary considerably, the alternative is to use 'historical data' in a similar manner to that used for process analysis. However, for this to give satisfactory predictions, three criteria must be met: there must be sufficient calibration samples from which to select a suitable subset; the elemental composition must be similar and the analyte concentrations used in the calibration should be accurate. It would be preferable to use certified concentration values, however, as this is seldom available in an industrial context, an alternative is to use concentration values that have been determined by some independent method and which have then been subsequently 'validated'. In this case all the analyte concentrations were independently analysed by a number of umpire laboratories and validated by comparison with a NiS fire assay.

The samples were first divided into 3 subsets based upon their analyte (Au, Ag or Pd) content, this ensured that only samples containing a specific analyte were used for the prediction of the validation samples containing the same analyte. It is then important to decide which samples, in each of the three subsets, are of similar composition with respect to their base metal content. This was performed using principal components analysis (PCA) and the Hotelling t-test.

Hotelling originally introduced the T^2 statistic in 1931 as a multivariate analogue of the univariate t^2 statistic of the Irish statistician, William Sealy Gossett (1876-1937), more familiar in the guise of his pseudonymic 'Student' t-test for the difference between two means (141). First consider a bivariate example in which two analytes, x and y are assumed to have normal distributions with means $\bar{x}$ and $\bar{y}$ variances (i.e. standard deviation squared) as in Eqns. 5.1 and 5.2:

$$S_x^2 = \sum_{i=1,n}(x_i - \bar{x})^2 /(n-1) \qquad \text{Eqn. 5.1}$$

$$S_y^2 = \sum_{i=1,n}(y_i - \bar{y})^2 /(n-1) \qquad \text{Eqn. 5.2}$$

and covariance as in Eqn. 5.3:

$$S_{xy} = \sum_{i=1,n}(x_i - \bar{x})(y_i - \bar{y})/(n-1) \qquad \text{Eqn. 5.3}$$

where n is the number of determinations on which the estimated values of $\bar{x}$, $\bar{y}$, S_x^2, S_y^2, and S_{xy} are based. The test statistic (T^2) is then given by Eqn. 5.4:

$$T^2 = c(Z_x^2 - 2\rho Z_x Z_y + Z_y^2) \qquad \text{Eqn. 5.4}$$

where $Z_x = (x - \bar{x})/S_x$ and $Z_y = (y - \bar{y})/S_y$ and $c = 1/(1-\rho^2)$ where $\rho = S_{xy}/(S_x S_y)$.

Equation 5.4 however, is only suitable for bivariate data and cannot be used with

multivariate data which requires the use of PCA.

The method of PCA has already been described in detail in section 1.4.2.2. If it is assumed that the scores are normally distributed it is possible to calculate confidence limits for the scores. The statistical confidence limit, the value for T^2 can be calculated by means of the F-distribution as shown in Eqn. 5.1:

$$T^2_{k,m,\alpha} = \frac{k(m-1)}{m-k} F_{k,m-k,\alpha} \qquad \text{Eqn. 5.1}$$

where m is the number of samples, k is the number of PCs, α is the confidence interval chosen (e.g. 95%) and F is the F-statistic for the appropriate degrees of freedom. For any two PCs, $T^2_{k,m,\alpha}$ describes an ellipse in the PC-space defined by PC_n and PC_{n+1} (142)(where $n > 1$). Each individual sample's t-value is given by Eqn. 5.2 (143):

$$T_i^2 = \boldsymbol{t}_i \lambda^{-1} \boldsymbol{t}_i' = \boldsymbol{x}_i \boldsymbol{P} \lambda^{-1} \boldsymbol{P}' \boldsymbol{x}_i' \qquad \text{Eqn. 5.2}$$

where $\boldsymbol{t}_i$ refers to the i-*th* row of the matrix of scores vectors from the PCA model and λ^{-1} is the diagonal matrix containing the inverse of the eigenvalues associated with the k eigenvectors retained in the model. For $T_i^2 > T^2_{k,m,\alpha}$, samples are situated outside of the ellipse indicating that they are outliers, as shown in Fig. 5.1.

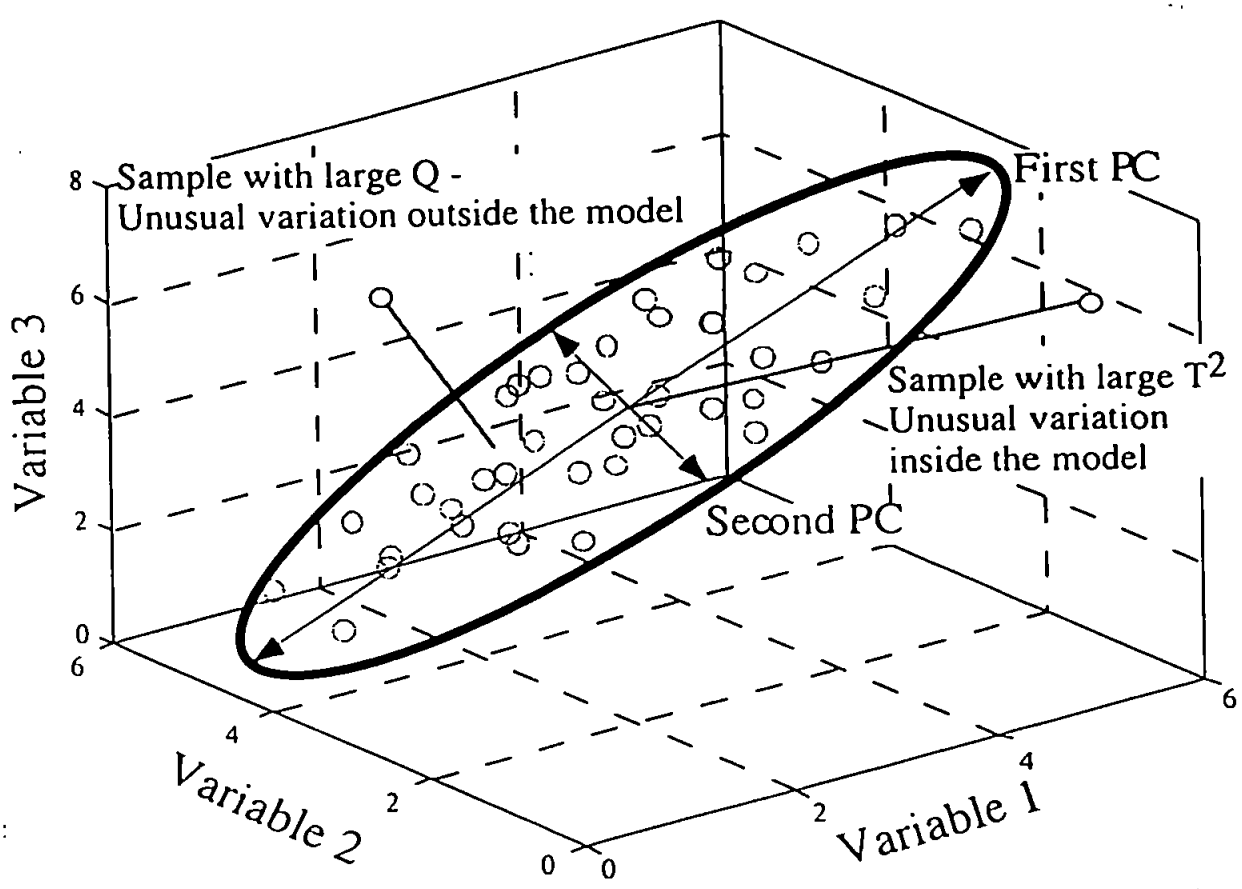


Figure 5.1 Graphical representation of Hotelling's T^2 confidence ellipse

It is important to note that the T^2 statistic makes the assumption that the distribution of the data, is multivariate normal. If $u = (u_1, \ldots u_k)$ is distributed such that the u_i are all independent normal univariate random variables, then u is said to have a multivariate normal distribution. Clearly data is often not normally distributed, however, the central limit theorem states that the sums of several different groups will tend to be normally distributed, regardless of the probability distribution of the individual groups (144) (145). This suggests that factor-based methods, such as PCA where the scores are a weighted sum of individual variables, will tend to produce measures that are more normally distributed then the original data (143).

This multivariate equivalent of the Student's t-test can be used to detect those samples with different composition (sample outliers) in a data set. Given a set of samples $\mathbf{S} = (s_1, \ldots s_2)$ the corresponding data matrix will be $\mathbf{X} = (x_1, \ldots x_k)$ and of size m x n (where m is the number of samples and n the number of variables). Prior to performing PCA the data set $\mathbf{X}$ is autoscaled. This has the effect of making the variance constant across the measurements (i.e. each variable has the same weighting in the model (146)). This procedure is very important for spectral interpretation because the intensity of a wavelength has no correlation with wavelength importance (147). The PCA model can then be built and the scores vectors, $\mathbf{t}$, for each PC plotted. Since the maximum amount of variance is captured in the early PCs it is only necessary to produce a plot of the early score vectors (Fig. 5.1) in order to detect unusual samples. The plot reveals how the samples are related to each other. Samples that are close together are similar with respect to the original measurements provided the plot displays a sufficient amount of variation. Since this mathematical proximity translates to chemical similarity if meaningful measurements have been made (46) unusual samples can be identified on the basis of their composition. The degree of confidence for the T^2 confidence region is then selected (e.g. 95%, $\alpha = 0.05$).

5.2.2.1.2 Procedure

The emission data from the peroxide fusion samples was first autoscaled ensuring that each variable was considered equally important (the advantages and disadvantages of different types of preprocessing prior to sample classification is given in section 5.3). The samples were first divided into 3 subsets based upon their analyte (Au, Ag or Pd) content and then subjected to PCA. The degree of confidence, α, was chosen to be 95% (i.e.

$\alpha = 0.05$, a one sided t-test was used). After each PCA model was built the T^2 was calculated and the correct number of outlying samples calculated according to $\alpha \times m$. If the number of outlying samples exceeded $\alpha \times m$ all outlying samples were deleted. A PCA model was then built using the new data and the process repeated until the number of outlying samples was statistically acceptable.

To select the calibration and validation samples, all samples were ranked in ascending order of analyte concentration. Validation samples were then selected at regular intervals, the size of the interval depending upon the total number of samples. This ensured that both the calibration and validation data set contained samples that were representative of the analyte concentrations available.

5.2.2.2 Variable Reduction and Multivariate Calibration

5.2.2.2.1 Theory

See Chapter 4 (Section 4.4).

5.2.2.2.2 Procedure

Application of the variable selection algorithms is described in Chapter 4. The data collected for the industrial fusion samples comprised data corresponding to known analytes and matrix elements in the samples. Because of this the amount of noise present in the corresponding subarrays was vastly reduced compared to the noise present when data from all subarrays was used as in Chapter 4 . It was therefore possible to obtain RMSECV plots for Au, Ag and Pd, and determine the approximate number of PCs to use in the UVE-PLS

algorithm, as opposed to simply using a range of PCs (c.f. Chapter 4).

5.3 Results and Discussion

The most common pre-processing methods and their effects and disadvantages are shown in Table 5.2. From this it is evident that no pre-processing step is without its problems when performing classification using PCA. By autoscaling, informative and uninformative (noisy) variables within the data set are scaled to unit variance and are thereby given the same importance. The use of mean centering, however, may lead to the situation shown in Fig. 5.2, where the high variability of informative variables 1 and 2 overshadow the variability of informative variables 3 and 4. This may give a high level of explained variance, but would omit important chemical information necessary for the correct classification of the samples on the basis of their composition.

Table 5.2 Summary of common pre-processing techniques, their effects and disadvantages

Pre-processing technique	Effects	Disadvantages
No pre-processing	Allows natural variability to influence model	Important variables with smaller variability ignored
Autoscaling	Allows all variables to influence model equally	Noise given same importance as relevant variables
Mean centering	Allows variables with greater variability to dominate model	Important variables with smaller variability ignored

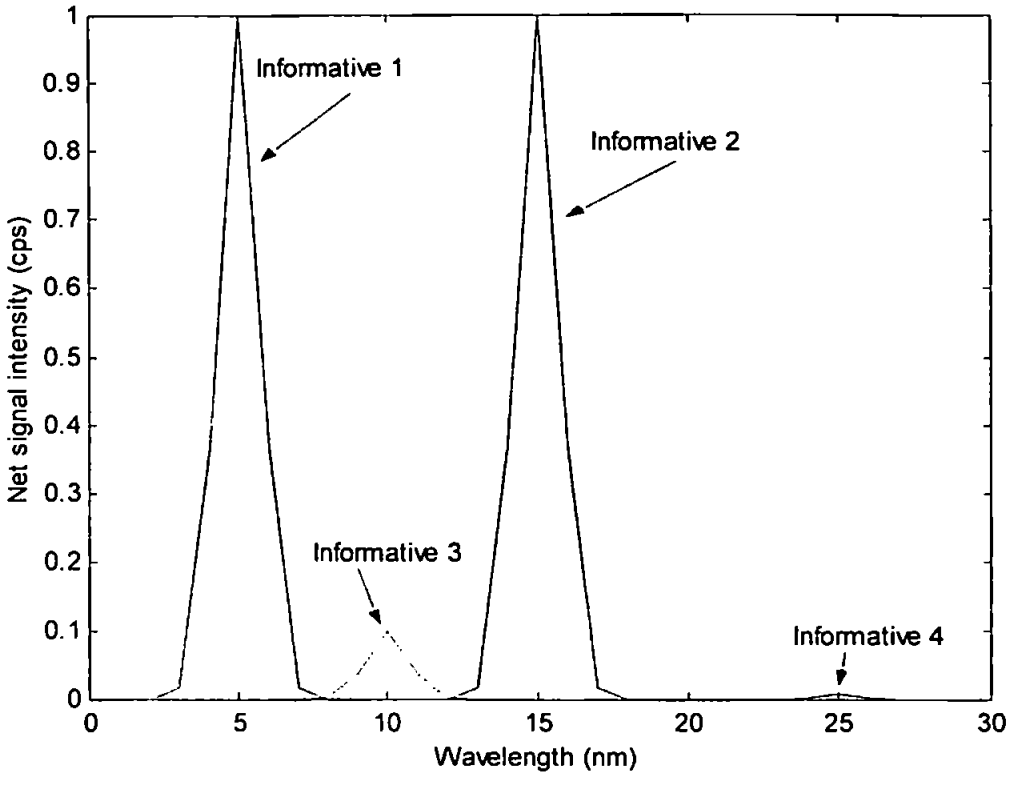


Figure 5.2 Dominance of uninformative variables over informative ones thereby giving a false level of high explained variance

Because of the large variability inherent with complex ICP-AES spectra, autoscaling was chosen as the pre-processing method. Plots of the first 2 PCs obtained after the application of PCA on the data subsets are shown in Figs. 5.3 - 5.5 for Au, Ag and Pd respectively. Also shown are Hotelling's T^2 confidence ellipses at the 95 % confidence interval.

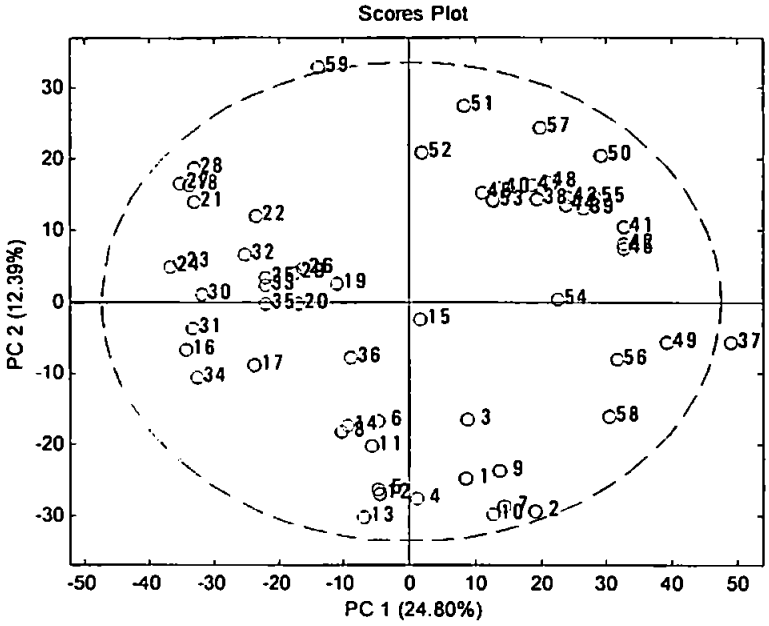


Figure 5.3 Scores plot (first 2 PCs) for Au (95 % confidence ellipse) also showing a statistically acceptable number of outliers (59 and 37).

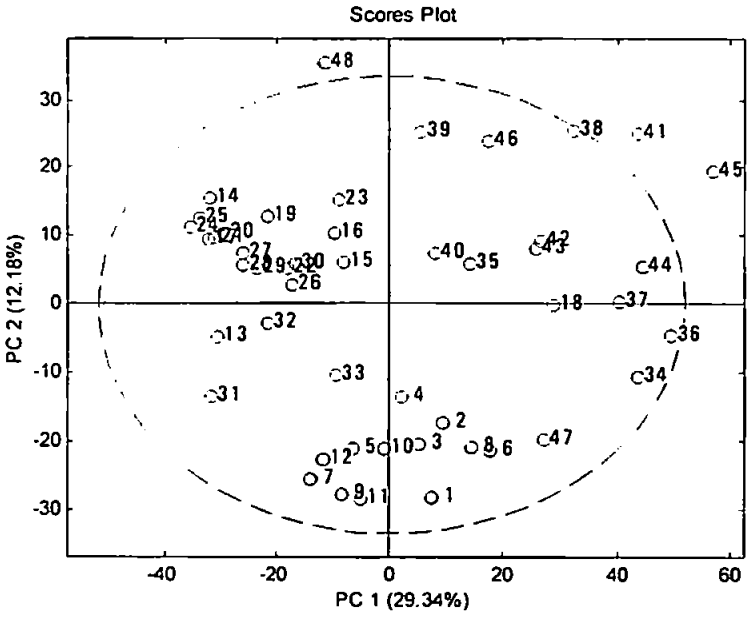


Figure 5.4 Scores plot (first 2 PCs) for Ag (95 % confidence ellipse) also showing a statistically acceptable number of outliers (48, 41 and 45).

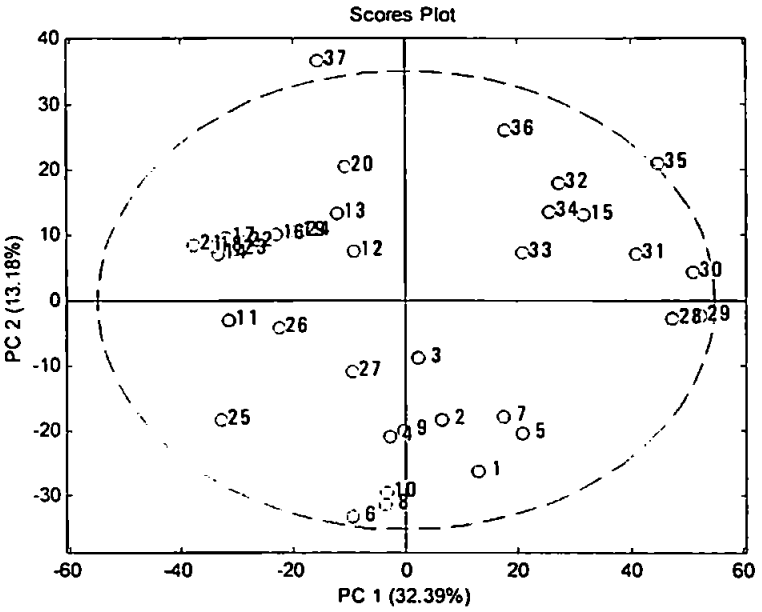


Figure 5.5 Scores plot (first 2 PCs) for Pd (95 % confidence ellipse) also showing a statistically acceptable number of outliers (37 and 35).

The amount of variance explained by the first 2 PCs, in the final PCA model, for Au, Ag and Pd was ~ 37%, ~ 42% and 46% respectively. Despite the majority of the samples being contained within the confidence ellipse, a large spread was evident within the data for all analytes indicating that large differences in the chemical composition of the samples remained. The final number of samples retained for Au, Ag and Pd was 58, 48 and 37 respectively, with 41, 41, and 63 % of the total number of samples being classified as outliers (sample rejection) for the same analytes.

5.3.1 Application of UVE-PLS and IVD-PLS Algorithms to the Industrial Fusion Samples

The UVE and IVD algorithms were applied to the available segmented spectrum obtained from the ICP-AES instrument and multivariate calibration models built for the prediction of Au, Ag and Pd in complex industrial fusion samples. The general efficacy of the variable selection algorithms are discussed, followed by a comparison of the results obtained for the multivariate calibration and prediction of Au, Ag and Pd.

Before the UVE-PLS algorithm can be implemented, one must enter the optimum number of PCs. In this particular instance this was possible by using the RMSECV plot. Because the UVE-PLS algorithm uses autoscaled data (see Chapter 4), the industrial fusion data was also autoscaled prior to determining the optimum number of PCs. This was accomplished by plotting the RMSECV for increasing numbers of PCs, the number of PCs which yielded the minimum value was chosen. Results for Au, Ag and Pd are shown in Figs. 5.6 - Fig. 5.8 respectively, using this data, 8, 5 and 9 PCs were chosen for Au, Ag and Pd respectively, resulting in RMSECV values of ~ 0.4. Application of the variable reduction algorithms resulted in the deletion of the majority of the original 2268 spectral data points (Table 5.3). It is evident that the UVE-PLS algorithm had the largest effect, eliminating between 95-90% of the spectral data.

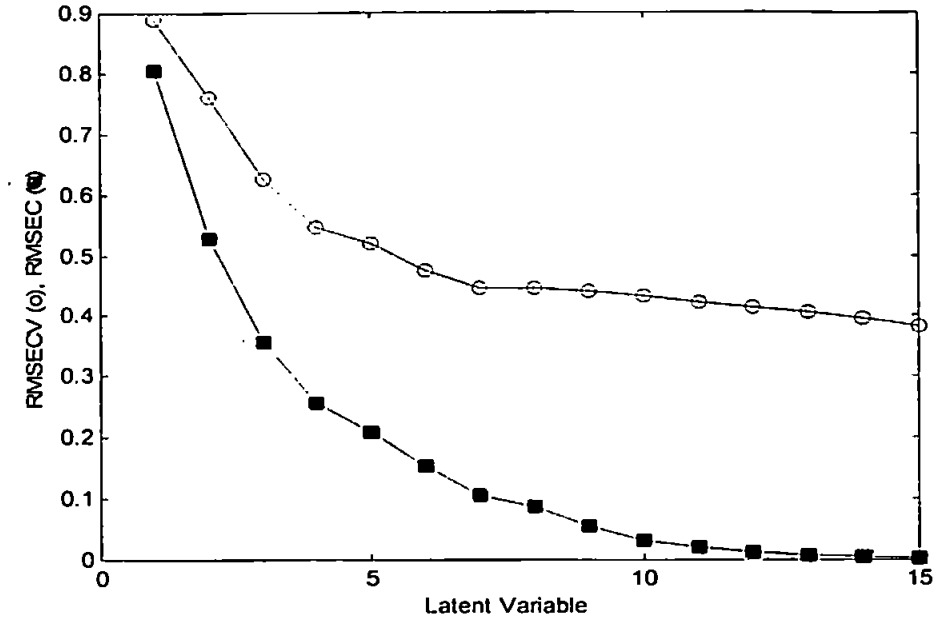


Figure 5.6 RMSECV / RMSEC plot for Au (autoscaled)

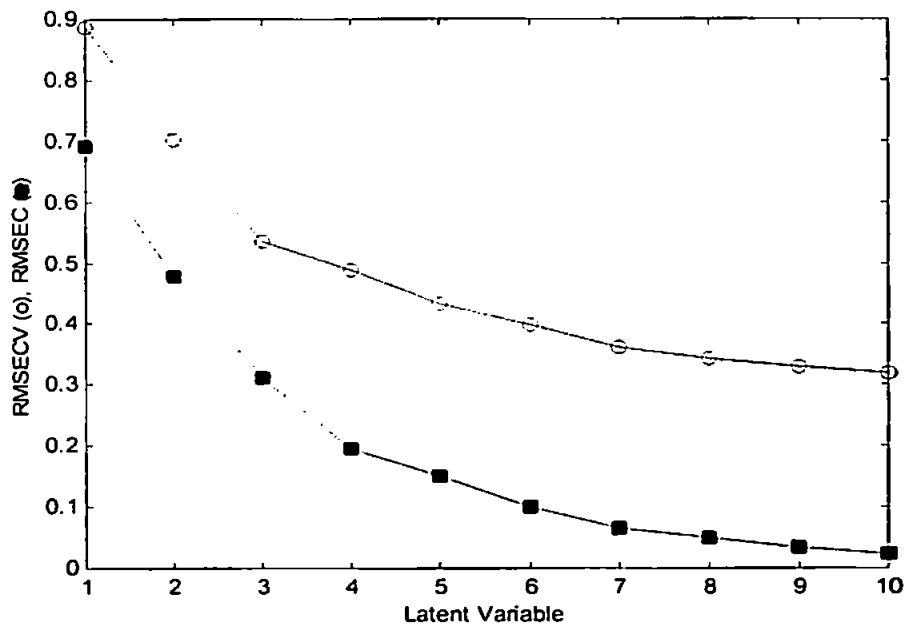


Figure 5.7 RMSECV / RMSEC plot for Ag (autoscaled)

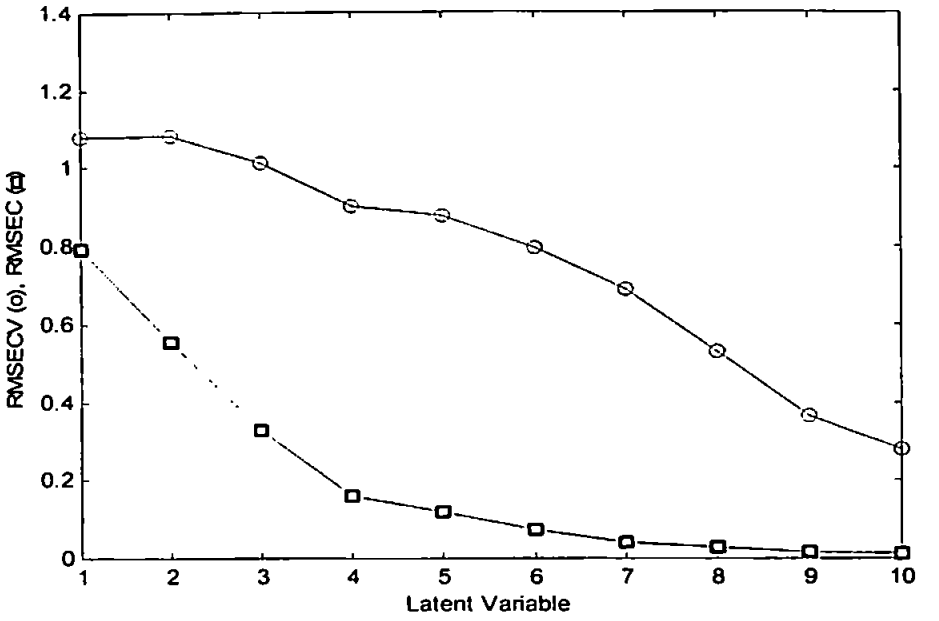


Figure 5.8 RMSECV / RMSEC plot for Pd (autoscaled)

Table 5.3 Effect of applying the UVE-PLS and IVD-PLS algorithms to the original 2268 variables in the data matrix for Au, Ag and Pd calibration.

	Au	Ag	Pd	Au	Ag	Pd
	UVE-PLS			IVD-PLS		
variables	229	166	110	83	55	30
αn	57	57	57	*	*	*
$\alpha n / n var s$	4.01	2.91	1.92	*	*	*

* No t-test used in IVD-PLS

The IVD-PLS algorithm then reduced the remaining data by between 64-72%, depending on the analyte. For Au and Ag the $\alpha n / n var s$ ratio >2.5, in the worst case, Pd, the ratio of type I error variables to truly informative variables was < 2. There was therefore an equal

probability of the IVD-PLS algorithm ranking either variable type within the first few programme iterations. The final Pd model was therefore highly dependent on the importance of the truly informative and Type I error variables (see section 5.3.2.2).

The emission spectrum for all calibration solutions is shown in Fig. 5.9. Also shown are the regions of the spectrum (i.e. the wavelength data points) which were selected by the variable reduction algorithms for the prediction of Ag. Explanations for the selection of those regions shown are similar to those given previously (section 4.7.1). However, in this particular instance the interferent lines include W, Ir, Mn, Cr and W.

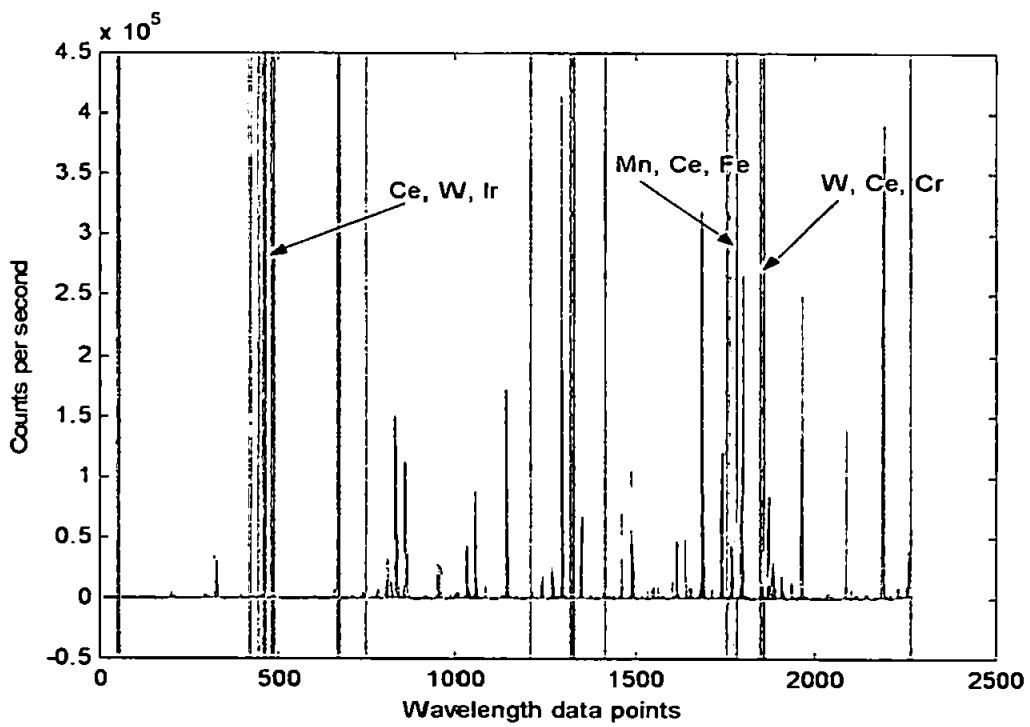


Figure 5.9 Partial emission spectrum for all calibration samples with those regions selected for the prediction of Ag (55 wavelength data points).

The effectiveness of the IVD-PLS algorithm is illustrated in Fig. 5.10, which shows the cumulative sum plot of the var_j^{ivd} ratios and the RMSECV values. It is evident from

Fig. 5.10a that a minimum RMSECV value of 0.025 was obtained when only 30 wavelength data points were included in the model (Fig 5.10b).

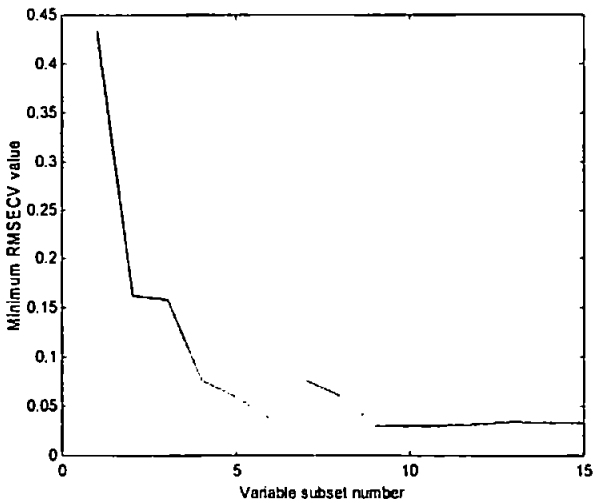


Figure 5.10a Plot of the minimum RMSECV (Pd) value for each variable subset defined by the cumulative sum of the IVD ratio cut off point

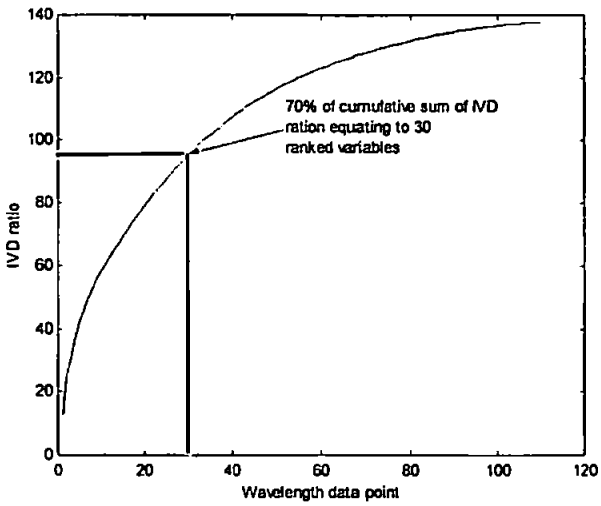


Figure 5.10b Plot of the regression coefficient magnitude / standard error of regression coefficient (IVD ratio) of each ranked wavelength data point for Pd

5.3.2 Multivariate Calibration and Quantitative Prediction for Fusion Samples

5.3.2.1 Detection of Calibration Outliers

Every effort was made, through the use of PCA and Hotelling's T statistic, to ensure that the data sets had the same general base-metal composition, however, large variations in the PLS scores for the first two PCs were still evident (Figs 5.3–5.5). The 95 % confidence interval simply shows the boundary at which any sample ceases to be a member of any particular sample population within the ellipse. It can be concluded from the large variations in Figs. 5.3 – 5.5, and the absence of any substantial groupings, that the majority of the samples have dissimilar metal content. Although from a modelling perspective it would be better to use samples of the same composition, PLS can still be used if a suitable method of detecting potential outliers is used. The detection of outliers during PLS can be accomplished using a concentration residual plot (Fig. 5.11). The abscissa is plotted as the predicted concentration, and the ordinate is plotted as the concentration residual (actual – predicted concentration). Outliers are then detected as those whose concentration residual is $>3\times\sigma_{conc.resid}$. Because of so few samples to start with, a 99 % confidence interval was used as opposed to 95 % which is more commonplace. Hence, outliers were detected and removed from the calibration dataset and the final number of samples for the calibration were 47, 38, and 31 for Au, Ag and Pd respectively.

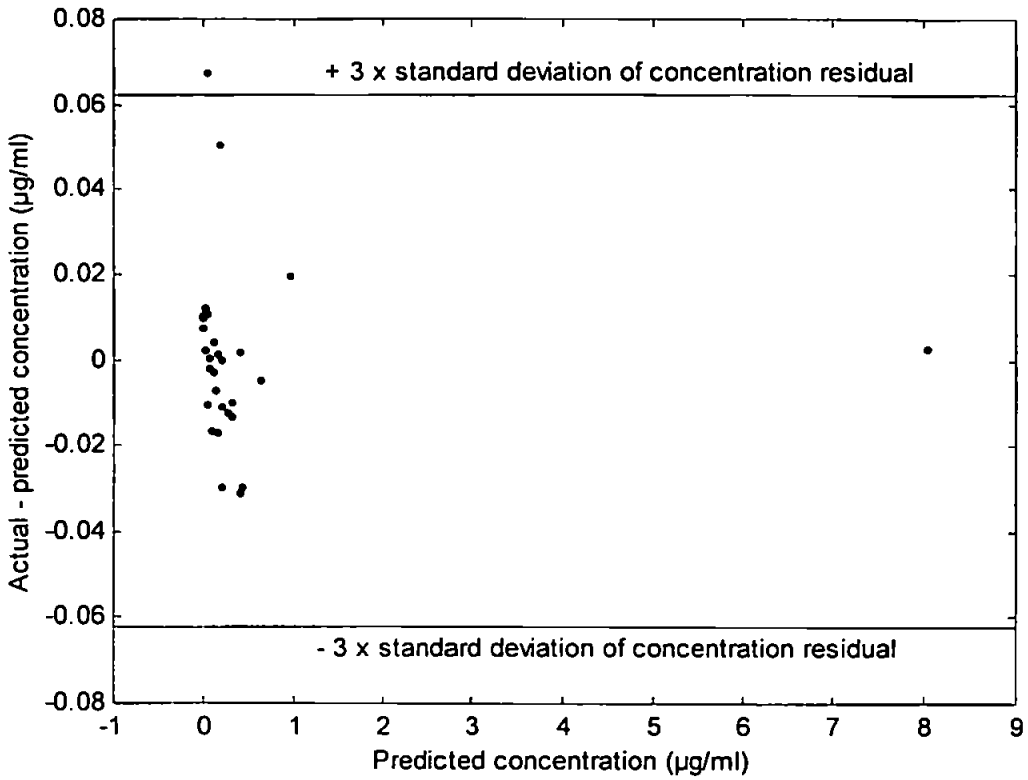


Figure 5.11 A typical concentration residual plot for Pd showing the $3 \times \sigma_{conc.resid}$ boundary and 1 outlier.

5.3.2.2 Analysis of Independent Test Samples

Predictive errors of Au, Ag and Pd in the fusion samples are shown in Table 5.4. In order to compare the effectiveness of the variable reduction routine, a PLS model was also built using the entire spectrum available with no variable selection (Table 5.4). For all analytes there was a significant increase in the accuracy of the predictions following variable selection (outliers removed). Only 83 variables out of 2268 were required to obtain a RRMSE value of 10.71 % for Au in the test set, compared to 38.31 % for the full spectrum. For Ag and Pd test data the RRMSE values were 8.80 and 12.10 % respectively, compared to 82.31 and 194.70 % using the full spectrum.

The $\alpha n / n var s$ ratio for Au, Ag and Pd respectively was 4.01, 2.91 and 1.92 (Table 5.3). Evidently the low ratio for Pd did not result in an elevated RRMSE value. One possible explanation for this is that the truly informative variables possessed far larger IVD ratios then the Type I error variables. This would result in them being placed into the final PLS model ahead of any Type I error variables, and a minimum RMSECV being found.

Table 5.4 Independent test RRMSE % values for Au, Ag and Pd using PLS1 with variable reduction and full spectra modelling.

	Au	Ag	Pd
Variable Selection			
RRMSE % [Mc]	74.23(10.71)	8.80	12.10
Number of variables after t-test	229	166	110
Final number of variables selected	83	55	30
Full Spectrum (2268 wavelength points)			
RRMSE % [Mc]	142.55	59.42	140.01
RRMSE % [As]	38.3	82.3	194.7

( ) Outlier removed (Table 5.3)
Mc Mean centered
As Autoscaled

5.3.2.3 Comparison of Methods

In the following discussion three sets of data, obtained for samples treated in three different ways, are compared as follows:

1. Samples were prepared using a NiS fire assay method and analysed using ICP-AES by univariate calibration, designated ***FA-UC.*** These were the consensus values which were used as in-house standards.
2. Samples were prepared using the peroxide fusion digestion method and analysed by ICP-AES using the variable reduction routine and multivariate calibration by PLS, designated ***F-VR-PLS.***

3. Samples were prepared using the peroxide fusion digestion method and analysed by ICP-AES using univariate calibration and interelement correction, designated ***F-UC-IEC.***

Plots of concentration predicted using the F-VR-PLS method versus concentrations obtained using the FA-UC method, for both the calibration and test samples, are shown in Figs 5.12 - 5.14 for Au, Ag and Pd respectively. Also shown are the 95 % confidence intervals. The correlation coefficients for both calibration and test data for all three analytes were > 0.99, showing the overall success of the variable selection method.

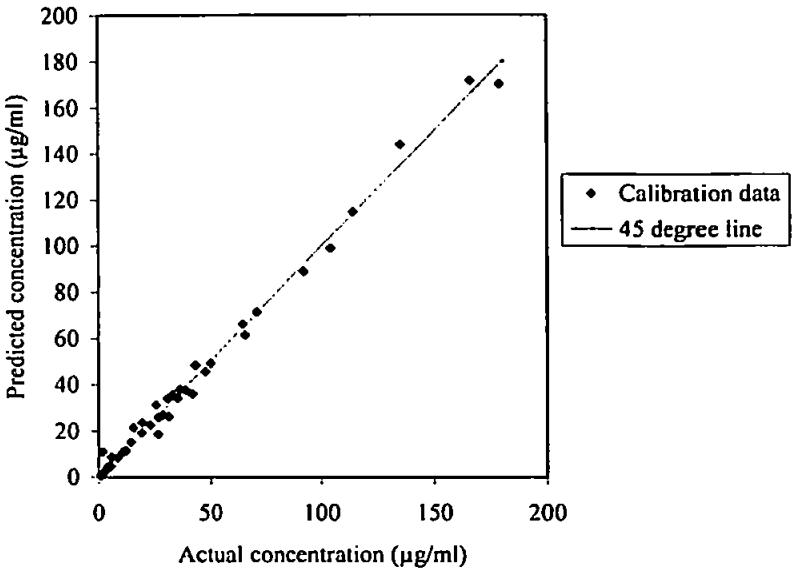
Predicted concentrations in the independent test samples obtained using the three methods, and relative percentage errors for the F-VR-PLS and F-UC-IEC methods compared to the FA-UC method (consensus values), are shown in Tables 5.5-5.7 for Au, Ag and Pd respectively.

Taking Au first, the RRMSE was 132 % using F-UC-IEC compared to 74 % using F-VR-PLS (Table 5.5), indicating a slightly better level of accuracy. As can be seen from Table 5.3, in some cases the former method yielded better accuracy, and in some cases the latter (Fig. 5.15), however, it should be noted that the former set of results were heavily biased by samples 11 and 12, and the latter by sample 11.

In the case of Ag RRMSEs were 8.5 % and 8.8 % for F-UC-IEC and F-VR-PLS methods respectively. It should be noted that, in the latter case, a large error was obtained for sample 1 for which the consensus value obtained using the FA-UC method was only 0.02 μg/ml. In general, the two methods gave comparable levels of accuracy (Fig. 5.16).

For Pd the RRMSE was 8.6 % using F-UC-IEC compared to 12.2 % using the F-VR-PLS, with the latter being heavily influenced by samples 1 and 2 with respective concentrations of 0.02 and 0.09 μg/ml (Fig. 5.17).

(a) Calibration



(b) Test

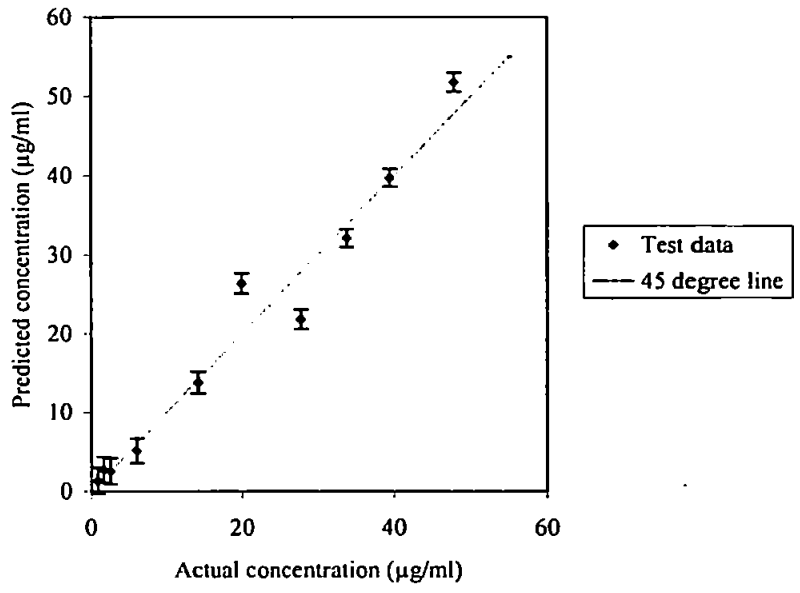
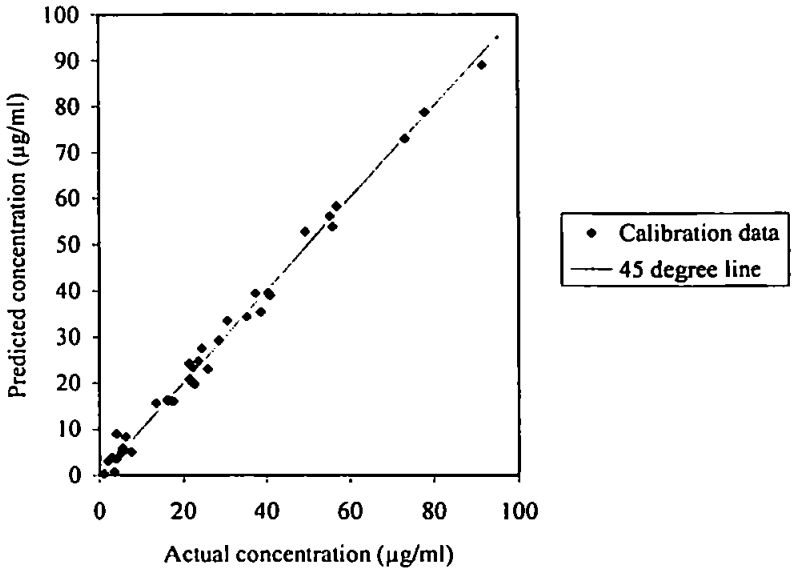


Figure 5.12 Concentration of Au predicted using the F-VR-PLS method (predicted concentration) vs the FA-UC method (actual concentration): (a) calibration samples; (b) independent test samples with 95% confidence interval.

(a) Calibration



(b) Test

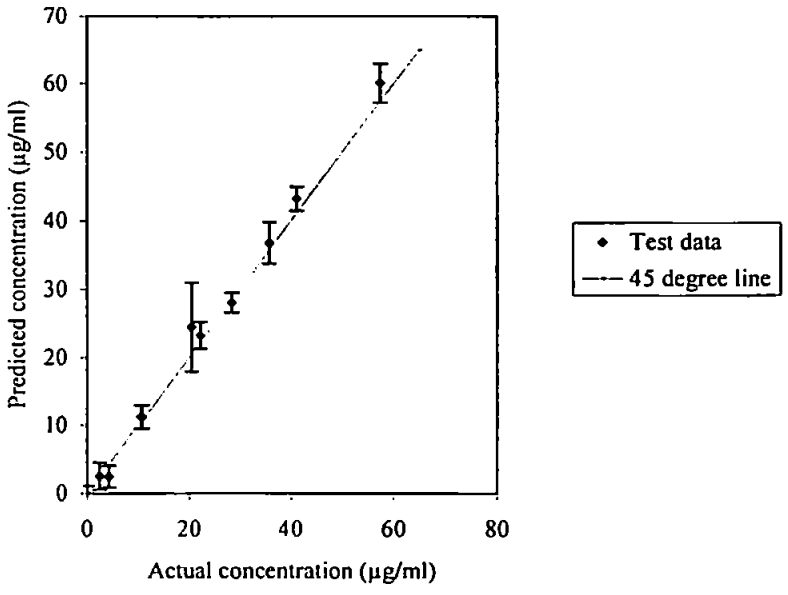
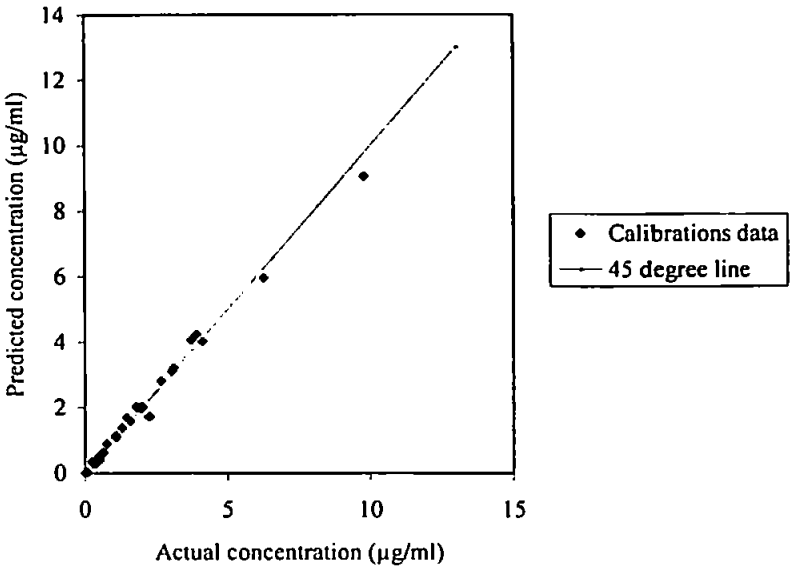


Figure 5.13 Concentration of Ag predicted using the F-VR-PLS method (predicted concentration) vs the FA-UC method (actual concentration): (a) calibration samples; (b) independent test samples with 95% confidence interval.

(a) Calibration



(b) Test

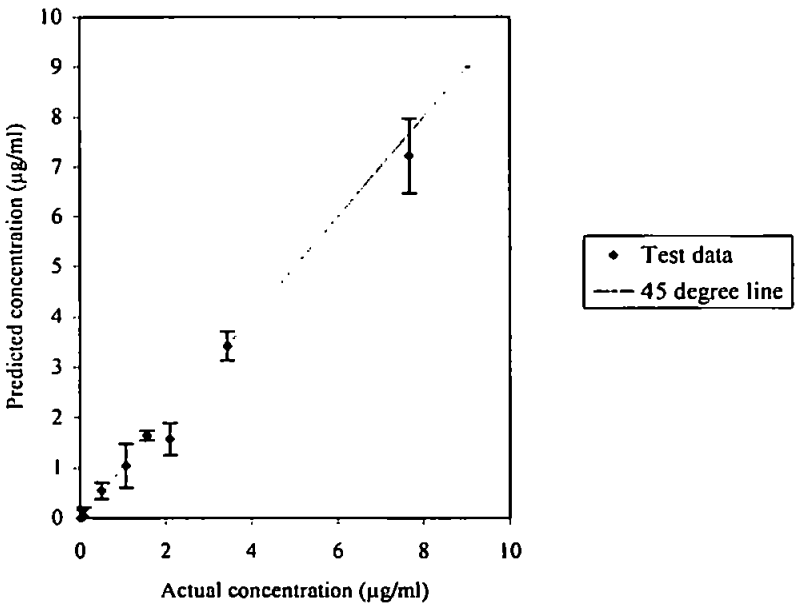


Figure 5.14 Concentration of Pd predicted using the F-VR-PLS method (predicted concentration) vs the FA-UC method (actual concentration): (a) calibration samples; (b) independent test samples with 95% confidence interval.

Table 5.5. Fire assay, fusion and variable selection method Au test sample concentrations with absolute % error values and 95% confidence limits.

Au concentration (μg/ml)			% error relative to FA-UC method	
FA-UC	F-VR-PLS	F-UC-IEC	F-UC-IEC	FVR-PLS
0.88	1.33	0.95	9.03	52.00
1.66	2.72	1.92	15.90	63.86
2.55	2.56	2.65	4.13	0.59
6.05	5.14	6.49	7.36	14.97
14.10	13.76	14.29	1.35	2.41
19.89	26.34	21.02	5.66	32.43
27.84	21.79	28.16	1.16	21.73
33.80	32.13	33.16	1.91	4.94
39.45	39.72	40.48	2.62	0.70
47.84	51.78	49.13	2.71	8.25
87.06	***-0.31***	**0.09**	**99.90**	**100.36**
129.52	126.29	0.02	99.98	2.49
RRMSE (%)			132	74

Table 5.6. Fire assay, fusion and variable selection method Ag test sample concentrations with absolute % error values and 95% confidence limits

Ag concentration (μg/ml)			% error relative to FA-UC method	
FA-UC	F-VR-PLS	F-UC-IEC	F-UC-IEC	FVR-PLS
0.01	-2.31	0.05	375.67	23200
2.43	2.52	1.99	18.09	3.92
4.33	2.45	4.44	2.63	43.35
10.79	11.27	11.59	7.38	4.45
20.44	24.46	20.39	0.23	19.67
22.27	23.24	24.32	9.21	4.38
28.43	28.05	27.27	4.09	1.34
35.73	36.78	38.23	7.00	2.95
41.17	43.21	44.44	7.94	4.96
57.51	60.10	61.12	6.29	4.51
RRMSE (%)			8.5	8.8

Table 5.7 Fire assay, fusion and variable selection method Pd test sample concentrations with absolute % error values and 95% confidence limits. (values also given for variable selection predictions with and without outliers)

Pd concentration (μg/ml)			% error relative to FA-UC method	
FA-UC	F-VR-PLS	F-UC-IEC	F-UC-IEC	FVR-PLS
0.02	0.001	0.01	16.00	93.33
0.09	0.042	0.09	4.89	53.33
0.51	0.546	0.51	0.46	8.12
1.08	1.043	1.11	3.06	3.43
1.56	1.648	1.63	4.42	5.64
2.10	1.575	2.19	4.56	24.82
3.44	3.417	3.32	3.35	0.52
7.68	7.217	8.15	6.23	5.97
RRMSE (%)			8.6	12.2

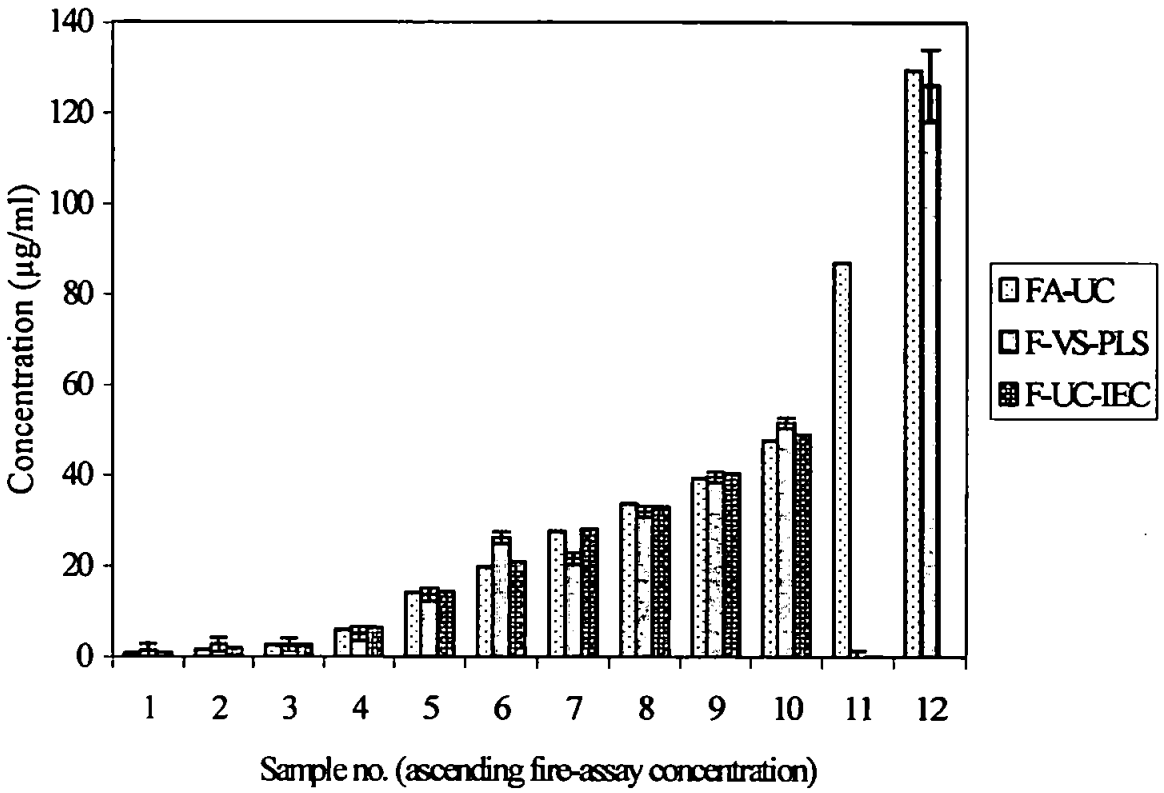


Figure 5.15 Fire-assay, variable selection and fusion estimation of the independent test samples for Au with 95% confidence limits.

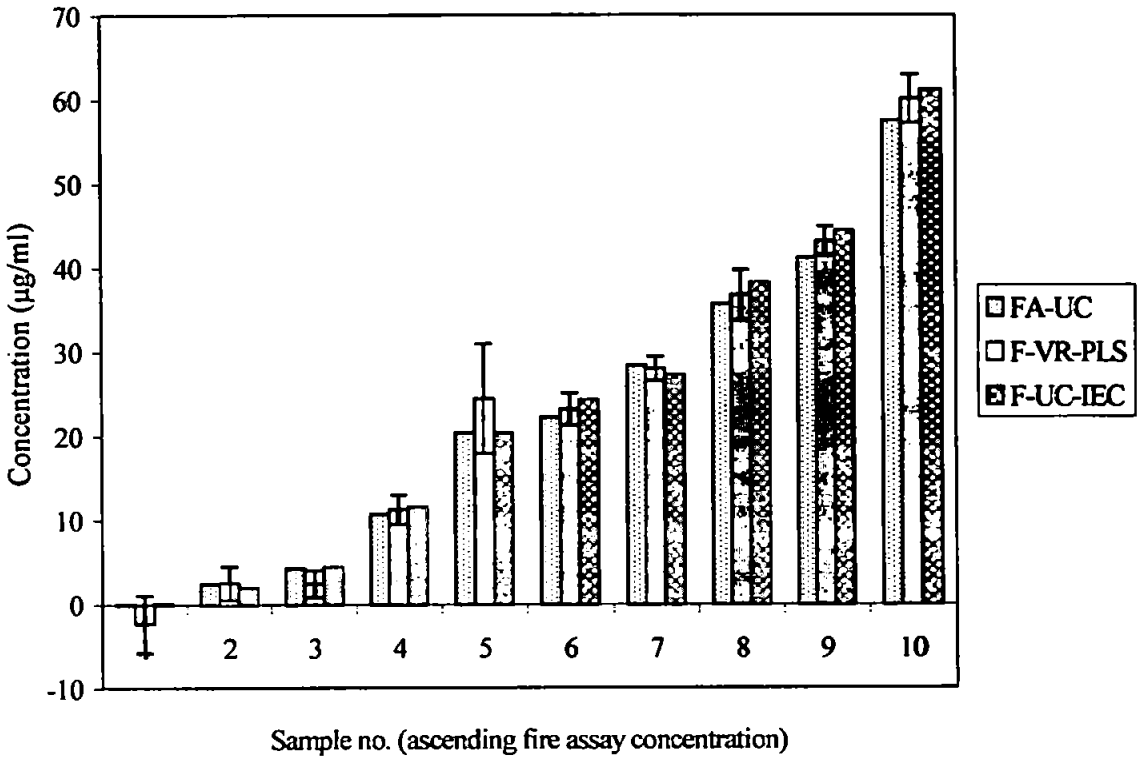


Figure 5.16 Fire-assay, variable selection and fusion estimation of the independent test samples for Ag with 95% confidence limits.

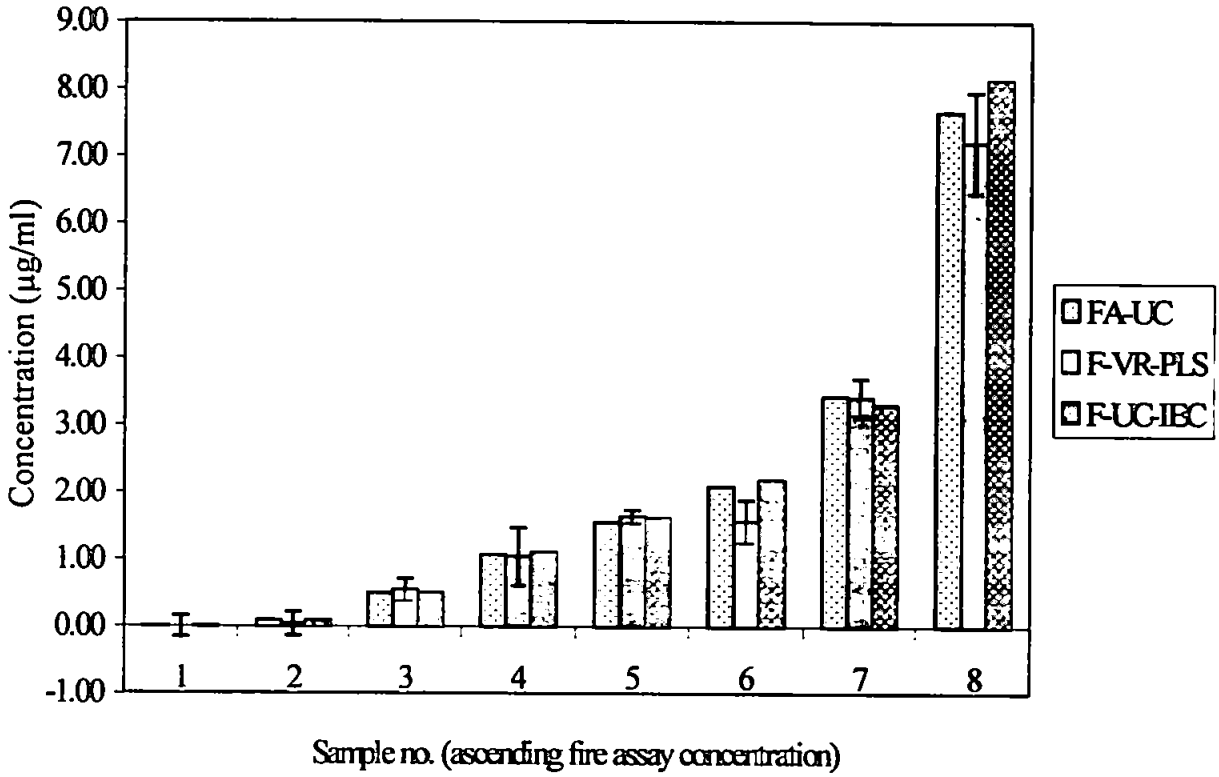


Figure 5.17 Fire-assay, variable selection and fusion estimation of the independent test samples for Pd with 95% confidence limits.

5.4 Conclusions

Spectral data obtained for the analysis of fusion digests has been used to build multivariate calibration models using PLS to predict the concentration of Au, Ag and Pd in test samples. In order to achieve this, variable elimination and selection algorithms were used to select the informative parts of the ICP-AES emission spectra without having to resort to line selection or the need to assign background correction points in order to obtain the net integrated line intensities for individual analyte or matrix lines. The model errors

for both the calibration and independent test data, have shown considerable improvement compared to the errors achieved when using all 2268 wavelengths thus reinforcing the fact that PLS benefits from selective variable reduction. The variable selection method and PLS multivariate calibration gave results comparable to those obtained using a more traditional univariate calibration approach with interelement correction. Calibration models were built using 47, 38 and 31 samples for Au, Ag and Pd respectively, hence, it is envisaged that an improvement in the accuracy of prediction would be obtained if more samples were used to build the model.

CHAPTER 6 - CALIBRATION TRANSFER OVER TIME USING PIECEWISE DIRECT STANDARDISATION (PDS) FOR COMPLEX SYNTHETIC SOLUTIONS

6.1 Introduction

The effort in building a multivariate model is often considerable in terms of both time and money, hence it would be beneficial if it were possible to use the models over a long period of time. However, changes in instrument response can occur due to temperature fluctuations, electronic drift, wavelength or detector instability e.t.c. If this happens following the calibration of an instrument, subsequent use of the calibration model will most probably produce erroneous results. This poses severe restrictions on the successful application of multivariate calibration models.

There are many publications detailing the development of calibration drift correction methods for broad spectrum techniques (113, 114, 139, 148-153), though none has been applied to atomic emission spectroscopy.

Work on univariate calibration was performed to eliminate the sample-to-sample difference, using either internal standards or the zeeman effect (154), but most other work in transporting calibration models has been published in the area of NIR analysis. Three publications are notable. Osborne and Fearn (155) investigated the affects of transferring single-wavelength calibration models between nine different instruments for the prediction of protein and moisture in wheat flour using NIR spectroscopy. Single wavelength bias correction terms for the two calibration equations on each instrument were determined and the long-term stability of the calibration was studied. Later Shenk *et. al.* (156) published results from a study where a large number of candidate calibration equations were

developed on a single instrument and then transferred to six other instruments. The "best" equation was adjusted for bias, offset, and wavelength selection on the other instruments and the standard error of prediction (SEP) was compared between the original and the other instruments for a set of 60 samples. Mark *et. al.* (157) published work describing the selection of wavelengths for NIR calibration based on their robustness toward wavelength shifts between instruments. These methods all involve calibration utilising a single, or sometimes a limited number of wavelengths, and are not generally applicable to multivariate calibration based on full spectral responses, or where variable selection has been applied to the full spectrum.

It is possible to solve the calibration transfer problem in a multivariate way by applying chemometric techniques which attempt to find a transformation function that makes the measured response obtained from one instrument the same as that which would be obtained on the same instrument at a later point in time. This method is commonly termed piecewise direct standardisation (PDS). There are two methods of correction, the first method transforms the calibration model itself, while the second transforms the response from the instrument at $t = 2$ to match that which would have been obtained if the sample had been measured at $t = 1$. The later of the two methods has been used in this study. All methods allow the full response of the instrument to be utilised without restriction on the number of wavelengths, which can be included in the calibration model.

6.2 Calibration Transfer Theory

Assume that the response matrix for a full calibration set R_1, has been measured, and $\overline{R}_1$ is a small subset of R_1. The response matrix of this subset is measured on the same instrument at a later time, and is denoted by $\overline{R}_2$. Through standardisation, it is hoped that the calibration information contained in R_1 could be transferred without measuring the response matrix (R_2) of the full calibration set at a later time.

6.2.1 Standardisation with the Classical Calibration Model

Let the concentration matrix, *m* (samples) x *n* (wavelengths), for the full calibration and standardisation subset be C and $\overline{C}$, respectively. A linear relationship between the response and concentration is assumed for simplicity, Eqn's. 6.1and 6.2:

$$R_1 = CK_1 \qquad \text{Eqn. 6.1}$$

$$R_2 = CK_2$$

$$= C(K_1 + \Delta K) \qquad \text{Eqn. 6.2}$$

where K and K_1 are the sensitivity matrices for time $t = 1$ and $t = 2$ (rows being the pure components spectra) and ΔK is the difference matrix between them. The same relationship should be true for the standardisation subset analysed at time, $t = 2$, as shown in Eqn's. 6.3 and 6.4:

$$\overline{R}_1 = \overline{C}K_1 \qquad \text{Eqn. 6.3}$$

$$\overline{R}_2 = \overline{C}K_2$$

$$\overline{\mathrm{R}}_2 = \overline{\mathrm{C}}(\mathrm{K}_1 + \Delta\mathrm{K}) \qquad \text{Eqn. 6.4}$$

Solving Eqn's. 6.3 and 6.4 for ΔK gives Eqn. 6.5:

$$\Delta K = \overline{C}^{+}(\overline{R}_2 - \overline{R}_1) \qquad \text{Eqn. 6.5}$$

where $\overline{C}^{+}$ is the pseudoinverse of $\overline{C}$. Substituting ΔK into Eqn. 6.2 and using Eqn. 6.1, R_2 is estimated as Eqn. 6.6:

$$\hat{R}_2 = R_1 + C\overline{C}^{+}(\overline{R}_2 - \overline{R}_1) \qquad \text{Eqn. 6.6}$$

With $\hat{R}_2$ and C, a new calibration model can be built for prediction at time, $t = 2$. Two assumptions are implied in this method: that the linear relationship is the same at time, $t = 1$ and at time $t = 2$, and that the concentrations for all elements contributing to the response must be known.

6.2.2 Standardisation with the Inverse Calibration Model

Instead of using the classical model, the inverse calibration model corresponding to Eqn's. 6.1 - 6.4 is used, i.e. Eqn's 6.7-6.10:

$$c = R_1 b_1 \qquad \text{Eqn. 6.7}$$

$$c = R_2 b_2$$

$$= R_2(b_1 + \Delta b) \qquad \text{Eqn. 6.8}$$

and

$$\bar{c} = \bar{R}_1 \bar{b}_1 \qquad \text{Eqn. 6.9}$$

$$\bar{c} = \bar{R}_2 \bar{b}_2$$

$$= \bar{R}_2 (\bar{b}_1 + \Delta \bar{b}) \qquad \text{Eqn. 6.10}$$

where b_1 and b_2 are the two regression vectors for an analyte at time, $t = 1$ and time $t = 2$, $\bar{b}_1$ and $\bar{b}_2$ are the corresponding regression vectors calculated from the subset, c and $\bar{c}$ represent the concentration vectors for this analyte in the full set and subset respectively. Using Eqn. 6.7 to calculate b_1 and combining Eqn. 6.9 with Eqn. 6.10 to estimate Δb, a standardised regression vector can be calculated as follows:

$$\hat{b}_2 = b_1 + \Delta b$$

$$= b_1 + (\bar{b}_2 - \bar{b}_1)$$

$$= \mathbf{R}_1^{+}\mathbf{c} + (\bar{\mathbf{R}}_2^{+} - \bar{\mathbf{R}}_1^{+})\bar{\mathbf{c}} \qquad \text{Eqn. 6.11}$$

With the standardised regression vector $\hat{b}_2$ (an estimate of b_2), prediction at time, $t = 2$ can be made. The major advantage, of course with the inverse modelling approach, is that only the concentration of the analyte of interest is required.

6.2.3 Direct Standardisation (DS)

The previous sections corrected the calibration model, built at time, $t = 1$, so that it

could be applied on spectral responses collected at time, $t = 2$. Another approach called direct standardisation is available in which the spectra (future samples) measured at time, $t = 2$, are corrected to match spectra (the calibration model), built at time, $t = 1$, hence the calibration model remains unchanged. In direct standardisation, response matrices on both instruments are related to each other by a transformation matrix $\boldsymbol{F}$, i.e.

$$\overline{\boldsymbol{R}}_1 = \overline{\boldsymbol{R}}_2 \boldsymbol{F} \qquad \text{Eqn. 6.12}$$

where $\boldsymbol{F}$ is a square matrix whose dimensions are wavelength by wavelength. From Eqn. 6.12, the transformation matrix $\boldsymbol{F}$ is calculated as:

$$\boldsymbol{F} = \overline{\boldsymbol{R}}_2^{+} \overline{\boldsymbol{R}}_1 \qquad \text{Eqn. 6.13}$$

And the response vector of an unknown sample measured at time, $t = 2$, $\boldsymbol{r}_{2,un}^T$, is standardised to the response vector $\boldsymbol{r}_{1,un}^T$, expected from the instrument at time, $t = 1$ according to:

$$\hat{\boldsymbol{r}}_{1,un}^T = \boldsymbol{r}_{2,un}^T \boldsymbol{F} \qquad \text{Eqn. 6.14}$$

If there are many wavelengths present, this is simply performed for each wavelength. With $\boldsymbol{r}_{1,un}^T$, the model constructed using $\boldsymbol{R}_1$ and $\boldsymbol{C}$ (at time, $t = 1$) can now be used for the prediction of samples analysed at time, $t = 2$.

6.2.4 Piecewise Direct Standardisation (PDS)

In the previous section, the number of subset samples must be at least equal to the rank of $\boldsymbol{R}_1$, to ensure that the inverse of $\boldsymbol{R}_1$ is stable, and hence an adequate standardisation is achieved. In real applications this could lead to large numbers of subset samples being needed. Also, it is noticed that in DS, the whole spectrum at time, $t = 2$ is used to fit each spectral point at time, $t = 1$. For real spectroscopic data, however, spectral variations are often limited to a smaller region. Therefore, each spectral point at time, $t = 1$, would more likely be related to the spectral measurements at nearby wavelengths than the full spectrum at time, $t = 2$. On the basis of these considerations, a piecewise standardisation method has been developed to reconstruct each spectral point at time $t = 1$ from several measurements in a small window at time $t = 2$. For subset measurements $\boldsymbol{r}_{1,i}$, at wavelength index i at time $t = 1$ subset measurements at time $t = 2$, $\boldsymbol{r}_{2,i-j}, \boldsymbol{r}_{2,i-j+1}, \ldots\ldots \boldsymbol{r}_{2,i+k-1}$, and $\boldsymbol{r}_{2,i+k}$ at nearby wavelengths from index point $i - j$ to $i + k$ are chosen and put into a matrix

$$\boldsymbol{X}_i = \boldsymbol{r}_{2,i-j}, \boldsymbol{r}_{2,i-j+1}, \ldots\ldots \boldsymbol{r}_{2,i+k-1}, \boldsymbol{r}_{2,i+k} \qquad \text{Eqn. 6.15}$$

For a nominal shift and linear intensity change from time $t = 1$ to $t = 2$, $\boldsymbol{r}_{1,i} \alpha \boldsymbol{r}_{2,l}$ with $i - j \le l \le i + k$. For the case where l lies between two nominally denoted index points with a non-linear intensity change, this equation will not hold. It is then possible to establish a local multivariate regression model in the form of

$$\boldsymbol{r}_{1,i} = \boldsymbol{X}_i \boldsymbol{b}_i \qquad \text{Eqn. 6.16}$$

which can perform interpolation and provide a reasonable approximation to the nonlinear intensity change provided that it is not too severe. Each regression vector b_i can be calculated by means of PCR or PLS regression. These regression vectors are arranged along the main diagonal of the transformation matrix F while the rest of the elements are zero, which results in a banded diagonal matrix

$$F = diag(b_1^T, b_2^T, ...b_i^T) \qquad \text{Eqn. 6.17}$$

When compared to the DS method, PDS is in fact a calculation of the transformation F (Eqn. 6.12) by setting most of the off-diagonal elements to zero (i is the number of spectral channels included).

It is the fact that rank $(\mathbf{X}_i) \leq$ rank $(\mathbf{R}_i)$ which makes it possible to reduce the number of subset samples, and hence avoid an ill-conditioned problem where the number of variables is much larger than the number of calibration subset samples. The transformation matrix is subsequently used to transfer $r_{2,un}^T$ piece by piece into the spectrum as if it were measured at time, $t = 1$.

The version of PDS used in this study not only provides multiplicative correction, but correction for additive differences also. An example of an additive difference would be if the source drifted slightly and gave an increased background reading at time, $t = 2$ compared to the signal at $t = 1$. If the additive background is large enough, multiplicative corrections alone will not accomplish the correction (158).

The method used in this study removes the additive term from the transfer standards prior to standardisation, estimates the correct multiplicative model, and then estimates the additive correction required. The removal of the additive term is achieved by

mean-centering the transfer sample spectra prior to application of the PDS function.

6.2.4.1 Selection of Subset for Standardisation

The subset used in the standardisation must contain enough information to describe the difference between the spectra at time, $t = 1$ and $t = 2$. A stepwise procedure is employed here to select the sample with the highest leverage (maximum h_i) according to

$$\boldsymbol{H} = \boldsymbol{R}_1^{+}\boldsymbol{R}_1 \qquad \text{Eqn. 6.18}$$

(assuming that outliers have been detected and deleted from the calibration set).The information contained in the sample just selected is then removed from the set of samples. This procedure continues until the desired number of samples have been included in the subset.

6.3 Experimental (Instrumentation and Reagents)

See Chapter 3 (section 3.2.1).

6.3.1 Procedures

Figure 6.1 illustrates the procedure for the application of spectra standardisation using PDS with variable selection. Prior to PDS standardisation, outliers in the original data set are removed via the inspection of residual concentration plots (outliers tend to produce high leverage values). Because the subset response matrices, $\overline{\boldsymbol{R}}_1$ and $\overline{\boldsymbol{R}}_2$, should contain the same spectral regions for the calculation of the PDS transformation matrix, $\boldsymbol{F}$, (see Eqn. 6.15) the variable reduction routines (UVE-PLS and IVD-PLS) must be performed on the original and drifted data sets taken at time, $t = 1$ and $t = 2$ respectively,

prior to the application of the PDS routine. To accomplish this, alterations were made to both the UVE-PLS and IVD-PLS routines which enabled the simultaneous selection of informative wavelengths, based upon data set 1, from data set 2. Thus, not only does the method standardise the spectra in terms of differences in signal drift, but in doing so it also standardises the spectra from $t = 2$, in terms of variable importance, making those variables which were informative at $t = 1$, also informative at $t = 2$.

To perform the UVE-PLS and IVD-PLS routines the number of PC's must be chosen. Because the PDS study used the same calibration and test data from Chapter 4 for the calibration model, i.e. data at $t = 1$, the number of PC's chosen was based upon the work in chapter 4, section 4.6.1, Table 4.8. Because the RRMSE % values did not differ greatly with the number of PC's chosen or the $^{\alpha n}/_{n var s}$ ratio, 6 PC's were used in the UVE-PLS routine for Pt, Pd and Rh. For all analytes the $^{\alpha n}/_{n var s}$ ratio $\approx> 4$ giving a low probability of Type I error variables being entered into the PLS model within the first few iterations. The number of PC's for the IVD-PLS routine was then determined by using a cross-validation procedure. After the variable subsets had been obtained the PDS routine is applied and synthetic sample predictions are then made. This consists of four separate steps:

i). The calibration transfer samples are determined from the data obtained using the variable reduction and selection algorithms (section 6.2.6).

ii). The transform matrix is calculated (section 6.2.4) using those calibration samples obtained in step ii). This data is not pre-processed in any way.

iii). The PLS calibration model is obtained using the original ($t = 1$) mean-centred calibration data, and the synthetic test samples from time $t = 2$ scaled appropriately.

iv). Predictions are then made for the standardised synthetic samples from step iii).

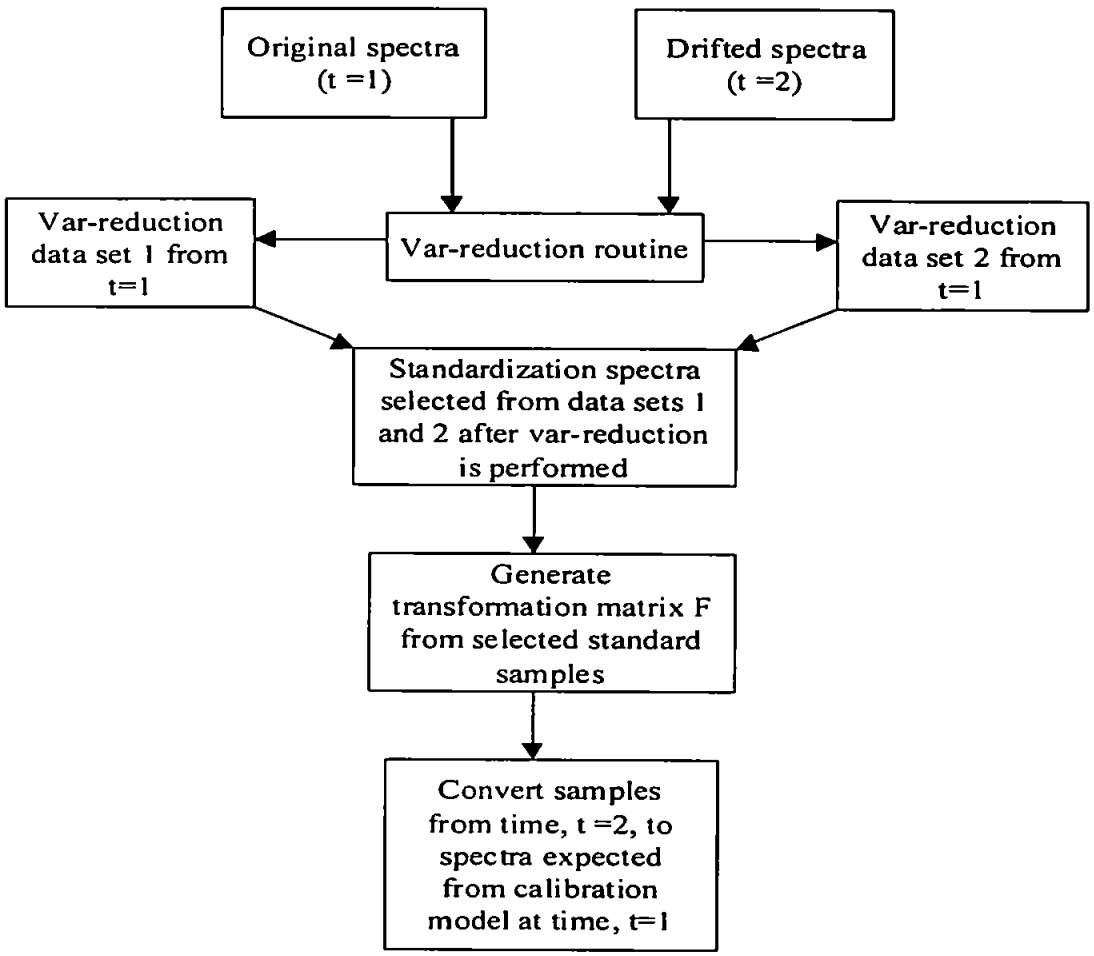


Figure 6.1 Flow-diagram for the process of standardisation with variable selection.

An *m*-function, *stdgen*, implemented in the PLS Toolbox to spectroscopic instrument standardisation has been used to obtain the calibration transfer matrix. The inputs were as shown in Tables 6.1-6.3 which give the RRMSE % values obtained for the independent test samples with PDS applied. It should be noted that the *m*-function *stdgen*, implemented in the PLS Toolbox, requires the number of PC's to be allocated for the construction of the transformation matrix F, this is because of the use of singular value decomposition in the PLS procedure. For this study the number of PC's allowed was set equal to the number of calibration samples. In the event that unsatisfactory RRMSE %

values were obtained with any particular combination of window size and calibration subset sample size, the number of PC's could be optimised also.

The data sets used in this study were collected 11 days apart (data set $t = 1$ and $t = 2$ refers to data collected on 16/7/99 and 26/7/99 respectively) thereby giving sufficient time for the instrument to drift.

6.4 Results and Discussion

6.4.1 Instrumental Drift

In order to illustrate the amount of drift the instrument experienced between analysing the two data-sets, a solution containing the middle concentration of all elements was analysed at every 10^{th} sample for each data-set (Fig. 6.2). The line chosen was In 325.609 nm as this experienced the least spectral interference of any of the lines. It is evident from Fig 6.2 that the drift for this line was considerable indicating that the prediction of data-set 2 samples using a calibration derived from data-set 1 would give erroneous results.

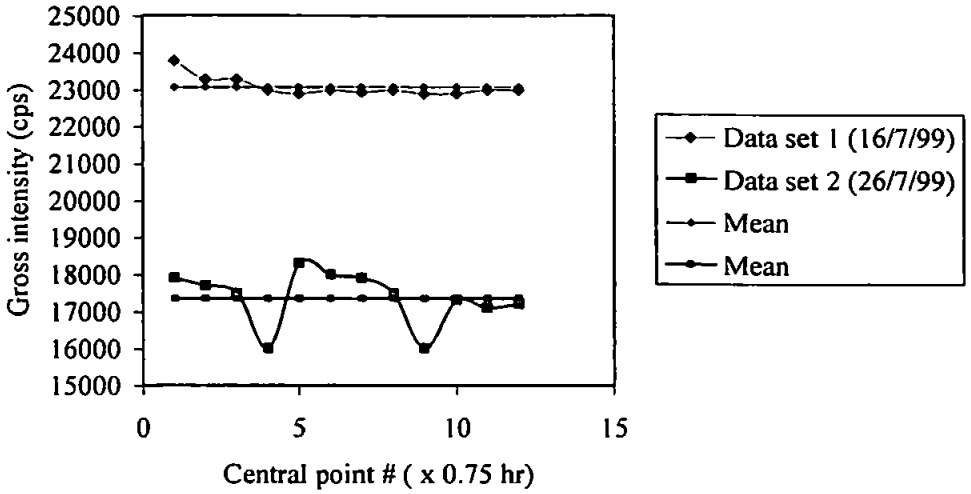


Figure 6.2 Central point In 325.609 nm concentration (μg ml^{-1}) over time for data-sets 1 and 2 (16/7/99 and 26/7/99 respectively) using gross intensity.

6.4.2 Calibration Subset Optimisation

The effect of calibration subset sample number and spectral window size on the RRMSE of the calibration data for Pt, Pd, and Rh is shown in Figs. 6.4-6.6, a summary of the optimum parameters is shown in Figure 6.3. For Pt and Pd the lowest RRMSE % values were 4.14 and 3.03 % respectively using only 3 calibration subset samples (Fig 6.3 and Tables 6.1 and 6.2). For Rh the number of samples increases to 5 giving a RRMSE % value of 1.88. Generally, for all three analytes the RRMSE % value does not decrease with increasing numbers of calibration subset samples, which may indicate a lack of intrinsic modelling ability (lack of fit). The effect of varying the window size was also very clear, with a distinct minimum RRMSE value for all three analytes. This was to be expected, too few wavelengths would not give enough information for standardisation and too many would result in the noise component predominating. The large number of wavelengths required for Pt (11 compared to 7 and 5 for Pd and Rh respectively), may indicate a large non-linear response for this analyte at the wavelength regions modelled (139) or a lack of model fit.

Because the RRMSE% values (4.14, 3.03 and 1.88 % for Pt, Pd and Rh respectively (Tables 6.1-6.3)) obtained were acceptable PC optimisation within the *stdgen* procedure was not carried out and was simply set equal to the maximum number of calibration subset samples.

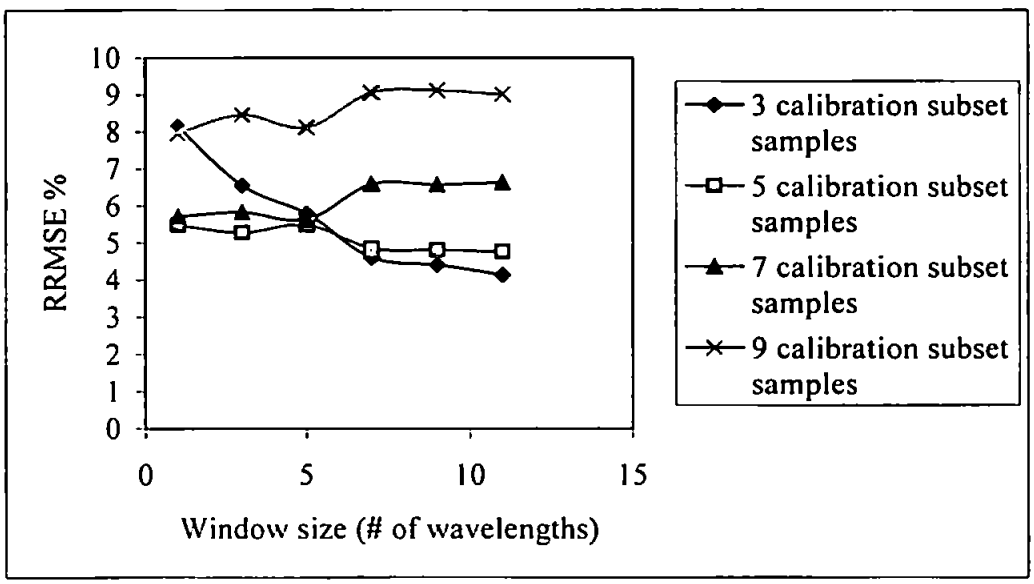


Figure 6.3 RRMSE % values for Pt with 3, 5, 7, and 9 calibration subset samples

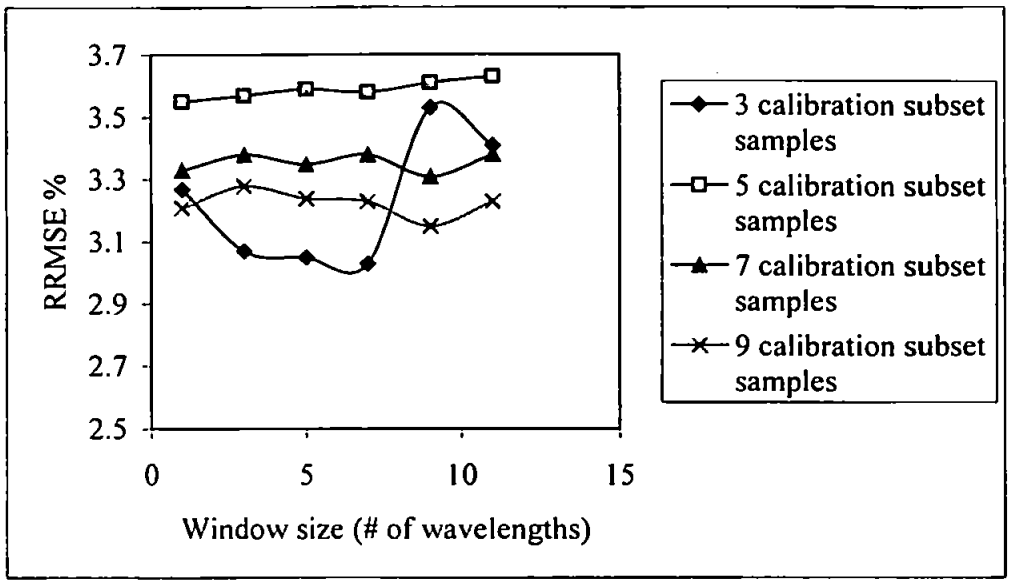


Figure 6.4 RRMSE % values for Pd with 3, 5, 7, and 9 calibration subset samples

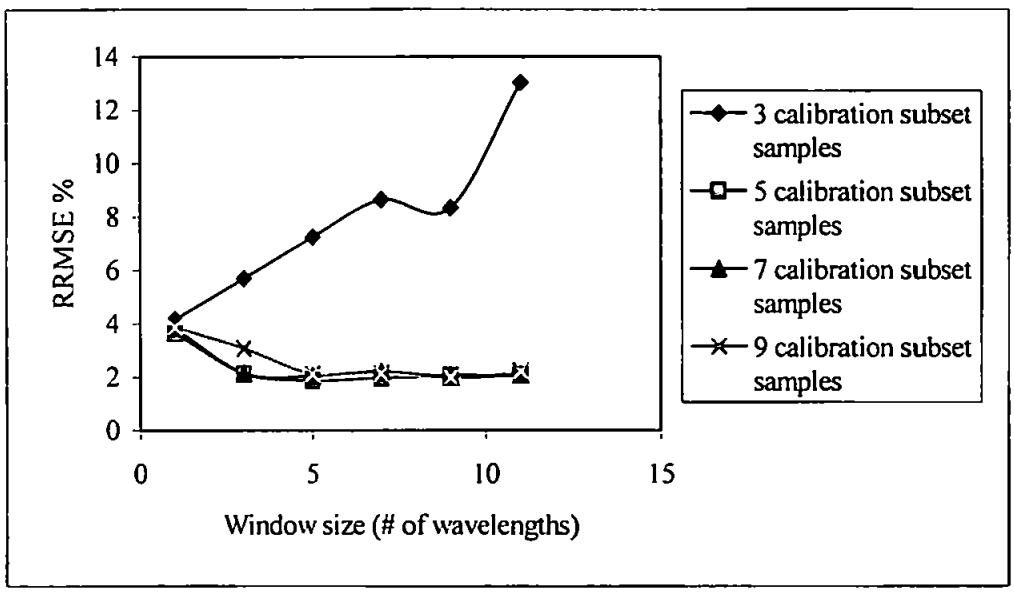


Figure 6.5 RRMSE % values for Rh with 3, 5, 7, and 9 calibration subset samples

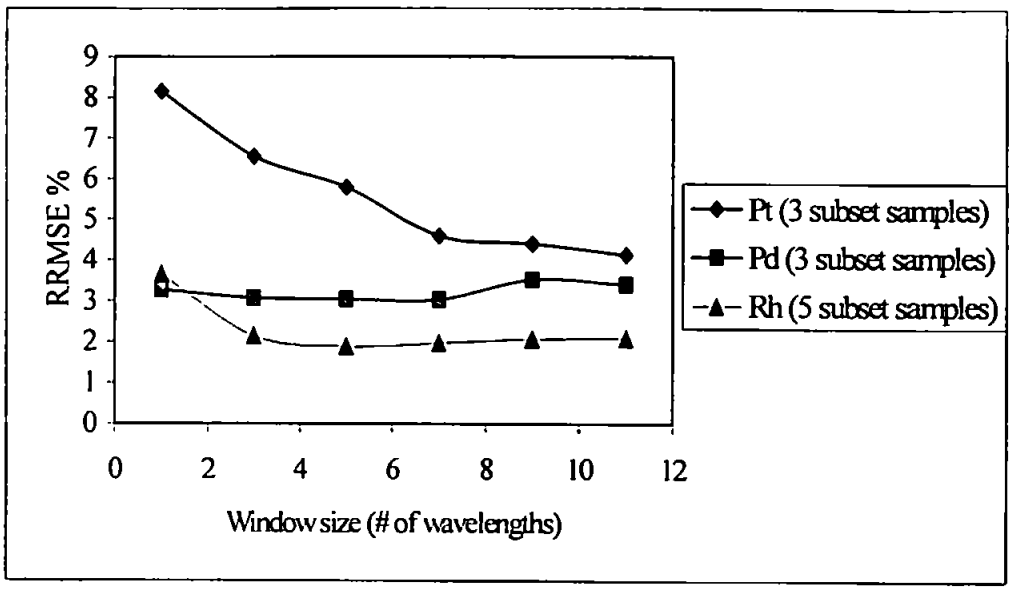


Figure 6.6 Lowest RRMSE % for Pt, Pd and Rh

Table 6.1 RRMSE % values for Pt test samples using different combinations of subset sample number, in addition to the window size and maximum number of principal components used in the stdgen *m*-function (calibration transfer function).

Pt test sample (RRMSE %)							
Sample size	Window size						Maximum number of PC's
	1	3	5	7	9	11	
3	8.15	6.56	5.8	4.6	4.41	**4.14**	3
5	5.48	5.28	5.49	4.86	4.81	4.77	5
7	5.72	5.83	5.64	6.6	6.58	6.64	7
9	7.96	8.46	8.12	9.06	9.12	9.02	9

Table 6.2 RRMSE % values for Pd using different combinations of subset sample number, in addition to the window size and maximum number of principal components used in the stdgen *m*-function (calibration transfer function).

Pd test sample (RRMSE %)							
Sample size	Window size						Maximum number of PC's
	1	3	5	7	9	11	
3	3.27	3.07	3.05	**3.03**	3.53	3.41	3
5	3.55	3.57	3.59	3.58	3.61	3.63	5
7	3.33	3.38	3.35	3.38	3.31	3.38	7
9	3.21	3.28	3.24	3.23	3.15	3.23	9

Table 6.3 RRMSE % values for Rh using different combinations of subset sample number, in addition to the window size and maximum number of principal components used in the stdgen *m*-function (calibration transfer function).

Rh test sample (RRMSE %)							
Sample size	Window size						Maximum number of PC's
	1	3	5	7	9	11	
3	4.17	5.7	7.24	8.64	8.34	13.02	3
5	3.65	2.14	**1.88**	1.97	2.06	2.08	5
7	3.83	2.15	2.04	2.2	2	2.08	7
9	3.89	3.08	2.12	2.14	1.99	2.23	9

6.4.3 Multivariate Calibration and Quantitative Prediction for Synthetic Samples

The results of the application of PDS correction to calibration and test samples analysed 10 days apart (i.e. using $t = 1$ calibration data with standardised $t = 2$ test samples) are shown in Figures 6.7, 6.9 and 6.11 and Tables 6.4-6.6 for Pt, Pd and Rh respectively. As a comparison, results are also given where prediction is performed using $t = 2$ calibration data and test data (Figs. 6.8, 6.10 and 6.12). It is quite evident that without correction, the accuracy of the test sample concentrations was much improved when correction was applied. For Pt, Pd, and Rh respectively the RRMSE % value with correction was 4.14, 3.03 and 1.88% (Tables 6.4-6.6), compared with 73.04, 44.39 and 28.06 % without correction. It is evident that there was a clear bias in the uncorrected concentrations for all three analytes (Figs. 6.7, 6.9 and 6.11). For all three analytes the RRMSE % values for the standardised $t = 2$ test samples, using the $t = 1$ calibration data were approximately the same as the errors when the $t = 2$ calibration data was used to predict the $t = 2$ test samples, indicating that a satisfactory transformation of the $t = 2$ samples had taken place.

Confidence intervals (95% ~ $2 \times \sigma$) for the three analytes also showed a significant improvement with the application of PDS correction. This is most noticeable for Rh, Pt, shows a moderate improvement, with Pd showing no significant difference in the confidence intervals (Figs. 6.7, 6.9 and 6.11). The relatively large confidence intervals shown by Rh (Fig. 6.11) without correction may be correlated to the much lower concentration range of the Rh test samples (between 0 – 10 $\mu g\ ml^{-1}$), compared with both Pt and Pd which varied between 0 – 40 $\mu g\ ml^{-1}$.

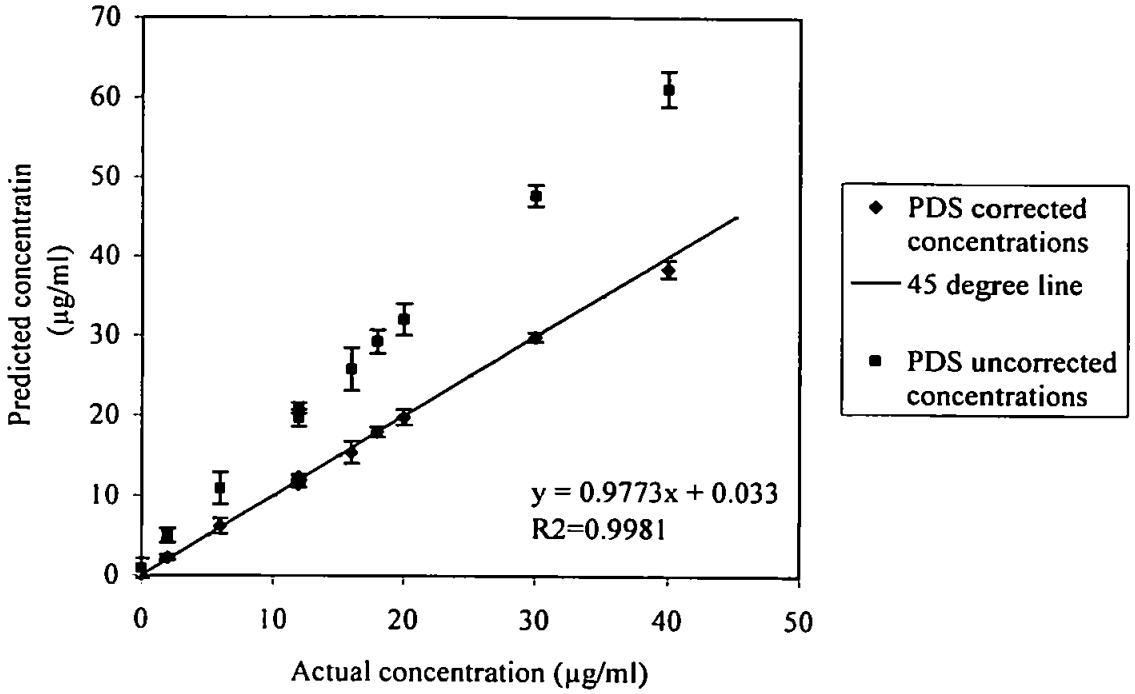


Figure 6.7 Actual vs predicted concentrations for Pt test solutions with and without PDS correction, using $t = 1$ calibration data and $t = 2$ standardised test data (95 % c.i.)

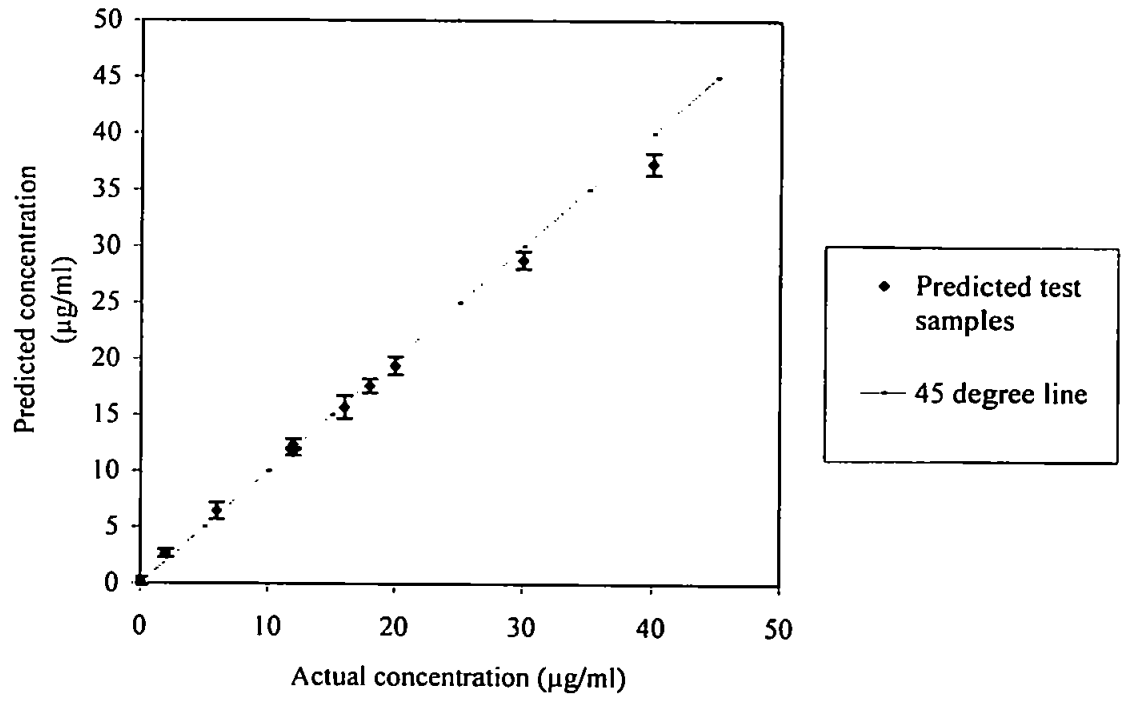


Figure 6.8 Actual vs predicted concentrations for Pt test solutions using $t = 2$ calibration and test data (95 % confidence interval)

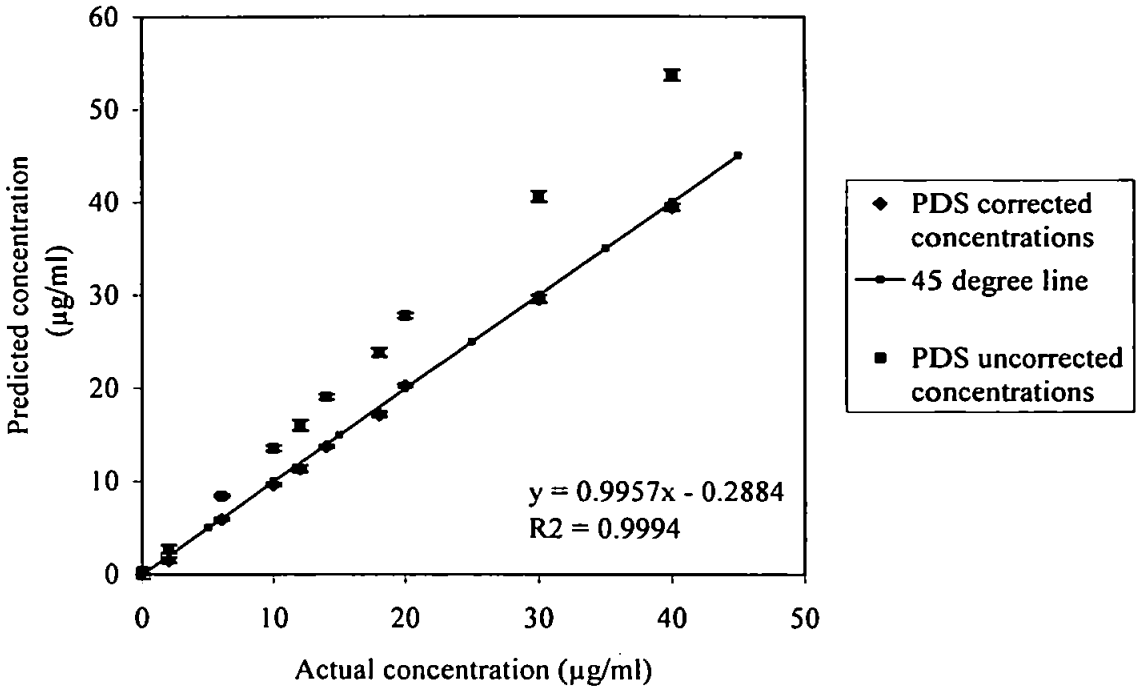


Figure 6.9 Actual vs predicted concentrations for Pd test solutions with and without PDS correction, using $t = 1$ calibration data and $t = 2$ standardised test data (95 % c.i.)

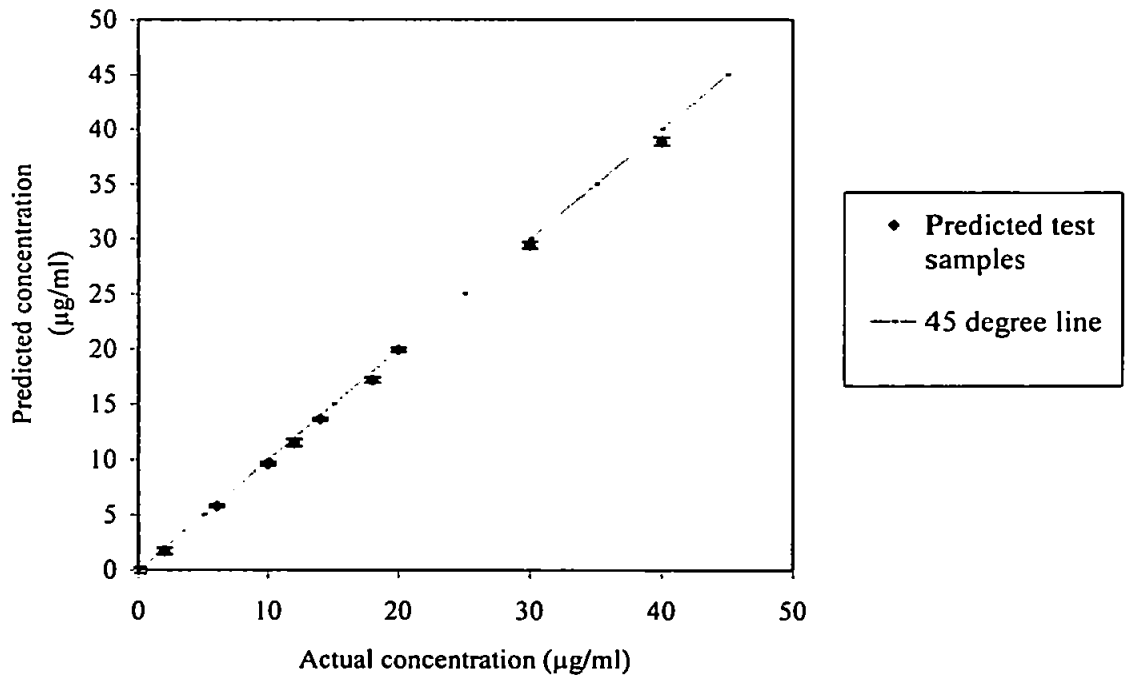


Figure 6.10 Actual vs predicted concentrations for Pd test solutions using $t = 2$ calibration and test data (95 % confidence interval)

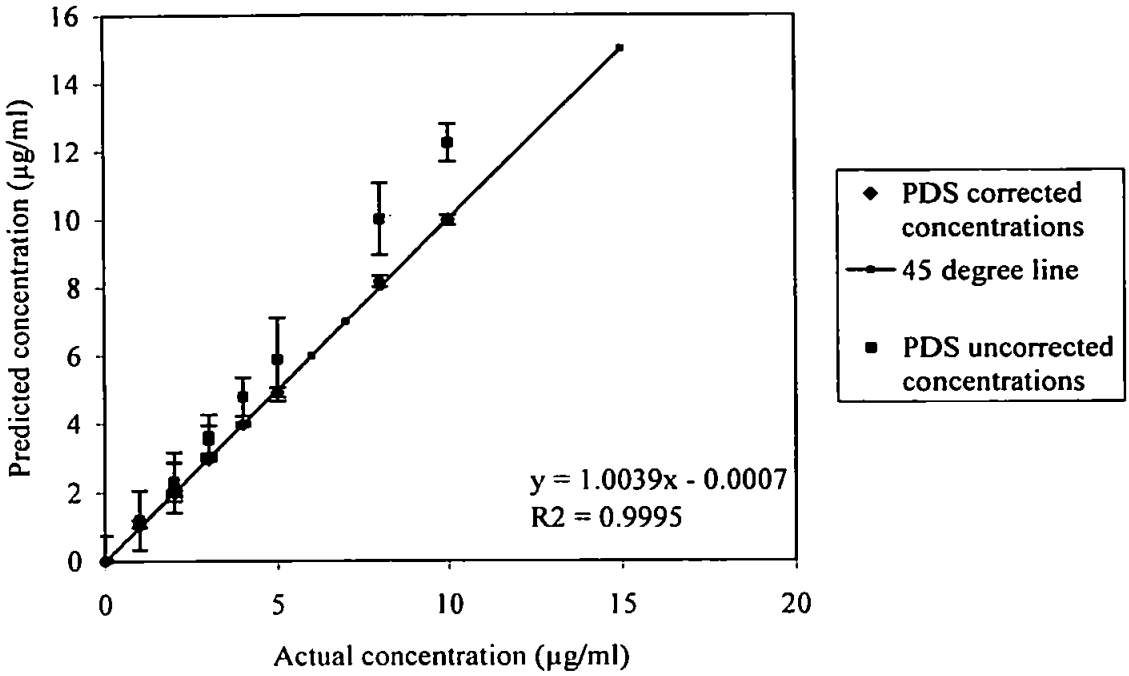


Figure 6.11 Actual vs predicted concentrations for Rh test solutions with and without PDS correction, using $t = 1$ calibration data and $t = 2$ standardised test data (95 % c.i.)

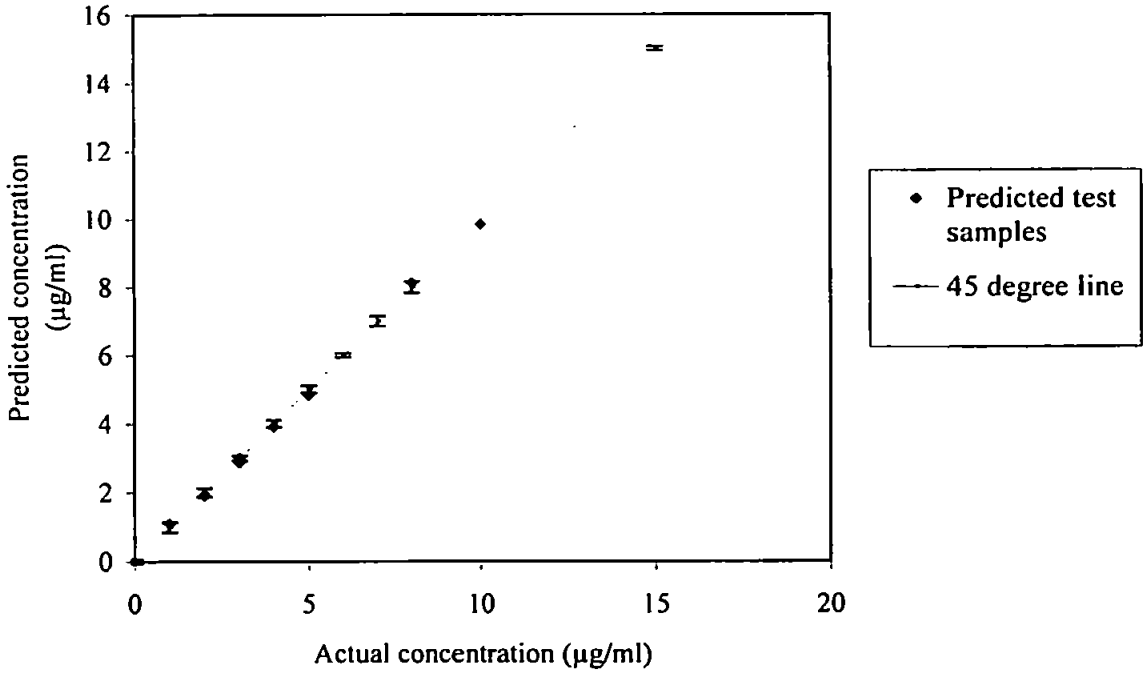


Figure 6.12 Actual vs predicted concentrations for Rh test solutions using $t = 2$ calibration and test data (95 % confidence interval)

Table 6.4 Actual and predicted concentrations (μg/ml), including 95% confidence intervals, for Pt test samples with and without PDS correction

	Pt			
Pred	Actual (corrected)	C. I. (corrected)	Actual (uncorrected)	C. I. (uncorrected)
12	12.34	0.32	20.96	0.65
16	15.43	1.38	25.82	2.67
20	19.81	0.96	32.04	1.99
12	11.46	0.39	19.71	1.07
18	18.01	0.62	29.27	1.46
6	6.18	0.98	10.90	1.98
2	2.21	0.34	4.97	0.89
40	38.43	1.08	61.10	2.20
30	29.83	0.58	47.64	1.34
0	-0.92	0.53	0.88	1.19
RRMSE %	4.14		73.04	

Table 6.5 Actual and predicted concentrations (μg/ml), including 95% confidence intervals, for Pd test samples with and without PDS correction

	Pd			
Pred	Actual (corrected)	C. I. (corrected)	Actual (uncorrected)	C. I. (uncorrected)
20	20.30	0.19	27.79	0.28
12	11.35	0.38	16.04	0.57
18	17.20	0.29	23.85	0.46
14	13.76	0.18	19.09	0.27
10	9.64	0.19	13.56	0.29
30	29.58	0.38	40.58	0.59
6	5.89	0.15	8.40	0.15
2	1.51	0.30	2.69	0.44
40	39.44	0.37	53.69	0.58
0	-0.21	0.28	0.32	0.38
RRMSE %	3.03		44.39	

Table 6.6 Actual and predicted concentrations (μg/ml), including 95% confidence intervals, for Rh test samples with and without PDS correction

	Rh			
Pred	Actual (corrected)	C. I. (corrected)	Actual (uncorrected)	C. I. (uncorrected)
3	2.96	0.05	3.53	0.43
5	4.92	0.14	5.89	1.22
2	1.98	0.11	2.28	0.88
4	3.99	0.06	4.79	0.57
3	3.04	0.09	3.63	0.63
1	1.08	0.11	1.19	0.86
2	1.99	0.05	2.30	0.55
8	8.18	0.16	10.01	1.05
10	9.98	0.15	12.24	0.55
0	0.01	0.05	-0.13	0.89
RRMSE %	1.88		28.06	

6.5 Conclusions

The application of PDS, in conjunction with variable elimination and selection, has shown that it is not only possible to use large amounts of information in the ICP-AES emission spectrum without having to resort to line selection but that it is also possible to use the same calibration model over a period of time.

Compared to the errors observed when synthetic test samples were modelled with calibration data analysed at the same time, standardised predictions gave errors that were comparable, indicating that successful transformation of the spectra from $t = 2$ to $t = 1$ had been accomplished.

The following general conclusions can be draw from the work discussed in the preceding chapters:

7.1 Final Conclusions

1 Interferences can range from relatively simple background shifts which are easily corrected for, to more complex spectral overlaps. The number and nature of interferents has been shown to affect the accuracy and precision of IEC factors, with accurate prediction only possible with the correct positioning of the BCPs or when matrix effects are minimal. The contribution to the analyte signal by a relatively small interference can lead to a degradation in precision of the IEC factor compared with the use of a mixed solution containing the same interferents. However, this can be overcome by using mixed interferent solutions.

2 A range of univariate calibration techniques, including interelement correction and matrix matching, were shown to be severely limited in their application due to the presence of interferences, both spectral and non-spectral. Overall, matrix matching was the most accurate of the univariate calibration methods and was comparable to, or better than, IEC correction. The accuracy of prediction in the autocatalyst samples was not as good as for the test solutions, which was most probably due to unaccounted for interferents. Multivariate calibration was initially performed using net signal intensities. In order to obtain the net signal from the instrument a 2-point background correction was performed by the instrument software, however, the selection of optimum background correction points for all 248 lines proved problematical due to the presence of adjacent spectroscopic interferences. It was therefore decided to model the data using gross intensities which did not require the use of background correction. The RRMSE values for Pt, Pd and Rh in the

test solutions were 5.77, 2.96 and 3.46 % respectively (Table 3.4), indicating that, overall, this calibration strategy was as good as matrix matching.

3 The application of variable elimination and selection algorithms has shown that it is possible to use the complete available ICP-AES emission spectrum for multivariate modelling without having to resort to line selection or the need to assign background correction points in order to obtain the net analyte signal. Indeed, one of the benefits of this approach is the selection of parts of the spectrum which appear uninformative, such as continuum background, but which can be highly informative to a bilinear modelling technique such as PLS, which is able to detect useful variation in all parts of the spectrum. The new method has several desirable properties: it is computationally simple, it has significance tests of multivariate model parameters and allows the calculation of independent test data confidence intervals. The model errors for the independent test data, have shown considerable improvement compared with the errors achieved when using all 5684 wavelengths and marginal improvement compared with the more traditional individual wavelength data set consisting of intense analyte and matrix lines. Of greater interest is the important advantage of being able to utilise the full spectrum because the need for individual line selection is no longer required. In situations where such systems are not fully understood the selection of pertinent individual lines may prove impossible. The variable selection techniques have also been used to develop a multivariate confidence interval which can be used not only for the PLS parameters, but also for independent point predictions. The main limitations of the technique are, firstly that large numbers of t-tests are required resulting in uninformative variables being incorrectly classified as significant, i.e. informative. And secondly that the correct number of PCs are required for the UVE and IVD-PLS algorithms. Successful use of the technique requires that these two criteria be carefully considered.

4 Spectral data obtained for the analysis of fusion digests has been used to build multivariate calibration models using PLS to predict the concentration of Au, Ag and Pd in test samples. In order to achieve this, variable elimination and selection algorithms were used to select the informative parts of the ICP-AES emission spectra without having to resort to line selection or the need to assign background correction points in order to obtain the net integrated line intensities. The model errors for both the calibration and independent test data, have shown considerable improvement compared with the errors achieved when using all 2268 wavelengths thus reinforcing the fact that PLS benefits from selective variable reduction. The variable selection method and PLS multivariate calibration gave results comparable to those obtained using a more traditional univariate calibration approach with interelement correction.

5 A simulation study has demonstrated that piecewise direct standardisation is potentially useful to transfer calibrations over time when recalibration using the entire calibration set is not a viable alternative, providing that optimisation of both window size and the number of calibration subsets samples is carried out.

7.2 Suggestions for Future Work

The work described in the preceding chapters could be developed in a number of ways. Possible areas for further investigation are summarised below:

7.2.1 Short Term Projects (Industrial Data using the Full Spectrum)

1. Only 36 % of the full segmented spectrum were obtained for the industrial samples in this study. Further investigations using data from all 201 subarrays would enable the variable selection algorithms to select any spectral area covered by the subarrays. This has the potential to reduce predictive error even further because the information inputted into the selection algorithms is not limited in any way.

2. The use of variable selection with the industrial samples was limited in that only 111 samples were available in total. Further work is required with a larger sample database which would enable sample clustering, on the basis of their composition, to be identified more easily. This may reduce predictive errors even further.

3. Application of the selection routines to other forms of spectroscopy, such as broad spectrum techniques.

4. The application of calibration transfer was limited to that of complex synthetic samples only. The use of real industrial samples would enable a full validation of PDS.

7.2.2 Short Term Projects (UVE and IVD-PLS Algorithm Adaptations)

1. Both the UVE and IVD-PLS routines used the $\overline{\beta}_j$ in order to obtain an estimate of β. Further investigation into the effects of β estimators are required to determine the relationship of such estimates with the Jackknife standard error For example, the use of the optimum number of PCs with all calibration samples.

2. The UVE-PLS algorithm has the limitation in that the optimum PC (Aopt) is decided upon by 'bracketing'. Further work is required in order to determine Aopt objectively.

3. The Jackknife has now been improved upon and the use of Bootstrap techniques have been shown to give more accurate estimations of standard error. Both the UVE and PLS algorithms may benefit from the use of Bootstrap.

4. Allied to the use of the Bootstrap is the estimation of prediction confidence intervals. The use of the Jackknife does give a crude estimate, however Bootstrap methods are now being developed which have been shown to outperform the Jackknife confidence interval estimate. In addition, the Bootstrap could be used to overcome the deficiency of the Jackknife for estimating calibration confidence limits.

5. In performing multiple t-test's the problem arises that $\alpha \times \#\ var$ will be identified as significant even when they are not. Investigations looking at whether the inclusion of an uninformative variable gives a poorer model than the elimination of that same variable if it were informative would then reveal how useful the 'Bonferroni' correction (159) would be. Allied to this is an investigation into Type II errors and whether additional informative variables can be identified from those rejected by UVE-PLS.

7.2.3 Short Term Projects (Instrumental Drift and Multivariate Limit of Detection)

1. Adaptations to the PDS algorithm may allow its use in correcting for short term instrumental drift. This would be of use when 'clean' internal standard lines are not

available which is often the case with complex matrices.

2. Of interest in any quantitative determination is the limit of detection. The construction of a multivariate limit of detection, however has received very little attention. One possible solution may be to use the sum of each PCs score vector in place of the net analyte signal as in Eqn. 7.1:

$$c_L = \frac{z \times \sigma_{b,\sum_{i=1}^{n} PC}}{b_1}$$

where z is a statistical constant, usually 2, $\sqrt{2}$, or 3; $\sigma_{b,\sum_{i=1}^{n} PC}$ is the standard deviation (standard error) for the sum of the score vectors for PCs *1-n*, derived from a number of replicate blank solutions and b_1 is a composite sensitivity.

APPENDICES

(I-IV)

APPENDIX I

Matlab Program for the Ranking of the Signal Standard Deviation.

```
Standard deviation ranking-PLS. Inputs are the predictor block (x),
predicted block (y),the number of latent variables to be
calculated(maxlv) and an optional variable (out) to suppress intermediate
output [out=0 suppresses output]. Optional Outputs are the vector of
regression vectors(m),x loadings (p), y loadings (q), x weights (w)x
scores (t), y scores (u), and inner relation coefficients (b). The main
output is the ranked emission matrix (samples x ranked wavelengths).


[mx,nx]         =size(x);
[newmx,newnx]   =size(x);
[min,nin]       =size(x);
[mm,nm]         =size(x);
[my1,ny1]       =size(x);
msj             =size(x,2);
mw              =size(x,2);

n1=mx*nx;
sj=std(x);
sj=sj';
[y1,in]=sort(sj);
[min,nin]=size(in);

n=0
for j=min:-1:1;
   n=n+1;
   ww1(n,:)=in(j);
end

new_model_headers=model_headers(:,ww1);
newauto=auto(:,ww1);
newtest=test(:,ww1);
test=newtest;

newx=x(:,ww1);
x=newx;

if pre==0;
   y=y;
   test=test;
elseif pre==1;
   [mcx,nnx]=mncn(x);
   [mcy,nny]=mncn(y);
   mc_test=scale(test,nnx);
   x=mcx;
   y=mcy;
   test=mc_test;
elseif pre==2
   [ax,mmx,stdx]=auto(x);
   [ay,mmy,stdy]=auto(y);
   auto_test=scale(test,mmx,stdx);
   x=ax;
   y=ay;
```

```
   test=auto_test;
end
[mx,nx]=size(x);
lv=input('How many latent variables are required?');
if lv>nx
   error ('Too many latent variables !')
end

[mx,nx] = size(x);
[my,ny] = size(y);
q       = zeros(ny,lv);
t       = zeros(mx,lv);
u       = zeros(my,lv);
bin     = zeros(1,lv);
rankx = rank(x);
olv = lv;
rankx = rank(x);
if rankx < olv
   lv = rankx;
end

j=input ('What jump do you require (<nx)?');
var=input('Do you want calibration (3) or test (4) rrmse?');
co=input('What cut-off for rrmsec do you want?');
XX=x;
TT=test;
counter=0;
n=0;
for jumpn=1:1:fix(nx/j);
   CO=fix(co/j);
   n=n+1;
   counter=counter+j
   xnew=XX(:,1:counter);
   tnew=TT(:,1:counter);

   YY=y;
   x=xnew;
   if var==3;
      X=xnew;
   else
      X=tnew;
   end
   p=zeros(counter,lv);
   w=zeros(counter,lv);

   for i = 1:lv;
      [pp,qq,ww,tt,uu] = plsnipal(x,y);
      bin(1,i)   = uu'*tt/(tt'*tt);
      x          = x - tt*pp';
      y          = y - bin(1,i)*tt*qq';
      t(:,i)     = tt(:,1);
      u(:,i)     = uu(:,1);
      p(:,i)     = pp(:,1);
      w(:,i)     = ww(:,1);
      q(:,i)     = qq(:,1);
      yprdn=nipals_pred(X,bin,p,q,w,i);

      if pre==0;
         reyprdn=yprdn;
      elseif pre==1;
         reyprdn=rescale(yprdn,nny);
      elseif pre==2
         reyprdn=rescale(yprdn,mmy,stdy);
```

```
      end
      le=length(reyprdn);
      if var==3;
         [ym,yn]=size(true_c);
         ymean=sum(true_c(:,1))/ym;
         rrmse =(sqrt((sum((true_c-reyprdn).^2))/le))*(100/ymean);
      else
         [ym,yn]=size(test_conc);
         ymean=sum(test_conc(:,1))/ym;
         rrmse =(sqrt((sum((test_conc-reyprdn).^2))/le))*(100/ymean);
      end
      srrmse(n,i)=rrmse;
   end
   y=YY;
   if CO==jumpn
      plot(srrmse)
      xlabel('Number of variables ')
      ylabel('RRMSE (%)')
      title('RRMSE (%) versus number of wavelengths')
      return;
   end
end

function yprdn = nipals_pred(X,bin,p,q,w,i);
x=X;
[mx,nx] = size(x);
[mq,nq] = size(q);
[mw,nw] = size(w);
that    = zeros(mx,i);
yprdn   = zeros(mx,mq);

if i>nw
   error(sprintf('Maximum number of latent variables exceeded (Max =
%g)',nw));
end
for ii=1:i
   that(:,ii) = x*w(:,ii);
   x  = x - that(:,ii)*p(:,ii)';
end
for ii=1:i
   yprdn = yprdn + bin(1,ii)*that(:,ii)*q(:,ii)';
end
```

Matlab Program for Uninformative Variable Elimination by Partial Least Squares

```
UVE-PLS

Algorithm computes t-test statistic of b for each wavelength and
deletes those whos value = 0 at the stated significance level.
Inputs are the predictor block (x), predicted block (y)
the number of latent variables to be calculated(maxlv)
and an optional variable (out) to suppress intermediate
output [out=0 suppresses output].Optional Outputs are the the vector of
regression vectors(m),x loadings (p), y loadings (q), x weights (w)
x scores (t), y scores (u), and inner relation coefficients (b).
predominant outputs are the x, y and test x data for the IVD-PLS routine.


[mx,nx]=size(x);
[mmy,ny]=size(y);
XX=x;
YY=y;
TT=test;

for loo=1:mx;
   count=loo
   x(loo,:)=[];
   y(loo,:)=[];
   if pre==0
      test=test;
      x=XX;
      y=YY;
   elseif pre==1
      [mcx,nncx]=mncn(x);
      x=mcx;
      [mcy,ny]=mncn(y);
      y=mcy;
   elseif pre==2
      [ax,amx,stdx]=auto(x);
      x=ax;
      [ay,amy,stdy]=auto(y);
      y=ay;
   end
   [mx,nx] = size(x);
   [my,ny] = size(y);
   p       = zeros(nx,lv);
   q       = zeros(ny,lv);
   w       = zeros(nx,lv);
   t       = zeros(mx,lv);
   u       = zeros(my,lv);
   b       = zeros(1,lv);
   olv = lv;
   rankx = rank(x);
   if rankx < olv
      lv = rankx;
      disp('  ')
      sss = sprintf('Rank of X is %g, which is less than lv of
%g',lv,olv);
```

```
      disp(sss);
      sss = sprintf('Calculating %g LVs only',lv);
      disp(sss);
   end
   for i = 1:lv
      [pp,qq,ww,tt,uu] = plsnipal(x,y);
      b(1,i)   = uu'*tt/(tt'*tt);
      x        = x - tt*pp';
      y        = y - b(1,i)*tt*qq';
      t(:,i)   = tt(:,1);
      u(:,i)   = uu(:,1);
      p(:,i)   = pp(:,1);
      w(:,i)   = ww(:,1);
      q(:,i)   = qq(:,1);
   end

   m = zeros(lv,nx);
   m(1:lv,:) = conpred(b,w,p,q,lv);
   m= cumsum(m,1);
   m_opt=m(lv,:);
   summ_opt(loo,:)=m_opt;
   x=XX;
   y=YY;
end

[mx,nnx]=size(XX);
S=summ_opt;
for m=1:mx
   summ_opt(m,:)=[];
   mu(m,:)=mean(summ_opt);
   summ_opt=S;
   MMM(m,:)=mean(summ_opt)-summ_opt(m,:);
end

summ_loo_squared=sum(MMM.^2);
jk_stderror=sqrt(((mx-1)/mx)*summ_loo_squared);
jk_mean=(sum(mu))/mx;
fc=(jk_mean./jk_stderror);

alpha=input('What alpha level do you want?');
pause;
dof=input('What are the degrees of freedom');
z=2;
ttest=ttestp(alpha,dof,z);
[i,k]=find(abs(fc)<ttest);
unscaled_xone=XX;
unscaled_xone(:,k)=[];
Utest=TT;
Utest(:,k)=[];
[mx,nx]=size(Utest);
nmh=(1:nnx);
amh=model_headers;
amh(:,k)=[];
nmh(:,k)=[];
nmtwo=mtwo;
nmtwo(:,k)=[];
nmtwotest=mtwotest;
nmtwotest(:,k)=[];
nmauto=auto;
nmauto(:,k)=[];

[mx,nx]=size(XX);
t=ttestp(alpha,mx,2);
```

```
ci=t.*jk_stderror;

function m = conpred(b,w,p,q,lv)
[mq,nq] = size(q);
[mw,nw] = size(w);
if nw ~= lv
   if lv > nw
      s = sprintf('Original model has a maximum of %g LVs',nw);
      disp('  '), disp(s)
      s = sprintf('Calculating vectors for %g LVs only',nw);
      disp(s), disp('  ')
      lv = nw;
   else
      w = w(:,1:lv);
      q = q(:,1:lv);
      p = p(:,1:lv);
      b = b(:,1:lv);
   end
end
m = zeros(lv,mw);
m = (w*inv(p'*w)*diag(b))';
```

APPENDIX III

Matlab Program for Informative Variable Degradation by Partial Least Squares (IVD-PLS).

```
IVD-PLS
Algorithm computes b/se(b) ratio for each wavelengths IVD value
and ranks in descending order Inputs are the predictor block (x),
predicted block (y)the number of latent variables to be calculated(maxlv)
and an optional variable (out) to suppress intermediate
output [out=0 suppresses output].Optional Outputs are the vector of
regression vectors(m),x loadings (p), y loadings (q), x weights (w)
x scores (t), y scores (u), and inner relation coefficients (b).
Predominant output are the x, y and test x matrices for final PLS1
modelling.

[mx,nx]=size(x);
XX=x;
YY=y;

for loo=1:mx;
   count=loo
   x(loo,:)=[];
   y(loo,:)=[];

   if pre==0
      x=XX;
      y=YY;
   elseif pre==1
      [mcx,nncx]=mncn(x);
      x=mcx;
      [mcy,ny]=mncn(y);
      y=mcy;
   elseif pre==2
      [ax,amx,stdx]=auto(x);
      x=ax;
      [ay,amy,stdy]=auto(y);
      y=ay;
   end

   [mx,nx] = size(x);
   [my,ny] = size(y);
   p       = zeros(nx,lv);
   q       = zeros(ny,lv);
   w       = zeros(nx,lv);
   t       = zeros(mx,lv);
   u       = zeros(my,lv);
   b       = zeros(1,lv);
   olv = lv;
   rankx = rank(x);
   if rankx < olv
      lv = rankx;
      disp('  ')
      sss = sprintf('Rank of X is %g, which is less than lv of
%g',lv,olv);
      disp(sss);
```

```
      sss = sprintf('Calculating %g LVs only',lv);
      disp(sss);
   end
   for i = 1:lv
      [pp,qq,ww,tt,uu] = plsnipal(x,y);
      b(1,i)   = uu'*tt/(tt'*tt);
      x        = x - tt*pp';
      y        = y - b(1,i)*tt*qq';
      t(:,i)   = tt(:,1);
      u(:,i)   = uu(:,1);
      p(:,i)   = pp(:,1);
      w(:,i)   = ww(:,1);
      q(:,i)   = qq(:,1);
   end

   m = zeros(lv,nx);
   m(1:lv,:) = conpred(b,w,p,q,lv);
   m= cumsum(m,1);

   m_opt=m(lv,:);
   summ_opt(loo,:)=m_opt;
   x=XX;
   y=YY;
end

[mx,nx]=size(XX)
S=summ_opt;
for m=1:mx
   summ_opt(m,:)=[];
   mu(m,:)=mean(summ_opt);
   summ_opt=S;
   MMM(m,:)=mean(summ_opt)-summ_opt(m,:);
end

summ_loo_squared=sum(MMM.^2);
jk_stderror=sqrt(((mx-1)/mx)*summ_loo_squared);
jk_mean=(sum(mu))/mx;
B=abs(jk_mean./jk_stderror);
[c_rank,in]=sort(B);

[mx,nx]=size(jk_mean);
N=0;
for n=nx:-1:1
   N=N+1;
   b_in(:,N)=in(:,n);
   b_rank(:,N)=c_rank(:,n);
end

unscaled_xtwo=XX(:,b_in);
UUtest=Utest(:,b_in);
[mx,nx]=size(jk_mean);
famh=amh(:,b_in);
fnmh=nmh(:,b_in);
fci=ci(:,b_in);
fnmtwo=nmtwo(:,b_in);
fnmtwotest=nmtwotest(:,b_in);
fnmauto=nmauto(:,b_in);

slevel=input('What start level do you want');
pause
step=input('What step level do you want');
pause
pcstep=input('What % punish factor do you want');
```

```
[ooptmin,rmsecv,csx] = opt(b_rank,unscaled_xtwo,slevel,pcstep,step,YY);
mini=min(ooptmin);
[i,j]=find(ooptmin==mini);
pcl=slevel+((i-1)*step);
ul=pcl*max(csx);
vars=(find((ul-csx)>0));
v=sprintf(' Use %g variables',max(vars));disp(v);
function [ooptmin,rmsecv,csx] =
opt(b_rank,unscaled_xtwo,slevel,pcstep,step,YY)
x=unscaled_xtwo;
y=YY;
n=0;
csx=cumsum(b_rank);
for m=slevel:step:1.0
   n=n+1;
   ul=m*max(csx);
   loc=(find((ul-csx)>0));
   optx=x(:,1:max(loc));
   [mx,nx]=size(optx);
   if nx>10
      lv=10;
   else
      lv=nx;
   end
   [press,cumpress,rmsecv] =
crossval(optx,y,'nip','loo',lv,[],[],1,1,[]);
   optmin=min(rmsecv);
   [i,j]=find(optmin==rmsecv);
   pcs=j;
   if pcs ==1;
      ooptmin(n,:)=optmin;
   else
      while optmin+(pcstep*optmin)>rmsecv(:,(pcs-1));
         optmin=rmsecv(:,(pcs-1));
         pcs=pcs-1;
         if pcs<2
            break
         end
      end
   end
   ooptmin(n,:)=optmin;
   clear optmin;
end
function m = conpred(b,w,p,q,lv)
[mq,nq] = size(q);
[mw,nw] = size(w);
if nw ~= lv
   if lv > nw
      s = sprintf('Original model has a maximum of %g LVs',nw);
      disp('  '), disp(s)
      s = sprintf('Calculating vectors for %g LVs only',nw);
      disp(s), disp('  ')
      lv = nw;
   else
      w = w(:,1:lv);
      q = q(:,1:lv);
      p = p(:,1:lv);
      b = b(:,1:lv);
   end
end
m = zeros(lv,mw);
m = (w*inv(p'*w)*diag(b))';
```

Matlab Program for the Jackknife confidence interval.

```
Jackknife confidence interval-PLS. Algorithm computes b/se(b) ratio for
each wavelength and ranks indescending ordder Inputs are the predictor
block (x), predicted block (y)the number of latent variables to be
calculated(maxlv)and an optional variable (out) to suppress intermediate
output [out=0 suppresses output].Optional Outputs are the the vector of
regression vectors (m),x loadings (p), y loadings (q), x weights (w)
x scores (t), y scores (u), and inner relation coefficients (b).
Predominant output is the Jackknife standard error of the test samples.


[mx,nx]=size(x);
[my,ny]=size(y);
XX=x;
YY=y;

for loo=1:mx;
   count=loo
   x(loo,:)=[];
   y(loo,:)=[];

   [mcx,nncx]=mncn(x);
   x=mcx;
   [mcy,nncy]=mncn(y);
   y=mcy;
   sx=scale(test,nncx);


   [mx,nx] = size(x);
   [my,ny] = size(y);
   p       = zeros(nx,lv);
   q       = zeros(ny,lv);
   w       = zeros(nx,lv);
   t       = zeros(mx,lv);
   u       = zeros(my,lv);
   b       = zeros(1,lv);
   olv = lv;
   rankx = rank(x);
   if rankx < olv
      lv = rankx;
      disp('  ')
      sss = sprintf('Rank of X is %g, which is less than lv of
%g',lv,olv);
      disp(sss);
      sss = sprintf('Calculating %g LVs only',lv);
      disp(sss);
   end
   for i = 1:lv
      [pp,qq,ww,tt,uu] = plsnipal(x,y);
      b(1,i)    = uu'*tt/(tt'*tt);
      x         = x - tt*pp';
      y         = y - b(1,i)*tt*qq';
      t(:,i)    = tt(:,1);
      u(:,i)    = uu(:,1);
      p(:,i)    = pp(:,1);
```

```
      w(:,i)    = ww(:,1);
      q(:,i)    = qq(:,1);
   end
   m = zeros(lv,nx);
   m(1:lv,:) = conpred(b,w,p,q,lv);
   m= cumsum(m,1);
   predn=sx*m(lv,:)';
   predns=rescale(predn,nncy);
   fpred(:,count)=predns;
   x=XX;
   y=YY;
end
fpred=fpred';

[mx,nx]=size(fpred);
F=fpred;
for m=1:mx
   fpred(m,:)=[];
   mu(m,:)=mean(fpred);
   fpred=F;
   MMM(m,:)=mean(fpred)-fpred(m,:);
end

[mx,nx]=size(XX);
summ_loo_squared=sum(MMM.^2);
jk_stderror=sqrt(((mx-1)/mx)*summ_loo_squared);
jk_mean=sum(mu)/mx;
fc=(jk_mean./jk_stderror);

alpha=input('What alpha level do you want?');
pause;
dof=input('What are the degrees of freedom');
z=2;
ttest=ttestp(alpha,dof,z);
test_ci=jk_stderror.*ttest;

function m = conpred(b,w,p,q,lv)
[mq,nq] = size(q);
[mw,nw] = size(w);
if nw ~= lv
   if lv > nw
      s = sprintf('Original model has a maximum of %g LVs',nw);
      disp('  '), disp(s)
      s = sprintf('Calculating vectors for %g LVs only',nw);
      disp(s), disp('  ')
      lv = nw;
   else
      w = w(:,1:lv);
      q = q(:,1:lv);
      p = p(:,1:lv);
      b = b(:,1:lv);
   end
end
m = zeros(lv,mw);
m = (w*inv(p'*w)*diag(b))';
```

A stand alone Guided User Interface (GUI) incorporating all of those algorithms shown in appendices i-iv is included on a CD-ROM together with operating guidelines.

REFERENCES

(1) Boumans, P. W. J. M. *Inductively Coupled Plasma Emission Spectroscopy Part 1: Methodology, Instrumentation, and Performance*; John Wiley & Sons: New york, Chichester, Brisbane, Toronto, Singapore, 1987.

(2) Thompson, M.; Walsh, J. N. *Handbook of Inductively Coupled Plasma Spectrometry*; Blackie: Glasgow and London, 1989.

(3) Boumans, P. W. J. *Inductively Coupled Plasma Atomic Emission Spectroscopy: Part 2: Applications and Fundementals*; John Wiley & Sons: New York, Chichester, Brisbane, Toronto, Singapore, 1987.

(4) Pimentel, M. F.; Pasquini, B. d. B. N. C.; Araujo, M. C. U. d. *Spectrochimica Acta* **1997,** *52B*, 2151.

(5) Babat, G. I. *Vestn. Elektroprom.* **1942,** *2*, 1.

(6) Babat, G. I. *Vestn. Elektroprom* **1942,** *3*, 2.

(7) Reed, T. B. *J. Applied Physics* **1961,** *32*, 821.

(8) Reed, T. B. *J. Applied Physics* **1961**, 2534.

(9) Reed, T. B. *Int. Sci. Technol.* **1962,** *6*, 42.

(10) Fassel, V. A. *ICP Information Newsletter* **1976,** *1*, 267.

(11) Fassel, V. A. *Analytical Chemistry* **1979,** *51*, 1290A.

(12) Greenfield, S. *ICP Information Newsletter* **1975,** *1*, 3.

(13) Greenfield, S. *Analyst* **1964,** *89*, 713.

(14) Wendt, R. H.; Fassel, V. A. *Analytical Chemistry* **1965,** *37*, 920.

(15) Dickenson, G. W.; Fassel, V. A. *Analytical Chemistry* **1969,** *41*, 1021.

(16) Fassel, V. A. In *Proc. 16th Coll. Spectr. Int. Heidelberg 1971, Plenary lectures and reports*; Adam Hilger: London, 1972, pp 63.

(17) Wendt, R. H.; Fassel, V. A. *Analytical Chemistry* **1966,** *38*, 337.

(18) Greenfield, S.; Jones, I. L.; Berry, C. T. *Analyst* **1964,** *89*, 713.

(19) Greenfield, S.; Berry, C. T.; Bunch, L. G. *Spectroscopy with a High Frequency Plasma Torch*, Wokingham, England 1965; Radyne International.

(20) Greenfield, S.; Jones, I. L. W.; Berry, C. T. In *U. S. Patent 3,467,471*: United States, 1969.

(21) Greenfield, S.; Jones, I. L. W.; Berry, C. T.; Bunch, L. G. *Proc. Soc. Anal. Chem.* **1965,** *2*, 385.

(22) Barnes, R. M. *Emission Spectroscopy*; Dowden, Hutchinson & Ross: Stroudsburg, PA, 1796.

(23) Ebdon, L.; Evans, E. H.; Fisher, A. S.; Hill, S. J. *An Introduction to Analytical Atomic Spectrometry*; John Wiley & Sons: Chichester, New York, Weinheim, Brisbane, Singapore, Toronto, 1998.

(24) Boss, C. B.; Fredeen, K. J. *Concepts, Instrumentation, and Techniques in Inductively Coupled Plasma Optical Emission Spectroscopy*, 2nd Edition ed.; Perkin-Elmer Corporation, 1997.

(25) Earle, C. W.; Baker, M. E.; Denton, M. B.; Pomeroy``, R. S. *Trends in Analytical Chemistry* **1993,** *12*, 395.

(26) Falkin, D.; Vosloo, M. *Spec. Europe* **1993,** *5*, 16.

(27) Harnly, J. M.; Fields, R. E. *J. Applied Spectroscopy* **1997,** *51*, 334A.

(28) Botto, R. I. *Spectrochimica Acta* **1983,** *38B*, 129-149.

(29) Boumans, P. W. J. M.; Vrakking, J. J. A. M. *Spectrochimica Acta* **1984,** *39B*, 1239.

(30) Peterson, C. A. PhD thesis, Iowa State University, Ames, 1977.

(31) Larson, G. F.; Fassel, V. A. *J. Applied Spectroscopy* **1979,** *33*, 592.

(32) Boumans, P. W. J. M. *Spectrochimica Acta* **1988,** *43B*, 173.

(33) Harrison, G. R. *M. I. T. Wavelength Tables*; M. I. T. Press, Cambridge, MA, 1969.

(34) Phelps, F. M. *Wavelength Tables, Vol. 2 : Wavelength by Element*; M. I. T. Press, Cambridge, MA, 1982.

(35) Meggers, W. M.; Corliss, C. H.; Scriber, B. F. *Table of Spectral Line Intensities,*; N. B. S. monograph 145, U. S. Government Office, Washington, DC, 1975.

(36) Wohlers, C. C. *ICP Information Newsletter* **1985,** *10*, 593.

(37) Church, S. E. *Geostandards Newsletter* **1981,** *2*, 133.

(38) Parsons, M. L.; Foster, A.; Anderson, D. *An Atlas of Spectral Interferences in ICP Spectroscopy*; Plenum Press, New York, 1980.

(39) Boumans, P. W. J. M. *Line Coincidence Tables for ICP-AES*; Pergamon Press, 1984.

(40) Griffiths, M. L.; Svozil, D.; Worsfold, P. J.; Denham, S.; Evans, E. H. *J. Analytical Atomic Spectroscopy* **2000,** *15*, 697.

(41) Schmidt, G. D.; Slavin, W. *Analytical Chemistry* **1982,** *54*, 2491.

(42) Myers, S. A.; Tracey, D. H. *Spectrochimica Acta* **1983,** *38B*, 1227.

(43) Ramsey, M. H.; Thompson, M. *J. Analytical Chemistry* **1987,** *2*, 497.

(44) Miller, J. C.; Miller, J. N. *Statistics for Analytical Chemistry*, 3rd Edition ed.; Ellis Horwood: Chichester, 1993.

(45) Mendenhall, W.; Sincich, T. *A second course in statistics: regression analysis*, 5th Edition ed.; Prentice-Hall Inc.: Upper Saddle River, New Jersey, 1996.

(46) Beebe, K. R.; Pell, R. J.; Seasholtz, M. B. *Chemometrics, a practical guide*; John Wiley and Sons, 1998.

(47) Thomas, E. V.; Haaland, D. M. *Analytical Chemistry* **1990,** *62*, 1091-1099.

(48) Martens, H.; Naes, T. *Multivariate Calibration*; John Wiley & Sons: Chichester, 1989.

(49) Tabachnick, B. G.; Fidell, L. S. *Using multivariate statistics*; Harper & Row: New York, 1989.

(50) Meglen, R. R. *J. Marine Chemistry* **1992,** *39*, 217.

(51) Wold, S.; Esbensen, K.; Geladi, P. *Chememetrics and Intelligent Laboratory Systems* **1987,** *2*, 37.

(52) Howery, D. G. *Am. Lab.* **1976,** 14.

(53) Haaland, D. M.; Thomas, E. V. *Analytical Chemistry* **1988,** *60*, 1193.

(54) Geladi, P.; Kowalski, B. R. *Analytica Chimica Acta* **1986,** *185*, 1.

(55) Hoskuldsson, A. J. *J. Chemometrics* **1988,** *2*, 211.

(56) Blanco, M.; Coello, J.; Iturriaga, H.; Maspoch, S.; Redon, M.; Riba, J. *Analytica Chimica Acta* **1992,** *259*, 219.

(57) Lehrmann, R.; Burck, J.; Ache, H.-J. *Process Qualilty Control* **1993,** *4*, 139.

(58) Lew, R.; Balke, S. T. *J. Applied Specoscopy* **1993,** *47*, 1747.

(59) Small, G. W.; Arnold, M. A.; Marquardt, L. A. *Analytical Chemistry* **1993,** *65*, 3279.

(60) Haaland, D. M.; Thomas, E. V. *Analytical Chemistry* **1988,** *60*, 1202.

(61) Frank, I. *J. Chemometrics* **1987,** *1*, 233.

(62) Wold, S.; Martens, H.; Wold, H. *Lecture Notes in Mathematics*; Springer Verlag: Heidelberg, 1983.

(63) Jong, S. d. *Chemometrics and Intelligent laboratory Systems* **1993,** *18*, 251.

(64) Peralta-Zamora, P.; Cornejo-ponce, L.; Nagata, N.; Poppi, R. J. *Talanta* **1997,** *44*, 1815.

(65) Luis, M. L.; Garcia, J. M.; Jimenez, F.; Jimenez, A. I.; Arias, J. J. *J. AOAC International* **1999,** *82*, 1054.

(66) Roncevic, S.; Siroki, M. *J. Analytical Atomic Spectrometry* **1994,** *9*, 99.

(67) Luterotti, S.; Sikovec, M.; Bicanic, D. *Talanta* **2000,** *53*, 103.

(68) Heras, L. A. d. l.; Bocci, F.; Betti, M.; Actis-Data, L. O. *Fresnius Journal of Analytical Chemistry* **2000,** *368*, 95.

(69) Lelievre, C.; Hennequin, D.; Lequerler, J. F.; Barillier, D. *J. Atomic Spectroscopy* **2000,** *21*, 23.

(70) Gurleyuk, H.; Tyson, J. F.; Uden, P. C. *Spectrochimica Acta* **2000,** *55B*, 935.

(71) Panayi, A. E.; Spyrou, N. M.; Ubertalli, L. C.; White, M. A.; Part, P. *Biological Trace Element Research* **1999,** *71-2*, 529.

(72) Chetcuti, A. F.; Wong, D. K. Y.; Stuart, M. C. *Analytical Chemistry* **1999,** *71*, 4088.

(73) Daskalova, N.; Boevski, I. *Spectrochimica Acta* **1999,** *54B*, 1099.

(74) Bae, Z. U.; Lee, S. H.; Lee, S. H. *Talanta* **1996,** *44*, 47.

(75) Danzaki, Y.; Takada, K.; Wagatsuma, K. *Fresnius Journal of Analytical Chemistry* **1998,** *361*, 410.

(76) Naes, T.; Isaksson, T.; Kowalski, B. R. *Analytical Chemistry* **1990,** *62*, 664.

(77) Wang, Y.; Veltkamp, D. J.; Kowalski, B. R. *Analytical Chemistry* **1991,** *63*, 2750.

(78) Centner, V.; Massart, D.-L. *Analytical Chemistry* **1996,** *68*, 3851-3858.

(79) Frenich, A. G.; Jouan-Rimbaud, D.; Massart, D. L.; Kuttatharmmakul, S.; Galera, M. M.; Vidal, J. l. M. *Analyst* **1995,** *120*, 2787-2792.

(80) Walmsley, A. D. *Analytica Chimica Acta* **1997,** *354*, 225-232.

(81) Shaw, A. D.; Camillo, A. d.; Vlahov, G.; Jones, A.; Bianchi, G.; Rowland, J.; Kell, D. B. *Analytica Chimica Acta* **1997,** *348*, 357.

(82) Bechmann, I. E.; Sturup, S.; Kristensen, L. V. *Fresnius Journal of Analytical Chemistry* **2000,** *368*, 708.

(83) Glick, M.; Brushwyler, K. R.; Hieftje, G. M. *J. Applied Spectroscopy* **1991,** *45*, 328.

(84) Morales, J. A.; Veen, E. H. v.; Loos-vollebregt, M. T. C. d. *Spectrochimica Acta* **1998,** *53B*, 683-697.

(85) Veen, E. H. v.; Bosch, S.; Loos-Vollebregt, M. T. C. d. *Spectrochimica Acta* **1997,** *52B*, 321-337.

(86) Spiegelman, C. H.; McShane, M.; Goetz, M. J.; Motamedi, M.; Yue, Q. L.; Cote, G. L. *Analytical Chemistry* **1998,** *70*, 35-44.

(87) Westad, F.; Martens, H. *J. Near Infrared Spectroscopy* **2000,** *8*, 117-124.

(88) Brown, C. D.; Vega-montoto, L.; Wentzell, P. D. *J. Applied Spectroscopy* **2000,** *54*, 1055.

(89) Creaser, C. S.; Hutchinson, W. E.; Stephenson, R. G. *J. Applied Spectroscopy* **2000,** *54*, 1624.

(90) Kalivas, J. H.; Kowalski, B. R. *Analytical Chemistry* **1981,** *53*, 2207.

(91) Udelhoven, T.; Schutt, B. *Chemometrics and intelligent laboratory systems* **2000,** *51*, 9.

(92) Baunsgaard, D.; Munck, L.; Norgaard, L. *J. Applied Spectroscopy* **2000,** *54*, 1684.

(93) Huang, J.; Esbensen, K. H. *Chemometrics and Intelligent Laboratory Systems* **2000,** *54*, 1.

(94) Bauer, G.; Rehana, I.; Wegscheider, W. *Spectrochimica Acta* **1988,** *43B*, 971.

(95) Yang, J.; Piao, Z.; Zeng, X. *Spectrochimica Acta* **1991,** *46B*, 953.

(96) Lorber, A.; Goldbart, Z.; Harel, A. *Analytical Chemistry* **1985,** *57*, 2537.

(97) Lorber, A.; Harel, A.; Goldbart, Z.; Brenner, I. B. *Analytical Chemistry* **1987,** *59*, 1260.

(98) Hassan, S. M.; Loux, N. T. *Spectrochimica Acta* **1990,** *45B*, 719.

(99) Veen, E. H. v.; Loos-Vollebregt, M. T. C. d. *Spectrochimica Acta* **1990,** *45B*, 313.

(100) Veen, E. H. v.; Oukes, F. J.; Loos-Vollebregt, M. T. C. d. *Spectrochimica Acta* **1990,** *45B*, 1109.

(101) Veen, E. H. v.; Loos-Vollebregt, M. T. C. d. *Analytical Chemistry* **1991,** *63*, 1441.

(102) McShane, M. J.; Cote, G. L.; Spiegelmen, C. H. *J. Applied Spectroscopy* **1998,** *52,* 878.

(103) Rupprecht, M.; Probst, T. *Analytica Chimica Acta* **1998,** *358,* 205.

(104) Zhang, P.; Littlejohn, D. *Spectrochimica Acta* **1995,** *50B,* 1263.

(105) Depiesse, M.; Biraud, Y.; Lesage, A.; Richou, J. *Quant. Spectrosc. Radiat. Transfer* **1995,** *54,* 539.

(106) Daskalova, N.; Velichkov, S.; Slavova, P. *Spectrochimica Acta* **1996,** *51B,* 733.

(107) Hirokawa, T.; Masukawa, F.; Ito, K.; Shoto, E. *Analytical Sience* **1996,** *12,* 13.

(108) Sadler, D. A.; Littlejohn, D. *J. Analytical Atomic Spectrometry* **1996,** *11,* 1105.

(109) Saxberg, B. E. H.; Kowalski, B. R. *Analytical Chemistry* **1979,** *51,* 1031.

(110) Kalivas, J. H.; Kowalski, B. R. *Analytical Chemistry* **1982,** *54,* 560.

(111) Martens, H.; Jensen, S. A. , Prague 1983; Elsevier; 607.

(112) Mcshane, M. J.; Cameron, B. D.; Cote, G. L.; Spielgelman, C. H. *J. Applied Spectroscopy* **1999,** *53,* 1575-1581.

(113) Sjoblom, J.; Svensson, O.; Josefson, M.; Kullberg, H.; Wold, S. *Chemometrics and Intelligent Laboratory Systems* **1998,** *44,* 229.

(114) Thygesen, L. G.; Lundqvist, S. O. *J. Near Infrared Spectroscopy* **2000,** *8,* 191.

(115) Otto, M. *Chemometrics, Statistics and Computer Application in Analytical Chemistry*; Wiley-VCH: Weinheim, New York, Chichester, Brisbane, Singapore, Toronto, 1999.

(116) Kowalski, B. R. *Chem. & Ind.* **1978,** *Nov.*, 882.

(117) Massart, D. L.; Vandeginste, B. G. M.; Deming, S. N.; Michotte, Y.; Kauman, L. *Chemometrics: a Texbook*; Elsevier: Amsterdam, 1988.

(118) Brereton, R. G. *Analyst* **1987,** *112,* 1635.

(119) Beebe, K. R.; Kowalski, B. R. *Analytical Chemistry* **1987,** *59,* 1007A.

(120) Brereton, R. G. *Analyst* **1997,** *122,* 1521.

(121) Araujo, P. W.; Brereton, R. G. *Trends in Analytical Chemistry* **1996,** *15*.

(122) Taguchi, G. *System of Experimental Design*; Kraus: New York, 1987.

(123) Barnard, T. W.; Crockett, M. I.; Ivaldi, J. C.; Lundberg, P. L. *Analytical Chemistry* **1993,** *65*, 1225.

(124) Barnard, T. W.; Crockett, M. I.; Ivaldi, J. C.; lundberg, P. L.; Yates, D. A.; Levine, P. A.; Sauer, D. J. *Analytical Chemistry* **1993,** *65*, 1231.

(125) Raghu, K. *J. Quality Technology* **1985,** *17*, 4.

(126) Ross, P. J. *Taguchi Techniques for Quality Engineering*; McGraw-Hill: New York, 1978.

(127) Rossi, D. T.; Pardue, H. L. *Analytica Chemica Acta* **1985,** *175*, 153.

(128) Brown, C. W.; Lynch, P. F.; Obremski, R. J.; Lavery, D. S. *Analytical Chemistry* **1982,** *54*, 1472.

(129) Goetz, M. J.; Spielgelman, C. H.; Cote, G.; Motamedi, M. ; Departmant of Statistics, Texas A&M Univeristy, 1995.

(130) Bangalore, A. S.; Schaffer, R. E.; Small, G. W.; Arnold, M. A. *Analytical Chemistry* **1996,** *68*, 4200.

(131) Navarro-Villoslada, F.; perez-Arribas, L. V.; Leon-Gonzalez, M. E.; Polo-Diez, L. M. *Analytica Chimica Acta* **1995,** *313*, 93.

(132) Jouan-Rimbaud, D.; Walczac, B.; Massart, D. L.; Last, I. R.; Prebble, K. A. *Analytica Chimica Acta* **1995,** *304*, 285.

(133) McShane, M. J.; Cote, G. L.; Spielgelman, C. *J. Applied Spectroscopy* **1997,** *51*, 1559.

(134) Lepage, R.; Billard, L. *Exploring the Limits of Bootstrap*; John Wiley & Sons, 1992.

(135) Krzanowski, W. J. *Principles of Multivariate Analysis*; Oxford University Press, 1998.

(136) Shao, J.; Tu, D. *The Jackknife and Bootstrap*; Springer, 1995.

(137) Faber, N. M. *Analytical Chemistry* **2000,** *72*, 4675-4676.

(138) Leardi, R.; Gonzalez, A. L. *Chemometrica and Intelligent Laboratory Systems* **1998,** *41*, 195-207.

(139) Wang, Y. D.; Veltkamp, D. J.; Kowalski, B. R. *Analytical Chemistry* **1991,** *63*, 2750.

(140) Frimpong, A.; Fryer, B. J.; Longerich, H. P.; Chen, Z.; Jackson, S. E. *Analyst* **1995,** *120*, 1675.

(141) Howarth, R. J.; Coles, B. J.; Ramsey, M. H. *Analyst* **2000,** *125*, 2032.

(142) Massart, D. L.; Vandeginste, B. G. M.; Buydens, L. M. C.; jong, S. D.; Lewi, P. J.; Smeyers-Verbeke, J. *Handbook of Chemometrics and Qualimetrics: Part A*; Elsevier: Amsterdam, Lausanne, New York, Oxford, Shannon, Singapore, Tokyo, 1997.

(143) Wise, B. M.; Gallagher, N. B. *PLS_Toolbox 2.0 Manual: for use with Matlab (TM)*; Eigenvector Research: Washington, 1999.

(144) Anderson, T. W. *An introduction to Multivariate Statistical Analysis*: New York, 1984.

(145) Box, G. E. P.; Hunter, W. G. *Statistics for Experimenters*; John Wiley & Sons: New York, 1978.

(146) Kokot, S.; Phuong, T. D. *Analyst* **1999,** *124*, 561.

(147) Perkins, J. H.; Hasenoehrl, E. J.; Griffiths, P. R. *Chemometrics and Intelligent Laboratory Systems* **1992,** *15*, 75.

(148) Geladi, P.; Barring, H.; Dabakk, E.; Trygg, J.; Antti, H.; Wold, S.; Karlberg, B. *J. Near Infrared Spectroscopy* **1999,** *7*, 251.

(149) Anderson, C. E.; Kalivas, J. H. *J. Applied Spectroscopy* **1999,** *53*, 1268.

(150) Despagne, F.; Walczak, B.; Massart, D. L. *J. Applied Spectroscopy* **1998,** *52*, 732.

(151) Bouveresse, E.; Casolino, C.; Massart, D. L. *J. Applied Spectroscopy* **1998,** *52,* 604.

(152) Swierenga, H.; Haanstra, W. G.; Weijer, A. P. d.; Buydens, L. M. C. *J. Applied Spectroscopy* **1998,** *52,* 7.

(153) Wang, Z. Y.; Dean, T.; Kowalski, B. R. *J. Applied Spectroscopy* **1995,** *67,* 2379.

(154) Robinson, J. W. *J. Atomic Spectroscopy*; Marcel Dekker, Inc.: New York & Basel, 1990.

(155) Osborne, B. G.; Fearn, T. J. *J. of Food Technology* **1983,** *18,* 453.

(156) Shenk, J. S.; Westerhaus, M. O.; Templeton, W. C. *Crop Science* **1984,** *25,* 159.

(157) Mark, H.; (Jr.), J. W. *J. Spectroscopy* **1988,** *3,* 28.

(158) Wang, X.; Dean, T.; Kowalski, B. R. *Analytical Chemistry* **1995,** *67,* 2379.

(159) Newton, R. R.; Rudestam, K. E. *Your Statistical Consultant*; Sage Publications: Thousand Oaks, London, New Delhi, 1999.

PUBLICATIONS

Comparison of traditional and multivariate calibration techniques applied to complex matrices using inductively coupled plasma atomic emission spectroscopy

JAAS
Full Paper

Mike L. Griffiths, Daniel Svozil, Paul J. Worsfold, Sue Denham and E. Hywel Evans*

Department of Environmental Sciences, Plymouth Environmental Research Centre, University of Plymouth, Drake Circus, Plymouth, Devon, UK PL4 8AA

Received 28th February 2000, Accepted 20th June 2000
Published on the Web 25th July 2000

Univariate and multivariate (partial least squares 1, PLS1) calibration techniques are compared for the determination of Pt, Pd and Rh in a complex sample matrix. The univariate techniques utilised either pure standards, pure standards with interelement correction factors applied, or matrix matched standards. Univariate calibration yielded relative root mean square errors (RRMSEs) of prediction for Pt, Pd and Rh of 19, 15, and 54% and 13, 18, and 88% for test solutions and real samples, respectively. Univariate calibration with matrix matching proved the most accurate method with RRMSEs for Pt, Pd and Rh of 2.5, 3.7 and 2.4% for a series of synthetic test solutions, and 12, 2.3 and 8.0% for real samples, respectively. In comparison, the multivariate calibration method yielded relative root mean errors for Pt, Pd and Rh of 5.8, 3.0 and 3.5% in the test solutions, and 32, 7.4 and 76% in the real samples. The relative error for the matrix matching and PLS1 techniques was dependent upon analyte concentration.

Introduction

Inductively coupled plasma atomic emission spectrometry (ICP-AES) is now well established as a powerful technique for multielement analysis,[1] but can suffer from both spectral and non-spectral interferences which limit the accuracy, repeatability and reproducibility of the information obtained. The nature of the interference is often complex and it is not always possible to apply the required corrections in order to achieve accurate and precise analysis. In contrast, the information obtained by molecular spectroscopic techniques has been greatly enhanced by the application of data handling tools.

Traditionally, for quantitative analysis in atomic spectroscopy, a single spectral line is chosen, based upon the criteria of line sensitivity and freedom from spectral interferences. Many attempts have been made to correct for spectral interferences in ICP-AES, including standard additions, matrix matching, inter-element correction, and optimisation of line selection. However, these methods suffer from serious limitations when a sample with a complex matrix is presented for analysis. For example, the use of interelement correction factors requires interferent lines that can be used in a univariate fashion to determine the concentration of the interferents themselves, which is not always possible with complex samples.

The use of chemometric approaches to correct for interferences in ICP-AES has emerged as an attractive alternative and various multivariate calibration techniques have been proposed.[2–5] The least complex and most widely available of these assume that there is a linear relationship between response signal and analyte level, such as multiple linear regression (MLR), stepwise multiple linear regression (SMLR), principle components regression (PCR) and partial least squares (PLS).

Atomic emission spectra are well suited to the application of multivariate methods of calibration and analysis because intensity data are recorded at multiple wavelengths.[6] It has been demonstrated that multivariate methods yield better analytical results compared with the more traditional methods in cases where the analyte signal is complicated by spectral line overlap from an interfering species.[7–15] One requirement of multivariate methods, however, is that the factor space defined by the multi-element standards used for model calibration encompasses all likely constituents (analytes and interferents) and concentrations of the real sample matrices. To ensure that this requirement is met, it is necessary to acquire data for the multivariate calibration model using an appropriate experimental design.

In this paper multivariate calibration is compared to traditional univariate calibration with interelement correction and matrix matching for the determination of platinum group metals in autocatalyst digests.

Experimental

Instrumentation and reagents

All data were collected using a simultaneous échelle inductively coupled plasma atomic emission spectrometer (Perkin-Elmer Optima 3000 ICP, Norwalk, USA) equipped with a segmented charge-coupled array detection system.[16,17] Instrumental operating conditions were optimised using simplex optimisation and are given in Table 1.

Single and multielement solutions were prepared by serial dilution of ultra-pure stock standards (10 000 and 1 000 μg ml^{-1}, Johnson Matthey plc, Royston, Hertfordshire, UK). Water was deionised, double distilled (18 MΩ quality) and acids were of Aristar grade (Merck-BDH, Poole, Dorset, UK). All glassware was acid washed in 10% v/v nitric acid for 24 h then rinsed thoroughly with 18 MΩ water. All plasticware was metal-free, high-density polypropylene (Anachem, Luton, Bedfordshire, UK). Calibration and test solutions containing varying concentrations of Pt, Pd, Rh, Al, Mg, Ce, Zr and Ba, plus the internal standards In, Sc and Y were prepared from the stock solutions and stored in high-density polypropylene tubes. Digests of autocatalyst samples (Johnson Matthey plc, Royston, Hertfordshire, UK) which had been independently analysed using an alternative method, and which had been validated by comparison with a NiS fire assay, were used for method validation.

DOI: 10.1039/b001582m

Table 1 Optimised instrumental parameters used for the collection of all data

Carrier gas flow/l min^{-1}	0.93
Auxiliary gas flow/l min^{-1}	0.5
Plasma gas flow/l min^{-1}	16
Viewing height above the load coil/mm	9
Power/W	1286
Spray chamber	Ryton, double-pass
Nebuliser	Seaspray, glass concentric
Resolution	High
Read time/integration time/s	3/0.2

Calibration

Univariate calibration. Univariate calibration was performed by single point calibration using a zero standard containing 10% v/v *aqua regia* (3 : 1 HCl : HNO_3) and a single element calibration solution for each analyte at the concentrations expected in the samples (*i.e.*, Pt 50 μg ml^{-1}, Pd 50 μg ml^{-1}, Rh 10 μg ml^{-1}). A calibration graph was prepared using the net integrated peak areas for the analyte lines (Pt 214.423 nm, Pd 248.892 nm and Rh 343.489 nm) and the concentrations of the analytes in the samples were determined by interpolation.

Interelement correction (IEC). Interelement correction factors for each of the suspected interfering elements were determined. A calibration graph was first constructed in the usual way, and solutions of the suspected interferents (1000 μg ml^{-1}) were aspirated while monitoring the analyte line(s) of interest. The apparent analyte concentration was then determined at the line of interest, and an IEC factor calculated by dividing the apparent analyte concentration at that particular wavelength by the interferent concentration. Eqn. 1 was then applied to obtain the corrected analyte concentration.

$$C_c = C_u - \sum_{i=1}^{i} (F_i \times I_{ic}) \quad (1)$$

Where C_u, C_c, F_i and I_{ic} are the uncorrected analyte concentration, corrected analyte concentration, IEC factor, and interferent concentration respectively for each interfering element, *i*.

Matrix matched calibration. Calibration was carried out using a zero standard containing 10% v/v *aqua regia* (3 : 1 HCl : HNO_3). A multi-element calibration solution containing each element at the concentrations expected in the samples (Table 2) was used also. The internal standards were included in the solutions but were subsequently not used in any of the calibrations. The calibration standard contained both the analyte and matrix elements at the highest concentrations expected in the autocatalyst matrix. The composition of these matrix-matched standards is given in Table 2.

Multivariate calibration. Several multivariate-modelling algorithms were studied, including principal components regression, partial least squares 1 and 2 and multiple linear regression analysis, using Matlab Software Version 5.0, and the PLS_Toolbox 2.0 (Mathworks Inc.). Of these four techniques PLS1 gave the best predictive results and only those are reported here. All data were mean centred, and full-set random cross validation was used throughout this work to minimise systematic error in forming the models.

Cross-validation [the equation for which is given in eqn. (2)] is a measure of the model's ability to predict new samples and consists of leaving out one calibration sample from the calibration set, building the model on the remaining calibration samples, then predicting the value(s) for the left out calibration sample [$\hat{y}$, eqn. (2)] and computing the prediction residuals. The process is repeated until all calibration samples have been left out once; then all prediction residuals are combined to compute the root mean square of cross-validation. To assess the model fit to the calibration data the RMSEC value was used. The format is the same as that used in eqn. (2) except that $\hat{y}$ represents the values of the predicted variables when all calibration samples are left included in the model.

Table 2 Concentrations (μg ml^{-1}) of the standards for the matrix matched calibration

Mixed calibration standard	Zero standard	Calibration standard
Pt[a]	0	40
Pd[a]	0	40
Rh[a]	0	10
Ba[b]	0	25
Ce[b]	0	300
Zr[b]	0	100
In[c]	50	50
Sc[c]	25	25
Y[c]	25	25
Al[b]	1000	1000
Mg[b]	500	500

[a]Analyte. [b]Matrix element. [c]Internal standard.

$$RMSECV = \sqrt{\frac{\sum (\hat{y}_i - y_i)^2}{N}} \quad (2)$$

where y is the known concentration, $\hat{y}$ is the predicted concentration, and N is the number of experiments.

Estimation of errors

The errors of prediction relative to the known values, for individual test solutions and autocatalyst samples, were compared using relative standard error (RSE) values, calculated as shown in eqn. (3):

$$RSE(\%) = 100 \times \frac{(\hat{y}_i - y_i)}{y_i} \quad (3)$$

The overall efficacy of the different calibration methods was compared using the relative root mean square error (RRMSE), defined in eqn. (4), which gives a general idea of the error of prediction for a range of concentrations:

$$RRMSE(\%) = 100 \times \frac{1}{mean(y)} \sqrt{\frac{\sum (\hat{y}_i - y_i)^2}{N}} \quad (4)$$

Experimental design

In a working laboratory it is desirable to maximise the time spent analysing samples compared to the calibration step. Traditionally, multivariate calibration data-sets have been acquired using experimental designs based on a factorial or partial factorial approach. However, for the 8-factor problem studied here, such approaches would result in an impracticably large number of experiments if more than a few levels were used [*e.g.*, 65 536 for a design with 4 concentration levels and 8 factors (4^8)]. Hence, the calibration set for multivariate analysis in this work was prepared using a Taguchi orthogonal array design[18] in order to cover the required factor space with the minimum number of experiments. The concentration ranges of the elements were determined from historical data on the composition of autocatalyst digest samples. The orthogonal array contained 8 factors at 7 levels with a total of 49 experiments, represented as $OA_{49}(7^8)$ (*cf.*, 5 764 801 experiments for a full factorial design). The levels and factors in the design are shown in Table 3. The orthogonality of the array ensured that the effect of every factor assigned to the column in the orthogonal matrix could be estimated while all the

Table 3 Concentration levels (μg ml^{-1}) and factors in the orthogonal array design

	Level						
Factor	1	2	3	4	5	6	7
Pt	0	5	10	20	30	40	50
Pd	0	5	10	20	30	40	50
Rh	0	1	2	4	6	8	10
Ba	0	1	5	10	50	100	200
Ce	0	1	10	50	100	300	500
Zr	0	1	10	50	100	300	500
Mg	0	1	10	50	100	300	500
Al	0	1	10	100	200	500	1000

remaining factor effects were zero. The multivariate models were built using intensity data for 248 lines which included all the of the most intense Pt, Pd and Rh lines and many of the most intense matrix lines.

Results and discussion

Individual RSE values for the test solutions are shown in Fig. 1(a)–(c), and RRMSE values are shown in Table 4. Individual RSE values for the autocatalyst digests are shown in Fig. 2(a)–(c).

Univariate calibration

When univariate calibration was used the RRMSEs of prediction for the test solutions for Pt, Pd and Rh were 19, 15, and 54%, respectively (Table 4). The high errors for Rh were probably due to matrix induced suppression or a combination of matrix-induced suppression and spectral interference. Predicted concentrations of Rh were lower than their actual values in all cases (Table 5). There were no direct

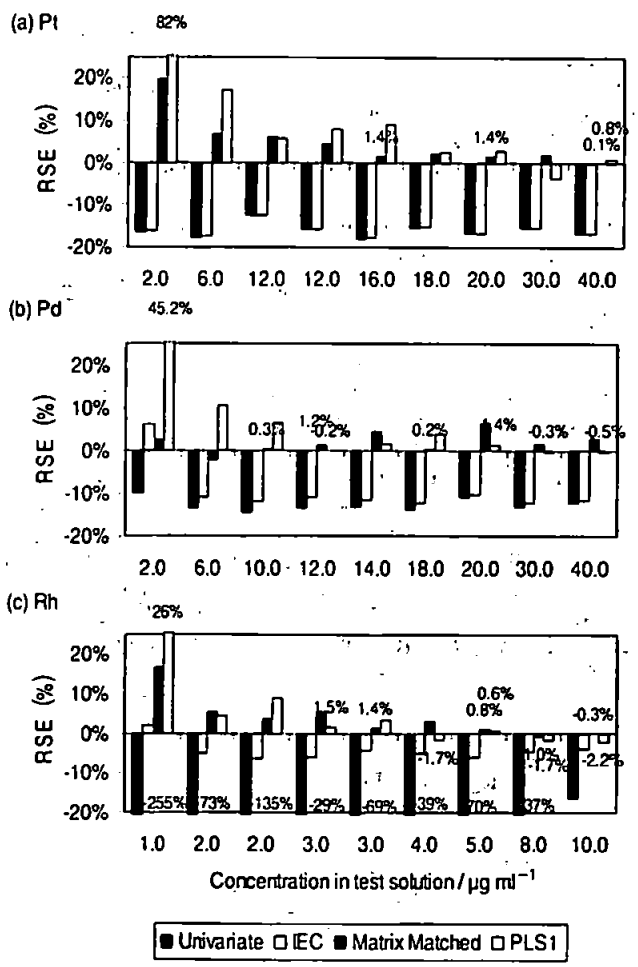


Fig. 1 Comparison of RSE of prediction obtained using univariate, univariate with IEC, matrix matched and PLS1 calibration for: (a) Pt; (b) Pd; (c) Rh; in synthetic test samples.

Table 4 RRMSE of prediction for the concentration of Pt, Pd and Rh in the synthetic test samples and autocatalyst digests

	Test solutions			Autocatalyst digests		
Calibration method	Pt	Pd	Rh	Pt	Pd	Rh
Univariate	19	15	54	13	18	88
Univariate with IEC	19	14	5.7	12	17	2.1
Matrix matched	2.5	3.7	2.4	12	2.4	8.0
PLS1	5.8	3.0	3.5	32	7.5	76

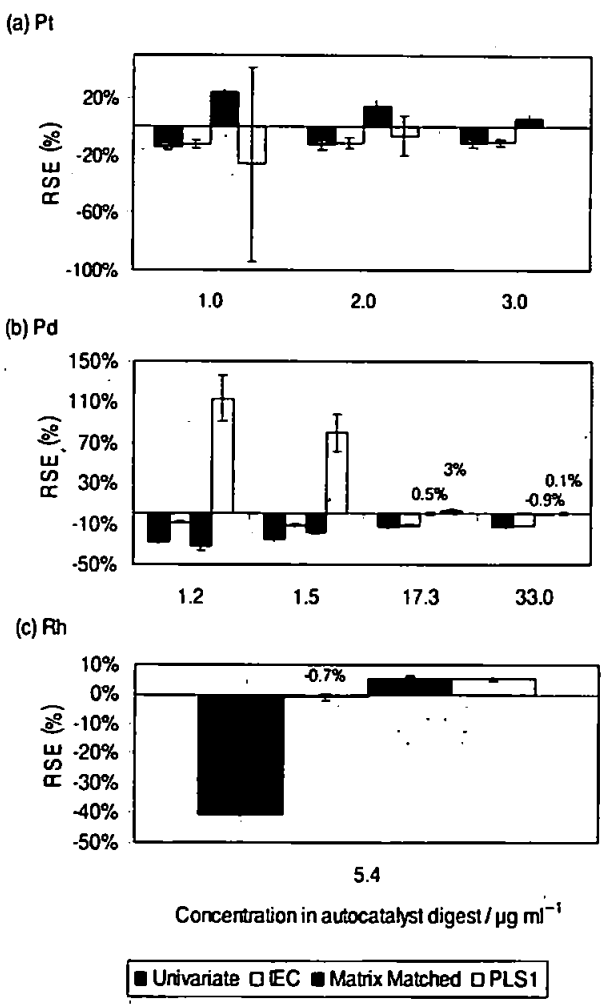


Fig. 2 Comparison of RSE of prediction obtained using univariate, univariate with IEC, matrix matched and PLS1 calibration for: (a) Pt; (b) Pd; (c) Rh; in autocatalyst samples. Each result is the mean of three replicate analyses, and the error bars represents ± one standard deviation.

Table 5 Known and predicted concentrations (μg ml^{-1}) of the test solutions after the application of univariate calibration on net signal intensity obtained using 2-point background correction

	Known			Predicted		
Test solution	Pt 214.423	Pd 248.892	Rh 343.489	Pt 214.423	Pd 248.892	Rh 343.489
Te1	12.00	20.00	3.00	10.48	17.82	2.14
Te2	16.00	12.00	5.00	13.13	10.38	1.52
Te3	20.00	18.00	2.00	16.66	15.49	−0.69
Te4	12.00	14.00	4.00	10.10	12.16	2.45
Te5	18.00	10.00	3.00	15.24	8.54	0.93
Te6	6.00	30.00	1.00	4.94	26.04	−1.55
Te7	2.00	6.00	2.00	1.67	5.18	0.55
Te8	40.00	2.00	8.00	33.32	1.80	5.03
Te9	30.00	40.00	10.00	25.41	35.10	8.34
Te10	0.00	0.00	0.00	−0.01	−0.19	−1.33

Table 6 Magnitude of the IEC factors used to correct for spectroscopic interferences

	IEC factor at analyte line/nm		
Interfering element	Pt 214.423	Pd 248.892	Rh 343.489
Pt	—	9.86×10^{-3}	-2.23×10^{-6}
Pd	3.67×10^{-4}	—	5.45×10^{-5}
Rh	-7.15×10^{-5}	-6.52×10^{-4}	—
Al	-1.47×10^{-5}	-8.32×10^{-5}	3.18×10^{-5}
Mg	6.38×10^{-6}	-3.66×10^{-4}	2.59×10^{-5}
Ce	1.08×10^{-5}	-2.3×10^{-4}	-6.68×10^{-3}
Zr	-4.16×10^{-5}	-8.75×10^{-5}	8.03×10^{-6}
Ba	-3.26×10^{-5}	-1.37×10^{-4}	1.67×10^{-5}
In	-1.16×10^{-5}	-7.23×10^{-5}	-5.53×10^{-6}
Sc	-6.78×10^{-5}	-5.64×10^{-4}	4.75×10^{-4}
Y	-4.22×10^{-6}	-5.06×10^{-5}	-1.33×10^{-3}
Sr	-2.95×10^{-4}	-1.08×10^{-4}	1.02×10^{-5}

spectral overlap interferences on the Rh 343.489 nm line which would cause this high error but there was a baseline shift due to the presence of Zr [Fig. 3(c)]. Net intensity data were calculated by baseline subtraction, using background correction points (BCPs) positioned either side of the line, so the baseline shift should be accounted for; however, the positioning of the BCPs for Rh 343.489 was hampered by the proximity of the Ce 343.521 nm line [Fig. 3(c)], resulting in failure of the background correction routine.

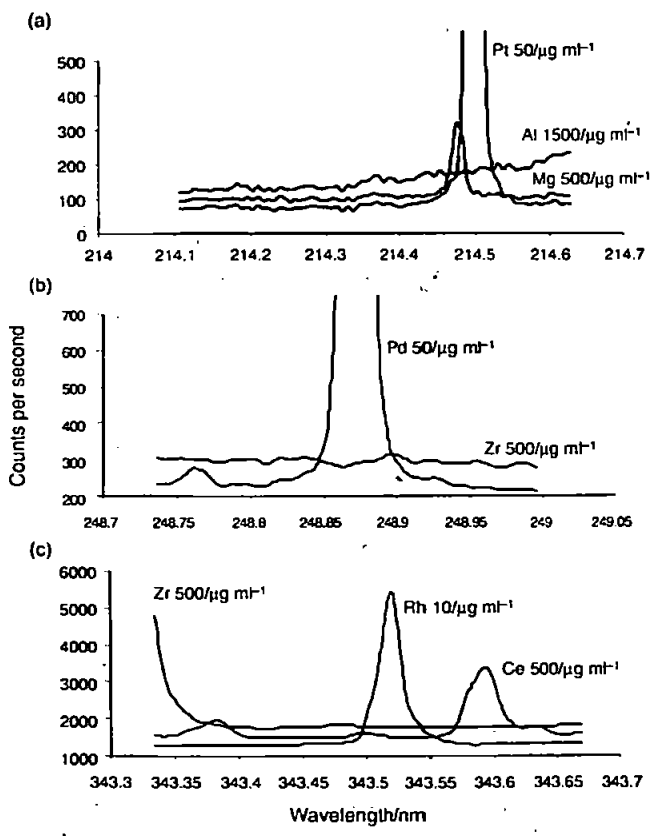


Fig. 3 Interferences on the: (a) Pt 214.423 nm; (b) Pd 248.892 nm; (c) Rh 343.489 nm lines.

Predicted concentrations for Pt and Pd were much closer to the known values and there was no trend in the relative error with concentration [Fig. 1(a) and (b)], despite the presence of some small spectral interference and baseline shift [Fig. 3(a) and 3(b)]. For example, even at a relatively high Mg concentration of 500 μg ml^{-1} the spectral interference was small relative to the peak for Pt [Fig. 3(a)], and the baseline shift caused by 500 μg ml^{-1} of Zr was relatively minor at the Pd 248.892 line [Fig. 3(b)].

The RSEs for the prediction of Pt, Pd and Rh in the autocatalyst samples are shown in Fig. 2(a)–(c). Each point represents the average of three replicate sample digests corrected for mass of sample. Concentrations of Pt, Pd and Rh were predicted to be lower than their actual values. The RRMSEs for Pt, Pd and Rh were 13, 18 and 88%, respectively, again confirming that Rh was not predicted as well as the other analytes.

Interelement correction (IEC)

The magnitude of the IEC factors varied considerably from -5.5×10^{-6} to 9.9×10^{-3} (Table 6). For the prediction of

Table 7 RMSEC and RMSECV values for the PLS1 model calibration data set constructed using gross mean centred data

	RMSEC	RMSECV
Pt	0.77	1.05
Pd	0.40	0.48
Rh	0.12	0.24

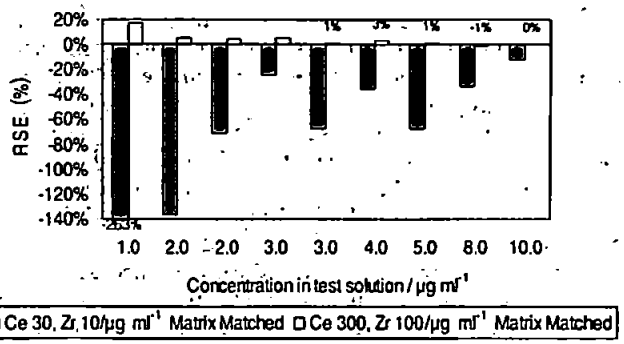


Fig. 4 Effect on the RRSE for Rh in the test samples when matrix matching is applied with different concentrations of the main interferents (Ce and Zr) on the line chosen (Rh 343.489 nm).

analyte concentrations in the synthetic test solutions the application of IEC generally resulted in an improvement in the accuracy of prediction. The greatest improvement was observed for Rh, with the RRMSE falling from 54 to to 5.7% when IEC was applied (Table 4). The RRMSE values for Pt and Pd did not change significantly with values of 19 and 14%, respectively, after the application of IEC (Table 4), which is also reflected in the magnitude of the RSE values for the individual solutions [Fig. 1(a) and (b)]. Results for the autocatalyst samples followed the same trend where the RRMSEs for Pt, Pd and Rh were 12, 17 and 2.1%, respectively (Table 4). Hence, IEC had a significant effect only for the correction of spectral interferences on Rh in this instance.

Matrix matched calibration

Results obtained using a matrix matched standard are shown in Fig. 1(a)–(c) and Fig. 2(a)–(c) for the test solutions and autocatalyst samples, respectively. Overall, this was the most accurate of the univariate calibration methods for the prediction of Pt and Pd in the test solutions, and was comparable to, or better than, IEC for Rh for all but the lowest concentration (Fig. 1). This overall improvement in accuracy of prediction for the test solutions is reflected in RRMSE values of 2.5, 3.7 and 2.4% for Pt, Pd and Rh, respectively (Table 4).

For the autocatalyst samples, the RRMSEs for Pt, Pd and Rh were 12, 2.3 and 8.0% respectively (Table 4). Evidently, the accuracy of prediction for Pt and Pd in the autocatalyst samples was not as good as for the test solutions, which was probably due to the concentrations of the matrix matched elements not being exactly the same as their concentrations in the autocatalyst samples. Matrix-induced suppression of the analyte signal will only be effectively corrected for by matrix matching the standards to the samples. The relatively low RRMSEs for all three analytes, achieved when matrix matching was used, suggests that matrix-induced suppression had a greater effect on the accuracy of the results than spectral interferences at the lines chosen for analysis. The main disadvantage of matrix matching is that the matrix matched standards contain matrix elements at fixed concentrations, whereas the concentrations in the samples may vary quite considerably. This is demonstrated in Fig. 4, which shows the RSEs obtained for Rh in the individual test solutions when the concentrations of Ce and Zr were reduced by a factor of ten, to 30 and 10 μg ml^{-1}, respectively. As can be seen, the RSEs increased considerably at all concentrations when incorrect matrix matching was employed, and a comparison of Figs. 4 and 1(c) reveals that the results were very similar to those obtained when non-matrix matched univariate calibration was used.

Multivariate calibration: partial least squares 1 (PLS1)

Multivariate calibration was initially performed using net signal intensities. In order to obtain the net signal from the instrument, a 2-point background correction was performed by the instrument software, however, the selection of optimum background correction points for all 248 lines proved problematical due to the presence of adjacent spectroscopic interferences. Hence, it was decided to model the data using gross as well as net signal intensities. All data were translated along the co-ordinate origin by mean-centering each variable. Autoscaling was also tried but gave much worse results, probably due to the noise being scaled equally with the

Table 8 Known and predicted concentrations (μg ml^{-1}) of Pt, Pd and Rh in the synthetic test solutions (Te) and autocatalyst samples, predicted using PLS1

	Pt		Pd		Rh	
Sample	Known	Predicted	Known	Predicted	Known	Predicted
Te1	12.00	12.71	20.00	20.27	3.00	3.04
Te2	16.00	17.45	12.00	11.98	5.00	5.03
Te3	20.00	20.55	18.00	18.68	2.00	2.18
Te4	12.00	12.95	14.00	14.21	4.00	3.93
Te5	18.00	18.43	10.00	10.64	3.00	3.10
Te6	6.00	7.04	30.00	29.90	1.00	1.26
Te7	2.00	3.64	6.00	6.64	2.00	2.09
Te8	40.00	40.33	2.00	2.90	8.00	7.86
Te9	30.00	28.93	40.00	39.81	10.00	9.78
Te10	n/d	2.20	n/d	0.44	n/k[a]	0.36
S1 R 1	1.01	1.26	1.22	2.46	n/k[a]	−0.11
S1 R 2	1.02	−0.03	1.19	2.85	n/k[a]	0.14
S1 R 3	0.99	0.98	1.22	2.45	n/k[a]	0.10
S2 R 1	1.99	1.55	1.51	2.77	n/k[a]	0.30
S2 R 2	1.88	1.97	1.48	2.36	n/k[a]	0.10
S2 R 3	1.99	1.97	1.49	2.93	n/k[a]	−0.34
S4 R 1	n/d	1.06	17.62	18.19	5.36	5.63
S4 R 2	n/d	1.11	17.86	17.92	5.36	5.62
S4 R 3	n/d	1.36	17.36	17.59	5.33	5.65
S5 R 1	n/d	1.47	33.70	32.63	0.05	2.83
S5 R 2	n/d	0.25	34.13	33.84	0.03	2.93
S5 R 3	n/d	0.63	33.73	32.70	0.03	2.95

[a]n/k = not known.

informative data. Results obtained using net signal intensity data were worse than those obtained when using gross signal data, probably due to the incorrect assignment of BCPs, so only results obtained using the gross signal data are presented here.

Results for the model validation are shown in Table 7, indicating that the model for Rh had the lowest error (*i.e.*, lowest RMSECV and RMSEC values). The RMSEs values for Pt, Pd and Rh in the test solutions were 5.8, 3.0 and 3.5%, respectively (Table 4), indicating that, overall, this calibration strategy was as good as matrix matching. The individual results and RSEs for the synthetic test solutions are shown in Table 8 and Fig. 1, respectively. In the majority of cases, the RSEs for the analytes after multivariate treatment were lower than the corresponding values when univariate and interelement correction was applied. It is evident from Fig. 1 that the predictive accuracy of the PLS1 model was highly dependent on analyte concentration. For example, the RSE for Pd in the test solutions decreased from 45% for 2 μg ml^{-1} to 11% for the 6 μg ml^{-1} solution, and to only −0.5% for 40 μg ml^{-1}. This pattern was repeated for the autocatalyst samples, with the RSE for Pd changing from 114% at *ca.* 1.2 μg ml^{-1} to 0.1% at *ca.* 33 μg ml^{-1} (Fig. 2).

For the autocatalyst samples the RRMSEs for Pt, Pd and Rh were 32, 7.5 and 76%. The relative failure of the model at low concentrations can be partly explained by the fact that the lowest concentrations used in the multivariate calibration data set were 5, 5 and 1 μg ml^{-1} for Pt, Pd and Rh, respectively, so the lower end of the concentration range was not modelled sufficiently well to enable accurate prediction of analyte concentration below these concentrations. This hypothesis is lent credibility by the fact that nearly all of the synthetic test solutions were predicted with lower RSE error values than the autocatalyst samples, indicating that interfering elements may have been present in the autocatalyst samples which were not included in the model. This would have a relatively greater influence on the accuracy of prediction at lower concentrations. One possible solution to this would be to model the lower concentration range, where noise is likely to have a greater influence, separately from the higher concentration range.

Conclusions

A number of calibration methods have been compared for the simultaneous determination of Pt, Pd and Rh in test solutions containing a synthetic matrix and autocatalyst samples containing varying concentrations of these analytes. Traditional calibration showed that for the elements in the test samples the predominant interferences on the Pt 214.423 nm, Pd 248.892 nm and Rh 343.489 nm lines were caused by matrix induced suppression and spectroscopic interference, or a combination of both. Several calibration methods were compared, with the best being matrix matching and multivariate calibration using PLS1. Matrix matching failed when the standards were not matched correctly, which will often be the case with variable matrices, and PLS1 yielded good results at high concentrations but was less effective at low concentrations due to the noise contribution. Further work is proceeding to improve the accuracy of multivariate calibration at low concentrations.

References

1 A. Montaser and D. W. Golightly, *Inductively Coupled Plasma on Analytical Atomic Spectrometry*, VCH, New York, 1992, 2nd. edn.
2 J. C. Ivaldi, D. Tracy, T. W. Barnard and W. Slavin, *Spectrochim. Acta, Part B*, 1992, **47B**, 12.
3 G. Bauer and W. Wegscheider, *Spectrochim. Acta, Part B*, 1992, **47B**, 1.
4 P. Zang, D. Littlejohn and P. Neal, *Spectrochim. Acta, Part B*, 1993, **48B**, 1517.
5 D. A. Sadler and D. Littlejohn, *J. Anal. At. Spectrom.*, 1996, **11**, 1105.
6 K. R. Beebe and B. R. Kowalski, *Anal. Chem.*, 1987, **59**, 1007A.
7 E. H. van Veen and M. T. C. de Loos-Vollebregt, *Anal. Chem.*, 1991, **63**, 1441.
8 E. H. van Veen, F. J. Oukes and M. T. C. de Loss-Vollebregt, *Spectrochim Acta, Part B*, 1990, **45B**, 1109.
9 E. H. van Veen and M. T. C. de Loos-Vollebregt, *Spectrochim. Acta, Part B*, 1990, **45B**, 313.
10 G. Bauer, W. Wegscheider and H. M Ortner, *Fresnius' Z. Anal. Chem.*, 1990, **340**, 135.
11 G. Bauer, W. Wegscheider and H. M Ortner, *Spectrochim. Acta, Part B*, 1991, **46B**, 1185.
12 A. Lorber, A. Harel, Z. Goldbart and I. B. Brenner, *Anal. Chem.*, 1987, **59**, 1260.
13 D. F. Wirsz and M. W. Blades, *Anal. Chem.*, 1986, **58**, 51.
14 P. Taylor and P. Schutyser, *Spectrochim. Acta, Part B*, 1986, **41B**, 81.
15 M. Glick, K. R. Brushwyler and G. M. Heiftje, *Appl. Spectrosc.*, 1991, **45**, 328.
16 T. W. Barnard, M. I. Crockett, J. C. Ivaldi and P. L. Lundberg, *Anal. Chem.*, 1993, **65**, 1225.
17 T. W. Barnard, M. I. Crockett, J. C. Ivaldi, P. L. Lundberg, D. A. Yates, P. A. Levine and D. J. Sauer, *Anal. Chem.*, 1993, **65**, 1231.
18 K. Raghu, *J. Qual. Technol.*, 1985, **17**, 4.