The University of Hull

The Development and Application of Chemometrics to Process Analysis in an Industrial Environment

Being a Thesis Submitted for the Degree of

Doctor of Philosophy

In the University of Hull

By

James Moffatt, BSc.

January 1999

Acknowledgements

I would like to thank Dr. Tony Walmsley for all the continual assistance to me throughout the development of this project and thesis, without his guidance there would be no work to report. I would also like to thank Dr. Tony Walmsley and Dr. S.J. Haswell for their assistance and friendship during all the aspects of academic and non academic life that I encountered during my time at the university of Hull.

This work was sponsored throughout by Smith & Nephew Hull, who I would like to thank for their support, especially Barry Cunningham without whom this project would never have been started, and Richard Wilkins, Tom Allan and David Lloyd Jones for their continual support through the three years. Charlie Shaw and Bill Mortimer have provided constant advice and information for which I am very grateful.

Lorna Nelstrop needs special thanks for all the help and support she has given me. Also Steve Todd, Gareth Owen and Sue Neville for putting up with me over the years.

Finally I would like to thank all the members of the chemistry department who have helped me in my time at Hull, in whatever way, and without whom I would not have completed this project.

Table of Contents

Abstract

This thesis describes two main sections of work, an examination of a commercial product, Intrasite Gel, and the development of an algorithm for variable selection using projected latent structures.

Following on from the successful development of a variable selection procedure for multivariate linear regression this work looks at transferring this idea for use with projected latent structures. The first part of this thesis will show how the variable selection algorithm was developed and used with three different data sets. The algorithm will be shown to be superior to standard projected latent structures, for linear multi-component data. Although the final algorithm developed requires considerable computing resources to carry out this is compensated for by significantly improved model predictions and robustness. The final algorithm developed is written to run using MATLAB ® on any computer platform that supports this application, though the principles of operation could be transferred to another method of execution, for example custom code written in C or Pascal. The approach used in the development of this method is that the ability of the model to predict unknown samples is of far greater importance than the internal performance of the model. All the assessments of the procedures developed are based on the ability of the model to predict accurately and precisely samples that were not presented to the model during the training stage.

The second section of this thesis is concerned with the study of Intrasite Gel, produced by Smith & Nephew Ltd. Hull. The material in question is a medical device intended to assist in the treatment and healing of wounds that are necrotic, sloughy or

granulating. The product is characterised by its ability to maintain moisture equilibrium in a wound environment and to provide a suitable medium to encourage the growth of new cell tissue. Medical devices require registration, and as part of that registration a number of tests are made on samples to ensure that the material meets the required specifications. There was some concern at Smith & Nephew that the tests they were required to carry out as part of the device registration were not providing appropriate information about the product. Of particular interest was the fluid absorption property as it was suspected that the test has a large amount of random error associated with it and an investigation was required to examine this test and to provide an alternative procedure should the fluid absorption test prove inadequate. Also of interest to Smith & Nephew was the issue of sampling frequency, as it was felt that this should also be examined to determine whether the correct rate of sampling to ensure product quality was being carried out. The work reported here shows that the fluid absorption test as it stands is insufficient to the task of monitoring this property of Intrasite gel and that an alternative test should be considered. This work also showed that current sampling rate was too high and that the high sampling rate may in fact cause misleading assumptions as to the stability and quality of the product.

3Glossary of Terms

Terminology

ANOVA

ANalysis Of VAriance, a standard test to examine the influences of variance within a data set.

Autocorrelation

The internal correlation between samples within a variable, either as a time function or a space function.

Autoscaling

Setting the mean and standard deviation of a matrix to zero and one respectively. This removes the effect of magnitude in a system, and reduces the influence of noise between variables.

Calibration

The determination of the relationship between two (or more) data matrices, normally called the independent matrix (X-Block) and the dependent matrix (Y-Block)

CLS

Conditional Least Squares, a variation on MLR where the coefficients are required to meet certain properties.

Cluster Analysis

Examination of the grouping or class of a group of objects

Chi-squared test

From a given mean and standard deviation the chi-squared test can be used to determine the normal expected distribution for that population, which can the be compared with the observed distribution.

Collinearity

Collinearity is a linear or nearly linear relationship between variables within a independent data matrix. Collinearity causes problems with some methods of inverting a matrix, and reduced the predictive ability of a callibration

Correlation

A quantitative term describing the linearity between two variables

Correlation Coefficient

The correlation scaled between –1 and +1, +1 indicating a strong positive relationship, zero, no relationship, and –1 indicating a strong negative relationship

CUSUM

The CUmulative SUMation of a vector. A method for examining the way in which the mean of a variable changes over time

Data Set

Term used to describe the data that relates to a particular problem, a data set can be more than one data matrix

Dependent Data Matrix

The response variable or variables, for spectral information the dependent data matrix would be the component information. The response data can be quantitative or qualitative, with qualitative information the calibration carried out is for classification.

Dixons Q Test

This is a test for outlying values, the value calculated for the test is compared to a table of values to determine whether the specified point is outlying. This method is mostly used for small vectors.

Eigenvalue

When decomposing a matrix into two other matrices with the constraint to capture maximal variance in consecutive vectors the eigenvalue shows how much variance is captured by the corresponding eigenvector.

Eigenvector

An eigenvector is the vector of coefficients that rotate a data matrix onto the axis that form the principal components.

GLS

Generalised Least Squares, a variation on MLR that deals with heteroscedastic

residuals.

Grubbs Test

Grubbs test detects outliers by their effect on the standard deviation of a group of samples.

Heteroscedastic

Heteroscedastic residuals are ones in which the error is not normally distributed across the span of the data space.

Homoscedastic

Homoscedastic residuals occur when the error in a model is normally distributed.

Independent Data Matrix

The independent data matrix is the matrix of descriptive data pertaining to a system, in spectroscopy the independent matrix would normally be the matrix containing the spectra.

ITTFA

Iterative Target Testing Factor Analysis is a method to extract real world information from a matrix of data, for example with UV data ITTFA can be used to extract the molar extinction coefficients for the pure components.

Kalman Filter

Factor analysis method for removing noise from a signal

KNN

K-Nearest Neighbour, a classification technique that assigns a class to a sample based on its relationship to similar samples.

LWR

Locally Weighted Regression is a linear regression method that can be used to model non-linear data by examining the curve in short segments where the assumption can be made that a sufficiently short curve behaves as a line.

Mean Centring

A method for removing the influence of magnitude from a variable by subtracting the mean of the vector from each point in the vector.

MLR

Multivariable Linear Regression, a least squares method for determining the coefficients that relate an independent data matrix to a dependent data matrix.

NIPALS

Non-Iterative Partial Least Squares, a method for calculating PLS

NLS

Non-linear Least Squares, a variation on MLR that used a non-linear function to map the X-Block to the Y-Block

PCA

Principal Components Analysis is used to decompose a matrix into two other matrices with the constraint that the vectors produced describe maximal variance of the original matrix.

PCR

Once a matrix has been decomposed using PCA the resultant vectors can be regressed against a response variable to form a Principal Components Regression model.

PEP

Percentage Error of Prediction, a method of comparing models developed using different methods, or form different data.

PLS

Projected Latent Structures, also known as Partial Least Squares, a factor analysis method that extracts new vectors on the basis of their correlation with a target vector and is used to build calibration models.

PRESS

Predicted Residual Error Sum of Squares, a method of determining the predictive ability of a model, normally used where small differences in models are being examined for the same data set. PRESS cannot be used to compare different components of a data set.

Range Scaling

A method of reducing the effect of magnitude on a data matrix, the matrix is divided by the largest absolute value in the matrix

SIMCA

Soft Independent Modelling Class Analogy models the classification of samples by considering groups of samples as independent models, assigning a class to a sample according to the model which it best fits. A sample can be assigned to more than one class.

Smoothing

Any method with is used to reduce the effect of randomly distributed noise within a vector. This includes moving average smoothing and Savitzy Golay smoothing.

SNV

Standard Normal Variate is autoscaling carried out by sample rather than variable.

SVD

Single Value decomposition is a non-iterative method for extracting eigenvalues and eigenvectors from a matrix.

TFA

Target Factor Analysis can be used to detect the presence of signals within a more complex signal, for example TFA can be used to detect the presence of a particular metal in a UV spectrum based on the pure component spectrum of that metal.

WLS

Weighted Least Squares, a variation on MLR that can try and account for heteroscedastic residuals.

X-Block

The independent data matrix.

Y-Block

The dependent data matrix.

1. Introduction

1.1. Chemometrics

Chemometrics can be seen as the use or study of mathematics and its use in chemical systems. Many of the techniques are little different from those found in standard statistics or Biometrics, others such as variable selection techniques are associated mainly with the chemistry side of statistics. This thesis considers closely variable selection techniques, as can be found in Walmsley [1] and Walmsley *et. al.* [2].

A definition of chemometrics taken from Chemometrics: A Textbook [3] "*The chemical discipline that uses mathematical, statistical, and other methods employing formal logic (a) to design or select optimal measurement procedures and experiments, and (b) to provide maximum relevant chemical information by analysing chemical data.*".

Another definition is that by Malinowski [4], "*The use of mathematical and statistical methods for handling, interpreting and predicting chemical data.* Yet a third by Svant Wold [5], *chemometrics is the art of extracting chemically relevant information from data produced in chemical experiments.*".

These definitions cover most of chemometrics, its use in both experimental design and in data analysis. Much of the work in chemometrics has taken place around techniques used in a laboratory, methods to examine the results of experiments on a small scale. These include regression and calibration based upon spectra results, and work towards the optimisation of experiments. There is, however, a large area of

chemometrics devoted to process analysis. This looks at a chemical process as a whole, relating the conditions of the process and its input to the properties and qualities of the outputs.

The use of chemometric techniques in everyday chemistry is becoming increasingly prevalent, the range of problems to which chemometric techniques can be successfully applied is increasing rapidly. Largely this can be put down to an evolution in the use of chemometrics, a sort of natural selection takes place, i.e. techniques that provide robust reproducible and useful results proliferate, while less robust or poorly defined techniques become neglected. It takes considerable time for theoretical work to be converted into practical applications in any discipline and this is no different in chemistry. Many areas can be quite conservative, which is because after applying and developing new techniques the methods can be very expensive, as well was time consuming. As in any field, a few people will champion new ideas, as they receive the benefits, more researchers will begin to use the techniques and true growth will begin.

Much of the needed fundamental work has been done, current methods have been shown to be successful and applicable. Chemometrics is doing well in shedding its roots, it began life as a few obscure statistical tools useful to psychologists in the early twentieth century, much of this work was published by people like Hotelling [6, 7, 8, 9], Bartlett [10, 11] and Thurstone [12] (though many of the mathematical premises date from the nineteen century). These roots are apparent even in many of the more modern and respected works on the subject for example the one by Malinowski, Factor Analysis in Chemistry, Wiley, 1992 [4].

While a proper definition of chemometrics shows its roots in all the uses of statistics with chemistry, this work will consider two distinct sections. The first section of techniques considered are those to do with calibration and regression. While this includes ordinary linear regression and multivariate linear regression, together with non-linear variations, the section will be considered as factor analysis techniques, though none of these two techniques properly belongs to this category. A definition of factor analysis can be obtained from Malinowski [4], *"Factor analysis is a multivariate technique for reducing matrices of data to their lowest dimensionality by use of orthogonal factor space and transformations that yield predictions and / or recognisable factors"*. Malinowski, in this definition, considered factor analysis to be a single technique however many methods can be considered factor analysis.

The second section considered is that of process analysis, and process control. These techniques look at data from mostly large-scale processes. The analysis of process analysis data can make use of factor analysis techniques, however it is mainly concerned with techniques such as ANOVA, CUSUM, t-tests, F-tests and Shewhart charts, among others.

Chemometric techniques are widely applied to a variety of problems, including comparison of methods, experimental design, calibration and modelling, outlier detection and class separation.

1.2. History of Chemometrics

The techniques that form the core of chemometrics today (factor analysis techniques) did not start in the chemical field. The first few steps towards modern factor analysis techniques were slow and took many years of development and refinement.

Factor analysis techniques can be traced back to 1901, and the paper by Pearson [13]. Pearson's paper is not the first work to examine the axis of an ellipsoid but his work was the first to describe the lines and planes of best fit through such a system. Also, unlike any earlier work his method did not assume two or three dimensions, but was equally applicable to multi-dimensional space. Pearson did not describe lines or planes other than the principal one. It was left to Hotelling in 1933 [6] to provide the necessary rigorous definition of procedure to extract the principal axes and those axes that successively describe information within the data space. Pearson also assumed that any such data space would be ellipsoid, ignoring the possibility of non-symmetrical data. This flaw was quite serious, and one consequence of this was that the order in which the rows and columns of a data set were presented to the algorithm affected the results produced. This problem seriously affected how this technique was received, and it also followed through to much of the work over the next few years. Many of the papers produced between 1901 and 1954 were to do with this problem, various authors argued over the merits of their various techniques. L.L. Thurstone and his son T.G. Thurstone argued for many years in print about each other's variations, though they also collaborated on papers as well [12]. Although they both did much to advance both the techniques of factor analysis and the coverage that those techniques received by chemists, they also damaged the view many held of these techniques. Most people of this time held that the univariate techniques of the time were superior to the new factor analysis methods. Fisher and MacKenzie in 1923 [14] argued that PCA was sometimes a superior method to ANOVA for examining the causes of variation within a data set. In the same paper they also proposed a modification to principal components that was to form the basis for Projected Latent

Structures. In the most part these techniques were ignored by chemists, and the method that would become PLS proposed by Fisher and MacKenzie was also ignored.

Psychologists in the '20's and '30's provided most of the work in factor analysis of that time. Psychological research of that period was concerned with the underlying properties of intelligence, which they referred to as factors, and they looked at techniques that extracted these factors.

H. Harmon [15] was one of the first chemists to consider the applications of factor analysis to chemistry, publishing in the 1960's, though others such as Higman published at the same time. In 1964 C. Radhakrishna Rao published a review of factor analysis used in chemistry [16] , he suggested that the tide of public opinion towards factor analysis had turned in its favour. He also showed that many of the methods were clearly superior to the univariate methods otherwise used. The problems due to inconsistencies in calculations mentioned above were mostly resolved by this point. The mainstream maths of such routines as PCR, PCA, cluster analysis and related techniques were all well documented, understood and respected. PLS was fully developed by H. Wold in 1964 [17], from the paper in 1923 by Fisher and MacKenzie [14], and then further modified by his son, S. Wold [18]. Excellent papers in the subject was later written by Paul Geladi and Bruce Kowalski [19] (1986) and S. Wold again in 1989 [20].

Much of current work is in developing variations or complimentary processes rather than entirely new methods, and in determining the optimum chemometric approach to use for the mathematical analysis of data, [21, 22, 23].

1.2.1. Application of Chemometrics

Chemometrics has a very wide application, and grouping different techniques can be problematic. Possibly the easiest grouping, and also one of the more useful is that into supervised and unsupervised methods.

1.2.1.1. Unsupervised Methods

Unsupervised methods are those where there is only an X-block data set, i.e. no calibration / quantification information is provided. Such a data set might be from the spectroscopic analysis of a group of petrol samples, where only the samples have been provided, and no qualitative or quantitative information about the petrol being examined. Unsupervised methods can be used here to remove noise, provide smoothing, and examine the relationships between the different petrol samples, which might come from several different manufacturers or be of differing octane ratings or grades. Unsupervised methods are often used for exploratory data analysis, where the relationships between the rows or between the columns of the data set can provide useful information.

1.2.1.2. Supervised Methods

Supervised methods are used where both an X-block and a Y-block data set are provided, i.e. the spectra of the petrol samples could be provided together with the octane rating, the manufacturer and the concentration of some of the additives. Supervised methods can be used to group the petrols by manufacturer, or by grade, or octane rating, and they can be used to determine the relationship between the spectral information and for example the octane rating of the petrol. This would enable a new petrol sample to be analysed by the same method, and its octane rating calculated

rather than tested. If the method used to examine the samples was a cluster analysis method the new petrol sample could be assigned to a manufacturer, or grade type.

1.2.1.3. Spectroscopy

Chemometrics can be used for calibration in most spectroscopic methods [24]. Linear Regression, Multivariate Linear Regression, Principal Components Regression, and Projected Latent Structures have all been used for direct calibration / prediction of spectroscopic results. Spectroscopic methods include near infrared spectroscopy (NIR), ultraviolet spectroscopy (UV), ultraviolet-visible spectroscopy (UV-Vis), infrared spectroscopy (IR), diffuse reflectance infra-red Fourier transform spectroscopy (DRIFTS). These techniques all examine the relationship between a matrix of independent data (X-block or spectral information) and a matrix of dependent data (Y-block or concentration information). These terms will be used interchangeably depending on circumstances. These methods will be examined in detail later in the chapter. These methods are not exclusive to spectroscopic data, and further applications will be discussed.

Also used with spectroscopic data, though again not exclusively, are other techniques,

1.2.1.3.1. Kalman filter

The Kalman filter was developed in the field of electronics by Rudolf Emil Kalman, as a useful method of removing noise from signals. The Kalman filter is an iterative least squares estimator that can be used to determine the correct linear response in a system perturbed by Gaussian noise [25]. This is useful in spectroscopy to remove

noise from spectra and to compensate for the effects of drift. The Kalman filter is a supervised method.

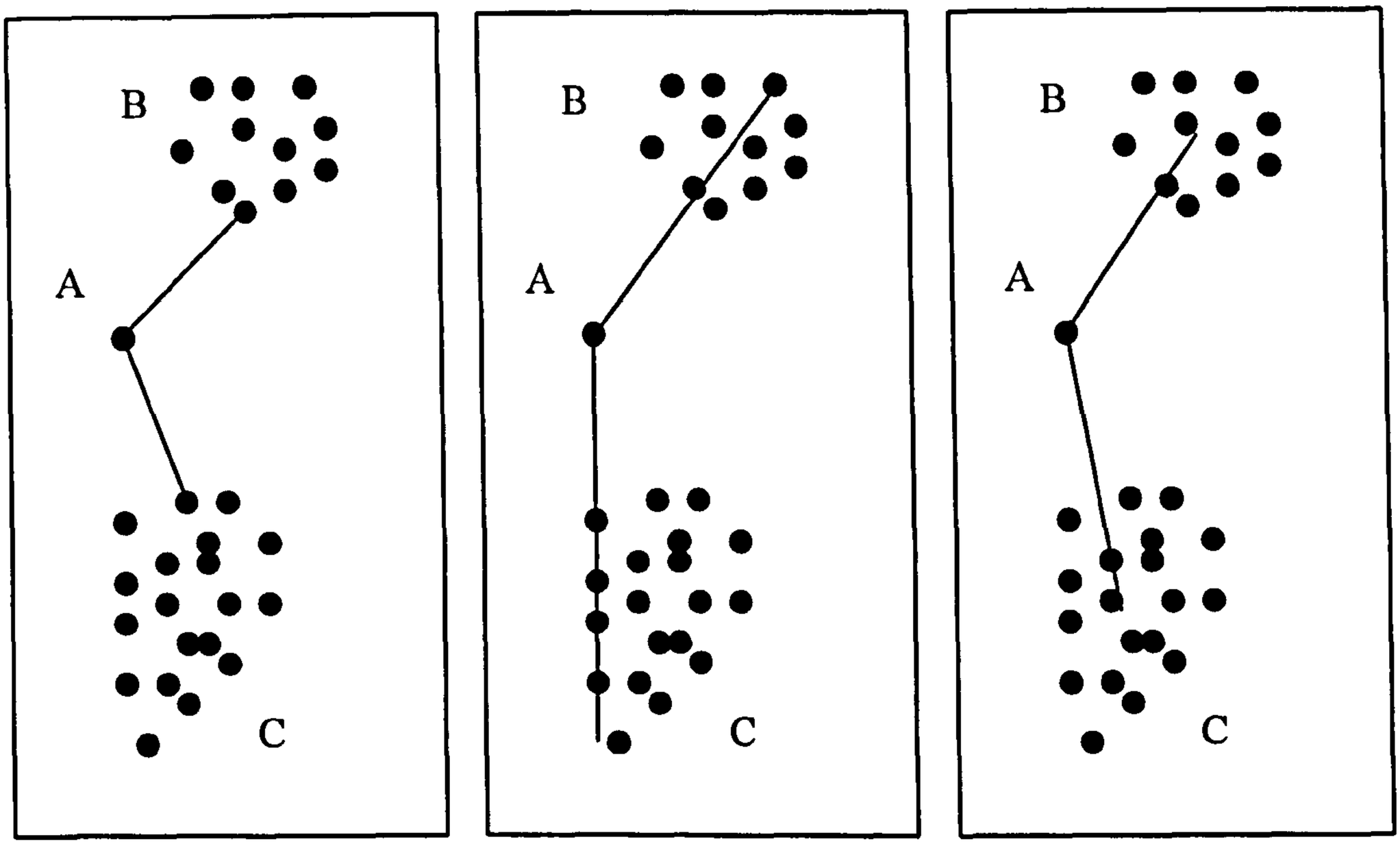
1.2.1.3.2. Cluster Analysis

Cluster analysis refers to techniques that examine the relationships between rows or between columns of a data set. This provides information about underlying factors that explain the information in a data space. In the petrol example used earlier cluster analysis would group the samples according to the biggest underlying source of variation. The clustering could be by manufacturer though as this is an unsupervised method there is no way to determine which manufacturer, or even if it is the manufacturer that has caused the clustering seen. Cluster analysis is normally carried out using a factor analysis technique as clusters can be seen to be related to underlying factors in a data set. Principal Components Analysis (PCA) could be used to carry out this task.

1.2.1.3.3. Hierarchical Cluster Analysis (HCA)

Hierarchical cluster analysis is a method by which the similarities between different rows or samples of a data set can be determined. This is performed by taking a vector from the data set and comparing it to the other vectors in the data set. The way in which the similarity is determined is dependent on the type of clustering carried out, most methods are based on the standard deviations and correlation's between vectors. The measure of the distance between two vectors is know as the Mahalanobis [26] distance, and is normally scaled to between one and zero, one being identical samples, zero being orthogonal. Vectors are examined one at a time and assigned to clusters, a cluster can consist of a single vector. The method by which a vector is assigned a

cluster can have a large effect on the type of clustering observed (Figure 1.1). The three most common method of assigning clusters are nearest neighbour, furthest neighbour and cluster centroid.



1. Nearest Neighbour. The vector A is considered, the cluster to which it is linked is the cluster that has the point closest to A in it.

2. Furthest Neighbour. The vector A is assigned to the cluster in which the furthest point from A within that cluster is closest to A.

3. Cluster Centroid. A is assigned to the cluster with the centre of gravity closest to A.

Figure 1.1 Three types of clustering used in Hierarchical Cluster Analysis

In all these cases (Figure 1.1) if the **point A** falls out side a certain clustering criteria it forms a new cluster. HCA can be carried out on raw data, pre-processed data or on principal components or latent vectors. The output of a HCA is either a table of distances, or a dendrogram (Figure 1.2).

In the case of the petrol example the histogram may group the samples by manufacturer, then by grade, or the other way round, or may show an entirely different clustering system. The way in which the samples are clustered could be

greatly affected by the clustering method chosen. The difference could cause a change between clustering by grade to clustering by manufacturer.

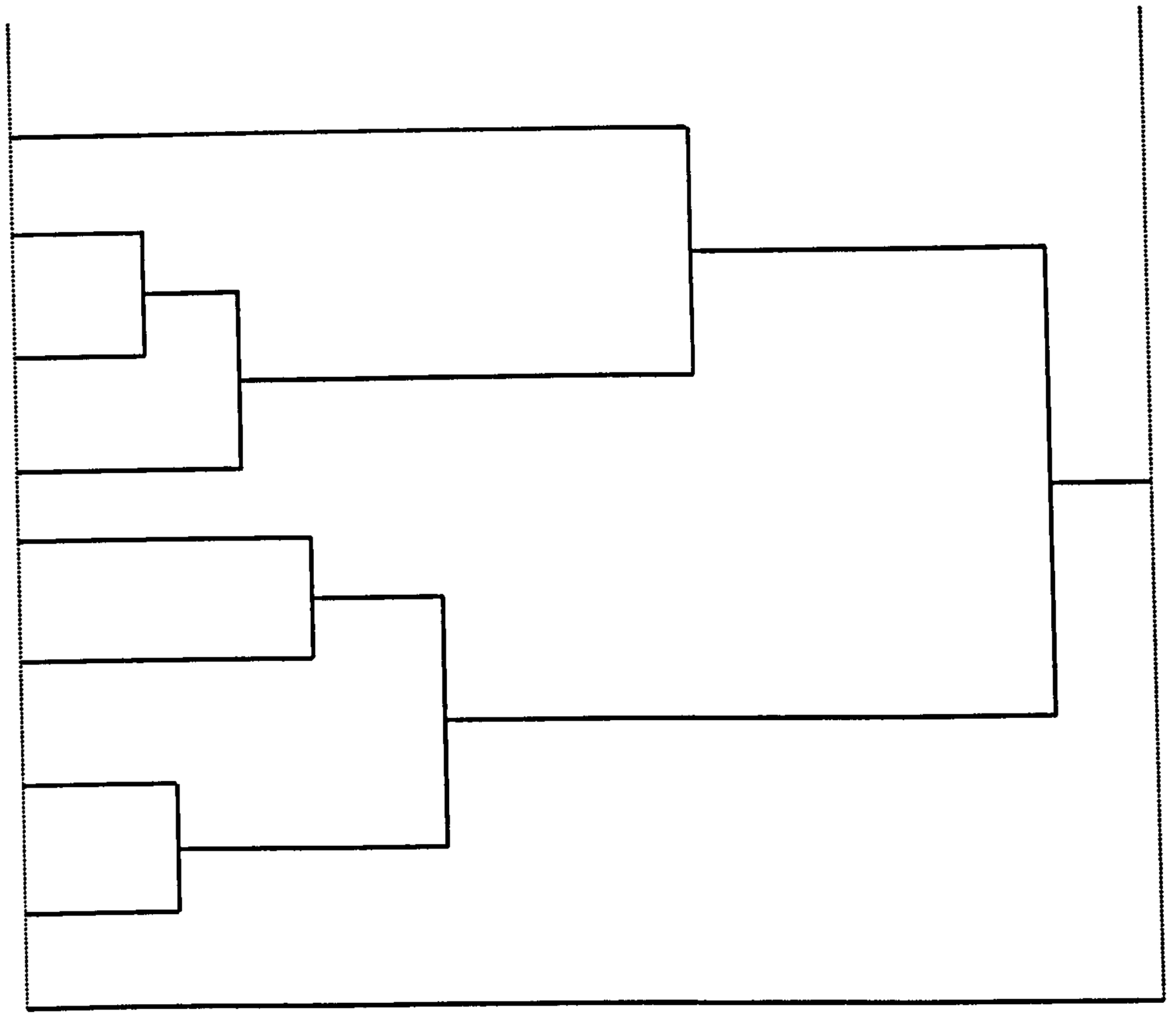


Figure 1.2 An example of a dendrogram showing random data

1.2.1.3.4. *Discriminate Analysis*

Discriminate Analysis is the name given to the group of techniques that look at assigning classes to groups of objects based on information about the data set [3]. Discriminate analysis is a supervised method. Class modelling can be seen to arise when the attribute to be predicted is discrete rather than continuous. In discriminate analysis a training set is used where the correct class assignments for each sample is known, a model is constructed, and an unknown sample can then be projected against the model to determine its correct class. This is commonly used in the food and drinks business, where for example a sample of wine analysed by UV spectroscopy

can be modelled to show which grapes were used to make it, and within grape types which growing region the wine came form.

SIMCA [27] is an old class determination method, and while it is still used it has several flaws. SIMCA uses factor analysis on each cluster to build a model of all the clusters present, new samples presented to the model are matched to each cluster until the cluster is found with the smallest amount of residual error. SIMCA cannot work effectively when the number of samples in each cluster is too small, so the data sets required to calibrate can be quite large. SIMCA is not robust when the clusters have distinct sub groupings, or when the clusters are too close together. SIMCA will however indicate when a sample could belong to more than one group, and it can also indicate when a sample does not belong to any group.

Another key discriminate analysis method is KNN [3], unlike SIMCA KNN is non-parametric, which means that detail within clusters is not as significant, and that the clusters can be any size. With KNN there is no within class modelling, the determination is entirely done using the distance measure (Mahalanobis distance). When an unknown sample is presented to a KNN model its distance measure is calculated, and this measure is compared to the K distance measures closest to it, the new sample is assigned to the cluster which has the most members within the K nearest measures. KNN will work with sample poor data, and if weights are used can assign unknowns to a cluster with only a single member. However KNN will always assign a new sample to an existing cluster, which can be an important consideration.

Using the petrol example a training set of petrol spectra together with information about who manufactured each sample could be used to develop a model, the spectral information from an unknown sample could then be taken and used to determine which manufacturer produced that particular sample of petrol.

1.2.1.3.5. Iterative Target Testing Factor Analysis (ITTFA)

ITTFA [4] extracts factors from a data set that contain information. These factors can be rotated to correlate to real world properties (Rotation is examined in the chapter on factor analysis 1.6.4). The number of factors within a data set is fixed however the alignment of those factors within the data set is open with an infinite number of possible orientations for any non-singular data matrix. (The problem of singular matrices will be examined in the chapter on factor analysis 1.6). As an example, with an UV spectroscopy data set of a solution of metals ITTFA can be used to extract the molar extinction coefficients for the metals in the solution. This is easy with UV spectroscopy because the UV signal for each component is Gaussian and linear as long as the solution obeys Beer's law. ITTFA can be seen to be a search method, examining a data space for factors that obey certain criteria, in the petrol octane example, ITFA would extract the spectra for the pure components of the petrol samples.

1.2.1.3.6. Target Factor Analysis (TFA)

TFA [4] uses many of the same principles as ITTFA, namely that are extracting factors according to real world rules. In the case of TFA factors can be extracted from a data space that correspond to input vectors. With spectroscopic data the input vector can be a pure spectra, the data space is decomposed into factors that include the target vector. This can be used to determined whether the target vector exists in the data space. In the spectroscopic example the data space can be tested to determine whether the component that formed the target vector is present within the sample that produced the data set.

This method would enable spectra of petrol samples to be tested for the presence of a certain additive.

1.2.1.3.7. Principal Component Analysis (PCA)

Any non-singular matrix can be decomposed into two different matrices, the combination of which will reproduce the original data matrices exactly [28]. The exact details of two possible methods of carrying out this operation will be covered in detail in the chapter on factor analysis. In principal components analysis the decomposition of the data matrix is constrained such that the variance within the data set is described by the new theoretical axes produced in order of decreasing variance. Thus the first principal component will contain the greatest amount of variance, the second component the next greatest and so forth.

PCA is a useful tool for exploratory data analysis because by plotting the data set on the new axes formed by the principal components the relative distribution of variance within the data set can be discerned. Groupings of objects of similar structure will occur, and data points that are outliers can become more clearly recognisable.

Likely clustering in the case of petrol samples would be by grade, or manufacturer, information about which portions of the spectra were important for the clustering shown could be obtained as well.

1.3. Calibration & Regression

Calibration is the name given to the process of relating one data matrix to another. This could be in the form of an input to a system to an output or property of the system, or of an output of a system to a property of a system. Calibration depends on

the relationship $X \propto Y$, this can be a direct linear relationship, an inverse relationship, a non-linear relationship, an inverse non-linear relationship, or any combination of these. If there is no relationship between the two (or more) data sets then no information transfer can take place. By convention the X data set is referred to as the Independent data set, and the Y as the Dependent data set. The X data set (or X-block) is normally a measured value, the Y data set (or Y-block) can also be a measured value but in chemical systems is as likely to be a calculated value, e.g. weight, concentration, percentage.

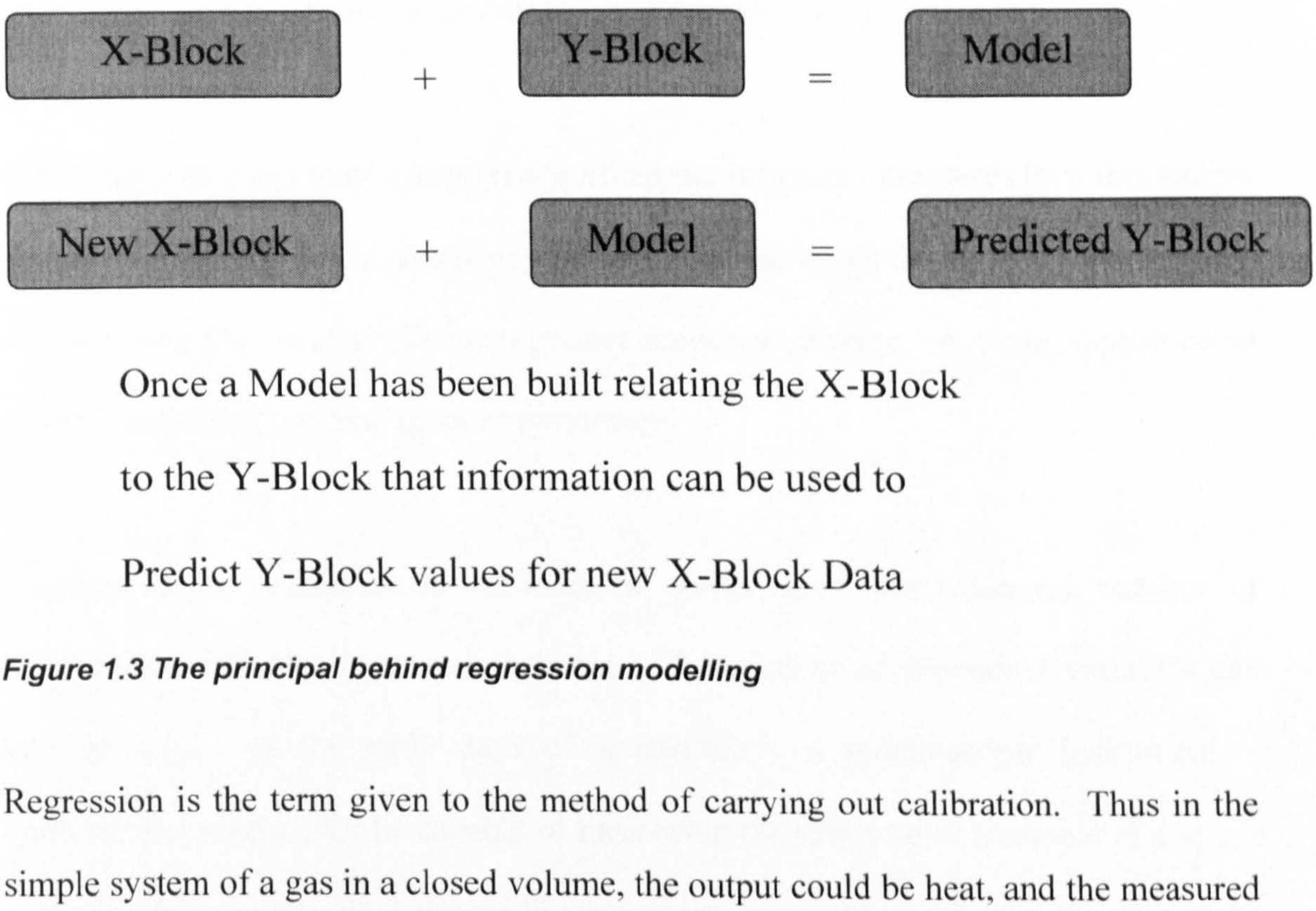


Figure 1.3 The principal behind regression modelling

Regression is the term given to the method of carrying out calibration. Thus in the simple system of a gas in a closed volume, the output could be heat, and the measured property the pressure. The calibration would be the process of relating the heat in the system to the pressure. Regression would be the method by which the relationship between the heat input and the pressure is determined (See Figure 1.3).

In a spectroscopic example the X-block would be the measured spectra of a sample, the Y-block could be the calculated concentrations of the components of the sample, calculated from the known mass of material that formed the sample.

Although the above statement implies that calibration examines two analogue values, the term calibration can also be used in the discrete sense of assigning a class to a sample as would occur in cluster or discriminate analysis.

There are great number of possible regression methods, the choice of the regression method is determined by the type of system being examined, and the information that is sought. In the case of an ideal gas in a closed system, the relationship between the temperature and pressure is a very simple one (equation 1.1),

$$Y = mX + C. \qquad \textbf{(1.1)}$$

With a non-ideal gas over a large range of temperatures and pressures then this simple equation will not provide an accurate solution, as the range of values becomes more extreme then the results will have greater amounts of error. A more sophisticated model is required, containing more parameters.

Standard linear regression is sufficient in some cases, but often the number of independent variables is greater than one. The number of dependent variables can also increase. In the early days of spectroscopy a spectroscopic instrument, a spectrometer might only be capable of measuring the response of a sample at a single wavelength at a time, thus simple linear regression would enable a calibration to be performed. This is a workable solution as long as the sample being measured is simple, there are no interfering matrix elements, and the instrument carrying out the measurement is reliable and free from drift. It was quickly found that calibrations carried out using measurements from several different wavelengths were more reliable and provided results with less error. Thus, linear regression was no longer

enough as more variables needed to be included in the calibration. Multivariate Linear Regression (MLR) is a method of including more than one variable into the X-block of a regression calculation, MLR involves minimising the least squares solution to equation 1.2

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots\ldots \beta_n X_n \qquad \textbf{(1.2)}$$

Using MLR the response or more than one frequency from a spectrometer could be included in a calibration. MLR has many problems, which will be discussed more fully in the section on MLR (1.5), and other methods were developed to removed these weaknesses. These methods include locally weighted regression (LWR), principal components regression (PCR) and projected latent structures (PLS).

In the three cases mentioned above, PLS, PCR and LWR, the principle equation is identical, however the terms corresponding to X are linear combinations of the original variables rather than the variables themselves. In PLS and PCR the functions that relate the linear combinations to the dependent data set are constant, in LWR they vary across the response surface to account for non-linearity's.

LWR is used only for calibrations using non-linear response data, by modelling linear sections of the data, and the whole calibration is then composed of many smaller linear models. PCR and PLS can be modified to model non-linear data, however this takes the form of a non-linear function relating the linear combinations of the original variables to the dependent variables. PLS and PCR can also be used on non-linear data when the data set itself has been linearised.

1.4. Data Pre-treatment

Chemometric techniques are all mathematical algorithms, they will all work on any appropriate sets of numbers. The results obtained from the use of the chemometric technique selected will vary with the quality of the data used. An essential task in any chemometric analysis is the determination of the appropriate technique to use, the use of an inappropriate technique will not provided any useful information. Often the difference between a high quality data set and a poor one, or the correct selected of technique and the wrong one can be affected by any pre-treatment that the data set undergoes.

Pre-treatment involves modifying a data set so that the useful characteristics are enhanced.

There are three main ways of pre-treating data, detecting outliers, smoothing the data set, and scaling the data set. All these processes can be applied to a single data set, though it is unusual for more than one method of a particular type to be used on a single data set.

1.4.1. Outliers

Outliers are samples or values within a data set that do not appear to come from the same population as the rest of the samples or values. Outliers are not extreme values within a data set, they are values from a different population. This means that it is not sufficient to calculate that a point is extreme for a data set and remove it, some valid reason must exist to exclude a sample from a data set. Samples and values within a data set can be tested to determine whether they qualify as outliers, but they must also be examined to determine the reason that they are outliers. The most common form

of outliers are from measurement errors, particularly when there is a human transfer of information. Typographical errors account for the vast majority of outlying samples found in process analysis data sets. Typographical errors can be rectified if the original source document exists.

1.4.1.1. Dixon's Q test

The Dixon Q [29, 30] test is a simple test to determine whether a single point is an outlier. The test is carried out by examining the relationship between the suspected outlier and its nearest neighbour, and the span of data including the suspect value (equation 1.3).

$$Q = (x_1 - x_2)/(x_n - x_1) \qquad \textbf{(1.3)}$$

where x_1 is the suspect value, and x_2 is the nearest neighbour to the suspect value. The Q result cross-referenced with the Q table of expected values, if the Q value exceeds the tabulated value the sample is considered an outlier. The test is easily carried out, however it is not effective when there are several suspect points. The test is best used for small data sets or single vectors.

1.4.1.2. Grub's Test

The Grub's test [29][30] is an examination of the standard deviation of a vector with and without the suspect value, or can be calculated as simple value derived from the mean of the vector together with the suspect value and the standard deviation. The two forms are as follows, equation 1.4 & equation 1.5.

$$G = \frac{(x_i - \bar{x})}{s} \qquad \textbf{(1.4)}$$

where G is the test value, x_i is the suspect value, $\bar{x}$ is the mean value, and s is the standard deviation.

Or,

$$R = 100\left(1 - \frac{S_1}{S}\right) \qquad (1.5)$$

from equation 1.5 where R is the percentage reduction in the standard deviation when the suspect sample is removed from the vector, S_1 is the standard deviation without the suspect sample, S is the standard deviation with the suspect sample.

The Grubs test is useful however its value decreases with increasing size of test vector.

1.4.1.3. Standard Deviation

Where there is a large vector to be tested, or many outlier are suspected then another useful method of examining outliers is by taking the standard deviation [29, 30] for the data set, and examining points that exist beyond a pre-determined limit, often ± 3 times the standard deviation. This method is useful for examining large data sets, and will often remove values that might distort a model however, care must be taken to examine removed values to determine their true status.

When factor analysis methods are applied to a data set there are other opportunities to examine the data space for outlying points. If principal components are taken from the data space, the original data set can be redrawn onto the new axis, and the values for the scores and loadings examined for outliers rather than the original values. This has the advantage that values are considered as outliers on the basis of their response within several vectors as opposed to just on the basis of a single extreme point. The methods already examined can be used with the scores and loadings from the data set.

Iterative target testing factor analysis can be used to determine whether a vector within a data set is taken from the same population as the rest of the data set. The vector to be tested is used as its own test vector, the resultant vector is compared to the original vector using an F-test. The significance level shows whether the two vectors are the same, if they differ the test vector is not from the same population as the rest of the data set and could be considered an outlier.

In all cases of outlier detection the vectors can be row or column vectors, though there should a valid reason for testing the data set by row or column. There is little reason to test for outliers within spectra, as opposed to testing by sample.

Samples should not be removed as outliers just because these tests indicate that mathematically they appear to be outliers. Each potential outlier should be considered to examine its reason for appearing different to the rest of the population.

1.4.2. Smoothing

Smoothing is the process of distributing random noise across a data set, the principle being that the random nature of the errors present will cause the error to cancel each other out. Smoothing also works for some type of systematic error, such as when there is a baseline drift on a spectroscopic instrument. Smoothing does tend to broaden peaks, so if the peak maxima have shifted on an instrument the smoothed spectra will tend to help counter this because the peak maxima will be spread across several wavelengths. This means that neighbouring wavelengths will provide the same information as might be found at the normal peak maxima.

Smoothing tends to hide unique events, spreading them across a sample. This means that smoothing is totally inappropriate for certain types of data, this includes process analysis data, and process measurement data. In the analysis of samples from a

process, the readings taken for different samples must remain discrete, this also applies to instrument readings from a process, individual readings will be from different times, sensors and even batches.

The most frequently used type of smoothing is moving average. Here a window of values is taken, and the average of the values in the window replaces the value at the centre of the window. The windows moves through the vector from one end to the other. This size of the window taken reflects the amount of smoothing required, the larger the window the greater the smoothing and the more information lost.

1.4.3. Scaling

Scaling a data set can be used to correct certain types of problems with a data set, or to adjust a data set to highlight features of interest. Scaling a data set can included modifications made to account for non-linearity. Scaling is used to counter non-linearity in a data set, and it is used to enhance features.

Any vector will have a standard deviation (equation 1.6),

$$S_K = \left[\frac{1}{NP-1}\sum_{i=1}^{NP}\left(x_{iK} - \overline{x_K}\right)^2\right]^{1/2} \qquad (1.6)$$

and a mean (equation 1.7),

$$\overline{x} = \frac{\sum x}{n} \qquad (1.7)$$

the standard deviation is an indication of the degree of variance in a vector and its magnitude within the vector, and the mean will indicate the average values for the vector. When the vector is from a single population and is normally distributed these two statistics are useful descriptors for the data set. When there is more than one vector being considered, any large differences in the standard deviations and means of

the vectors can have a significant effect on the results of any calculations carried out, particularly any factor analysis techniques. Factor analysis techniques look to extract useful information from a data set, large variation in the standard deviations and means of a data set can mask the variation that is required to produce the model. One possible method of reducing the influence of variations in magnitude and variation between vectors is to use scaling methods.

1.4.3.1. Range Scaling

Range scaling is carried out by dividing the values in a vector by the maximum absolute value in the vector (equation 1.8).

$$x_i = \frac{x_i}{x_{max}} \qquad \textbf{(1.8)}$$

this has the result of scaling a vector between 1 and –1. This does not centre the mean to zero however, though the mean may coincidentally be zero. Range scaling removes the effect of magnitude from a set of vectors, however it has no effect on the variance within the vectors.

1.4.3.2. Mean Centring

Mean centring sets the mean of a data set to zero. This is carried out by subtracting the mean of a vector from each value in the vector (equation 1.9).

$$x_i = x_i - \bar{x} \qquad \textbf{(1.9)}$$

mean centring enhances the variance in a vector, this can have unexpected results on a data set when the vectors have similar magnitudes but widely differing means.

1.4.3.3. Autoscaling

Autoscaling sets a vectors mean to zero, and its standard deviation to one. This removes the influence that magnitude and extreme variation might have. Magnitude distorts factors extracted from a data set because either the coefficients must overcompensate for the extreme values, or the factors selected are biased towards the vectors with large magnitudes. Overcompensated factors lead to greater noise in predictions. Extreme variation overemphasises vectors containing noise in comparison with vectors containing information.

Autoscaling can be carried out using the following formula (equation 1.10).

$$x'_{iK} = \frac{x_{iK} - \overline{x_K}}{S_K}. \qquad \textbf{(1.10)}$$

1.4.3.4. Linearisation

It is often easier to linearise the data set rather than try and develop an non-linear model. If the non-linearity within a variable or data set is constant across the range then it may be possible to linearise the variable. This requires that the type of non-linearity be determined, such as a log term, cube term e.t.c. The best method for this is to make use of knowledge about the system being examined. Chemical knowledge can often indicate what the non-linear term within a system might be. An examination of the normality of the variable can provide information, the chi-squared test can be used to determine how the variable differs from normal, which can give information about non-linearity. Herteroscedatic residuals can also indicate non-linearity, though heteroscedatic residuals are also symptoms of systematic noise or the result of other types of scaling.

Once the type of non-linearity has been determined then the vector can have the appropriate function applied to it.

If the type of non-linearity cannot be determined then either trial and error may lead to the correct solution, or a non-linear model could be developed.

1.4.4. Chi-Squared Calculations

The Chi-squared calculation [30] is designed to provide information about the distribution of points within a data set. This is useful both as a test for normality, and also as a method of determining the profile of the population that a series of points comes from. The Chi-squared calculation requires the mean and standard deviation to be calculated, this can be calculated from the data set, or used as a target for an analysis. The distribution of points within a data set can be compared with the distribution that should exist for that data set with its given mean and standard deviation.

1.5. MLR

Multivariate Linear Regression (MLR) is essentially a method of solving a system of simultaneous equations (equation 1.11). The aim is to find the coefficients for the independent variables that will allow the calculation of the dependent variable.

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \beta_3 X_{13} \cdots + \beta_i X_{1i} + e_1 \\ y_2 &= \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \beta_3 X_{23} \cdots + \beta_i X_{2i} + e_2 \\ &\vdots \\ y_2 &= \beta_0 + \beta_1 X_{m1} + \beta_2 X_{m2} + \beta_3 X_{m3} \cdots + \beta_i X_{mi} + e_m \end{aligned} \qquad \textbf{(1.11)}$$

β represents the linear regression parameters, these can only be determined in limited cases, normally ***b*** is taken, the linear estimation of β. β can be calculated when the

number of coefficients is the same as or less than the number of samples (rows), the mean of the random errors is zero, and they are normally distributed.

Where no exact solution is possible, a solution is obtained to satisfy equation 1.12 where the error term is minimised: -

$$E = \| xb - y \| \tag{1.12}$$

When the number of variables is greater than the number of coefficients, the over determined case, and when the number of variables is less than the number of coefficients, the under determined case, no exact solution is possible as there are an infinite number of possible solutions. In most modern chemical systems, either one situation or the other exists. In spectroscopy modern scanning instruments can measure several thousand variables with ease, measuring the same number of samples would be problematic. In large scale chemical processes measurements can be taken every second for a number of variables, quickly building a data set with many times the number of samples compared to variables.

The solving of simultaneous equations is essentially a matrix manipulation problem. MATLAB is the ideal tool to use in this situation since MATLAB is designed specifically to handle matrix manipulation. The regression coefficients can be calculated using simple matrix operators.

Starting with the term to be minimised, the squared length of the error term (equation 1.13),

$$\begin{aligned} E^2 &= (xb-y)^T(xb-y) \\ E^2 &= x^T xb^2 - 2x^T yb + y^T y \end{aligned} \tag{1.13}$$

this is minimised by taking the derivative with respect to ***b***, setting ***b*** to zero (equation 1.14)

$$\frac{dE^2}{db} = 2x^T xb - xy^T y = 0 \tag{1.14}$$

which gives equation 1.15,

$$b = \frac{x^T y}{x^T x} \tag{1.15}$$

because of the relationship with equation 1.16

$$\frac{x^T}{x^T x} = (x^T x)^{-1} x^T \tag{1.16}$$

this becomes equation 1.17,

$$b = (x^T x)^{-1} x^T y \tag{1.17}$$

and from this,

$$(x^T x)^{-1} x^T \tag{1.18}$$

is the right pseudo inverse (equation 1.18), MATLAB ® uses the right pseudo inverse when the forward divisor (/) is used, the calculation of the regression coefficients can achieved using a single line of commands written in MATLAB ®.

MLR has a number of drawbacks. In the over and under determined systems the coefficients, ***b***, can vary widely with just a small variation in the data set, this is particularly true when the independent data matrix is nearly singular. This is when the rank of the matrix is less than the dimensionality, in practise this is when one column of a matrix is collinear with another variable, or combination of variables. This problem can be demonstrated very simply.

$$x=\begin{bmatrix}1 & 2 & 3\\ 2 & 4 & 6\\ 3 & 4 & 7\\ 4 & 8.0001 & 12\end{bmatrix}, y=\begin{bmatrix}2\\4\\6\\8\end{bmatrix}, b=\begin{bmatrix}2\\0\\0\end{bmatrix} \quad \textbf{(1.19)}$$

x is not quite singular, and the calculation of ***b*** appears to give reasonable answers (equation 1.19), however if ***y*** is changed slightly (equations 1.20 & 1.21), the coefficients can vary widely.

$$x=\begin{bmatrix}1 & 2 & 3\\ 2 & 4 & 6\\ 3 & 4 & 7\\ 4 & 8.0001 & 12\end{bmatrix}, y=\begin{bmatrix}2\\4\\6\\8.0001\end{bmatrix}, b=\begin{bmatrix}12\\10\\-10\end{bmatrix}, \quad \textbf{(1.20)}$$

and,

$$x=\begin{bmatrix}1 & 2 & 3\\ 2 & 4 & 6\\ 3 & 4 & 7\\ 4 & 8.0001 & 12\end{bmatrix}, y=\begin{bmatrix}2\\4\\6\\7.9999\end{bmatrix}, b=\begin{bmatrix}-8\\-10\\10\end{bmatrix}, \quad \textbf{(1.21)}$$

In ICP and UV spectroscopy this is a particular problem as the peaks produced are normally simple Gaussian curves. Thus any group of samples with little error and only one component is likely to produce a data set that is nearly singular, even some of the more complicated data sets can suffer from this.

Although there are a number of modifications to MLR that can be made to accommodate these problems, the best solution is to move to a different approach. One of the more successful group of methods are factor analysis techniques, these include PCR and PLS.

1.6. Factor Analysis [4]

In the calibration of equation 1.22,

$$bX + E = Y, \tag{1.22}$$

the data set X can be considered to be composed of an information term and an noise term (equation 1.23),

$$x = i + e. \tag{1.23}$$

Unmodified it is difficult to separate the information from the noise. Smoothing can be used to help with this, but most forms of smoothing assume normal and random distribution for error, in many cases, error is neither normally distributed nor random. Smoothing also makes no allowances for noise within the data set that is information not useful to the model being built. If the variables that form the data set can be recombined into a form where the information is already separated from the noise and error then this problem can be solved.

Factor analysis is a method of producing a linear combination of the original variables where the noise term is separated from the information term. The way in which this occurs depends on the actual factor analysis technique being used.

All factor analysis method relies on the basic principal that any non-singular matrix can be decomposed into two other matrices (equation 1.24).

$$\underset{n*m}{X} = \underset{n*m}{U^t} \underset{m*m}{V}, \tag{1.24}$$

The rules that are used to generate the two matrices determine the type of information produced. If the factor analysis method being considered is Principal Components Analysis or Principal Components Regression, then the equation to be considered is;

$$DP = \lambda P, \tag{1.25}$$

In this case (equation 1.25), p_i is an eigenvector, and λ_i is its corresponding eigenvalue, and $\boldsymbol{D}$ is the covariance matrix from the data set. In principal components analysis or regression $\boldsymbol{p}$ is known as the loadings, and provides information about the columns (variables) of $\boldsymbol{D}$. Information about the samples or rows of $\boldsymbol{D}$ can be found by calculating equation 1.26.

$$t = D\ p\ , \qquad \textbf{(1.26)}$$

the matrix $\boldsymbol{t}$ is known as the scores.

$$DP = \lambda\ P \qquad \textbf{(1.27)}$$

In equation 1.27 the values for λ are calculated individually by iteration. The result of this is that the first vector of $\boldsymbol{p}$ will describe the most variance from the data set. The second vector of $\boldsymbol{p}$ will describe the next greatest axis of variance.

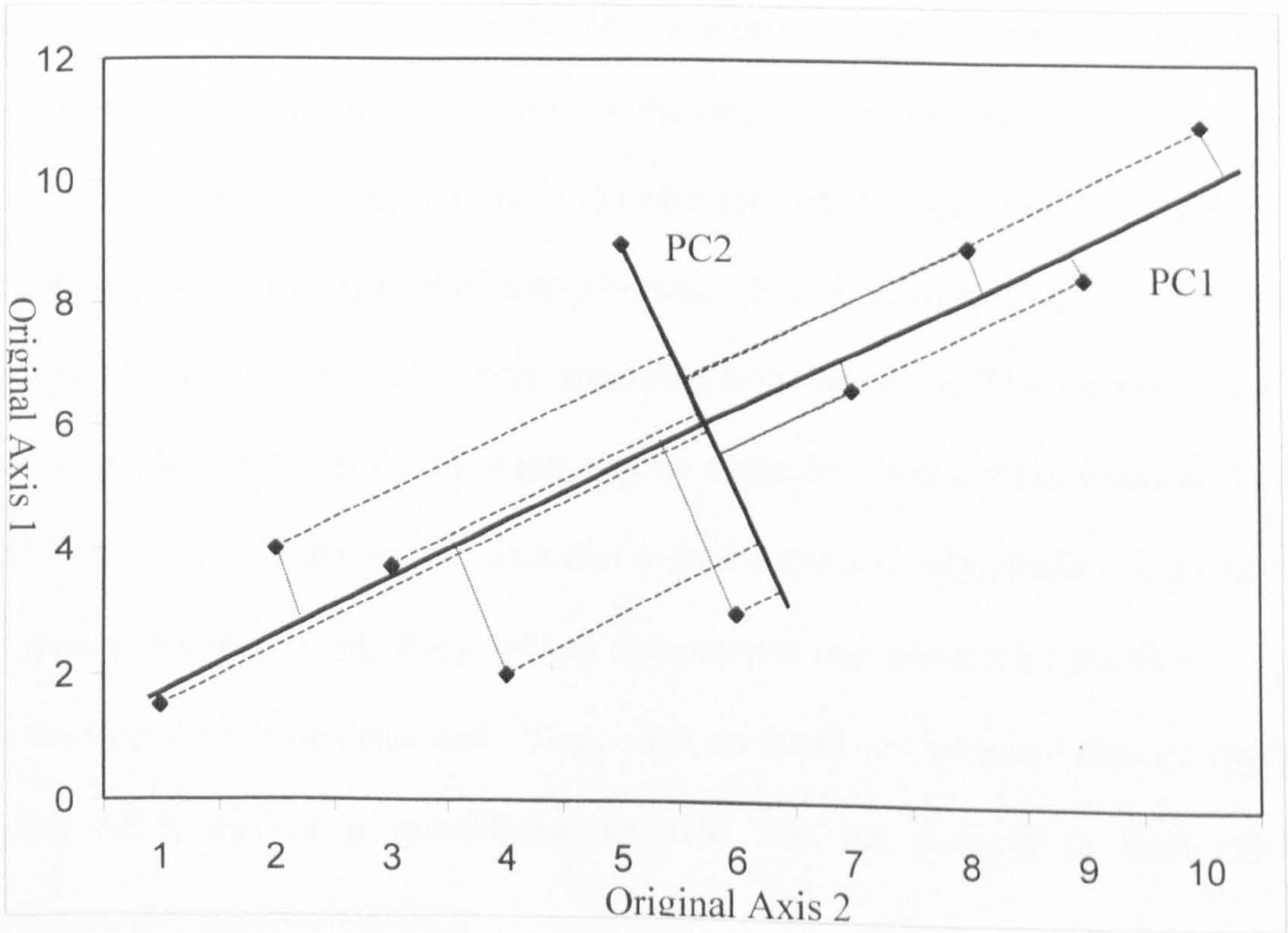


Figure 1.4 Graphical representation of how two principle components could be derived

The above figure (Figure 1.4) shows ten points plotted at random. These points could be re-plotted on new orthogonal axis that pass through successive quantities of variance. The line labelled PC1 passes though the greatest amount of variance, with only two dimensions only one other axis is possible, and this passes through the next greatest variance with the constraint that it is orthogonal to PC1. If this were a calibration, PC1 would represent the least squares best fit between the points [1.5, 4.0, 3.7, 2.0, 9.0, 3.0, 6.6, 9.0, 8.5, 11.0] (X) and the numbers 1 through 10 (Y). If it is assumed that the relationship between these two sets of number is linear, and that the error is entirely in the X axis, then vertical distance between PC1 and the points would represent error in the measurement of X.

If information about the relationship between the original axis and the two new axis is retained (the loadings) and the positions of the points on these new axis (the scores) is produced then no information is lost and the original data set can be recreated with no loss. If only the information about the positions of the points on PC1 were taken, which represents the true relationship between X and Y, then it can be seen that the information in the data set has been separated from the noise. This example uses only two variables, however the principle can be expanded to many dimensions. In PCA the data set is redrawn on new axis that describe successively smaller portions of the variance. By taking only the principal components that contain information the noise in the data set can be discarded. The points on these new axis can then be regressed using MLR against a quantifying variable and the process is then principal components regression (PCR).

In the decomposition of the data matrix that leads to PCA and PCR there are two key properties of the resultant matrices that force a single maximal solution to the result. Starting from equation 1.28,

$$Dp_i = \lambda_i p_i, \tag{1.28}$$

or equation 1.29,

$$\left|D - \lambda I\right| = 0, \tag{1.29}$$

where $\underset{n*n}{D}$ is non-singular, and has ***n*** real non-negative roots (equation 1.30) (eigenvalues),

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \lambda_n, \tag{1.30}$$

$$p_i^T * p_i = 1, \tag{1.31}$$

that is the variance captured by each component is maximal (equation 1.31), and from equation 1.32 and equation 1.33 it follows that

$$t_i = D\ p_i, \tag{1.32}$$

$$t_i^T * t_j = 0, i \neq j, \tag{1.33}$$

that is that the new vectors (equation 1.2) are orthogonal (equation 1.33).

Using these equations the eigenvalues and eigenvectors of the matrix ***D*** can be determined by successive approximation, an approximation for p_1 is entered, λ_1 is determined, then p_1 is recalculated. This is repeated until there is no change in the value for p_1. The first loading is multiplied out by the data matrix, to give the scores for the first component (equation 1.34),

$$t_1 = D\ p_1, \tag{1.34}$$

the data matrix is recomposed from the first principal component and this is subtracted from the original data matrix (equation 1.35),

$$D^{*} = D - (t_1^T * t_1), \tag{1.35}$$

the second principal component can then be extracted by the same procedure from D^{*}. This is repeated until n components have been removed, all that remains in D^{*} should be the electronic error. With MATLAB on a Pentium][computer the electronic error is the value $2.2*10^{-16}$.

This method is classical eigenvector decomposition, and is computationally exhaustive, it also can produce unstable solutions, and fails to converge with some data sets. This type of matrix decomposition is more usually carried out by Single Value Decomposition (SVD) to avoid these problems.

1.6.1. SVD

SVD [4] is a non-iterative method of decomposing a matrix that fulfils all the requirements of PCA (equations 1.30, 1.31, 1.32 and 1.33).

Starting with a data set $\underset{p*q}{X}$, this can be expressed as equation 1.36,

$$X = L\begin{bmatrix}\Delta & 0\\ 0 & 0\end{bmatrix}M^T, \tag{1.36}$$

where,

L & M are orthonormal, Δ is a diagonal matrix where the non-zero elements are the square roots of the eigenvalues of equation 1.37.

$$X\,X^T \,\&\, X^T X\,, \tag{1.37}$$

these are called singular values, and where equation 1.38 and equation 1.39hold,

$$L^T X\ X^T L = \begin{bmatrix} \Delta^2 & 0 \\ 0 & 0 \end{bmatrix}_{p*p}, \tag{1.38}$$

$$M^T X\ X^T M = \begin{bmatrix} \Delta^2 & 0 \\ 0 & 0 \end{bmatrix}_{q*q}, \tag{1.39}$$

thus, if

$$Xp = \lambda p, \tag{1.40}$$

$$(X - \lambda I)p = 0 \tag{1.41}$$

SVD is mathematically identical to PCA, and the properties of the resultant matrices remain the same however SVD is calculated in a single step from the original data matrix, rather than using an iterative method. In an examination of the scores and loadings matrices produced by these two methods, it can be seen that the numerical values in these matrices are not identical for the two techniques.

1.6.2. PLS

Projected Latent Structures (PLS) [4] is a method of decomposing an X block matrix and a Y block matrix into vectors such that the resultant vectors from the X block are highly correlated with the vectors from the Y block.

The result of this is that the coefficients of the X block variables that provide information relating to the Y block increase, while the coefficients for variables with no information tend towards zero.

The PLS algorithm used in this work is the NIPALS [4] algorithm, which is based on the PLS2 procedure. The PLS2 procedure is different from the PLS1 procedure in

that it allows the calculation of coefficients for data sets with more than one vector to the Y-Block. This has important implications for this work since the data sets concerned all have four vectors in the Y-Block, if PLS1 were used the coefficients would have to have been calculated individually for each vector. The implication of this is that any multiple collinearity or interactions would be ignored during the calculations, and the variables selected would be calculated independently for each vector, this would lead to both redundancy in the variables selected and to the possible loss of information concerning overlapping peaks.

NIPALS [4] relies on the mathematical fact that seen in equation 1.42,

$$D_j = \sum u_j s_j v_j^t \qquad \textbf{(1.42)}$$

Where **D** is the Data matrix, ***u*** & ***v*** are vectors, and s is a scalar for all ***D*** where ***D*** is non-singular (A singular matrix has no inverse, and so cannot be used for these calculations).

This expression can be seen in equation 1.43,

$$D\ v_1 = u_1 s_1 \qquad \textbf{(1.43)}$$

Here a randomly selected vector v_1 is selected and used to calculate s_1 & u_1 this is an approximation of u_1, a better approximation can then be found by recreating v_1 using equation 1.44

$$u_1^T D = s_1 v_1 \qquad \textbf{(1.44)}$$

This is repeated until convergence for a value of v_1. This allows the calculation of D_1 the first approximation. The residual matrix is then calculated from equation 1.45,

$$E_1 = D - D_1 \qquad \textbf{(1.45)}$$

The next eigenvector v_2 can then be extracted from the residual matrix. In each stage of the calculation of the vectors u_j and v_j the vectors are normalised to unit length to ensure orthogonality between the vectors.

NIPALS describes the decomposition of a matrix into eigenvalues and eigenvectors however this is for one matrix and does not allow for a relationship between two matrices. NIPALS can effectively be used to carryout PCA however this can more effectively be done using SVD. NIPALS is useful in that it allows for the possibility of relationship between two matrices. If the eigenvectors are calculated simultaneously for two different matrices (equation 1.46 & equation 1.47),

$$Y\ p_i = q_i a_i \tag{1.46}$$

$$D\ v_i = u_i s_i \tag{1.47}$$

then a relationship can be found between p_i & v_i and q_i & u_i

such as is seen in equation 1.49 and equation 1.50,

$$w_i q_i = u_i \tag{1.49}$$

$$t_i p_i = v_i \tag{1.50}$$

thus for the first latent variable, an estimation of v_i would be made, then an estimation of p_i, then an estimation of t_i, and so on, this process is cycled until convergence. The residual matrices are then calculated and the next eigenvector generated. This process can be stopped when the required amount of information has been extracted from the matrices. One of the major advantages of PLS is that this process can be carried out for more than one Y Block vector, this process needs to be carried out for each Y Block vector, producing a vector of weights for each. This can increase the time taken for the calculations considerably, the number of calculations required is multiplied by the number of Y Block variables.

PLS provides both predictive information, allowing calibration of an X-block against a Y-block, and it also provides descriptive information about how the Y-block data affects the Y-block data. This diagnostic information is useful for fault diagnosis and error detection. One of the faults of any variable selection process is the loss of descriptive information in the X-block and that relationship with the Y-block, and a consequent loss of fault detection. The routine for variable selection presented in this paper is less susceptible to this problem than many other techniques because it does not concentrate on highly correlated variables or variables at the centre of peaks as most of the other techniques tend to do. This will be cover further later on in the paper.

1.6.3. PLS vs. PCR

PLS and PCR are possibly the most commonly used factor analysis techniques for regression analysis of two-dimensional data. Which technique to use is an important decision to make. In simple terms PCR maximises variance, and PLS maximises correlation. This will affect the choice of appropriate technique to select. When the information required for a calibration is a small part of the total variability of a matrix then PCR will have trouble modelling. This is because PCR selects principal components according to variation, the first components selected will not contain useful information, the required information will be in the smaller components. When the number of components in the matrix exceed the number of components for which there are Y-block variables then PCR will also have problems. This is due to unwanted variation for unknown components being captured in principal components

that also contain the information for wanted components. These considerations mean that for all but the simpler problems PLS is likely to provide an equal or better model.

A general rule is that PLS will capture the required information for modelling with fewer latent vectors than PCR would require principal components and will have lower error. PLS can require more calculation to find a solution which means on old computers it may be a slower algorithm, this last point should not be a consideration with modern computers. PLS can also require more memory in the computers carrying out the calculation as a larger number of matrices are required simultaneously to carry out the calculations, this is also only a minor consideration with modern computers.

1.6.4. Rotation

Factor analysis methods can be considered as a form of rotation, the original axis that the data is displayed on are rotated so they have new properties more closely related to the problem being examined. This rotation produces abstract factors, that is, factors that have no meaning to the real world, there may be a requirement to change this however. Once the factors have been extracted they can be further rotated to align them with real properties. As an example, in ITTFA the factors are rotated till they equal the molar extinction coefficients of the components present, remaining factors beyond the number of components present a noise.

When produced, the factors are formed in orthogonal pairs, that is that they are at right angles to each other, there are two types of rotation commonly used, rotation that retains that orthoganality of the factors, two methods of which are Quartimax and Varimax [31], and rotation that does not, this is known as oblique rotation, there are

many methods that are oblique, such as Oblimax, Quartimin, Biquartimin and Promax [32]. ITTFA is an oblique rotation method since it cannot be assumed that the molar extinction coefficients will be orthoganal to each other.

1.7. Variable selection

Variable selection has been a recurring theme in chemometrics from since multivariate techniques were developed. The reasons for choosing variable selection vary, but three important reasons should be discussed.

1. Although modern instruments are capable to recording thousands of wavelengths in a very short period of time, depending on the technique from milliseconds to seconds, this often comes at a price. Instruments capable of scanning large numbers of frequencies are often expensive, considerably more so that an instrument designed to scan just a few wavelengths. If the wavelengths of a spectrum that can be used to solve a problem can be identified then an instrument can be purchased to examine just those wavelengths.
2. When thousands of wavelengths are scanned the problem of calibration becomes more difficult, in order to carry out a prediction a computer is required to carry out the thousands of calculations needed. If a satisfactory model can be developed with a small number of wavelengths then the problem becomes one that can be dealt with with a calculator. Predictions based on calibrations can be made by recording the responses at the selected wavelengths and simply multiplying by the appropriate coefficients.
3. If there is only a single analyte of interest in a complex matrix the measurements of the responses of frequencies not of interest will introduce error into the model. It is possible to reduce this problem by selecting sections of a spectra to examine,

however this is a crude method in most circumstances and is not a precise as calculating the correct wavelengths to record.

Most variable selection procedures are computationally expensive. The exception are those based on examination of either loadings coefficients or on correlations between the dependent and the independent data matrices[33] [34]. These two methods can be carried out rapidly since they only require one calculation of the relevant coefficients, though if the loadings are used to select variables then the calculations must be repeated after variable selection to ensure that removing certain variables does not unduly perturb the system. Removing a variable from a multivariate system can have a large effect on the corresponding coefficients, though the effect is larger with methods such as MLR and PCR compared with PLS. The computational expense of variable selection is rarely a reason not to carry out variable selection as modern computers can carry out most calculations in a reasonable period of time.

An argument against variable selection is that reducing the dimensionality of a data set reduces the ability of the model to detect faults in the system being modelled. This is based on the fact that in process modelling the stability of the process under examination is often measured by examining the noise in the system. When the system is stable the noise will tend to be stable. Changes in the noise will indicate a change in the system that may require correction. This approach is common and works for any stable process, however if the process is examined by examination of the active components then this information is not entirely lost. Variable selection can make detection of unexpected matrix elements more difficult and if the situation calls for the detection of foreign materials in a process stream for example then

variable selection may not be appropriate. When the matrix is known to vary and it is still only the response of one of mode components of the matrix that is of interest that variable selection will allow a more robust model as it will not include error introduced by matrix elements that were not present during modelling.

When the reason for a model is purely the prediction of one or more components then variable selection will invariably allow a better model to be built than could be constructed without.

In most spectroscopic situations an ordinary MLR model will be hugely underdetermined, the number of variables will be greater than the number of samples for most situations. Variable selection can correct this.

1.8. Model Building

Model building is the process by which the relationship between the independent data matrix and the dependent data matrix is determined. Here the assumption will be that both matrices are two-dimensional. Methods that use three dimensional or greater matrices will not be covered here. An examination of the process required to model a data set against reference materials can be seen in the appropriate BS ISO document [35].

There are a series of steps to model building that should be followed, some of the steps have a greater importance than others.

Initially it is important to know what questions will be answered by the model. This would normally be considered before all else as it will determine the type of model

built and the techniques used. This is normally a consideration as to whether quantitative information is required from the model, or qualitative. Some form of regression will be required for quantification and some form of classification will be required for qualification. Only regression models will be considered here.

What information is available to build the model? This question is best asked, and answered, before any data is collected allowing experimental design to be used to optimise the whole modelling process. If this question is asked before the data is collected then experiments can be designed to collect the required number of samples of sufficient variation. Ideally the samples will span the possible range of responses required of the model. Any model required to predict beyond the range of input data will lack robustness. The range to be calibrated must be determined and samples collected to span that range for all components of interest, and ideally including information about possible interference and matrix effects.

Once these steps have been carried out where possible, an initial examination of the data will give an idea of which of the appropriate techniques would be the best starting place, and the consideration of any pre-processing can be made.
Unsupervised clustering analysis would provide useful information, indicating highly correlated variables, and any deviation from normality within the data matrix that might effect the modelling or type of modelling required.

The initial analysis will indicate whether any pre-processing is appropriate. Spectral information may require correction for baseline problems, and some spectra will require modification to highlight the features of interest. For example NIR spectra

would normally require some form of derivative to be taken as the variation in the spectra that holds the required information will normally be only a small part of the variability within that spectra.

Pre-processing will then be carried out. Pre-processing can involve several stages, dealing with missing values, with large data sets that can often be carried out by eliminating samples or variables that contain omissions. If it is considered inappropriate to remove whole rows or columns then the results must either be obtained by new experiments or calculated in some way, either interpolation, imputation or extrapolation. Extrapolation will reduce the robustness of the model. Interpolation can be carried out when there is strong autocorrelation within the variable, and imputation is carried out to retain certain properties of the data set. Interpolation carried out by calculating the missing values from the other X-Block variables is a step that will reduce robustness since this implies a degree of collinearity, and thus means that the matrix will be singular or nearly singular.

Modelling with the appropriate technique is the next step. A good guide here is Occams Razor, the simplest model is the best, thus for most regression purposes the progression of techniques should be LR, MLR, PCR, PLS. That is, a linear regression model where it provides a sufficiently high quality answers is all that is needed. Obviously, a univariate approach is very limited and only appropriate for a very small number of possible cases, but it can always be considered as a starting point. MLR provides many advantages over linear regression, and is still a remarkably good method, particularly if some form of variable selection is used. Many of the limitations of MLR have been addressed in other texts, and solutions to MLR can be

found to solve most irregularities in a data set. The effort required to optimise a least squares method means that moving onto a factor analysis approach is normally a better solution. If a least squares method is required then there are variation such as GLS, WLS, CLS, NLS, [29] and then several versions of least squares that consider the sources of error in a model and attempt modelling without the assumption that all the error is in the X-Block. If MLR is insufficient to model the data then a factor analysis method can be used, PCR is a useful technique, in many cases it can provide a far superior model to MLR. Where PCR does not work, PLS can be tried.

Any appropriate pre-processing is then carried out.

When a method has been selected and the data pre-processed for the initial modelling the data available for the model must be arranged. Some form of validation will be required for any serious model. The data set ideally would be separated into a training set, a test set for standard methods, and a validation set will be required for factor analysis methods. The training set is the set of data from which the relationships between the X-Block and the Y-Block will be derived. The test set is the data which will be used to optimise the model during factor analysis model building, and to determine the error with simple model building. The validation set is the data set that will be used after the model has been built to determine the model error. Ideally the three sections of the data will occur by random selection. The data set as a whole should have no replicate samples in it.

With LR and MLR and other non-factor analysis methods the regression coefficients are calculated, the error is calculated using the validation set and some assessment is made as the models quality.

With factor analysis methods the number of factors be used in the model must be optimised. It is important to recognise that if the number of factors chosen is equal to the number of variables used to build the model then there will be no difference between the MLR model and the factor analysis model. Methods used to determine the number of factors to use include block validation, leave one out validation and venetian blinds validation. Block validation methods tend to be superior to other methods in terms of determining how good a predictor a model will be, cross validation methods tend to suggest too few factors be kept in the model for the model to be robust. The reason to leave out a test set is to allow for block validation during the factor selection stage.

The error of prediction for the completed model is then used to determine its quality.

Usual techniques to determine predictive ability are PRESS and PEP.

1.9. Process analysis

Process analysis covers the use of statistics to analyse the data produces by industrial processes, in this context chemical ones, though that is not the only context where this type of maths is appropriate. Any system that produces large amounts of data can benefit from the use of chemometric techniques. Although the simple methods such as t-tests and F-tests have their place here, they are not common, regression modelling methods are more frequently used together with trend analysing tools.

When processes are examined rather than spectra there are several key considerations. First scaling methods are often used, it is the nature of measurements on differing physical properties that they are likely to be measured on different scales. Temperature in Kelvin may be several orders of magnitude lower than a pressure in Pascals. Scaling methods are chosen that reflect this, and typically range scaling will be selected.

Second smoothing is almost never applied. Each sample is a discrete segment of information often spread over the time domain. With data from batch systems there can be no smoothing between batches as this will compromise the clear distinction between batches. Even with continuous flow systems there are strong reasons not to use smoothing methods, there may well be important events in the process where information will be critically distorted by smoothing, an example of such an event may be the activation of pumps that are responding to unusual occurrences such as a thermal runaway. Smoothing such data will lead to the effects of such activation being spread both forward and backwards in the time domain, an unlikely occurrence in the actual process. There are smoothing methods that smooth only forwards, or are weighted to smooth in the direction of the time line however they are beyond the scope of this document.

The use of regression tools in process analysis is to examine the relationships between the process parameters. These relationships can be used to examine or control a dependent function that could be related to the speed or efficiency of a reaction or some property of the manufactured material. The goal of process analysis is to completely model the process so that the output of the process is under control and meets the required specification at all times. For this to be possible it is normal to

consider the properties of the raw materials used as part of the data used for the process model.

Process analysis techniques such as Shewhart charts and CUSUM charts are used to examine the stability and current performance of a process. Shewhart charts are used to examine the stability of a process variable with a stable mean and give an indication of the degree and frequency with which a process exceeds its operating parameters. CUSUM charts are used to examine processes where the mean of a process variable is not stable and can be used to look at reasons why the mean may have changed. CUSUM charts also play a role in examining collinearity between variables, trends can be examined to see if they occur in the same manner in other parameters of the process.

1.10. Statistical Process Control

Statistical process control is a situation where the process under consideration has been successfully modelled and by modifying parameters that are under operator control any change in the required performance or output of a process can be modified by the appropriate changes to the process parameters.

1.10.1. Correlation Coefficient

The correlation coefficient [30] is a scaled version of covariance, and is also known as the product moment correlation coefficient, which is the product moment of two vectors about their means. Variance is calculated from equation 1.51,

$$s^2 = \sum \frac{\left(x - \bar{x}\right)^2}{\left(n-1\right)}, \tag{1.51}$$

this is related to the covariance by equation 1.52,

$$COV = \frac{1}{n-1}\sum\left(y-\bar{y}\right)\left(x-\bar{x}\right), \tag{1.52}$$

the covariance can vary between $-\infty$ & $+\infty$, and is of little use in comparing the relationship between groups of different lines, a scaled version of this number would give an indication of the relative relationship between different pair of vectors, and can be seen in equation 1.53

$$r = \frac{\sum_i\left(\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)\right)}{\left(\left[\sum_i\left(x_i-\bar{x}\right)^2\right]\left[\sum_i\left(y_i-\bar{y}\right)^2\right]\right)^{0.5}}, \tag{1.53}$$

the correlation coefficient varies between –1 & 1, and gives an indication of the relationship between two vectors. The correlation term can be misleading, indicating a strong or weak relationship where none exists; a correlation term should never be used without a visual inspection of the data.

1.10.2. ANOVA

ANOVA [30] is analysis is variance, ANOVAs are very useful tools for making comparisons between data matrices and can be used to consider the differences in error between different groups of data. ANOVA calculations are used to compare means to determine whether any significant differences that occurred. ANOVAs can be used to compare more than two means, and are useful in distinguishing between different sources of error or variation. As an example and ANOVA can be used to examine the difference between repeatability and reproducibility for an experiment.

1.10.3. CUSUM Charts

The cumulative summation of variation [36, 37] from the mean of a data set is a useful method of monitoring the way in which the underlying trend of a process varies

over time. CUSUM charts are useful to show where the mean of a process has changed.

CUSUMS are calculated by taking the mean of the vector from each point in the vector, and summing each point successively with the points before it, thus in an example using 10 points,

Sample	Sample –Mean	CUSUM
10	-6	-6
15	-1	-7
20	4	-3
20	4	1
20	4	5
25	9	14
13	-3	11
14	-2	9
13	-3	6
10	-6	0
Mean = 16	Mean = 0	

Table 1.1 Example of CUSUM calculations

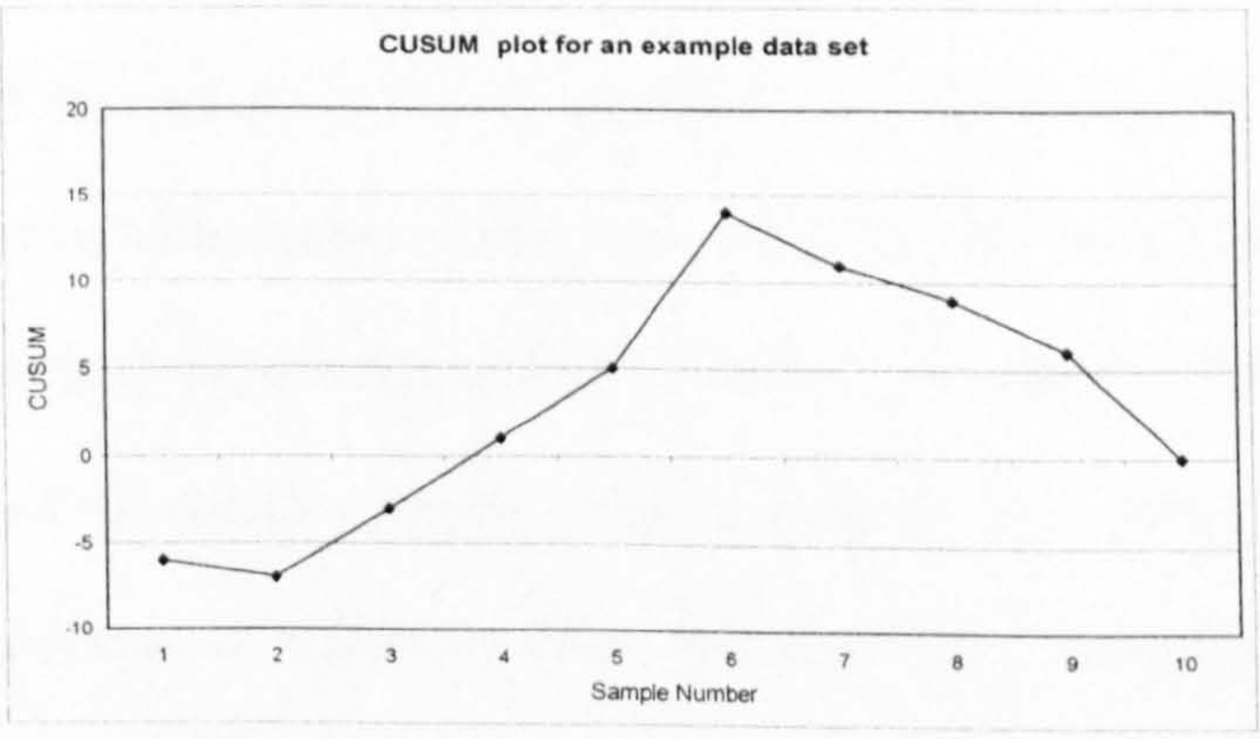


Figure 1.5 Example of a CUSUM plot for random data

The CUSUM shows how a vector changes over time, regions of constant slope indicate time when the values in the vector are constant, flat lines indicate that the vector values are equal to the mean for the vector, positive slopes show periods where the vector values are greater than the mean, and negative slopes show periods where the vector values are lower than the mean. The magnitude of a slope indicates the degree to which the process is deviating from the mean. Uneven portions of the graph

show where the value for the process is changing, changes in slope above a certain degree indicate significant changes in the process. Thus a constant negative slope as an example does not indicate that the process is out of control, just that the process is currently running lower than average.

1.11. Current Research in Chemometrics

Chemometrics research has developed from its early days, when the arguments were about the relative merits of multivariate methods compared with univariate methods, Rao, C.E., [16], to the modern arguments about the merits of different multivariate methods, such as MLR, PCR, PLS, and ridge regression. Comparisons of the various techniques appear fairly regularly, and the general consensus is that while ridge regression is slightly ahead [39] it is the quality and type of data that has a significant bearing on the results of the various methods tried [40, 41]. Kowalski and Seasholtz, wrote a paper outlining available chemometric methods [42], and since then Wold has published several letters and papers describing in detail collections of the methods available at one time and the relative merits of these methods, two good papers were both published in the Journal of Chemometrics and Intelligent Laboratory systems [43, 44], these papers give indications of the current developments in chemometrics, and also consider important issues such as data pre-processing [41], in each case the later papers have a much greater range of techniques to pull from than the previous ones. Other recent developments are again by Wold, orthogonal signal correction (OSC) is a method based on PLS that is designed to replace other smoothing methods for spectra that remove information relating to the Y-block [45] , and this method has been looked at as an approach to remove the traditional problem of transferring a calibration from one instrument to another, with some success [46]. The pros and

cons of each of these methods are well understood. With clean data, well-defined peaks and no overlapping they will each provide very similar solutions. This can be seen in the UV data set in Walmsley's paper [1], where three of the components are relatively error free, and provide good results with any of the methods tried. The improvement seen with the variable selection techniques is due to correcting the rank of the matrix by removing unwanted variables, and also removing the error contribution from these variables. The difference is when the data set becomes more complicated and noisy. The addition of noise quickly reduces the effectiveness of MLR, and methods that are designed to compensate for noise, OSC, PCR, PLS become more useful. If the problem of noise is further compounded by having the component of interest as only a small percentage of the signal then the effectiveness of PCR is reduced, this can be seen in the iron component.

With the increase in computer power cheaply available the complexity of the techniques considered also increases, the ability to collect large data sets (1000 x 1000), three dimensional data sets (1000 x 1000 x 1000) and sets of even higher dimension mean that the methods required to deal with them also become more complex. Parallel factor analysis (Parafac) which was originated by Harshman in the '70's [47, 48, 49] is explained in great detail by Bro [50]. Parafac is a type of Trilinear decomposition (TLD) and deals with large three dimensional data sets by trying to maintain the three dimensional arrangement rather than using an approach based on unfolding the data space. [51, 52]. Three-dimensional matrices are becoming more common, and it is easy to imagine how they are generated, take for example a GC connected to a UV detector, running many samples. With spectra being recorded at time intervals for each sample a three dimensional array is created.

These methods have been compared with standard two dimensional methods, [52] and in general the three way methods often allow easier interpretation of the results, but with a slight penalty in increased error in modelling. Modelling three-dimensional data with a two dimensional method such as PCA or PLS may seam nonsensical but the matrix is simply "unfolded", take an ***I x j x k*** three-dimensional matrix, this would unfold into an ***I x JK*** matrix. The expense with using a 2D method is the increased complexity of the model, a far greater number of factors will be generated than for the 3D method [50], and consequently interpretation can be far more complex.

With variable selection there is some argument as to the usefulness of variable selection [personal communication with McKelvey and Wold, 1998 & 1999, Appendix V], and these arguments are expanded in sections 1.7, 3.1, 3.2, 3.5 and 5.1. For the work reported here the NIPALS algorithm was selected, this algorithm is a general PLS algorithm that is useful for all types of data sets (tall thin, short fat, tall fat, e.t.c.) and discussions on the various different PLS method available can be seen in many of the papers by De Jong, who has published prolifically on this subject [53, 54, 55, 56]. Variable selection has been attempted using many different techniques, there are the fast methods which tend to produce results quickly without iteration, these tend to be based on either selection of variables from the correlation of the predictor variables to the Y-Block [57, 58] or on the magnitude of the coefficients produced during the modelling [59, 60]. These types of methods have two flaws. First with multi-component data they perform very poorly, either wavelengths / coefficients are selected based on there performance with individual components and thus include a surplus of variables, or wavelengths are included based on there

multiple correlation with all the components, this means many of the best variables are excluded. Secondly they make the assumption that either the correlation coefficients or the coefficients of regression give a true indication of the best additions to the model, as this work shows this is not always the case and that low correlation variables, or variables that might have small coefficients can provide important information. This can be seen in the example of the UV data set [1]. A simple correlation approach examining each component individually will select wavelengths centred on each peak for the three clean components, for the iron component there are no correlations greater than ±20%, and either no wavelengths will be selected or the selection will be very poor. Selecting on the basis of the coefficients will have, in many respects, a bigger problem. The first PC or LV extracted from the data set will describe the average of the spectra, and the second will be for the fourth component, copper. These two components provide the balance of the variation in the data set, and coefficients dealing with the iron component will actually rank lower than coefficients for the noise, so may not be included in the model at all.

The next group of methods are based on genetic algorithms (GA) or simulated annealing (SA) These rely on weighted random chance to throw together the correct variables for prediction [61, 62]. While these methods are also effective, they have their own problems. Genetic algorithms take variables from a pool of variables, assess their use, and throw poor variables back into the pool. There is also a chance that good variables will be returned to the pool randomly. This means that these methods are poor solutions to the problem of searching the data space for appropriate variables. While on the surface they may appear similar to the method proposed here

there is the difference that within a given search variables have only one chance of entering a model, and two of getting removed, so unlike the GA and SA useful variables are unlikely to be removed and poor variables are less likely to be included in the model. Both SAs and GAs tend to take longer to calculate as they are not efficient at searching the data space. The method proposed here [2] looks at all the problems associated with the other variable selection techniques. Its main flaw is that although it is faster than the SA and GA methods it is still slow, especially compared with the correlation and coefficient methods. This method does however select variables on the basis of overall improvement to the model, as determined by predictive error, this has advantages over the other methods which either concentrate on individual chemical components or compromise by selecting variables correlated with all components. The different data sets presented here show differing features, clean data with high signal to noise, non-linear data, very noisy data with small responses, and noisy data with good responses, this gives an indication of how the algorithm will perform in a variety of situations, not all of them ideal for this method. With the data set containing non-linear data the linear component was modelled well, at the expense of the other three (non-linear) components. The correlation between the linear component and the concentration information was small however it was modelled by taking information about the variation in the other components to allow the contribution from these components (effectively noise) to be removed.

The problems associated with the development of new tools are that as they become more specialised and sophisticated the knowledge required to not only select the correct method but to implement it correctly increases daily. Simple UV scans on clean samples [1] suit simple methods, such as MLR, however once the complexity of

the system being modelled increases and the signal to noise ratio drops more sophisticated methods are required. In a spectral environment variable selection is nearly always of assistance, removing sections of the data with no information produces no great danger. In a process environment where some of the variables may be only partially (or not at all) understood this is not such a safe option, variables that are removed are not modelled, and a variable that is thought to contain noise or provide no information may contain vital information critical to the operation of the plant when its value changes. As always great care must be taken in selecting a method, and in applying it to any data set

1.12. Software

All work carried out was done on a Pentium computer, running Windows 95. This was chosen for its cost, ease of use and availability. The software chosen was thus limited by this operating system. The University of Hull supports Microsoft Office, and site licenses were available so this was chosen for general word-processing and spreadsheet applications.

1.12.1. Excel

The spreadsheet Excel was used for general manipulation of data, the file format is reasonably transferable, and the Dynamic Data Exchange structure is relatively bug free. Excel was used mainly for its graphing functions, all mathematical calculation were carried out in MATLAB ®.

1.12.2. Word

Microsoft Word is a moderately good word processing package and all the features expected in a modern word processor are present. All reports, papers and this thesis were written with Word.

1.12.3. PowerPoint

PowerPoint is a general drawing tool, with a reasonable group of drawing tools. PowerPoint was sufficient to draw all the diagrams used in this thesis, and was used to prepare any presentations given.

1.13. Maths Software

The maths software used was MATLAB by Mathworks, Version 4.02 in the first two years and version 5.02 in the final year.

For research into chemometric techniques, and for flexible application of chemometric techniques flexible software needs to be used. This limits the choice of software. Suitable software requires the ability to describe exactly how mathematical techniques should be carried out to allow variation in methods such as PLS and PCR. Many of the standard chemometric tools are thus unsuitable.

Pirouette is spectral modelling tool designed to allow factor analysis techniques to be used easily, with a wide range of options for scaling, methods and other data handling, however because of its graphical nature it tend to be relatively slow in calculating results and more critically, it does not allow modification of the maths used to carry out the various techniques. Pirouette does not facilitate the easy export

of results into other formats and does not allow easy access to coefficients produced by calculations, making it inappropriate to the development of chemometric methods.

Unscrambler is another standard chemometric tool, in recent years it has been developed extensively to allow more flexibility however it is still restricted to a relatively small list of factor analysis techniques. Unscrambler is also flawed in many respects due to bugs in the coding, possibly due to the speed of recent developments. Although the graphical interface it very different to Pirouette it is otherwise very similar in terms of modifications to code. Unscrambler does have a rudimentary scripting tool, however it is not very flexible and will not allow modification to the code used to carry out the calculations.

Spectracalc is an old chemometrics package, and while its maths tools for calculating results do allow modification to is greatly hindered by its user interface, manipulating raw data is difficult, and its design as a spectral tool is quite rigid making its use for other reasons difficult.

These problems also persist in software developed by equipment manufacturers for analysis of data produced by specific spectrometers, most are capable of carrying out chemometric analysis of data to a reasonable standard but are not useful for varying methods or data not produced on the machine they were developed for, they will not be discussed further.

1.13.1. Mathcad & Mathmatica

Mathcad and Mathmatica are both tools useful for writing reports using maths. They both carry out scripted math formulae and a powerful programs for calculating the

results of mathematical equations. They are both principally built around their user interface and are designed to allow the easy inclusion of mathematical equations and results into documents. This is the main reason why they are inferior to MATLAB ® for this work. Both tools carry out calculations from within the documents they are written in, displaying the results within the reports themselves. This makes both programs very difficult to use with large amounts of data, effectively limiting their use with spectral or process analysis. They are also both based around building equations from standard mathematical tools which can make it difficult to script the equations needed for complex matrix manipulation.

1.13.2. MATLAB ®

MATLAB ® is a scripting language for maths, particularly matrix manipulation. MATLAB ® can be used in both a command line interface mode, where operations are carried out on matrices directly command by command and in a batch file mode, where strings of commands can be written for be followed in sequence to allow more complex tasks to be carried out. MATLAB ® has been designed specifically to process matrix maths, and as such is significantly faster than most other applications in carrying out these tasks. MATLAB ® has a large array of built in functions, however these form the building blocks to construct other functions. Groups of specially written functions are known as toolboxes, and normally have a focus on a particular field, Mathworks has written toolboxes for neural networks, chemometrics and statistics among others. While the chemometrics toolbox is useful, a toolbox written by another company is the one principally used for this research. The PLS Toolbox, by Eigenvector Research contains a large group of tools specifically for carrying out chemometric calculations and they are better organised and designed than the ones in the Mathworks toolbox.

MATLAB ® is a batch processing tool, most of the tools in a toolbox are written as "m" files, which are flat ACSII files containing commands. M files can be written to carry out most functions, and it is "m" files that are used to carry out the steps required for the development outlined in this thesis.

1.14. Intrasite Gel

1.14.1. Confidentiality

Intrasite is a commercial product produced by Smith and Nephew, certain aspects of the gel and its manufacture cannot be published. It should be noted that the following aspects are omitted from this thesis for these reasons.

1. The dry polymer that is used to make Intrasite Gel is manufactured by another company; that company requires that their name not be published.
2. Specific details of the specifications of the dried polymer may not be published.
3. Specific details of the gels manufacture may not be published, including a description of the exact formulation and specifications of manufacture.
4. One of the variables used in the analysis of Intrasite will be referred to as SC1, this variable is an important parameter to the properties and manufacture of Intrasite Gel and contains information that is commercially sensitive.
5. Detailed examination of the dried polymer is not permitted.
6. Reverse engineering of the dried polymer is not permitted.

1.14.2. Introduction to Intrasite Gel

Intrasite Gel is the product name given to a specific formulation of a carboxymethyl cellulose hydrogel (the structure can be seen in Appendix I). Intrasite Gel is used to assist in the healing of severe wounds, usually severe lacerations. The gel is a viscous paste and is packed into the wound once the wound has been cleaned. Carboxymethyl cellulose gels, particularly Intrasite gel, act in several ways in a wound:-

1. Acts as a barrier to prevent micro-organisms from entering the wound
2. Maintains a constant level of moisture (moist wound healing)
3. Assists in *sloughing* where dead cells are removed by the body
4. Assists in *granulation* when new skin cells form.

The main marketing feature of the product is its ability to maintain moisture equilibrium within a wound. The Gel is sterile but contains no drugs or medication of any kind.

Intrasite gel was originally purchased by Smith & Nephew as Sherisorbe from Sheerings Ag. The gel has two basic formulations, a starch based polymer (originally Sherisorbe) which is the original form of the gel, and a carboxymethyl cellulose polymer (Akucell X181) which is the more modern formulation. The starch based polymer is still produced but in a reduced quantity. Hydrogel is produced from the dry powdered carboxymethyl cellulose polymer which is produced in bulk approximately once a year by an outside company. The dry powdered polymer is

then made up in water in smaller batches as required. Originally there were several formulations of the gel for export to different countries however these have been merged into the one formulation over the years.

The dried polymer is delivered approximately once a year, this is termed a bulk batch, the polymer is then made up into the gel about once a week, and that batch is packaged and sterilised on a daily basis.

Although each bulk polymer batch conforms to the same set of standards there are significant differences in the properties of individual batches made up from different bulk polymer batches, thus there is variation between each individual batch and a greater variation between batches made up from different bulk polymer batches. The specifications for the bulk polymer batches are very broad, and the only information supplied by the manufacturer is that the batch conforms to the specifications, no other information is recorded so bulk batch variation in properties cannot be used to assist in building a global model for this polymer.

Once made up in water the gel is packaged and sterilised. There are five standard types of packaging, 10ml & 20ml sachets, and 8ml, 15ml and 20ml "appli-packs". Appli-packs are bulb shaped dispensers with a nozzle that can be used with one hand. Sterilisation occurs after packaging and is carried out on a small batches. Once sterilised, samples are taken to the laboratory for analysis.

The carboxymethyl cellulose polymer that forms up Hydrogel is highly absorbent, this absorbency is the basis of the useful properties of the gel. The use of this type of

product in medicine is relatively recent, there has been little research into this type of application of absorptive gels. Current knowledge suggests that its usefulness is based entirely on is physical absorptive properties, maintaining a moisture equilibrium with the wound.

1.14.3. Fluid Absorption

The fluid absorption test is one of the key test carried out on Intrasite Gel, the properties of the gel that it is marketed on are based around the fluid absorption characteristics of the gel. The test is a simple test, as described in the test procedure sheet, [63]. Although this test takes 24 hours to allow for equilibrium it is known that this period is not sufficient for true equilibrium. Equilibrium time is based on the vigour and length of initial mixing with saline solution. It is also known that with centrifuging more liquid can be measured above the gel layer than was added for the test. This is evidence that the moisture retention is absorption, not adsorption. The gel is also known to be soluble in both water and saline solution, the solubility is variable, and is normally between 20% and 40% by weight.

1.14.4. Intrasite Tests

During laboratory analysis seven tests are carried out, the hydrogel must meet the specification for all of them;

1. Identification of propylene glycol, this is to determine that the gel has been made up in propylene glycol [64].

2. Identification of carboxymethyl cellulose, this test ensures that the material being tested performs to the chemical characteristics of carboxymethyl cellulose [65].

3. pH, the pH of the material is critical since it is intended for medical use on open wounds, if the the pH is not within the specified limits then there can be severe reactions to the application of the gel [66].

4. Elasticity, this is a rheological test, the elasticity of the gel affects the ease with which the gel can be applied [67]

5. Viscosity, this is a rheological test, the viscosity of the material affects whether the material remains in the wound [68].

6. SC1: [69]

7. Fluid Absorption, one of the key properties of Intrasite gel is its ability to maintain an equilibrium of the moisture present in a wound, this test is intended to have an indication of the fluid transfer that occurs between the wound and the gel [63]

2 Experimental

The Experimental section is divided into two parts, the first deals with the development of the variable selection PLS algorithm, outlining the five important stages that occurred during the production of the final algorithm. The second part of the experimental deals with the examination of Intrasite Gel, initially looking at the measurements made in the laboratory as a whole, then focusing on the fluid absorption parameter.

2.1 Variable Selection PLS

The variable selection section of this thesis covers the development of an iterative method to select variables from a data matrix based on their ability to improve the predictive ability of the generated model. This development took place in five main stages, outlining the important decisions that were made during the development of the final algorithm to carry out variable selection using PLS.

In the following text the notation is as follows, all indexing is relative to the matrix being indexed unless stated otherwise.

---

j is the number of rows

i is the number of columns

k is the number of variables

r is the number of components being predicted

q is the number of samples

h is the loop number

N is the matrix of actual values

P is the matrix of predicted values

T is matrix of training data

V is matrix of validation data

C^1 is matrix of training concentration information

C^2 is matrix of validation concentration information

S is matrix of selected variables (initially is empty)

s is the number of selected variables

Model prediction in this section is based on the Predicted Residual Error Sum of Squares (PRESS) (equation 2.1),

$$PRESS = \sum_{i=1}^{r} \sum_{j=1}^{q} (N_{ij} - P_{ij})^2 \qquad \textbf{(2.1)}$$

this is calculated from a validation set. After each variable is added, the PLS model is built using the training set, it is then used to predict on the validation set, which is completely independent of the training process.

2.1.1 Data Sets

Three data sets were used, one UV spectra data set [1], and two synthetic data sets.

2.1.1.1 The UV Data set [1]

The data consisted of 52 spectra of 4 transition metal ions (Fe, Co, Ni and Cu) run on a Varian DMS90 UV/VIS spectrometer, over the 190-890nm range, at a varied concentration ranges. The entire spectra range was digitised, with a data spacing of 3.3nm, giving 211 spectral points. The concentration of the iron was miscalculated at the sample preparation stage and is present only at the limit of detection. The iron response has a large amount of error and calibration is difficult.

The data was split to give 40 training samples and 12 'unknowns'

2.1.1.2 Synthetic Data Set 1

Sixty samples of forty points with four overlapping components of random concentration. 4% normally distributed random noise added to each data point. The non-linear response components were two squared terms, and a logarithmic term.

All the components are have a normally distributed concentration range. This data set was produced in MATLAB ® using a Gaussian curve generator and a random number generator. The concentration of each component in any one spectra was determined using a linear random number sequence. The noise added to the spectra was generated using a normal distributed random number generator.

2.1.1.3 Synthetic Data Set 2

Eighty samples of two hundred and fifty points with four overlapping components of random concentration. Up to 10% randomly distributed noise added to each data point. All the components are linear and normal. The concentration of each component in any one spectra was determined using a linear random number sequence. The noise added to the spectra was generated using a normal distributed random number generator.

2.1.1.4 Data Pre-treatment

Data was treated using autoscaling, producing data sets where the variance in the variables has a mean of zero. Autoscaling was determined by ordinary PLS to give the best response to the UV data set. Mean centring and range scaling were also tried.

2.1.2 Single Addition Mode, SVA-PLS

The first attempt at prediction based variable selection started with a single randomly selected variable, a PLS model was calculated using this variable and one latent vector. Another variable from the pool of remaining variables is then selected at random and added into the model. The PLS model is recalculated and the optimum number of latent vectors for this model determined. An improvement in the model leads to the variable being selected, no improvement or a worse model and the variable is removed from both the model and the pool. This is in contrast to many genetic algorithm methods where the variable is returned to the pool of unselected variables. The random selection of an individual variable from the pool continues, each time the number of latent vectors is re-calculated. This process stops once all the variables have been added into the model and been either selected or discarded. Once the selection of variables has been determined the selected variables are recorded, together with the value for the PRESS those selected variables produced. The whole process is repeated.

The order in which a variable is added into the model affects how that variable changes the PRESS produced. This is one of the key reasons why selecting variables

by either their correlation with the determinant or the magnitude of their loading produces models that can lack robustness and include excess variables. This is also why the modelling process must be repeated. The variables that produce the lowest PRESS might not be found on the first attempt.

The algorithm that this procedure follows can be seen below

Calculate PLS using T_{qk} and C^1_{qr}

Predict using V_{qk} and C^2_{qr}

Determine correct number of latent vectors for minimum press, l

$$BASEPRESS = \sum_{i=1}^{r} \sum_{j=1}^{q} (N_{ij} - P_{ij})^2$$

select 1 random variable and put it into S

Start loop (h)

Calculate PLS using $[S_{qs}\ T_h]$ and C^1_{qr}, using l latent vectors

Predict using $[S_{qs}\ V_h]$ and C^2_{qr}

Determine correct number of latent vectors for minimum press, l

$$PRESS = \sum_{i=1}^{r} \sum_{j=1}^{q} (N_{ij} - P_{ij})^2$$

If BASEPRESS > PRESS then add T_h to S and BASEPRESS = PRESS

Stop loop when h is equal to k

Loop

Record the variables in S and the final value for BASEPRESS, and l

Repeat the whole process at least $2 * \sqrt{k}$ times

Determine the iteration with the lowest BASEPRESS, these are the variables to keep, together with l

---

Although this algorithm produced a reasonable model there were several flaws. The main issue was that due to the algorithm of necessity starting with a single latent vector the error in the model was too high. The result of this is that the first variables presented to the model, up to the number equal to the number of latent vectors required to describe the model, will be accepted as improving the model. This process leads to a poorer model than could be produced without the first few variables in the model, as these variables invariably had little or no information to add.

This problem is compounded when several dependent variables are being calibrated simultaneously. The models produced tend to be very unstable with a very wide spread of potential PRESS results.

2.1.3 Multiple Variable Addition Single Pass MVA-PLS

Taking into account the problems associated with the first attempt the method was adjusted. Initially the optimum number of latent vectors is determined using block validation PRESS. The number of latent vectors is taken as the number of starting variables, chosen at random. Variables are then added in blocks equal to this number, for reasons of ease of programming.

The blocks added are considered for addition or removal as a whole, either the whole block is added or the whole block is removed.

These changes produced better models than the first algorithm, although this method produced more stable models also included a great number of surplus variables.

Calculate PLS using T_{qk} and C^1_{qr}

Predict using V_{qk} and C^2_{qr}

Determine correct number of latent vectors for minimum press, l

$$BASEPRESS = \sum_{i=1}^{r} \sum_{j=1}^{q} (N_{ij} - P_{ij})^2$$

select l random variables and put them into S

Start loop (h:l); (increase the value of h by l each iteration)

Calculate PLS using $[S_{qs}\ T_{h-h+l}]$ and C^1_{qr}, using l latent vectors

Predict using $[S_{qs}\ V_{h-h+l}]$ and C^2_{qr}

$$PRESS = \sum_{i=1}^{r} \sum_{j=1}^{q} (N_{ij} - P_{ij})^2$$

If BASEPRESS > PRESS then add T_h to S and BASEPRESS = PRESS

Stop loop when h is equal to k

Loop

Record the variables in S and the final value for BASEPRESS, and l

Repeat the whole process at least $2 * \sqrt{k}$ times

Determine the iteration with the lowest BASEPRESS, these are the variables to keep, together with l

2.1.4 Single Variable Addition, Single Variable Removal, SVA-SVR-PLS

By adding variables initially as a group and then singly the flaws in the first two attempts were removed. However the models produced contained too many variables.

Rather than try and restrict the addition of variables, the next approach considered was to remove unwanted variables. This is done by taking the selected variables, then running through the addition procedure in reverse, remove one variable, test the model, if the model is worse, add the variable back in to the model, otherwise, discard it.

Calculate PLS using T_{qk} and C^1_{qr}

Predict using V_{qk} and C^2_{qr}

Determine correct number of latent vectors for minimum press, l

$$BASEPRESS = \sum_{i=1}^{r} \sum_{j=1}^{q} (N_{ij} - P_{ij})^2$$

select l random variables and put them into S

Start loop (h)

Calculate PLS using $[S_{qs}\ T_h]$ and C^1_{qr}, using l latent vectors

Predict using $[S_{qs}\ V_h]$ and C^2_{qr}

$$PRESS = \sum_{i=1}^{r} \sum_{j=1}^{q} (N_{ij} - P_{ij})^2$$

If BASEPRESS > PRESS then add T_h to S and BASEPRESS = PRESS

Stop loop when h is equal to k

Loop

Record the variables in S and the final value for BASEPRESS, and l

Set T equal to S, Set S to empty.

Start loop (h)

Calculate PLS using $[S_{qs}\ T_{k-h}]$ and C^1_{qr}, using l latent vectors

Predict using $[S_{qs}\ V_{k-h}]$ and C^2_{qr}

$$PRESS = \sum_{i=1}^{r} \sum_{j=1}^{q} (N_{ij} - P_{ij})^2$$

If BASEPRESS > PRESS then add T_{h+1} to S and BASEPRESS = PRESS

Stop loop when h is equal to $k-1$

Record the variables in S and the final value for BASEPRESS, and l

Loop

Repeat the whole process at least $2 * \sqrt{k}$ times

Determine the iteration with the lowest BASEPRESS, these are the variables to keep, together with l

---

2.1.5 Single Variable Removal, SVR-PLS

By using the addition mode followed by removal mode the model was improved significantly. This raised the possibility that the modelling process might be superior if just the removal mode is used. Instead of adding variables into a group, the whole spectra could be taken and then variable could be removed individually.

Calculate PLS using T_{qk} and C^{l}_{qr}

Predict using V_{qk} and C^{2}_{qr}

Determine correct number of latent vectors for minimum press, l

$$BASEPRESS = \sum_{i=1}^{r} \sum_{j=1}^{q} (N_{ij} - P_{ij})^2$$

Start loop (h)

Calculate PLS using $[S_{qs}\ T_{k-h}]$ and C^{l}_{qr}, using l latent vectors

Predict using $[S_{qs}\ V_{k-h}]$ and C^{2}_{qr}

$$PRESS = \sum_{i=1}^{r} \sum_{j=1}^{q} (N_{ij} - P_{ij})^2$$

If BASEPRESS > PRESS then add T_{h+1} to S and BASEPRESS = PRESS

Stop loop when ***h*** is equal to ***k-1***

Loop

Record the variables in ***S*** and the final value for BASEPRESS, and ***l***

Repeat the whole process at least $2 * \sqrt{k}$ times

Determine the iteration with the lowest BASEPRESS, these are the variables to keep, together with ***l***

2.1.6 Single Variable Removal Duel Pass with Squashing Function, SVR-DP-PLS

The removal mode works better than the methods tried before, however due to the random order of selection, and the issues of co-linearity, surplus variables may still remain. Once the unwanted variables have been removed once, a second pass is made through the algorithm, starting again with the shuffling of the variables. The second pass removes variables that were added in, then found to be inferior to variables already added.

Calculate PLS using T_{qk} and C^1_{qr}

Predict using V_{qk} and C^2_{qr}

Determine correct number of latent vectors for minimum press, ***l***

$$BASEPRESS = \sum_{i=1}^{r} \sum_{j=1}^{q} (N_{ij} - P_{ij})^2$$

Start loop (***h***)

Calculate PLS using $[S_{qs}\ T_{k-h}]$ and C^1_{qr}, using ***l*** latent vectors

Predict using $[S_{qs}\ V_{k-h}]$ and C^2_{qr}

$$PRESS = \sum_{i=1}^{r} \sum_{j=1}^{q} (N_{ij} - P_{ij})^2$$

If BASEPRESS > PRESS then add T_{h+1} to S and BASEPRESS = PRESS

Stop loop when h is equal to k-1

Loop

Record the variables in S and the final value for BASEPRESS, and l

Set T equal to S, Set S to empty.

Start loop (h)

Calculate PLS using $[S_{qs}\ T_{k-h}]$ and C^{l}_{qr}, using l latent vectors

Predict using $[S_{qs}\ V_{k-h}]$ and C^{2}_{qr}

$$PRESS = \sum_{i=1}^{r} \sum_{j=1}^{q} (N_{ij} - P_{ij})^2$$

If BASEPRESS > PRESS then add T_{h+1} to S and BASEPRESS = PRESS

Stop loop when h is equal to k

Loop

Record the variables in S and the final value for BASEPRESS, and l

Repeat the whole process at least $2 * \sqrt{k}$ times

Determine the iteration with the lowest BASEPRESS, these are the variables to keep, together with l

---

2.1.6.1 Squash Function

The current method still has a tendency to include too many variables into a model, a large number on the first pass, a significantly smaller number on the second pass. This can be adjusted with a squashing function. The squashing function (known mathematically as a cost function) is active when the calculation is performed as to whether a variable is included into a model. The standard calculation is;

If BASEPRESS > PRESS then add T_h to S and BASEPRESS = PRESS

This can be modified, here, χ is the squashing value.

If BASEPRESS > (PRESS * χ) then add T_h to S and BASEPRESS = PRESS

In this case the PRESS value must smaller than the BASEPRESS by a factor of χ.

Thus, a value for χ smaller than 1 will have the effect of reducing the number of variables added to a model and a value greater than 1 will increase the number of variables added.

The values for the squashing function need to be chosen with care for each data set modelled. There are two squashing functions, the first controlling addition of variables during the first pass. The second squashing function controls the addition of variables during the second pass.

The squashing function can greatly affect the quality of the model produced by the algorithm, correctly used the function will produce a more stable model. Incorrect balanced and magnitude of the two squashing functions leads to unstable modelling as either too many or too few variables are included in the model.

2.1.7 Selected Variables Histograms

The variables selected for each iteration are different for each iteration when spectral data is analysed (or any data with a high degree of collinearity). The selection however will be centred on sections of the spectra. By collecting the selected variables of each iteration and plotting them as a histogram of frequency overlaying the spectra itself key information about important sections of the spectra can be

gained. This information can be used in several ways, first the information is useful for determining which sections of the spectra are useful. Secondly this information could be used in a variable ranking method to alter the way in which variables are selected, or to weight the value given to a particular wavelength. This would be useful where variable selection as outlined here is not providing the sort of model required by the data, either due to large amounts of interference or some consideration to differing standard of robustness.

2.1.8 Number of Iterations

Each iteration of any of the above methods produces a different value for the minimum PRESS under normal circumstances. If the data set is composed of spectra then there will almost always be variation in the variables selected. A data set from a process may be different from this as there may be only a small number of variables within the data set that provide information for a calibration. The difficulty is in determining the number of iterations to use in the modelling. The more iterations carried out the lower the PRESS produced is likely to be. This takes longer to compute. One way of examining the problem is to carry out the iterations X times, recording the value for the PRESS on each occasion. The values for the PRESS can be charted as a histogram using appropriate bins, and a χ^2 test carried out to determine the shape of the peak being generated. From this the chance of producing a lower PRESS value can be easily calculated, and the modelling can be stopped when the chance of finding a lower PRESS value falls below a pre-determined limit.

2.1.9 Final MATLAB ® Code

The final MATLAB ® code can be found in Appendix II, this code requires MATLAB ® 5.2.1 and the PLS Toolbox 1.

2.2 *Intrasite Gel*

Intrasite Gel is a registered medical device produced by Smith & Nephew Hull, and is used in the care of wounds. Intrasite Gel is used to treat necrotic, sloughy, and granulating wounds, in its role in wound care Intrasite Gel provides a moist wound environment which aids healing. The measurements made on the material in Smith & Nephew's laboratories were examined in this thesis to look at the stability of the production, and there was a detailed examination of the fluid absorption measurement to consider whether the test for fluid absorption should be replaced or modified.

2.2.1 Initial Data

The Intrasite data set is composed of two sections, the product analysis results and the sterilisation parameters. The data for Intrasite gel dates back to January 1993, when the current formulation was initially developed, and continues to the present day. In this work, only data up to December 1997 was included in this analysis. Prior to January 1993 the polymer was starch based and as the starch based polymer was not covered by this research, this data has not been considered. The historical data before January 1996 is held on paper records, this information was typed into Microsoft Excel to allow its inclusion into the research.

After batch production the gel is tested twice, initially with the SC1 test as soon as the gel is produced. If the test meets specification then the batch is then packaged into its delivery unit, either an apli-pack (8ml, 15ml, 25ml) [70] or a sachet (10ml, 20ml) [71]. The Apli-packs are hard plastic dispensers designed for one-handed use, as is often convenient in the environment where Intrasite Gel is often used. After the gel

has been packaged it is sterilised in batches of between 400 and 8000 units (a unit being either an apli-pack or a sachet) depending on the unit size.

Sterilisation occurs according to the F_0 procedure, where F_0 is the integral of the time the batch spends above 121°C. Random samples are taken from the sterilised batch and sent for analysis. Part of the sample will remain in storage to allow re-testing later if required. F_0 is used as an indication of biological activity, items sterilised to the F_0 standard are assumed sterile for the purposes of medical devices and dressings. The F_0 value must exceed 22 for the item to be considered sterile. Further information about F_0 can be found in Appendix I.

During laboratory analysis seven tests are carried out, the hydrogel must meet the specification for all of them;

1. Identification of propylene glycol, this is to determine that the gel has been made up in propylene glycol [64]
2. Identification of carboxymethyl cellulose, this test ensures that the material being tests performs to the chemical characteristics of carboxymethyl cellulose [65]
3. pH, the pH of the material is critical since it is intended for medical use on open wounds, if the the pH is not within the specified limits then there can be severe reactions to the application of the gel [66]
4. Elasticity, this is a rheological test, the elasticity of the gel affects the ease with which the gel can be applied [67]
5. Viscosity, this is a rheological test, the viscosity of the material affects whether the material remains in the wound [68]
6. SC1 [69]

7. Fluid Absorption, one of the key properties of Intrasite gel is its ability to maintain an equilibrium of the moisture present in a wound, this test is intended to have an indication of the fluid transfer that occurs between the wound and the gel [63]

For historical reasons, test number 6 [69], the test SC1 is carried out twice, though this is not part of the test method. The results for only one of these tests (the first one listed) has been used as there is negligible difference between them.

Intrasite Gel was purchased as a complete product, and the product arrived with its registration. However, Fluid absorption and viscosity coefficient measurements were added to the list of required measurements. Fluid absorption was added when Intrasite gel was acquired, viscosity coefficient was added in February 1994.

The variables recorded for the sterilisation of Intrasite gel are recorded directly from the sterilisation equipment (Table 2.1) , they are recorded on paper. One years worth of data, January 5th 1995 through December 18th, was entered into a spreadsheet for an initial examination. The variables recorded are seen in table 2.1

There were a number of occasions during the period examined that the batch failed to sterilise properly, when this occurred the batch was simply re-sterilised. Unfortunately although it is thought that the total time above 121°C is important to the properties of the Intrasite gel, the instrument reading of the sterilisation attempts before the batch was successfully sterilised were not recorded.

Graphs of these data sets can be seen in the appendix. The sterilisation data set was collected at a later data than the analysis data set.

1	Quantity	Number of units in a batch
2	Heat up Time	The time required to reach 121°C
3	Pressure	the Pressure in the steriliser
4	Minimum Temperature	The lowest temperature the batch reached once F_0 had been reached. If the temperature dropped below 121°C for any reason before sterilisation was complete then sterilisation had to be repeated
5	Maximum Temperature	The highest temperature reached during sterilisation
6	Hold Time	The time the batch was held at 121°C or greater
7	Cool Time	Once sterilisation has occurred the batch is allowed to cool slowly
8	F_0	The integral of the temperature above 121°C

Table 2.1 Variables Taken from the Sterilisation Process of Intrasite Gel

2.2.1.1 Intrasite Experiment 1

The statistics of the data set was examined. The historical analysis data and the sterilisation data where typed into the computer by hand. The data set was checked initially by examining the spread of data, values outside the expected range were considered outliers and checked against the source and corrected. Where the source contained the same value and the value was found to lie outside the possible range of values, that data point was removed. All incomplete rows were deleted from the data set, with process data of such a large quantity there was seen no value in imputing or interpolating values.

The examination of the analysis data set was split into four parts. Initially the data was examined as a whole and a global model was looked for. The data was considered when split by bulk batch. The data for the year 1997 was examined. Finally the data set was split according to the analyst that carried out the testing. Information about the analyst that carried out the testing was only available for data after December 1996. The work concerning the variation in analyst results is reported in a report for Smith & Nephew that can be seen in Appendix VII

While it was hoped that a general model for the analysis of Intrasite gel might be developed, it was recognised that the between bulk batch variation might make this impossible. This is the reason why the data set was examined on an individual bulk batch basis as well as using the entire data set. If separate models for the fluid absorption could be developed, some form of transfer function might be developed that would allow the model to be transferred between bulk polymer batches.

2.2.2 Initial Examination

The fluid absorption test [63], by the settling volume method, was an essential issue in the initial project. This test is flawed in several respects, first it does not really represent the environment that the gel would be used in, the test involves the examination of the saturated gel, not the gel in equilibrium with a moist environment, and second the test method contains a high degree of error due to two factors, first the equipment used does not allow accurate measurement of results, and second the gel is up to 40% soluble in water. This solubility is known to vary constantly from batch to batch, and is not constant for one bulk batch of polymer. There are a variety of reasons why this might be the case, most likely due to particle size variation due to

the bulk batch being incompletely homogenised. The fluid absorption test [63] is examined in detail, including details of the solubility of the gel in QA3174 [72], and in QGM\137 [73].

2.2.2.1 Normality, Intrasite Experiment 2

Histograms were calculated for the frequency of values in the analysis variables. The expected distributions were also calculated for each variable, based on the population mean and standard deviation of each variable. For the full data sets the variables were found to depart from normal distribution. All the variables show evidence of a binomial distribution. A non-normal distribution might be expected when the product is produced from differing batches of starting material.

Histograms of value distribution were then calculated for the variables, taking data collected during 1996 and 1997, and the results compared with the expected distributions, calculated on the new population mean and standard deviation. The results show that the variables now follow a normal distribution. The Chi-squared test is used to examine whether the distribution of values is normal. In all the variables the distribution of values taken from 1996 and 1997 show normal behaviour.

2.2.2.2 Correlation, Intrasite Experiment 3

The purpose of the initial examination was to determine whether the test could be replaced with a simple calibration based on the other analysis variables. This would allow prediction of the result of the settling volume method based on parameters that are measured with a greater precision and accuracy. The first examination was of the simple correlation between the analysis variables to determine if the fluid absorption variable closely matched any of the other recorded parameters. The correlation showed a moderate level of correlation between fluid absorption and the solids content, and a slightly better one between fluid absorption and the viscosity coefficient.

The correlation was then examined for sections of the variables where the distribution was known to be normal. Two sections where taken, one composed of data from the years 1995 through 1997, and the other section was the data from 1997 only. The results showed little difference compared to the results taken from the full data set.

2.2.2.3 Regression Modelling, Intrasite Experiment 4

Despite the high correlation between the viscosity coefficient and the solids content, the relatively low correlation between fluid absorption and the other variables, and the non-normal distribution of the full data set, an MLR model was attempted. This was done for the full data set and two bulk batches, one from 1996, the other from 1997, where the distribution is normal.

All the models produced showed more error than might be expected from the levels of error present in the measurements. The modelling was repeated with Projected Latent Structures (PLS) in order to reduce the error of modelling.

2.2.3 Intrasite Experiment 5, Inclusion of the sterilisation data

The MLR and PLS models while poor did show a relationship between fluid absorption and the other variables. With carefully selected process measurements there will be little or no correlation between recorded variables, thus the required information to improve the model might be missing from the selected variables. Data from a different source could provide an improved model. The sterilisation data was the only other source of information available about Intrasite gel, so this was collected. No attempt was made to transfer all the historical sterilisation data into spreadsheet format as there was no evidence that this contained any useful information. If the sterilisation data proved useful then the remaining data could be transferred at a later date.

The sterilisation data was added directly to the data set that already existed, taking only those batches that matched. On a number of occasions the data for a particular batch might be recorded twice, usually from a re-test, occasionally from a typographical error in the batch numbering. In cases where the discrepancy could not be resolved the first instance of a batch was taken and the other instance was deleted.

Correlation analysis was done, although the results showed that there was very poor correlation between the fluid absorption and the new variables. MLR and PLS

calibrations were carried out to determine whether the model error had been reduced by the addition of the new variables. It is possible that some of the solubility of Intrasite gel in water might be explained by the conditions during sterilisation, and thus the model might be improved. The models produced were still very poor, showing at least 40% error.

2.2.4 Intrasite Experiment 6, Effect of pH on Measured Fluid Absorption

The equipment used to carry out the settling volume test is one of the major reasons for the large error in the measurement, however the solubility of the hydrogel in the saline solution is of great importance as well. In this and all other experiments calling for saline solution, saline solution means a solution in pure water of 0.142 mol l^{-1} sodium chloride and 0.0025 mol l^{-1} calcium chloride. The effect of the pH of the saline solution was investigated to determine whether there is any noticeable effect on the result of the settling volume test. The solubility of a material is affected by the ionic strength of the solution into which is dissolving, and the pH of the solution reflects this. A series of experiments were carried out to examine the effect of the pH. Due to sampling limitations, it is only possible to examine about 100 ml of Intrasite gel. If the sample is greater than this then archive samples must be used and this is unacceptable under the guidelines for the registration of medical devices. The guidelines call for samples of the batches analyses to be held in storage for at least five years in case re-testing is required. The experiment to examine the effect of pH

was intended to have five replicates for each of six levels of the pH, spanning the pH range limits described in the specifications. It was not possible to take the 300ml required for this series of test from one batch so the experiment was run with 5 different batches, each batch tested at six different pH.

The results of the test were examined for between and within batch variation using ANOVA. The variation between the batches was found to be greater than the variation due to pH changes by a significant amount. It was decided that trying to account for the error in the settling volume method was not going to significantly reduce the error in the test.

2.2.5 Examination of Process Control, Intrasite Experiment 7

Fluid absorption was initially examined because the test to measure it is the one considered most flawed, and the results are the least reliable. Smith and Nephew's interest however is wider than that, clearly they wish to ensure that their product is produced to a high and constant standard. Excessive testing to show this is not of value. Smith and Nephew would like to reduce the level of testing they carry out and still be certain that the product they produce is still to the same standard. The results from the analysis of the sample can be used to examine the production of Intrasite gel.

2.2.5.1 CUSUM Charts

In producing a CUSUM chart there is no assumption of a normal distribution. The CUSUMS were plotted for all the whole data set. Solids content, viscosity coefficient

and fluid absorption show approximately the same profile, the pH shows a different profile. The elasticity profile appears to have a profile corresponding to the combined effects of the pH profile and the solids content profile.

The CUSUM profiles were also plotted for the data from the year 1997. These CUSUM profiles showed similar relationships to the profiles from the full data sets. One of the possible causes for changes in population mean shown in the CUSUM charts is the change in analyst. The routine analysis of Intrasite gel is carried out by a undergraduate student on placement with Smith & Nephew, this requires that each year the person carrying out the analysis changes. Periods of holiday and training of new staff affect which analyst carries out the tests. The CUSUMS for the variables were plotted for time periods where the analyst was know and was constant for long periods of time to investigate these this possible influence. The analysts designated 65, 67, 68 and 76 suited the requirements for this examination, each person analysed Intrasite gel for a period of more than six months. The results show that for each analysis the profiles of the variables, particularly viscosity coefficient, elasticity and SC1 show a remarkable correlation. This work is reported in Appendix VII, a report for Smith & Nephew.

The CUSUM charts were used to examine the sampling frequency employed for the analysis of Intrasite gel, the Shewhart charts are not appropriate for this task as there is uncertainty about the exact error present in each measurement. If the process is under control the product can be assumed to be within specification. The minimum level of analysis to determine the process state is all that is needed. By calculating the CUSUM charts for the variables using fewer points than are available the effect of

reduced sampling can be determined. The CUSUM charts for each variable were calculated using every second, fifth, tenth and twentieth point. The profiles present in the charts from the full data set can still clearly be seen, suggesting that the process can be monitored using fewer sampling points.

2.2.6 Reference Data

The models produced so far are not useful as a replacement to the actual settling volume test, so a new approach was considered. A method that measured the fluid absorption accurately and precisely would allow the relationship between fluid absorption and the other variables to be studied. The results from this method could then be related to the results from the settling volume method, and the settling volume results predicted. This new prediction would not contain the error present in the actual measurement. If this approach turn out not to be feasible then the new method could replace the settling volume method as the standard test. Prediction of the settling volume test results from the reference method is required, however the reference method cannot be used directly. The relationship between the reference method and the settling volume test, and the relationship between the reference method and the analysis variables need to be established. This will allow the development of a model predicting the settling volume method,

1. Measure Fluid absorption, or a closely related property
2. Precision
3. Accuracy
4. Reproducibility
5. Ease of measurement
6. Speed of Measurement

7. Low Skill requirements
8. Low Materials and Equipment Cost

Ideally any method would not just measure fluid absorption, but also have some information about the rate at which any equilibrium was reached. Four methods were examined as to their suitability.

The first method considered monitoring the viscosity coefficient of the Intrasite gel as saline solution was added to it. This would show not only the total fluid absorbed by the gel, but also give some indication of the rate at which the fluid was absorbed. This test would satisfy all the requirements except 5 and 6. The measurement would not be particularly easy due to the requirement of measuring the viscosity of a material with a changing volume. This would cause problems with the design of the equipment since the probe would need to remain at constant height relative to the surface of the gel. Early tests showed that the test would also be very slow as the gel would have to reach full equilibrium to each level of saline added or the viscosity coefficient results would be unstable. Further reading suggested other reasons why the test would fail. M. Dolz *et. al.* [74, 75] showed that the viscosity of carboxymethyl cellulose polymers was affected by shear stress, the viscosity was found to reduce by 40 to 50% after five minuets of stirring. Since shear stress would be a factor in mixing the gel once saline had been added and a factor in the test itself the results, the results would be unreliable.

The second method considered was the standard method for examining the fluid absorption of solids gels and foams. The material to be tested is placed under a petri dish and saline is slowly pumped in. The amount of fluid pumped in before the petri

dish leaks is the fluid absorbed by the material being tested. This test would require modification to be used with Intrasite gel, as air would not be displaced in this case, so arrangements would be needed to deal with the expansion of the gel. The flow rate of the saline would need to be considerably slower than standard. Of the requirements, only 1 and 8 are satisfied by this test. The precision, accuracy and reproducibility of the test would be suspect due to difficulty in ensuring that the gel is homogenous beneath the petri dish. The test would require too much preparation, and due to the requirements of equilibrium would take too long. There is a high degree is skill required in setting the test up and monitoring it during the testing period.

The third method is one of the standard methods for measuring the fluid absorption of a material, and is covered by the British Pharmacopoeia Appendix VI, Physical Test Methods. The test is known as the tea bag method where a known mass of the sample of the material to be tested is placed in a semi-permeable membrane, and suspended in a solution, in the case of Intrasite gel, saline solution. After equilibrium has been reached the sample is re-weighed and the fluid absorbed can be calculated. This method was discounted immediately due to the solubility of Intrasite gel [73] in saline solution and water.

The forth Method was referred to as "the Paddington Cup" method, and the Agar Plate method [76]. The Paddington cup method was developed as a method to examine the differences between competitors products and Intrasite gel. The method examined the different properties of the various hydrogels in absorbing and releasing saline solution. To carry out this test a sample of the hydrogel to be measured is left

in contact with a suitable material. To examine the water absorbing properties of the hydrogel the material used is Agar gel, more than one concentration of Agar is used, 1%, 2%, 4% & 6%. For the examination of the water donating properties Gelatine gel was used, at several different concentrations, 10%, 20% and 30%. This range of substrates is used because the different hydrogels produced by the various manufacturers can have markedly different properties, some are intended for slightly different end uses and this will effect the water transfer properties. The span of materials used allows the relative properties of the different gels to be examined.

The main drawback to this method is that it is time consuming, taking approximately four days to complete. The method is also unable to provide any information about the rate at which fluid is transferred. This method however does not posses the flaws evident in the other methods considered, and it measures directly the transfer of fluid between different mediums, relating directly to its end use. This test was selected as the one most likely to provide a useful replacement to the settling volume method despite its long time requirement.

2.2.6.1 The "Paddington Cup" Method

This method has shown its usefulness for the comparison of different hydrogels, the fluid absorption is calculated from the change in weight of either an agar disk of varying concentration or of a gelatine disk of varying concentration.

The work is carried out in a 60ml syringe with the nose cut off, leaving a wide opening with a smooth edge. The plunger of the syringe is withdrawn to leave a suitable volume of space, approximately 30ml. The syringe is weighed. 10g of substrate are introduced into the syringe, the syringe is then sealed and is left to

solidify and equilibrate at 25°C for 24 hours. The syringe is re-weighed. 10g of the hydrogel sample being tested are added, and the syringe weighed again, then sealed. The experiment is left for 48 hours at 25°c. After 48 hours, the syringes are unsealed then weighed again, and the hydrogel removed, care being taken to ensure that the substrate surface is not disturbed. Either the removed hydrogel or the syringe with substrate can then be weighed again, for practical reasons it is easier to weigh the syringe and substrate. The percentage change in weight can be calculated from these measurements and used to give a relative measure of the fluid absorption of the sample.

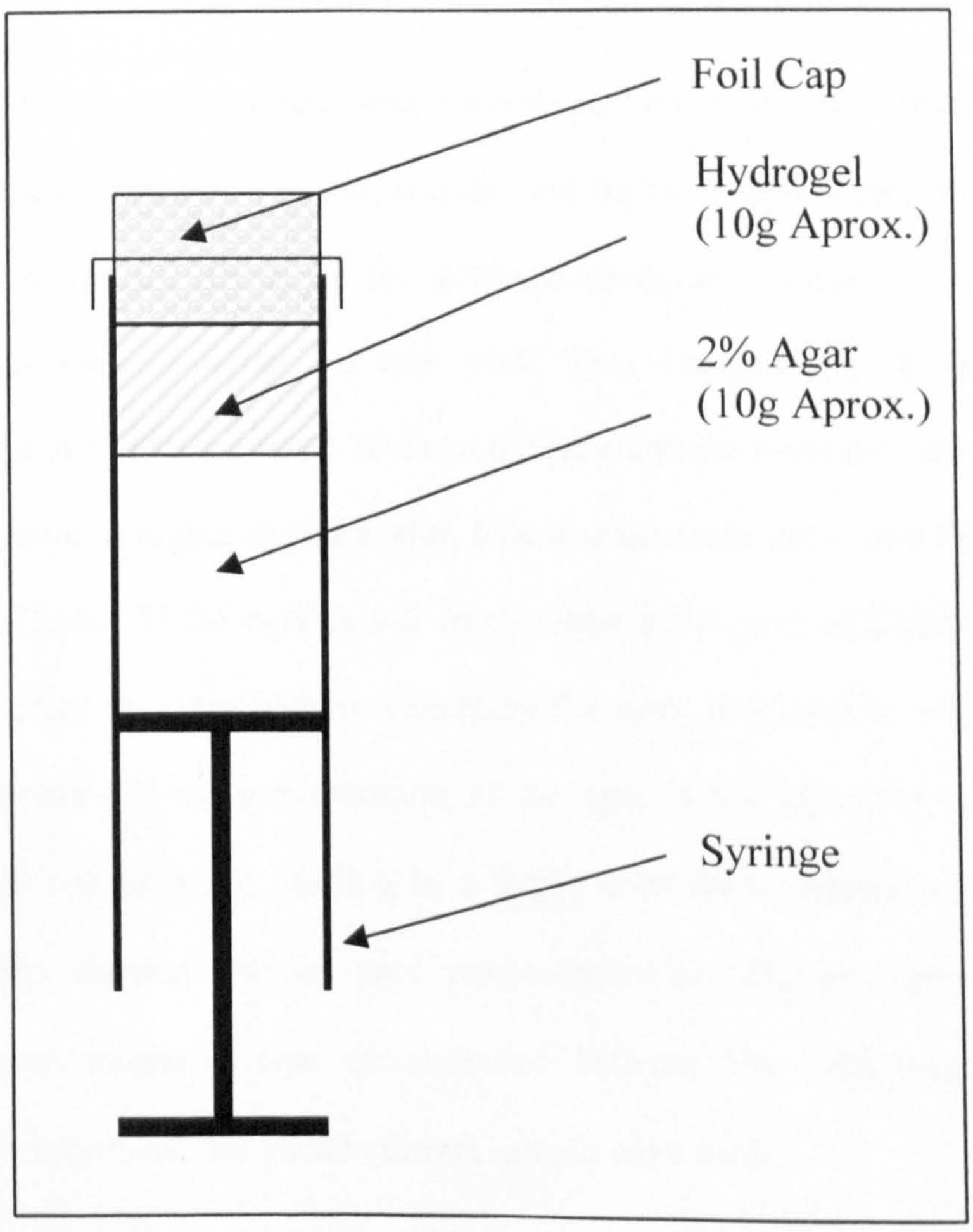


Figure 2.1 Diagram Showing the Arrangement of Apparatus for the Analysis of Hydrogels Using the "Paddington Cup" Method

2.2.6.1.1 *Intrasite Experiment 8, The "Paddington Cup" Method*

For the comparison of competing hydrogels all the various substrates are used. For the analysis of Intrasite gel only one substrate is needed. The variation in fluid transfer between batches of Intrasite gel is less than the variation between competitors products and Intrasite gel, making the different substrates unnecessary.

2.2.6.1.2 *Intrasite Experiment 9, Selecting the Correct Substrate*

The Paddington cup method was carried out on a all the different substrates recommended in the test method, and the best for the intended purpose was selected. After examining the results for the different substrates 2% agar was selected as the most appropriate to carry out this work. The concentration of agar affects the percentage of fluid transferred between the agar and the hydrogel, lower percentages of agar lead to a higher fluid transfer, however the trade-off is that the agar gel will not be as firm. If the agar is too fragile then it becomes difficult to remove the hydrogel from the agar without damaging the agar, this leads to large error in the measurements. If the concentration of the agar is too high, the volume of fluid transferred can be small, leading to a larger error from imprecision in measuring. Initial tests showed that an agar concentration of 2% w/w provided the best compromise, ranges of agar concentration between 1% - 6% were tried. Three replicate experiments for each hydrogel sample were made.

2.2.6.1.3 Intrasite Experiment 10, Generating the Reference Data

The Paddington cup method was carried out on 42 different batches of Intrasite gel over a six week period, using three replicates of each sample. The results showed that the fluid transfer between the agar and the Intrasite gel were fairly stable over the test period, this would be expected if the procedure to manufacture the gel was stable. The relatively constant results could be seen as an indication that the test is not really measuring what it is intended to. To demonstrate that the test was in fact measuring a changing property the test was carried out in its entirety (using all the recommended substrates) on a variety of competitor samples to compare the results.

2.2.6.2 Analysis of the Paddington Cup Data

The data produced was examined for its relationship to the fluid absorption variables measured using the settling volume method, and any relationship with the other recorded variables. No real results were expected as the Paddington cup method produced very stable results. A relationship between these new measurements and the old variables was found, though further work is required to improve on this.

3. PLS Results and Discussion

3.1. Reasons for Variable Selection

Much of the work described here has been reported in a paper submitted to Chemometrics & Intelligent Laboratory Systems [2], and can be seen in Appendix VI.

Multivariate regression techniques produce coefficients between an independent matrix and one or more dependent vectors. The coefficients minimise the influence of variables that do not positively contribute to the model, and maximise the contribution of variables that provide useful information. The coefficients produced by MLR are hindered when the problem is either over-determined or under-determined; the result is overcompensation of the coefficients. A large positive coefficient to a variable with no information is compensated by a large negative coefficient from another variable that contains no information. PLS and PCR solve this problem by regressing against new vectors themselves products of the independent vectors. The new vectors are created to minimise the contribution from variables with no information. The new vectors do however contain a contribution from all the variables used in the model. When a new test vector is introduced to the model for prediction the prediction is based on the coefficients spanning all the input vectors. If the vectors for those variables not containing any information contain values that are outside the ranges used in the calibration then the calibration will have error introduced from those variables. This is more extreme for MLR where coefficients for variables that contain no useful information can be quite large, however even in the factor analysis techniques the coefficients for vectors without information will not be at zero, they will just tend to zero. The contribution for a large number of unwanted variables provides a large part of the error present in a model. One possible solution to this is

to remove the variables that do not have information in them. Removing variables can be done as part of pre-processing, with spectroscopic calibration sections of spectra are routinely removed when it is known that there is only noise in that section. This is not an ideal method. Manual deletion of variables suffers from two main flaws

1. The judgement of the analyst must be considered, no two people will remove exactly the same sections of spectra, and the sections that are removed may not be the best one to remove. When examining complex spectra most people will remove sections where there is high noise, and sections were there is a low response, retaining those sections that contain the peaks. This can be counterproductive, information about the background noise in a spectra is important to a model, and the sections between overlapping peaks will often provide the information required to separate peaks.

2. With many spectra, particularly noise free spectra the largest source of error in prediction can be caused by collinearity between neighbouring wavelengths in a peak. Neighbouring wavelengths tend to provide the same information as each other, and are thus collinear, leaving these variable in the model will influence the matrix towards singularity, and this can be seen to strongly influence the coefficients.

Several possible routines have been considered to allow for the selection of the correct variables to build a model, they are all based on within model predictions, that is modelling the calibration set. Models that are optimised to predict from the calibration set often perform poorly when it come to the prediction of new results.

Using the predictive ability of the model to select the appropriate variables is a better solution.

3.2. *Reasons to Avoid Variable Selection*

Variable selection is not always appropriate; the most likely reason for this is when the object of the analysis is not the direct calibration of the data set with the aim of producing a model to predict component values, often component concentrations. This is usually associated with process analysis situations where the noise level present in the data set is often used as an indicator of the stability of the process. It is assumed that if the noise is stable across the variables recorded then the process is also stable. It is also important to consider the application of variable selection when the presence of unusual events within the data set must be considered important, that is that an event that does not occur in the training set in a section of the variables that otherwise has no information. An example of this might be the presence of an unexpected contaminant in a flow stream that does not affect the section(s) of the spectra that contain information about the analyte, where the presence of this contaminant must be detected. If a variable selection procedure is used the presence of this contaminant may not be detected as it will not effect the prediction results, of course even if a variable selection procedure is not used this does not mean that the contaminant will be detected as it may not have a significant effect on the predicted result of the component(s) of interest. It is important to note that variation in the prediction of the concentration of components should never be used as a method to detect contaminants, other methods should be used, such as more direct monitoring across all the variables, unless the model has been built expressly for that purpose.

In the situation where matrix effects need to be ignored, for example if a water sample is examined for lead only and other materials present are of no interest, then a variable selection routine will provide some level of robustness towards the matrix effects that may be present. Obviously for variables where a contaminant directly overlaps the variables selected to model a component there will be the same problems with a variable selection method as there would be experienced with a model built without variable selection.

3.3. Variable Selection MLR

Of the common multivariate techniques, MLR, PCR, and PLS, MLR will normally has the greatest error, this is due to the overcompensating coefficients, and under determination in most cases. The paper by Walmsley [1] concerning VS-MLR examined variable selection in several stages. Once the limitations of the existing techniques had been determined then the best approach was considered. The main issue with the other variable selection methods developed was the model building on the basis of calibration performance not predictive ability. By deciding to select variables on the basis of their predictive ability the two most popular methods had to be discounted immediately, variable selection by examining the correlation coefficients will always provide the same solution, and no modification based on predictive ability is practical. Variable selection based on the coefficients (effectively partial multiple correlation coefficient) also cannot be modified by predictive ability. These two methods are also weak when more than one dependent vector is considered simultaneously, and both have difficulty eliminating the problems caused by

collinearity since they both effectively act to enhance collinearity by selecting highly correlated variables.

The method first used to select variables was a single pass addition method; a single variable was selected initially by choosing the most highly correlated variable. The PRESS produced using this one variable was used as a baseline to compare the effect of adding in variables. Variables were added into the model individually on a random basis, a new PRESS was produced and if the model was an improvement with the added variable then it was retained, otherwise the variable was discarded. As the model improves the PRESS is updated so that each new PRESS produced is compared to the current best PRESS produced up to that point. Variables were added on a random basis as with most spectra the wavelengths within a peak are highly correlated, thus if the variables were presented to the model in sequence the first wavelength in a peak would be selected as being important to the model, the next wavelengths of the peak are likely to be reject as information about the peak is already present in the model and collinearity will cause the model performance to degrade with the added variable, the result of this is that sequential addition of variable neglects the most important variables as they have already been encountered in some form. This selection procedure was run a set number of times, and the group of selected variables that produced the best model were kept. The model building process must be repeated because of the random addition of the variables. The MLR coefficients vary according to the variables present in a model, certain groups of variables can produce unpredictable effects. If a small group of variables are selected in the early stages of building that provide a good solution to the problem then there may not be any single other variables that will provide an improvement in the model

and no more variables will be selected, the model will have peaked early, below its best. The additions of variables can also result in a very slow increase in the performance of the model, the result of this will be the selection of a large number of variables that are not actually required in the model.

This method of selecting variables produced a strong improvement on standard MLR, however there was still a tendency to retain too many variables. The reason for this is that if a variable that is retained in the model provided poorer quality information to the model than another variable that has not already been selected, then both variables will end up selected as there was no procedure to discard variable that became redundant. The VS-MLR procedure was modified to have a removal mode, once all the variables had been tested the process was run in reverse, a model was built from all the selected variables, and the PRESS produced used as a new baseline, variables were then removed individually and randomly and the model re-tested. Random chance still operates, so the model building must be repeated a number of time to obtain a good solution, however variables that had been added in the first pass that were then exceeded by subsequent variables will be removed from the model producing a better model.

Once these two stages were developed a squashing function was added to the model, the squashing function either encourages the addition of variables, or hinders them. The squashing function is a multiplier for the current best PRESS for a model. If the squashing function acts to reduce the current best PRESS then the addition of variables will be reduced as any improvement to the model for the addition of a variable will have to be greater than normal for the variable to be selected, and the

reverse will be true when the squashing function makes the current best PRESS larger.

The variable selection procedure described worked very well with many data sets, providing an 80% reduction in error on average. This method was transferred to PLS to determine whether it would be useful.

3.4. Comparison with VS-MLR

This work stemmed from work in developing a variable selection MLR algorithm; the development of the variable selection routine for MLR was eventually published [1]. During the development of the VS-MLR routine this work was begun, one of the data sets used in the two different routines (VS-PLS and VS-MLR) is the same, the UV-Data set. This can be seen as a link between the two methods for comparison purposes. The final version of the VS-MLR code can be seen to outperform VS-PLS for all components through all the VS-PLS versions with the exception of the final VS-PLS method (SVR-DP-PLS, section 3.10) where VS-PLS is superior to VS-MLR for the first component, iron. There are various reasons why VS-PLS is inferior to MLR for the other three components. First the UV data set is a very good data set, with little noise and peaks with minimal overlap, except for the iron peak, which is near the limit of detection. The UV data set for the other components is ideally suited to analysis using VS-MLR, once the number of variables is reduced the error is low, and the system is over determined, this allows VS-MLR to extract very good coefficients to describe this data set. VS-PLS has a much greater noise overhead for the "clean" components where by necessity noise is added in, both through the

various calculations, of which there are significantly more for the PLS algorithm than for the MLR algorithm, and through the use of latent vectors which tends to spread noise that cannot be removed from the system across all the components being predicted. This is due to the MLR calculations having separate coefficients for each component relating back to the separate variables, while the PLS coefficients are calculated for each component individually back to the same latent vectors, which contain contributions from all the selected variables. This can be seen with the iron component where the advantage that PLS has with regards to removing noise outweighs the advantages MLR has with limiting noise across components, giving a better prediction for the iron component. While there have been no other comparative uses with other data sets it is hypothesised that this advantage that VS-PLS has means that for noisy data, data where the peaks overlap to a significant degree, and data where the peaks of interest are not the major influences in the data set , VS-PLS will show strong predictive superiority over VS-MLR.

3.5. Variable Selection PLS

The development of the variable selection PLS algorithm took place in approximately five stages, the results of each of those stages are outlined below. The histograms examined in each section were only developed after the final stage had been produced, so each stage was repeated to get the information needed to produce the histograms. The results reported are all shown using PRESS values this is because the PRESS values can be used to compare results between components in a model and between different models of the same data, but not to compare models of different data.

3.6. Variable Selection Histograms

The process of variable selection used is an iterative one using randomized variable orders. As has been discussed this is to reduce to effect of collinearity on the selection of variables. On any one pass through the algorithm there will normally be variations in the variables selected, no two models are likely to be identical unless the collinearity of the data is very low or the data set is very small. Although the various models produced will be developed using differing selections of variables, variables that contain particularly important information will be selected with a greater frequency than variables with lower information or worse signal to noise ratios. By examining every run through the algorithm that a data set makes, not just the one with the lowest error, patterns can be built up about which variables contain more information than others, this will tend to indicate sections of the spectra that contain useful information. The frequency that a variable has been selected over the whole course of training is recorded so that this information can be incorporated into the model evaluation, in the figures that display this information only a few of the spectra used in the model have been shown to avoid crowding of the graphs.

3.7. Single Addition Mode, SVA-PLS

The flow chart for the first variable selection method can be seen in figure 3.1, and shows the stepwise procedure followed in this algorithm.

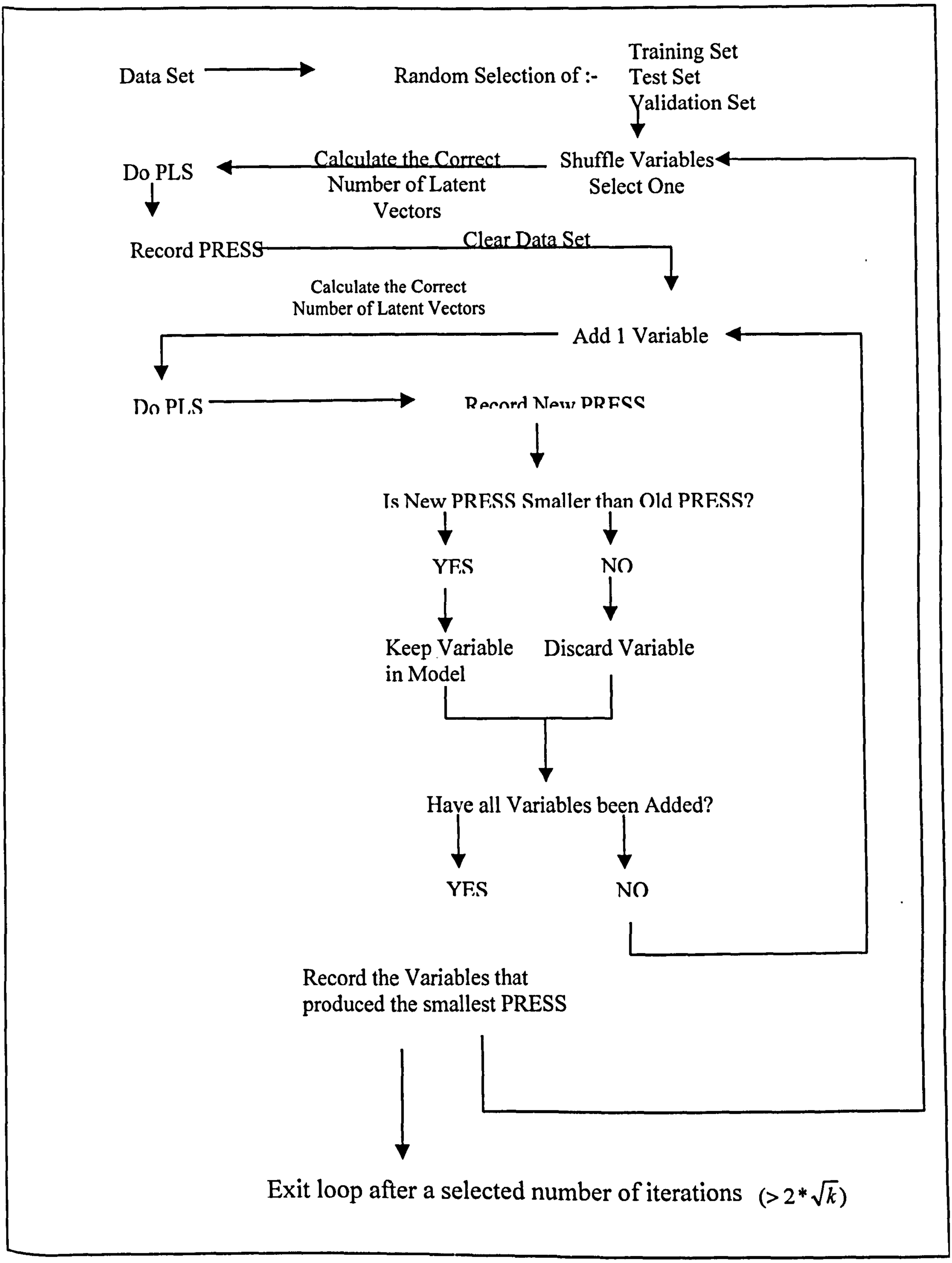


Figure 3.1, Flow chart for the first variable selection method, SVA

3.7.1. UV Data Set

Two hundred iterations were trained and the best model at that point was examined. The average number of variables selected over 200 iterations was 23, and the number of variables selected for the model with the lowest PRESS was 45.

Component	PRESS
Fe	21.8267
Co	0.7054
Ni	0.5974
Cu	0.2126

Table 3-1 PRESS Results for ordinary PLS using the UV data set, 7 LVs were used, and the base PRESS was 23.34

Component	PRESS
Fe	16.8267
Co	0.6054
Ni	0.5674
Cu	0.116

Table 3.2 PRESS values for the model developed for the UV data set using SVA. 7 LVs were used, and the base PRESS was 16.20

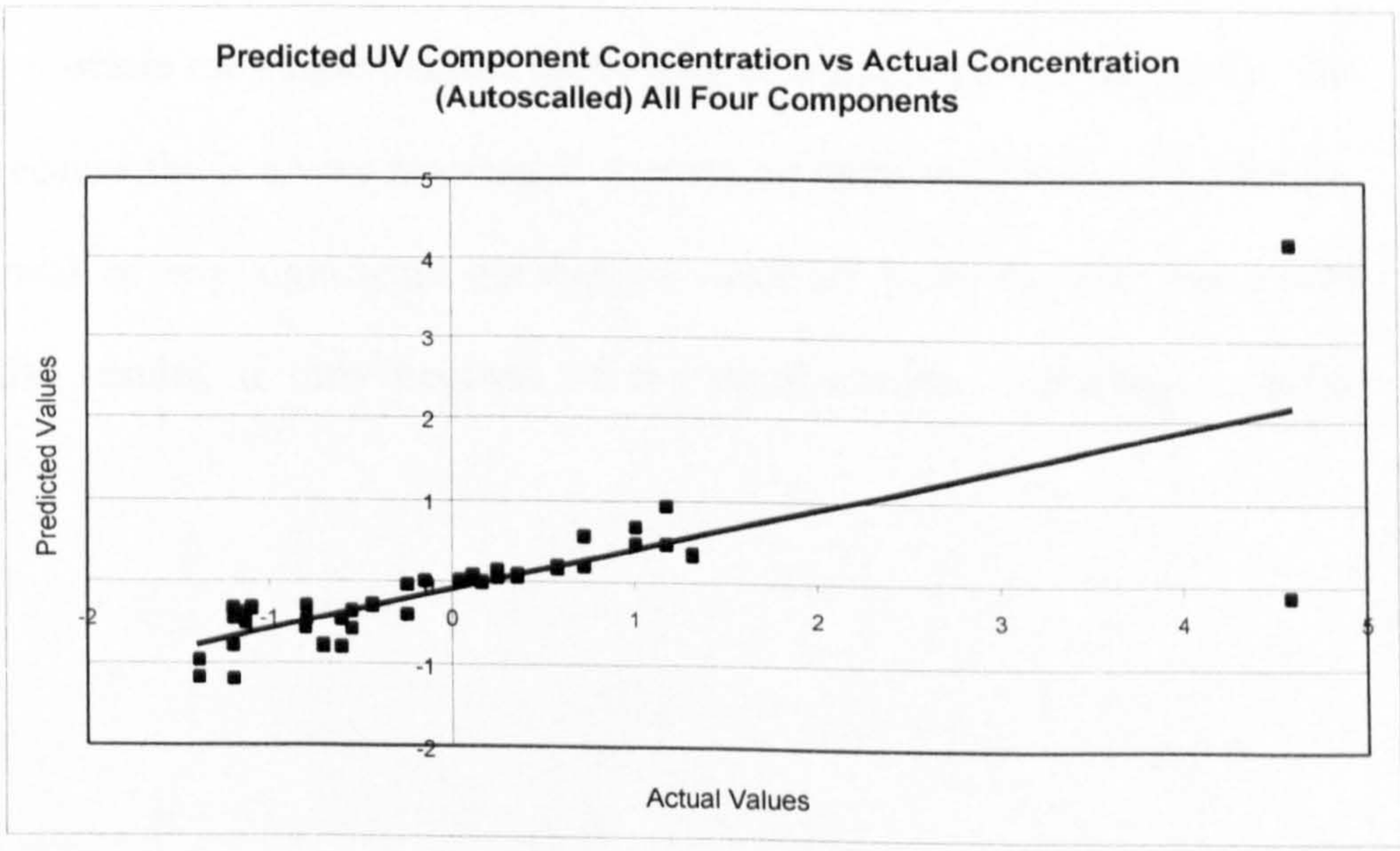


Figure 3.2 Prediction results for the UV data set using SVA

On average this method retained about 10% of the available variables, and improved over ordinary PLS by a reduction in error of about 33% (Table 3.1 and Table 3.2), this compares poorly with the equivalent VS-MLR method, which achieved reductions in error of about 80% with this method. The PLS variable reduction techniques was not expected to outperform the MLR method initially due to the limitations of the PLS algorithm when used in this way. The algorithm starts with a single variable, and this requires the PLS model to initially use a single latent vector, this is a very poor situation for a multi-component mixture and results in each variable examined being selected until sufficient variables are present in the PLS model for a stable solution. As the variables added may have little or no relevance to the model this can mean the addition of a significant number of variables before the solution stabilises. Figure 3.2 shows the predicted results for this model, all the components have been plotted as a single data set as it is the overall results that are of interest in this work not just the results from a single component. It should be noted that although the model has been considered as a whole the single biggest improvement is in the prediction for Fe, this likely to be because the is a very high noise component in the information for the Fe, and the removal of any significant number of variables from the data set would achieve similar results, if only because of the commensurate reduction in noise present.

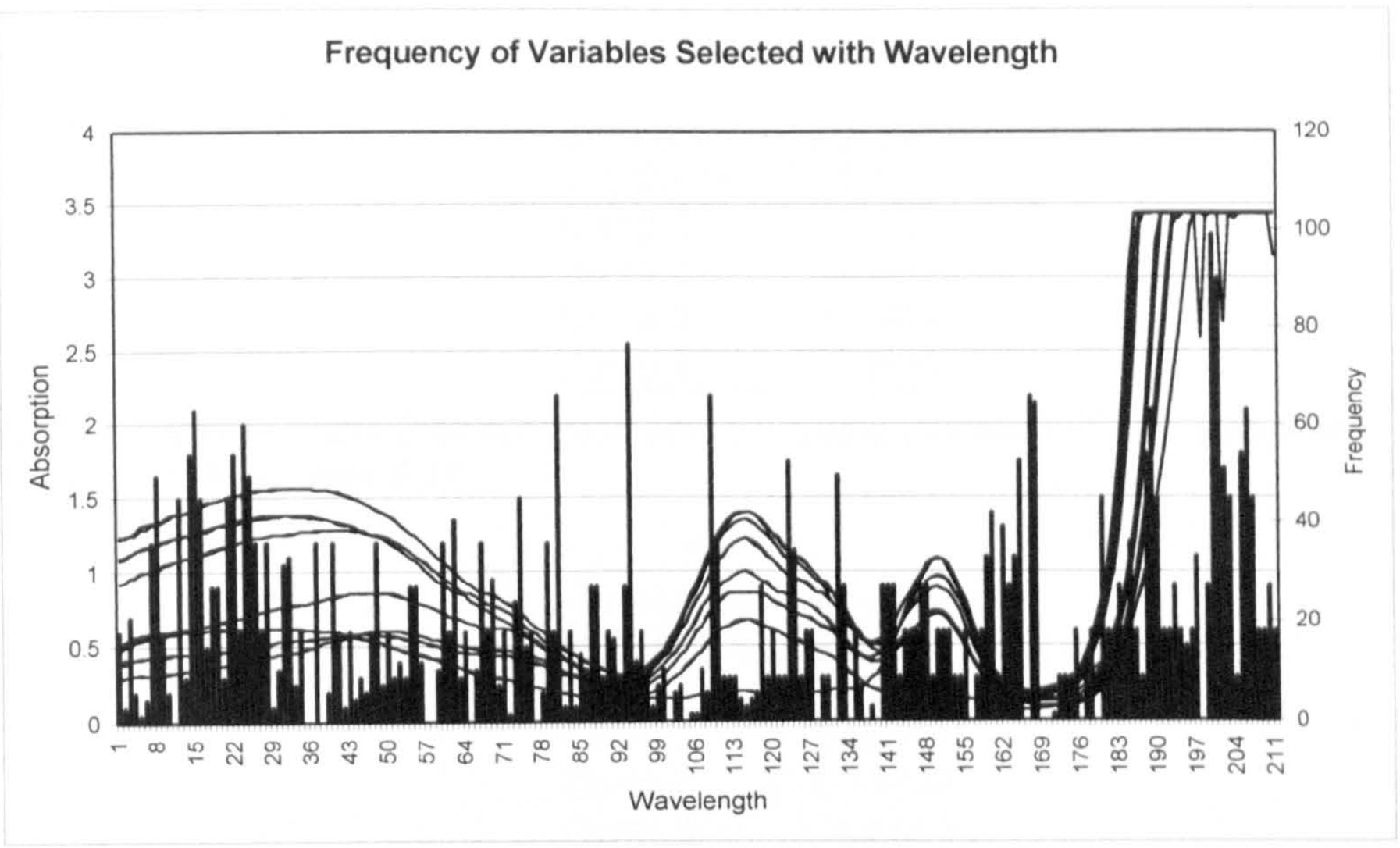


Figure 3.3 Frequency with which a particular wavenumber is selected from the UV data set by SVA

Figure 3.3 shows that there is structure to the selection of variables, although no wavenumber has been selected constantly by the model, this is expected as many of the wavenmbers carry the same information. A feature of interest is that the centres of the peaks have not been picked more frequently than other sections of the spectra, this was hypothesised earlier as is likely to be due to the fact that more information is available in the overlapped areas concerning the contribution from differing components.

3.7.2. Artificial Data Set 1

Two hundred iterations were trained and the best model at that point was examined. The average number of variables selected over 200 iterations was 15, and the number of variables selected for the model with the lowest PRESS was 14.

Component	PRESS
Comp 1	2.3795
Comp 2	2.2962
Comp 3	0.21709
Comp 4	1.8912

Table 3.3 PRESS values for the first artificial data set using ordinary PLS, 6 LVs were used and the base PRESS was 6.57

Component	PRESS
Comp 1	3.3795
Comp 2	3.2962
Comp 3	0.11709
Comp 4	4.8912

Table 3.4 PRESS Values for the first artificial data set using SVA, 6 LVs were used, and the base PRESS was 11.68

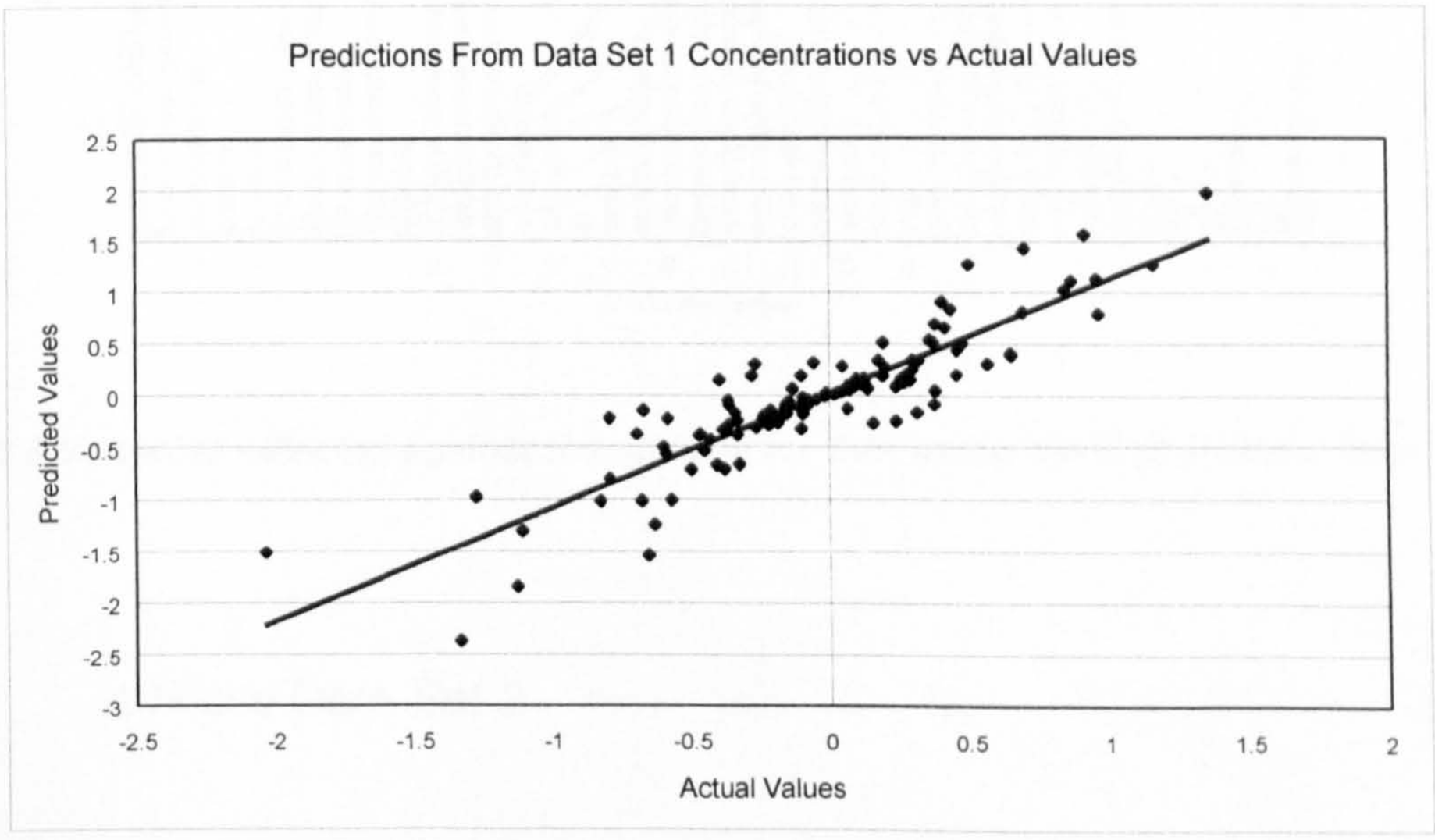


Figure 3.4 Predicition results for SVA on the first artificial data set

The PRESS results seen with this data set shows that the model is actually inferior in most respects to the original PLS one, the overall PRESS is lower, and it is only the PRESS for component three that shows any improvement. Component three is the only linear component in this model. This could be because for the non-linear components a stable model cannot be built up using the addition of stable models, and

any variable that could improve the model could only do so as part of an interruption with another variable. Figure 3.5 shows that there is structure to the variable selection and this could be due to the selection of variables that allows the contribution of the three non-linear components to be removed from the calculation for the linear component.

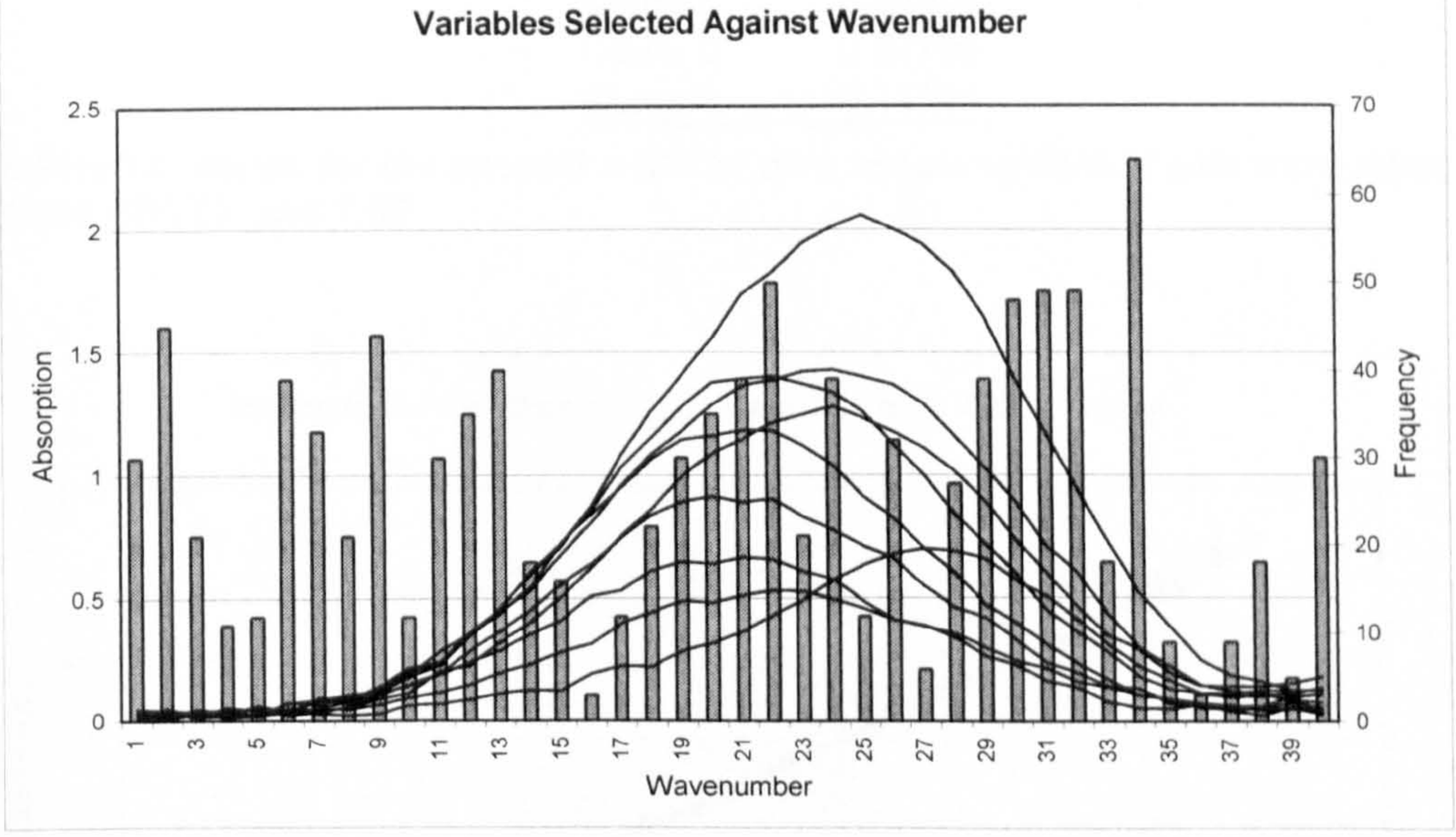


Figure 3.5 Variables selected against the spectra for SVA using the first artificial data set

3.7.3. Artificial Data Set 2

Two hundred iterations were trained and the best model at that point was examined.

The average number of variables selected over 200 iterations was 73, and the number of variables selected for the model with the lowest PRESS was 73.

Component	PRESS
Comp 1	0.16282
Comp 2	0.36906
Comp 3	0.39796
Comp 4	0.17581

Table 3.5 PRESS results for the second data set using ordinary PLS, 4 LVs were used, and the base PRESS was 1.11

Component	PRESS
Comp 1	0.14282
Comp 2	0.34906
Comp 3	0.37796
Comp 4	0.15581

Table 3.6 PRESS results for the second artificial data set using SVA, 4 LVs were used, and the base PRESS was 1.00

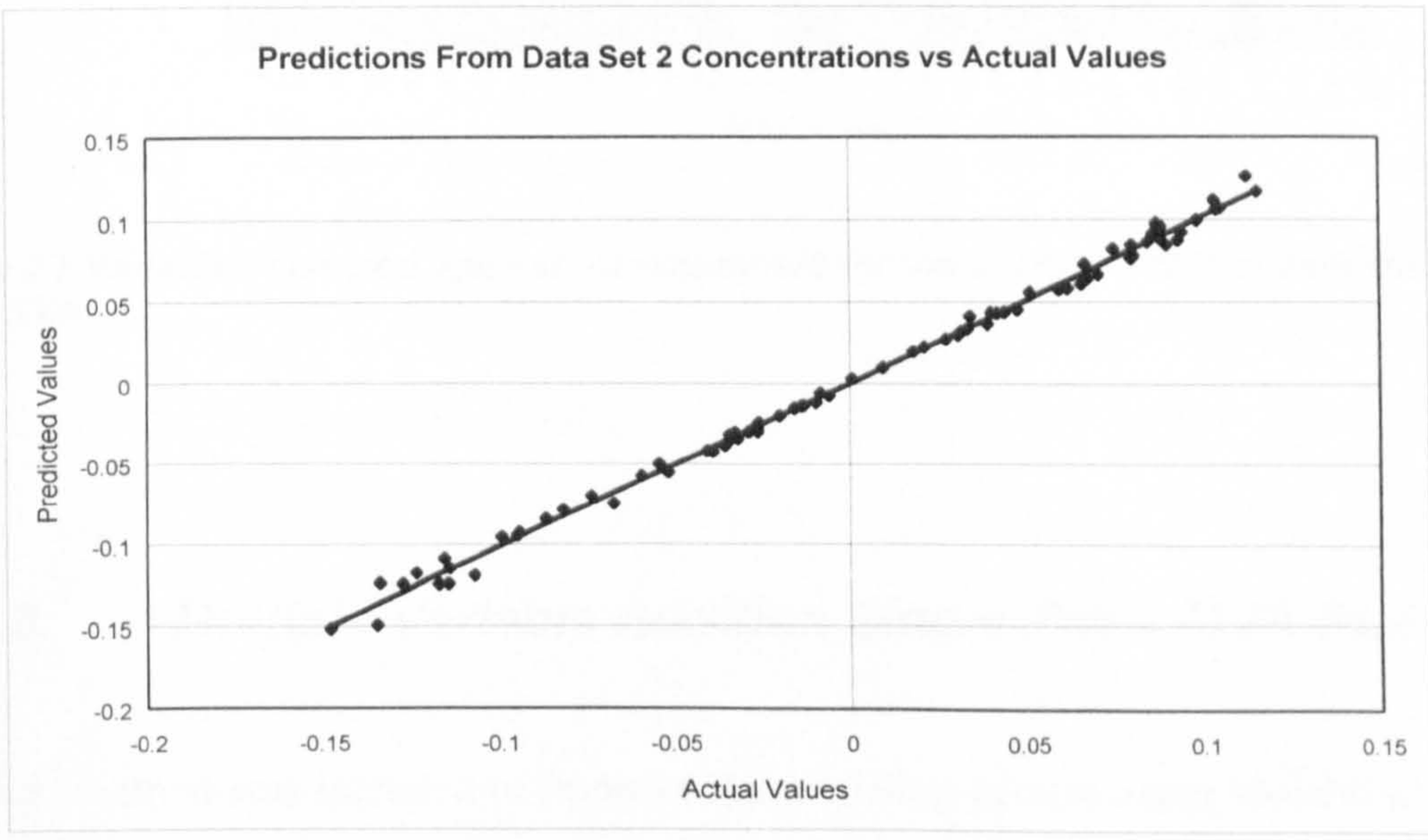


Figure 3.6 Predicted results for SVA on the second artificial data set

Ordinary PLS can predict using this data set very well, the improvement using the variable selection routine is only minor, figure 3.7 shows that although there is some structure to the variables selected this does not show any clear features. It is possible that there is too little variation and too little noise for individual variables to be significantly better than any other variable. There is a clear indication from the

amount of noise in evidence that far too many variables are being selected, the model could probably perform very well with a very small number of variables.

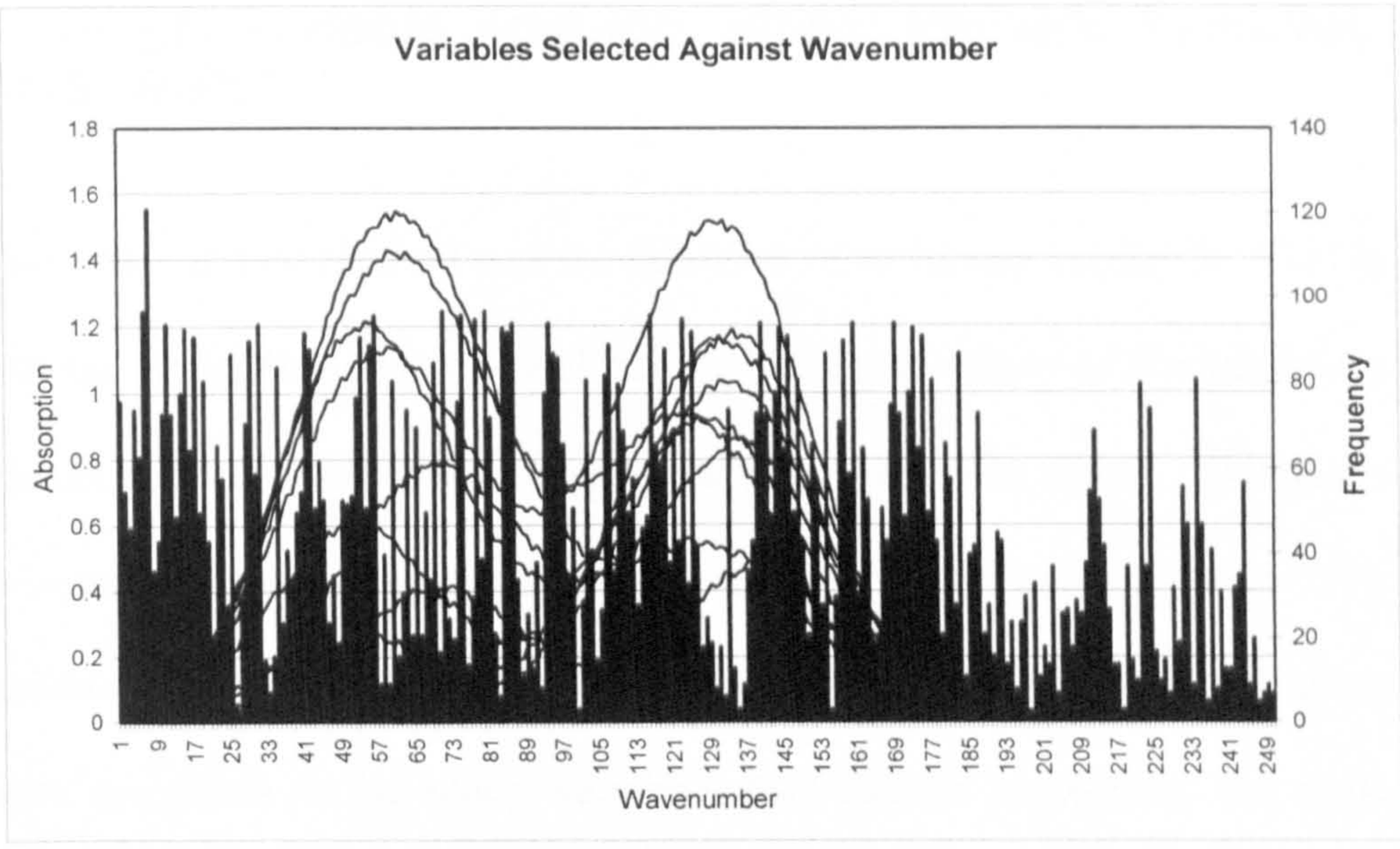


Figure 3.7 Variables selected against wavenumbers for the second artificial data set using SVA

3.8. Multiple Variable Addition Single Pass, MVA-PLS

The next method was intended to improve the modeling of non-linear variables, and to reduce the number of variables added initially when the stable PLS model was being built. Variables were added to the model groups of size equal to the number of components in the data set so hear the number of variables added was four for each data set. This reasoning was flawed, as variables were accepted or rejected as blocks of four, this resulted in very poor variable selection and a large number of variables being selected to no benefit. The results showed that about 45% of variables available were selected in each case and although there was a slight improvement in the modelling for the non-linear components this was due to the increased number of

variables being selected bringing the model closer to the ordinary PLS model. There was no improvement for the linear components.

3.9. Single Variable Addition, Single Variable Removal, SVA-SVR-PLS

The main flaw with the first method was the selection of too many variables. This is due mainly to the instability of the model during the initial stages of the modeling when insufficient variables have been selected for the model to be stable. This gives rise to the selection of any variable regardless of suitability. Another key reason for a surplus of selected variables is again the issue of co-linearity. In each of the three data sets there are likely to be many variables that contain essentially the same information with only slight differences in the signal to noise ratio If a variable is selected initially with a low signal to noise ratio, any variable with the same information but a better signal to noise ratio will be selected as well at a later stage. The variable selection routine needs some procedure to remove redundantly selected variables. The logical method to do this is to test the suitability of variables after they have been selected by examining the performance of the model when they are removed. This method showed improvements over the previous two methods, and the flow chart showing the algorithm can be seen in Figure 3.8

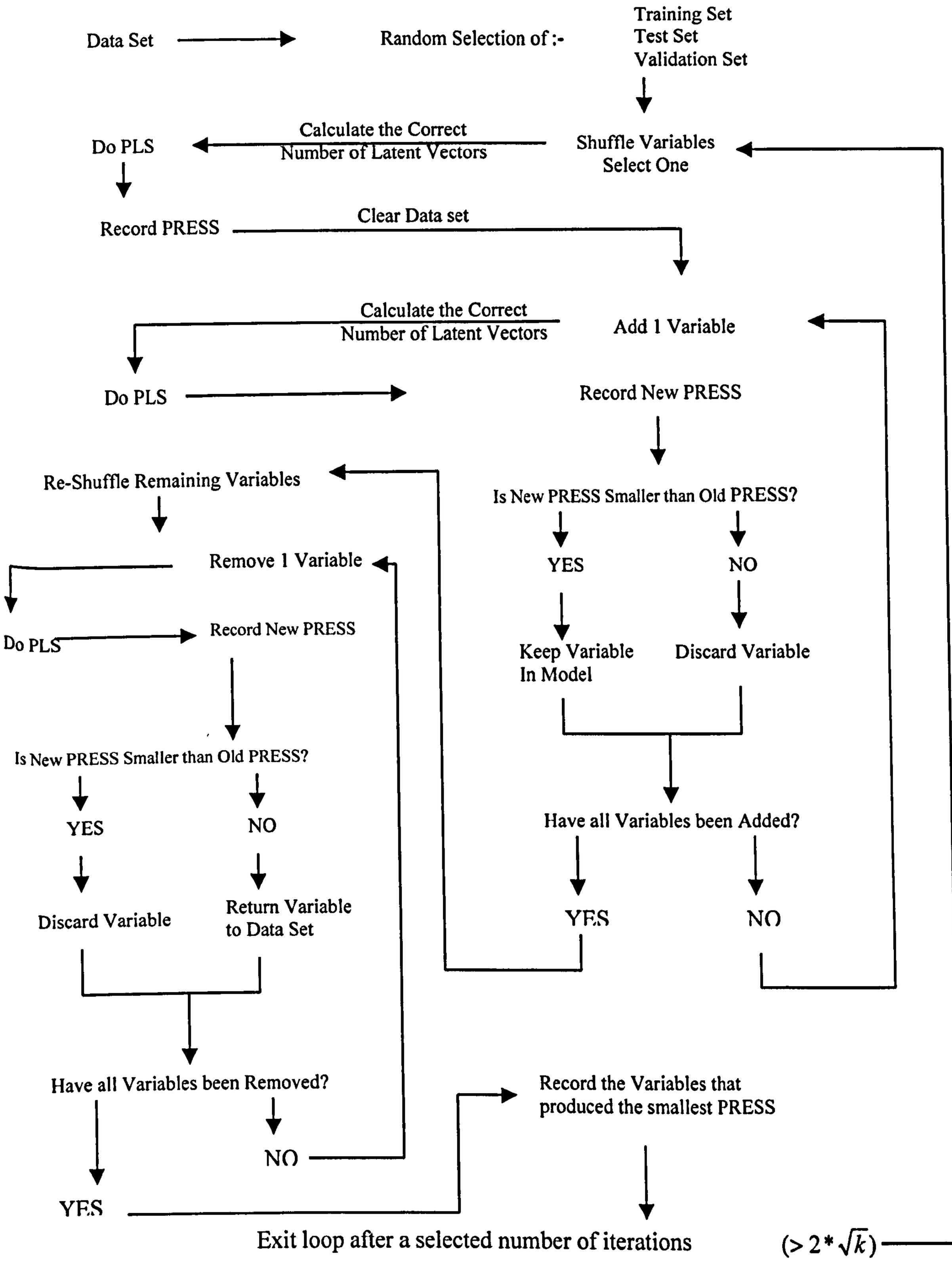


Figure 3.8 Flow chart showing SVA-SVR

3.9.1. UV Data Set

Two hundred iterations were trained and the best model at that point was examined. The average number of variables selected over 200 iterations was 16, and the number of variables selected for the model with the lowest PRESS was 15. The PRESS values can be seen in Table 3.7.

Component	PRESS
Fe	11.2546
Co	0.54923
Ni	0.3235
Cu	0.2

Table 3.7 PRESS values for the model developed for the UV data set using SVA-SVR. 7 LVs were used, and the base PRESS was 12.32

This algorithm shows a good improvement over the previous two with this data set, one key point is both a reduction in the number of selected variables and a reduction in the predictive error. The structure to the frequency of variable selection (Figure 3.10) is also more pronounced, again clearly emphasising the variables that are away from the peak centres. This shows the importance of examining the contribution of variables that provide information about the overlap between peaks.

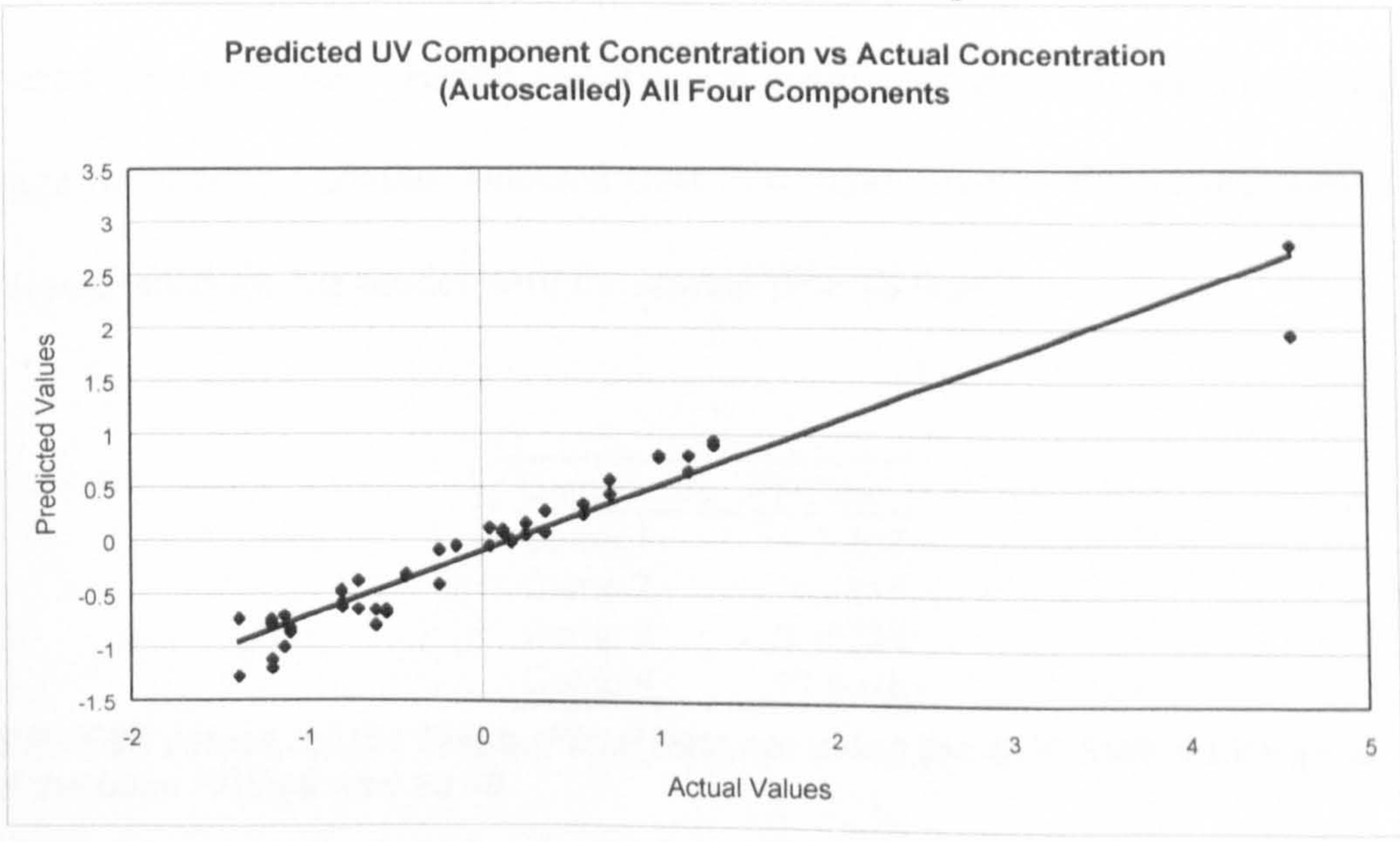


Figure 3.9 Prediction results for the UV data set using SVA-SVR

In this case the high end noise seen in the spectra has had variables selected from it fairly constantly, this is likely to be a requirement to examine the base line noise across the spectra.

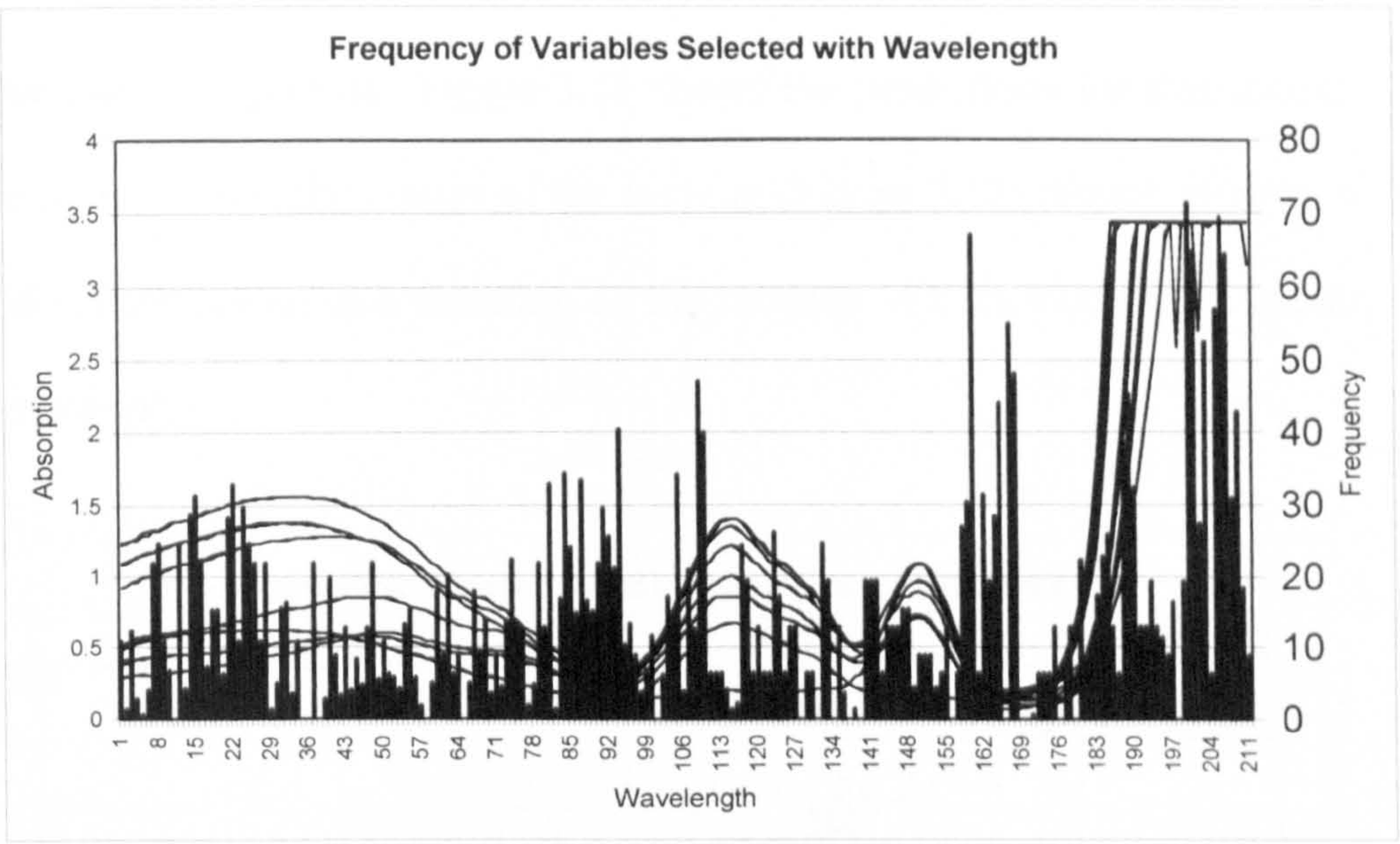


Figure 3.10 Frequency of variables selected against the spectra, UV data set using SVA-SVR

3.9.2. Artificial Data Set 1

Two hundred iterations were trained and the best model at that point was examined. The average number of variables selected over 200 iterations was 6, and the number of variables selected for the model with the lowest PRESS was 2.

Component	PRESS
Comp 1	2.468
Comp 2	8.2316
Comp 3	0.10324
Comp 4	10.0015

Table 3.8 PRESS values for the first artificial data set using the SVA-SVR, 6 LVs were used and the base PRESS was 20.80

With the exception of the third component of this data set this method was inferior to any method tried so far, ordinary PLS, and the two VS-PLS methods performed better. As intended this method reduced the number of variables selected however the variables that remained appear to have been those that contained information concerning the third component. Figure 3.11 shows the predictions for this model. The variables selected over the course of the training (Figure 3.12) remain similar to those selected initially, again concentrating on the sections of the spectra that explain the linear component.

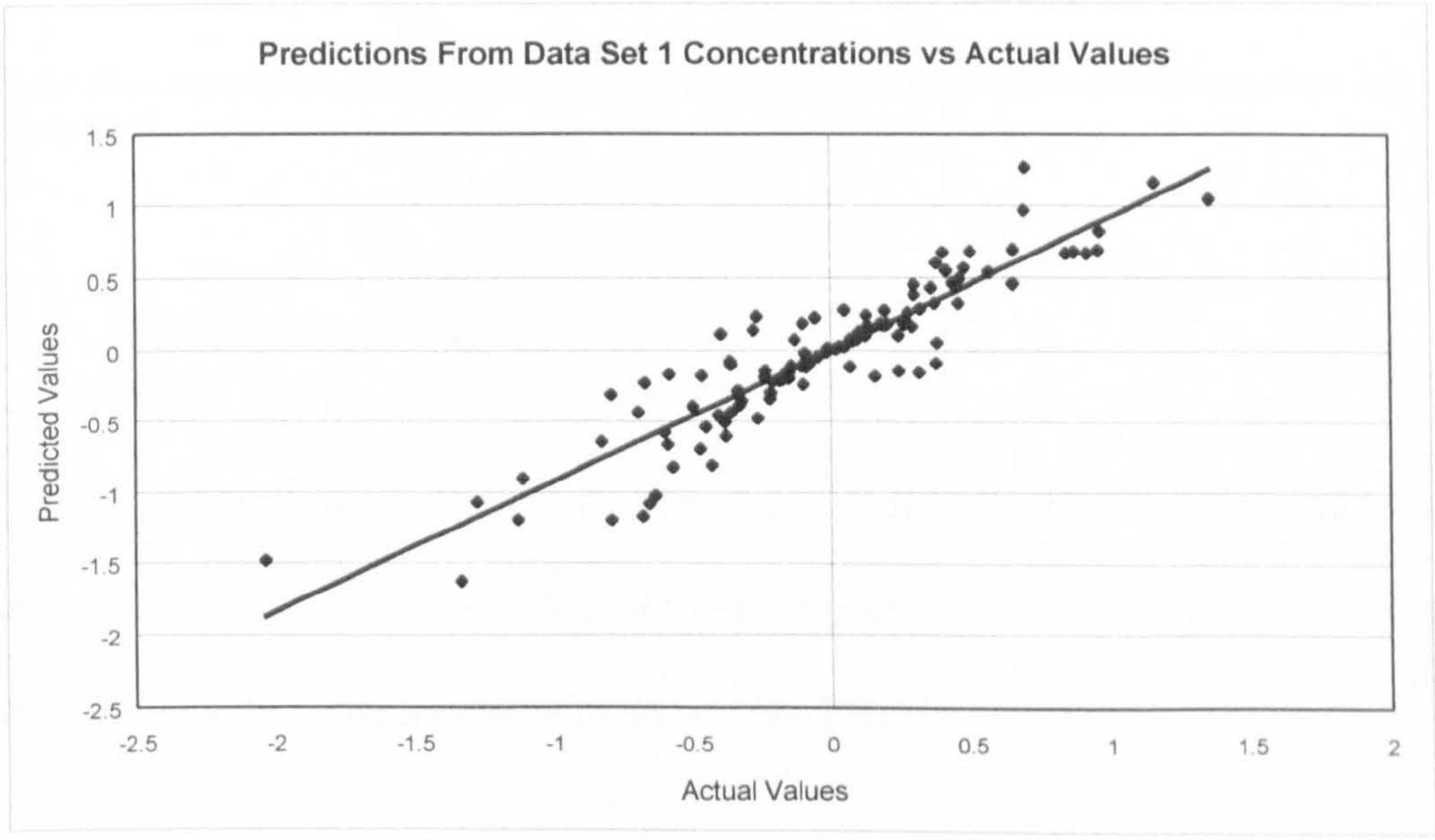


Figure 3.11 Prediction results for the fist artificial data set using SVA-SVR

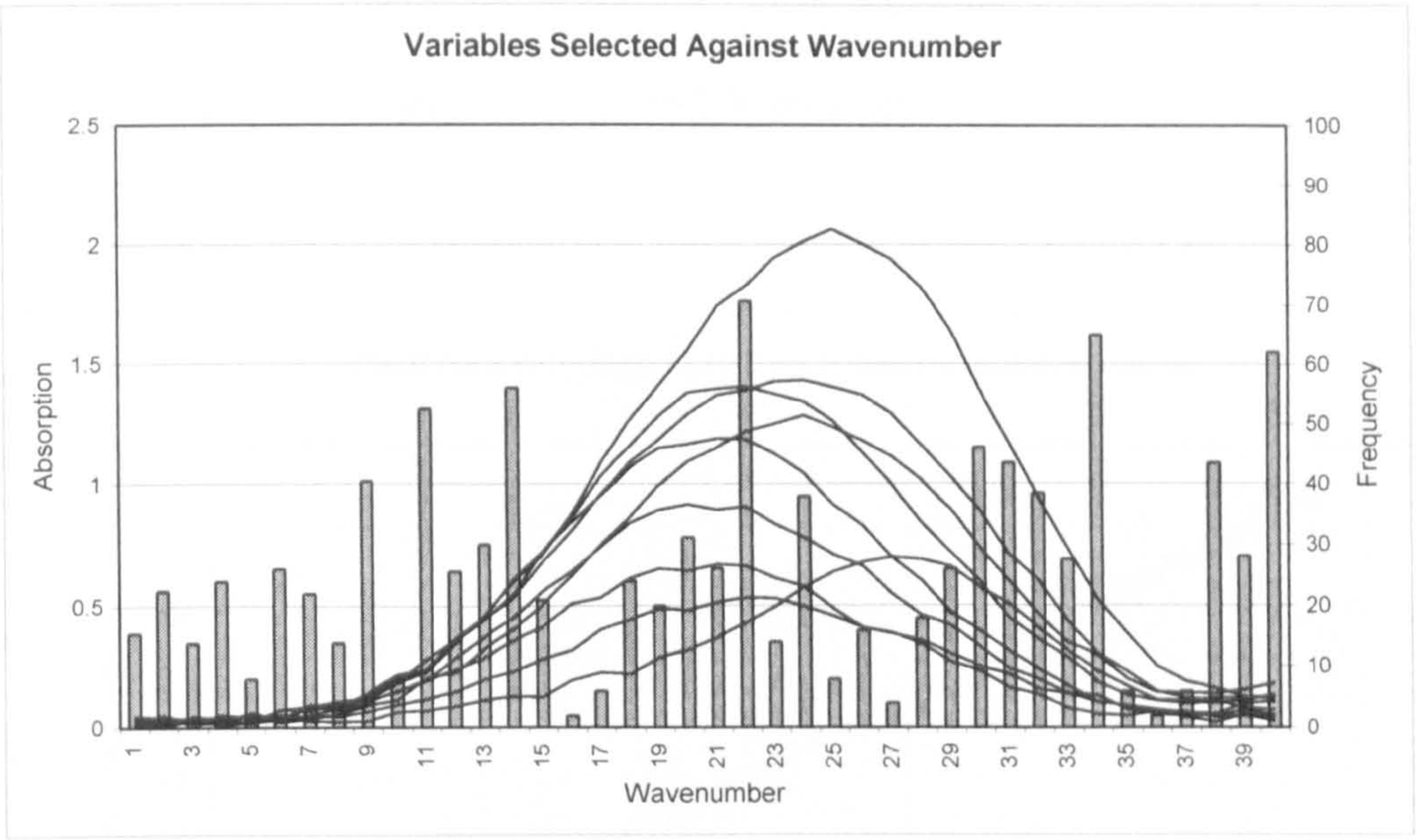


Figure 3.12 Frequency of variables selected vs. spectra for the first artificial data set using SVA-SVR

3.9.3. *Artificial Data Set 2*

Two hundred iterations were trained and the best model at that point was examined. The average number of variables selected over 200 iterations was 29, and the number of variables selected for the model with the lowest PRESS was 22.

Component	PRESS
Comp 1	0.06832
Comp 2	0.36022
Comp 3	0.1096
Comp 4	0.0758

Table 3.9 PRESS values for the second artificial data set using SVA-SVR, 4 LVs were used and the base PRESS was 0.61394

This method did significantly better than the first and second methods, this was expected as there is a large number of variables in this data set, and unlike the UV data set there is a high percentage of collinearity. The PRESS values (Table 3.8) show that the improvement is equal across all the components in contrast to the other

two data sets, again this is due to the non-linear variables in the first artificial data set and the high noise in the Fe component of the UV data set.

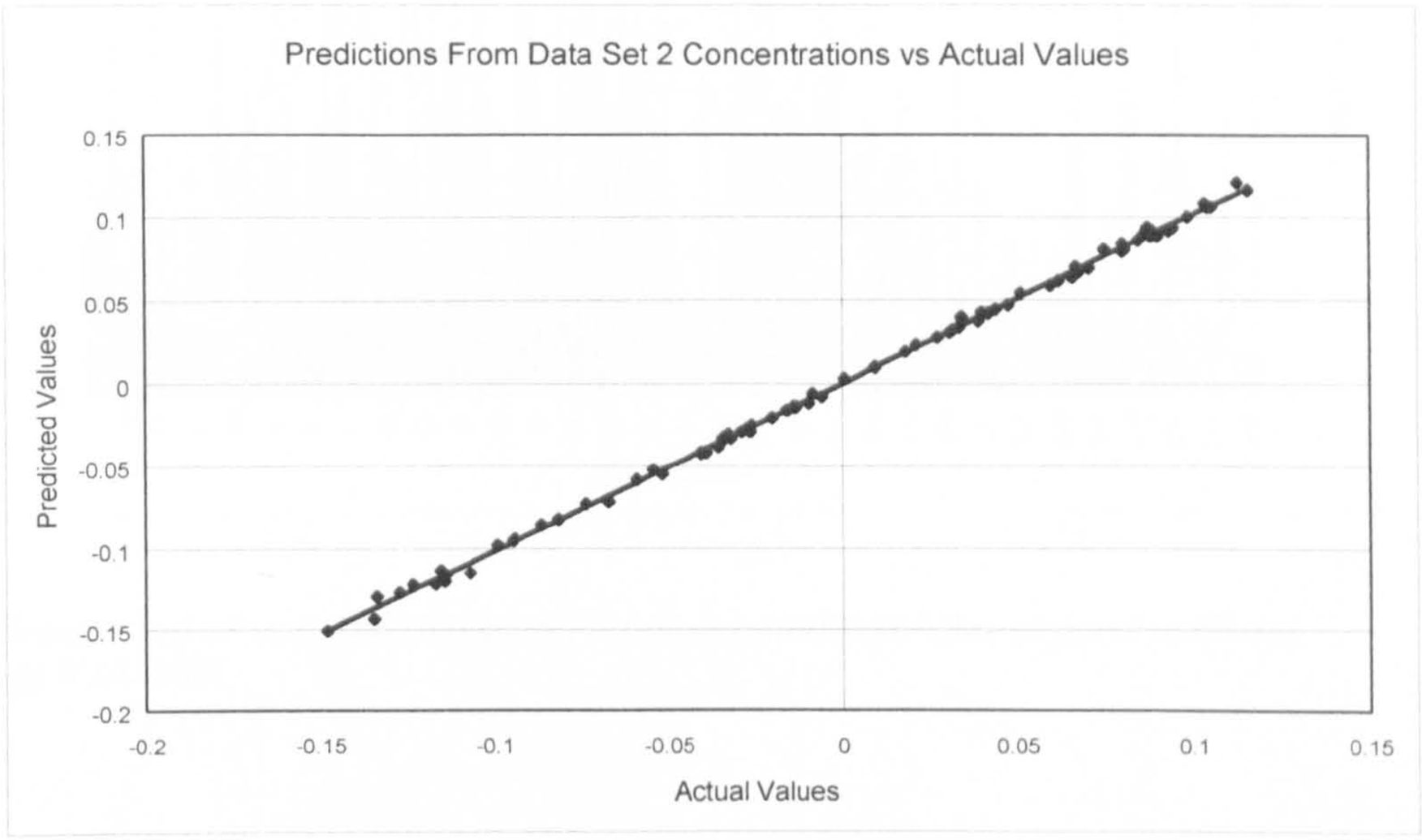


Figure 3.13 Prediction results for the second artificial data set using SVA-SVR

Figure 3.13 does not show any huge improvement over any of the previous methods, however this is not expected as the modelling for this data set was very good anyway. Figure 3.14 shows a little more structure compared with the previous histogram (Figure 3.7), an although it cannot be conclusive again it shows that the information about the overlaps between the peaks is again ranked higher than the information contained in the peak maxima.

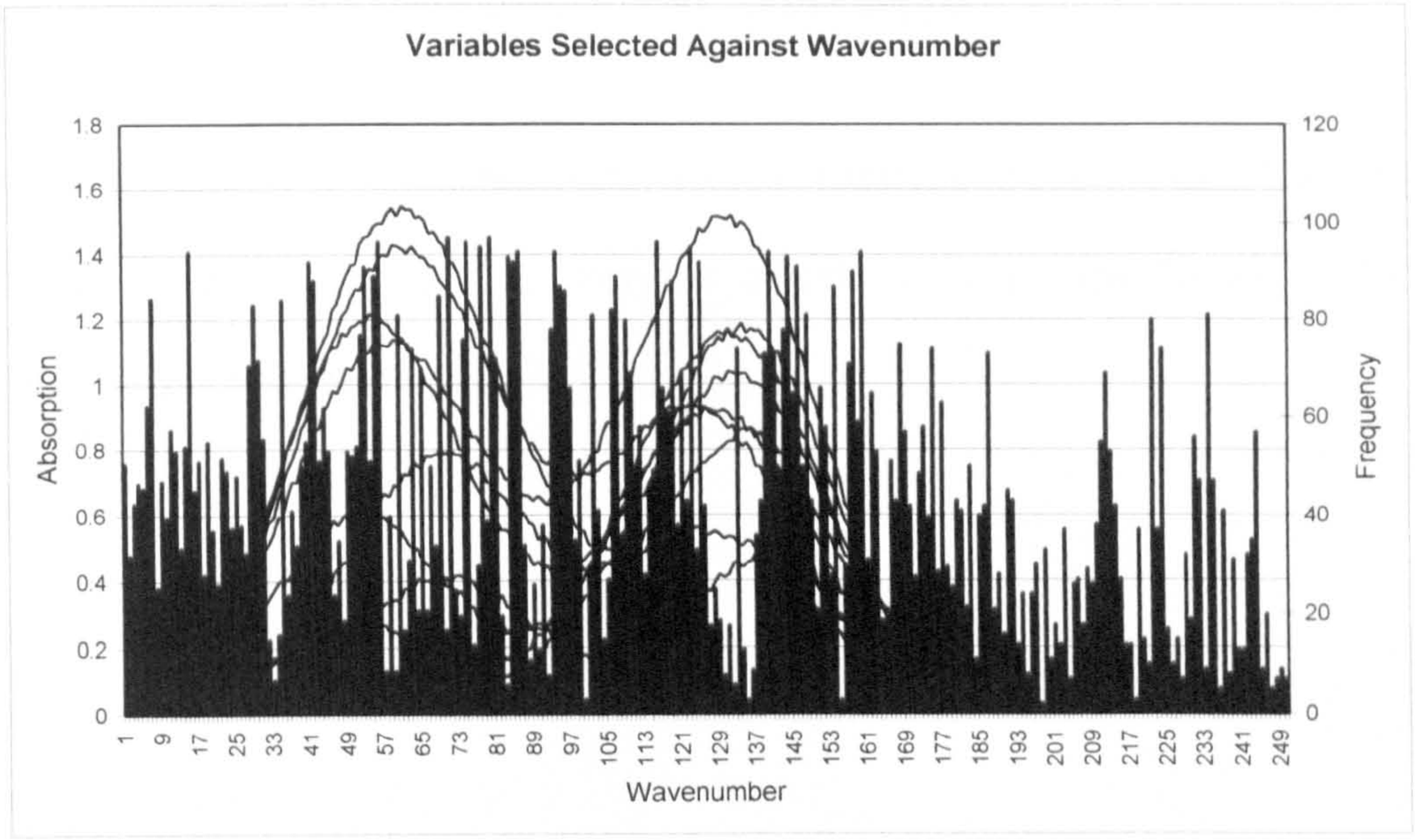


Figure 3.14 Frequency of variables selected against spectra for the second artificial data set using SVA-SVR

3.9.4. Summary

The overall result of the third VS-PLS method is that it successfully solves the major problem with the first method, that of selecting too many variables. This is not true of the non-linear components, but the method does improve the modelling of linear components when there are non-linear components present in the data set. The presence of non-linear components overlapping with a linear component can cause severe problems with conventional calibration methods.

One interpretation of the Addition-Removal algorithm is that the Addition section "thins" out the variables that may be of use, and the Removal section sorts through to files the selection. This raises the possibility that the algorithm could function perfectly well with just the Removal section, the addition stage could well be superfluous. This hypothesis is tested in the next section.

3.10. Single Variable Removal, SVR-PLS

The single variable removal algorithm is intended to examine the possibility that the initial variable addition step in the previous algorithm was superfluous. The algorithm used to test this can be seen in Figure 3.15.

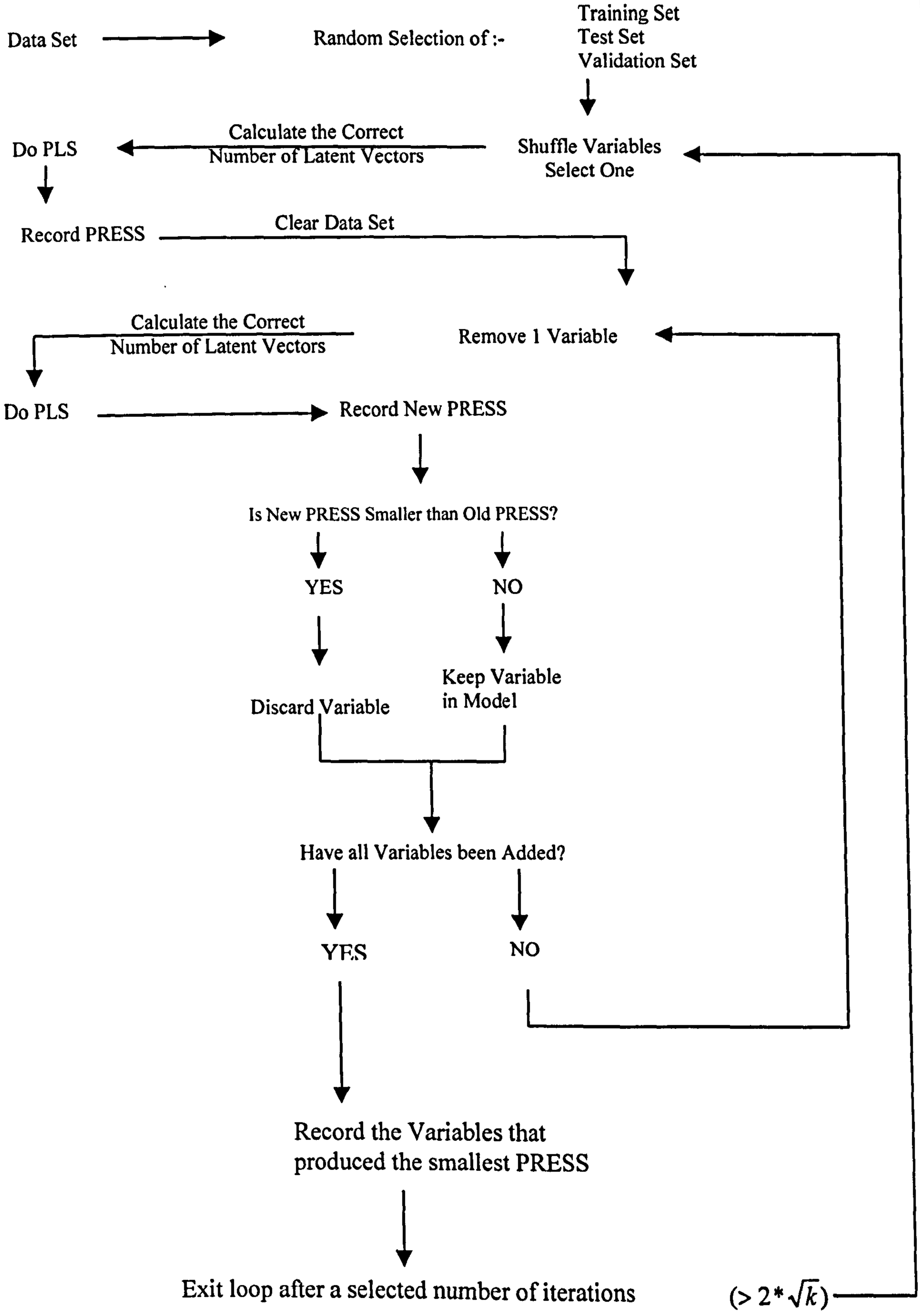


Figure 3.15 Flow Chart for Single Variable Removal, SVR

3.10.1. UV Data Set

Two hundred iterations were trained and the best model at that point was examined. The average number of variables selected over 200 iterations was 16, and the number of variables selected for the model with the lowest PRESS was 13. The PRESS values can be seen in Table 3.7.

Component	PRESS
Fe	11.3487
Co	0.23269
Ni	0.20232
Cu	0.097441

Table 3.10 PRESS values for the UV data set using SVR, 7 LVs were used and the base PRESS was 11.88

What is significant about this method is the remarkable similarity between the results for the Addition-Removal method and this one, the PRESS values seen in Table 3.110 are lower than those in Table 3.7. This suggests as hypothesised that the initial addition step is in fact redundant.

The prediction results (Figure 3.16) are also similar to the results in Figure 3.9 as expected from the PRESS results, and the frequency of variable selected (Figure 3.17) is again very similar to Figure 3.10.

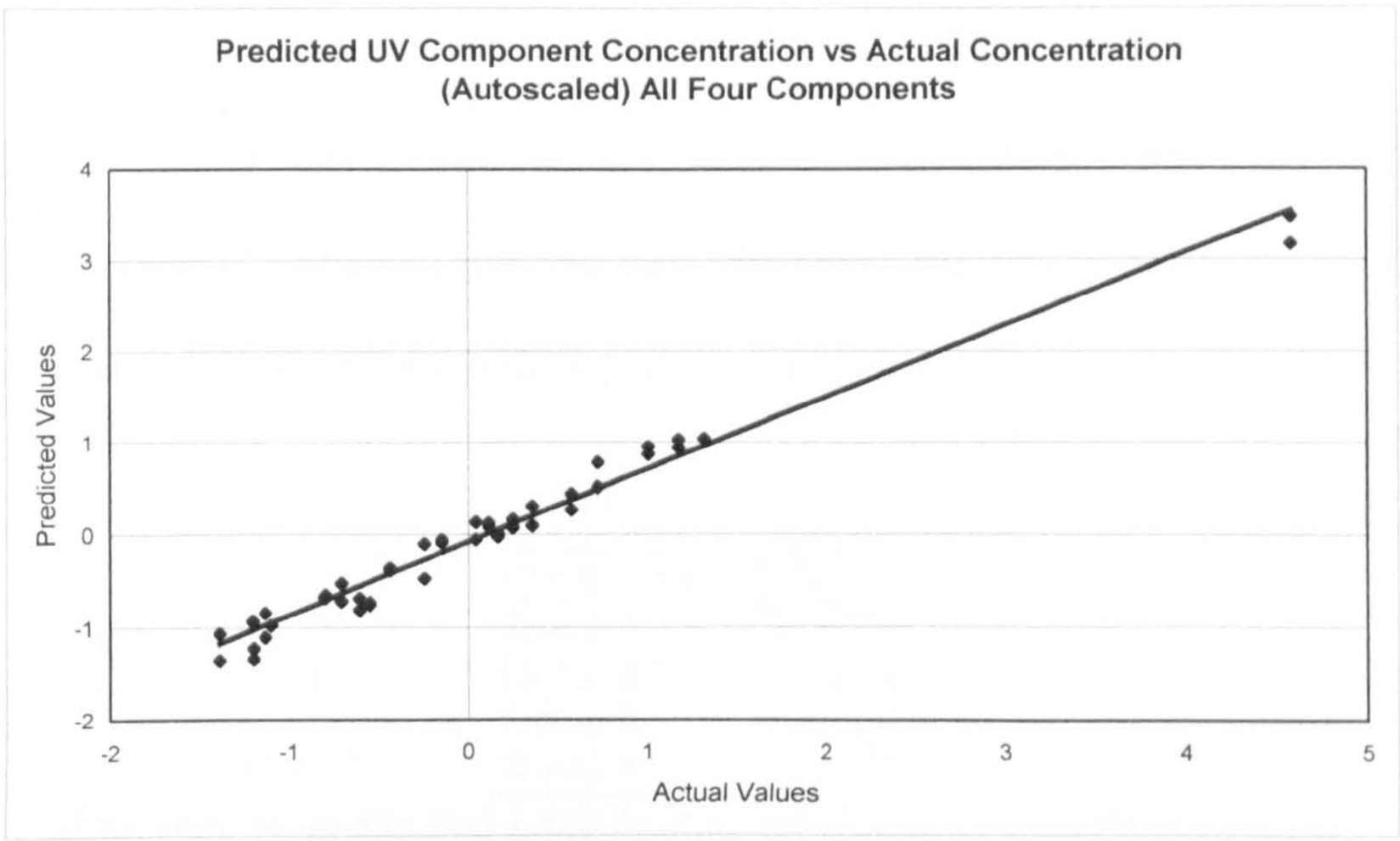


Figure 3.16 Prediction results for the UV data set using SVR

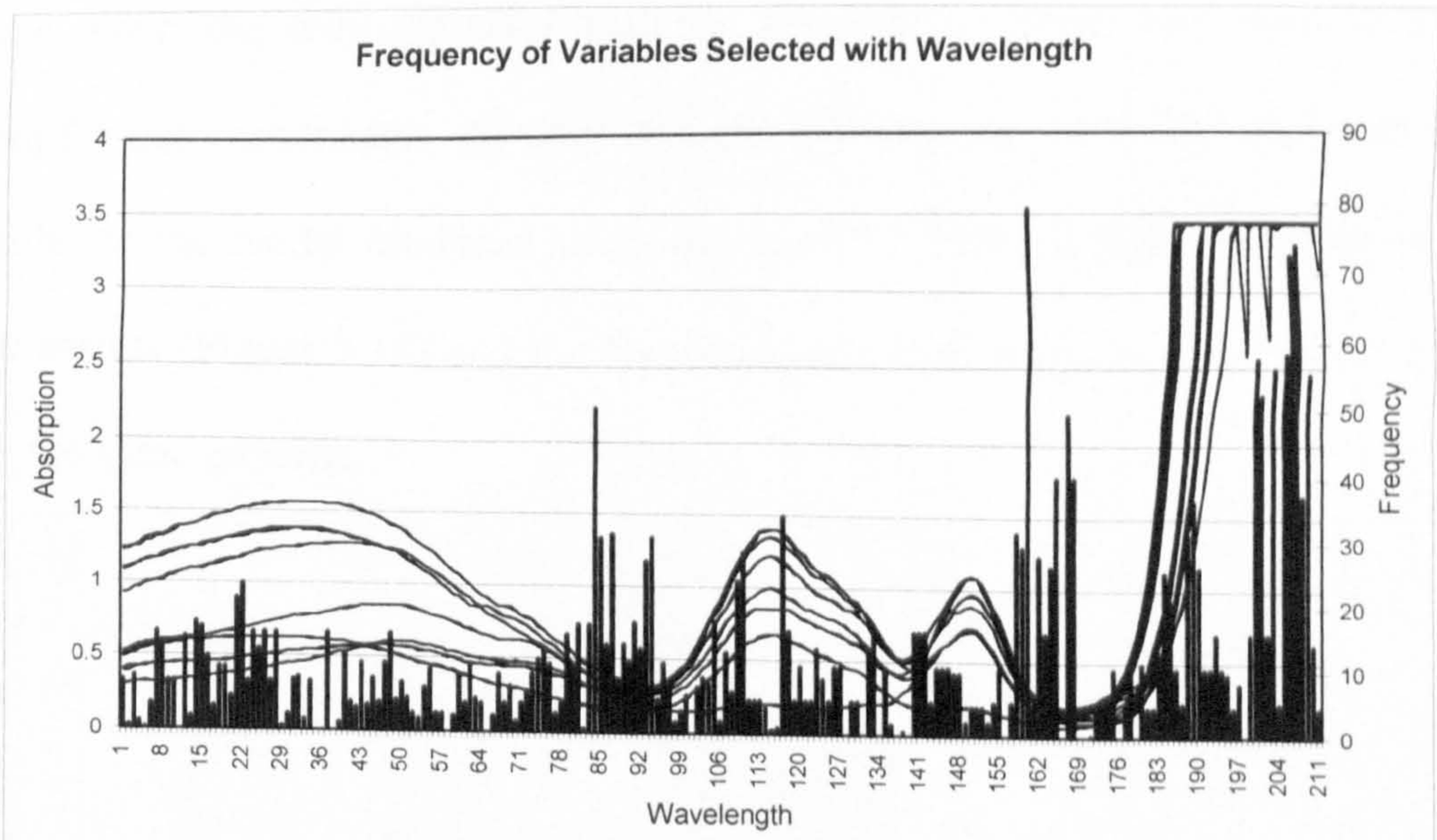


Figure 3.17 Frequency of variable selection vs. spectra for the UV data set using SVR

3.10.2. *Artificial Data Set 1*

Two hundred iterations were trained and the best model at that point was examined. The average number of variables selected over 200 iterations was 6, and the number of variables selected for the model with the lowest PRESS was 3.

Component	PRESS
Comp 1	3.01254
Comp 2	11.5584
Comp 3	0.05897
Comp 4	15.001

Table 3.11 PRESS results for the first artificial data set using SVR, 6 LVs were used, and the base PRESS was 29.63

This model is again worse for all the components apart from the linear third one, it is likely that again the only variables that are selected are those that improve the modelling for this component. As seen with the UV data set the model produced is very similar to the model produced using the third VS-PLS method, and both the predicted results (Figure 3.18) and the frequency of variable selection (Figure 3.19) are show the same patterns.

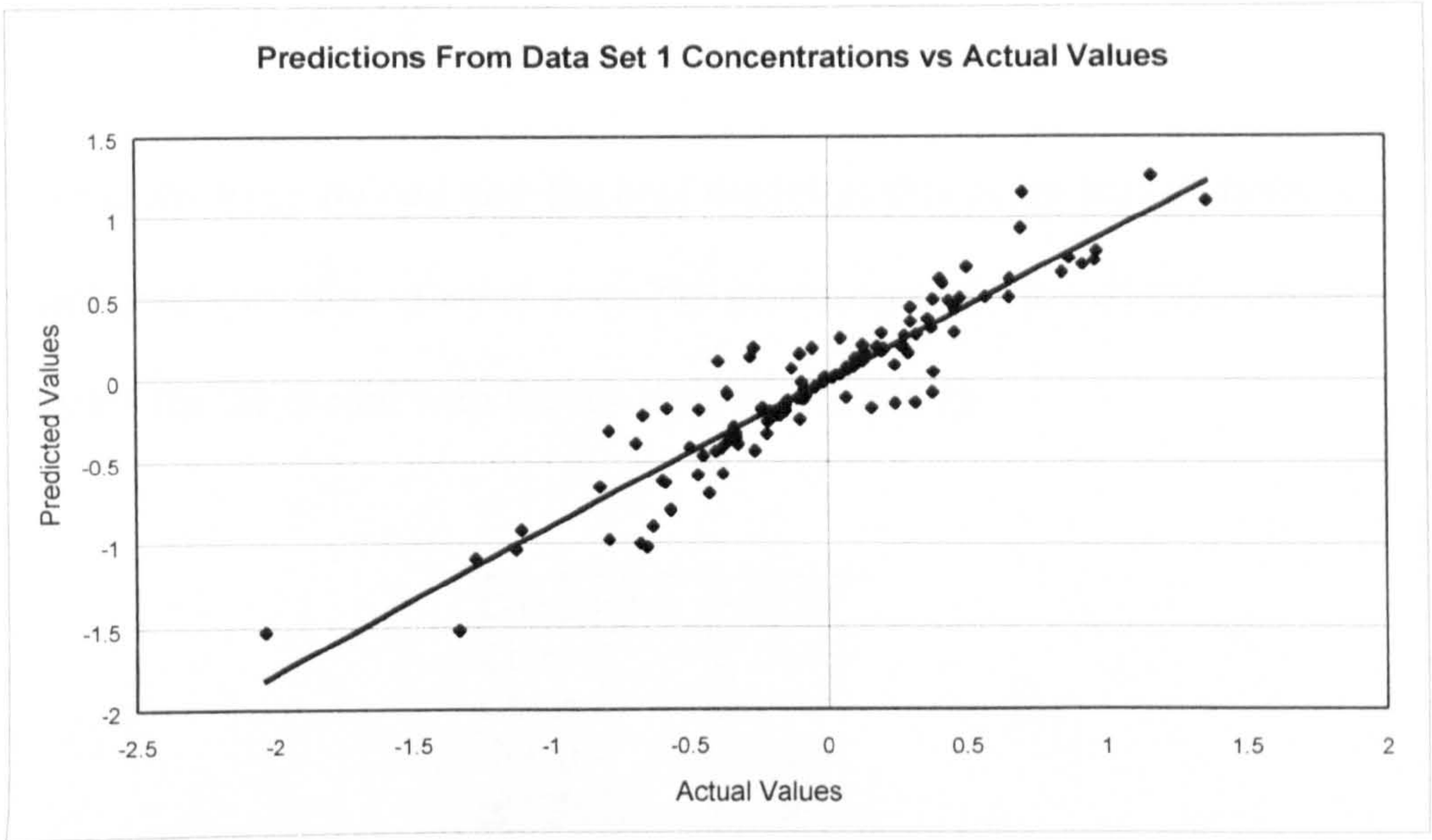


Figure 3.18 Prediction results for the first artificial data set using SVR

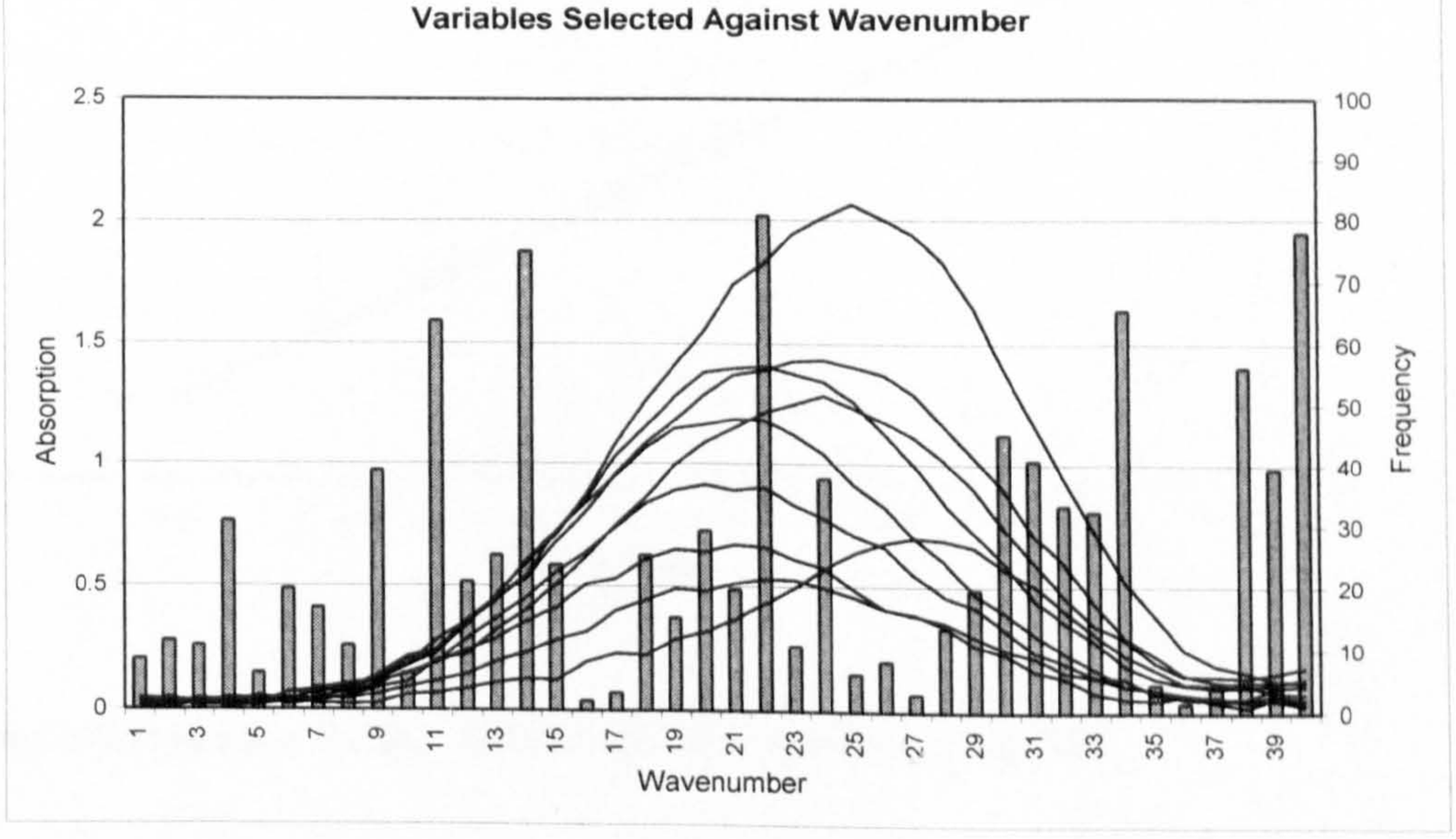


Figure 3.19 Frequency of variable selection vs. spectra for the first artificial data set using SVR

3.10.3. *Artificial Data Set 2*

Two hundred iterations were trained and the best model at that point was examined. The average number of variables selected over 200 iterations was 19, and the number of variables selected for the model with the lowest PRESS was 21.

Component	PRESS
Comp 1	0.07123
Comp 2	0.35441
Comp 3	0.11230
Comp 4	0.07400

Table 3.12 PRESS results for the second artificial data set using SVR, 4 LVs were used and the base PRESS was 0.62394

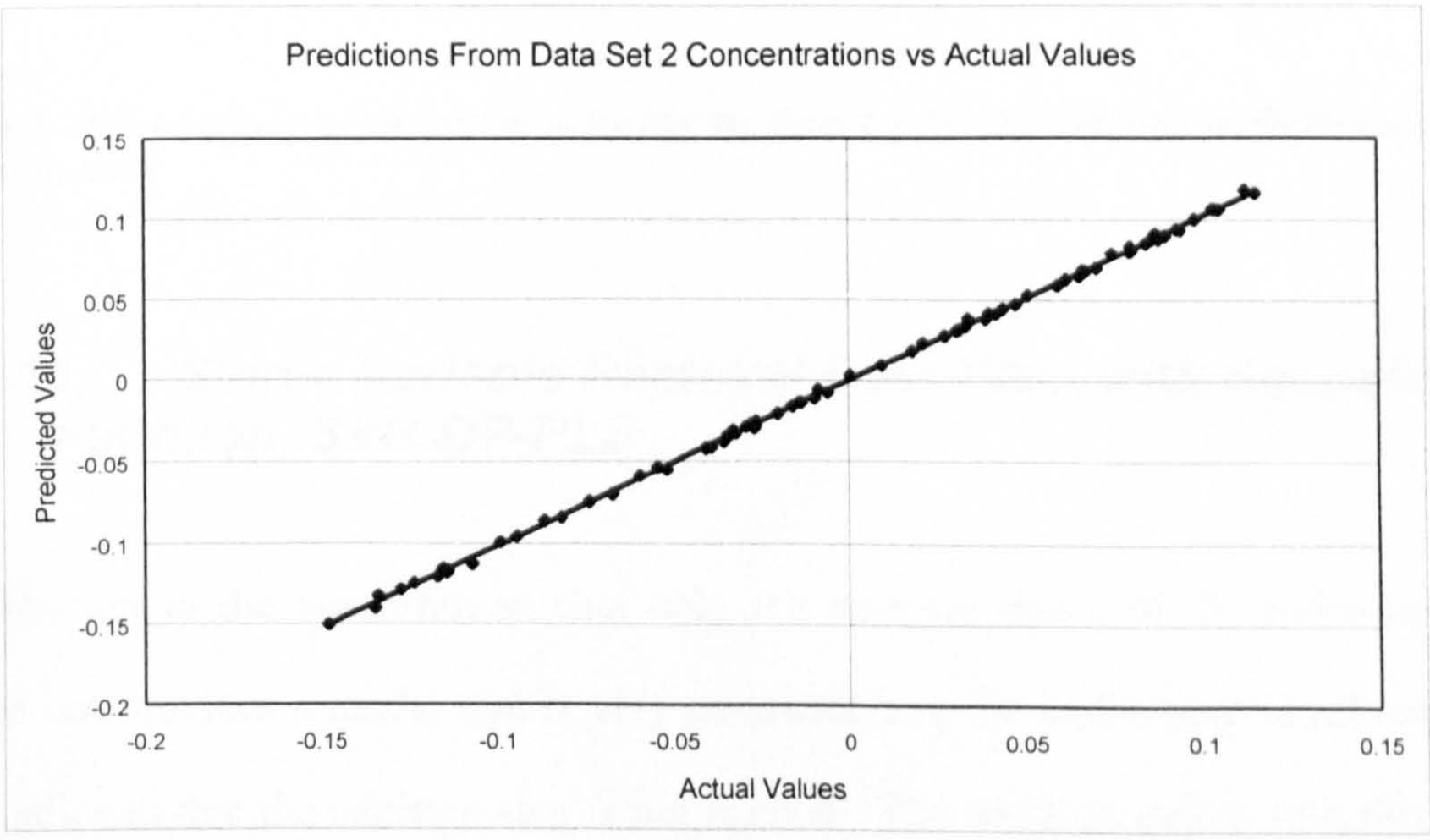


Figure 3.20 Predicition results for the second artificial data set using SVR

This data set shows the same pattern of results as the other data sets, this algorithm has produced very similar results to the addition-removal algorithm, the PRESS results (Table 3.12) the prediction results (Figure 3.20) and the variables selected appear to be very similar to the previous results. With this data set it may be difficult

to detect variation as the model is very good in all cases, however this does act to batch up the reasoning from the other two data sets.

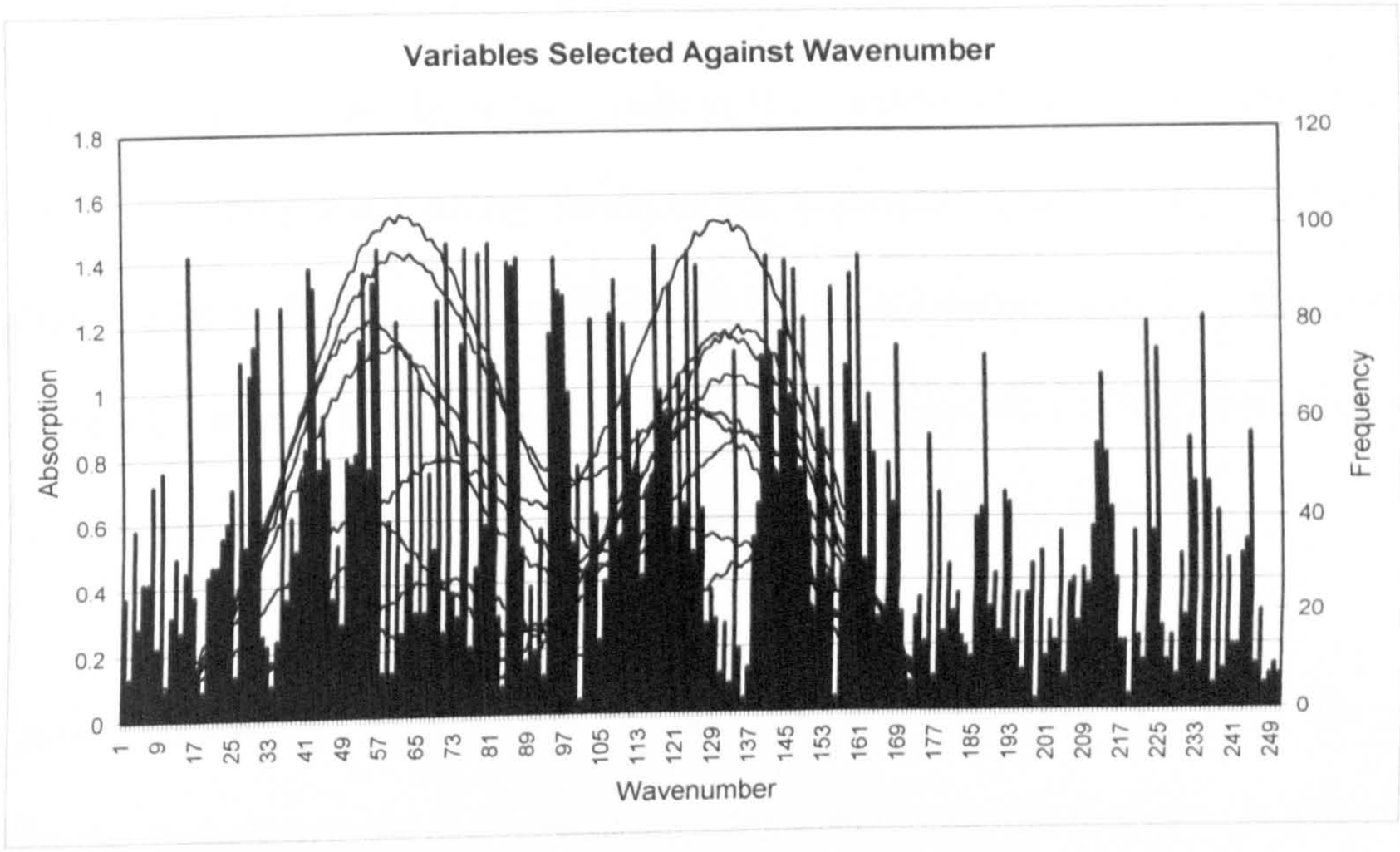


Figure 3.21 Frequency of variables selected vs. Spectra for the second artificial data set using SVR

3.11. Single Variable Removal Duel Pass with Squashing Function, SVR-DP-PLS

Modification to the algorithm so that only the removal phase of the procedure is carried out provides a model that is very comparable to the addition-removal model. This indicates that the addition step is not needed. The addition step is selecting the correct variables, as the variable the addition mode selects are the ones used for the removal mode, however the addition mode is selecting too great a number of variables. The removal stage is not removing enough variables. This is due to the sequence in which variables are presented to the model. The solution chosen for this is to re-shuffle the selected variables and repeat the variable removal procedure, by changing the sequence of the variables again this problem will be reduced. Another change added at the same time was a squashing function. When two nearly identical

variables are presented to the model the overall PRESS may be reduced by a small fraction, the reduction in model error may not be sufficient to warrant the inclusion of the second variable however the model has no procedure to reject the variable. A squashing function could be used to adjust the likelihood of a particular variable is selected. By using a squashing function the algorithm could be adjusted so that a variable is only included in the model when the model error drops by a fixed amount, a not when the model error is just a fraction smaller than the current best value for the PRESS. A squashing function is used in both variable removal stages.

The flow chart for this method can be seen in Figure 3.22 the squashing function is used in the comparison of the current best PRESS result and the new PRESS result, and is the scalar by which the new PRESS result must improve on the old one for the variable being tested to be included in the model. If the squashing function is less than one the new PRESS value will have to be that much smaller than the original for the variable to be added into the data set. There is little reason for a squashing function greater than one, this would tend to encourage variables to be included into the data set, which is not normally an issue.

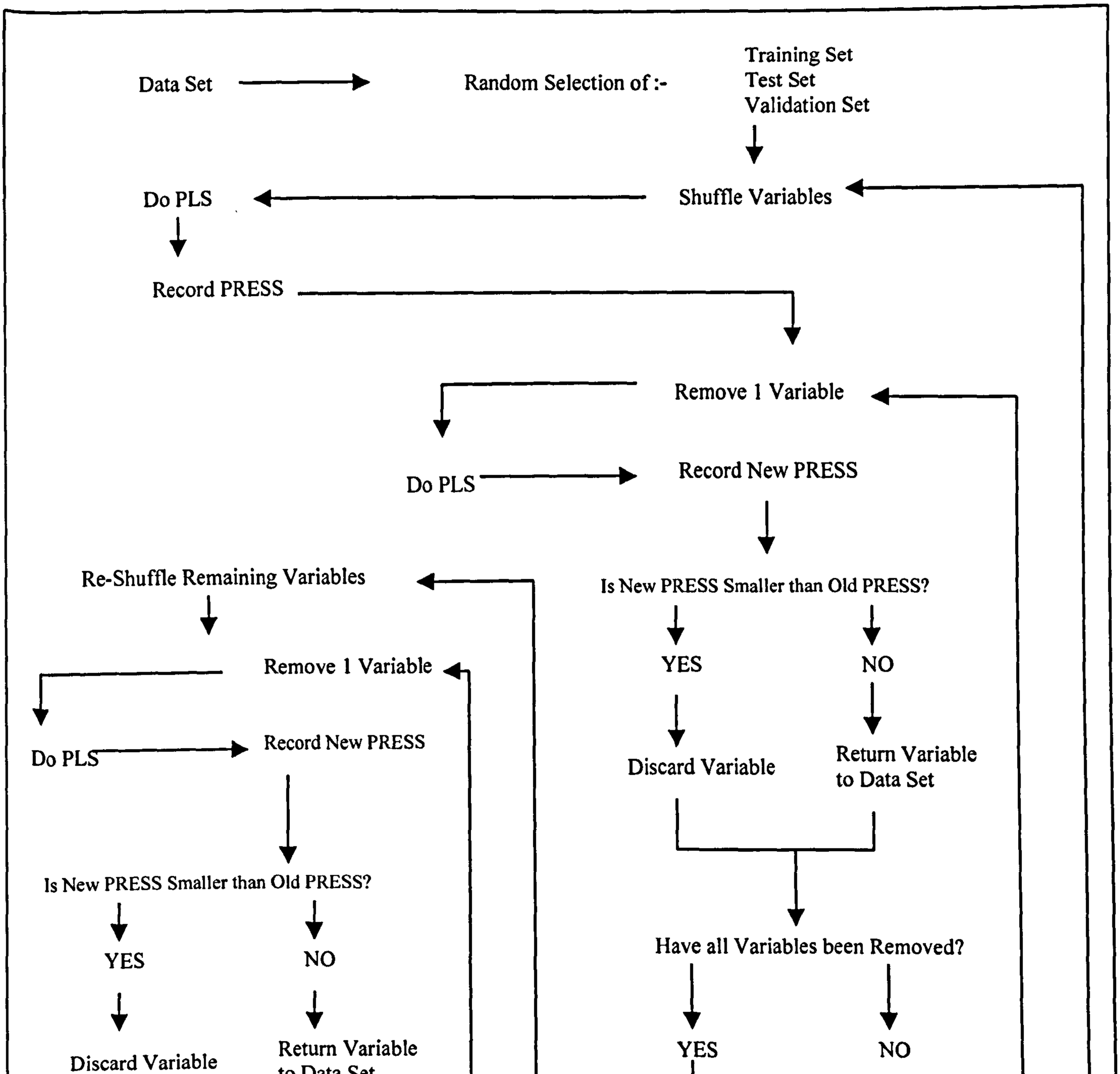


Figure 3.22 Flow Chart for Single Variable Removal Duel Pass, SVR-DP-PLS

3.11.1. Matlab Code for the final VS-PLS method, SVR-DP-PLS

The code for this algorithm in MATLAB ® can be seen in Appendix II. The code as displayed will work with MATLAB ® 5.2 provided that the PLS Toolbox 1 or PLS Toolbox 2 from Eigenvector Research is also available.

3.11.2. UV Data Set

After 200 iteration training the best model up to that point was examined.

Average number of variables selected over 200 iterations, 13. Variables selected for the model with the lowest PRESS, 11.

Component	PRESS
Fe	3.7648
Co	0.0440
Ni	0.1563
Cu	0.0046

Table 3.13 PRESS results for the UV data set using SVR-DP, 7 LV's were used, and the base PRESS was 3.7366

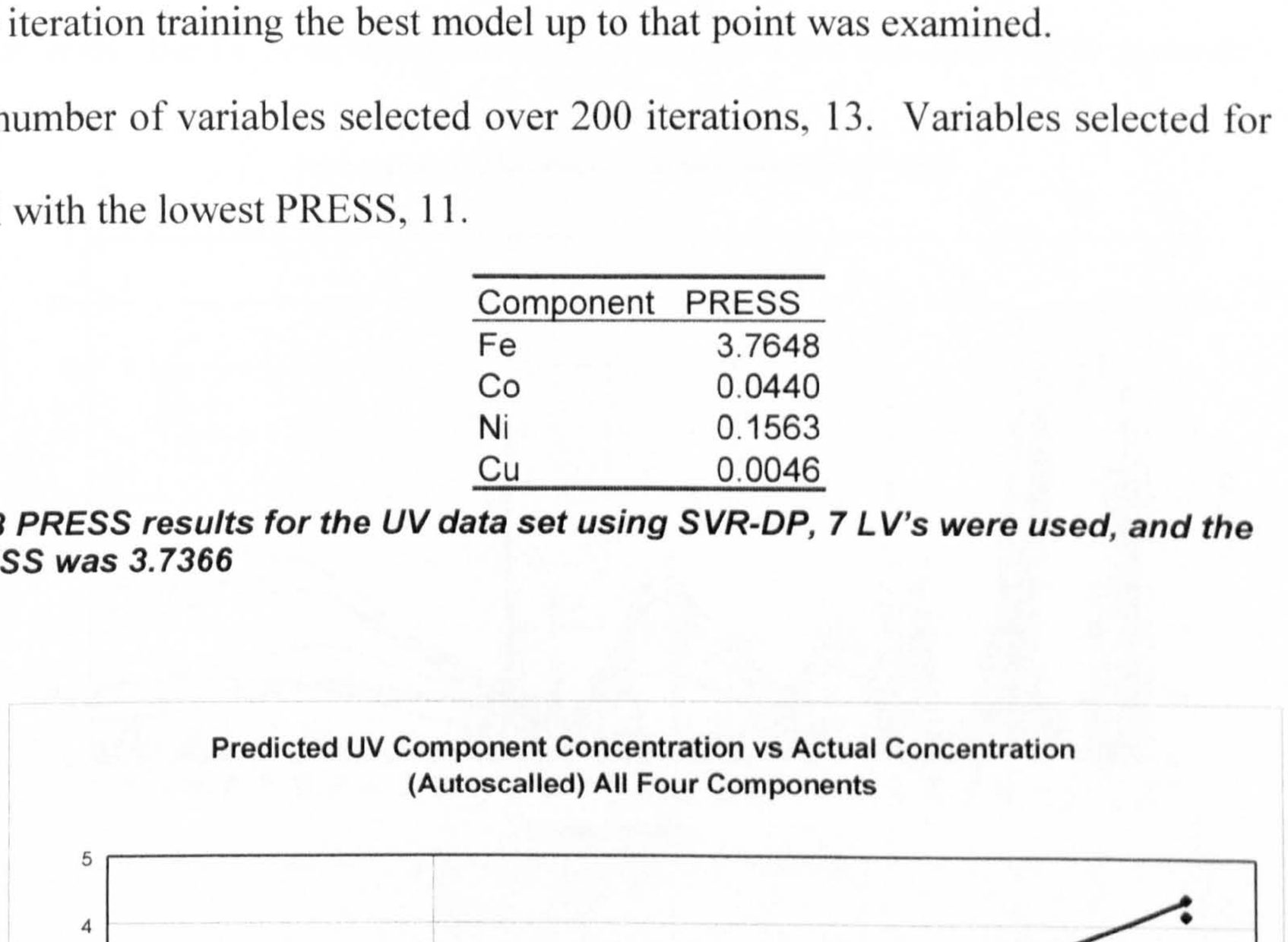


Figure 3.23 Prediction results for the UV data set using SVR-DP

This model shows a significant improvement over the previous method, PRESS values (Table 3.13) for all the components have dropped by a significant amount compared to previous values, the prediction results (Figure 3.23) also appear better. The biggest change however is in the plot showing variables selected (Figure 3.24), the "background" variables, those that are selected infrequently has dropped, leaving only the larger peaks behind. There are only three major groupings of variables, together with the grouping in the noise, it is likely that the variables containing the information about the Fe component are very similar and are still difficult to separate.

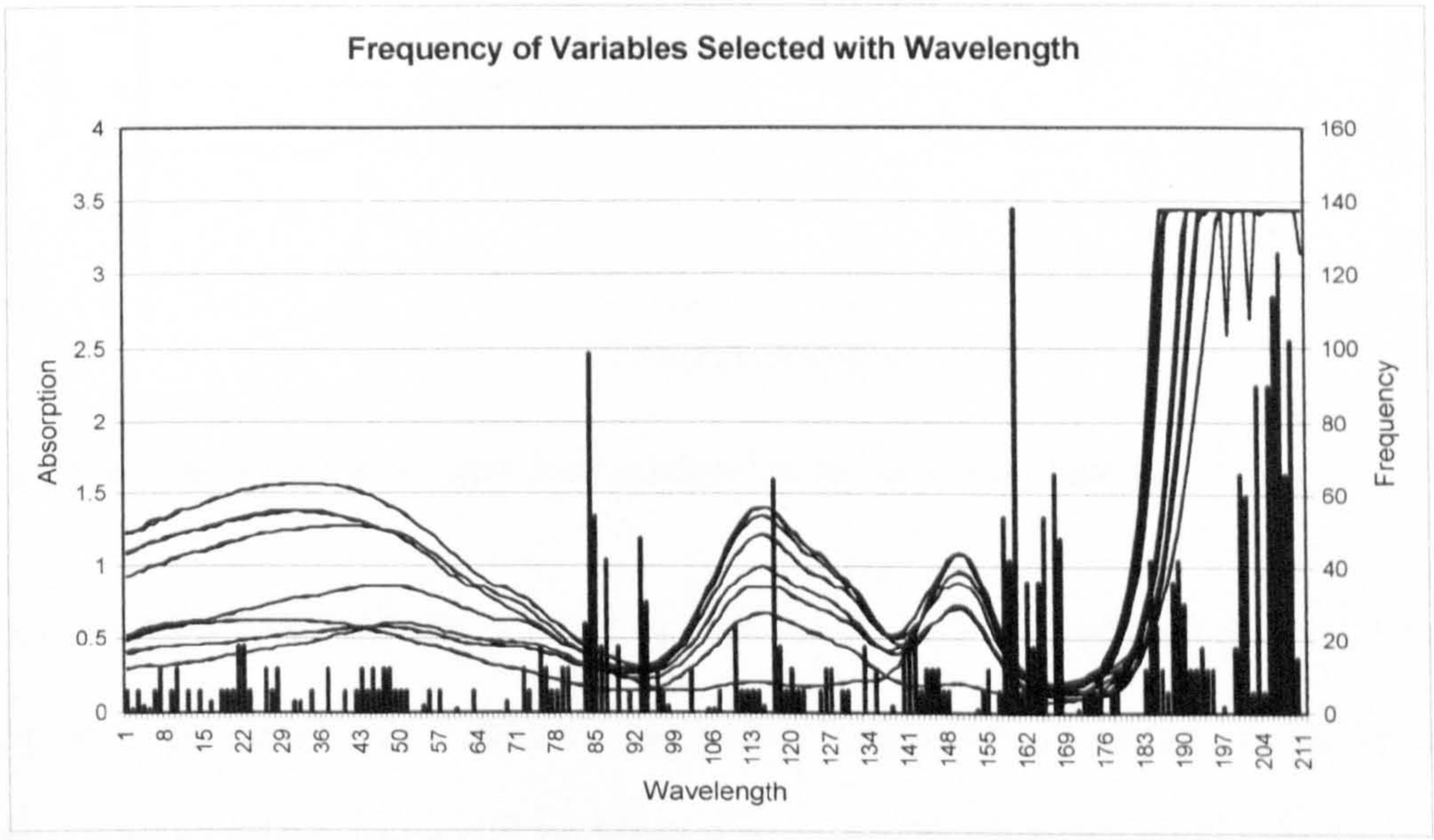


Figure 3.24 Frequency of variable selection for the UV data set using SVR-DP

3.11.3. Artificial Data Set 1

Two hundred iterations were trained and the best model at that point was examined. The average number of variables selected over 200 iterations was 5, and the number of variables selected for the model with the lowest PRESS was 2.

Component	PRESS
Comp 1	4.1828
Comp 2	16.9242
Comp 3	0.0002
Comp 4	24.6406

Table 3.14 PRESS results for the first artificial data set using SVR-DP, 6 LV's were selected, and the base PRESS was 45.7136

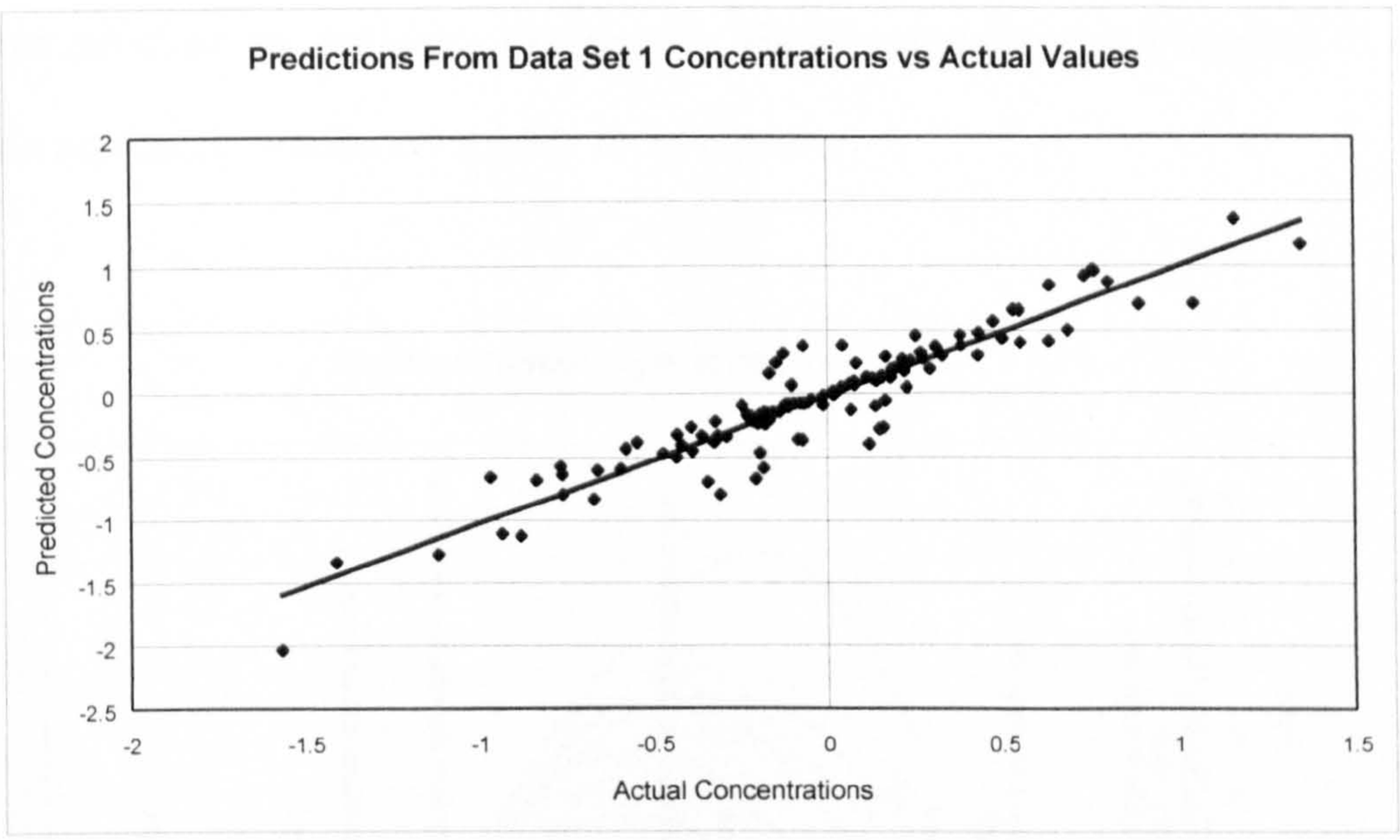


Figure 3.25 Prediction results for the first artificial data set using SVR-DP

In terms of overall PRESS (Table 3.14) this is the worst model produced so far. This would appear to be caused by an almost constant PRESS contribution form the non-linear components during the model building, the model being influenced only by the relatively small changes in the PRESS for the linear component. The linear component is modelled very well despite this (Figure 3.25), with the lowest error for any of the other models built with this data set. This model has been built with most of the influences from the non-linear variables removed, and suggests that the current method can resolve the influences from many different sources of error, this model behaves as if the non-linear components are a source of error for the linear component, which is one possible way of interpreting this data set. The histogram showing the frequency of variables selected (Figure 3.26) indicates that the variables

selected for this model were chosen from the edges of the peak for information concerning overlaps and from the centre of the peak for magnitude information. This does not show that same degree of organisation that the histogram for the UV data set shows however it does clearly indicate that even in a crowded peak such as is present in this data set there are variables that provide significantly more information to a model than apparently similar ones fairly close together.

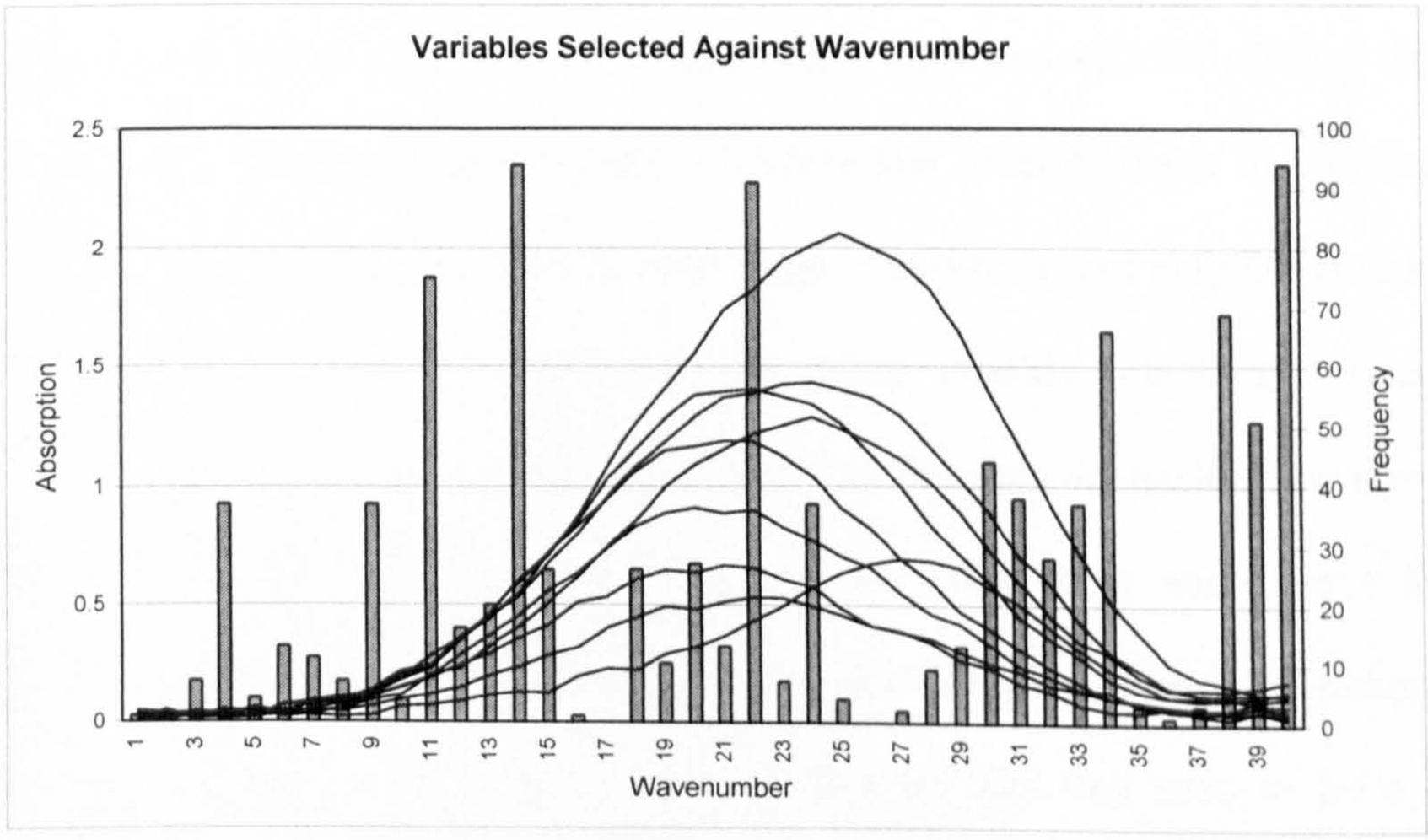


Figure 3.26 Frequency of variables selected for the first artificial data set using SVR-DP

3.11.4. Artificial Data Set 2

Two hundred iterations were trained and the best model at that point was examined. The average number of variables selected over 200 iterations was 12, and the number of variables selected for the model with the lowest PRESS was 12.

Component	PRESS
Comp 1	0.004063
Comp 2	0.003111
Comp 3	0.003802
Comp 4	0.004823

Table 3.15 PRESS results for the second artificial data set using SVR-DP, 4 LV's were used, and the base PRESS was 0.016

In comparison with the previous method (single variable removal) this method appears far more efficient in removing unwanted variables from the group of selected variables, with a data set hat contains a very high percentage of co-linear variables it is expected that there will be a lot of redundant variables selected during the first removal stage, this is increased by the relatively low level of noise in this data set. The PRESS results show that there is an average of two orders of magnitude reduction in the error of prediction for this data set using variable selection compared to ordinary PLS. This series of tests have shown that this data set has far less error than might be expected in any real data set but this does show some limit for the effectiveness of this algorithm, the reduction in error is comparable to the reduction in error seen for component four for the UV data set and can seen to be a useful comparison with real data sets built with high quality data. This model together with the model for the UV data set using this method (Section 3.10.2) show that the variable selection procedure described here is very efficient at determining appropriate variables to select for a robust model. The predicted results (Figure 3.27) give very little further information, however the variables selected (Figure 3.28) are showing that the selection is weighted to various sections of the data set.

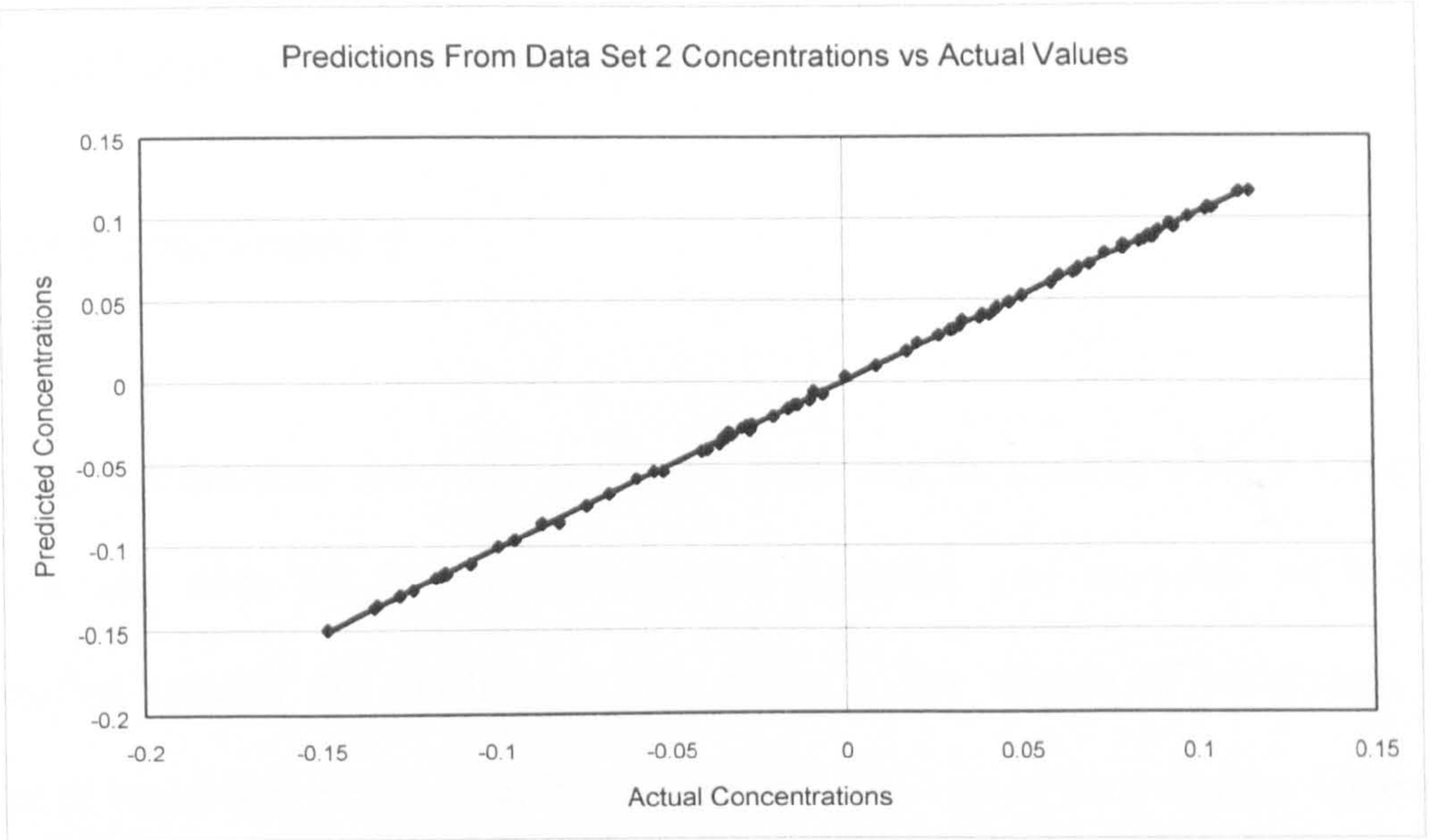


Figure 3.27 Prediction results for the second artificial data set using SVR-DP

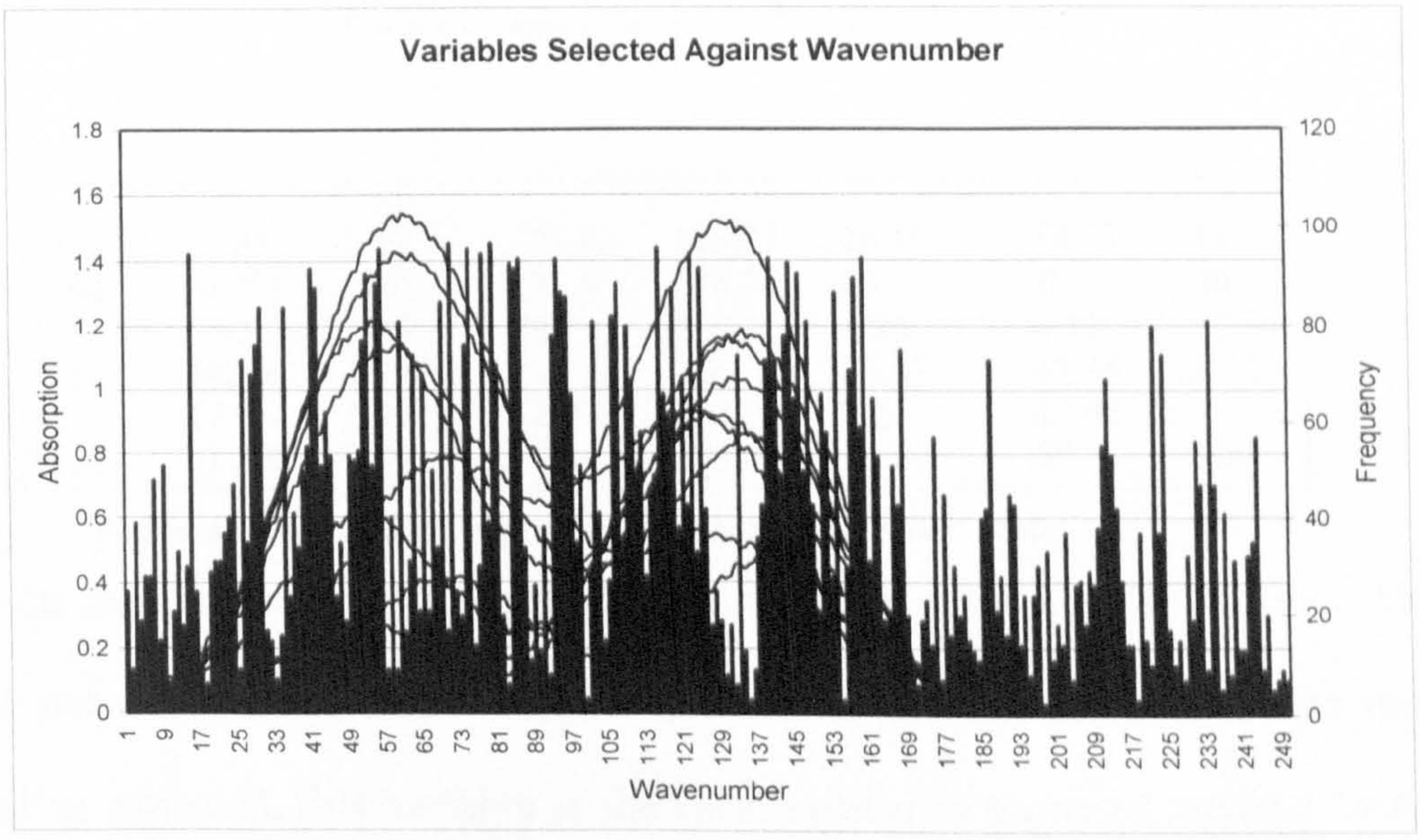


Figure 3.28 Frequency of variables selected for the second artificial data set using SVR-DP

4 Intrasite Gel Results and Discussion

4.1 Intrasite Experiment 1

The basic statistics of the data sets were examined, these can be seen in table 4.1 for the sterilisation data, and table 4.2 for the Intrasite Gel Analysis. As expected for a fully feedback controlled system, the sterilisation data shows a low degree of variability, the greatest amount of variability can be seen in the quantity. The rest of the variables appear to show low variability in comparison.

	QTY	*Heat up Time*	*Pressure*	*Min T°C*	*Max T°C*	*Hold Time*	*Cool Time*	*F(0)*
Mean	5244.73	33.85	3.12	121.42	123.23	26.55	18.63	51.96
Median	5888	34.35	3.202	121.4	123.2	30	17	56.85
Standard Deviation	1481.72	5.67	0.16	0.07	0.26	6.94	4.33	11.54
Range	7717	74.29	1.753	1.6	3.2	33.87	47.35	88.5
Minimum	441	17	2.037	121	122	7.2	12.05	7
Maximum	8158	91.29	3.79	122.6	125.2	41.07	59.4	95.5

Table 4.1 Basic Statistics for Batch Sterilisation Data

The analysis results show a high variability, the highest of which is the fluid absorption. This reflects the high degree of noise in the measurement. The pH measurement is the most highly controlled as expected, this variable is the most critical in terms of medical safety. The huge variation in means and variance that these variables display indicate that any modelling carried out should be preceded by autoscaling of the data set.

	pH	Elasticity	Viscosity Coefficient	Solids Content	Fluid Absorption
Mean	6.83	2107.30	356.73	2.76	158.92
Median	6.8	1978	369	2.8	110
Standard Deviation	0.17	570.34	103.65	0.29	91.52
Range	1.1	2584	415	1.7	315
Minimum	6.4	887	196	2	60
Maximum	7.5	3471	611	3.7	375

Table 4.2 Basic Statistics for Intrasite Gel Material Analysis

4.2 Intrasite Experiment 2

4.2.1 Fluid Absorption Distribution

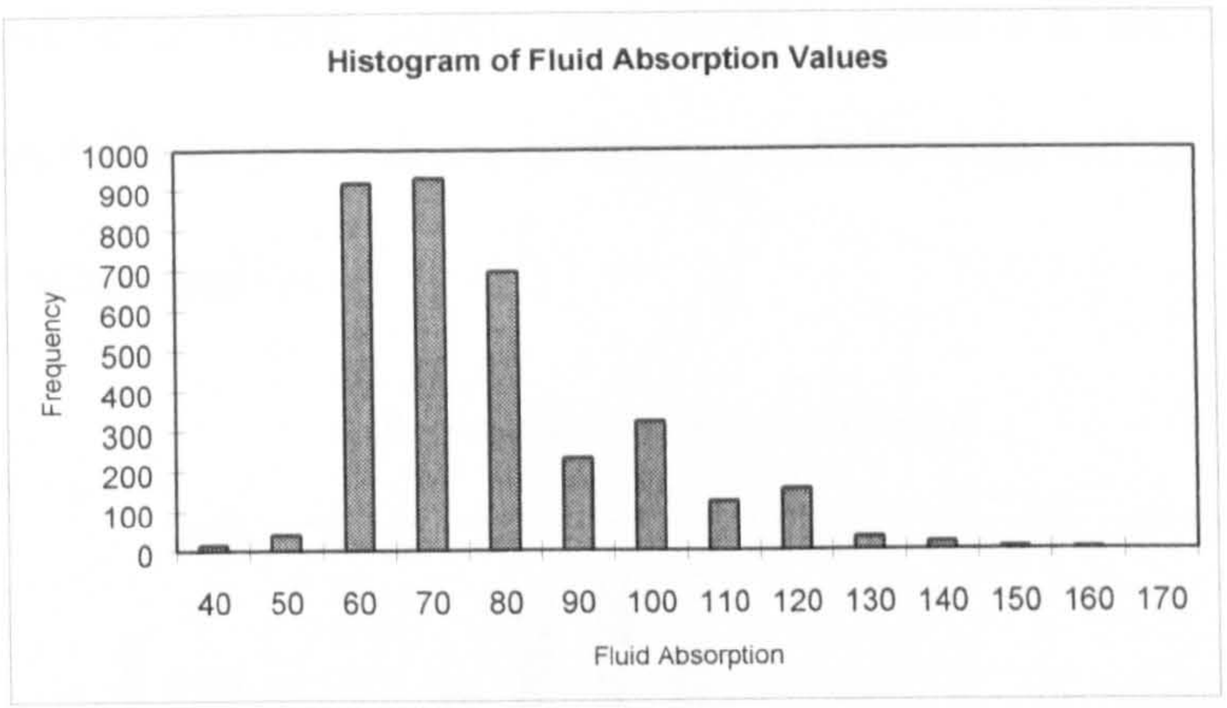


Figure 4.1 Bar Chart Showing Distribution of Values for Fluid Absorption

The fluid absorption (mls 0.9% NaCl / 100g) distribution (Figure 4.1) appears to be a combination of two separate means, one centred on a fluid absorption of 120, corresponding to the results obtained from the original formulation, and a second skewed distribution centred on a fluid absorption of 65. The reason for the two separate distributions is the formulation change that occurred in 1994. The exact details of the formulation change are covered by confidentiality agreements, however its effect can be clearly seen in all the distribution graphs. For each variable the expected distribution for the observed means and standard deviations are also plotted. The second, lower distribution appears to be skewed, one possible explanation is that since the value of 60 represents the lower limit for the specification for fluid absorption for Intrasite gel there is some pressure on analysts to determine that the value for fluid absorption is at least this value. There is no evidence to support this, and a more likely explanation is that this represents effect of producing the product to this specification. Figure 4.2 Shows the expected distribution for fluid absorption values given a normally distributed data set. A value of 55 for fluid absorption represents the lower limit for the release specification. One hypothesis could be that this lower limit is the

cause for the skewed distribution, either representing the a deliberate skewing of the data by the analyst to force the material to pass specification, or this is some feature caused by the manufacturing process. The former hypothesis is highly unlikely as Smith & Nephew adhere to strict control on analytical quality, and the SC1 variable is also subject to a specification limit of 4.5, and shows no evidence of this type of skewing, which would be expected if this was a feature of the analyst.

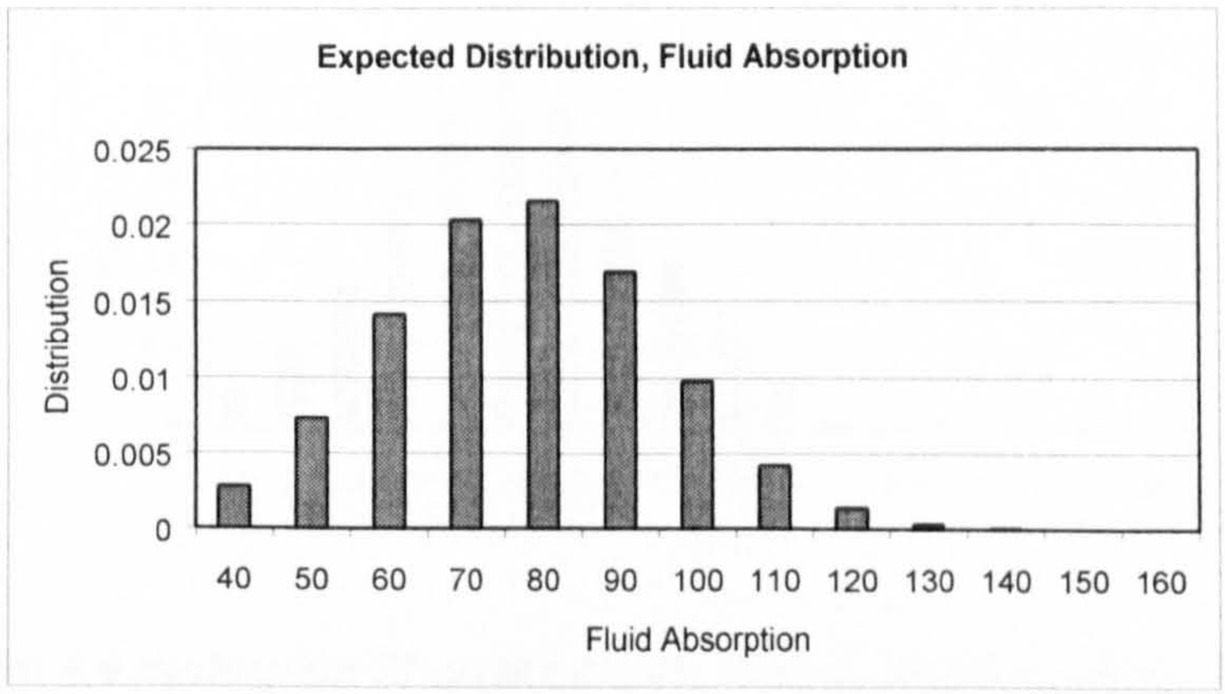


Figure 4.2 Histogram Showing Expected Distribution for Fluid Absorption Values

4.2.2 SC1

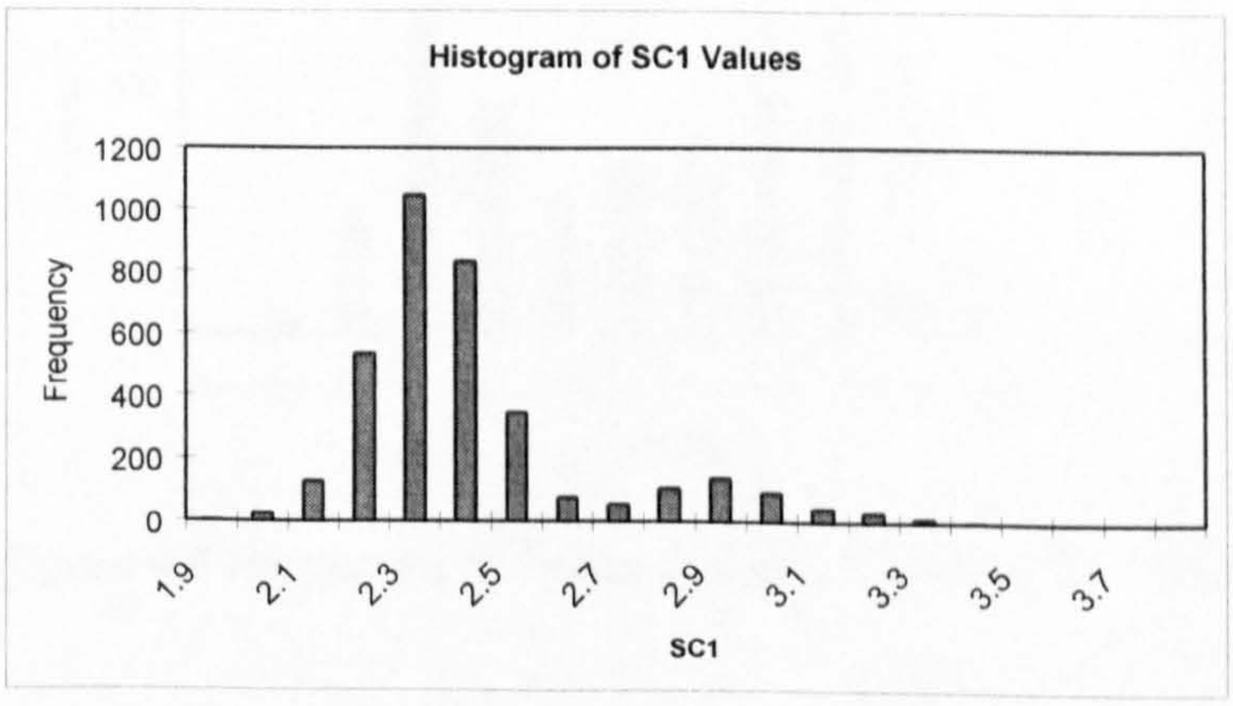


Figure 4.3 Histogram Showing the Distribution for SC1 Values

The bimodal distribution for SC1 (Figure 4.3) represents the effect of a change in the specification for the product, SC1. SC1 specification changed from 2.8 to 2.3 in 1994, originally Intrasite gel was manufactured to several different specifications to meet the

requirements of several different segments of the global market, however during the early 1990's the varying specifications were unified, meaning that only a single standard for the product existed after 1994. This change is the underlying reason for the bimodal distributions evident in all the distributions from the other variables. Figure 4.4 shows the expected distribution for this data.

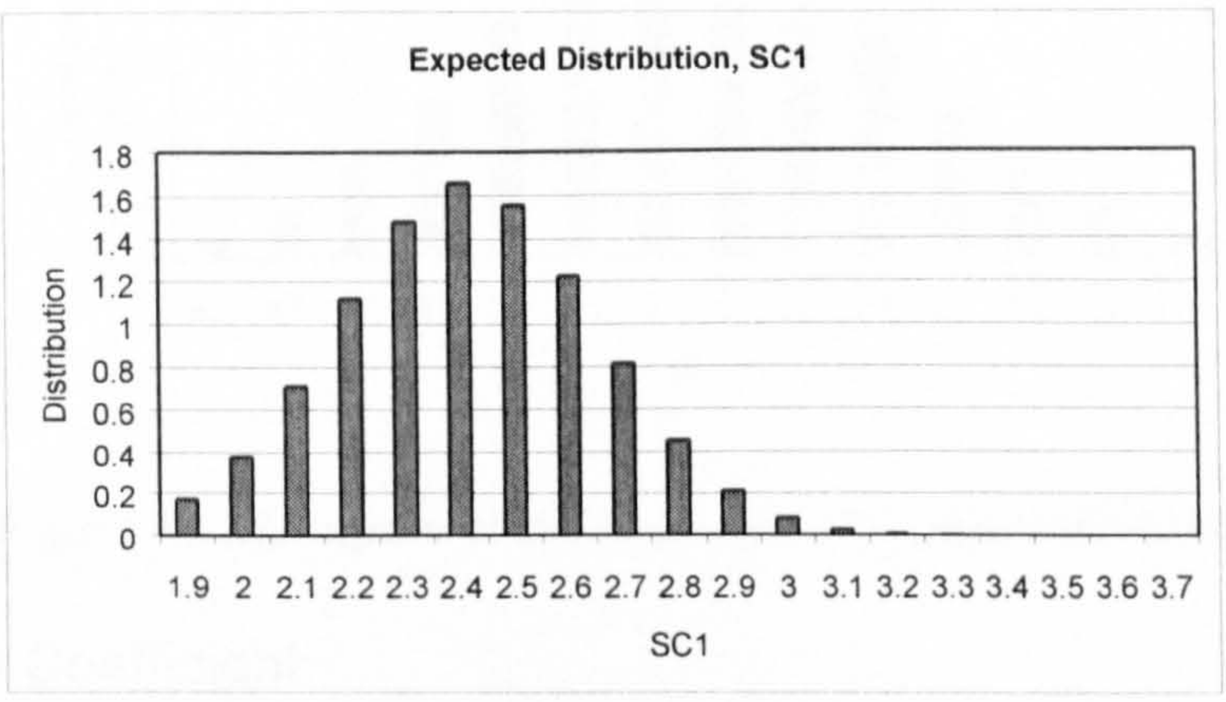


Figure 4.4 Histogram Showing the Expected Distribution for SC1 Values

4.2.3 pH

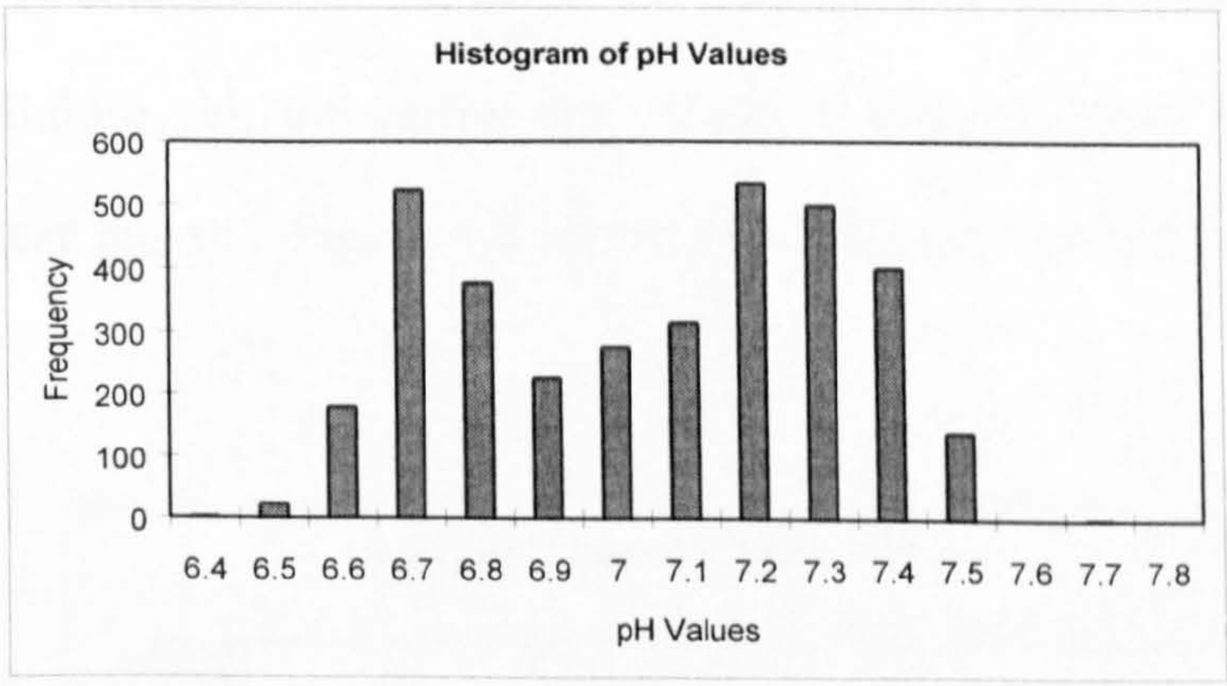


Figure 4.5 Histogram Showing the pH Distribution of Values

While it was not expected that the pH would be effected by the change in SC1, the distribution (Figure 4.5) shows that a bimodal distribution does exist in the pH variable. It should be noted however that the distribution appears to be even, unlike the distribution observed in the other variables. The pH variable does not show the clear change in means that occurred in the other variables. The pH may have been affected by a formulation change

in the manufacture of the raw polymer, which is known to change however no evidence is available to examine this possibility. Figure 4.6 shows the expected distribution for the pH values.

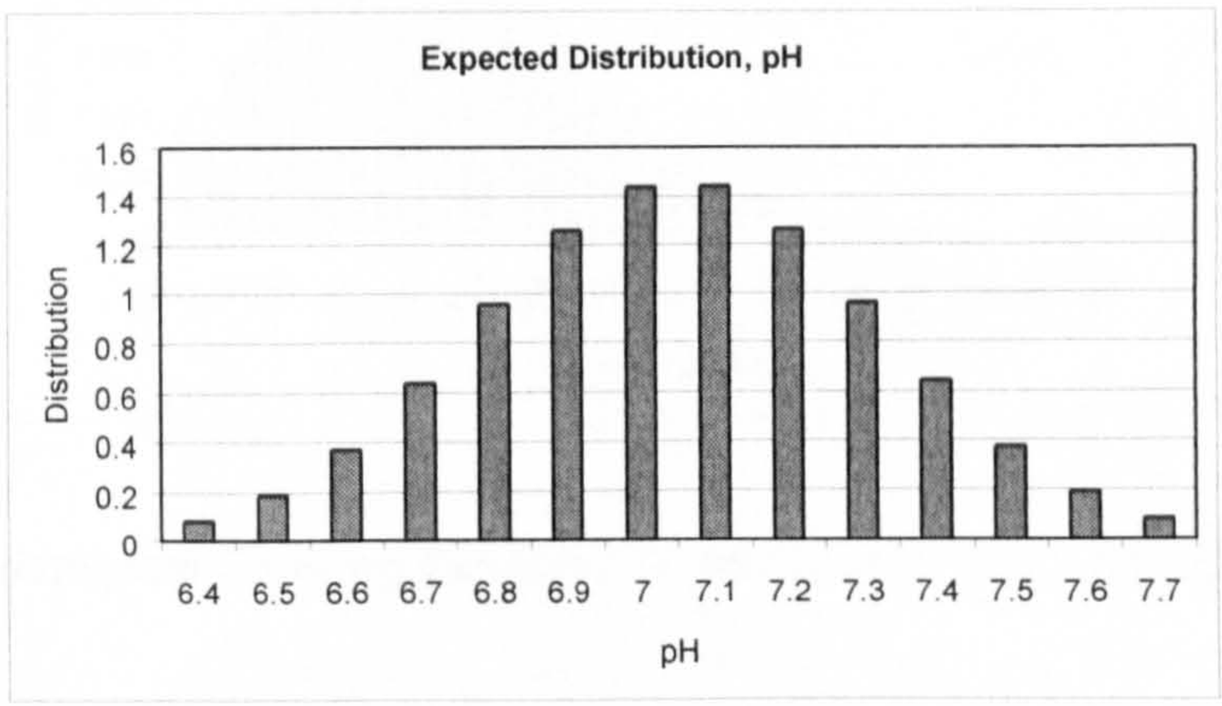


Figure 4.6 Histogram of the Expected pH Value Distribution

4.2.4 Viscosity Coefficient

The viscosity coefficient shows the same distribution as the other variables (Figure 4.7), though the viscosity coefficient is known to be close to the non-linear region for this type carboxymethyl cellulose gel, the earlier data shows a high viscosity coefficient possibly within the non-linear region. Figure 4.8 shows the expected viscosity distribution for this data set.

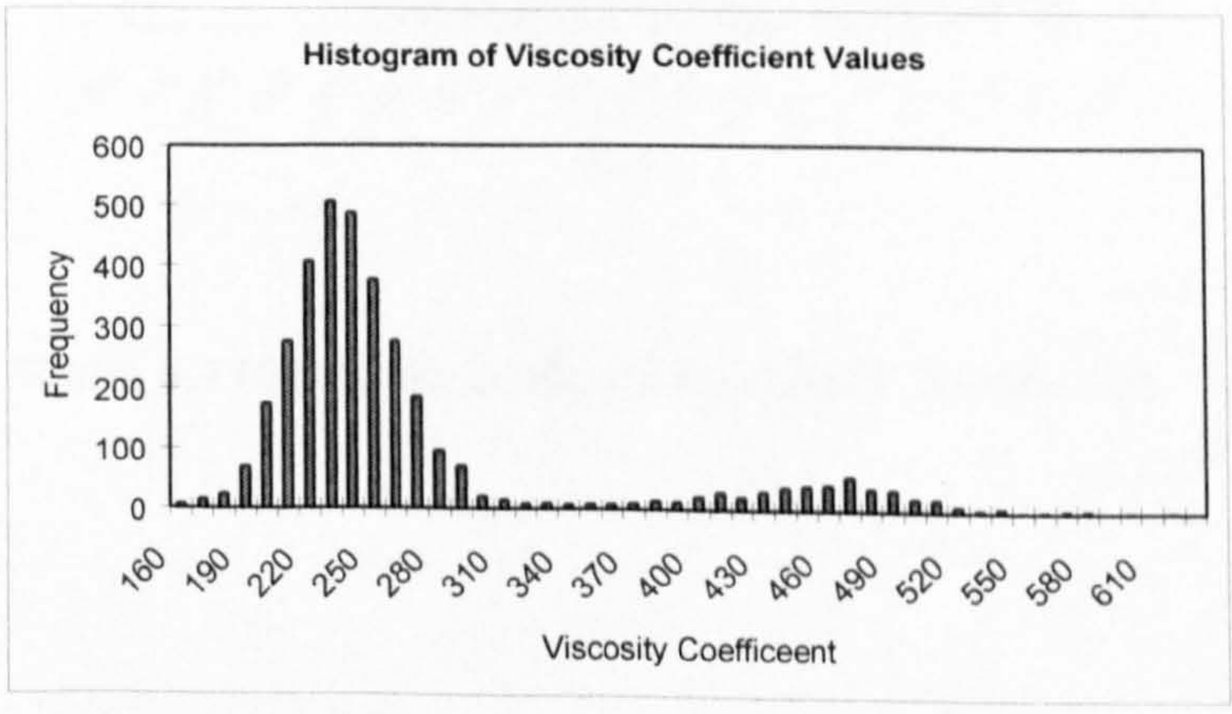


Figure 4.7 Histogram Showing the Distribution for Fluid Absorption Values

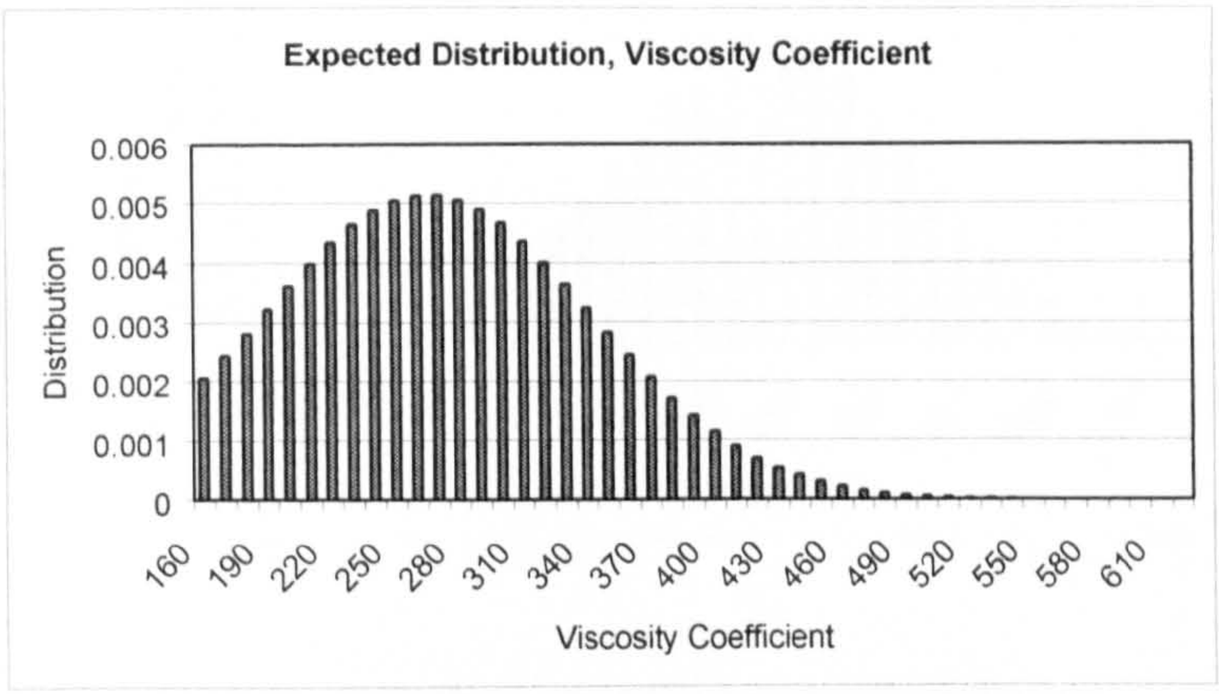


Figure 4.8 Histogram Showing Expected Distribution for Fluid Absorption Values

4.2.5 Elasticity

The elasticity (Figure 4.9) has the same distribution as the other variables although it is most similar to the viscosity coefficient (Figure 4.9) as might be expected. The expected elasticity distribution can be seen in figure 4.10.

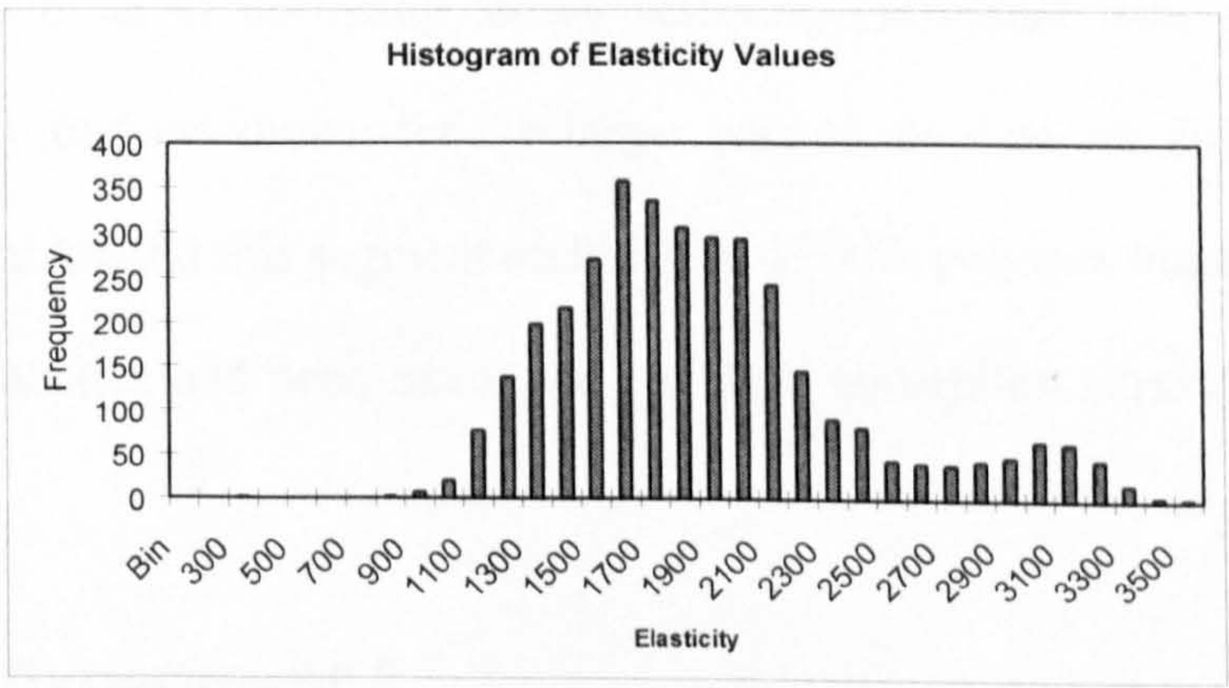


Figure 4.9 Histogram of Elasticity Values Distribution

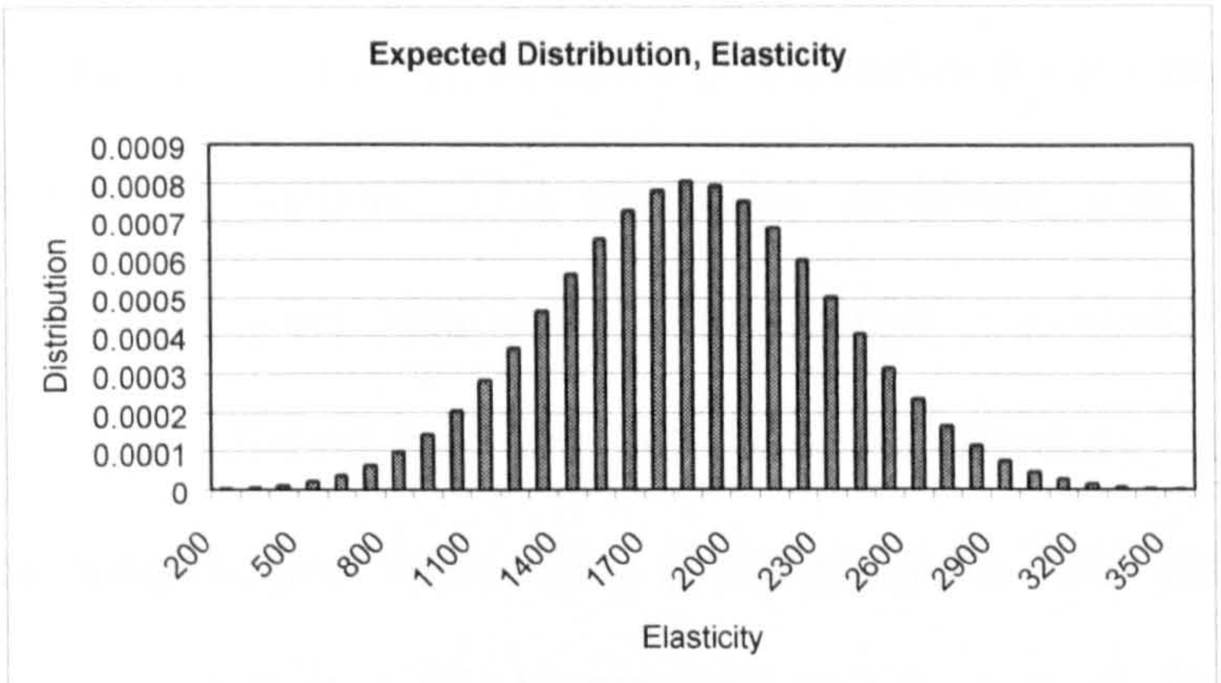


Figure 4.10 Expected Elasticity Distribution

The fact that four of the five variables show the same apparent distribution suggests that the same influences are affecting each variable in a similar manner, with the exception of the pH. The bimodal distribution does not preclude the possibility of a global model for all the available data however it does indicate that modelling would be more straight forward if a section of the data were taken that has a normal distribution. The data to be examined could be selected on the basis of normality, or by selecting individual bulk polymer batches to examine. The distributions shown for the larger part of the data set, July 1994 onwards is normal for all variables and this segment includes three bulk polymer batches.

After the initial statistics had been examined the fluid absorption variable was examined in detail.

4.3 Intrasite Experiment 3

	pH	*Elasticity*	*Viscosity Coefficient*	*SC1*	*Fluid Absorption*
pH	1				
Elasticity	-0.09	1			
Viscosity Coefficient	-0.39	0.84	1		
SC1	-0.40	0.62	0.90	1	
Fluid Absorption	-0.27	0.21	0.76	0.72	1

Table 4.3 Correlation Coefficients Between the Analysis Variables

At this stage the fluid absorption variable is the focus, and the correlation coefficients (Table 4.3) show an interesting disparity in values. The viscosity coefficient and the elasticity have

a relatively high correlation to each other, however that correlation does not transfer directly to either SC1 or the fluid absorption. This raises that possibility that the two variables combined may describe a significant portion of the information contained in either SC1 or in the fluid absorption. The relation between SC1 and the viscosity coefficient is also remarkably high considering the poor relationship exhibited by the other variables. This may indicate that a calibration may be possible between the viscosity coefficient and SC1. The very poor correlation between the pH and the other variables mirrors the disparity in the distributions. It is also possible that the correlation coefficients may be considered in a different way, the clear bimodal distribution may effect the correlation coefficients, as the two separate means could have the effect of large leverage values. This is examined by looking at the correlation between the variables in only the second, normal, section of the data (Table 4.4).

	pH	Elasticity	Viscosity Coefficient	SC1	Fluid Absorption
pH	1				
Elasticity	-0.2	1			
Viscosity Coefficient	-0.1	0.84	1		
SC1	-0	0.61	0.76	1	
Fluid Absorption	0.1	0.31	0.41	0.38	1

Table 4.4 Correlation Coefficients for Data from July 1994 - December 1997

The new values for the correlation coefficients are strong evidence that at least part of the high correlation coefficients experienced before was due to some form of leverage effect. The correlation coefficient between the elasticity and the viscosity has remained the same showing that any relationship between these two variables is similar in both the whole data set and the small set selected.

The correlation between the pH and SC1 shows that no relationship exists between these two variables. This together with the correlations with the other variables is further evidence that the bimodal distribution observed in the pH variable is a coincidence and not evidence of a possible relationship between the pH and any of the other variables.

4.4 Intrasite Experiment 4, Regression Modelling

Regression was performed against the fluid absorption variable. Fluid absorption is both the primary interest for assessing the properties of Intrasite Gel, it is also the variable that is known to contain the most error. Replacing the variable with a calculated result would useful. With the known error in the recorded variable, there is no expectation of a particularly high quality or robust model initially; the first regression calculations were carried out to determine whether there was any reason to continue to explore the possibility of calculating the fluid absorption rather than measuring it.

Standard MLR was carried out using the full data set available. This was then repeated using the data set that had been elected as normal in the examination of the distribution of the data, and finally MLR was carried out on a single bulk batch to examine any differences between the results for a single batch compared with the results from several batches that appeared to have a common mean.

4.4.1 MLR on the full data set

The full data set was shuffled randomly by sample and seventy percent used for the regression calculations, the remaining thirty percent was used as a validation set, that is 2700 samples used in calibration, 1200 samples used in validation. This data set spanned the period January 1993 through December 1997. The calculations were carried out in Matlab on the raw data. No smoothing is appropriate for process analysis data, and scaling methods are unlikely to effect the results of MLR on such an over determined data set.

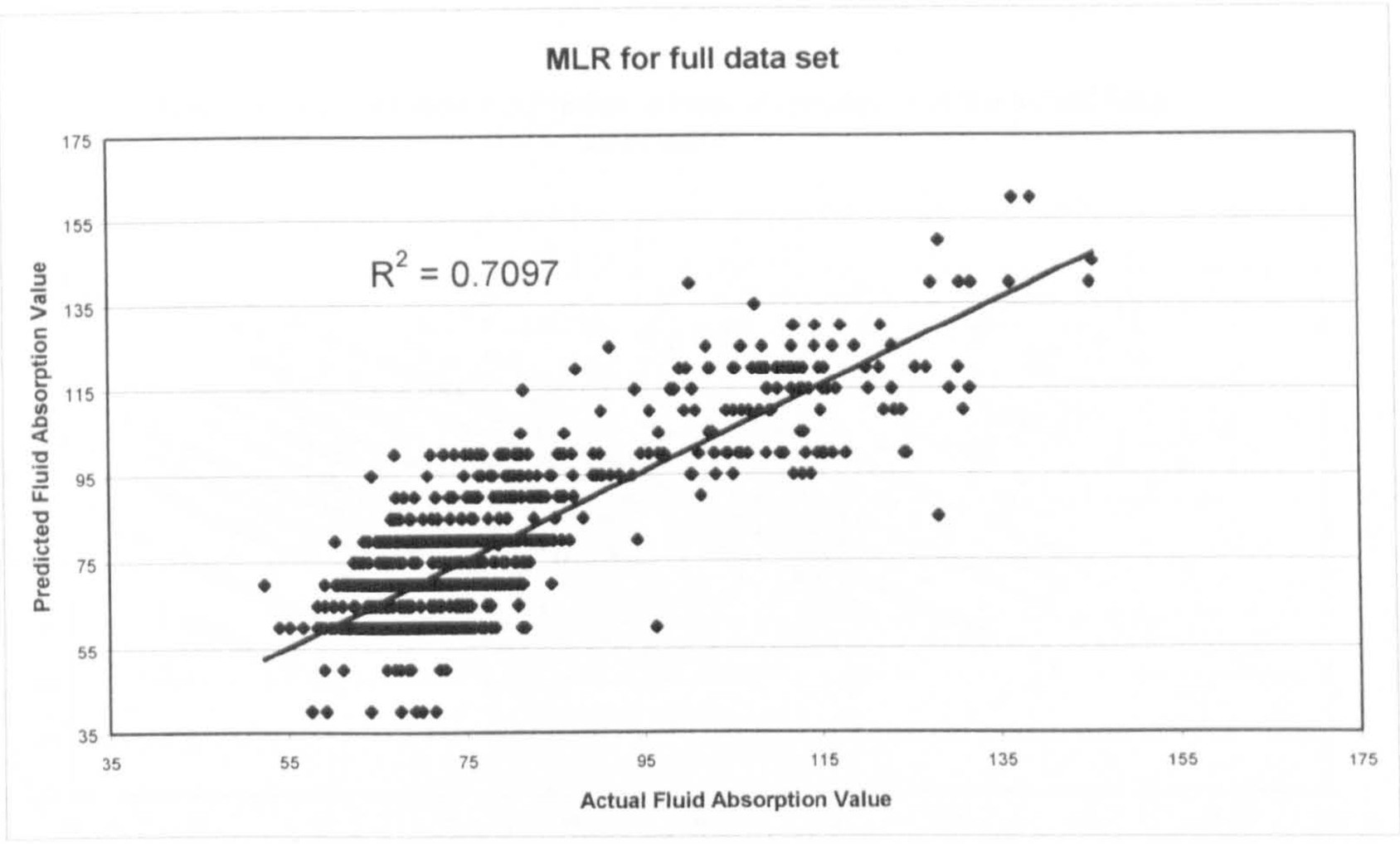


Figure 4.11 Scatter Plot of Predicted Fluid Absorption Values against Actual Fluid Absorption Values Using Standard MLR and the Full Data Set, R^2 of 0.7097

The result of the MLR calculation (Figure 4.11) could well be influenced by leverage values, though an examination of the residuals tends to suggest otherwise (Figure 4.12). Although it looks as if there are two separate populations for the error distribution in the residuals, a closer examination using a distribution plot (Figure 4.13) shows that there is in fact only one distribution evident. In this case the R^2 value is of little use, it indicates a fairly good calibration however this is most likely strongly influenced by the leverage effect of the two separate groups of data points. The residuals show normally distributed error within the two separate populations evident in the data.

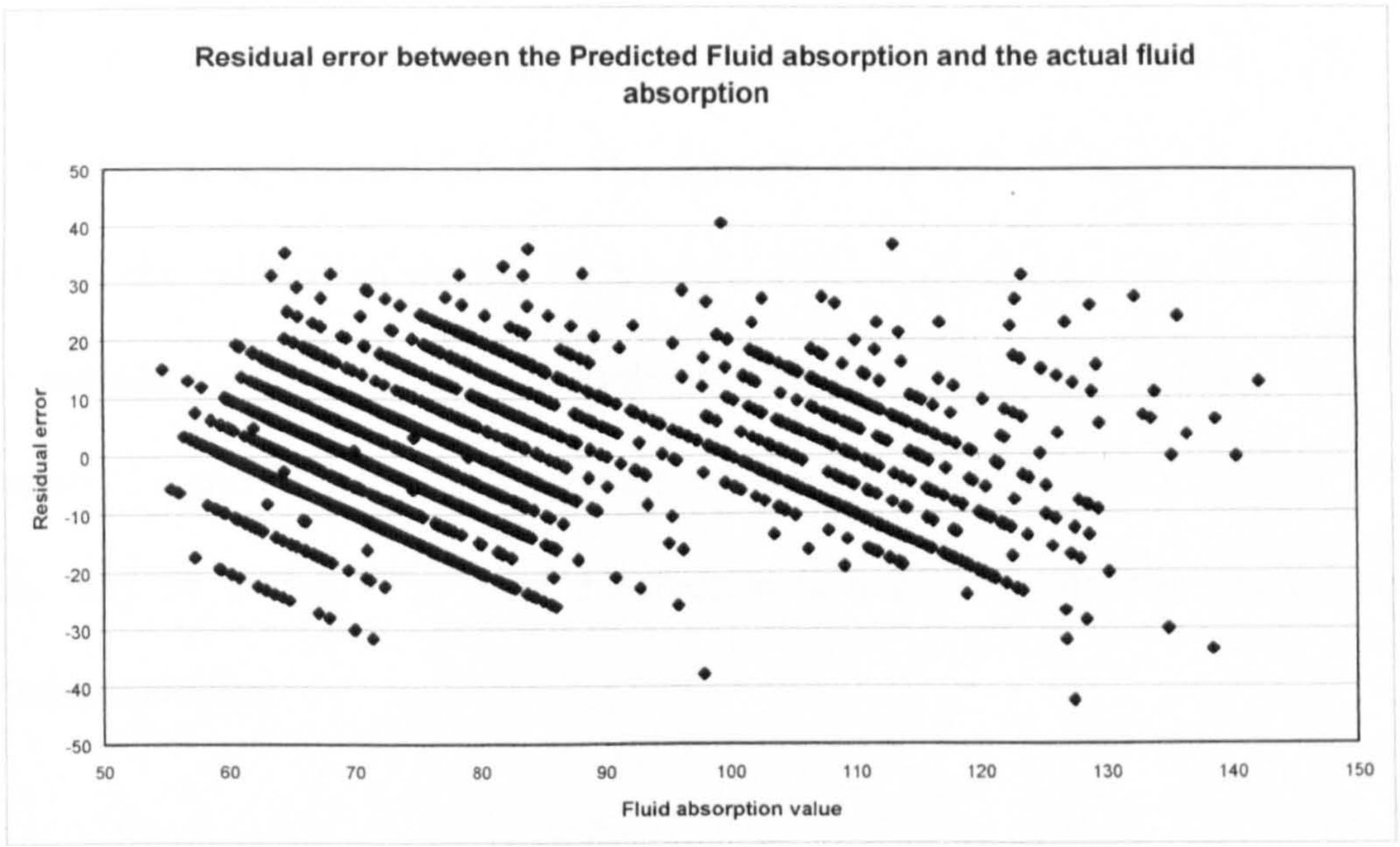


Figure 4.12 Plot of the Residual Error in the MLR predictions from Figure 4.11

The banding seen in both the prediction plot and the residual plot is the result of the measurement of the fluid absorption, which is carried out only to the nearest 5ml.

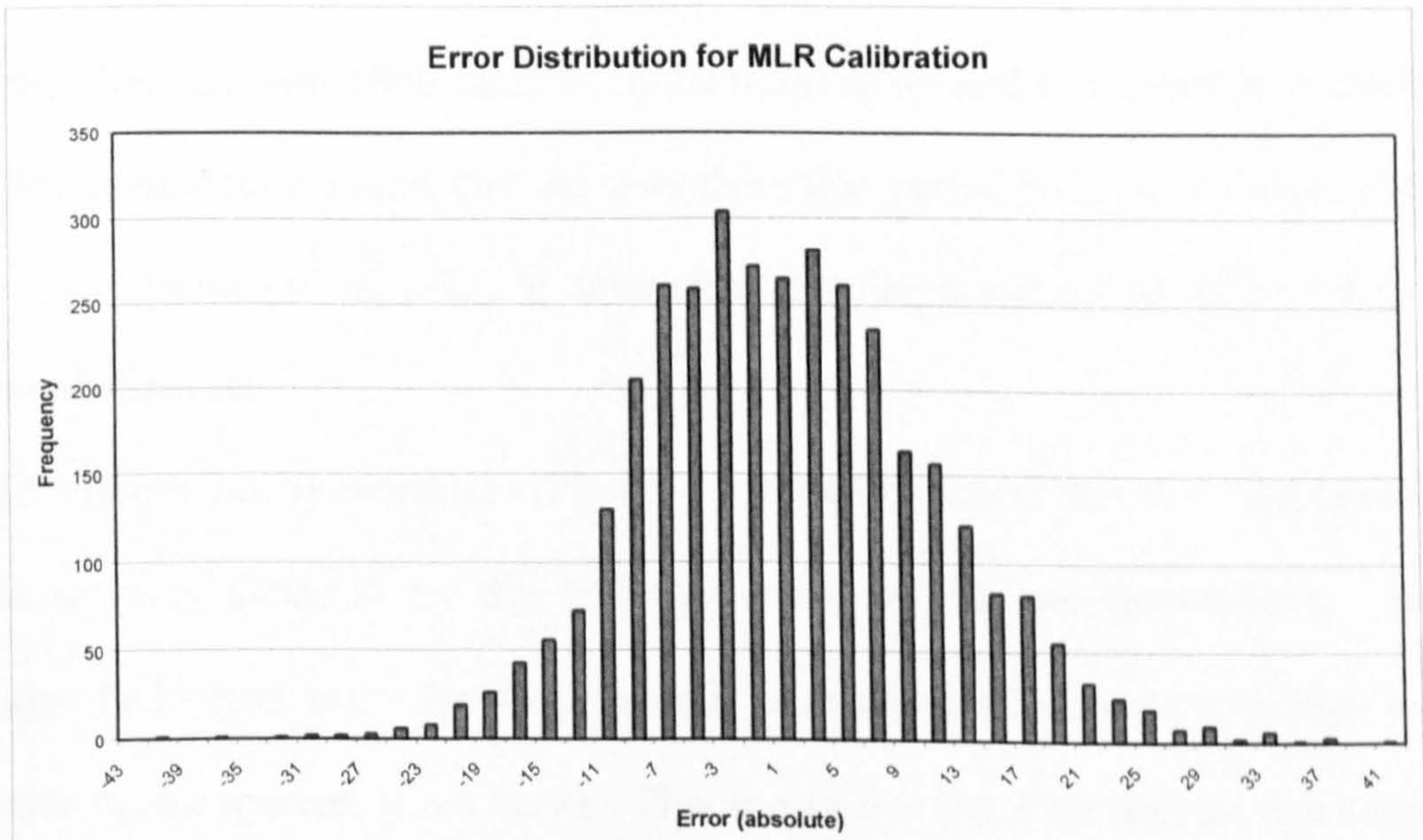


Figure 4.13 Residual Error Distribution Calculated from the Results for the MLR Prediction seen in Figure 4.11

At median distribution the error present in the predictions is forty percent. This error level is expected given the information known about the solubility of the material in aqueous media

and the known limitations of the test carried out. The error appears to be quite low when the two populations are taken into account and this tends to suggest that there is a relationship between the fluid absorption and the other variables that exists between the bulk polymer batches, not just an individual batch. The residual error here is too high for any practical application of the calibration model, for a model to be useful as a replacement to the actual test the residuals would have to be considerably smaller. This model does indicate that an investigation of a smaller portion of the data set where the distribution of values is known to be normal may be useful. The error distribution is indicative of either random results, or fairly robust modelling.

4.4.2 MLR on the normally distributed data

The data set was split into sections, the largest section being the period from the end of 1995 to the end of 1997. The selected data was shuffled by sample and divided into a training set and a validation set, with 1800 samples in the training set and 600 samples in the validation set. Earlier examination found that the data from this period follows a normal distribution. This can be examined using MLR to determine how the model compares to the model built for the whole data set.

This calibration is nearly worthless (Figure 4.14), and indicates that the calibration using the full data set was distorted by the leverage effect of the two populations. If the two populations had come from the same overall population a calibration similar to the first would have been expected, if not better. This shows that the error distribution seen before is the effect of random results as opposed to the error from a useful model.

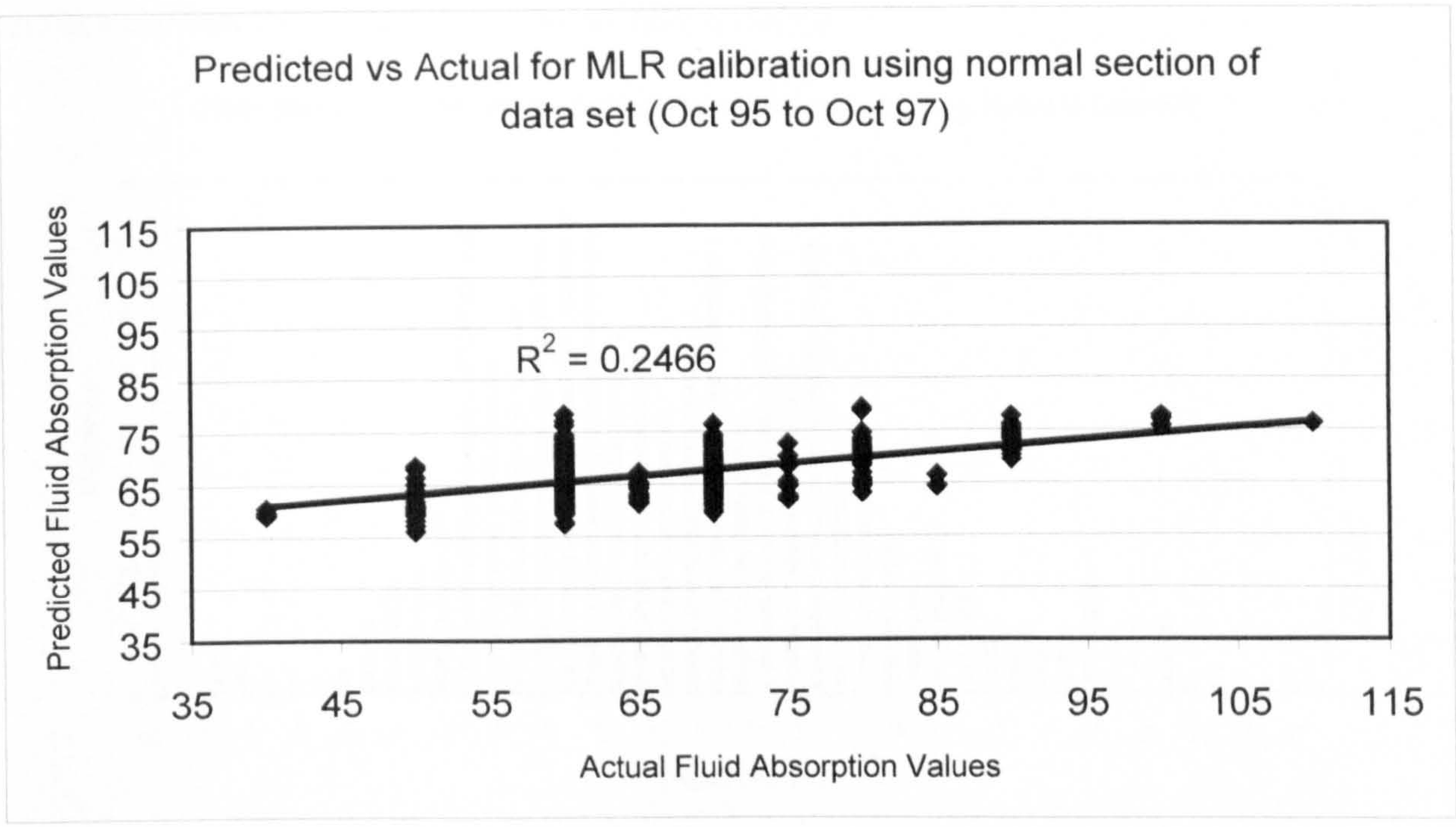


Figure 4.14 Scatter Plot of Predicted Fluid Absorption Values against Actual Fluid Absorption Values Using Standard MLR and the October 1995 to October 1997 Data set, R^2 of 0.2466

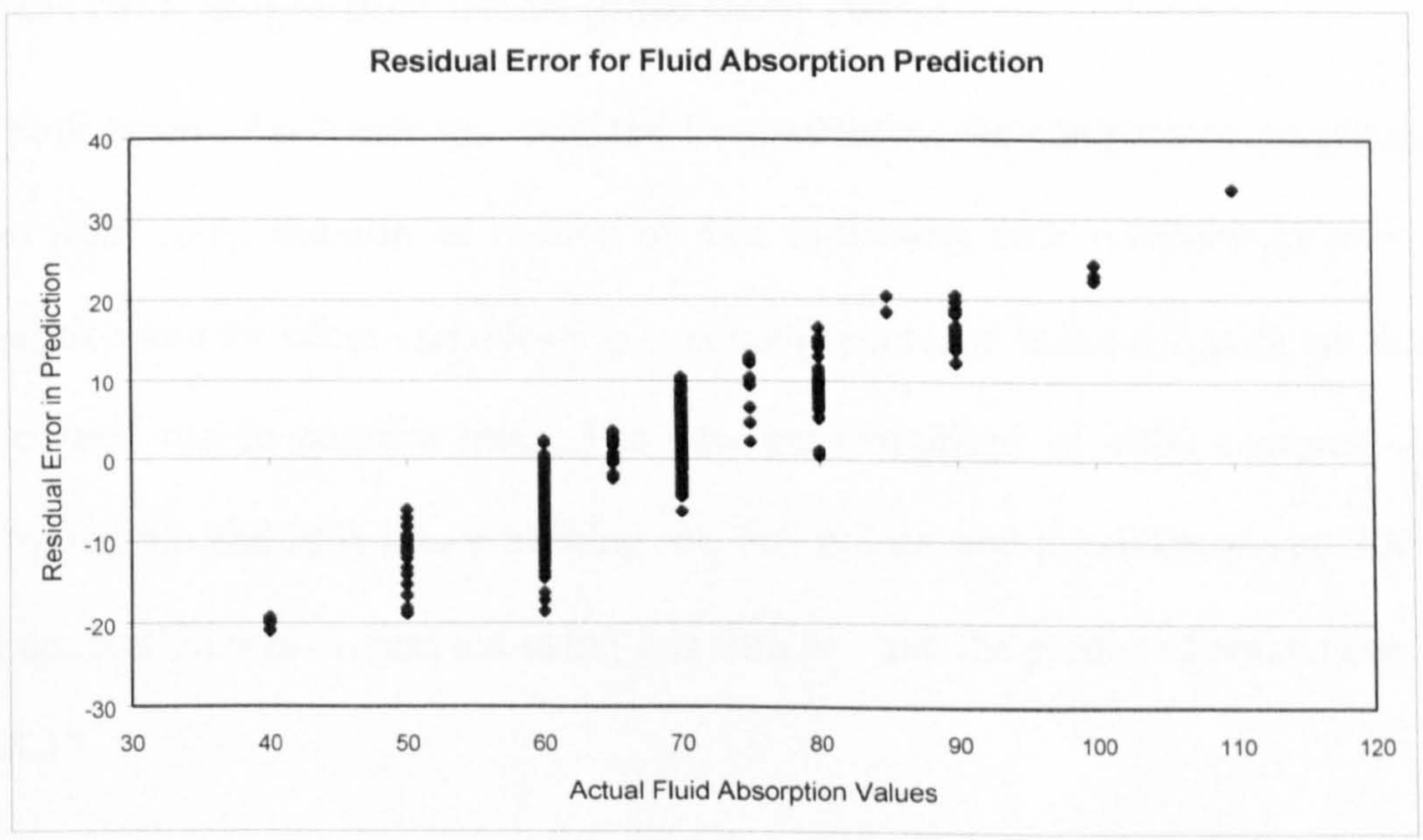


Figure 4.15 Residual Error for the Second MLR model, October 1995 to October 1997

The residual error shown here (Figure 4.15) is typical for a model where there is no relationship between the dependent and independent data matrices, the error is proportional to

the magnitude of the predicted value. The error distribution (Figure 4.16) is random, as expected when the predicted results are also random.

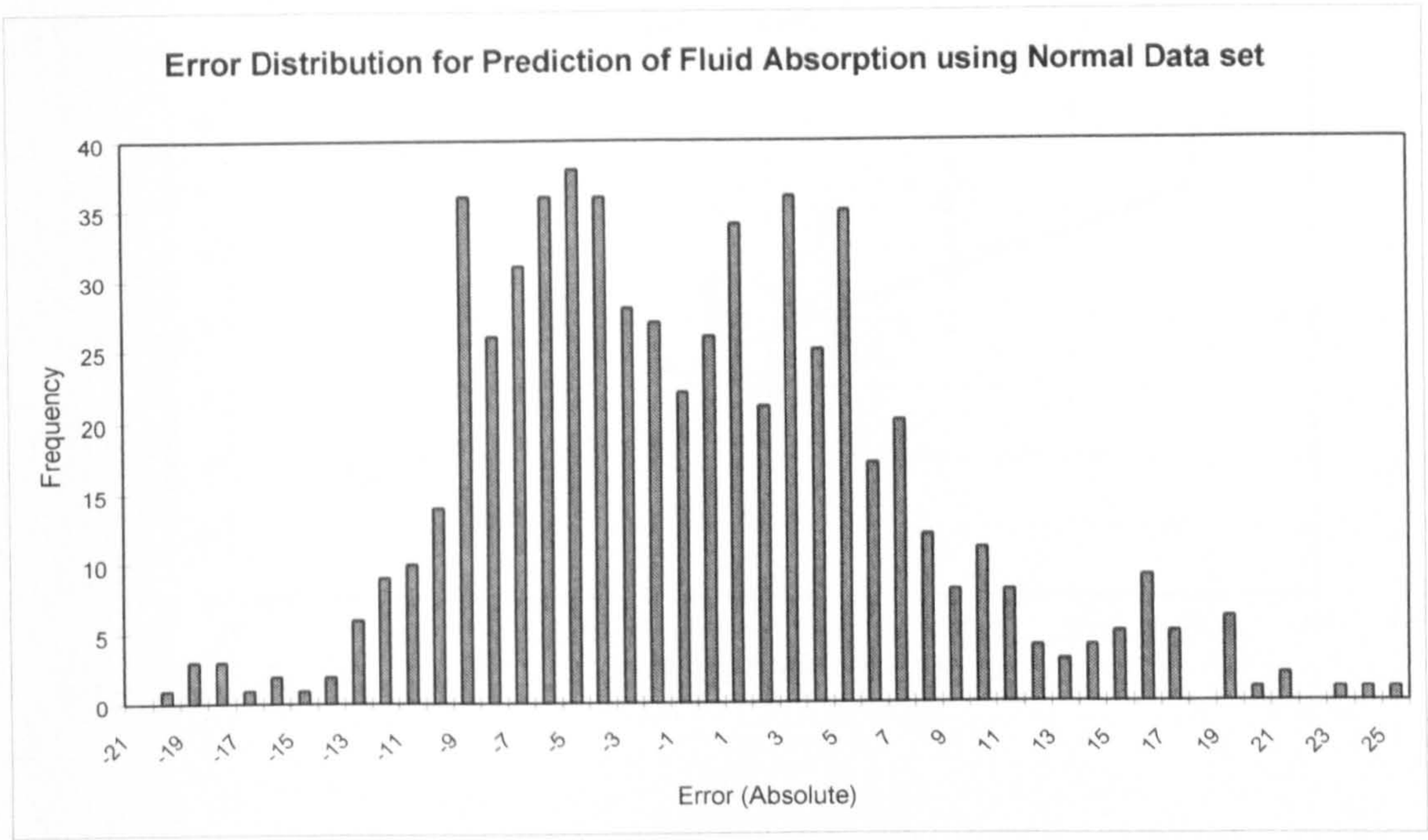


Figure 4.16 Error Distribution for Second MLR Calibration, October 1995 to October 1997

4.4.3 MLR on a single bulk batch (data from 1996)

A single bulk batch of polymer was selected for calibration for comparison purposes. With the results from using the normal section of data indicating little relationship between the fluid absorption and the other variables this was not expected to make a significant difference and was carried out to confirm this. The data set comprised of 1000 samples, and was shuffled by sample and split into a training set, 700 points, and a validation set, 300 points. An MLR calibration was carried out using this data set, and the predicted results can be seen in figure 4.17.

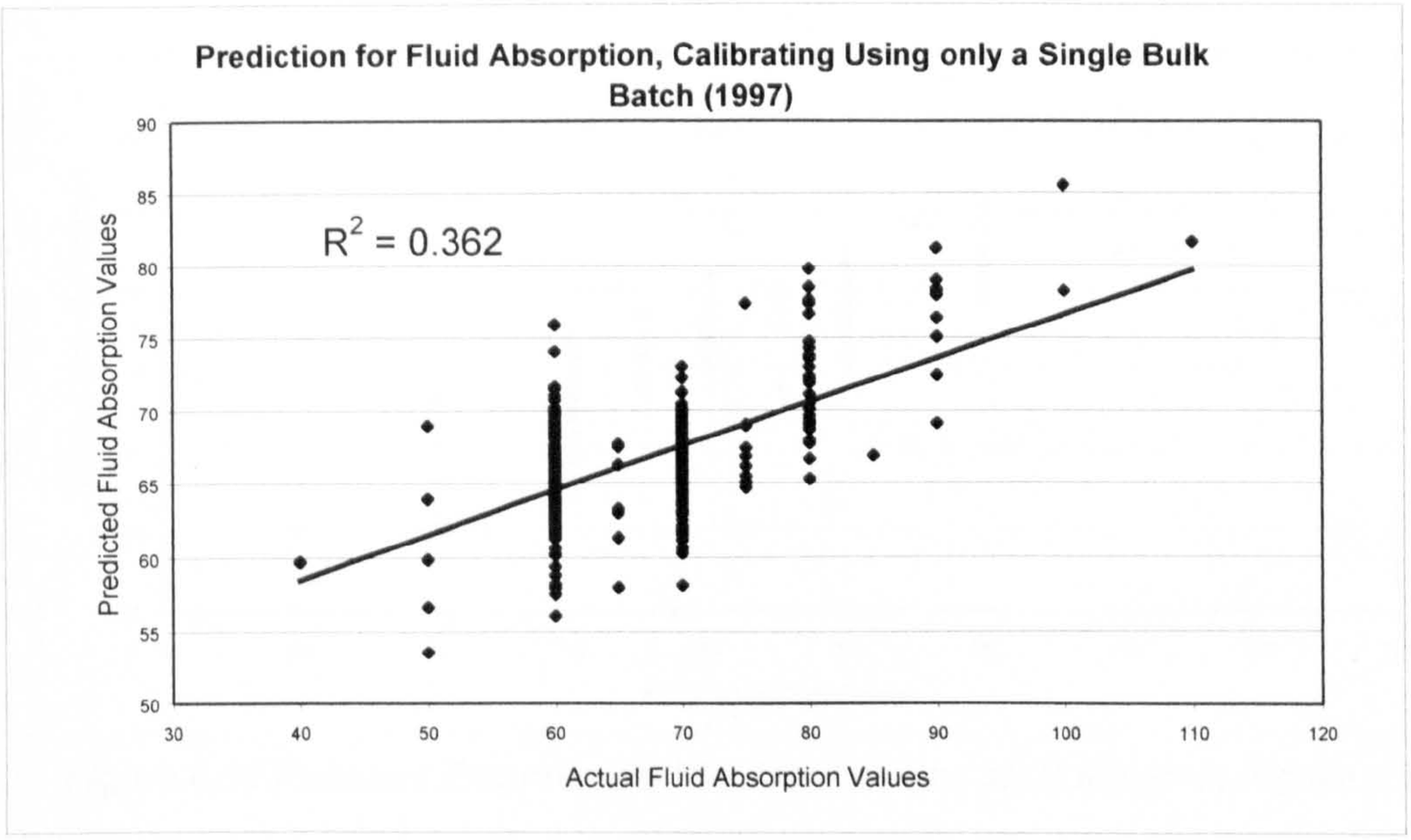


Figure 4.17 Predicted Fluid Absorption Values vs. Actual Fluid Absorption Values for the Single Bulk Polymer Batch from 1997, R^2 of 0.362

This model shows a minor improvement, however not enough for use. This suggests that there are differences between bulk polymer batches that affect the results of the analytical tests carried out. It is unlikely that further work with this data will lead to an improved model, and for the predictive error to drop further (Figure 4.18). The error plotted against the fluid absorption shown is indicative of no relationship existing, and the error is randomly distributed (Figure 4.19).

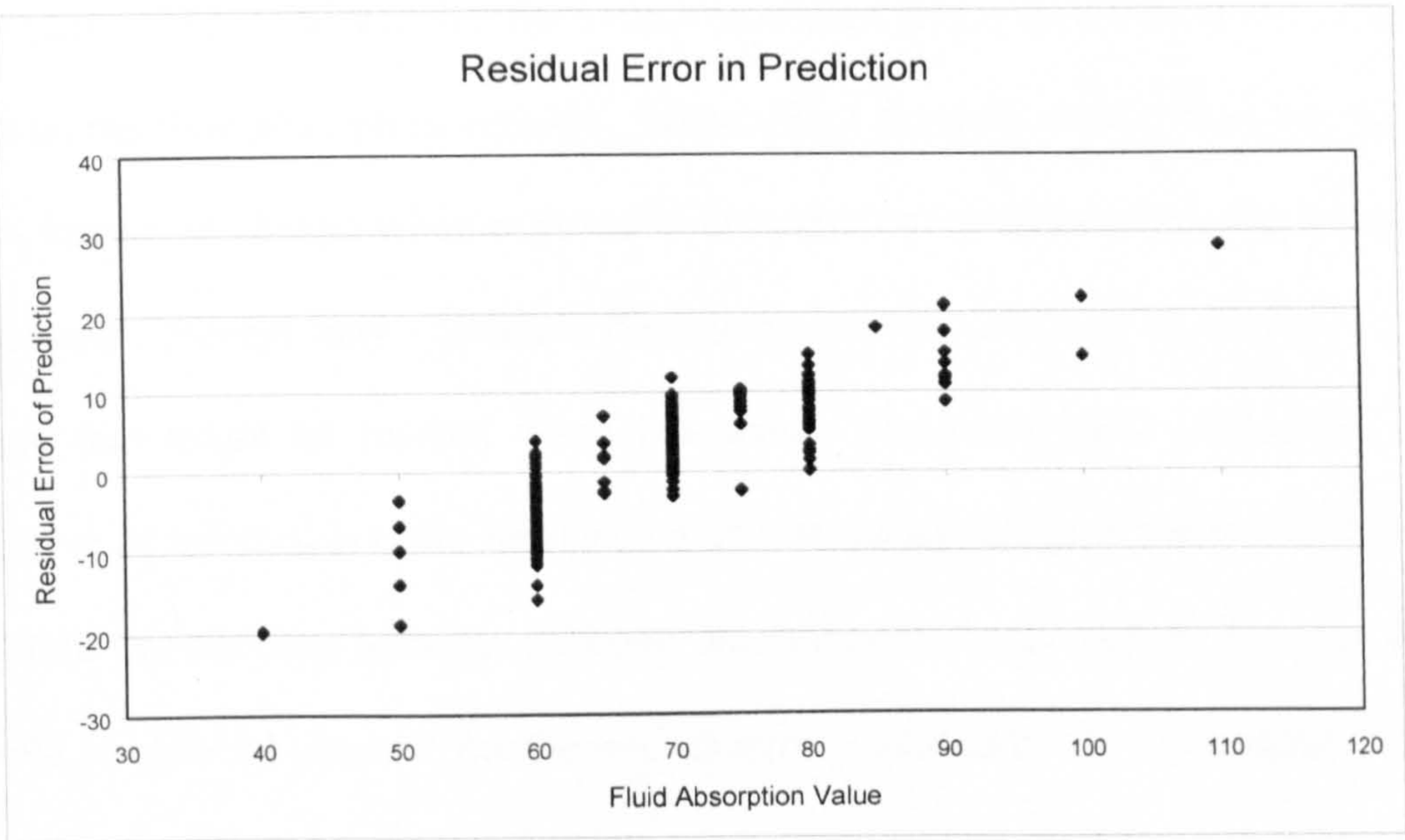


Figure 4.18 Residual Error for the Fluid Absorption MLR Model in Figure 4.17

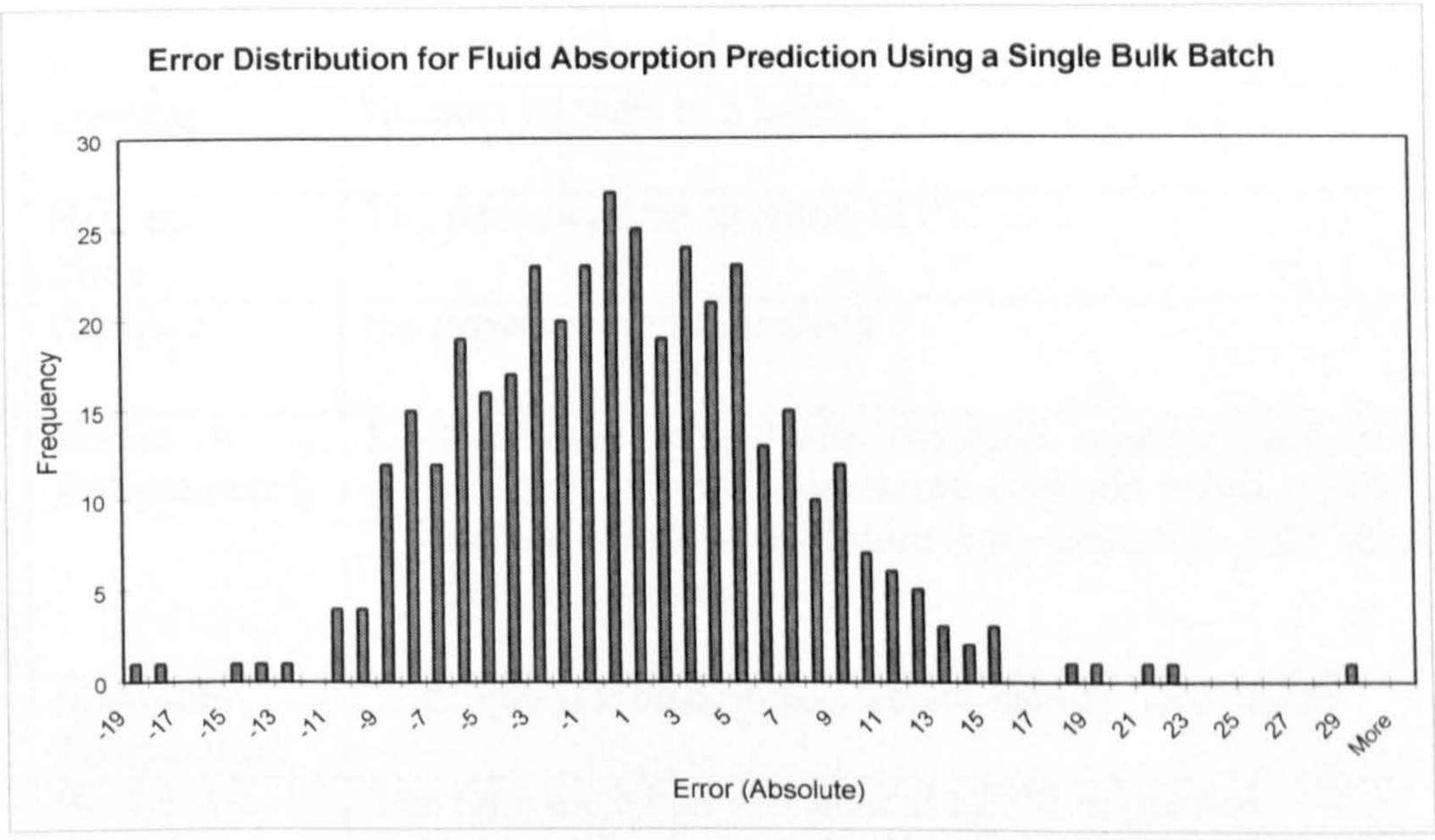


Figure 4.19 Error Distribution for the MLR Model shown in Figure 4.17

4.5 Intrasite Experiment 5, Inclusion of the Sterilisation Data

The current results give indication that it is unlikely that a model using the fluid absorption variable would be possible with less than 30% error. However the best model developed using a single bulk batch is far short of this. This means there must be other influences not described by the information available, and new information must be gained. At the time that this work was carried out the only sterilisation data available was for the year 1995 and

previous years. The data available for 1995 was copied into a spreadsheet and examined for its effect on the fluid absorption variable. Some effect from the sterilisation was expected as the gel is known to change when exposed to temperature, or when aged, the temperature of the sterilisation process may accelerate the ageing process and lead to changes in the fluid absorption that might be tracked using this data. The most likely reason for the fluid absorption to be affected is if the sterilisation affects the solubility of the finished product. If the variation in solubility between different batches could be accounted for then the model error could be reduced closer to the theoretical error produced by the measurement technique. The data that was copied into the spreadsheet was made up of the following variables (Table 4.5)

1	Quantity	Number of units in a batch
2	Heat up Time	The time required to reach 121°C
3	Pressure	the Pressure in the steriliser
4	Minimum Temperature	1. The lowest temperature the batch reached once Fo had been reached. If the temperature dropped below 121°C for any reason before sterilisation was complete then sterilisation had to be repeated
5	Maximum Temperature	The highest temperature reached during sterilisation
6	Hold Time	The time the batch was held at 121°C or greater
7	Cool Time	Once sterilisation has occurred the batch is allowed to cool slowly
8	Fo	The integral of the temperature above 121°C

Table 4.5 Sterilisation Variables Details

Graphs of these variables can be seen in Appendix III.

An MLR calibration for the data from the year 1995 was made, to compare with the results when the sterilisation data was added into the data set.

The correlation coefficients for the new variables were determined (Table 4.6),

	QTY	Heat Time	up Pressure	Minimum Temperature	Maximum Temperature	Hold Time	Cool Time	F(0)	Fluid Absorption
QTY	1.00								
Heat up Time	0.19	1.00							
Pressure	-0.06	-0.08	1.00						
Minimum Temperature	0.01	-0.03	-0.01	1.00					
Maximum Temperature	0.06	-0.34	0.04	0.00	1.00				
Hold Time	-0.07	0.30	0.01	-0.04	-0.10	1.00			
Cool Time	0.08	0.47	-0.03	-0.06	-0.41	0.11	1.00		
F(0)	-0.01	0.38	0.00	-0.01	0.09	0.70	0.25	1.00	
Fluid Absorption	0.02	-0.13	-0.02	0.03	0.10	0.00	-0.39	-0.07	1.00

Table 4.6 Correlation Coefficients for the Sterilisation Data

	pH	Elasticity	Viscosity Coefficient	SC1	Fluid Absorption
PH	1.00				
Elasticity	0.83	1.00			
Viscosity Coefficient	0.53	0.80	1.00		
SC1	0.09	0.30	0.57	1.00	
Fluid Absorption	-0.45	-0.36	-0.11	0.10	1.00

Table 4.7 Correlation Coefficients for the Analysis of Intrasite Gel During the Period from which the Sterilisation Data was Taken

These correlations are low, particularly with reference to the viscosity coefficient and SC1, and the correlations can also be seen to be low for the analysis data for the same period (Table 4.7). A calibration will be carried out for comparison purposes, first to examine just the analysis variables, then to examine the effect of adding in the sterilisation data.

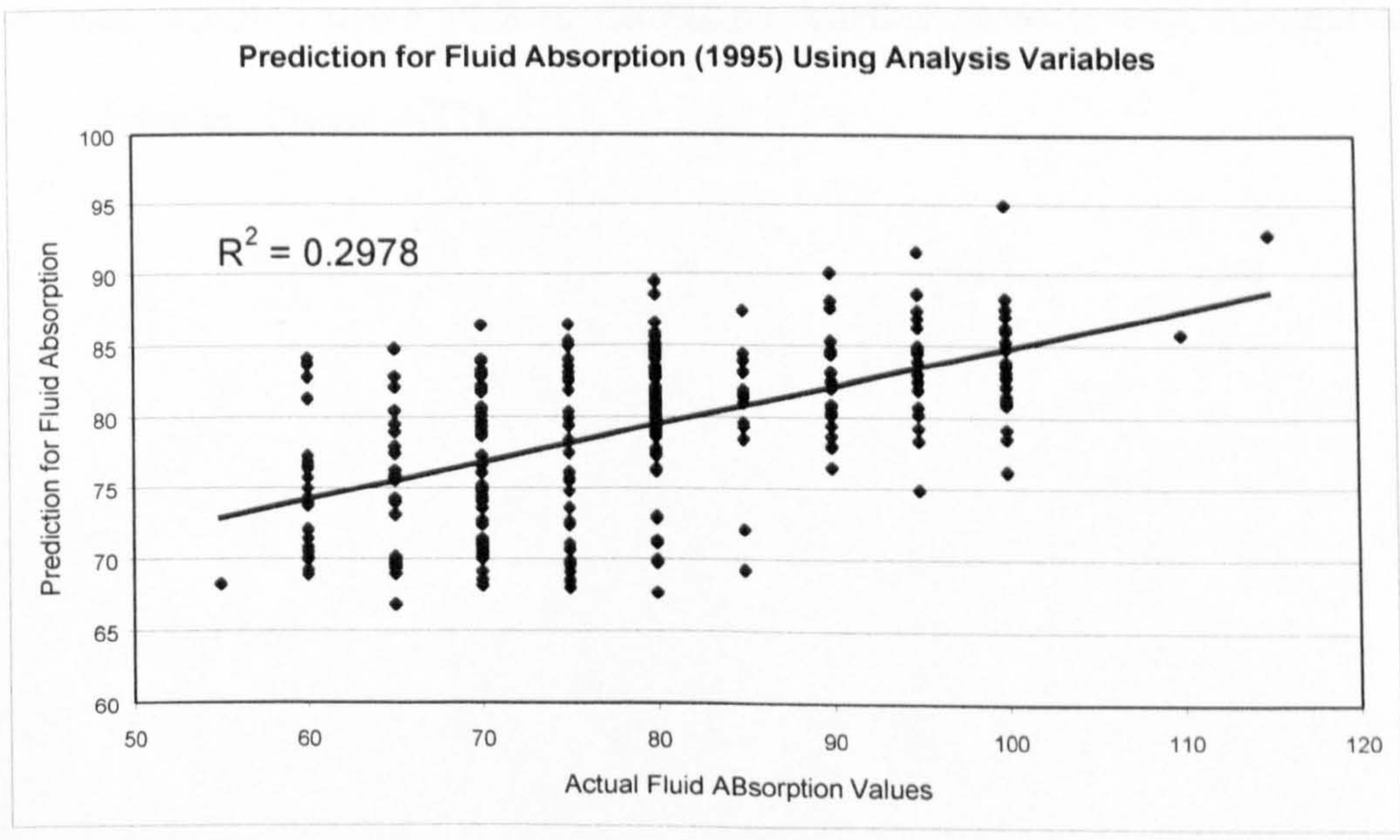


Figure 4.20 Predicted Fluid Absorption values vs. Actual Fluid Absorption values for the MLR Model Using data from 1995 using only the analysis variables, R^2 0.2978

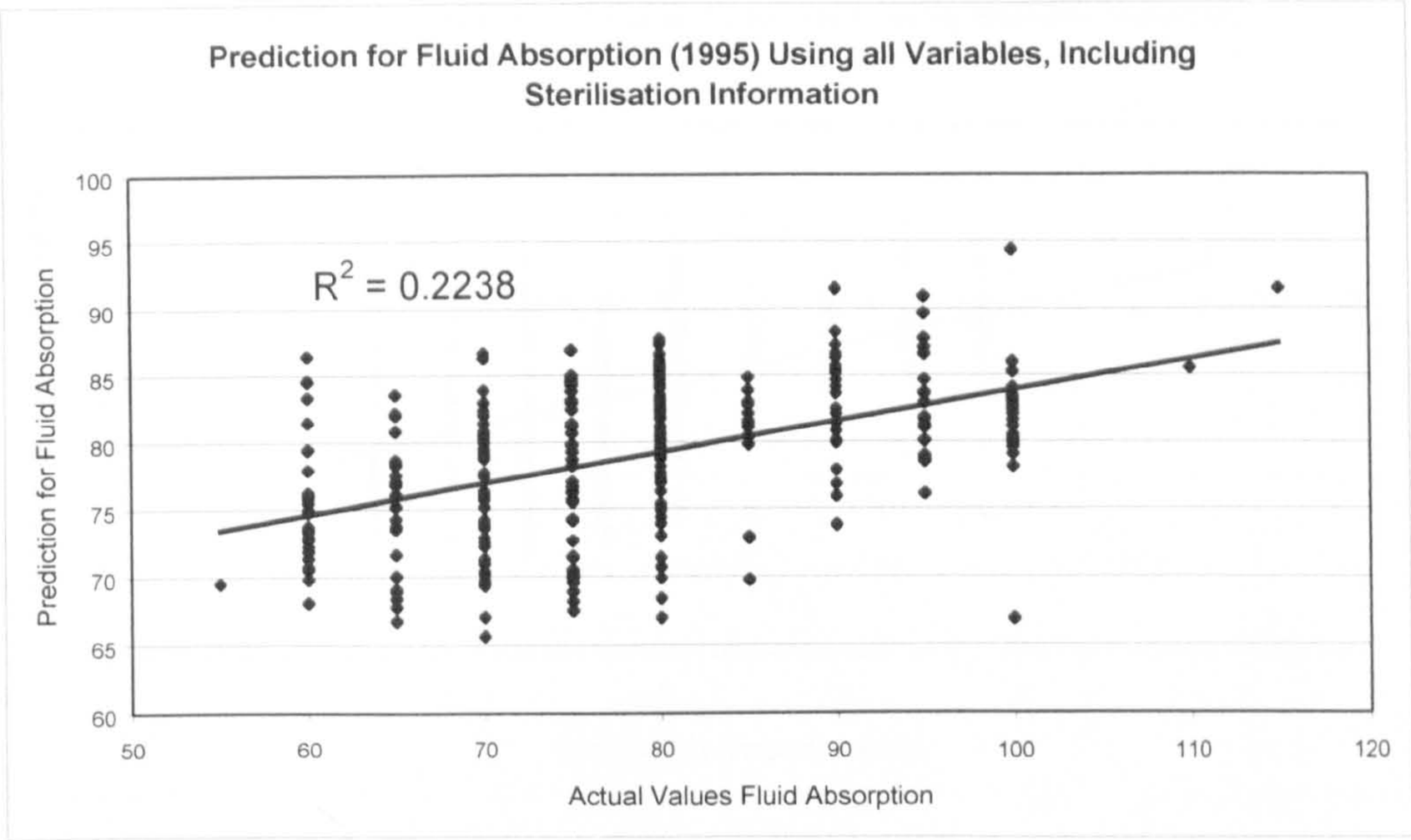


Figure 4.21 Predicted Fluid Absorption Values vs. Actual Fluid Absorption Values for the MLR Model Using Data From 1995, Including the Sterilisation Data. R^2 0.2238

These two calibrations (Figure 4.20 & Figure 4.21) both show a very poor calibration. The comparison between the calibration using just the analysis variables and the calibration using all the available information shows that the addition of the new variables has contributed only noise. If the sterilisation variables contain any information not supplied by the analysis variables this is hidden by the extra noise the variables introduce into the model. This calibration was repeated using PLS to determine whether there is any information in the sterilisation variables (Figure 4.22).

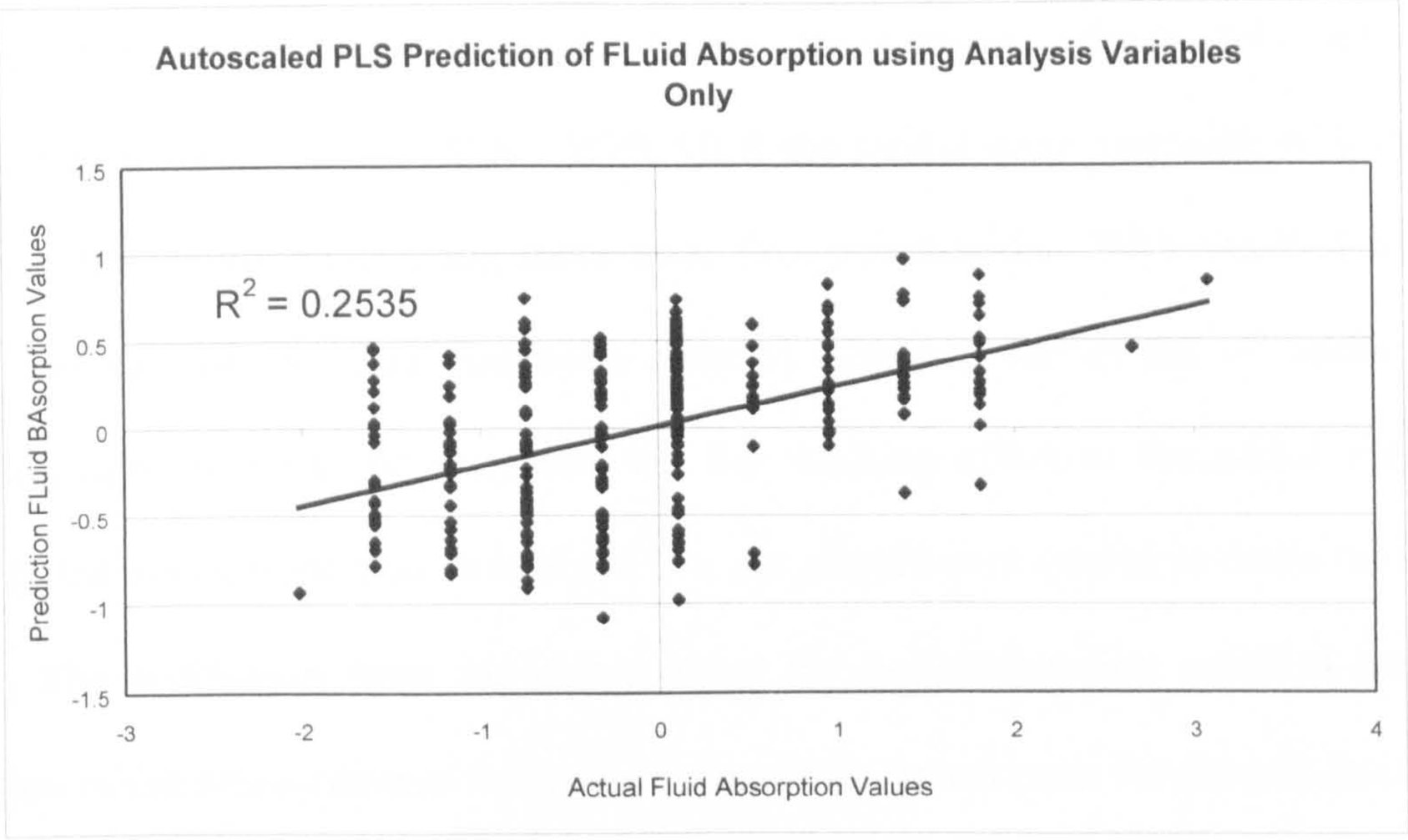


Figure 4.22 Predicted Fluid Absorption Values vs. Actual Fluid Absorption Values for the PLS Model of Data from 1995, Analysis Variable Only, R^2 0.2535, 2 lv's

There is a minor improvement to the model using two latent vectors for the PLS model over the MLR model, this does not make the model useful however. This is then compared with the change in the model when the sterilisation variables are introduced (Figure 4.23). Although the correlation results show that there is little information about the fluid absorption available from the sterilisation data PLS will allow that information to be separated for the noise of the model.

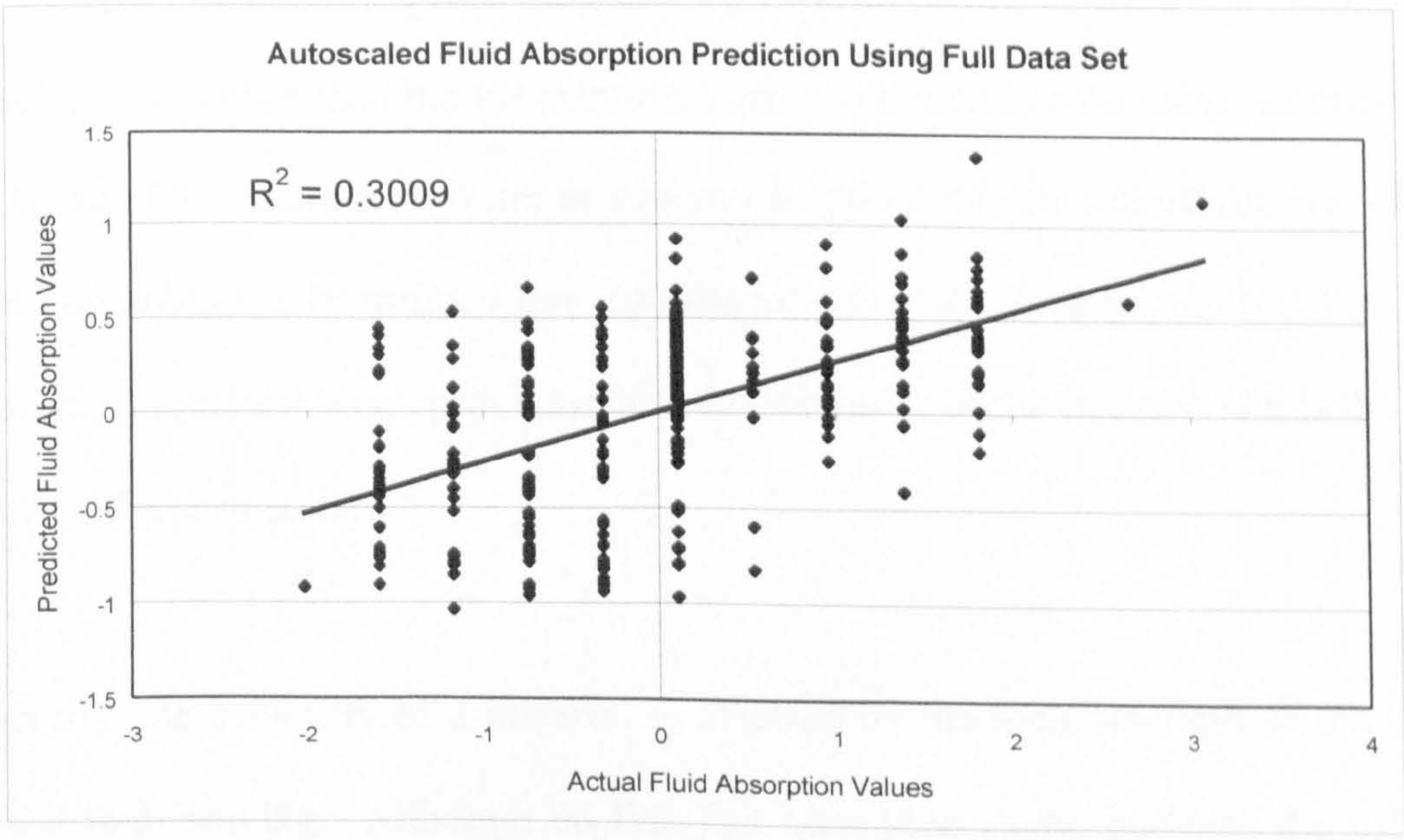


Figure 4.23 Predicted Fluid Absorption Values vs. Actual Fluid Absorption Values for the PLS Model of Data from 1995, all Variables, R^2 0.3009, 5 lv's

The significance of these results is that there is an improvement in the model compared with the attempt at modelling using MLR. With MLR the model error increased with the added variables as they were contributing more error than information. With the PLS model the error in the sterilisation data has been reduced allowing the effect of adding in the sterilisation information to be seen with out the masking effect of the added noise. The increase in the model performance is about 5%, not of sufficient quality to make the model of any use. The conclusion from this is that either the sterilisation data contains little useful information of the measurements themselves contain too much error for the information to be used.

4.6 Intrasite Experiment 6, Effect of pH on Measured Fluid Absorption

So far the models for fluid absorption in Intrasite Gel have fallen short of the level that should be theoretically possible given the expected level of error present in the measurement of fluid absorption. Other than the measurement error, the next known cause of error is due to the solubility of the material in water or aqueous solutions. If the reason for the solubility of Intrasite Gel could be determined this information could be used to improve the current model up to the theoretical level, providing that the solubility of the Intrasite Gel is the reason for the poor model seen so far.

It is known that the solubility of a material is affected by the ionic strength of the solution into which it is dissolving. Although no link has been seen so far between the pH of the Intrasite Gel and the fluid absorption value measured this may be because that link is being

masked by the experimental error. By examining the fluid absorption results obtained over a wider range of pH values than normally seen any effect of the pH, and thus the ionic strength can be assessed. If any link between the pH and the solubility is found then the data available can be re-assessed to determine if that information is already available in the data set.

Thirty experiments were carried out using five different Intrasite batches over thirteen different pH ranges, from 5.9 through to 9.1. pH values outside these ranges are known to break down the polymer chains. Initially there was intended to be only twelve sets of values, however by accident two sets of replicates were set up at pH 8.2 and these values were retained and another set produced for the missing value (9.1). The experiments were carried out using five batches due to the limited availability of large quantities of Intrasite Gel of a single batch, the information about the batches chosen for this experiment is in table 4.8. The results from the settling volume tests can be seen in table 4.9, this data is plotted as a graph in figure 4.24. The data was examined using ANOVA to determine whether the effect of the change in pH was greater than the effect of the change in batch, and the results of the ANOVA calculation can be seen in table 4.10. The batches in table 4.8 were selected on the basis of their original fluid absorption values and pH values being as close together as possible.

Batch Data	Batch Number	pH	Elasticity	Viscosity Coefficient	SC1	Fluid Absorption
2/4/97	970463	7.4	1469	210	2.2	70
2/4/97	970465	7.4	1414	210	2.3	70
2/4/97	970464	7.4	1419	210	2.3	70
2/10/97	970543	7.4	1459	220	2.2	70
2/10/97	970544	7.4	1445	230	2.2	70
2/10/97	970546	7.4	1494	220	2.3	70
2/10/97	970545	7.4	1601	240	2.3	70
2/12/97	970653	7.4	1454	210	2.3	70
2/12/97	970642	7.4	1553	230	2.3	70
2/12/97	970623	7.4	1633	240	2.3	70
2/14/97	970651	7.4	1469	210	2.3	70
2/20/97	970713	7.4	1541	220	2.2	70
2/20/97	970716	7.4	1492	220	2.3	70

Table 4.8 Batch Information for the Samples Used in the Experiment to Determine the effects of pH on Fluid Absorption

pH	Batch Number	Replicate 1	Replicate 2	Replicate 3	Replicate 4	Replicate 5
5.9	970463	80	90	85	75	75
6.4	970465	85	85	75	75	90
6.5	970464	65	70	70	75	85
6.7	970543	75	80	75	95	80
7.1	970544	90	85	90	90	90
7.2	970546	85	70	75	70	75
7.4	970545	80	75	70	80	75
7.7	970653	70	65	70	90	75
8.2	970642	80	85	85	100	90
8.2	970623	85	75	80	75	80
8.4	970651	70	75	75	65	70
8.8	970713	75	75	75	80	75
9.1	970716	85	80	80	80	80

Table 4.9 Fluid Absorption Values Using the Settling Volume Test on the Samples from Table 4.8, Using pH Values from 5.9 through 9.1

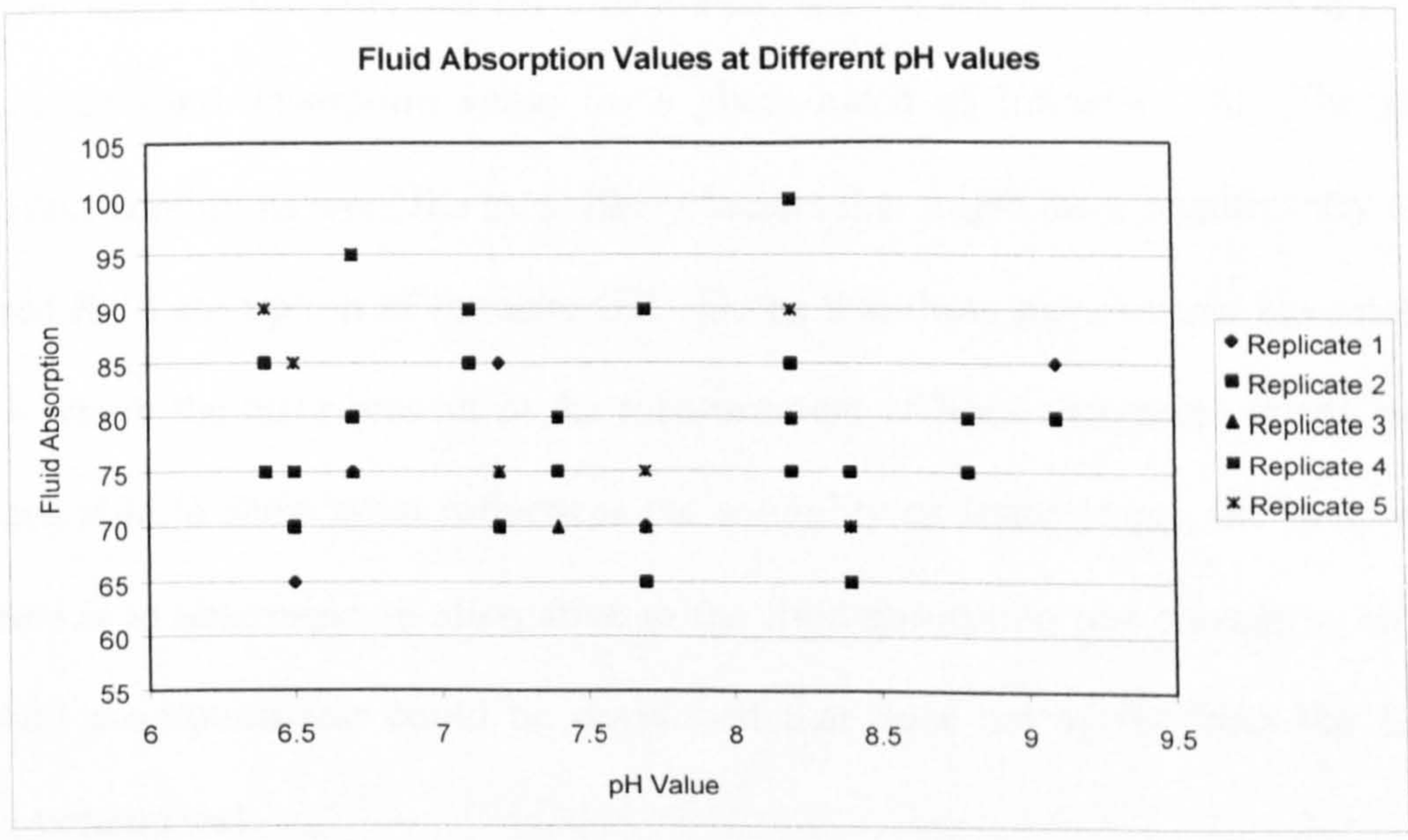


Figure 4.24 Graph Showing the Results of the Settling Volume Fluid Absorption Test at Different pH Values

The null hypothesis was that there is a significant between group difference at the 95% confidence limit, an ANOVA was carried out to examine this hypothesis.

ANOVA

Source of Variation	*Sum of Squares*	*Degrees of Freedom*	*MS*	*F*	*P-value*	*F critical*
Between Groups	1804.62	12	150.38	4.18	0.00014	1.94
Within Groups	1870	52	35.96			
Total	3674.61	64				

Table 4.10 ANOVA Results to Show that there is no Significant Effect of pH on the Settling Volume Test Results

The results of the ANOVA (Table 4.10) show that there is no between group difference at the 95% confidence limit, and that the null hypothesis should be rejected for this data.

This experiment has shown that the pH of the Intrasite Gel being examined is not a significant factor at the time that the tests are carried out and that is does not appear to affect the recorded fluid absorption value for a given batch of Intrasite Gel. The pH and the sterilisation conditions were the most likely factors that might have significantly affected the measured fluid absorption of Intrasite Gel. Given that these experiments have not been able to show where the error present in the measurement of fluid absorption arises, neither have they been able to show what influences the solubility of Intrasite gel, the best step forward from here is to determine an alternative to the fluid absorption test currently carried out. A new fluid absorption test could be developed that does not suffer from the flaws of the settling volume test.

4.7 Intrasite Experiment 7, Examining the Process using CUSUM charts

The CUSUM for each analysis variable was calculated, and the results were autoscaled to allow them to be easily plotted on the same graph for comparison purposes (Figure 4.25).

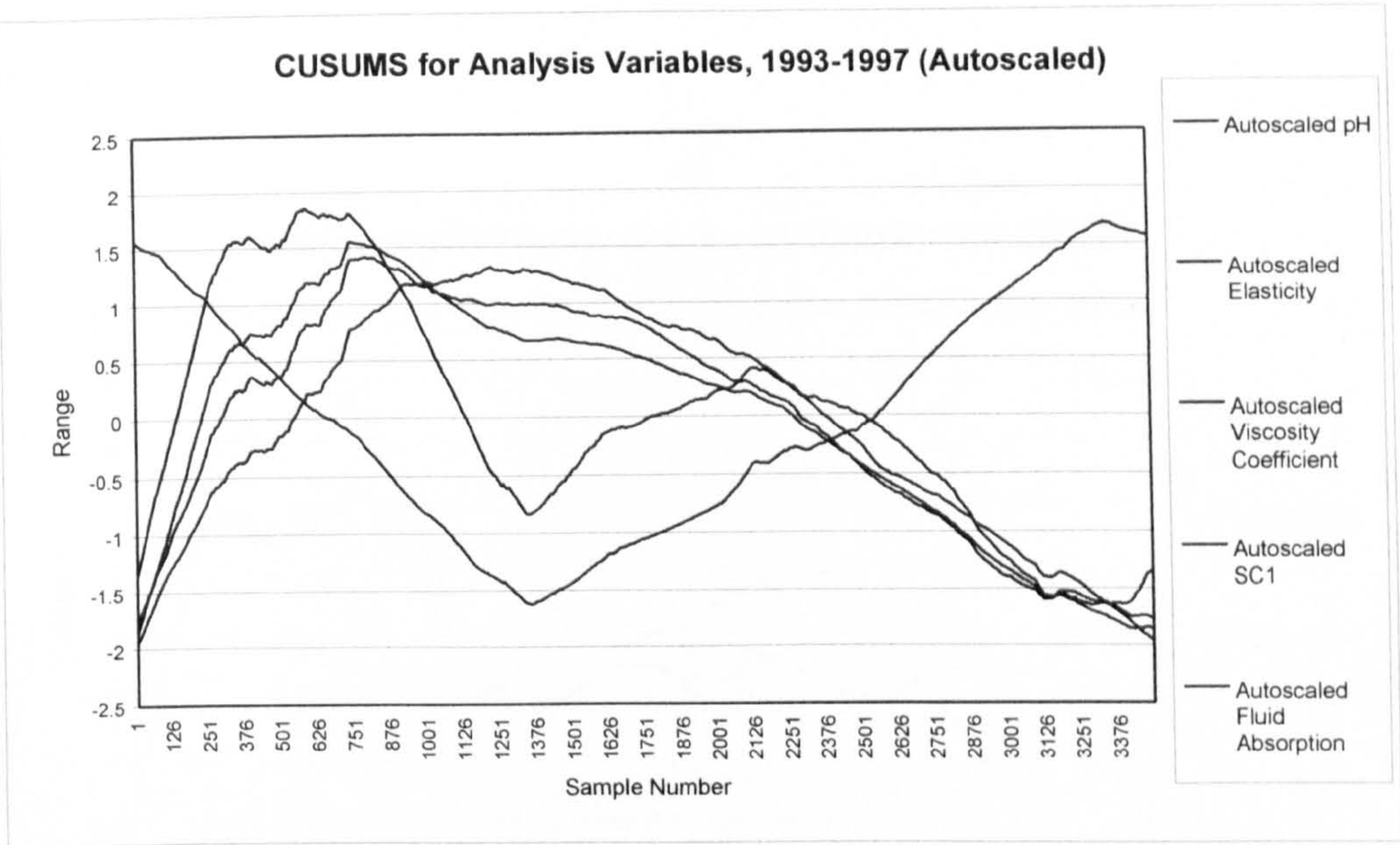


Figure 4.25 Single Point CUSUMS for the Analysis Data Set, 1993-1997

Given the poor correlation of the raw data these graphs show a surprising degree of correlation, this is confirmed by examining the correlation coefficients of the CUSUM data (Table 4.11).

	Autoscaled pH	Autoscaled Elasticity	Autoscaled Viscosity Coefficient	Autoscaled SC1	Autoscaled Fluid Absorption
Autoscaled pH	1				
Autoscaled Elasticity	-0.28	1			
Autoscaled Viscosity Coefficient	-0.75	0.80	1		
Autoscaled SC1	-0.86	0.67	0.97	1	
Autoscaled Fluid Absorption	-0.96	0.45	0.87	0.95	1

Table 4.11 Correlation Coefficients for the CUSUMS from Figure 4.25

These correlations are extremely high, and indicate that SC1, fluid absorption and viscosity coefficient are following the same trend. It can also be seen in figure 4.25 that the pH trend and the elasticity trend show many of the same features, although this is not seen in the raw data, or the correlations for the CUSUMs. The appearance is that the elasticity trend is the

sum of the pH trend and one of the other three parameters. This was examined by plotting the elasticity trend with the trend produced by adding the pH trend with the trend for SC1 (Figure 4.26).

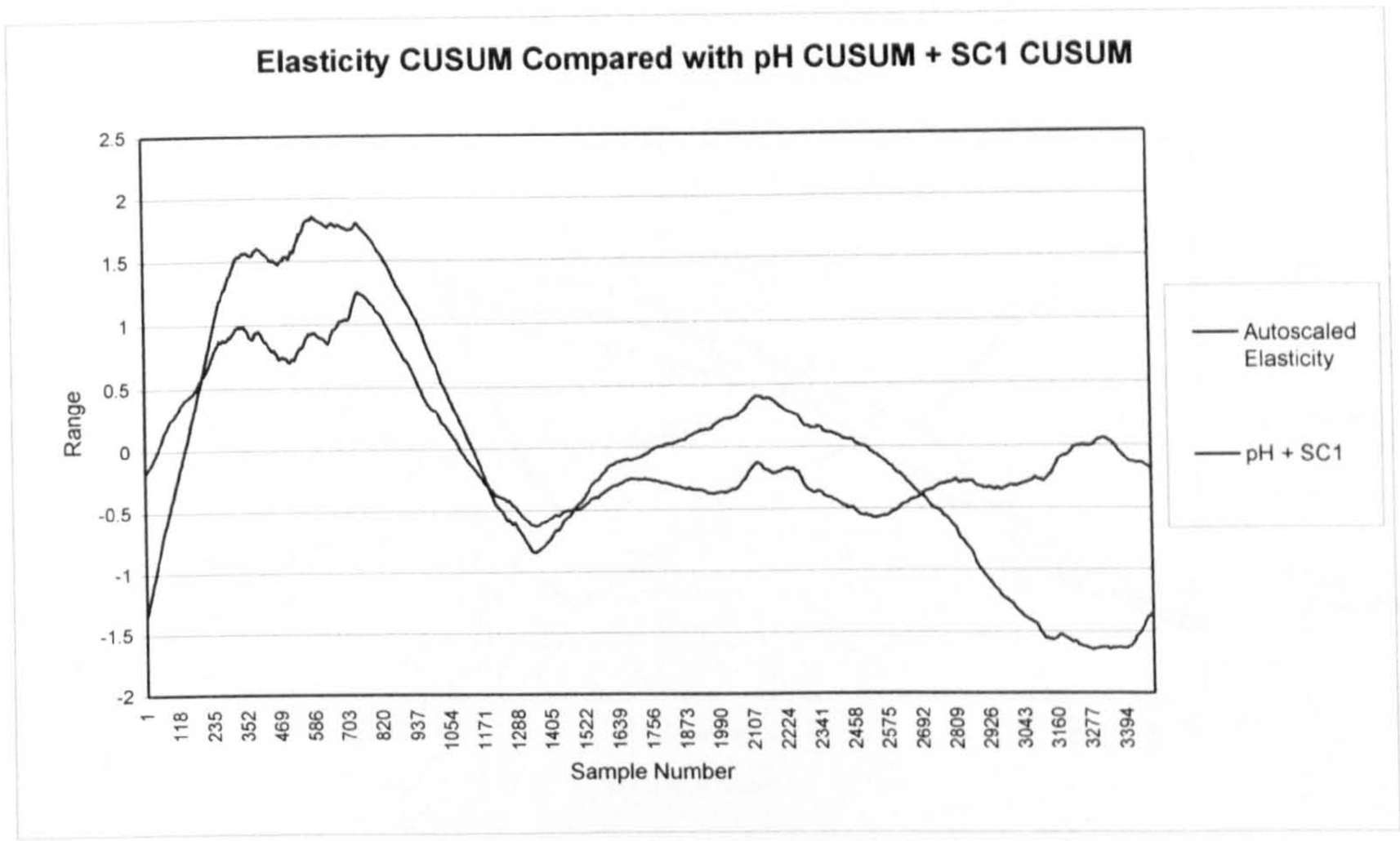


Figure 4.26 Comparison of the CUSUM for Elasticity and the Combined CUSUMs for SC1 and pH

This indicates that while the raw data shows no relationship between pH and any of the other variables, the pH is affecting the elasticity. A possible explanation for this is that the Intrasite gel material is broken down over long term periods at different rates according to the pH. This effect may be accelerated by the sterilisation of the material. If Intrasite experiment 6 had been left to equilibrate for a longer period, or had been heated, an effect from the pH may have been seen. These CUSUM plots also show that there is a high degree of noise in all the measurements made that is masking the relationships between the variables. This raises the possibility that measurements made at the current frequency may be misleading. Rather than measuring the variables at high frequency a better method of following the control of the process may be to reduce the frequency of measurements and examine the CUSUMS. The trends seen in the data should still be visible at much lower sampling rates. This was tested

by examining the CUSUMS that would be produced at lower sampling frequencies. The CUSMUS were generated for sampling at every 2nd, 5th, 10th and 20th point (Figure 4.27, 4.28, 4.29, and Figure 4.30).

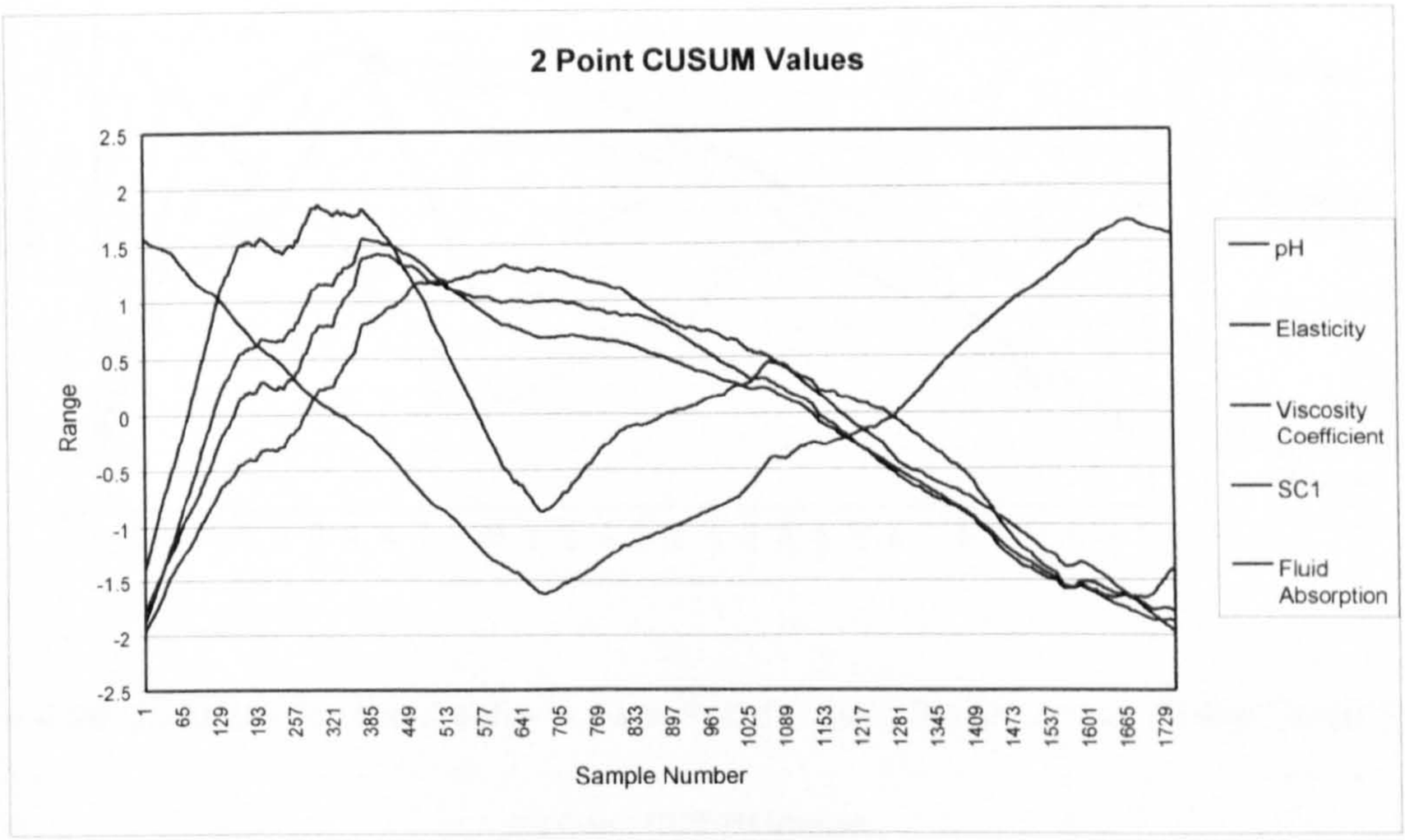


Figure 4.27 CUSUMS for the Analysis Data Set, CUSUM Derived from Every Other Sample

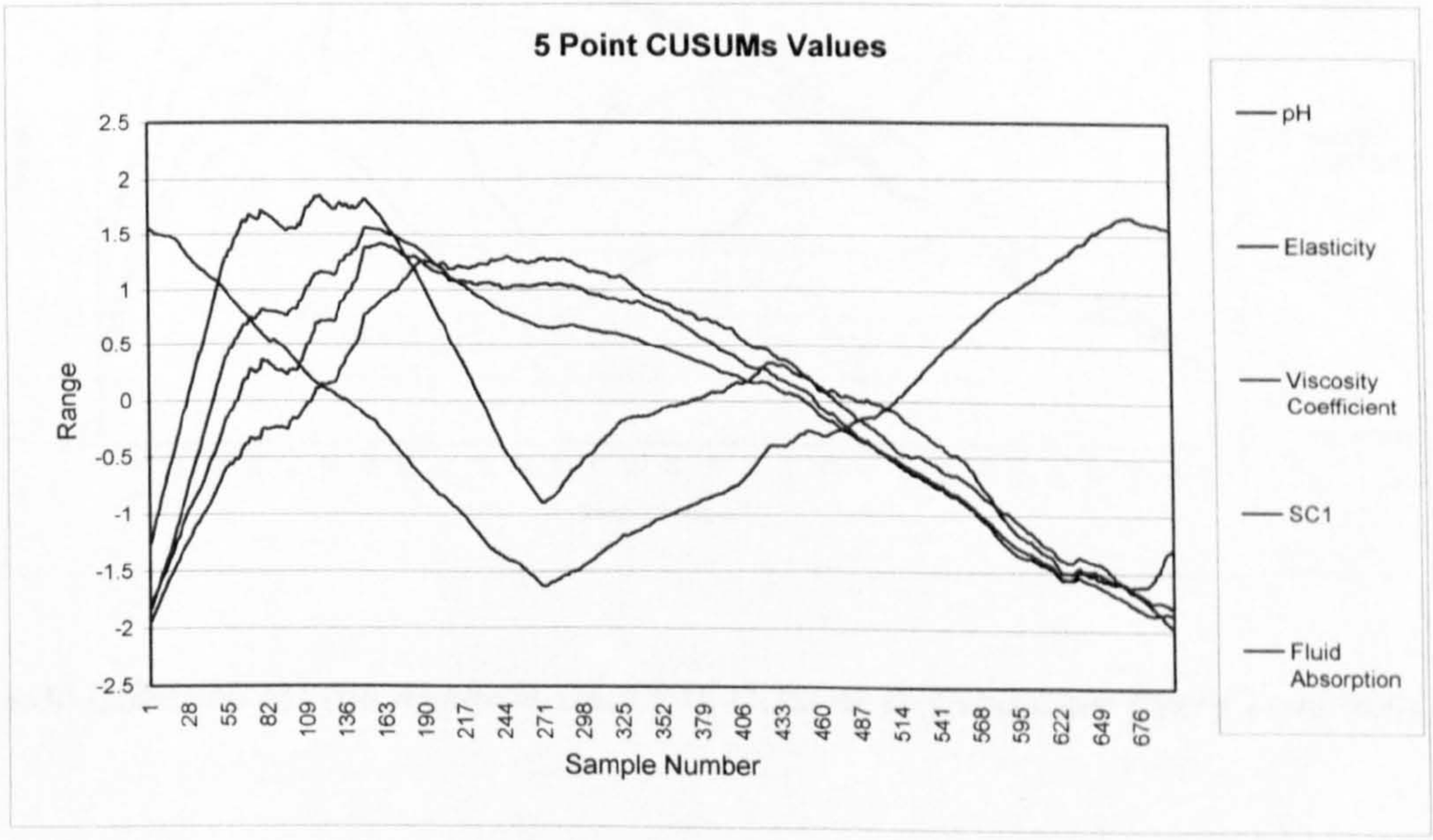


Figure 4.28 CUSUMS for the Analysis Data Set, CUSUM Derived from Every Fifth Sample

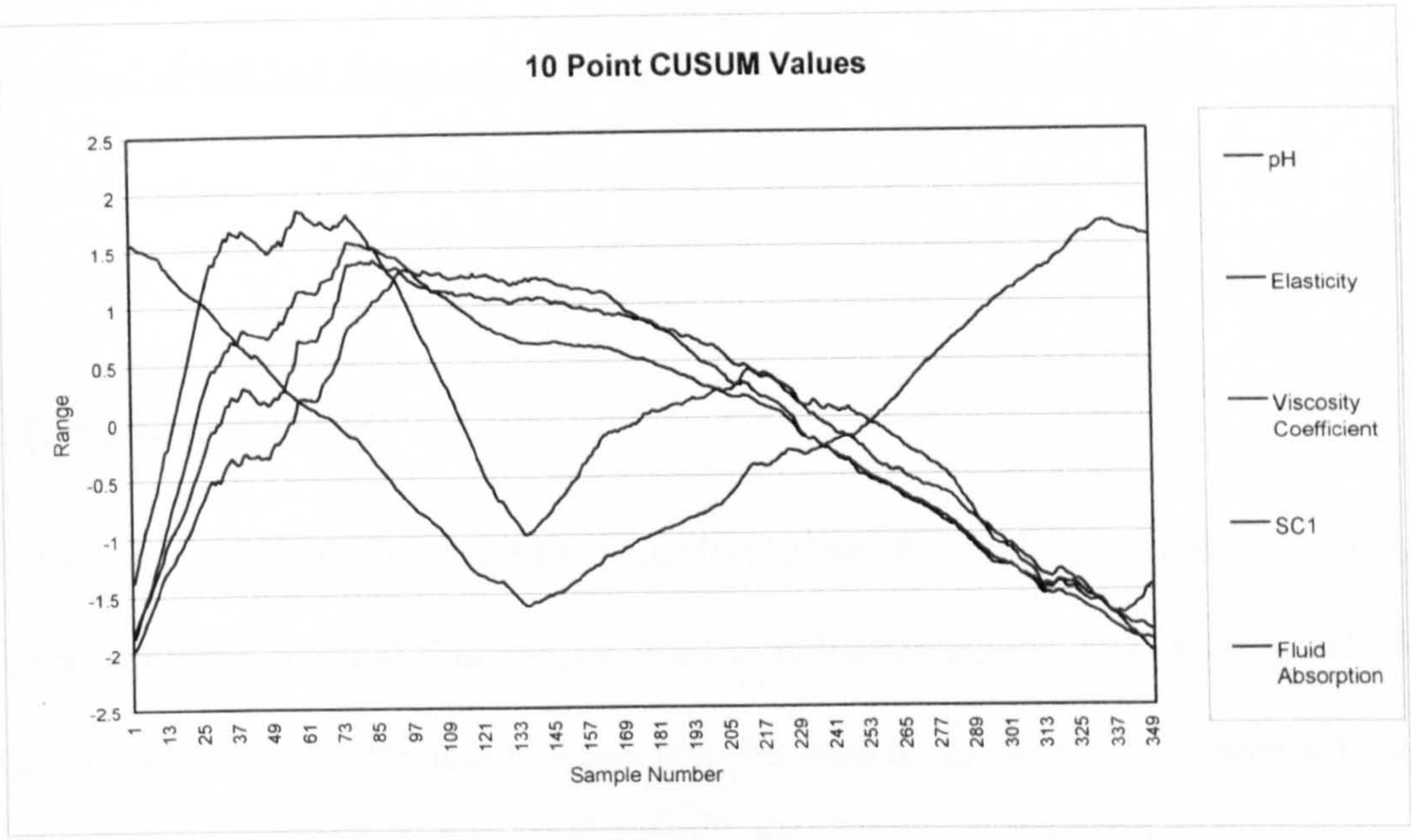


Figure 4.29 CUSUMS for the Analysis Data Set, CUSUM Derived from Every Tenth Sample

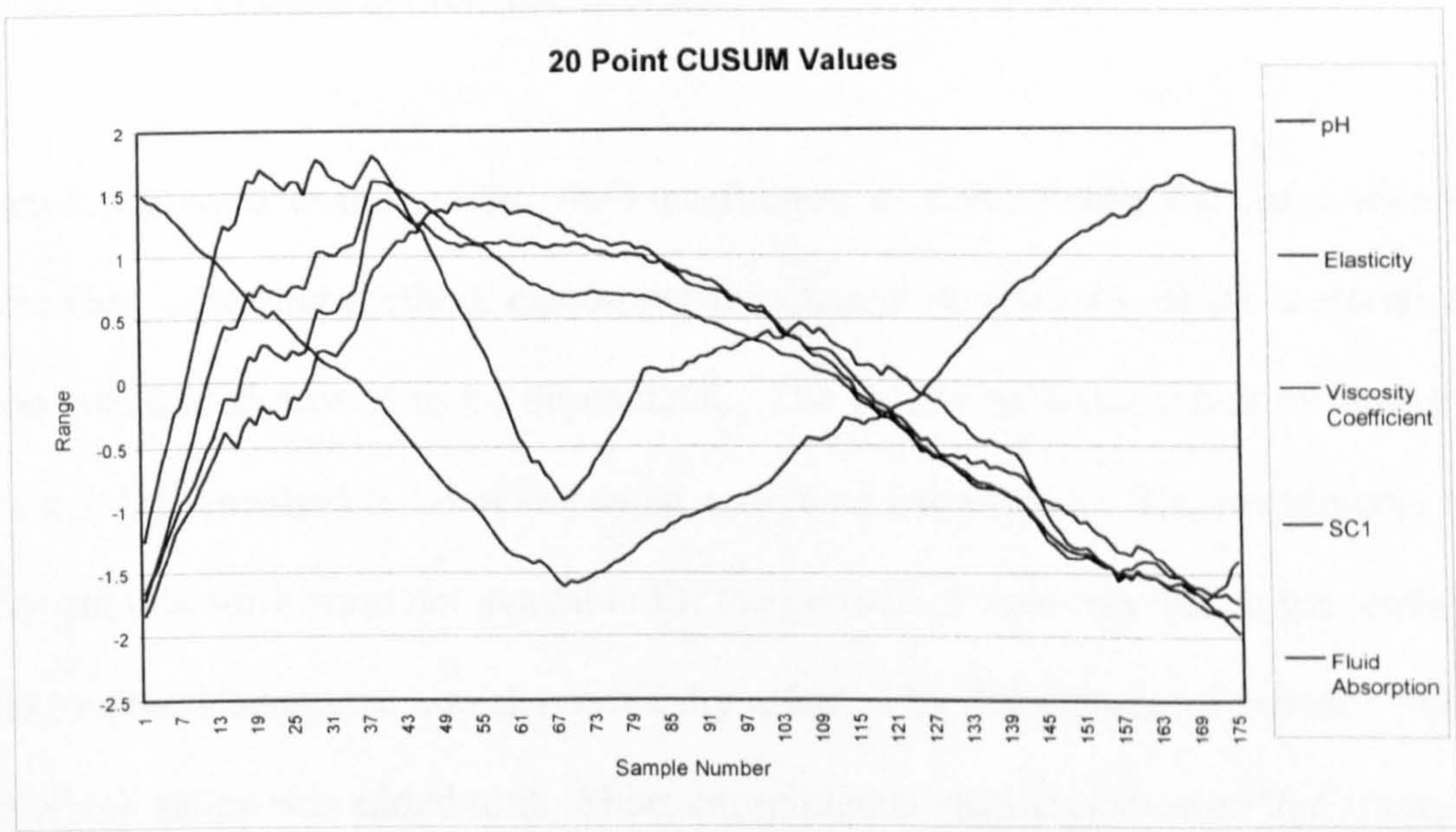


Figure 4.30 CUSUMS for the Analysis Data Set, CUSUM Derived from Every Twentieth Sample

These graphs all show that the trends observed in the Intrasite data set are visible at very reduced sampling frequencies. Despite the high random error in the measurements this shows that the underlying functions are all very well controlled and that change in the Intrasite Gel properties is a slow drift, the flat sections of each CUSUM curve correspond

very well with the known dates when the bulk polymer batches changed and means that the process is actually stable during production for each bulk polymer batch. the majority of the variation in the CUSUM charts could be explained by raw material changes.

4.8 Reference Data

The alternatives available now are to either to replace the settling volume test altogether or produce a reference method that can be used to calibrate against the settling volume test to find the correct values. The new test can also be used to provide a reference set that its self can be calibrated for, allowing the test to be discontinued. The selection of method will determine which of these approaches is best.

Four methods were examined for their usefulness in determining the fluid absorption of Intrasite Gel. The first method, examining the change in viscosity of the material as saline solution was added proved to be impractical. The uptake of water/saline by Intrasite gel is too slow for this method to be of use as an analytical experiment. The instruments available to carry out this work were not available for the periods of time required to get readings. The response of the viscometer was also critically affected by the volume of Intrasite gel, and the rate at which saline was added to it. Short experimental tests also showed that Intrasite Gel is thixotropic to quite a high degree, showing sheer thinning, and has a slow recovery. These factors would have made accurate testing difficult. The second method was based on the standard method for examining the fluid uptake by gels and foams, the material to be examined is placed into a sealed volume and the liquid who's uptake it to be examined is pumped into the gel or foam. Intrasite gel was found to be unsuitable for this as the flow rate of the saline would be prohibitively slow for accurate measurements to be made. The third

method considered was the standard method from the British Pharmacopoeia for the measurement of fluid uptake. Known as the tea bag method it involves immersing a known mass of the material in the liquid whose uptake is being examine until equilibrium is reached, removing the material and re-weighing it. The change in mass is related to the amount of fluid absorbed. This is not appropriate with Intrasite Gel as the material is known to be soluble in water and saline solution.

4.9 Intrasite Experiment 8, The "Paddington Cup" Method

The method selected for the examination of the fluid absorption of Intrasite Gel is known as the "Paddington Cup" method. The method was developed to examine the difference is fluid transfer properties of other materials designed to carry out a similar task as Intrasite Gel. The test was designed to examine the fluid transfer properties of materials that could have widely differing characteristics. Although the test takes a significant amount of time to carry out, much of that is waiting for equilibriums to occur and there is no need for an analyst to be present during this time. The test was initially carried out on a group of competing products for comparison purposes.

For this test only two media were used, 30% gelatine and 2% agar; these two materials had been selected as the best to highlight the differences of the materials being tested. Insufficient material existed for the tests to be run across the full range of substrates. Supplies limited the number of replicates that could be carried out and some of the materials could only be tested with a single substrate, agar 2%.

4.9.1 Results for 2% Agar

The products tested were the following materials.

1.x Sterigel LOT SG0196A
2.x Nu-Gel LOT 160196.18
3.x Solosite A50905B
4.x Curasol KKEI
5.x Carrasyn V 7 / 98 / AB
6.x Aquaform 1194 / 20
7.x Granugel 96050044
8.x Carrasyn F10 / 97

Tables 4.11a and 4.11b show the results of the fluid transfer test, and Table 4.12 explains the various column headings. All measurements were made in grams using a four figure balance. The results are graphed in figure 4.31. The experiment was straight forward to carry out however there are various stages during the experiment where experimental error is expected to have a significant effect. The hydrogel must be in clear contact with the substrate for good fluid transfer to take place and with these materials it is often difficult to ensure that no air is trapped. Also removing the hydrogel after equilibrium is also expected to introduce error since all the hydrogel must be removed and recovered for accurate results, this can sometimes be difficult due to the fragile nature of the agar substrate.

Material	W1 (g)	W2 (g)	W3 (g)	W4 (g)	W5 (g)	W6 (W2-W1)	W7 (W3-W2)	W8 (W4-W3)	W9 (W5-W2)	W10 (W5-W2 / W6 * 100)	W11 ((W4-W5)-W7)*100/W7)
										Change in Agar	Change In Hydrogel
1.1	32.38	43.19	53.21	53.04	43.01	10.81	10.02	-0.17	-0.18	-1.66512	0.0998
1.2	33.61	43.42	53.48	53.06	43.32	9.81	10.06	-0.42	-0.1	-1.01937	-3.18091
1.3	33.45	43.37	53.33	53.19	43.17	9.92	9.96	-0.14	-0.2	-2.01613	0.60241
1.4	33.53	43.3	53.29	53.18	43.14	9.77	9.99	-0.11	-0.16	-1.63767	0.500501
1.5	32.99	42.75	53.72	53.62	42.59	9.76	10.97	-0.1	-0.16	-1.63934	0.546946
2.1	33.59	43.5	53.53	53.38	40.85	9.91	10.03	-0.15	-2.65	-26.7407	24.92522
2.2	33.21	43.15	53.19	53.12	40.67	9.94	10.04	-0.07	-2.48	-24.9497	24.00398
2.3	33.3	43.21	53.16	53.08	40.74	9.91	9.95	-0.08	-2.47	-24.9243	24.0201
2.4	32.86	42.75	52.68	52.62	40.42	9.89	9.93	-0.06	-2.33	-23.5592	22.86002
2.5	32.91	42.84	52.83	52.72	40.26	9.93	9.99	-0.11	-2.58	-25.9819	24.72472
3.1	32.68	43.64	53.73	53.61	42.34	10.96	10.09	-0.12	-1.3	-11.8613	11.69475
3.2	33.46	43.27	53.3	53.19	42	9.81	10.03	-0.11	-1.27	-12.946	11.5653
3.3	33.49	43.48	53.48	53.37	42.31	9.99	10	-0.11	-1.17	-11.7117	10.6
3.4	33.19	43.04	53.16	53.12	41.88	9.85	10.12	-0.04	-1.16	-11.7766	11.06719
3.5	32.87	42.84	52.77	52.65	41.68	9.97	9.93	-0.12	-1.16	-11.6349	10.47331
4.1	33.09	43.05	53.02	52.96	41.44	9.96	9.97	-0.06	-1.61	-16.1647	15.54664
4.2	33.27	43.21	53.3	53.26	41.64	9.94	10.09	-0.04	-1.57	-15.7948	15.16353
4.3	33.52	43.47	53.44	53.37	41.86	9.95	9.97	-0.07	-1.61	-16.1809	15.44634
4.4	32.93	42.77	52.74	52.67	41.34	9.84	9.97	-0.07	-1.43	-14.5325	13.64092
4.5	33.04	42.88	52.85	52.8	41.37	9.84	9.97	-0.05	-1.51	-15.3455	14.64393

Table 4.11a First Half of Table 4.11 Fluid Transfer Test Results for Different Hydrogels Using 2% Agar as Substrate

Material	W1 (g)	W2 (g)	W3 (g)	W4 (g)	W5 (g)	W6 (W2-W1)	W7 (W3-W2)	W8 (W4-W3)	W9 (W5-W2)	W10 (W5-W2 / W6 * 100)	W11 ((W4-W5)-W7)*100/W7)
										Change in Agar	Change In Hydrogel
5.1	33.62	43.48	53.54	53.39	43.32	9.86	10.06	-0.15	-0.16	-1.62272	0.099404
5.2	33.27	43.23	53.17	53.02	43.08	9.96	9.94	-0.15	-0.15	-1.50602	0
5.3	32.26	42.21	52.18	52.09	42.07	9.95	9.97	-0.09	-0.14	-1.40704	0.501505
5.4	32.95	42.85	52.88	52.8	42.68	9.9	10.03	-0.08	-0.17	-1.71717	0.897308
5.5	32.32	42.23	52.13	52	42.03	9.91	9.9	-0.13	-0.2	-2.01816	0.707071
6.1	32.81	42.76	52.79	52.7	39.28	9.95	10.03	-0.09	-3.48	-34.9749	33.7986
6.2	33.57	43.45	53.43	53.37	39.74	9.88	9.98	-0.06	-3.71	-37.5506	36.57315
6.3	33.66	43.64	53.66	53.55	39.69	9.98	10.02	-0.11	-3.95	-39.5792	38.32335
6.4	33.61	43.63	53.72	53.66	39.7	10.02	10.09	-0.06	-3.93	-39.2216	38.35481
6.5	32.29	42.28	52.27	52.22	38.83	9.99	9.99	-0.05	-3.45	-34.5345	34.03403
7.1	33.5	43.48	53.55	53.49	40.44	9.98	10.07	-0.06	-3.04	-30.4609	29.59285
7.2	32.42	42.43	52.45	52.38	39.69	10.01	10.02	-0.07	-2.74	-27.3726	26.64671
7.3	32.45	42.47	52.56	52.5	39.67	10.02	10.09	-0.06	-2.8	-27.9441	27.1556
7.4	33.04	42.83	52.88	52.86	40.02	9.79	10.05	-0.02	-2.81	-28.7028	27.76119
7.5	32.55	42.53	52.5	52.48	39.66	9.98	9.97	-0.02	-2.87	-28.7575	28.58576
8.1	32.25	42.29	52.35	51.78	42.18	10.04	10.06	-0.57	-0.11	-1.09562	-4.57256
8.2	33.6	43.51	53.43	53.17	43.27	9.91	9.92	-0.26	-0.24	-2.4218	-0.20161
8.3	33.66	43.55	53.49	53.44	43.33	9.89	9.94	-0.05	-0.22	-2.22447	1.710262
8.4	32.9	42.86	52.79	52.74	42.62	9.96	9.93	-0.05	-0.24	-2.40964	1.913394
8.5	32.38	42.35	52.48	52.44	42.12	9.97	10.13	-0.04	-0.23	-2.30692	1.875617

Table 4.11b Second Half of Table 4.11 Showing Fluid Transfer Test Results on Different Hydrogels

Page numbering as original

Reference	Calculation	Meaning
W1		Weight of Syringe (g)
W2		Weight of Syringe + Substrate (g)
W3		Weight of Syringe + Substrate + Hydrogel (g)
W4		Weight after Equilibrium (48 Hours) (g)
W5		Weight of Syringe + Substrate (g)
W6	(W2-W1)	Mass of Substrate in Syringe
W7	(W3-W2)	Mass of Hydrogel Added
W8	(W4-W3)	Change in Mass of Whole Syringe After Equilibrium
W9	(W5-W2)	Change in Mass of Substrate After Equilibrium
W10	(W9/W6*100)	Percentage Change in Substrate Mass
W11	(((W4-W5)-W7)*100)/W7)	Percentage Change in Hydrogel Mass

Table 4.12 Index to Explain The Column Heading for Fluid Transfer Test Results

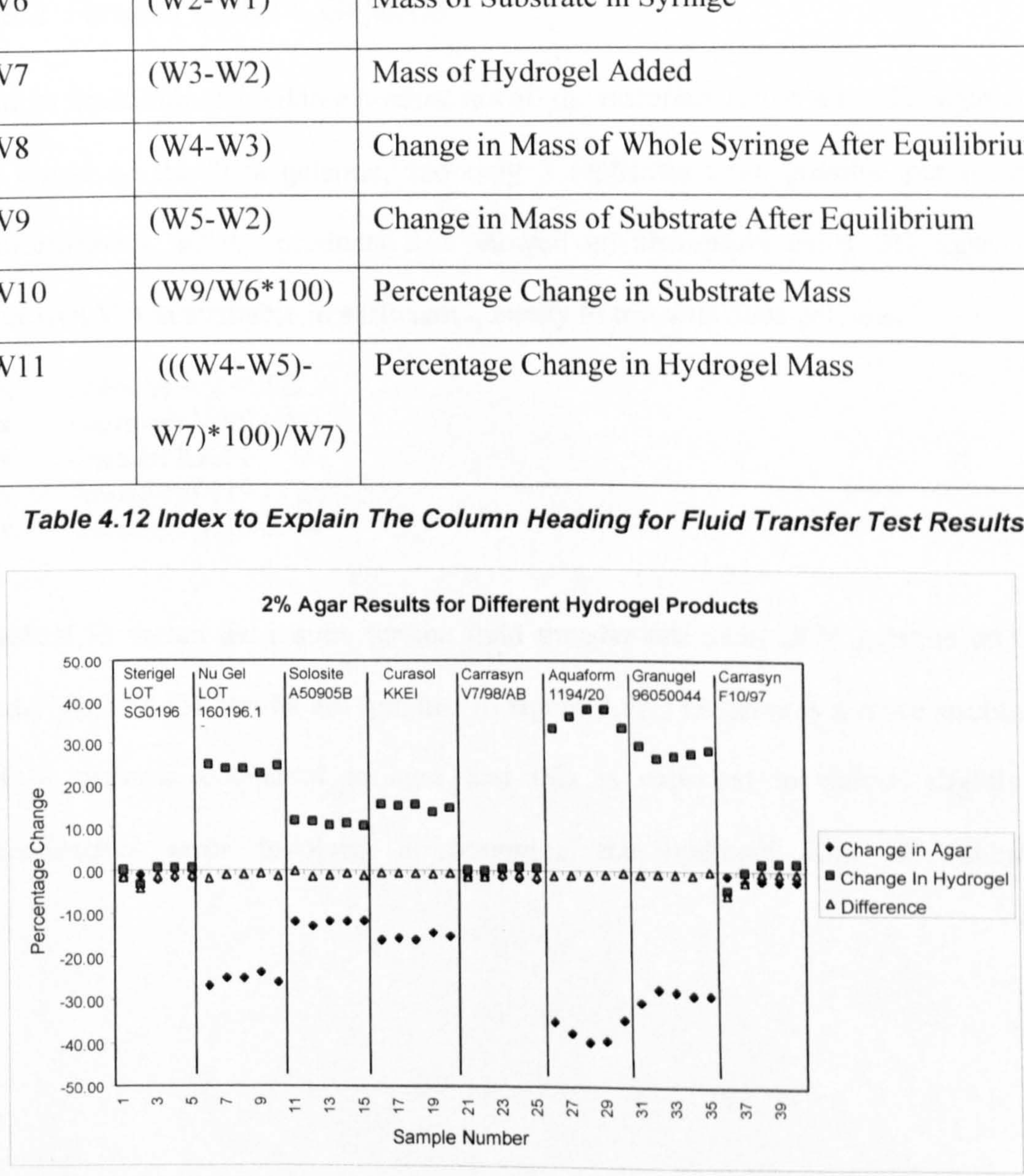


Figure 4.31 Graph Displaying the Results for the 2% Agar Fluid Transfer Test

Although there is no data regarding the expected results, the results clearly show that there are differences in the fluid transfer properties of the various materials tested (Figure 4.31). The results for Sterigel and the two Carrasyn products are due to the materials being very moist, and not absorbing fluid to any significant amount. Where the materials do absorb fluid, there are clear differences between the various products.

4.9.2 Results for 30% Gelatine

Due to limitation of available product not all the materials tested using 2% agar could be tested on the 30% gelatine, and only 3 replicates were possible per material. Unfortunately, of the products that showed no absorption using 2% agar, only Carrasyn V was available in sufficient quantity to test with 30% gelatine.

1.x Solosite A50905B
2.x Carrasyn V7/98/AB
3.x Curasol KKEI
4.x Aquaform 1194 / 20
5.x Granugel 96050044

Table 4.13 shows the results for the fluid transfer test using 30% gelatine on these materials, and the results are graphed in figure 4.32. Gelatine is a more structurally robust material compared to agar and this is expected to reduce slightly the experimental error involved in removing the hydrogel from the substrate.

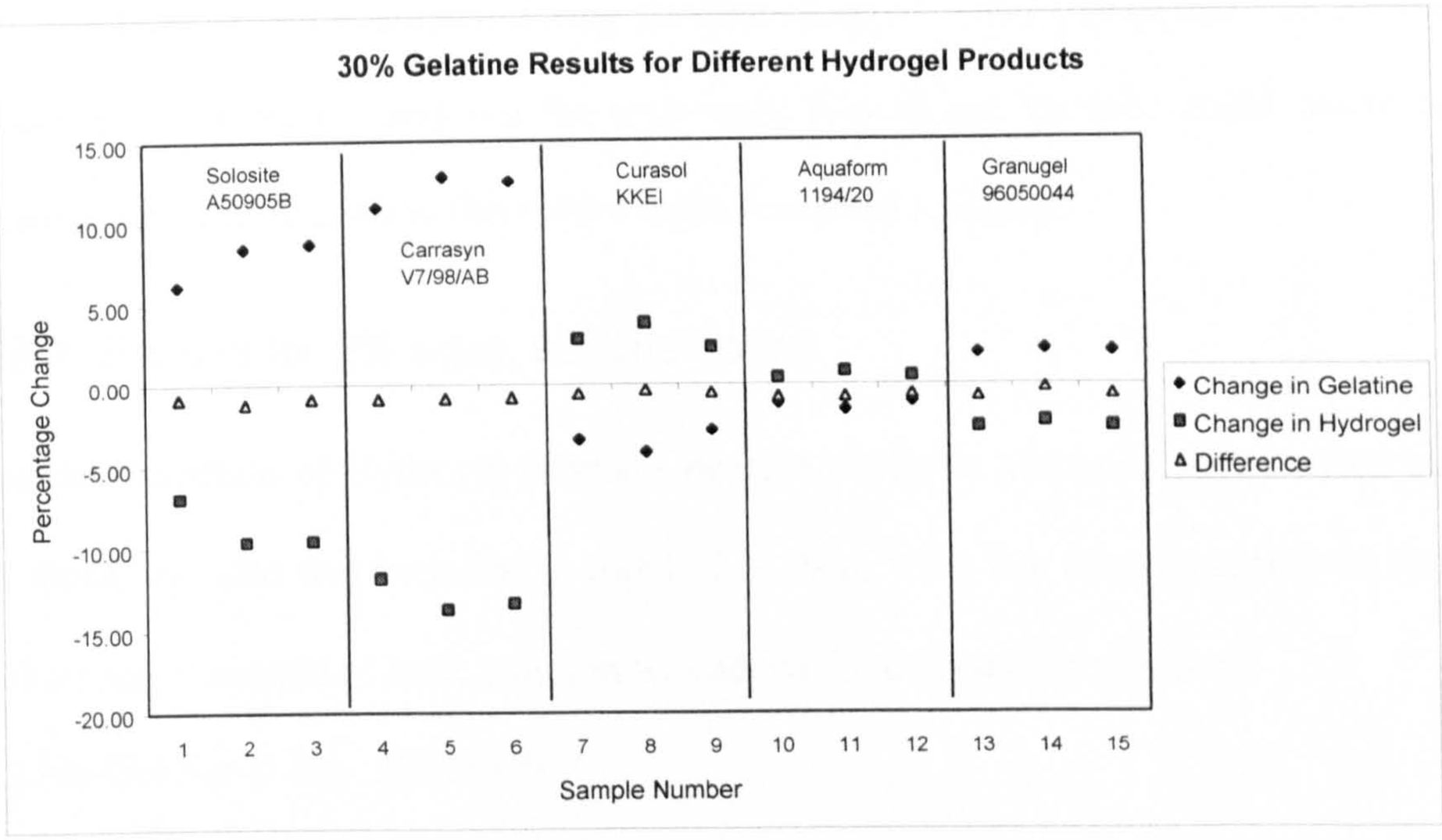


Figure 4.32 Graph Plotting the Results from Table 4.13, Fluid Transfer Test Results for Different Hydrogels on 30% Gelatine

This experiment shows the differences in fluid donation properties between the various products (Figure 4.32). Carrasyn clearly is a fluid donator under these conditions, as is Solosite. Solosite appears from these tests to have fluid transfer properties that fit between the fluid transfer properties of the 2% Agar, and the 30% Gelatine. Curasol, Aquaform and Granugel may possibly be fluid donators under these conditions, however it is not clear from these results. The overall response to gelatine produces fluid transfer values of a smaller magnitude to the values shown using 2% agar; this might be expected to effect the relative error of these measurements as the other factors are constant

The "Paddington Cup" test was designed to be used with four grades of agar, and four grades of gelatine, using all these materials would probably enable the various materials to be completely separated in terms of their fluid transfer properties. The possibility also exists that this test could be used to identify the various gels, however the test is not physically practical for that, and easier methods exist to do that task.

Intrasite Gel was not examined during the first series of tests for practical reasons, the resources available to carry out the tests were limited and Intrasite could easily be tested at another time when there were more resources available.

4.9.3 Results for 2% agar, second series

Another selection of Hydrogel products became available and tests were carried out on them, Intrasite was included in this test as there were few other materials to test, and a control sample of ionic solution was added for comparison purposes.

1.x Nu-Gel Serial No. 23019629

2.x Sterigel Serial No. SG1295a

3.x Serial No. 920900

4.x Intrasite Gel Serial No. 941215

5.x Ionic Solution (Control)

Table Appendix III.2 shows the results for this series of experiments and the results are plotted in figure 4.33.

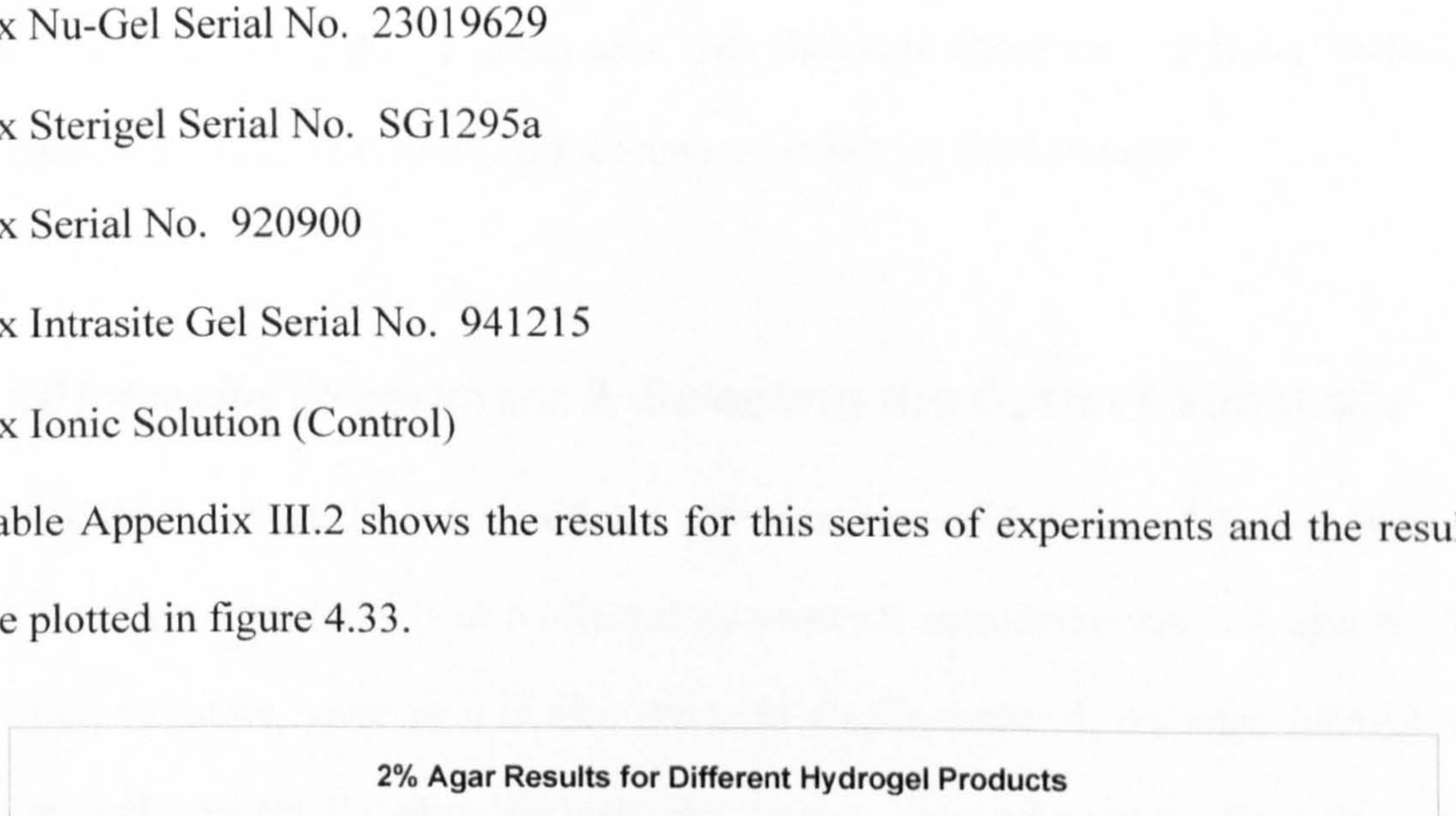


Figure 4.33 Graph Showing the Results from Table 3.14, Results of Second Series of 2% Agar Tests Using Different Hydrogels

The Nu-Gel and the Sterigel behaved in a similar manner to that seen before (Figure 4.33 cf. Figure 4.31), the variation could be put down to either variation in batches, or experimental error, no evidence exists to support one of these over the other. The spike in the Ionic Solution sample was produced by experimental error, some of the ionic solution leaked beneath the agar plug, and was weighed with the agar, not the rest of the ionic solution. This is not expected to be the reason for any great error with the other materials since they are viscous materials. The agar plug does shrink as fluid is removed from it, and this allows hydrogel to surround the plug. This may account for some of the variation seen with the more absorbent materials, including Intrasite Gel, as there is more surface area available for fluid transfer.

4.10 Intrasite Experiment 9, Selecting the Correct Substrate

The substrate selected has a significant effect on the performance of the test however the selection of material is also affected by practical considerations. 1% agar has the greatest response however it is also the most fragile material, the agar disintegrates quite easily, especially when the hydrogel is being removed from it. 2% Agar is also fragile, however not to the extent that it is unusable. Although gelatine is mechanically superior to the agar its fluid transfer is lower than agar, and it is more difficult and time consuming to prepare. Overall 2% agar appears to be the material that is most appropriate to use in the Intrasite reference data test series.

4.11 Intrasite Experiment 10, Generating the Reference Data Set

Over a five-week period 45 Intrasite Gel samples were tested using the fluid transfer test. Each Sample was carried out using three replicates, except the first four samples for which there was only sufficient material for two replicates. The tests were run concurrently with the normal Intrasite tests, as material was also required for archival purposes, this limited the number of replicates that were possible.

The samples taken and the results for the standard Intrasite tests show that this period was quite stable for all the variables, this information can be found in table 4.15, and the results for the fluid transfer tests can be seen in table 4.16. This data was plotted by individual sample in figure 4.34, and then plotted again as the average of the replicates in figure 4.35.

Date	Batch Number	PH	Elasticity	Viscosity	SCI	Fluid Absorption
28/01/97	970355	7.4	1637	240	2.5	80
28/01/97	970356	7.3	1606	230	2.4	70
28/01/97	970365	7.4	1771	240	2.3	85
28/01/97	970366	7.4	1612	230	2.3	70
29/01/97	970373	7.3	1333	200	2.2	60
29/01/97	970374	7.4	1392	210	2.2	70
29/01/97	970375	7.4	1515	220	2.2	70
31/01/97	970431	7.4	1358	200	2.2	70
31/01/97	970432	7.5	1400	200	2.2	70
01/02/97	970435	7.4	1431	210	2.2	60
31/01/97	970442	7.4	1414	210	2.2	60
31/01/97	970451	7.5	1583	220	2.3	70
01/02/97	970452	7.4	1514	220	2.3	60
10/02/97	970546	7.4	1494	220	2.3	70
10/04/97	971354	7.2	1218	190	2.1	60
10/04/97	971361	7.3	1154	180	2.1	60
10/04/97	971362	7.1	990	170	2.0	60
10/04/97	971363	7.4	1014	170	2.0	60
10/04/97	971364	7.4	1045	170	2.0	60
10/04/97	971412	7.1	1154	190	2.1	60
10/04/97	971413	7.3	1252	190	2.1	60
10/04/97	971425	7.4	1562	230	2.3	60
10/04/97	971511	7.4	1476	220	2.3	70
15/04/97	971513	7.3	1402	210	2.1	60
15/04/97	971521	7.2	1690	240	2.3	80
15/04/97	971522	7.5	1532	230	2.3	70
15/04/97	971523	7.3	1578	230	2.3	70
16/04/97	971531	7.5	1641	230	2.3	70
16/04/97	971533	7.1	1375	210	2.2	70
16/04/97	971541	7.4	1623	240	2.4	80
16/04/97	971543	7.4	1521	220	2.3	75
19/04/97	971612	7.3	1303	200	2.2	75
19/04/97	971613	7.4	1296	200	2.1	60
21/04/97	971614	7.4	1265	190	2.1	60
19/04/97	971615	7.4	1302	200	2.2	60
18/04/97	971621	7.3	1884	230	2.3	60
19/04/97	971622	7.3	1418	220	2.2	60
19/04/97	971623	7.3	1419	220	2.3	65
19/04/97	971624	7.4	1488	220	2.3	60
24/04/97	971633	7.2	1459	210	2.2	70
24/04/97	971634	7.3	1529	220	2.3	70
24/04/97	971635	7.4	1377	210	2.3	70
24/04/97	971641	7.3	972	170	1.9	50
24/04/97	971642	7.4	1006	170	2.0	60
24/04/97	971643	7.3	1316	200	2.1	70

Table 4.13 Batch Information for the Samples Used in the 2% Agar Fluid Transfer Tests to Generate Reference Data

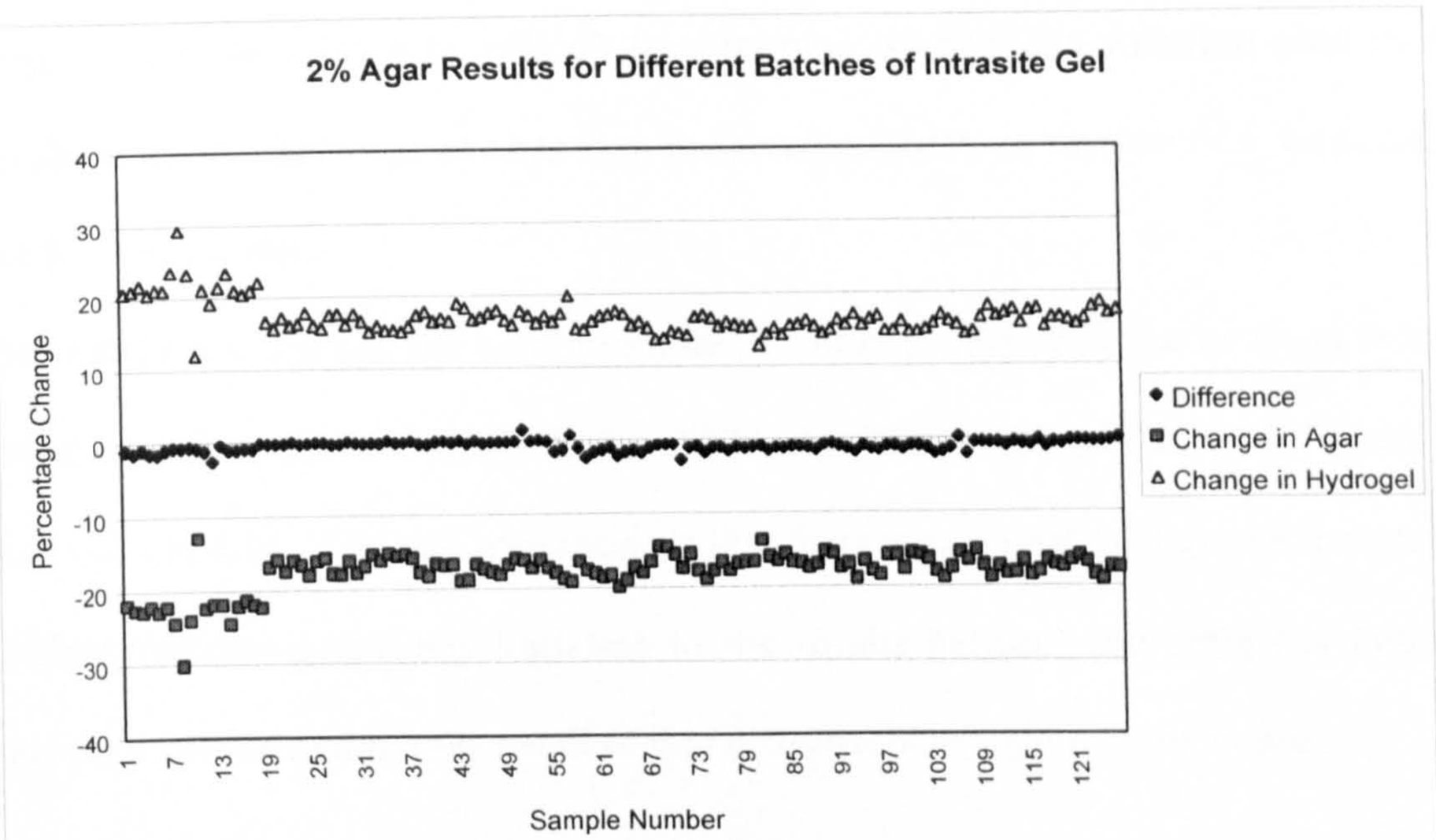


Figure 4.34 Graph Showing Results from Fluid Transfer Tests for Reference Data, Individual Values Plotted

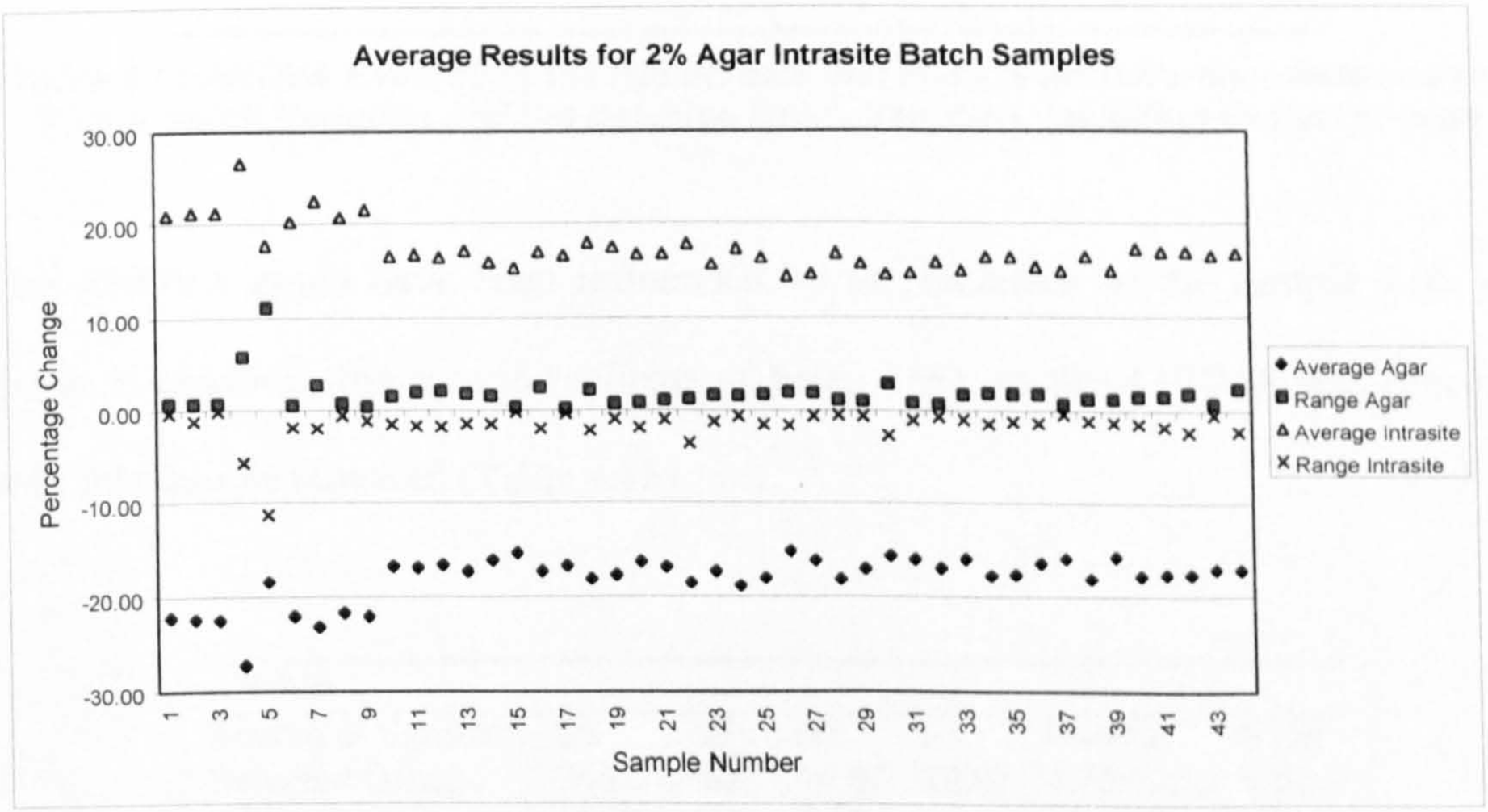


Figure 4.35 Average of Results for Replicates from the Fluid Transfer Test of Samples from Table 4.15

The large variations seen in the initial results is not explained, though a possible reason is that the Intrasite is introduced to the substrate via a syringe, and in the first few samples poor experimental technique may have enabled large air bubbles to form,

affecting the available surface are for the fluid transfer to take place. From the graphs (Figures 4.34 & 4.35) it is difficult to determine whether the variation seen in the results is greater between batches than between replicates so an ANOVA was carried out to examine this.

The ANOVA examined the within sample variation compared to the between sample variation, this required a single factor ANOVA. The results of the ANOVA can be seen in table 4.14. The null hypothesis is that there is no variation between the groups of data, indicating that any variation in the results between the groups is entirely random variation and not the result of real differences in fluid transfer values.

ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	846.31	44	19.23	12.53	1.68 E-23	1.51
Within Groups	138.14	90	1.53			

Table 4.14 ANOVA Evaluating the Hypothesis that there Is no Difference between the Within Batch Variation and the Between Batch Variation, Including Outlying Value

The ANOVA could have been influenced by the inclusion of the sample with the extreme variation (the second replicate of batch 356), so the ANOVA was repeated with this sample removed (Table 4.15)

ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	846.01	43	19.67	22.57	1.79 E-32	1.51
Within Groups	76.70	88	0.87			

Table 4.15 ANOVA Evaluating the Hypothesis that there Is no Difference between the Within Batch Variation and the Between Batch Variation, Excluding Outlying Value

In both cases the F value exceeds the F_{crit} value and the null hypothesis must be rejected, that is the variation between samples is greater than the variation between the replicates. This results indicates that the test is sensitive enough to detect the

differences in fluid absorption between different samples of Intrasite Gel. This test uses mass change not eye measurement of a graduated cylinder, and there is no chance of the gel being examined passing into solution, this means that the two biggest sources of error in the settling volume test are not present in this new test. This test is however time consuming to carry out, and is probably not a suitable replacement to the settling volume test by itself. An alternative is to change the registered test for fluid absorption to the fluid transfer test and then calibrate for this test, using the predicted values instead of the measured values.

The earlier calibration attempts have shown that the data set contains a high degree of noise, so for this examination PLS was selected as the calibration method immediately.

The variance captured by the PLS model is shown in table 4.16

-----X-Block-----			-----Y-Block-----	
LV #	This LV %	Total %	This LV %	Total %
1	73.68	73.68	63.29	63.29
2	23.42	97.10	2.03	65.33
3	0.77	97.87	0.67	66.00
4	2.13	100.00	0.11	66.11

Table 4.16 Table to Show the Information Captured by PLS using the Fluid Transfer Test Results and the Analysis Results

From this two LVs were selected as the appropriate number of factors to model with, the modelling was carried out twice, with two randomly selected data sets of thirty points used for the training set, and the remaining fifteen points used for validation.

The results for the PLS modelling of the two test sets can be seen in figure 4.36 and 4.37.

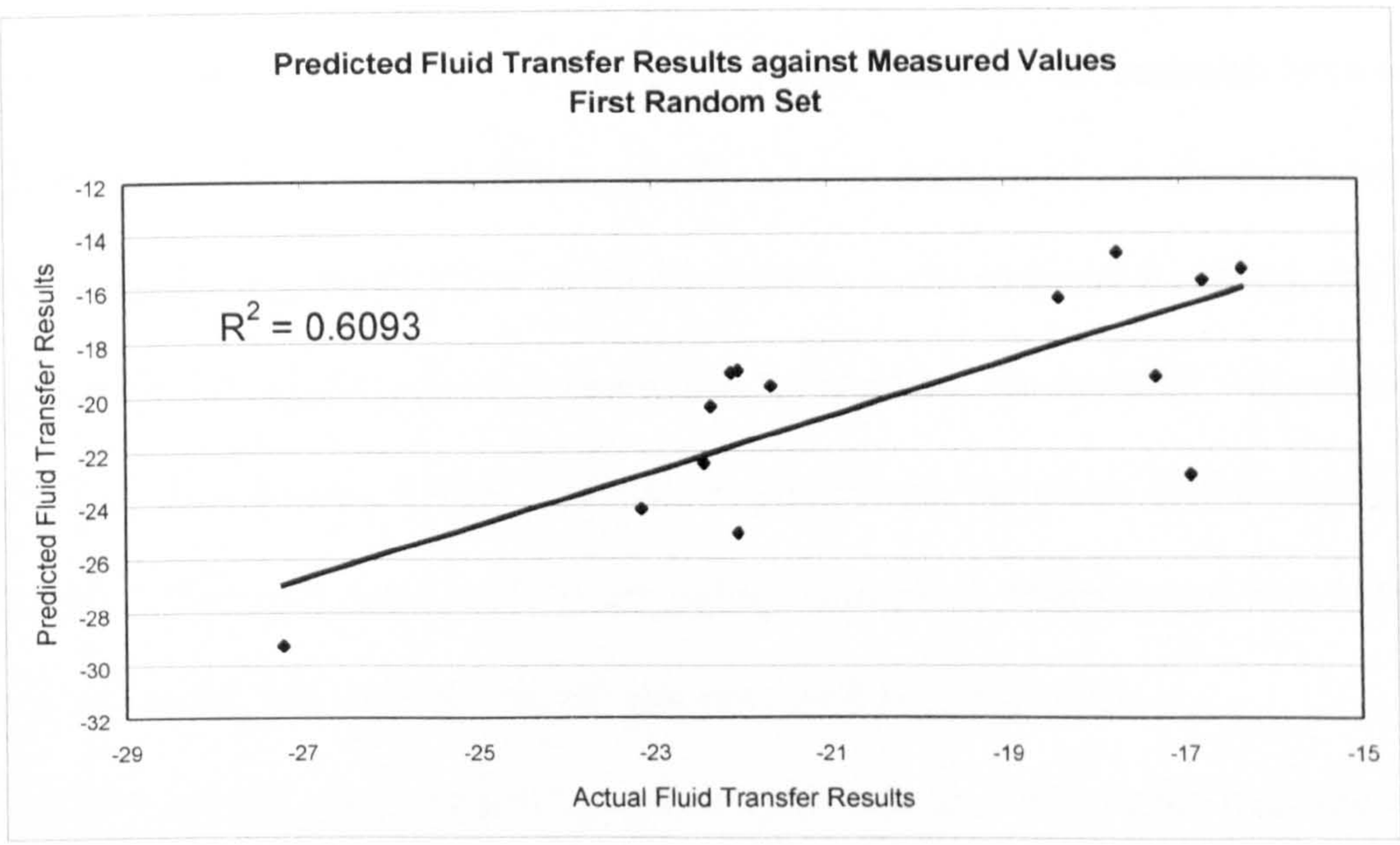


Figure 4.36 Predicted Fluid Transfer Results vs. Actual Fluid Transfer Results for the PLS Calibration of the Reference Data Set, First Random Selection of Samples, R^2=0.6093

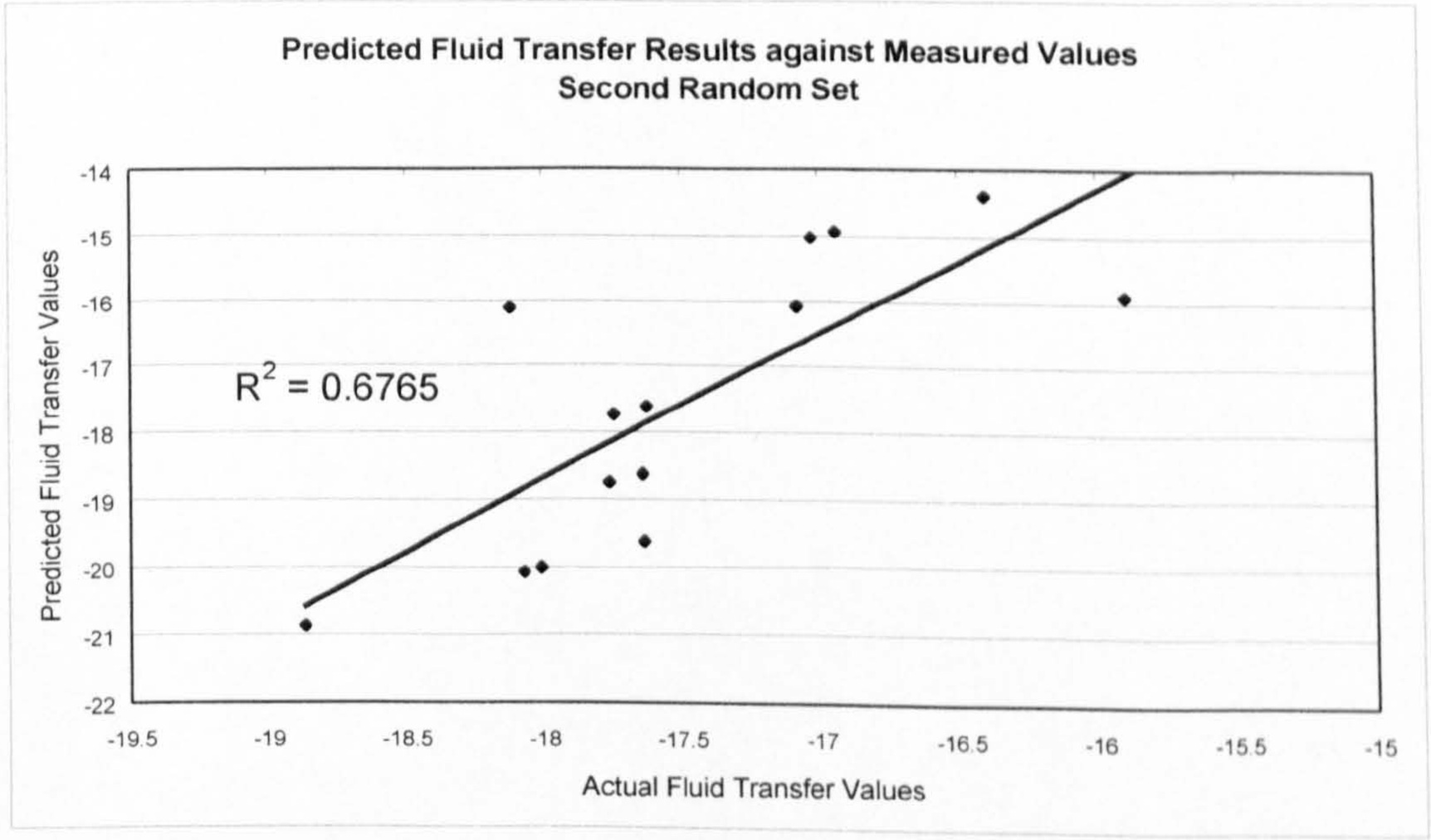


Figure 4.37 Predicted Fluid Transfer Results vs. Actual Fluid Transfer Results for the PLS Calibration of the Reference Data Set, Second Random Selection of Samples, R^2=0.6765

These correspond to an RSD of 11.8% (Figure 4.36) and 10.8% (Figure 4.37) respectively. These errors are large and do not suggest a particularly robust model. The sources of the error in the model require further investigation, it is likely that experimental error in the Y block is a big factor, the known variation between the replicates is ± 4.5%. This could account for a large fraction of the error present in the model. The error present in the method is mostly in the material handling, the fragile nature of the 2% agar substrate is one cause for concern, the material fragments easily and any fragmentation would give rise to significant variation in the results. The conditions of the test need to be more tightly controlled, temperature was monitored by not controlled for this experiment and may well be a big factor.

The indications are that this test produces more reproducible results than the settling volume test, and depending on the sources of error present in the analytical procedures as a whole a strong possibility exists for improving this method and model.

5. Conclusions

This thesis is broken down into two parts, the first aim of this document is to describe the development of a variable selection procedure for Projected Latent Structures (PLS), using MATLAB ™ and the PLS Toolbox (Barry Wise, Eigenvector Research) [77]. The development of this variable selection algorithm is broken down into five stages showing the important changes and decisions made during the development process. The second aim of this thesis is to record the examination of Intrasite Gel for Smith & Nephew, Hull. Intrasite gel is a Sodium Carboxymethyl Cellulose gel made with water and propylene glycol, and is registered in most countries of the world as a medical device. As a medical device there is a requirement that certain properties of the material are measured regularly to assess the suitability of the material to its intended purpose, and to ensure that it meets the specifications by which it is sold. An investigation was called for into the relationships between the various parameters recorded, and an investigation was required to examine the stability of the process used to manufacture Intrasite Gel.

5.1. Variable Selection Projected Latent Structures (VS-PLS)

This work was begun with the premise that the work that had been carried out on variable selection MLR [1] could be applied to PLS. There are strong reasons for variable selection using MLR, most important is the fact that most MLR systems will be underdetermined. PLS is often thought to removed most of the problems associated with MLR, and thus might be considered a poor candidate for variable selection, however in the calibration of a system to predict the concentrations of a

component there can be no reason to include in the model variables that provide absolutely no information to the model. In modern analytical instruments many thousands of variables can be recorded simultaneously, and only a very small percentage of these will have any information to provide. PLS uses coefficients to weight input variables into groups according to their importance towards the systems being calibrated. PLS requires that all variables have a weight, and thus all variables have a contribution towards the overall model however small. The greater the number of latent vectors that are included in the model the greater the overall contribution from variables with no information. This suggests that variables be removed from a data set to leave only those required to model the system being examined.

The procedure that is used to select these variables will have a large impact on the quality of the model produced and its ability to produce robust answers during prediction. This work looks at an iterative procedure that examined the effect on the prediction results for a model when the variables used to produce the model are changed.

The starting point for the development of the VS-PLS algorithm was the current state of the VS-MLR algorithm, a single variable addition procedure (SVA-MLR), which was an iterative method based on adding a randomly selected variables into a model to determine whether the added variable has a positive or a negative effect. The reason for random addition is to reduce the problems associated with collinearity, if two collinear variables are presented to the model one after the other there can be unpredictable effects, either both variables may be selected leading to redundancy in the model or the second variable, which may be superior to the first, will be rejected,

leaving a less robust model. This method was transferred directly to PLS to become single variable addition PLS (SVA-PLS) This method used a single stage, starting with a single variable and adding variables individually, testing the model after each addition to examine how the model performance had changed. This method was not expected to produce particularly good results with variable selection PLS as the initial starting position was with a single variable which would produce an unstable and poor PLS model. The result of this would be that the model would accept the addition of any variable added to the model until the point at which sufficient variables had been added to the model to produce a stable solution. If by chance the variables added to the model were all unsuitable in the early stages of modelling then a particularly poor model would be produced with a large number of variables added, it would be possible for such a model to perform worse than an ordinary PLS model. The tests using the three data sets showed that these expectations were true, and that while the models produced were often superior to the ordinary PLS models, there was evidence that there was significant levels of redundancy in the variable selected.

The failings of this first attempt (SVA-PLS), the unstable original model, and the large number of redundant variables, required that the method be improved. This was addressed by correcting the initial unstable model produced with a single variable. The new method started the modelling procedure with a number of randomly selected variables equal to the number of components in the Y-Block, and then adding in a number of variables in each iteration equal to the number of variables started with, this became multiple variable addition PLS (MVA-PLS). This method was found to be particularly poor. This is because the random addition of a group of variables just simulated the first few additions in the original method (SVA-PLS) without the option

of rejecting variables that were particularly poor, this was compounded by the addition of multiple variables after this, allowing variables to be accepted or rejected only in blocks. This led to the situation that several poor variables could be selected to allow the inclusion of a single good variable. This method produced models with more variables than SVA-PLS and was rejected immediately.

Following the failings of this second attempt the problem as approached from a different angle, SVA-PLS could produce reasonable models however they often contained variable that were unsuitable because they were selected in the early modelling stages when the model was unstable. The solution to this was to allow the opportunity for unsuitable variable to be removed. This was carried out by including a removal stage subsequent to the addition stage, once "candidate" variable had been selected they were tested by examining the model performance when these variables were removed individually. This became single variable addition single variable removal PLS (SVA-SVR-PLS). SVA-SVR-PLS appeared to solve many of the problems associated with the original methods, SVA-PLS and MVA-PLS, the prediction errors were smaller, and there was a significant reduction in the number of variable selected.

This method (SVA-SVR-PLS) performed well, improving over ordinary PLS, and the two previous variable selection methods, however this raised the question as to whether the use of the addition stage initially was actually improving the model or whether the algorithm would perform as well without the initial per-selection. This was tested by removing the addition stage entirely and writing the algorithm using only the single removal stage – singe variable removal PLS (SVR-PLS). As expected

when this method was used it was found there was very little change in the models produced. This confirmed that the initial variable addition stage was unnecessary and that the routine would perform well without it.

The original premise for variable selection was to remove variables with noise and variable containing only highly correlated information, SVR-PLS appeared to do this fairly well. There was still an issue regarding collinear variables however.

Consider two highly correlated variables, variable 1 contains information that is very useful to the model, variable 2 contains information that is slightly better than variable 1. If variable 1 is presented to the model first (random chance) then it will be selected. If variable 2 is subsequently presented, it will also be selected as a produces a slight improvement into the model. Thus there are now two collinear variables in the model, which is supposed to be produced without any collinear variables. The solution to this is to repeat the selection procedure on the variables that have already been selected, shuffling them again randomly. This reduces the chance that variables 1 & 2 will again be presented in the same order, thereby eliminating variable 1 from the model. This situation could occur with many collinear variables in spectral data, so several redundant variables could be selected. This method was referred to as single variable removal dual pass PLS (SVR-DP-PLS). As this algorithm was being developed a second method to deal with the selection of collinear variables was considered, that of a squashing function (mathematically a cost function). This would allow the addition of a variable to proceed only with a significantly smaller predictive error rather than a mathematically smaller error. The selection of an appropriate squashing function requires considerable thought, but was considered to be an overall

improvement in the model since it allows a large decrease in the number of selected variables at only a small penalty in increased predictive error. The squashing function was applied to both removal stages. SVR-DP-PLS showed significant performance improvements over the previous methods examined, producing improvements in both the number of variables selected and reducing the predictive error in the model. As the final algorithm was being developed some of the information generated during the procedures was also considered. While many iterations will be run, only one will be selected as producing the best result, however this does not mean that the other iterations do not have any information to provide. By recording the variables that are selected during each iteration a history can be built up of how frequently a variable has been selected and its position in the spectra. This provides information about the relative importance of particular sections of the spectra towards the model. This information was generated for each of the preceding methods, and charted. The histograms showed that with each successive generation of algorithm the location of the variable selected stabilised. Initially the frequency of variable selection showed a highly random pattern, however by the final method (SVR-DP-PLS), the histograms were showing that the variables selected were coming from quite rigidly defined sections of the spectra. This showed that frequently common sense when applied to variable selection would give misleading results as to the best variable to select, the variable selection methods tend to select variable that provide information about overlapping areas of the spectra, allowing individual peaks to be resolved. This histogram information could be used as a weighting method of for variables in situation where variable selection may be unsuitable or unwanted.

5.2. Intrasite Gel

Intrasite Gel is a Sodium Carboxymethyl Cellulose Gel, known as Sodium Carmallose in the British Pharmacopoeia. This material starts off as a powdered cross-linked polymer and is mixed with water and propylene glycol to produce the gel. The gel is sold in several different packs, flat sachets of 10g and 20g, and appli-packs, plastic bulbs that are designed to allow the gel to be dispensed with one hand; these come in three sizes, 8g, 15g and 25g. The containers of Intrasite Gel are sealed and sterilised, following the British Pharmacopoeia guidelines in Appendix XIII for steam sterilisation. The raw polymer is bought into Smith & Nephew according to specification, and the only analysis carried out on the polymer at this stage is identification tests to determine whether the material meets the specification. Once the polymer is made into the gel it is tested for SC1 [69], and if the batch meets the specification it is packed in the appropriate containers and sterilised.

The contents of sterilised batches are randomly sampled and the containers opened and analysed. The contents are required to meet the appropriate specifications [70] or [71] depending on whether the batch is appli-packs or sachets. Smith & Nephew wanted an overall examination of the Gel, and a closer examination of the fluid absorption test [63]. The variables were examined initially with respect to the amount of variation, the distribution of the samples within each variable and the correlations between the variables. The data produced during sterilisation was also considered for its relationship with the analysis, variables and finally there was an attempt to model the fluid absorption variable from the other available data.

The initial findings were that there was a large degree of variability in the various variables, (section 4.1) and that the variables showed a binomial distribution (section 4.2). The variability and the binomial distribution were put down to the same cause, that of changes in the raw material, which is produced externally to Smith & Nephew once a year. The initial correlations between the variables (section 4.3) showed that there were fairly strong relationships between the elasticity, the viscosity coefficient and SC1, with a lesser relationship between these variables and the fluid absorption measurement. When these variables were examined on sections of the data that showed normal distribution these correlations all decreased significantly, showing that part of the correlation seen earlier was due to leverage effects from the step changes in raw material properties. It should however be considered that with an infinite number of points any relationship greater than ±26 is significant (the Intrasite Gel data set contains in excess of 3000 points, which puts the calculation for the t-stat in the range of an infinite population). This shows that there are relationships between all the variables except pH, however they are not strong enough to suggest that a model could be built predict any one of them from the others with any precision and accuracy. The lack of a relationship between the pH and any of the other variables was of little surprise, the measured variables are all physical properties except the pH. Since the pH varies between 6.4 and 7.4 only very small variations in the cross-linking are required to produce changes in the free hydrogen ion concentration in this range.

The fluid absorption variable was of special interest at this point, the test for fluid absorption was known to contain up to 40% error, due to both the solubility of the material in saline solution [73] and the error associated with the test itself [72].

Modelling of this variable was carried out to determine whether the other variables could provide information about the source of the error in the fluid absorption test initially this was done on the whole data set for the analysis variables, then the sections of the data set where the data was normally distributed, and finally on data where information about the sterilisation process was also available. These models all showed error as great as the error already known to be in the measurement of the fluid absorption suggesting that the data available did not contain any information about the variability and error in the fluid absorption test. Although the pH variation is small this was considered as a possible reason for the error; pH is a representation of ionic concentration, and ionic concentration will effect solubility of materials. A series of tests were carried out on Intrasite Gal at different pH values, the range extending considerably outside the normal range of the pH. When these results were examined (section 4.6) it was found that there was no effect of the pH value on the fluid absorption value.

The overall relationships between the variable was still of interest, and there was also concern about the sampling rate for the analysis of Intrasite Gel. Given that there was no strong relationship between the raw variables the data was examined using CUSUM charts. The CUSUMs were calculated as normal however they were then autoscaled to allow direct comparison between the different CUSUM charts with very different magnitudes. When the CUSUMs were examined (section 4.7) a surprising degree of correlation was found between the variables, showing that despite the low correlations between the individual samples, the overall process trends were related. This is likely to be due to the high noise in the raw data that masks any relationship, once the aviation from the average is considered (CUSUMs) the relationships become

more evident. The fluid absorption can be seen to follow the same trends as both the viscosity coefficient and SC1, and there is a very strong negative correlation with the pH. It can be seen also that that there is a possible interaction between the elasticity and the pH. When the Elasticity CUSUM is plotted against the summation of the pH CUSUM and the SC1 CUSUM it can be seen that the elasticity follows a very highly correlated trend. This can also be seen with the pH CUSUM added to either the fluid absorption trend or the viscosity trend. It was thought that the evident relation ship between the pH and either the fluid absorption or the viscosity coefficient was a symptom rather than a cause and the true relationship is with SC1.

The hypothesis is that although experiments into the effect of the pH on fluid absorption showed no effect under the conditions used, it is likely that this is due to insufficient time or temperature. Thus if the experiments had been carried out at either an elevated temperature (as would occur during sterilisation) or for a significantly longer period of time, a relation ship between the pH and the fluid absorption would have been seen. The pH is likely to effect the cross-linking of the polymer, thus effecting the other measured variables.

The CUSUM calculations were used in the consideration of the sampling frequency for Intrasite Gel, they earlier plots had shown that despite the apparent lack of correlation in the raw data there was a very pronounced correlation between some of the variables in the CUSUM charts. This suggests that the process to produce Intrasite is actually far more stable than the analytical evidence suggests, the stability is masked by high error in the analytical measurements. The effect of reducing sampling on the process monitoring was investigated by plotting CUSUMs calculated

from different sampling rates. Rates of every point, every second point, every fifth, every tenth and every twentieth point were considered. When the charts were scaled appropriately, it was immediately apparent that the process trend could be seen to be identical in all the different sampling rates. The overall process trends were clearly visible in all the charts. This suggests that the current high sampling rate may give misleading information about the stability of the process due to high noise in the measurement. A reduced sampling rate together with process monitoring with CUSUMs could give much greater confidence in the performance of the process than examination of the individual measurements.

The fluid absorption test by the settling volume method [38] was still of interest, there was some doubt that the test was giving a true measure of the fluid absorption of the material. The other data available did not provide the information needed to determine the reasons for the high error in the test, so another approach was needed. The method considered was replacing the fluid absorption tests with another method to produce a reference data set, this reference data set could then be modelled to allow the prediction of the new test results from the other variables. Various methods were considered, and finally a fluid transfer test was selected as the most appropriate, this test went under the name "The Paddington Cup" method, for historical reasons. This method measured the fluid transferred from a hydrogel to a substrate of either gelatine or agar, of varying concentrations. As originally designed the test was used to measure the difference in fluid transfer properties between many different types of hydrogel wound dressings. The various substrates were required to differentiate between products that could have widely varying properties, as in this case the only material of interest was Intrasite Gel the test was used with only a single substrate.

The appropriate substrate was selected by experimentation to be the most suitable to characterise Intrasite Gel, which was found to be 2% agar, for the magnitude of the fluid transfer that occurred, and the structural stability of the agar. A series of experiments were carried out to measure the fluid transfer rates of a number of different batches of Intrasite Gel (section 4.11) and the results examined to determine whether the tests showed a significant difference between the batches. The results showed that the tests did show a significant difference between the batches, and this data was then modelled. Although the model was not particularly good it was significantly better than the models built using the old fluid absorption test. It was also considered that the test was moderately difficult to carry out. Experimental error in the physical measurement could account for a significant amount of the error, and better experimental techniques, with a more rigorously controlled environment might reduce this. Overall this method appeared to avoid many of the drawbacks of the settling volume test, but at the expense of greater testing time, and a more difficult experimental procedure. It was proposed that the settling volume test be replaced with the paddington cup method, that a suitably sized data set be generated and that the results of this test be predicted from the other variables rather than measured.

5.3. Future Work

5.3.1. Variable Selection

There are several areas from this work that need further investigation. Possibly the most interesting are the histograms generated during training iterations. The histograms show that the conventional wisdom that the peaks of a spectra are the most important may be misleading in many cases, and that the information found in the

overlap areas may be more useful. The histograms developed could be examined further to look into the possibility of using them as weighting criteria for use with spectral analysis where variable selection is not appropriate, they may also be of use in initialising weights for neural network training.

The squashing (cost) function used in SVR-DP-PLS also needs further investigation, selecting the correct quashing functions is a task that requires many attempts at optimisation for each data set and problem, some form of experimental design may be useful to examine the best values for these functions.

Variable selection has been show to be useful for both MLR and PLS, there is reason to believe that this may be true for other methods as well, the most likely candidate immediately is ridge regression. Ridge regression is a very useful technique that has bee shown to outperform both MLR and PLS [38], and a comparison with variable selection methods would be useful. Ridge regression is very time consuming to carry out on large data sets, some form of variable selection may not only improve the predictive results but also reduce the time required to carry out the calculations. There are other methods to look at, orthogonal signal correction, OSC [45], is one example and although this method looks like it has strong advantages, an investigation into the benefits of variable selection may be worthwhile.

5.3.2. Intrasite Gel

Intrasite Gel still has not been investigated fully, of critical interest is the relationship between the pH and the physical properties, an investigation is needed into the

possible effect of pH over long periods and at elevated temperatures. This is of special interest as the future of Intrasite gel may include the addition of a medicament, this will add further unknowns to the equation, and any interference from the p must be understood first.

The fluid transfer test needs further work, the experimental technique need to be refined to reduce experimental error, and a suitably sized reference data set needs to be generated to allow the settling volume test to be replaced.

References

1. Walmsley, A.D, *Improved variable selection procedure for multivariate linear regression*, Analytica Chimica Acta 354, 1997, 225-232
2. Moffatt, J.R., Walmsley, A.D., *Enhancements to PLS Using Prediction Based Variable Selection*. Submitted J. Chemometrics and Intelligent Laboratory Systems, Oct 1999
3. Massart, D.L., Vandeginste, B.G.M., Deming, S.N.Michotte, Y., Kaufman, L., *Chemomterics: A Textbook*, Elsevier, Amsterdam 1988
4. Malinowski, E.R., *Factor Analysis in Chemistry*, 2nd Edition, Wiley, 1992
5. Wold, S.,Lindberg, J.A.Persson, *Partial Least Squares Method for Spectrofluorimetric analysis of Mixtures of Humic Acid and Ligninsulfonate*, Anal. Chem., **55** (1983)
6. Hotelling, H., *Analysis of a complex statistical variable into principal components*, J. Educ. Psych, (1933) **26**, 417-441, 498-520
7. Hotelling, H., *The most predictable criterion*, J. Educ. Psych, **26**, 139-142.
8. Hotelling, H., *Simplified calculation of principal components*, Psychometrika, (1936), **1**, 27-35,
9. Hotelling, H., *Relations between two sets of variates*, Biometrika, 28, (1936), **28**, 321-377
10. Bartlett, M. S., *Tests of significance in factor analysis*, Brit. J. Psych. (1950) **3**, 77-85
11. Bartlett, M. S., *The effect of stamdardisation of a* χ^2 *approximation in factor analysis*, Biometrika, (1951) **38**, 337-344
12. Thurstone, L.L., Thurstone, G.T., *Factional studies of intelligence*, Chicago University Press, (1941)
13. Pearson, K., *On lines and planes of closest fit to systems of points in space*, Phil. Mag., **2** (Sixth Series) (1901), 559-572
14. Fisher, R., Mackenzie., *Studies in crop variation II. The manurial response of different potato varieties*, J. Agricultural Science, (1923), **13**, 311-320
15. Harmon , H.H., *Modern Factor Analysis*, University of Chicago Press, (1967)
16. Rao, C.R., *The use and interpretation of principal component analysis in applied research*, Sanhkya, Series A, (1964) **26**
17. Wold, H., *Nonlinear estimation by iterative least squares procedures*,(1966) (F.David – Editor) Research Papers in Statistics, Wiley, New York, (1966), 411-444
18. Wold, S., Jonsson, J., Eriksson, L., Hellberg, S., *A multivariate approach to saccharide quantitative structure-activity relationships exemplified by two series of 9-hydroxyellipticine glycosides*, Acta Chemica Scandinavica (1989) 286-289
19. Geladi, P., Kowalski, B.R., *Partial Least Squares Regression (PLS), A tutorial*, Analytica Chimica Acta, (1986)
20. Wold, S., Rõnnar, S., Lindgren, F., Geladi, P., *A PLS kernel algorithm for data sets with many variables and fewer objects. Part I: theory and algorithm*, Journal of Chemometrics (1994), 111-125
21. Burnham, A. J., Viveros, V., *Frameworks for Latent Variable Regression*, J. Chemometrics, (1996), 10, 31-45
22. Martinalverex, P.J., Herraiz ,T., Casal, V., *Comparative Prediction of the Retention Behaviours of Small Peptides in Several Reversed Phase High Performance E.T.C*, Analytica Chimica Acta, 326, (1996) **1-3**, 77-84

23. Lipp, M., Comparison of PLS, PCR and MLR for the quantitative determination of foreign oils and fats in butter fats, .e.t.c., Z Lebensm Unters Forsh (1996), 202, 193-198
24. Wu, W., Rutan, S.C., Baldovin, A., Massart, D., Analytica Chimica. Acta 335, (1996) 11
25. Sorenson, H.W., *IEEE Spectrum* July 1996
26. Mahalanobis, P.C., *On the generalised distance in statistics*, Proc. Nat. Inst. Sci. India, **12**, (1936), 49-55
27. Wold, S., Sjostrom, *SIMCA: A method for analysing chemical data in terms of similarity and analogy*, (B.R.Kowalski – Editor) Chemomitrics: Theory and Applications. ACS Symp. Series 52, Analytical Chemical Society, Washington DC, 1997
28. Wold, S., Esbensen, K., Geladi, P., *Principal Component Analysis*, Chemometrics and Intelligent Laboratory Systems, **2**, (1987), 37-52
29. Draper, N., Smith, H., *Applied regression analysis*, J Wiley & Sons, New York, 2nd edition (1981)
30. Miller, J.C., Miller, J.N., *Statistics for Analytical Chemistry*, 3rd Editon, Ellis Horwood (1993)
31. Kaiser, H.F., *The varimax criterion for analytical rotation in factor analysis*, (1960) Psychometrika, 23(3), 187-200
32. Cattel, R.B., Muerle, J.L., *The "Maxplane" program for factor rotation to oblique simple structure*, (1960) Educ. Psychol. Measurement, 20(3) 569-590
33. Heise, H.M., Bittnre, A., *Rapid and reliable spectral variable selection for statistical calibrations based on the PLS-regression vector choices*, Fresenius Journal of Analytical Chemistry, (1997), 93-99
34. Lindgren, F., Geladi, P., Ronnar, S., Wold, S., *Interactive variable selection (IVS) for PS. Part 1: Theory and algorithms*, Journal of Chemometrics, (1994), 349-363
35. BS ISO 11095: Linear calibration using reference materials (1996)
36. Massart, D.L., Vandeginste, B.G.M., Buydens, L.M.C., De Jong, S., Lewi, P.J., Smeyers-Verbeke, J., *Handbook of Chemometrics and Qualimetrics: Part A*, Elsevier (1998)
37. ISO 7870 (1993) (BS 7785 1993), Control Charts: General Guide
38. I.E.Frank and J.H.Friedman. *A Statistical View of some Chemometrics Regression Tools. With discussion,* Technometrics **35** (1993) 109-148
39. Sanchez, F.C., Rutan, S.C., Garcia, M.D., Massart, D.L., *Resolution of multicomponent overlapped peaks by the orthogonal projection approach, evolving factor analysis and window factor analysis*, Chemometrics and Intelligent Laboratory Systems, **36** (1997) 153-164
40. de Noord, O.E., *Multivariate calibration standardization*, Chemometrics and Intelligent Laboratory Systems **25** (1994) 85-97
41. de Noord, O.E., *The influence of data preprocessing on the robustness and parsimony of multivariate calibration models,* Chemometrics and Intelligent Laboratory Systems **23** (1994) 64-70
42. Kowalski, B.R., Seasholtz, MB., *Recent developments in multivariate calibration*, Journal of Chemometrics, **5**, 129-145 (1991)
43. Wold, S., Chemometrics; *What do we mean with it, and what do we want from it?* Journal of Chemometrics and Int. Lab. Systems, (1995), **30**, No.1, 109-115
44. Wold, S., Sjostrom, M., *Chemometrics, present and future success* , Journal of Chemometrics and Int. Lab. Systems, 1998, **44**, No.1-2, 3-14

45. Wold, S., Antti, H., Lindgren, F., Ohman, J., *Orthogonal signal correction of near-infrared spectra*, J.Chem & Int. Lab. Stystems. (1998), **44**, No.1-2, 175-185
46. Sjoblom, J., Svensson, O., Josefson, M., Kullberg, H., Wold, S., *An evaluation of orthogonal signal correction applied to calibration transfer of near infrared spectra,* Journal of Chemometrics and Int. Lab. Systems, (1998), **44**, No.1-2, 229-244
47. Harshman, R.A., *Foundations of the PARAFAC procedure: Model and conditions for an 'explanatory' multi-factor analysis*, UCLA Working papers in Phonetics, **16**, (1970)
48. Harshman, R.A., *Determination and proof of minimum uniqueness conditions for PARAFAC1*, UCLA Working papers in Phonetics, **22**, (1972)
49. Harshman, R.A., Berenbaum, S.A., *Basic concepts underlying the PARAFAC-CANDECOMP three way factor analysis model and its applications to longitudinal data*, Academic press, NY, 1981, 435-459
50. Bro, R., *PARAFAC, Tutorial and applications*, Chemometrics and Intelligent Laboratory Systems, **38**, (1997), 149-171
51. de Juan, A., Rutan, S.C., Tauler, R., Massart, D.L., *Comparison between the direct trilinear decomposition and the multivariate curve resolution-alternating least squares methods for the resolution of three-way data sets,* Chemometrics and Intelligent Laboratory Systems, (1998), 19-32
52. Smilde, A.K., Doornbos, D.A., *Three-way methods for the calibration of chromatographic systems: comparing PARAFAC and three-way PLS*, Journal of Chemometrics (1991), 345-360
53. Bro, R., de Jong, S., *A fast non-negativity-constrained least squares algorithm*, Journal of Chemometrics (1997), 393-401
54. de Jong, S., Braak, C.J.F., *Comments on the PLS kernel algorithm*, Journal of Chemometrics, (1994), 169-174
55. de Jong, S., Phatak, A., *Partial least squares regression*, SIAM, (1997), 25-36
56. de Jong, S., *SIMPLS: an alternative approach to partial least squares regression,* Chemometrics and Intelligent Laboratory Systems, (1993), 251-263
57. Jouan-Rimbaud, D., Massart, D.L., de Noord, O.E., *Random correlation in variable selection for multivariate calibration with a genetic algorithm* ,Chemometrics and Intelligent Laboratory Systems, (1996), 213-220
58. Adams, M.J., Allen, J.R., *Variable selection and multivariate calibration journal of models for X-ray fluorescence spectrometry,* Journal Of Analytical Atomic Spectrometry (1998)
59. Centner, V., Massart, D.L., de Noord. O.E., de Jong, S., Vandeginste, B.M., Sterna, C., *Elimination of uninformative variables for multivariate calibration* , Analytical Chemistry, (1996), 3851-3858
60. Heise HM;Bittner A, *Rapid and reliable spectral variable selection for statistical calibrations based on PLS-regression vector choices,* Fresenius Journal Of Analytical Chemistry, (1997), 93-99
61. Kubinyi, H., *Evolutionary variable selection in regression and PLS analyses,* Journal of Chemometrics (1996), 119-133
62. Bangalore, A.S., Shaffer, R.E., Small, G.W., Arnold, M.A., *Genetic algorithm-based method for selecting wavelengths and model size for use with partial least-squares regression: Application to near-infrared spectroscopy*, Analytical Chemistry, (1996) 4200-4212
63. SOP/QGM/028, Determination of the fluid absorption of Intrasite Gel, Smith & Nephew Hull, Internal Document

64. SOP/QGM/029, Identification of propylene glycol, Smith & Nephew Hull, Internal Document
65. SOP/QGM/135, Identification of carboxymethyl cellulose, Smith & Nephew Hull, Internal Document
66. SOP/QGM/01, pH determination of Intrasite Gel, Smith & Nephew Hull, Internal Document
67. SOP/QGM/038, Elasticity measurement of Intrasite Gel, Smith & Nephew Hull, Internal Document
68. SOP/QGM/039, Viscosity coefficient measurement of Intrasite Gel, Smith & Nephew Hull, Internal Document
69. SOP/QGM/136, SC1 measurement, Smith & Nephew Hull, Internal Document
70. A155, *Specification for apli-packs of Intrasite Gel*, Smith & Nephew Hull, Internal Document
71. A156, *Specification for sachets of Intrasite Gel*, Smith & Nephew Hull, Internal Document
72. QA3174 *Quality Assurance Report*, W. Mortimer, Smith & Nephew Hull, Internal Document
73. QGM\137 *Validation of 2 test methods to determine the percentage of soluable matter in Akucell X181 polymer*, W.Mortimer, Smith & Nephew Hull, Internal Document
74. Dolz, M., Roldan, C., Herraez, J.V., Belda, R., Sobrino, P., *Rheological Behaviour of Microcrystaline Cellulose Hydrogels*, Journal of Dispersion Science and Technology, 13(1), 95-113, (1992)
75. Mamdouh, T., Ghannam, M., Esmail, N., Rheological Properties of Carboxymethyl Cellulose, J. Applied Polymer Science, 1997, **64**, pp 289-301
76. SR/TW015/MS91-2, *Development of a method to Demonstrate that Intrasite Gel has the ability to absorb or release water,* Smith & Nephew Hull, Internal Document
77. Wise, B., Gallagher, N.B., *PLS_Toolbox 2.1 for use with MATLAB* ®, Eigenvector Research, 1998

Appendices

Appendix I

A. F0 Test

Appendix XIII of the 1998 British Pharmacopoeia deals with standards for sterilisation, two methods are recognised as first choice methods, steam sterilisation and gamma irradiation. These are the preferred method when they can be carried out on the sealed product (terminal sterilisation). Sterilisation can also be carried out using ethylene oxide, but this is only suggested when the other two methods are not suitable. When steam sterilisation is carried out a standard method is required to determine the level of sterilisation, this is monitored using the F0 value.

The F_0 value indicates the lethality of a process express as minutes at a temperature of 121°C, delivered by a process to a product in its final container.

The total F_0 figure takes into account the heating up and cooling down tha occurs during the process.

$$F_0 = D_{121}(\log N_0 - \log N) = D_{121} \log IF$$

where $D_{121} = D$ value of the reference spoors at 121°C

N_0 = initial number of viable micro-organisms

N = final number of viable micro-organisms

IF = inactivation factor

$$IF = N_0 - N = 10^{t/D}$$

$D = D$ value of micro-organism in exposure conditions

B. Structure of Carboxymethyl Cellulose

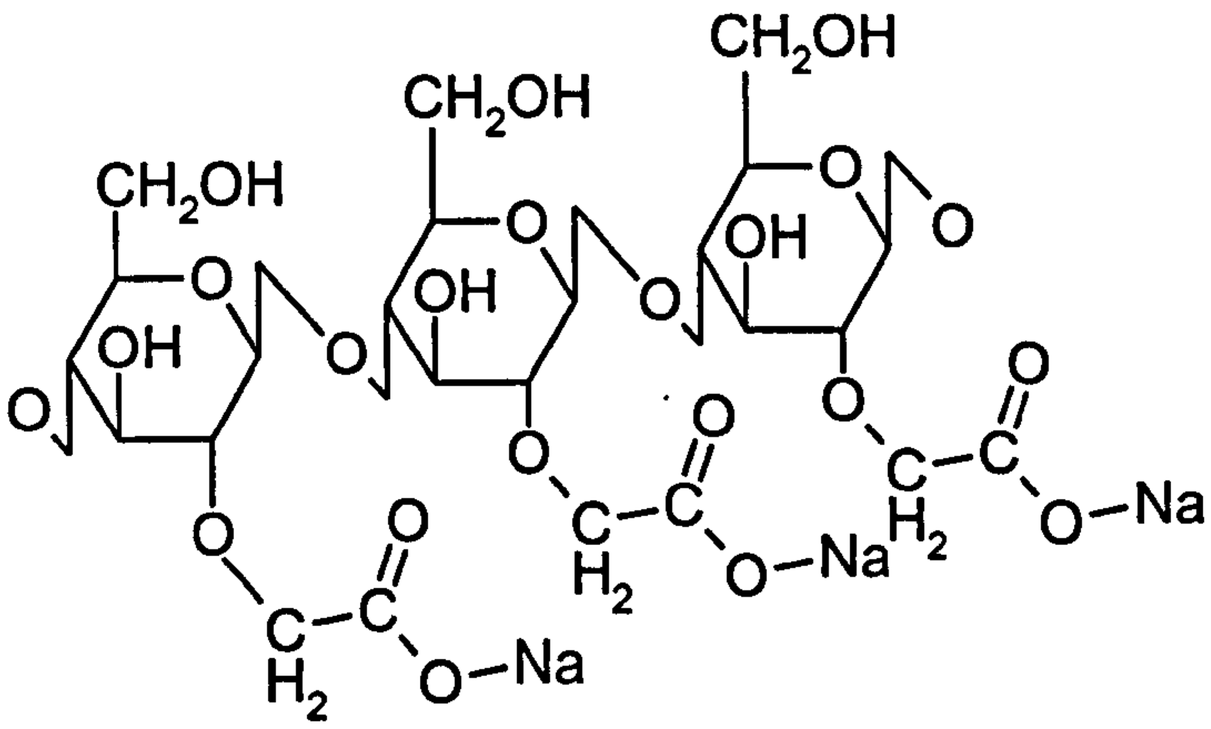


Figure I.1 Structure of Carboxymethyl Cellulose polymer repeat unit

Appendix II, Matlab Code for the Final VS-PLS Method

The following code is the final algorithm used in the VS-PLS method, this code will run under Matlab 5.2 using the PLS Toolbox 1, providing that CCCV.m is also available. This code will not run under Matlab 4.2 without modification as the use of the "Find" function changed between these two versions of Matlab. Although it is believed that this code will run with PLS Toolbox 2 this has not been validated.

```
function [ypred, press1, press2, selected, p, q, w, t, u,
b, ssqdif, mainselected, bestpress] = rempls(t_spect,
t_con, v_spect, v_con, iterations, squash1, squash2, lvs)

% This PLS function removes variables from the input
training matrix in
% order to improve the fit of the validation model, based
on PRESS.
% I/O [ypred, press1, press2, selected, p, q, w, t, u, b,
ssqdif,  mainselected,  bestpress]  =  rempls(t_spectra,
t_conc, v_spectra, v_conc, iterations, squash1, squash2,
lvs);
% Copyright 07/07/98 J.R.Moffatt
% v2.5
%
%
% ypred = prediction results for the validation data set
% press1 = the PRESS value for ordinary PLS
% press2 = lowest PRESS produced during modelling
% selected = variables used to produce the best model
% p,q,w,t,u = matrices used in the calculation of PLS
% ssqdif = information about model error
% mainselected = history information about variables
selected for all the iterations
% bestpress = history of PRESS values produced during
each iteration
%t_spectra, v_spectra, t_conc, v_conc, data sets for
modelling
% iterations = number of iterations to carry out
% squash1, squash2, values used for the squashing
functions
```

```
% lvs = number of latent vectors to use in the PLS
calculations

format long e

%-----------Check Inputs-------------

arguments = nargin;

if arguments < 7
   lvs=size(t_spect,2);
end
```

```
if arguments < 6
   squash=1;
end


if arguments < 5
   iterations = ceil(sqrt(size(t_spect,2)));
end

if arguments < 4
   [t_spect,  t_con,  v_spect,  v_con]  =  cccv(t_spect,
t_con);
   disp('Cross Validation used in model building');
end
```

```
cols = size(t_spect,2);


%-----------End Check Inputs-------------

%-----------Full PLS---------------------

[p,q,w,t,u,b,ssqdif] = pls(t_spect,t_con,lvs);
ypred = plspred(v_spect,b,p,q,w,lvs);

residuals = v_con - ypred;
residuals = residuals .* residuals;
basepress = sum(sum(residuals));
press1= sum(sum(residuals));
%-----------End Full PLS---------------------

%-----------Start Main Loop--------------------
mainselected=[];
for mainloop = 1:iterations

%-----------Reset Matrices-----------------

   cols    = size(t_spect,2);
```

```
order      = randperm(cols);
[Y,resort]   = sort(order);
t_spectra = t_spect(:,[order]);
v_spectra = v_spect(:,[order]);
t_conc     = t_con;
v_conc     = v_con;
t_pls      = [];
v_pls      = [];
t_var      = [];
v_var      = [];
selected   = [];
newlastpress = basepress;
currentlvs   = lvs;

%-----------End Reset Matrices-------------

   txt = sprintf('Now working on iteration %d',mainloop);
   disp(txt);

%------------Variable Removal Loop------------

   for loop = 1:cols-1

      t_spectra2 = [t_spectra(:,1:(cols-loop)) t_var];
      v_spectra2 = [v_spectra(:,1:(cols-loop)) v_var];

         if size(t_spectra2,2) < currentlvs
            currentlvs = size(t_spectra2,2);
            txt   =   sprintf('Number   of   LVs   is   now   %d
',currentlvs);
            disp(txt);
         end

      [p,q,w,t,u,b,ssqdif]                                    =
pls(t_spectra2,t_conc,currentlvs,1);
      ypred = plspred(v_spectra2,b,p,q,w,currentlvs);

      residuals = v_conc - ypred;
      residuals = residuals .* residuals;
      newpress(loop) = sum(sum(residuals));

         if newpress(loop) > newlastpress/squash1
            t_var = [t_spectra(:,(cols-loop+1)) t_var];
            v_var = [v_spectra(:,(cols-loop+1)) v_var];
            selected(cols-loop+1) = 1;
         else
            selected(cols-loop+1) = 0;
            newlastpress=newpress(loop);
         end
disp(loop)
disp(size(t_var,2))
```

```
disp(newpress(loop))

end
subplot(2,1,1)
plot(newpress)
drawnow
txt = sprintf('Iteration %d of %d',mainloop,iterations)
title(txt)
drawnow
   mainselected(mainloop,:)=selected(resort);
    bestpress(mainloop) = newlastpress;

%-----------End Variable Removal Loop------------

%----------Repeat Variable Removal-------------


   [Y I] = find(mainselected(mainloop,:));
    t_var=[];
   v_var=[];
   t_spectra2=[];
   v_spectra2=[];
    cols=size(I,2);
   neworder=randperm(cols);
   [Y resort]=sort(neworder);
    I2=I(neworder);
    t_spectra2=t_spect(:,I2);
   v_spectra2=v_spect(:,I2);
   newpress=[];
   selected4=[];

   for loop = 1:cols-1

      t_spectra3 = [t_spectra2(:,1:(cols-loop)) t_var];
      v_spectra3 = [v_spectra2(:,1:(cols-loop)) v_var];

         if size(t_spectra3,2) < currentlvs
            currentlvs = size(t_spectra3,2);
            txt   =   sprintf('Number   of   LVs   is   now   %d
',currentlvs);
            disp(txt);
         end

      [p,q,w,t,u,b,ssqdif]                                    =
pls(t_spectra3,t_conc,currentlvs,1);
      ypred = plspred(v_spectra3,b,p,q,w,currentlvs);

      residuals = v_conc - ypred;
      residuals = residuals .* residuals;
      newpress(loop) = sum(sum(residuals));
```

```
      if newpress(loop) > newlastpress/squash2
         t_var = [t_spectra2(:,(cols-loop+1)) t_var];
         v_var = [v_spectra2(:,(cols-loop+1)) v_var];

         selected4(cols-loop+1) = 1;
      else

         selected4(cols-loop+1) = 0;
         newlastpress=newpress(loop);
      end

   end
   disp(loop)
   disp(size(t_var,2))
   disp(newpress(loop))
   subplot(2,1,2)
   plot(newpress)
   drawnow

   %put selected4 into mainselected(mainloop)
   x=[];
   y=[];
   x=size(t_spect,2)-size(I2,2);
   x=zeros(1,x);
   y=(I2.*selected4);
   x=[y x];
   mainselected(mainloop,:)=x;
      bestpress(mainloop) = newlastpress;
   end

   %------------End Main Loop----------------------

   %------------Find Best Run----------------------

   [lowestpress,indexlowestpress] = min(bestpress);
   selected = mainselected(indexlowestpress,:);
   [Y I selected] = find(selected);
   numberselected=size(selected);

   if numberselected < lvs
      currentlvs=numberselected;
   else
      currentlvs=lvs;
   end

   [mx,nx]=size(mainselected);
   location=zeros(mx,nx);
   for x=1:mx;
      [i,j,k]=find(mainselected(x,:));
      location(x,k)=1;
```

```
end
bar(sum(location))
title('Postion of Most Frerquently Selected Variables')
drawnow


[p,q,w,t,u,b,ssqdif]                                    =
pls(t_spect(:,selected),t_con,lvs,1);
ypred = plspred(v_spect(:,selected),b,p,q,w,lvs);
figure
plot(v_con,ypred,'+');
title('Predicted vs Actual for VS-PLS Model')
dp;

residuals = v_con - ypred;
residuals = residuals .* residuals;
press2 = sum(sum(residuals));

txt = sprintf('Minimum Press %f in run %d using %d
variables',press2,indexlowestpress,size(selected,2));
disp(txt);
```

Appendix III

Graphs Showing the analysis variables for Intrasite Gel Laboratory Tests

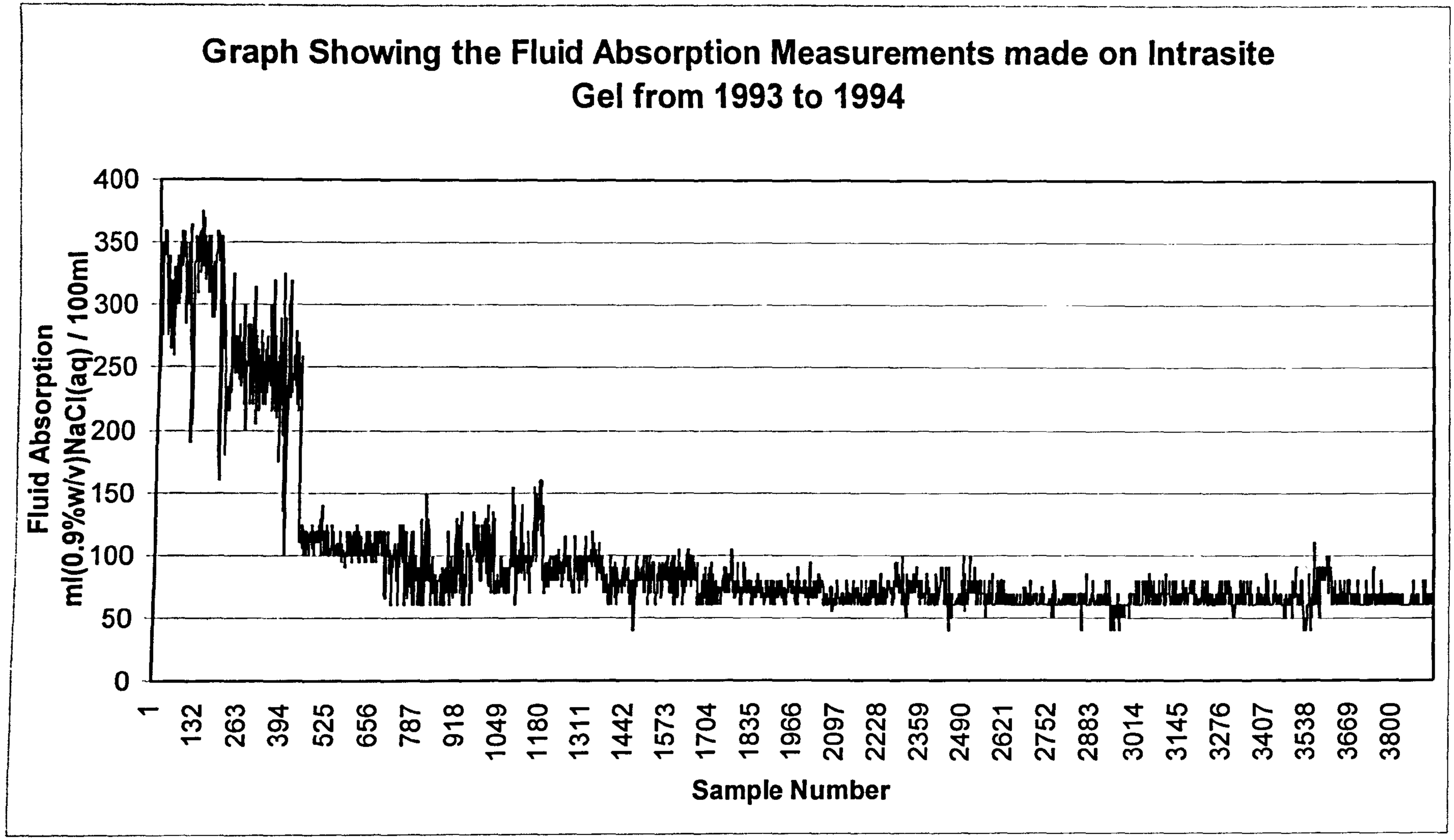


Figure III.1 Graph showing the Fluid Absorption measurements made on Intrasite Gel for the period of January 1993 through to December 1997

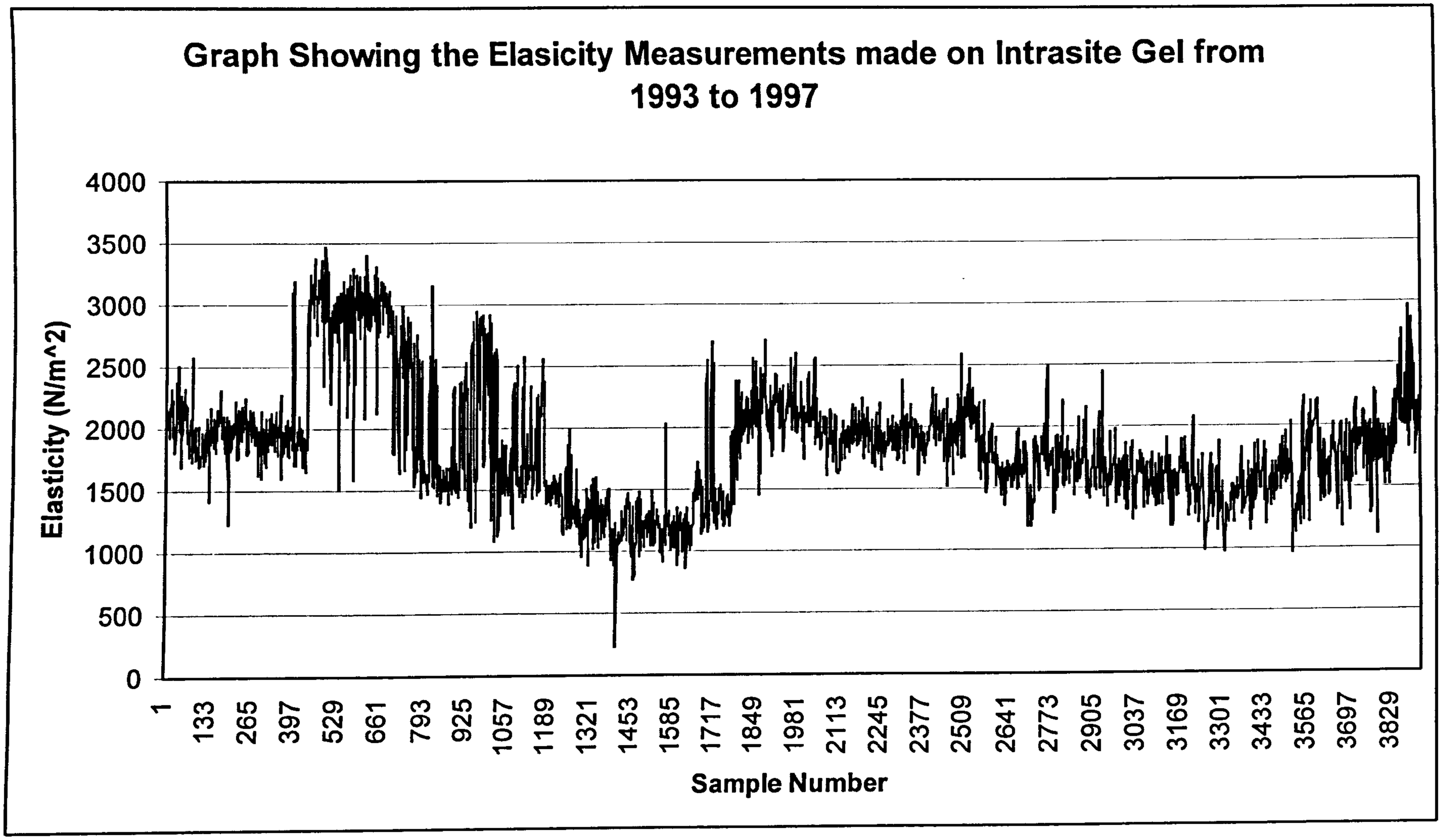


Figure III.2 Graph showing the Elasticity measurements made on Intrasite Gel for the period of January 1993 through to December 1997

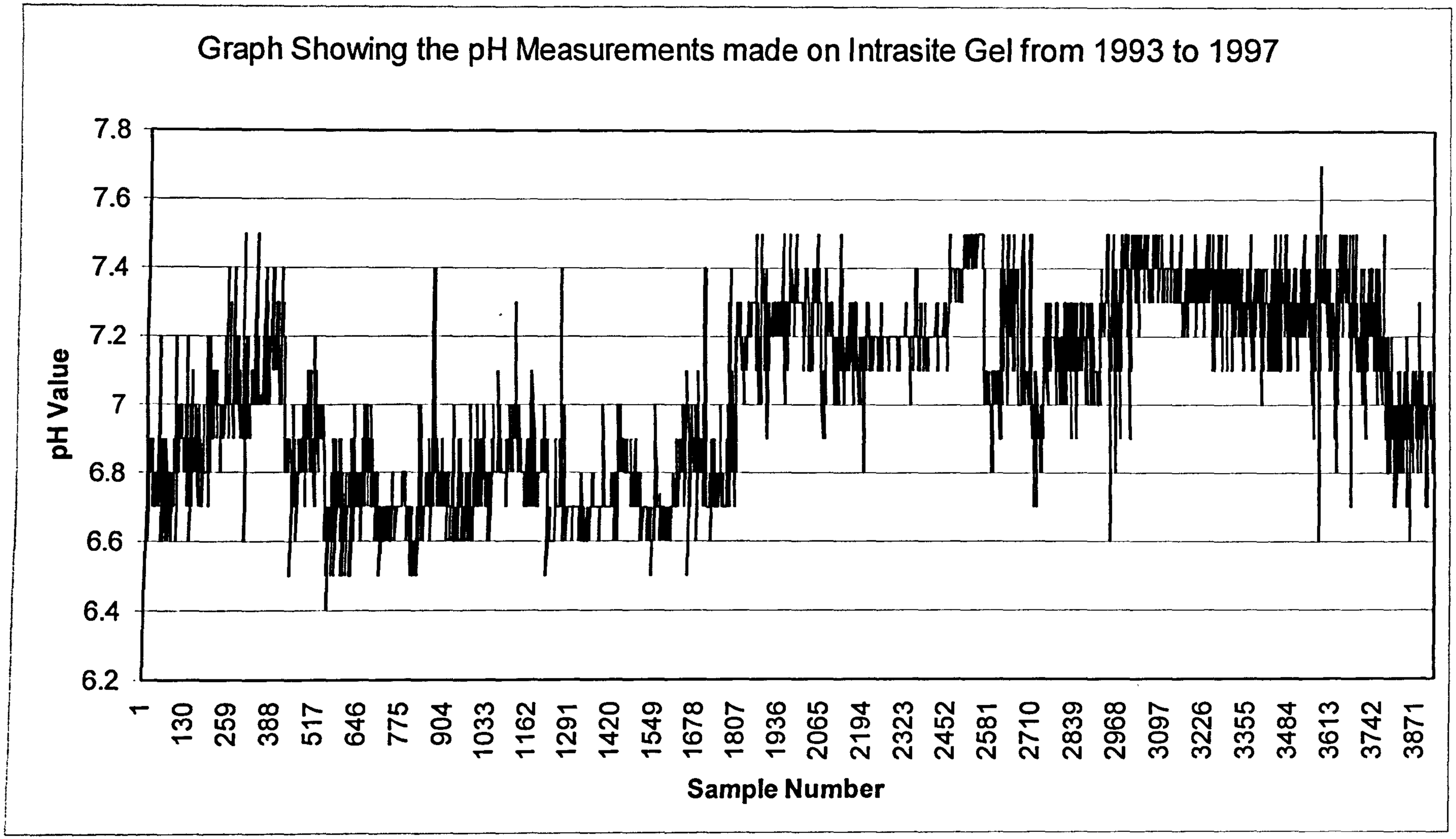


Figure III.3 Graph showing the pH measurements made on Intrasite Gel for the period of January 1993 through to December 1997

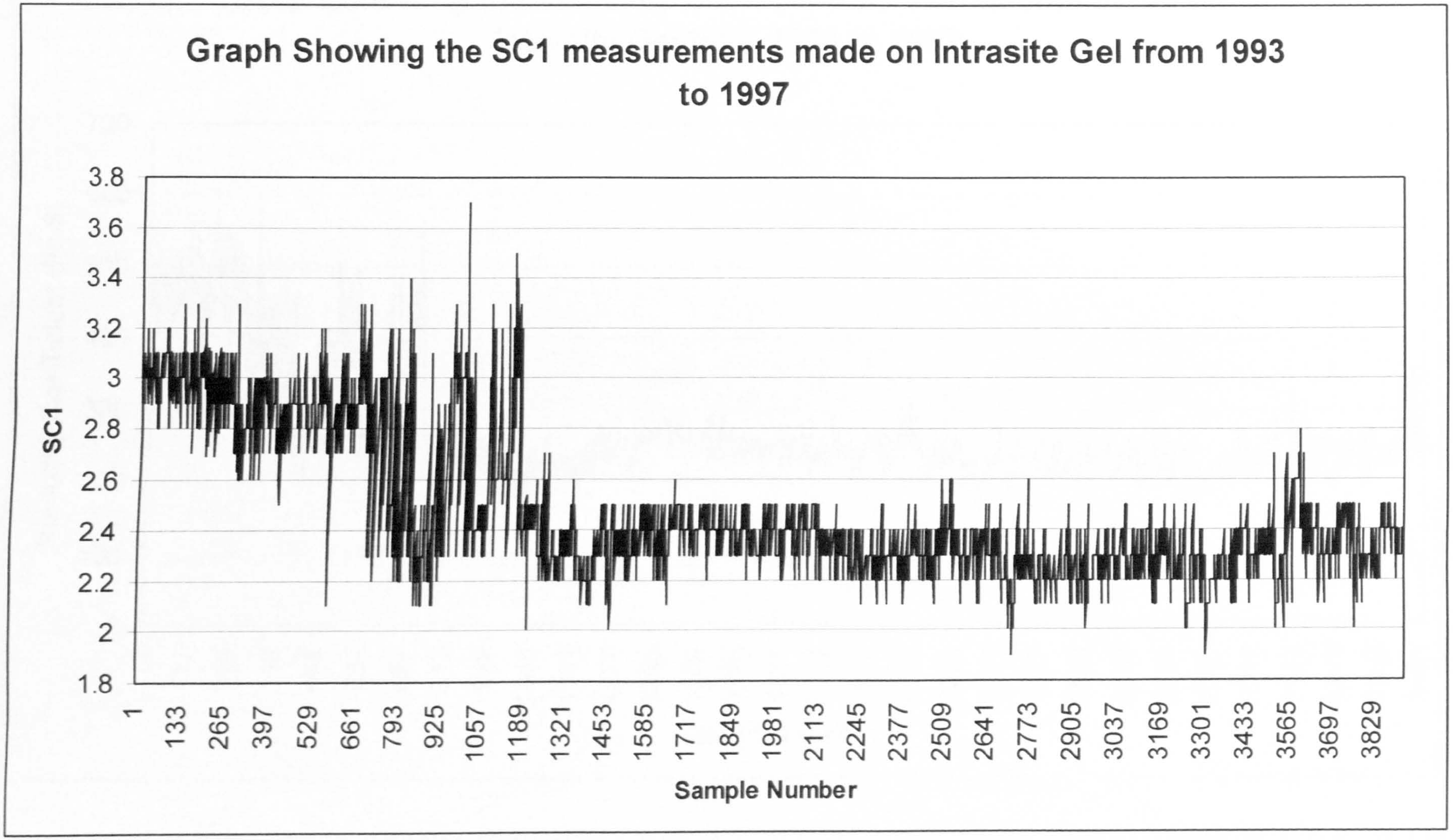


Figure III.4 Graph showing the SC1 measurements made on Intrasite Gel for the period of January 1993 through to December 1997

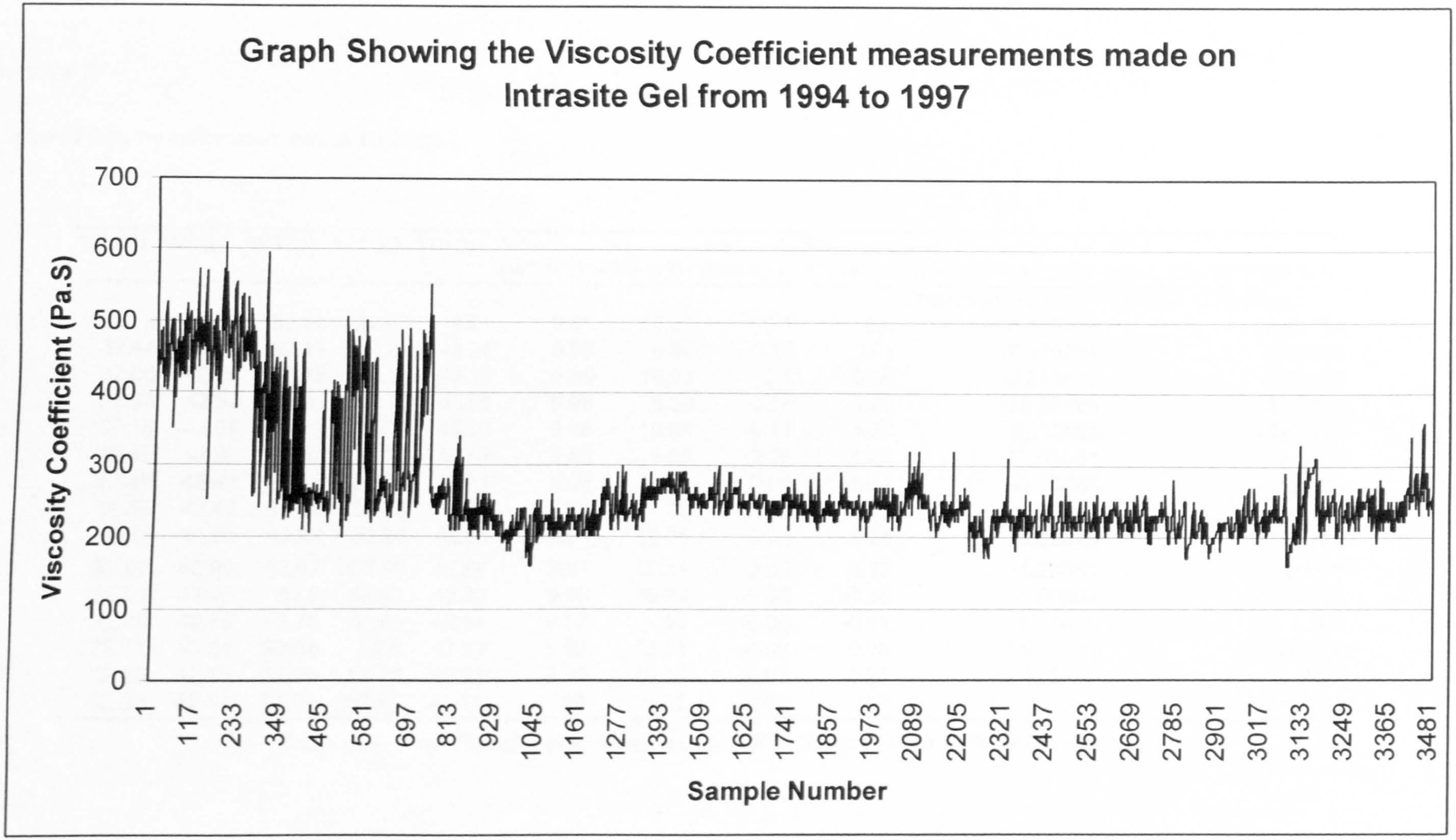


Figure III.5 Graph showing the Viscosity Coefficient measurements made on Intrasite Gel for the period of January 1993 through to December 1997

Appendix IV

Data collected from the fluid transfer tests on Hydrogels

W1 (g)	W2 (g)	W3 (g)	W4 (g)	W5 (g)	W6 (W2-W1)	W7 (W3-W2)	W8 (W4-W3)	W9 (W5-W2)	W10 (W5-W2 / W6 * 100)	W11 ((W4-W5)-W7)*100/W7)
									Change in Gelatine	Change in Hydrogel
32.69	42.5	52.57	52.47	43.1	9.81	10.07	-0.1	0.6	6.116208	-6.95134
32.44	42.42	52.41	52.29	43.26	9.98	9.99	-0.12	0.84	8.416834	-9.60961
32.36	42.25	52.28	52.18	43.11	9.89	10.03	-0.1	0.86	8.695652	-9.57129
32.88	42.86	52.85	52.76	43.95	9.98	9.99	-0.09	1.09	10.92184	-11.8118
33.18	43.06	53.12	53.01	44.33	9.88	10.06	-0.11	1.27	12.85425	-13.7177
32.47	42.35	52	51.95	43.59	9.88	9.65	-0.05	1.24	12.55061	-13.3679
32.46	42.44	52.45	52.39	42.1	9.98	10.01	-0.06	-0.34	-3.40681	2.797203
32.38	42.43	52.43	52.39	42.01	10.05	10	-0.04	-0.42	-4.1791	3.8
33.32	43.23	53.29	53.24	42.95	9.91	10.06	-0.05	-0.28	-2.82543	2.286282
32.99	42.96	53.07	52.99	42.84	9.97	10.11	-0.08	-0.12	-1.20361	0.395648
32.53	42.48	52.5	52.42	42.32	9.95	10.02	-0.08	-0.16	-1.60804	0.798403
32.88	42.75	52.75	52.69	42.64	9.87	10	-0.06	-0.11	-1.11449	0.5
32.71	42.64	52.68	52.6	42.83	9.93	10.04	-0.08	0.19	1.913394	-2.68924
32.26	42.15	52.28	52.25	42.36	9.89	10.13	-0.03	0.21	2.123357	-2.3692
32.94	42.94	53.04	52.97	43.14	10	10.1	-0.07	0.2	2	-2.67327

Table IV.1 Fluid Transfer Test Results Using 30% Gelatine With Different Hydrogel

Material	W1 (g)	W2 (g)	W3 (g)	W4 (g)	W5 (g)	W6 (g) (W2-W1)	W7 (g) (W3-W2)	W8 (g) (W4-W3)	W9 (g) (W5-W2)	W10 (%) (W5-W2 / W6 * 100)	W11 (%) ((W4-W5)-W7)*100/W7)
										Change in Agar	Change in Hydrogel
1.1	32.54	42.56	52.59	52.50	40.27	10.02	10.03	-0.09	-2.29	-22.85	21.93
1.2	33.63	43.55	53.58	53.50	40.92	9.92	10.03	-0.08	-2.63	-26.51	25.42
1.3	33.7	43.63	53.64	53.33	41.23	9.93	10.01	-0.31	-2.40	-24.17	20.88
2.1	33.48	43.32	53.39	53.33	43.10	9.84	10.07	-0.06	-0.22	-2.24	1.59
2.2	33.59	43.44	53.44	53.34	43.13	9.85	10.00	-0.10	-0.31	-3.15	2.10
2.3	33.10	42.97	53.02	52.86	42.76	9.87	10.05	-0.16	-0.21	-2.13	0.50
2.5	33.59	43.53	53.61	53.53	43.23	9.94	10.08	-0.08	-0.30	-3.02	2.18
3.1	33.11	43.09	53.15	53.08	40.30	9.98	10.06	-0.07	-2.79	-27.96	27.04
3.2	33.58	43.48	53.47	52.94	40.37	9.90	9.99	-0.53	-3.11	-31.41	25.83
3.3	33.63	43.40	53.35	53.25	40.43	9.77	9.95	-0.10	-2.97	-30.40	28.84
3.4	32.45	42.23	52.23	52.12	39.59	9.78	10.00	-0.11	-2.64	-26.99	25.30
3.5	33.49	43.45	53.53	53.44	40.49	9.96	10.08	-0.09	-2.96	-29.72	28.47
4.1	33.46	43.35	53.33	53.26	40.89	9.89	9.98	-0.07	-2.46	-24.87	23.95
4.2	33.51	43.49	53.58	53.49	40.98	9.98	10.09	-0.09	-2.51	-25.15	23.98
4.3	33.27	43.22	53.26	53.17	40.88	9.95	10.04	-0.09	-2.34	-23.52	22.41
4.4	33.65	43.52	53.54	53.38	40.99	9.87	10.02	-0.16	-2.53	-25.63	23.65
4.5	32.37	42.33	52.43	52.33	40.08	9.96	10.10	-0.10	-2.25	-22.59	21.29
5.1	32.11	42.08	52.11	51.90	42.11	9.97	10.03	-0.21	0.03	0.30	-2.39
5.2	33.60	43.53	53.60	53.44	43.54	9.93	10.07	-0.16	0.01	0.10	-1.69
5.3	32.95	42.74	52.71	52.62	42.79	9.79	9.97	-0.09	0.05	0.51	-1.40
5.4	32.87	42.97	52.94	52.39	45.90	10.10	9.97	-0.55	2.93	29.01	-34.90
5.5	33.52	43.43	53.49	53.26	43.59	9.91	10.06	-0.23	0.16	1.61	-3.88

Table IV.2 Fluid Transfer Test Results for Second Series of Experiments Using 2% Agar on Different Hydrogels

W1 (g)	W2 (g)	W3 (g)	W4 (g)	W5 (g)	W6 (W2-W1)	W7 (W3-W2)	W8 (W4-W3)	W9 (W5-W2)	W10 (W5-W2 / W6 * 100)	W11 ((W4-W5)-W7)*100/W7)	Batch Number
33.63	43.51	53.43	53.34	41.36	9.88	9.92	-0.09	-2.15	-21.76	20.77	354
33.57	43.54	53.60	53.48	41.30	9.97	10.06	-0.12	-2.24	-22.47	21.07	354
33.52	43.54	53.58	53.43	41.26	10.02	10.04	-0.15	-2.28	-22.75	21.22	355
32.62	42.53	52.48	52.39	40.34	9.91	9.95	-0.09	-2.19	-22.10	21.11	355
33.49	43.48	53.51	53.46	41.09	9.99	10.03	-0.05	-2.39	-23.92	23.33	356
32.89	42.86	52.90	52.83	41.58	9.97	10.04	-0.07	-1.28	-12.84	12.05	356
33.10	42.98	52.95	52.88	40.74	9.88	9.97	-0.07	-2.24	-22.67	21.77	361
33.61	43.51	53.56	53.45	41.33	9.90	10.05	-0.11	-2.18	-22.02	20.60	361
32.12	41.98	51.91	51.89	39.83	9.86	9.93	-0.02	-2.15	-21.81	21.45	362
32.41	42.22	52.21	52.15	39.82	9.81	9.99	-0.06	-2.40	-24.46	23.42	362
33.47	43.42	53.48	53.44	41.00	9.95	10.06	-0.04	-2.42	-24.32	23.66	363
33.16	43.06	53.07	53.03	40.08	9.90	10.01	-0.04	-2.98	-30.10	29.37	363
32.70	42.63	52.65	52.58	40.46	9.93	10.02	-0.07	-2.17	-21.85	20.96	364
32.38	42.24	52.25	52.27	40.05	9.86	10.01	0.02	-2.19	-22.21	22.08	364
32.31	42.25	52.22	52.12	40.06	9.94	9.97	-0.10	-2.19	-22.03	20.96	365
32.45	42.36	52.34	52.27	40.25	9.91	9.98	-0.07	-2.11	-21.29	20.44	365
32.29	42.16	52.22	52.15	39.96	9.87	10.06	-0.07	-2.20	-22.29	21.17	366
33.67	43.58	53.65	53.43	41.42	9.91	10.07	-0.22	-2.16	-21.80	19.27	366
32.98	42.85	52.91	52.91	41.08	9.87	10.06	0.00	-1.77	-17.93	17.59	373
32.30	42.31	52.34	52.34	40.69	10.01	10.03	0.00	-1.62	-16.18	16.15	373
32.68	42.60	52.64	52.64	40.84	9.92	10.04	0.00	-1.76	-17.74	17.53	373
32.91	42.99	53.08	53.08	41.37	10.08	10.09	0.00	-1.62	-16.07	16.06	374
33.59	43.37	53.36	53.36	41.74	9.78	9.99	0.00	-1.63	-16.67	16.32	374
33.68	43.59	53.62	53.62	41.81	9.91	10.03	0.00	-1.78	-17.96	17.75	374
32.47	42.38	52.39	52.39	40.73	9.91	10.01	0.00	-1.65	-16.65	16.48	375
32.68	42.71	52.77	52.77	41.02	10.03	10.06	0.00	-1.69	-16.85	16.80	375
33.09	43.01	53.08	53.08	41.35	9.92	10.07	0.00	-1.66	-16.73	16.48	375
32.43	42.32	52.36	52.36	40.65	9.89	10.04	0.00	-1.67	-16.89	16.63	412
32.42	42.21	52.17	52.17	40.65	9.79	9.96	0.00	-1.56	-15.93	15.66	412
33.61	43.47	53.49	53.49	41.75	9.86	10.02	0.00	-1.72	-17.44	17.17	412
32.37	42.28	52.31	52.31	40.62	9.91	10.03	0.00	-1.66	-16.75	16.55	413

W1 (g)	W2 (g)	W3 (g)	W4 (g)	W5 (g)	W6 (W2-W1)	W7 (W3-W2)	W8 (W4-W3)	W9 (W5-W2)	W10 (W5-W2 / W6 * 100)	W11 ((W4-W5)-W7)*100/W7)	Batch Number
32.98	43.01	53.08	53.08	41.41	10.03	10.07	0.00	-1.60	-15.95	15.89	413
33.59	44.57	54.60	54.60	42.79	10.98	10.03	0.00	-1.78	-16.21	17.75	413
32.85	42.72	52.81	52.81	41.01	9.87	10.09	0.00	-1.71	-17.33	16.95	425
32.91	42.82	52.86	52.86	41.06	9.91	10.04	0.00	-1.76	-17.76	17.53	425
33.21	43.12	53.14	53.14	41.33	9.91	10.02	0.00	-1.79	-18.06	17.86	425
32.42	42.31	52.28	52.28	40.71	9.89	9.97	0.00	-1.60	-16.18	16.05	431
32.12	42.13	52.21	52.21	40.55	10.01	10.08	0.00	-1.58	-15.78	15.67	431
33.68	43.48	53.50	53.50	41.73	9.80	10.02	0.00	-1.75	-17.86	17.47	431
33.61	43.49	53.57	53.57	41.82	9.88	10.08	0.00	-1.67	-16.90	16.57	432
32.81	42.71	52.78	52.78	41.19	9.90	10.07	0.00	-1.52	-15.35	15.09	432
33.28	43.19	53.28	53.28	41.59	9.91	10.09	0.00	-1.60	-16.15	15.86	432
32.92	42.96	52.98	52.98	41.43	10.04	10.02	0.00	-1.53	-15.24	15.27	435
33.50	43.39	53.46	53.46	41.85	9.89	10.07	0.00	-1.54	-15.57	15.29	435
33.07	42.98	53.03	53.03	41.46	9.91	10.05	0.00	-1.52	-15.34	15.12	435
32.26	42.24	52.25	52.25	40.66	9.98	10.01	0.00	-1.58	-15.83	15.78	442
33.65	43.46	53.51	53.51	41.72	9.81	10.05	0.00	-1.74	-17.74	17.31	442
33.56	43.36	53.44	53.44	41.57	9.80	10.08	0.00	-1.79	-18.27	17.76	442
33.58	43.46	53.38	53.38	41.76	9.88	9.92	0.00	-1.70	-17.21	17.14	451
33.19	43.17	53.12	53.12	41.56	9.98	9.95	0.00	-1.61	-16.13	16.18	451
33.29	43.22	53.23	53.23	41.51	9.93	10.01	0.00	-1.71	-17.22	17.08	451
33.48	43.39	53.30	53.30	41.52	9.91	9.91	0.00	-1.87	-18.87	18.87	452
33.53	43.38	53.48	53.48	41.53	9.85	10.10	0.00	-1.85	-18.78	18.32	452
32.89	42.94	53.00	53.00	41.26	10.05	10.06	0.00	-1.68	-16.72	16.70	452
32.46	42.14	52.14	52.04	40.41	9.68	10.00	-0.10	-1.73	-17.87	16.30	511
32.32	42.06	52.13	52.06	40.24	9.74	10.07	-0.07	-1.82	-18.69	17.38	511
33.62	43.43	52.40	52.31	41.56	9.81	8.97	-0.09	-1.87	-19.06	19.84	511
32.94	42.88	52.97	52.83	41.00	9.94	10.09	-0.14	-1.88	-18.91	17.24	513
32.31	42.14	52.18	52.07	40.45	9.83	10.04	-0.11	-1.69	-17.19	15.74	513
32.71	42.53	52.56	52.42	40.77	9.82	10.03	-0.14	-1.76	-17.92	16.15	513
32.92	42.74	52.68	52.59	41.13	9.82	9.94	-0.09	-1.61	-16.40	15.29	521
33.57	43.31	53.40	53.22	41.60	9.74	10.09	-0.18	-1.71	-17.56	15.16	521
33.68	43.51	53.59	53.47	41.74	9.83	10.08	-0.12	-1.77	-18.01	16.37	521

W1 (g)	W2 (g)	W3 (g)	W4 (g)	W5 (g)	W6 (W2-W1)	W7 (W3-W2)	W8 (W4-W3)	W9 (W5-W2)	W10 (W5-W2 / W6 * 100)	W11 ((W4-W5)-W7)*100/W7)	Batch Number
33.59	43.47	53.50	53.37	41.79	9.88	10.03	-0.13	-1.68	-17.00	15.45	522
33.19	43.09	53.09	53.03	41.58	9.90	10.00	-0.06	-1.51	-15.25	14.50	522
33.28	43.18	53.21	53.17	41.64	9.90	10.03	-0.04	-1.54	-15.56	14.96	522
33.60	43.44	53.41	53.32	41.82	9.84	9.97	-0.09	-1.62	-16.46	15.35	523
32.80	42.74	52.71	52.64	41.30	9.94	9.97	-0.07	-1.44	-14.49	13.74	523
33.27	43.43	53.45	53.36	41.95	10.16	10.02	-0.09	-1.48	-14.57	13.87	523
32.44	42.37	52.36	52.26	40.72	9.93	9.99	-0.10	-1.65	-16.62	15.52	531
32.69	42.43	52.45	52.34	40.71	9.74	10.02	-0.11	-1.72	-17.66	16.07	531
33.11	42.96	52.97	52.89	41.31	9.85	10.01	-0.08	-1.65	-16.75	15.68	531
32.84	42.70	52.77	52.65	41.13	9.86	10.07	-0.12	-1.57	-15.92	14.40	533
32.90	42.69	52.77	52.70	41.08	9.79	10.08	-0.07	-1.61	-16.45	15.28	533
33.19	43.03	53.03	52.93	41.50	9.84	10.00	-0.10	-1.53	-15.55	14.30	533
33.44	43.30	53.30	53.20	41.66	9.86	10.00	-0.10	-1.64	-16.63	15.40	541
33.50	43.39	53.46	53.38	41.75	9.89	10.07	-0.08	-1.64	-16.58	15.49	541
32.87	42.79	52.79	52.71	41.43	9.92	10.00	-0.08	-1.36	-13.71	12.80	541
32.94	42.92	52.88	52.81	41.37	9.98	9.96	-0.07	-1.55	-15.53	14.86	543
33.49	43.21	53.23	53.00	41.52	9.72	10.02	-0.23	-1.69	-17.39	14.57	543
33.07	42.89	52.89	52.80	41.37	9.82	10.00	-0.09	-1.52	-15.48	14.30	543
32.36	42.17	52.18	52.11	40.54	9.81	10.01	-0.07	-1.63	-16.62	15.58	546
32.98	42.91	53.00	52.92	41.24	9.93	10.09	-0.08	-1.67	-16.82	15.76	546
33.60	43.44	53.37	53.27	41.73	9.84	9.93	-0.10	-1.71	-17.38	16.21	546
32.26	42.07	52.05	51.99	40.33	9.81	9.98	-0.06	-1.74	-17.74	16.83	612
33.60	43.40	53.47	53.34	41.55	9.80	10.07	-0.13	-1.85	-18.88	17.08	612
33.25	43.06	53.05	52.97	41.32	9.81	9.99	-0.08	-1.74	-17.74	16.62	612
32.41	42.20	52.27	52.18	40.39	9.79	10.07	-0.09	-1.81	-18.49	17.08	613
32.11	41.97	51.97	51.89	40.17	9.86	10.00	-0.08	-1.80	-18.26	17.20	613
33.65	43.42	53.55	53.40	41.48	9.77	10.13	-0.15	-1.94	-19.86	17.67	613
33.62	42.87	52.92	52.86	41.21	9.25	10.05	-0.06	-1.66	-17.95	15.92	614
33.82	43.35	53.44	53.37	41.56	9.53	10.09	-0.07	-1.79	-18.78	17.05	614
33.27	43.05	53.10	53.02	41.34	9.78	10.05	-0.08	-1.71	-17.48	16.22	614
32.41	42.32	52.27	52.18	40.60	9.91	9.95	-0.09	-1.72	-17.36	16.38	615
32.31	42.15	52.21	52.12	40.48	9.84	10.06	-0.09	-1.67	-16.97	15.71	615

W1 (g)	W2 (g)	W3 (g)	W4 (g)	W5 (g)	W6 (W2-W1)	W7 (W3-W2)	W8 (W4-W3)	W9 (W5-W2)	W10 (W5-W2 / W6 * 100)	W11 ((W4-W5)-W7)*100/W7)	Batch Number
33.61	43.40	53.44	53.31	41.55	9.79	10.04	-0.13	-1.85	-18.90	17.13	615
33.59	43.52	53.53	53.50	41.63	9.93	10.01	-0.03	-1.89	-19.03	18.58	621
33.20	43.21	53.31	53.29	41.46	10.01	10.10	-0.02	-1.75	-17.48	17.13	621
33.29	43.23	53.22	53.23	41.49	9.94	9.99	0.01	-1.74	-17.51	17.52	621
32.37	42.23	52.21	52.20	40.65	9.86	9.98	-0.01	-1.58	-16.02	15.73	622
32.99	42.99	52.96	52.92	41.31	10.00	9.97	-0.04	-1.68	-16.80	16.45	622
33.60	43.54	53.49	53.45	41.71	9.94	9.95	-0.04	-1.83	-18.41	17.99	622
33.46	43.40	53.32	53.23	41.57	9.94	9.92	-0.09	-1.83	-18.41	17.54	623
33.52	43.49	53.51	53.50	41.70	9.97	10.02	-0.01	-1.79	-17.95	17.76	623
32.89	42.80	52.85	52.76	41.16	9.91	10.05	-0.09	-1.64	-16.55	15.42	623
32.84	42.77	52.78	52.74	41.07	9.93	10.01	-0.04	-1.70	-17.12	16.58	624
32.89	42.84	52.92	52.87	41.11	9.95	10.08	-0.05	-1.73	-17.39	16.67	624
33.20	43.16	53.17	53.15	41.51	9.96	10.01	-0.02	-1.65	-16.57	16.28	624
32.94	43.53	53.55	53.48	41.89	10.59	10.02	-0.07	-1.64	-15.49	15.67	633
33.48	42.66	52.67	52.60	41.15	9.18	10.01	-0.07	-1.51	-16.45	14.39	633
33.05	43.26	53.34	53.27	41.71	10.21	10.08	-0.07	-1.55	-15.18	14.68	633
32.96	42.83	52.81	52.74	41.29	9.87	9.98	-0.07	-1.54	-15.60	14.73	634
32.68	42.66	52.74	52.63	41.03	9.98	10.08	-0.11	-1.63	-16.33	15.08	634
32.42	42.34	52.30	52.21	40.77	9.92	9.96	-0.09	-1.57	-15.83	14.86	635
32.44	42.39	52.36	52.27	40.82	9.95	9.97	-0.09	-1.57	-15.78	14.84	635
33.66	43.44	53.45	53.35	41.72	9.78	10.01	-0.10	-1.72	-17.59	16.18	635
32.23	42.24	52.30	52.26	40.52	10.01	10.06	-0.04	-1.72	-17.18	16.70	641
33.62	43.50	53.42	53.37	41.65	9.88	9.92	-0.05	-1.85	-18.72	18.15	641
33.27	43.23	53.26	53.22	41.50	9.96	10.03	-0.04	-1.73	-17.37	16.85	641
32.93	42.88	52.89	52.81	41.23	9.95	10.01	-0.08	-1.65	-16.58	15.68	642
33.56	43.48	53.54	53.43	41.71	9.92	10.06	-0.11	-1.77	-17.84	16.50	642
33.70	43.57	53.63	53.51	41.75	9.87	10.06	-0.12	-1.82	-18.44	16.90	642
33.54	43.27	53.31	53.27	41.50	9.73	10.04	-0.04	-1.77	-18.19	17.23	643
33.48	43.42	53.38	53.34	41.62	9.94	9.96	-0.04	-1.80	-18.11	17.67	643
33.12	43.00	53.00	52.95	41.36	9.88	10.00	-0.05	-1.64	-16.60	15.90	643

Table IV.3 Fluid Transfer Results for the Samples from Table 4.15, Using 2% Agar

Appendix V, communication with McKelvey & Wold

Subject: PLS Code
Date: Wed, 2 Dec 1998 21:44:31 -0600
From: John McKelvey <mckelvey@NCSA.UIUC.EDU>
Reply-To: International Chemometrics Society <ICS-L@UMDD.UMD.EDU>
To: ICS-L@UMDD.UMD.EDU

Hello.. A first timer here.. so if i don't do it right please me know..

I am looking for a PLS procedure for use in fitting when the independent variables are more than a little collinear. Any suggestions would be appreciated.

Thanks!

John McKelvey
NCSA

Subject: Re: PLS Code
Date: Mon, 25 Jan 1999 12:14:34 +0000
From: James Moffatt <j.r.moffatt@chem.hull.ac.uk>
Organization: University of Hull
To: International Chemometrics Society <ICS-L@UMDD.UMD.EDU>

Hmmm,
I may get shot down in flames for this, but with collinear independent variables you really need some form of variable selection routine if you intend to use PLS, or use some method that is more robust towards rank deficient matrices, possibly a (p or b) Spline method.

One possibility that has worked quite well in the past without variable selection is to use Ridge Regression, this can give good results with this sort of data as the first step involves increasing the rank of the matrices.

Depending on the size of the data set another option is to put the collinear variables into the model as interations rather than the original variables.

I think more information about the data set you are considering might be useful to give a better answer

James Moffatt

Subject: Re: PLS Code
Date: Mon, 25 Jan 1999 08:20:39 -0600
From: John McKelvey <mckelvey@NCSA.UIUC.EDU>
Reply-To: International Chemometrics Society <ICS-L@UMDD.UMD.EDU>
To:

I had good luck in variable selection by using Ponder's QSAR code... I used his simulated annealing with his PLS.

John McKelvey
NCSA

Dear All:

To my great surprise I got the following message from our discussion group (NAmICS), where a gentleman called James Moffatt wrote:

>Hmmm,
> I may get shot down in flames for this, but with collinear independent >variables you really need some form of variable selection routine if you intend to use >PLS, or use some method that is more robust towards rank deficient matrices, >possibly a (p or b) Spline method.
>
> One possibility that has worked quite well in the past without variable >selection is to use Ridge Regression, this can give good results with this sort of data >as the first step involves increasing the rank of the matrices.
>

My question to James Moffatt: Have you ever tried PLS ?

According to the chemometrics literature and also a number of papers in the statistics literature, PLS (correctly implemented) is together with Ridge Regression the best available method to deal with collinear predictor variables in a regression situation. Don't call these X-variables "independent" since they obviously are not.

A good comparison is the Frank and Friedman paper in Technometrics:
I.E.Frank and J.H.Friedman. A Statistical View of some Chemometrics Regression Tools. With discussion. Technometrics 35 (1993) 109-148.

Read also:

A.Burnham, R.Viveros, and J.F.MacGregor
Frameworks for Latent Variable Multivariate Regression.
J.Chemometrics 10 (1996) 31-45

All the best, Yours
Svante Wold, Umea University

I guess I should reply to this,

I think I should state first that I have no interest in getting involved in a flame war about various regression methods, particularly with you Svante, since your experience and knowledge in this area greatly exceeds my own. Also I believe that any discussion about which method is always the best is meaningless since in chemometrics situation is everything.

I have read the papers you mentioned and I agree with many of their points in the context in which they are made, interestingly enough the I.E.Frank and J.H.Friedman paper clearly states that ridge regression is superior to PLS for this situation, for the reasons I stated in my original post, that of rank deficiency of the X-variables.

Yes I have tried PLS, certainly wouldnt have recommended a PLS variable selection approach without trying it first, and yes I still prefer ordinary PLS over some other methods.

Yes I agree that calling a group of collinear variables "Independent" is wrong, however in the context of the original posting (which I quoted) this is not what was said, the data set referred to was the Independent one, I assume this convention of calling the X-Block the Independent data set and the Y-Block the Dependent data set stems from the chemical engineering side of chemometrics, but that does not make it
wrong, just different. These terms of reference are used by many of the chemometric packages that come supplied with modern instruments. Pirouette, Spectracalc and Buhler's software are among those with this convention, and if memory serves, "Arthur" does as well.

As to PLS or ridge regression being the "best" I imagine this is true when you are comparing PLS to MLR or PCR, or where there are strong reasons for retaining the sections of a data set that contain relatively little information relating to the calibration, such as in process analysis where the background noise is often considered as important as the component information. However in a spectral calibration where the priority is the ability to predict the concentration of a component, using the full data set with collinear variables is not always the best choice. The same "Chemometrics literature" agrees with me on this, so I guess we must be both right. I would point out that a quick use of Rasmus Bro's web page search engine (http://www.optimax.dk/) with the arguments PLS & Variable Selection will return 25 hits with about 50% saying PLS or ridge regression is as good as variable selection, and 50% saying that variable selection is best, and if you refine your search you can make that balance come out anyway you wish.

Regards,

James Moffatt

Dear James:

With these qualifications in your answer, I get much less upset. You must forgive me, but I tend to get high blood pressure is somebody says that one needs to perform variable selection before PLS.

In various types of spectroscopy one indeed finds that sometimes variable selection before PLS (or PCR) gives better predictions, but sometimes not. It would be interesting to see whether ridge regression works better with variable selection for the same data sets.

Recently we have been looking at alternatives such as orthogonal signal correction (OSC) and that seems to reduce the need for variable selection substantially.

All the best // Sincerely // Svante

Svante Wold, Umea Univ.

Appendix VI, Enhancements to PLS Using Prediction Based Variable Selection.

James R. Moffatt and Anthony D. Walmsley*
Department of Chemistry, Faculty of Science and the Environment, University of

Hull, Cottingham Rd. HULL, HU6 7RX

*To whom correspondence should be addressed

Abstract

This paper describes a method for reducing the number of variables required to perform a spectral calibration using Projected Latent Structures (PLS). The predictive error is reduced, producing a more robust calibration. This method has been compared to ordinary PLS and Principal Component Regression (PCR) and was found to improve on both in terms of predictive ability of the resulting model.

The approach used is an iterative one, each variable is tested to examine whether its inclusion in the data set reduces the predicted error. The technique is excellent for data sets with a large number of variables, such as spectral data. More than one iteration is required to find the best error, but a consistent minimum error is obtained relatively quickly.

The procedure is computationally expensive, and so is unlikely to find uses in on-line spectral analysis, however for at-line or off-line data processing the results can be a significant improvement over the use of the full spectra.

Keywords

Chemometrics, Variable Selection, PLS, PCR, Spectroscopy, Multivariate Calibration,

Introduction

Chemometrics has been applied to spectroscopy for many years, with the recent rapid advance in computer power and the corresponding increase in spectroscopic technology data sets are becoming larger all the time. The chemometric tools commonly used in spectroscopy include linear regression (1), multivariate linear regression (1), principal component analysis (1), and projected latent structures (2).

Partial Least Squares (PLS) is a fairly old technique, it can be traced back to 1923, when R. Fisher & W. MacKenzie (3) first published an algorithm that was the precursor to the PLS normally used today. Some years later in 1966 H. Wold (4) published the paper that directly lead to PLS, this paper was later modified and improved by S. Wold in 1983 (5). PLS provides both predictive information, allowing calibration of an x-block against a y-block, and it also provides descriptive information about how the x-block data affects the y-block data. This diagnostic information is useful for fault diagnosis and error detection. One of the faults of any variable selection process is the loss of descriptive information in the x-block and that relationship with the y-block, and a consequent loss of fault detection. The routine for variable selection presented in this paper is less susceptible to this problem than many other techniques because it does not concentrate on highly correlated variables or variables at the centre of peaks as most of the other techniques tend to do. A good paper describing reasons why variable selection might not be appropriate in a particular case can be seen by S. Wold (6), and the importance of selecting the correct type of model is cover by E. Ronchetti (7).

Projected Latent Structures (PLS) is a method of decomposing an X block matrix and a Y block matrix into vectors such that the resultant vectors from the X block are highly correlated with the vectors from the Y block.

The result of this is that the coefficients of the X block variables that provide information relating to the Y block increase, while the coefficients for variable with no information tend towards zero.

NIPALS (2) relies on the mathematical fact that

$$\mathbf{D}_j = \sum \mathbf{u}_j s_j \mathbf{v}'_j$$

Where **D** is the Data matrix, **u** & **v** are vectors, and s is a scalar for all **D** where **D** is non-singular (A singular matrix has no inverse, and so cannot be used for these calculations)..

This expression can be expressed as: -

$$\mathbf{D}\mathbf{v}_1 = \mathbf{u}_1 s_1$$

Here a randomly selected vector $\mathbf{v}_1$ is selected and used to calculate s_1 & $\mathbf{u}_1$ this is an approximation of $\mathbf{u}_1$, a better approximation can then be found by recreating $\mathbf{v}_1$: -

$$\mathbf{u}_1'\mathbf{D} = s_1\mathbf{v}_1'$$

This is repeated until convergence for a value of $\mathbf{v}_1$. This allows the calculation of $\mathbf{D}_1$ the first approximation. The residual matrix is then calculated from this: -

$$\mathbf{E}_1 = \mathbf{D} - \mathbf{D}_1$$

The next eigenvector $\mathbf{v}_2$ can then be extracted from the residual matrix. In each stage of the calculation of the vectors $\mathbf{u}_j$ and $\mathbf{v}_j$ the vectors are normalised to unit length to ensure orthogonality between the vectors.

NIPALS describes the decomposition of a matrix into eigenvalues and eigenvectors however this is for one matrix and does not allow for a relationship between two matrices. NIPALS can effectively be used to carryout PCA however this can more effectively be done using SVD. NIPALS is useful in that it allows for the possibility of relationship between two matrices. If the eigenvectors are calculated simultaneously for two different matrices,

$$\mathbf{Y}\mathbf{p}_i = \mathbf{q}_i\mathbf{a}_i$$

$$\mathbf{D}\mathbf{v}_i = \mathbf{u}_i\mathbf{s}_i$$

then a relationship can be found between $\mathbf{p}_i$ & $\mathbf{v}_i$ and $\mathbf{q}_i$ & $\mathbf{u}_i$

such as

$$\mathbf{w}_i\ \mathbf{q}_i = \mathbf{u}_i$$

$$\mathbf{t}_i\ \mathbf{p}_i = \mathbf{v}_i$$

thus for the first latent variable, an estimation of $\mathbf{v}_i$ would be made, then an estimation of $\mathbf{p}_i$, then an estimation of $\mathbf{t}_i$, and so on, this process is cycled until convergence. The residual matrices are then calculated and the next eigenvector generated. This process can be stopped when the required amount of information has been extracted from the matrices. One of the major advantages of PLS is that this process can be carried out for more than one Y Block vector, this process needs to be carried out for each Y Block vector, producing a vector of weights for each. This can increase the time taken for the calculations considerably, the number of calculations required is multiplied by the number of Y Block variables.

It is the contribution from the unwanted variables that introduces a large proportion of the error in the calibration model, although the coefficients of unwanted variable tend to zero, they are rarely actually at zero. Thus in data sets with large errors, or samples with large matrix effects the contribution from unwanted variable can introduce a significant quantity of error. Removing these sources of error greatly reduces the predictive error of the model.

Several methods have been proposed to improve PLS, the three most common are variable selection by examining the correlation's between the variables of the independent matrix and the target matrix (8,9), examining the magnitude of the loadings coefficients (10,11), and using genetic algorithms to select variables (12,13,14). These methods are workable under certain circumstances, however they all have flaws.

Selecting variables by correlation is only useful were there is only one dependant (y-block) variable. Where the number of variables is greater than this there is no benefit obtained since a large number of variables will be selected, and many will have large quantities of noise associated with one or more of the other dependant variables. Even in the case of only one dependant variable this is quite an inefficient method as the variables selected in spectra tend to be from the centre of peaks, which captures little information about contaminants, and often leads to poor performance in prediction.

Selecting variables by examining the loading coefficients makes the assumption that a small coefficient indicates a variable that adds nothing to the model. This is often an

invalid assumption as variables with small coefficients can contain information about contaminants and noise that will improve the predictive ability of a model. In general models produced b this method tends to lack predictive robustness and are easily affected by unexpected contaminants or unusually high noise for a sample.

Using genetic algorithms to select variables is potentially a very good approach however in most approaches there is some trouble identifying the variables that are selected to produce the best model. Genetic algorithms also do not have very positive discrimination towards retaining a variable that is useful to a model, any selected variables can be discarded during the modelling process regardless of its usefulness, and there is no certainty that it will register as an important variables and be re-selected later.

Many of these approaches use leave-one-out cross validation, this approach can be misleading with regards to the error in the model, usually suggesting a lower number of latent vectors and a better predictive error than is found using a pure test/validation set. For this reason cross validation has not been used in this paper, instead a validation data set is used. Much of this work can be seen applied to Multivariate Linear Regression in the paper by Anthony Walmsley (15)

This paper suggests that a possible approach to the problem of calibration error is to force the coefficients for unwanted variables to zero. Removing these variables from the data set has the same effect. The problem remains of how to effectively remove unwanted variables from the model, quite often an attempt of this is made in the data pre-treatment stage, many spectral calibration software packages offer the chance to

exclude portions of the spectra that are known to be unimportant for the calibration. This approach risks the removal of variable important to the model, and also leaves a large proportion of the data set untouched. A better approach would be to find a method of selectively eliminating each variable on the basis of merit. The method outlined here uses the approach of including a variable in the data set only if it actually reduces the error in the predictive ability of the model, not just the error in the modelling of the training set.

Matrix singularity can be a problem with factor analysis techniques, especially where a lot of co-linear variables are present, it is quite easy to produce a singular or near singular matrix. By removing surplus variables, often co-linear ones, this problem can be minimised, and even ill conditioned or poorly scaled data can be used to produce low error models.

All the data sets used were pre-treated with autoscaling, the function for autoscaling can be seen below and is taken from Chemometrics: A textbook (2). The autoscaling was carried out by variable, such that the mean of each variable is zero and the standard deviation is one. The number of latent variables required for the models were determined in advance using cross validation, it was found that the optimum number of latent variables were unchanged by the variable reduction. This was expected since the model improvement is based on removing error contributed by unwanted variables.

Autoscaling.

$$x'_{iK} = \frac{x_{iK} - \overline{x_K}}{S_K} \qquad \text{where} \quad S_K = \left[\frac{1}{NP-1} \sum_{i=1}^{NP} \left(x_{iK} - \overline{x_K} \right)^2 \right]^{1/2}$$

Experimental

The Data Sets:

Three data sets were used, one UV spectra data set, and two synthetic data sets.

The UV Data set:

The data consisted of 52 spectra of 4 transition metal ions (Fe, Co, Ni and Cu) run on a UV/VIS spectrometer, over the 190-890 nm range, at a varied concentration range The entire spectra range was digitised, with a data spacing of 3.3nm, giving 211 spectral points. The data was then split to give 40 training samples and 12 'unknowns' Figure 1 in the appendix shows a plot of the spectra before any pre-treatment, figure 10 shows the data set after autoscaling.

Synthetic Data Set 1

Sixty samples of two hundred and fifty points with four overlapping peaks of random concentration, 4% normally distributed random noise added to each data point, 100% peak height systematic noise added to first 40 points, three non-linear response components, one linear response component. The non-linear response components were two squared terms, and a logarithmic term.

Figure 2 in the appendix shows a plot of the data set before any pre-treatment, figure 11 shows the data set after autoscaling.

Synthetic Data Set 2

Eighty samples of two hundred and fifty points with four overlapping components of random concentration. Up to 10% randomly distributed noise added to each data point.

Figure 3 in the appendix shows a plot of the data set before any pre-treatment, figure 12 shows the data set after autoscaling.

Data Pre-treatment

Data was treated using autoscaling, producing data sets where the variance in the variables has a mean of zero.

The algorithm

j is the number of rows
i is the number of columns
k is the number of variables
r is the number of components being predicted
q is the number of samples
h is the loop number
N is the matrix of actual values
P is the matrix of predicted values
T is matrix of training data
V is matrix of validation data
C^1 is matrix of training concentration information
C^2 is matrix of validation concentration information
S is matrix of selected variables (initially is empty)
s is the number of selected variables

Calculate PLS using T_{qk} and C^1_{qr} (1)

Predict using V_{qk} and C^2_{qr} (2)

$$BASEPRESS = \sum_{i=1}^{r} \sum_{j=1}^{q} (N_{ij} - P_{ij})^2 \quad (3)$$

Start loop (h)

Calculate PLS using $[S_{qs}\ T_{qk-h}]$ and C^1_{qr} (4)

Predict using $[S_{qs}\ V_{qk-h}]$ and C^2_{qr} (5)

$$PRESS = \sum_{i=1}^{r} \sum_{j=1}^{q} (N_{ij} - P_{ij})^2$$

If BASEPRESS < PRESS then add the removed variable to ***S*** and BASEPRESS changes to PRESS

Stop loop when ***h*** is equal to ***k***
Loop

Randomly shuffle the variables in ***S***

Repeat the above loop, replacing the contents of ***T*** with ***S*** and setting ***S*** to empty

Record the variables in ***S*** and the final value for BASEPRESS

Repeat the whole process at least $\sqrt{k}$ times

Determine the iteration with the lowest BASEPRESS

A flow chart of the variable removal procedure can be seen in Diagram 1 in the appendix.

SET-UP

Set-up involves randomly sorting the samples, then splitting them into a training set and a test set, then randomly shuffling the variables. PLS is carried out on all the variables in the training set (1) and a prediction produced on the test set (2). The PRESS (3) from this prediction is used as a BASEPRESS (3) for the model.

FIRST TRAINING STAGE

Initially only one training stage was used with no squashing function, but this resulted in an excessive retention of variables. A squashing function was then added, this reduced this problem, however the algorithm was found to be very sensitive to the squashing function and several attempts were needed with each training set to find an appropriate value. A second training stage together with a second squashing function

were added with the result that the number of final variables was reduced and the algorithm became less sensitive to picking a correct squashing function.

One variable is removed from the data set, a new model produced, and the test set used to produce a PRESS(3). The new PRESS is compared to the BASEPRESS. If the PRESS is greater, the model is producing more error in prediction, the variable is re-introduced into the data set (4), is marked as important to the model and the base PRESS changed to this new lower PRESS. If the PRESS is smaller the model is producing less error on prediction and the variable is discarded. The way in which the BASEPRESS and the new calculated PRESS is compared is determined by a squashing function. A squashing function of 1 means the two values are compared directly, there is no bias towards removing or keeping variables. If the squashing function is less than one the new PRESS must be a significant improvement over the BASEPRESS (the significance determined by the actual value of the squashing function), this will cause variables to be discarded more frequently. A squashing function greater than one will cause variables to be retained because the new PRESS will have to be significantly smaller than the BASEPRESS. In practice a squashing function smaller than one is normally used.

This process is repeated until all variables have been tested. After all the variables have been tested the variables that have been marked as important to the model are passed onto the second stage.

SECOND TRAINING STAGE

The remaining variables are again shuffled randomly and variables are removed individually from the remaining data set to determine whether they are important to the model, again a squashing function is used to gauge the significance of a variable

to the model. Once this second stage is completed the lowest press produced is recorded, together with the identity of the variables that produced it. This is the end of one iteration.

REPEAT ITERATIONS

The whole loop is repeated, the variables are shuffled each time, the samples are not. Once the required number of iterations has been carried out the best PRESS is determined from all the iterations and the variables that produced that PRESS are displayed. A number of iterations are required to find a statistical minimum PRESS, the actual number is dependent on the size of the data set.

Results & Discussion

For the variable selection stage of model development the PRESS is used to calculate the model error, however as this is not a useful comparison of the ability to model different components so the percentage error of prediction (PEP) is used for this. This enables the comparison of different components and different models.

The two stage variable removal is required for two purposes. First it improves the selection of a suitable squashing function. Secondly, during the initial selection procedure a variable may be selected that reduces the error in the model, but later a second variable may be retained which provides the same information to the model but with less error, only one of the two would be required. The second step serves to remove these surplus variables.

UV Data

The concentration of Fe in this data set was known to be at or below the limit of detection, this means that the spectral information referring to the Fe is almost entirely noise. The other three components contain a far higher signal to noise ratio. These results show a comparison between ordinary SIMPLS and variable selection PLS. Both the PRESS and the PEP are shown for the whole model and for the individual components. In all cases it can be seen that variable selection PLS outperforms ordinary PLS. Also shown is a histogram for the PRESS obtained in each iteration of the variable selection routine, and the number of variables used for each iteration. Table 1 shows the comparison between the two PLS methods.

UV DATA		7 LV's
PLS Model		
Base PRESS		23.3374
	PRESS	**Mean PEP**
Fe	21.8267	79.3353
Co	0.7054	32.7349
Ni	0.5974	15.9384
Cu	0.2126	9.0851
VS-PLS		
Base PRESS		3.7366
	PRESS	**Mean PEP**
Fe	3.7648	62.7932
Co	0.0440	19.6281
Ni	0.1563	12.1503
Cu	0.0046	3.6246

Table 1: Comparison of PLS and VS-PLS

Base PRESS is the PRESS for al four components together, the PEP and PRESS were then calculated for each individual component. Figure 4 shows the histogram of PRESS, the graph shows normal distribution, so the chance of getting a lower PRESS than any already achieved can be calculated.

Figure 5 shows the number of variables used in each iteration to produce the minimum error for that iteration.

Synthetic Data Set 1

This data set provided the most problems, three of the four components are non-linear, a logarithmic term, and two squared terms were used to define the way the concentration varied with the spectra. In this case table 2 shows that using variable selection PLS was inferior to ordinary PLS for all but the linear component. As with the previous case the four components were calculated simultaneously, here the algorithm could best reduce the press in each iteration by ignoring the contribution from the three non-linear components can only reducing the error for the linear component. This is illustrated by figure 6, where the PRESS remains constant for a majority of the iterations, any improvement in the PRESS for the third component is masked by the large error in the other four components.

Figure 7 shows the number of variables required to produce minimum errors for each iteration.

When the data set was recalculated for individual components, variable selection PLS was superior to ordinary PLS.

Synthetic Data Set 1		7 LV's
PLS Model		
Base PRESS		6.5669
	PRESS	**Mean PEP**
Comp 1	2.3795	57.2687
Comp 2	2.2962	137.7124
Comp 3	.21709	1.79
Comp 4	1.8912	23.2486
VS-PLS		
Base PRESS		45.7136
	PRESS	**Mean PEP**
Comp 1	4.1828	60.4518
Comp 2	16.9242	240.7574
Comp 3	0.0002	.001
Comp 4	24.6406	101.0593
PLS Model 2		
	PRESS	**Mean PEP**
Comp 1	2.1345	45.1235
Comp 2	1.9810	113.1178
Comp 3	.21652	1.5440
Comp 4	1.7714	21.4573
VS-PLS 2		
Base PRESS		45.7136
	PRESS	**Mean PEP**
Comp 1	2.0004	33.8901
Comp 2	1.1105	103.1035
Comp 3	2.16E-28	1.08E-12
Comp 4	1.1908	18.1123

Synthetic Data Set 2

Four latent structures were required to model this data with minimum error. This is expected as the data set is linear and does not contain any irregularities.

The data used here has no non-linearity, the error added is normally distributed. This means that the PEP shown for all but the third of the four components using variable selection PLS is at the minimum possible for this data set.

With all components the variable selection PLS performed better than the ordinary PLS.

Synthetic Data Set 2		4 LV's
PLS Model		
Base PRESS		1.1056
	PRESS	**Mean PEP**
Comp 1	0.16282	25.059
Comp 2	0.36906	30.546
Comp 3	0.39796	40.955
Comp 4	0.1758	9.8782
VS-PLS		
Base PRESS		
	PRESS	**Mean PEP**
Comp 1	0.004063	2.2349
Comp 2	0.003111	2.9662
Comp 3	0.003802	8.2871
Comp 4	0.004823	2.4565

References

(1) D.L.Massart, et al, Chemometrics: A Textbook, Elsevier,1988

(2) E.R.Malinowski, Factor Analysis in Chemistry, Wiley Inter-Science, 2nd edition 1991

(3) R. Fisher & W. MacKenzie, Studies in Crop Variation. II. The Manurial Response of Different Potato Varieties, Journal of Agricultural Science, 13 (1923), 311-320

(4) H. Wold, Nonlinear estimation by iterative least squares procedures, in F.David (Editor), Research Papers in Statistics, Wiley, New York, 1966, 411-444

(5) S. Wold, W. Lindberg, J.A.Persson, Partial Least Squares Method for Spectrofluorimetric analysis of Mixtures of Humic Acid and Ligninsulfonate, Anal. Chem., 55 (1983) 643

(6) S. Wold, N. Kettaneh, K. Tjessem, Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection, JOURNAL OF CHEMOMETRICS, 1996, Vol.10, No.5-6, pp.463-482

(7) E. Ronchetti, Robustness aspects of model choice, STATISTICA SINICA, 1997, Vol.7, No.2, pp.327-338

(8) TW. Ogorman, RF. Woolson, Using Kendall Tau(B) correlations to improve variable selection methods in case controlled studies, Journal of Biometrics, 1995, Vol 51, No 4, 1451-1460

(9) MJ. Adams, JR. Allen, Variable selection and multivariate calibration journal of models for X-ray fluorescence spectrometry, JOURNAL OF ANALYTICAL ATOMIC SPECTROMETRY, 1998, Vol.13, No.2, pp.119-124

(10) NM. AlKandari, IT. Jolliffe, Variable selection and interpretation in canonical correlation analysis, COMMUNICATIONS IN STATISTICS-SIMULATION AND COMPUTATION, 1997, Vol.26, No.3, pp.873-900

(11) HM. Heise, A. Bittner, Rapid and reliable spectral variable selection for statistical calibrations based on PLS-regression vector choices, FRESENIUS JOURNAL OF ANALYTICAL CHEMISTRY, 1997, Vol.359, No.1, pp.93-99

(12) D. JouanRimbaud, D. Massart, OE. DeNoord, Random Correlation in variable selection for multivariate calibration with a genetic algorithm, Chemometrics and Intelligent Laboratory Systems, 1996, Vol 35, No 2, 213-220

(13) K. Hasegawa, Y. Miyashita, K. Funatsu, GA strategy for variable selection in QSAR studies: GA-based PLS analysis of calcium channel antagonists, JOURNAL OF CHEMICAL INFORMATION AND COMPUTER SCIENCES, 1997, Vol.37, No.2, pp.306-310

(14) D. Broadhurst, R. Goodacre, A. Jones, JJ. Rowland, DB. Kell, Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with applications to pyrolysis mass spectrometry, ANALYTICA CHIMICA ACTA, 1997, Vol.348, No.1-3, pp.71-86

(15) A.D. Walmsley, Improved variable selection procedure for multivariate linear regression, Analytica Chimica Acta, 1997, 225-232

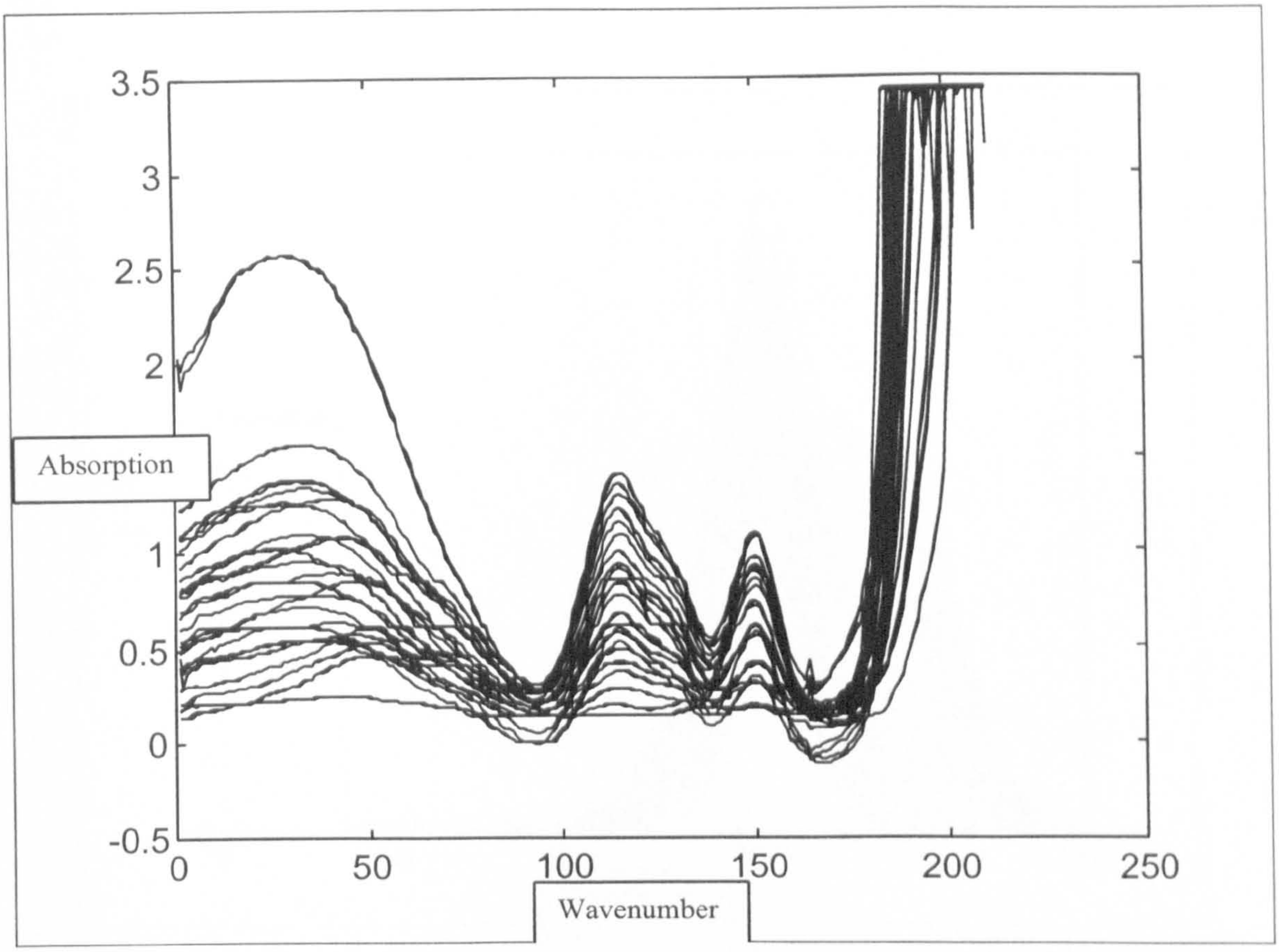


Figure 1: UV Data Set

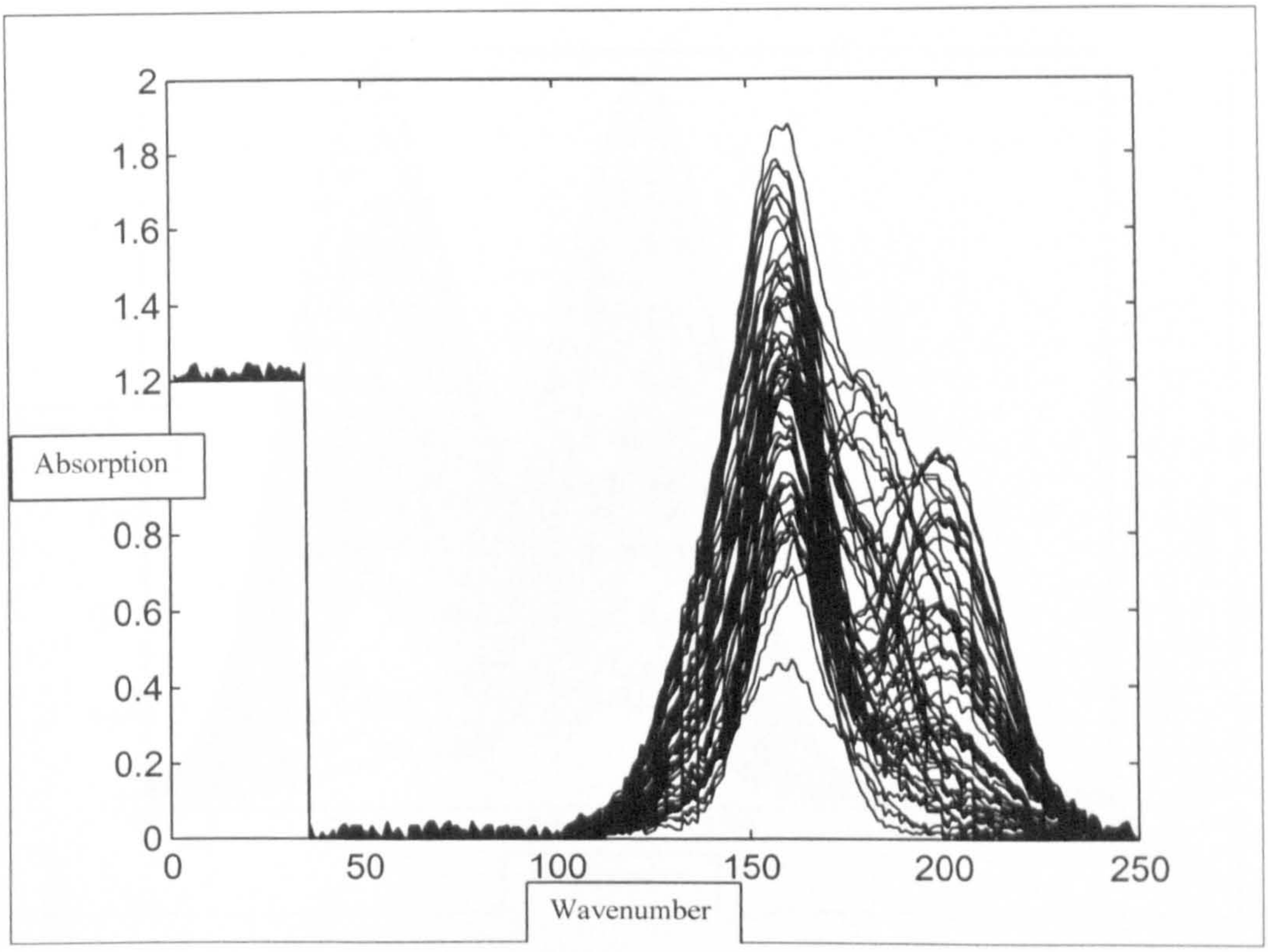


Figure 2: Synthetic Data Set 1

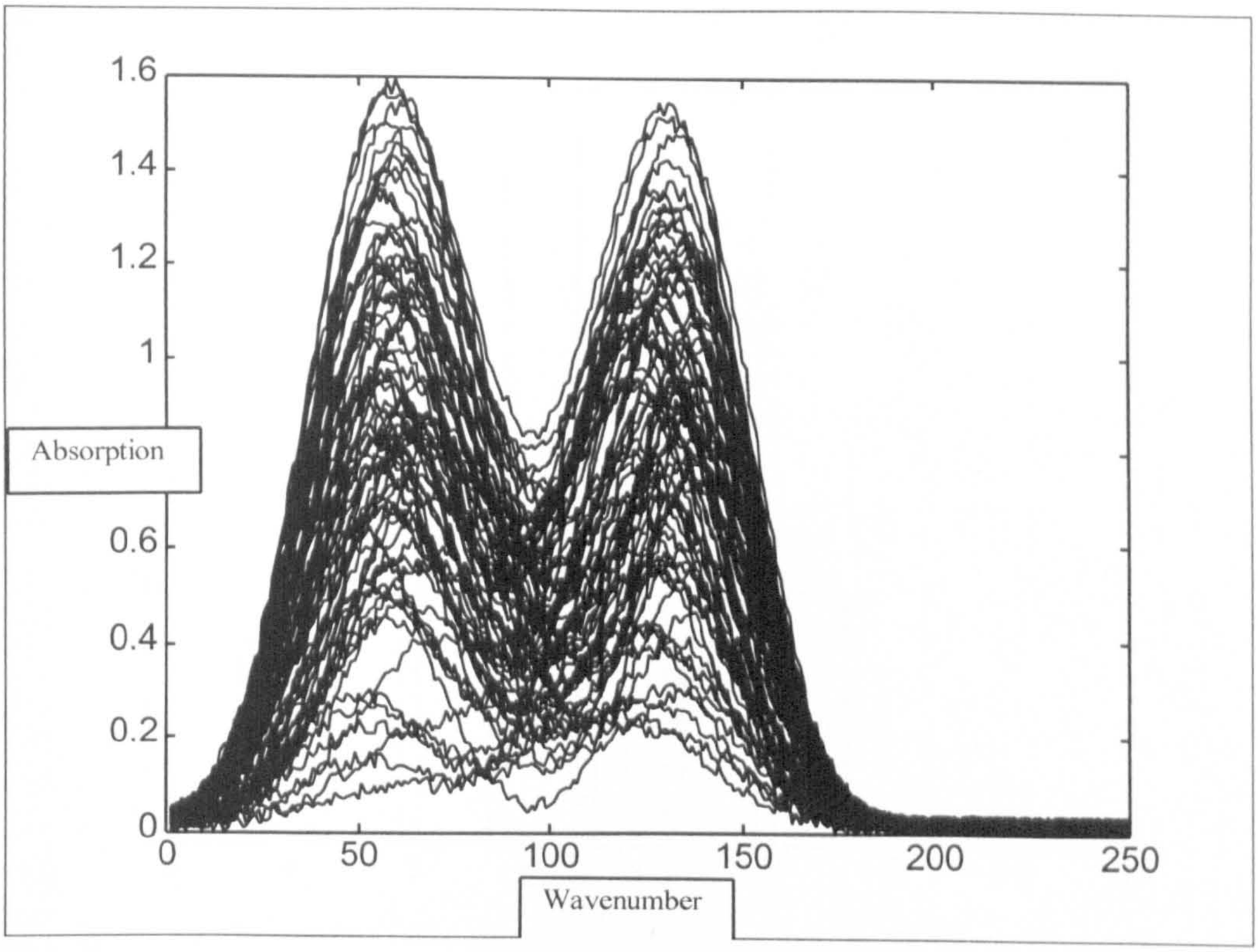


Figure 3: Synthetic Data Set 2

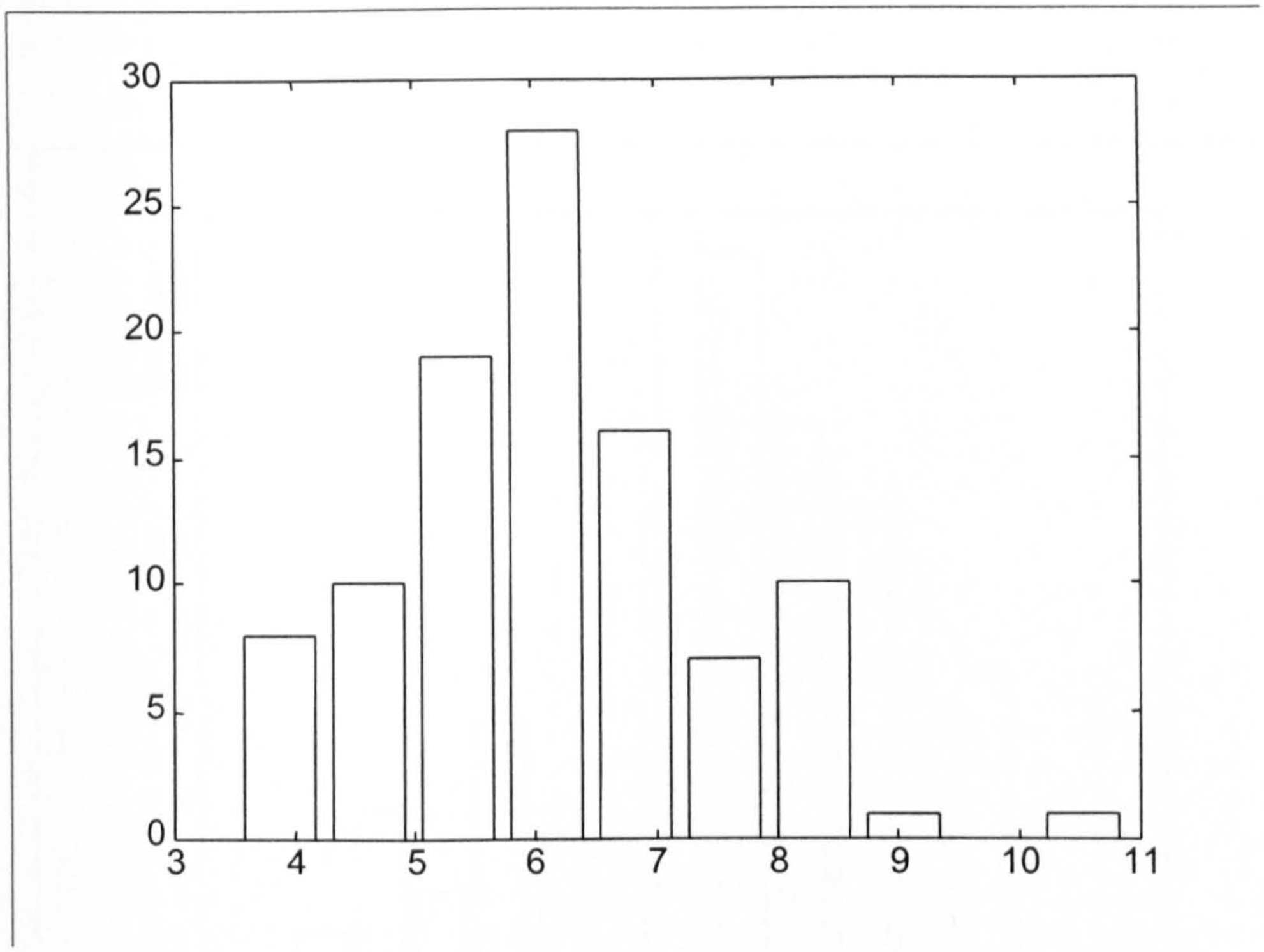


Figure 4: Histogram of PRESS for UV Data Set

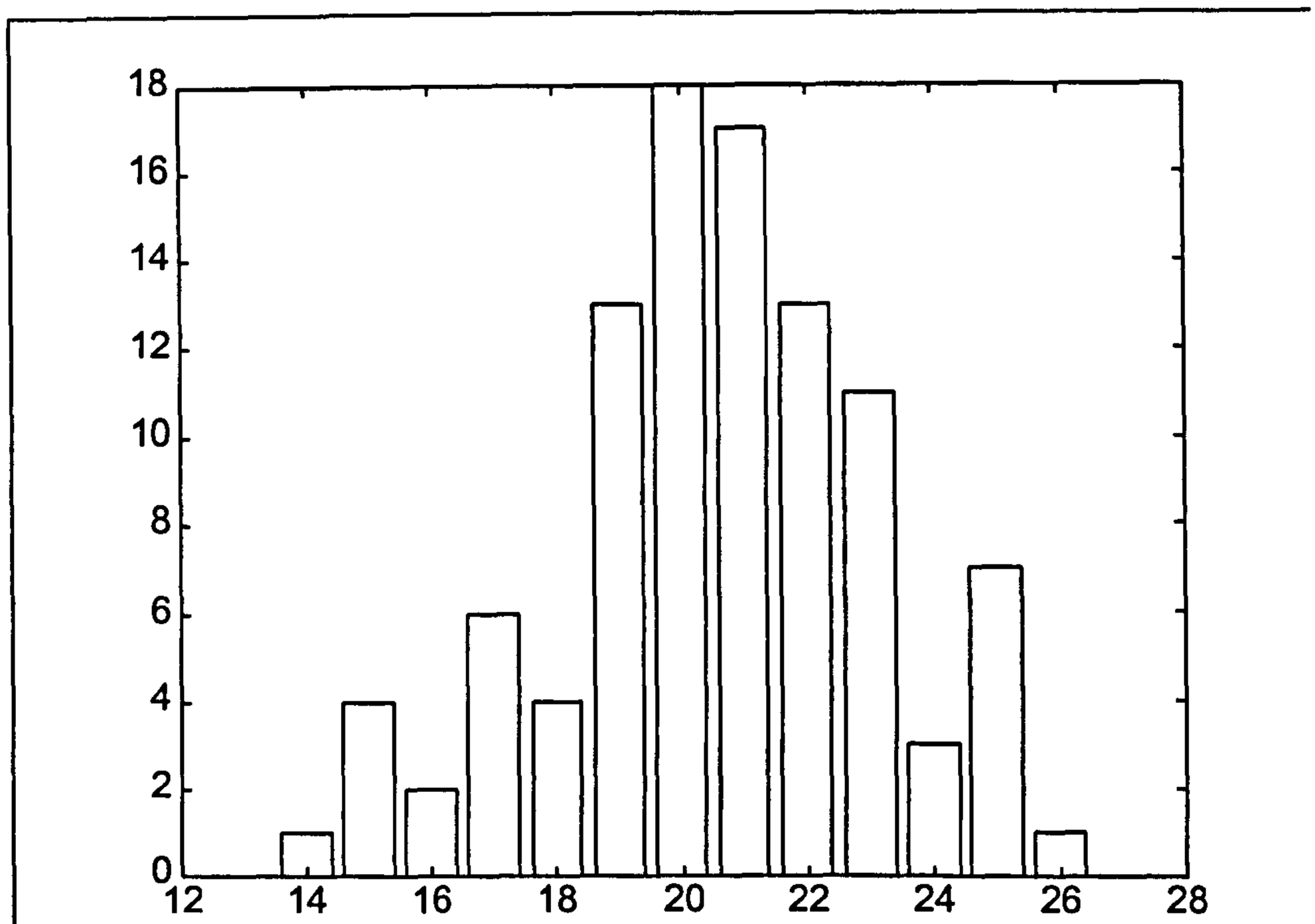


Figure 5: Histogram of Number of Variables Selected for UV Data Set

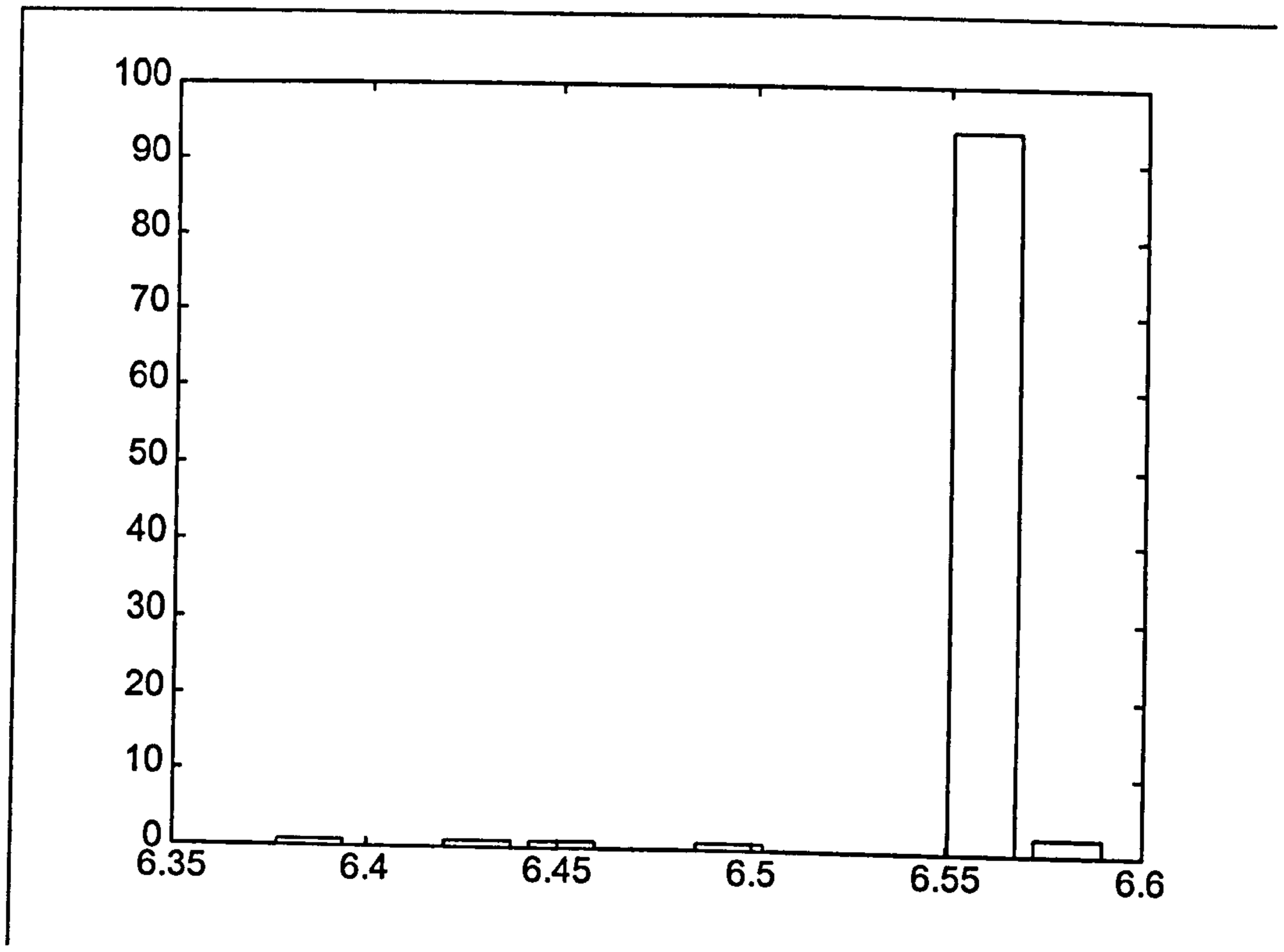


Figure 6: Histogram of PRESS for Synthetic Data Set 1

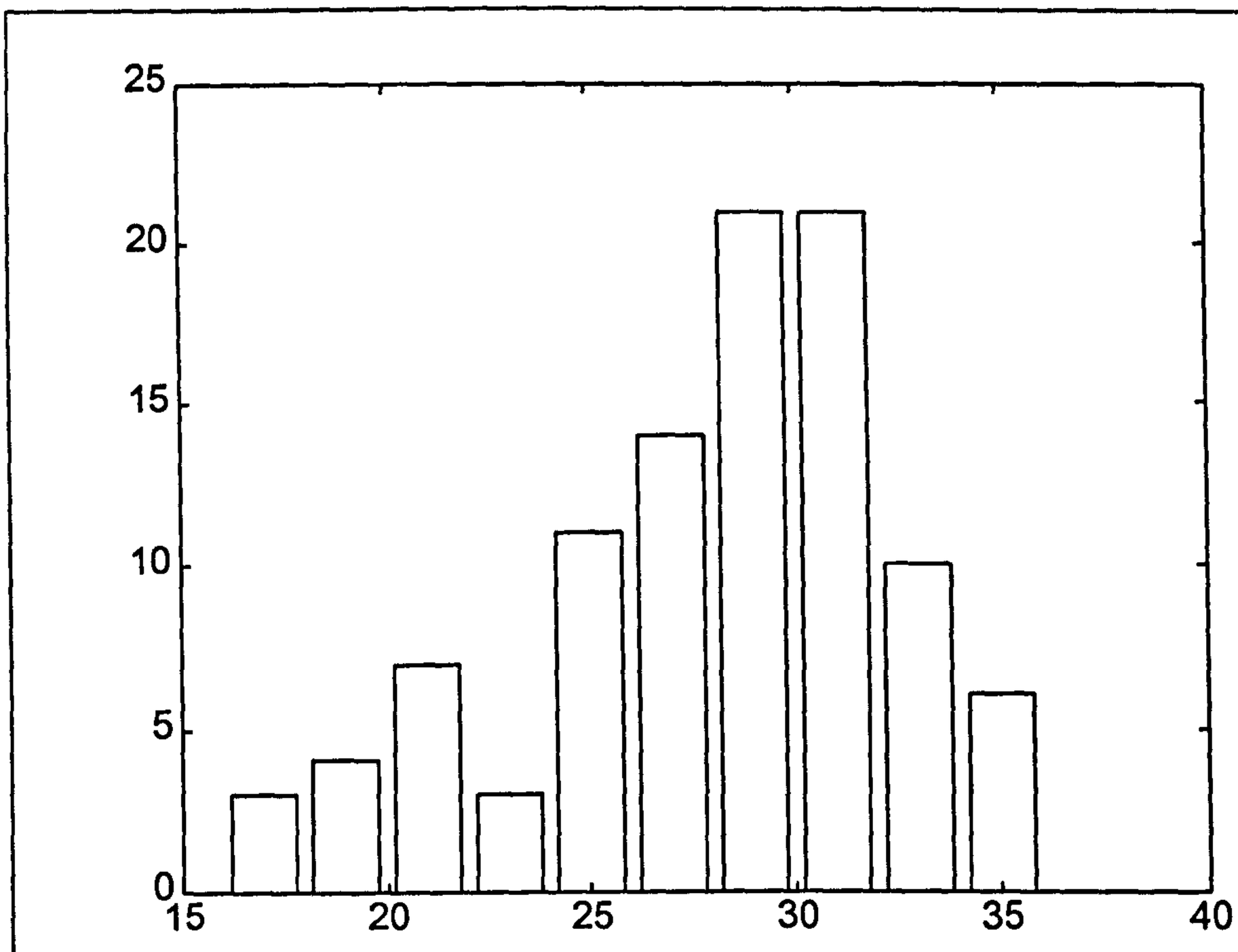


Figure 7: Histogram Showing Number of Variables Selected for Synthetic Data Set 1

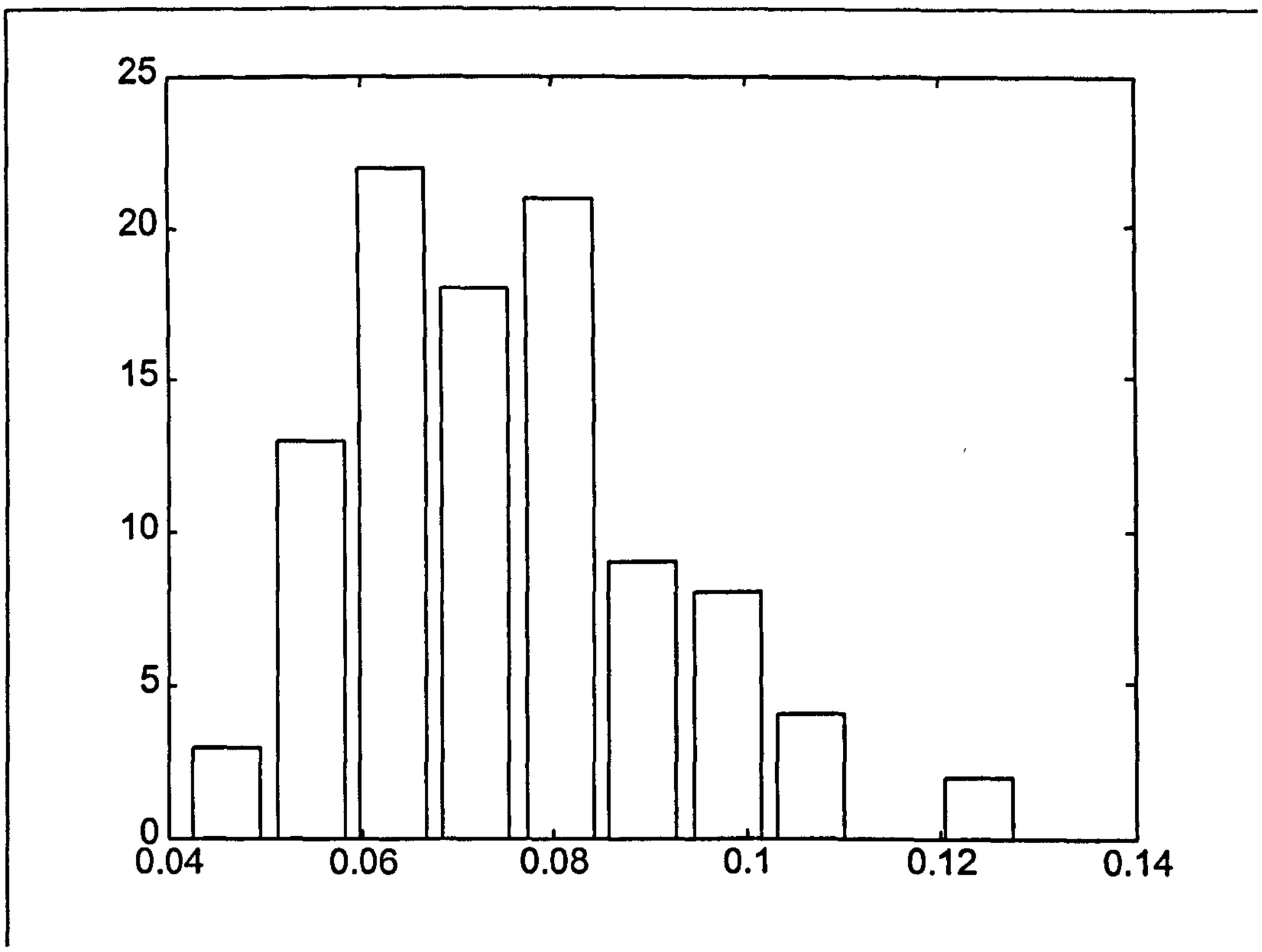


Figure 8: Histogram of PRESS for Synthetic Data Set 2:

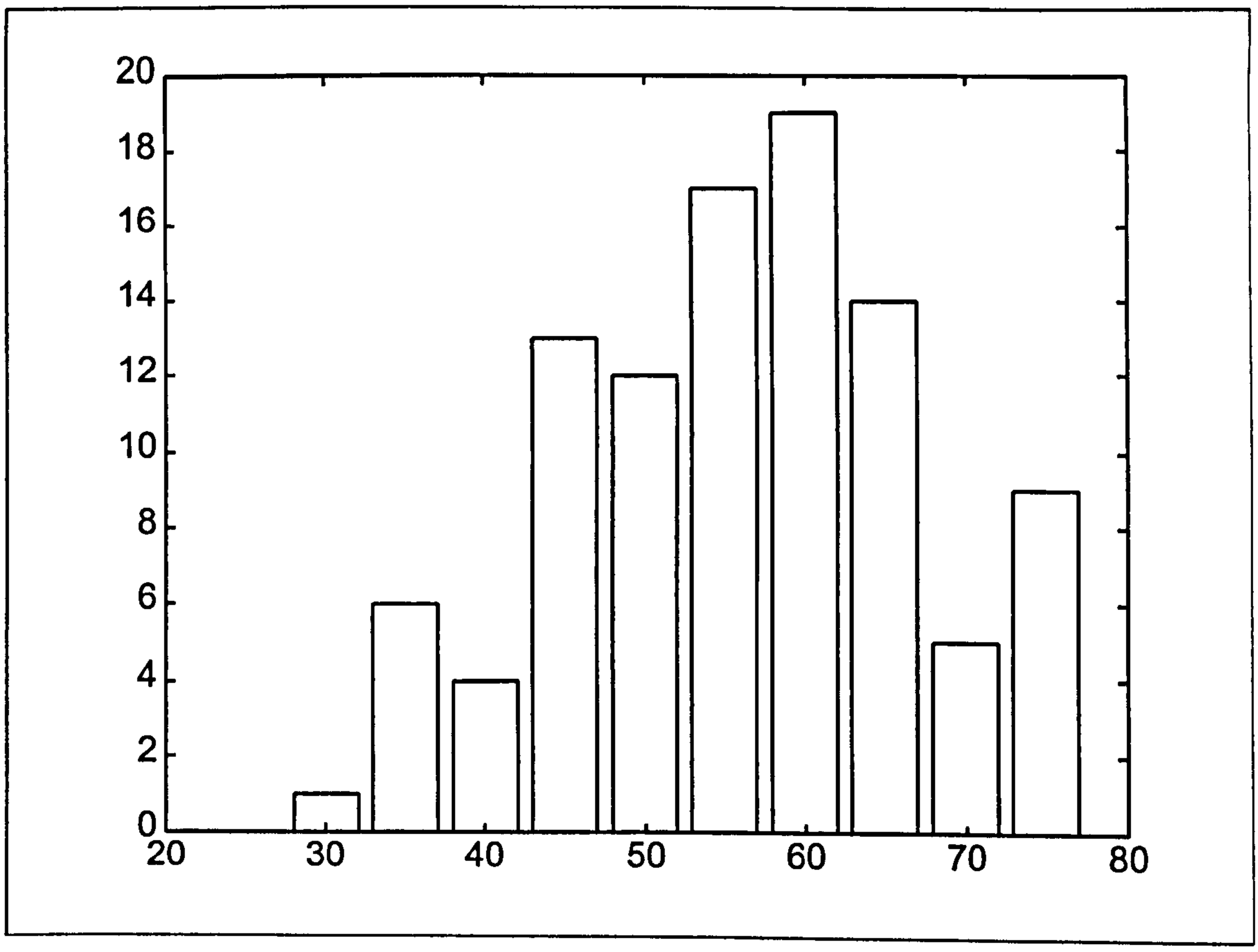


Figure 9:Histogram Showing Number of Variables Selected for Synthetic Data Set 2

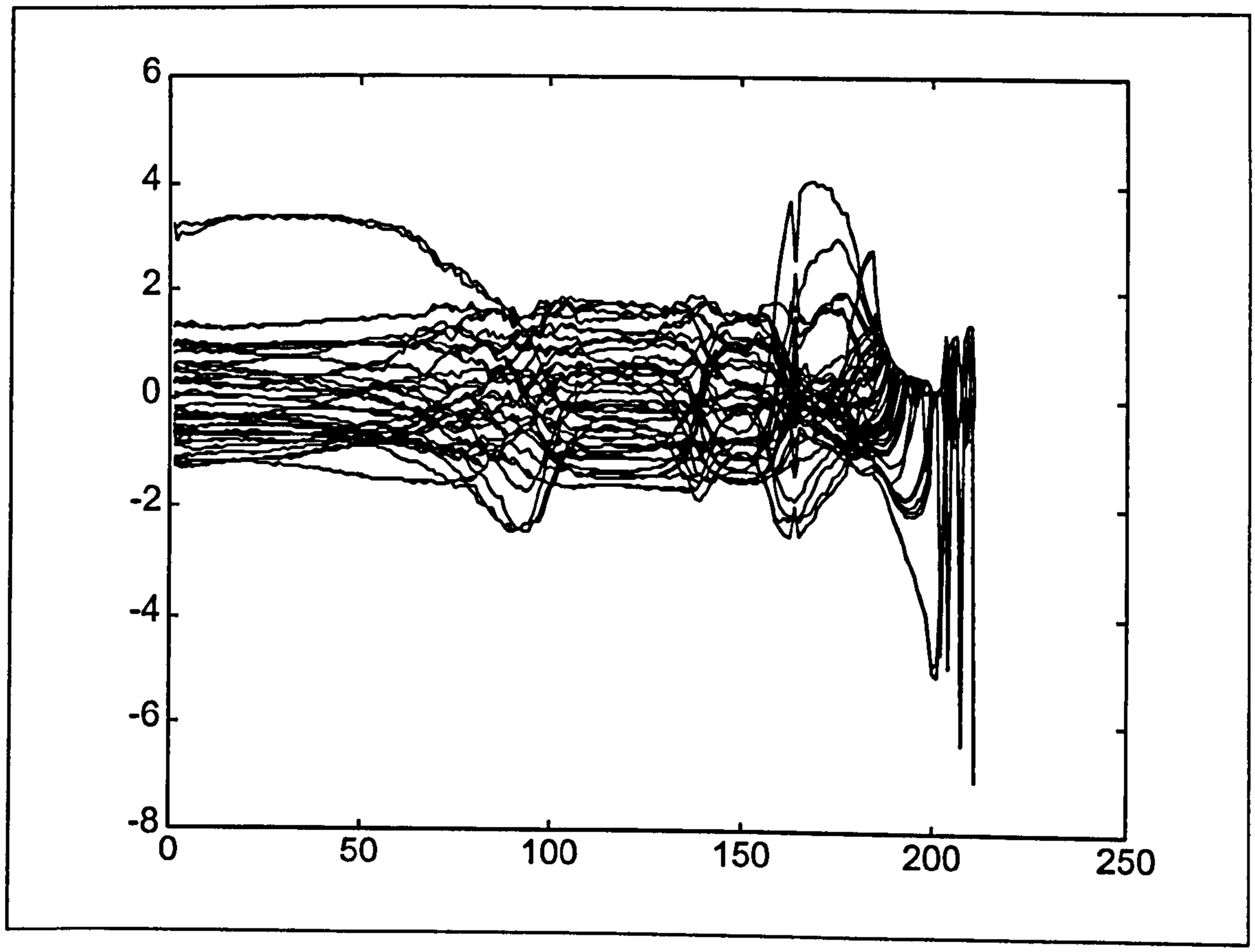


Figure 10: Autoscaled UV Data Set

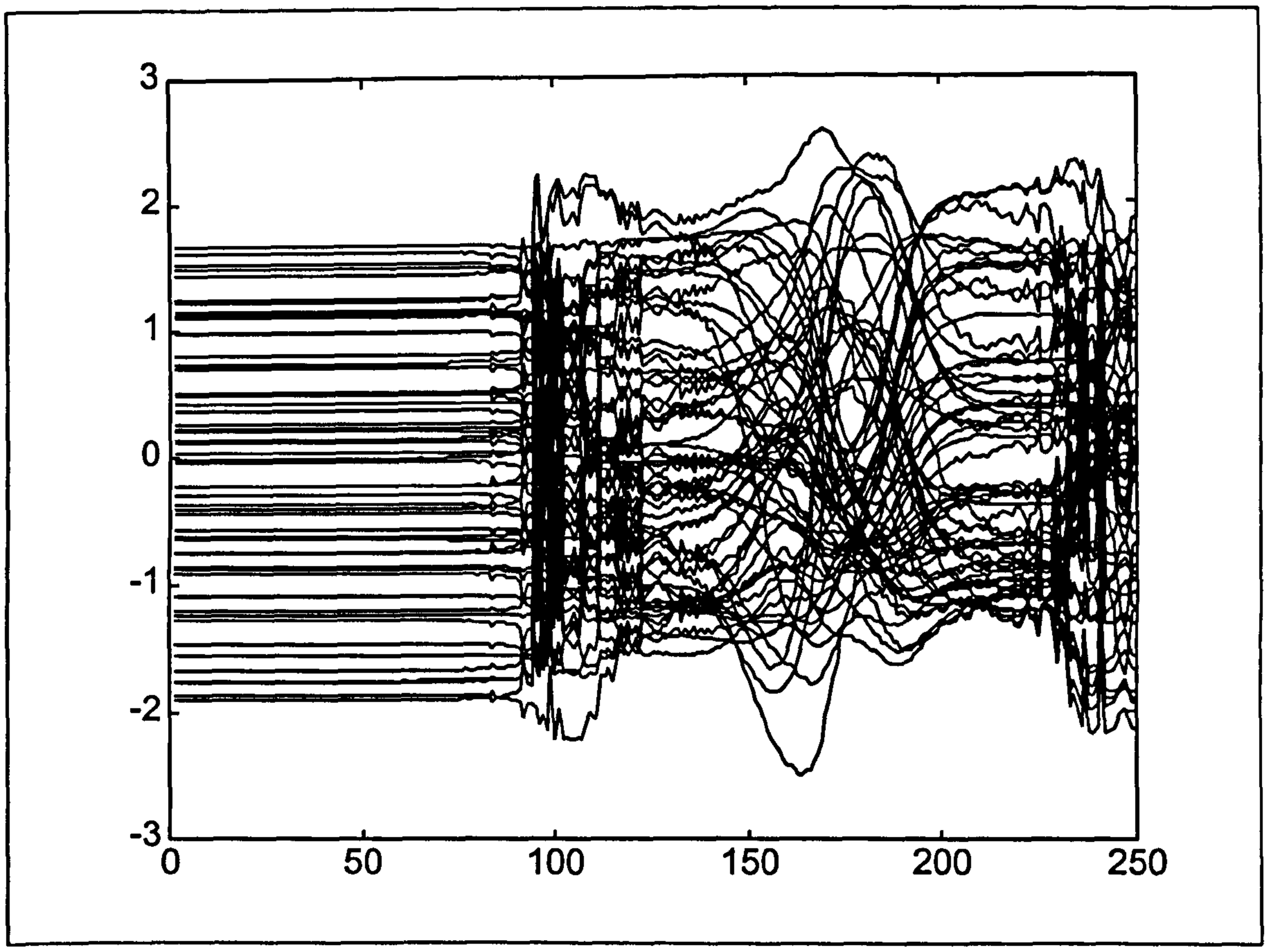


Figure 11: Autoscalled Synthetic Data Set 1

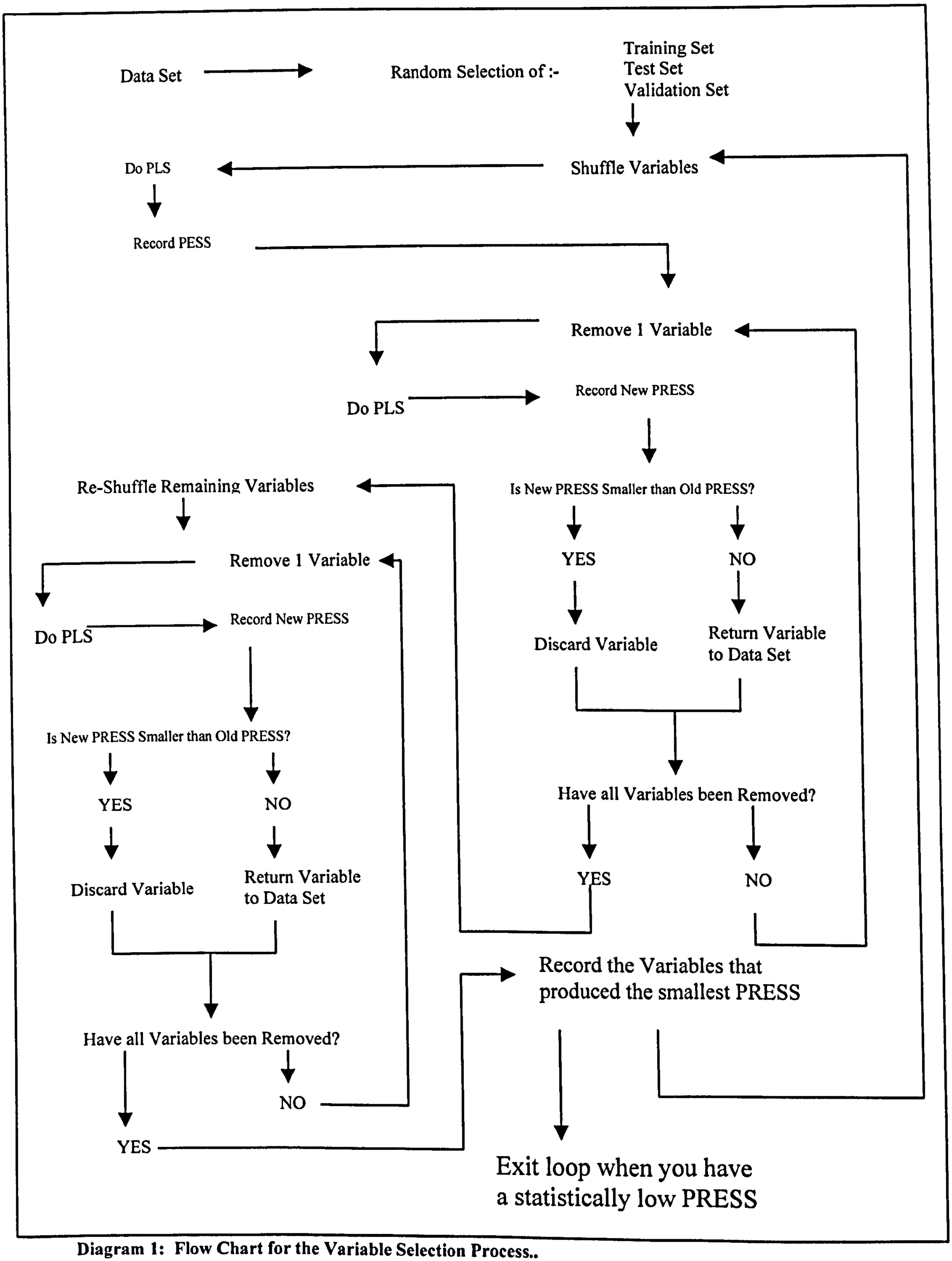


Diagram 1: Flow Chart for the Variable Selection Process..

Appendix VII

Intrasite Study Results

Abstract

This report is a detailed description of the analysis of the data produced during the routine testing of Intrasite Gel. The report looks at the currently available analysis data (the results generated by the daily batch analysis of Intrasite gel) and the data generated using the Paddington cup method. The data sets were examined to determine their reliability and error, the degree to which the process is under control was looked at, with particular attention to the issue of over sampling. The measured variable fluid absorption was also examined in detail to determine its value as an analytical measurement.

Introduction

Intrasite Gel is a carboxymethyl cellulose polymer gel, 2.3% by weight the remainder being water and propylene glycol. Intrasite Gel is made up from the powdered polymer slurried with propylene glycol and then mixed with water. The powdered polymer is produced in large quantities, and one batch is sufficient for at least a year's production. The powder is made up into smaller batches of the gel on a daily basis, and these small batches are further divided into six or so sub-batches. These small batches are then packaged into the delivery system (sachets or "appli-packs") and sterilised. Following sterilisation samples are taken for analysis. The current formulation of Intrasite has been used for several years and there is a significant quantity of data going back four years concerning the analysis of this formulation.

The measurements made on Intrasite Gel are two identities, identification of propylene glycol and the identification of sodium carboxymethyl cellulose, and measurements of pH, elasticity, viscosity, solids content, and fluid absorption.

The details of the tests carried out can be found in the following documents, obtainable from Smith & Nephew Ltd. Hull.

Identification of propylene glycol: SOP/QGM/029

Identification of carboxymethyl cellulose: SOP/QGM/135

pH: SOP/QGM/01

Elasticity: SOP/QGM/038

Viscosity: SOP/QGM/039

Solids Content: SOP/QGM/136

Fluid Absorption: SOP/QGM/028

The data set examined was for the X181 formulation of Intrasite Gel, and covers the time period from the 18th January 1995 through to the 10th December 1997. Most of the data was supplied on a spreadsheet, some had to be entered into a spreadsheet manually. The two identities were ignored for this analysis as all samples complied with these tests. The solids content test is carried out twice for each sample due to repeatability problems.

Work Carried Out

Initially the data set was examined for correlation between the variables, and for autocorrelation (7) within each variable, this was done both for the entire data set, and

for sections of the data set corresponding to individual polymer batches. Initially attempts were made to calibrate the data set against the fluid absorption variable. It was determined that this measurement (fluid absorption) was the one with the highest degree of error, and was thus providing the least information when analysed. The aim was to be able to predict future values of fluid absorption from the other variables, and thus have a degree of confidence that the fluid absorption of the product was within the specification range. A new method of determining the fluid absorption of Intrasite Gel was also developed as the quality of any calibration model is only as good as the errors in the reference data, and clearly these errors were initially quite high. The stability and control of the process were also examined using Cusum charts (7) and control charts (7) and the sampling frequency was examined to determine whether the material was being over or under sampled with respect to process control. The sampling frequency was considered using the Cusums, autocorrelations and control charts.

The Fluid Transfer Test

The current fluid absorption test (SOP/QGM/028) is the settling volume method, which involves monitoring the change in volume of a quantity of Intrasite Gel once a quantity of saline solution has been added to it. This test produces very poor results both from the issue of solubility and because of problems associated with reproducible measurement. One possible solution is to replace it with a fluid transfer test (also known as the Paddington cup method).

The fluid transfer test involves using weight measurements to monitor the transfer of fluid between two competing mediums. The test was originally developed to compare the hydrogels produced by different companies, this means that the test has some redundancies that were removed for testing just one material. Because the test was

developed to test the fluid transfer of a range of different hydrogels two different fluid transfer mediums were used, at different concentrations. The full test involves comparing the fluid transfers of each hydrogel between 30% gelatine, 20% gelatine, 10% gelatine, 4% agar, 3% agar, 2% agar and 1% agar. These materials range from strongly fluid absorbing (the 30% gelatine) which measures the ability of the hydrogel to donate water, and strongly fluid donating (1% agar) which measures the ability of the hydrogel to absorb fluid. These variations are required in order to compare different hydrogels, which might have widely different fluid transfer rates. When testing only one type of hydrogel there is no requirement to compare different fluid transfer mediums. Prior to running the full series of tests the correct medium to use was determined by testing each to determine maximum response. Intrasite gel is quite balanced between donating and accepting fluid in comparison with many other hydrogels available and thus either the high concentrating gelatine or the low concentrating agar would have been suitable. The 2% agar solution was selected as the best medium to use. An agar base was selected because preparing the agar was easier and faster operation compared with setting up the gelatine. The 1% agar would theoretically have given a better response, however 1% agar is a very fragile material, and physical distortion has a large effect on the results. This leads to larger levels of experimental error that outweighs the gain from the improved response.

The test operates by allowing a layer of Intrasite gel of known mass to equilibrate with a layer of 2% agar of known mass, in a sealed environment. After equilibrium the agar layer is re-weighed and the change in weight is expressed as a percentage change. Either layer could be weighed as a measure of change, however the agar layer is solid and is easier to handle during the experiment. The test was carried out on three replicates for each batch of gel and the results can be seen in figure 46, the

replicate variation can be seen in figure 47. An ANOVA was performed and showed that the variation between sample was more significant than the variation between replicates despite the large variation in the replicates. The very large initial values are due to inexperience with the fluid transfer test. If the test were to be introduced as a standard test the variation between replicates could be reduced significantly by better control of the environment the test is carried out in and better preparation of the agar.

Results & Discussion

The data Set

The solids content property of Intrasite Gel is measured twice. Due to the low variability, low standard deviation, high correlation between the two replicates, and poor correlation between these variables and the others in the data set, no advantage was seen for including both variables in the analysis and the variable with the fewest missing values was taken.

The full data set can be seen graphed in figures 1 through 5 in the appendix (only one of the solids contents variables is graphed). The table of correlation between the full variables in the data set can be seen in table 1, and the correlation between the variables for the time period of January 1997 through to December 1997 can be seen in table 2. With the possible exception of viscosity and elasticity there is a very poor correlation between the variables, and it should be noted that the correlations are worse when the shorter time span is selected. The poor correlation for the shorter period of time is due to the high error in each measurement, this acts to mask any correlation, with the longer time series the underlying trend is more apparent and the correlations can be seen.

The autocorrelation for each variable can be seen in figures 6 through 10. Autocorrelation is a technique that looks at the correlation between any one current point in a series and compares it to neighbouring points and short series of neighbouring points. Autocorrelation is useful for showing periodic trends in a time series, as an example, autocorrelation would highlight the seasonal variation in recorded air temperature as a periodic cycle. With Intrasite Gel the autocorrelation over thirty points show that there is little immediate correlation between any two neighbouring readings, however the level of correlation is quite high and does not change rapidly over time. This shows in all cases that the process is stable over the sixty-day window examined with only random noise distorting the autocorrelation. The autocorrelations shown in figure 6 through 10 are typical for a stable process with a high degree of random noise in the measurements. This indication of stability is also displayed in the Cusum charts where the effects of sampling frequency have been examined (figures 26 through 45), reducing sampling frequency has no effect on the process shown in the Cusum charts.

Process Control and Stability

The process stability for the production of Intrasite has been examined for the period of January 1997 through December 1997. The control stability was examined using control charts. The charts for each variable can be seen in figures 11 through 15.

The control limits set on the graphs represent two and three times the standard deviations of the data set, and even at the points where the readings have passed the action limit the material being tested is still well within the specification of the product. All the control charts show good stability except the pH chart. The periods where the control charts show instability match the periods when the analyst carrying

out the measurement changes. This is most clear at about the end of june 1997 and the end of august 1997.

The Cusum charts that can be seen in figures 16 through 20 show the general trend of the process, with large amounts of random variability in a measurement it can be difficult to determine trends in the process, but these can be more easily determined by looking at Cusums. The Cusums for the measurements made on Intrasite Gel all show the same trend. The process can be seen to change in the second half of the control charts, and this is mirrored in the Cusum charts where it can be seen that the process appears to change significantly. This change can actually be seen to be linked to a change in analyst at Smith & Nephew, and not to a real change in the process. The process appears to be less stable in the second half of the Cusum and control charts, this is likely to be due to the fact that the analyst changes quite frequently after this time, where before the analyst was constant for a large period of time. If the Cusum charts are compared for the period of time where the analyst was constant (figures 21 through 25) they can be clearly seen to be very similar. It should be noted that this is not in any way an indication of the quality of the analyst carrying out these tests, this merely indicated that there is a slight difference in the way in which each analyst reads and records results. It is also likely that the period where the process appears out of control on the control charts is caused by the change of analysts as well, towards the end of 1997 the analyst changes frequently. When the Cusum charts are examined in this manor the solids content chart, the elasticity chart and the viscosity chart are all very closely matched. The pH chart and the fluid absorption charts are not, this is due to these charts showing variation within the test, not displaying any real variation the process.

The sampling frequency for each variable has been examined by reducing the number of points used in each Cusum, as can be seen from figures 26 through 45. The trend shown by each Cusum shows the same features as the Cusums constructed using all the available data points. Obviously this technique is not appropriate to control charts where it is the individual values that are of interest, not the process trend.

Conclusions and Recommendations

From examining the autocorrelations, Cusums and control charts it is apparent that the process to produce Intrasite Gel is fairly stable over the long term, however due to error introduced from the measurement procedures, and variation introduced from different operators, predictions of future values are inaccurate. The measurement containing the greatest degree of error is the fluid absorption measurement (6). The measurement of pH also contains a large amount of random error. While the solids content, elasticity and viscosity reading also contain error these measurements all show a good indication of the general trend of the production process, as shown when comparing the Cusum charts produced when examining measurements made by a single operator at a time (Figures 21 through 25).

Process Recommendations

If the process to produce Intrasite Gel can be kept stable and under full control there is no reason to expect that the product will leave specification, to this end it is important to know how the process is behaving on an individual batch basis. The process for the production of Intrasite Gel appears to be under good control, based on both the control charts (figures 11 through 15) and the respective Cusum charts (Figures 16 through 20). There are two groups of recommendations that can be made from this

Medium Risk Proposal

The process control for the production of Intrasite Gel can be followed using the elasticity measurement, with the exception of the pH of the product the other properties follow the same trend as the elasticity. The basis for this is that the measurement for elasticity also shows the state of the other variables, when elasticity is within specification all the other measurements are in specification as well. The elasticity test is a fast test to carry out and could be carried out at line, giving a fast feedback as to process problems. Measurement of elasticity should be made for each batch produced (estimated at 4 to 6 measurements a day). If the elasticity control chart indicates that elasticity has moved into the action zone (which is still within the product specification) the other measurements should be carried out to ensure that no other problems exist. Measurements of viscosity, solids content and fluid absorption and pH should still be made every 20^{th} measurement of elasticity. The Cusums for these variables should then be compared on a regular basis with the Cusum for elasticity, with a marked deviation all measurements should resume at their previous frequency (one made per batch).

Low Risk Proposal

Measurement of elasticity and pH should be made for every batch, off line. When a measurement for either property moves into the warning zone of the control chart, measurements of the other properties should be resumed. If pH and elasticity remain stable then measurements of viscosity, solids content and fluid absorption should be made for every tenth sample. The Cusums for all variables should be compared at regular intervals to ensure that the process trends remain constant between the variables (the trend for all variables remain the same).

Specification Recommendations

The issue of specification is more difficult to address. It is not possible currently to predict individual measurements of analysis results based on any of the other analysis results, however this is due in the main part to the high random error in each of the measurements. This inability to predict measurements could cause problems as far as meeting requirements for reporting. From the data recorded for 1997 it is clear that with the exception of pH all the measurements follow the same trend. From this it can be assumed that if one measurement is out of specification or breaches the action limits then it is likely that other measurements will also fall out of limits. However without the ability to accurately predict individual measurements this assumption is difficult to prove in terms of analytical reported results. What is clear however is that for any one reported analytical result, the reported value is more likely to exceed specification due to error in measurement than it is due to real variation. Thus a more reliable way of assessing product quality could be by monitoring process trends not analytical results.

High Risk Proposal

None of the measurements currently made can show with any certainty exactly what the true value for any one of the properties really is. Thus it would be more efficient to follow the production of Intrasite Gel to determine that the production is under control, and select another measurement to ensure the product meats specification. A possible option is to stop the current testing and switch to an entirely new test for elasticity. It is quite possible to measure the elasticity of Intrasite gel without removing it from the apli-pack or sachet, several sonic interments for measuring elasticity are currently available, and these would appear to be clearly suited to the

task of measuring the elasticity of Intrasite Gel. The process is fast and is non-destructive. A much larger sampling rate could be taken, in a shorter period of time, and specification limits could be observed. Much of the work on elasticity determination using ultrasonic has been in the medical field relating to tissue elasticity, there is no reason why this work might not be adapted to examine the elasticity of Intrasite gel.

Medium Risk Proposal

The process trend can be used to determine product quality. The pH of Intrasite Gel will need to be monitored following the standard SOP. The pH does not follow the trend of the other properties, and is potentially the most critical in terms of health and safety, however the stability of the other parameters can be assessed using just the elasticity measurement. If the Cusum for elasticity suggests that the process leaving control then measurement of the other variables should be resumed until the process becomes stable again.

Low Risk Proposal

The low risk proposal assumes that analytical measurement can be reduced without compromising the required reporting level for Intrasite Gel. The frequency of analytical reporting for viscosity, fluid absorption, and solids content should be reduced to one-tenth their current level, measurement of elasticity and pH should remain at their current levels.

Appendix

References

References 1, 2, 3, 4, & 5 refer to internal Smith & Nephew reports

1. SR\ET002\MS93-2 Effect of propylene glycol and saline on the properties of Intrasite Gel
2. SR\TW015\MS91-2 Development of a method to demonstrate that Intrasite Gel has the ability to absorb or release water
3. QGM\137 Validation of 2 test methods to determine the percentage of soluble matter in Akucell X181 polymer
4. QA3174 Investigation of the fluid absorption capacity of pre-mixed hydrogel wound dressing
5. QA3390 Rheological Evaluation of Intrasite Gel
6. The Application of Chemometric Techniques to Products with Absorptive Properties, J.R.Moffatt, (1st year report)
7. Chemometrics: A Textbook, Volume 2, D.L.Massart, *et al* Elsevier press, 1988

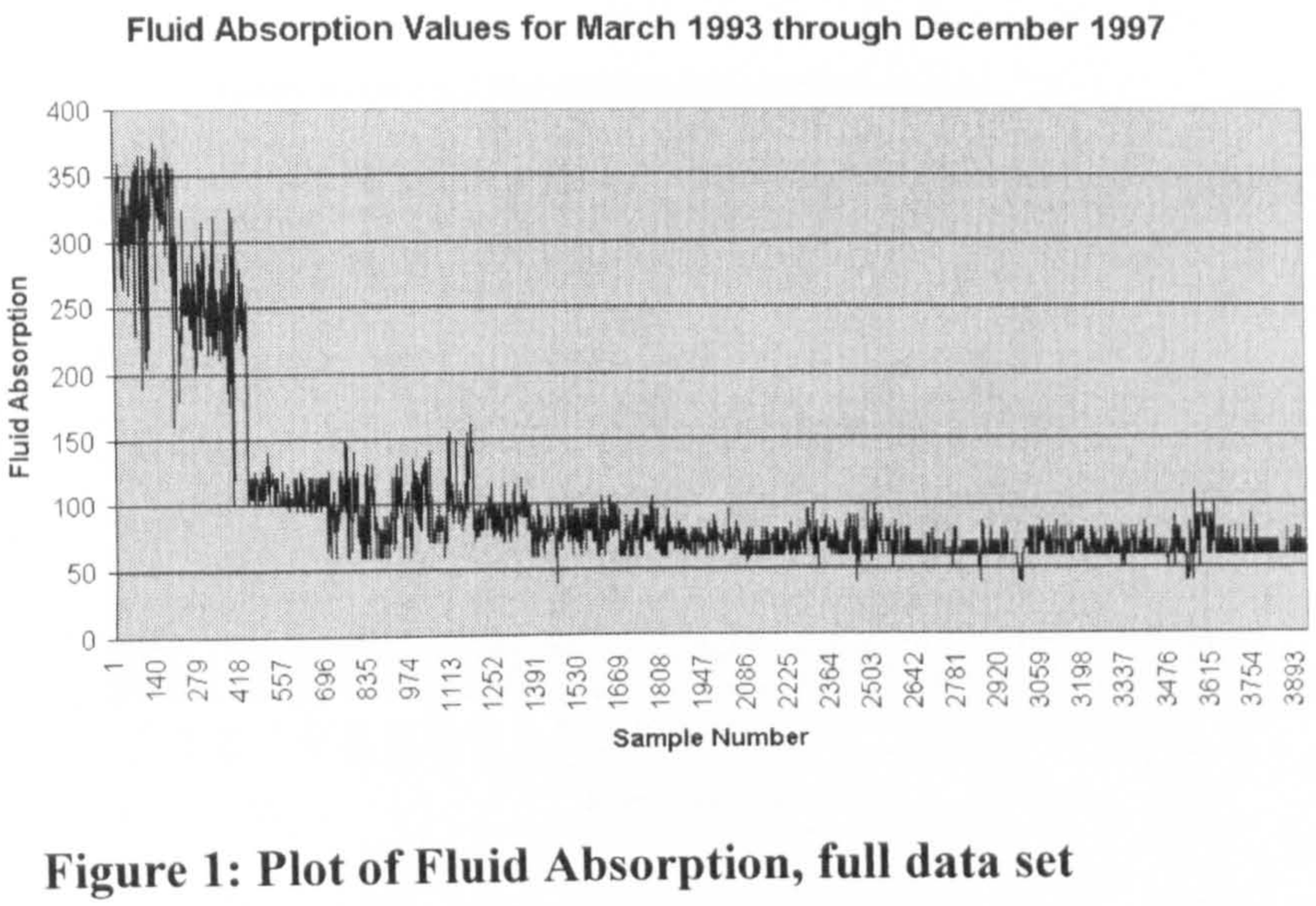


Figure 1: Plot of Fluid Absorption, full data set

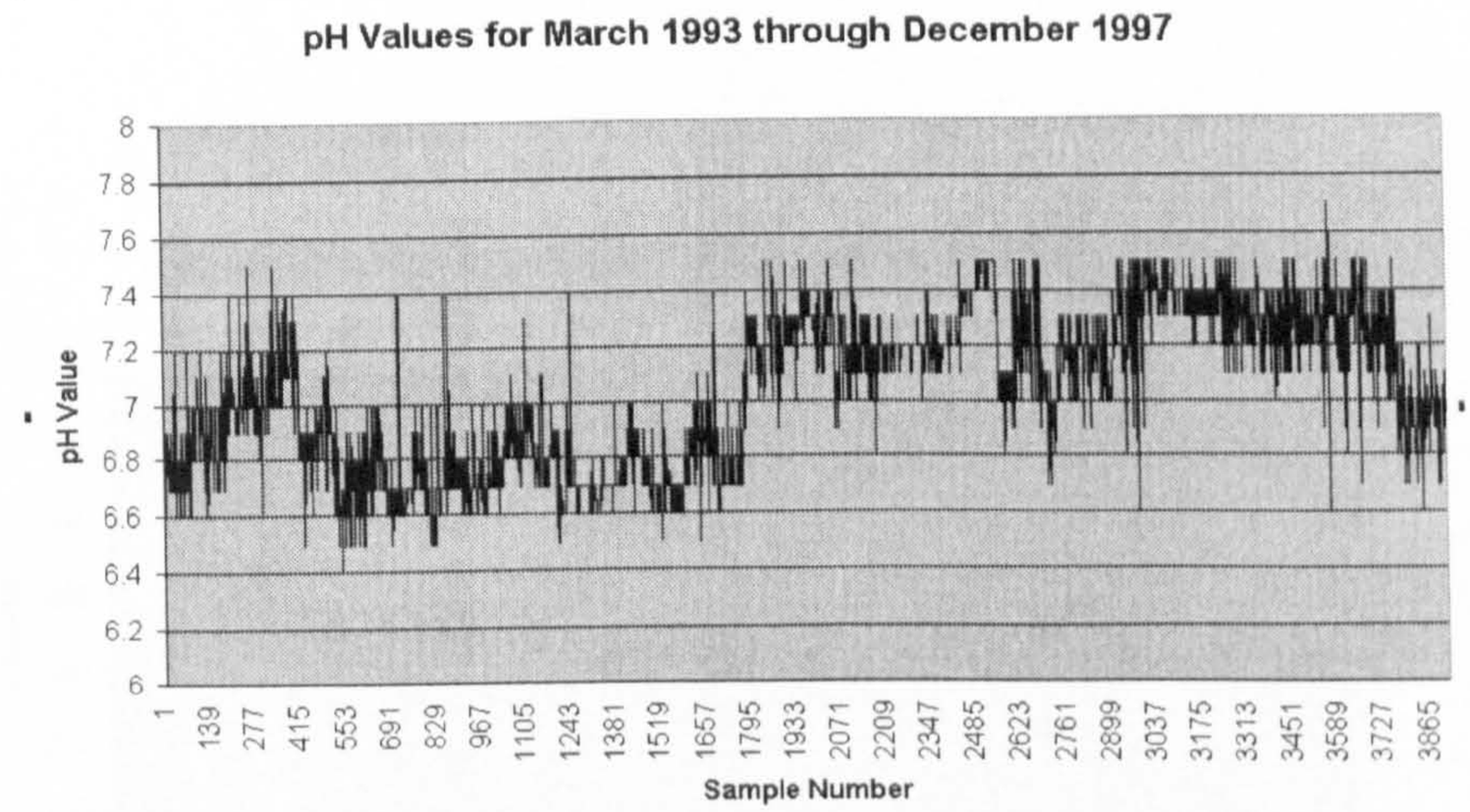


Figure 2: Plot of pH values, full data set

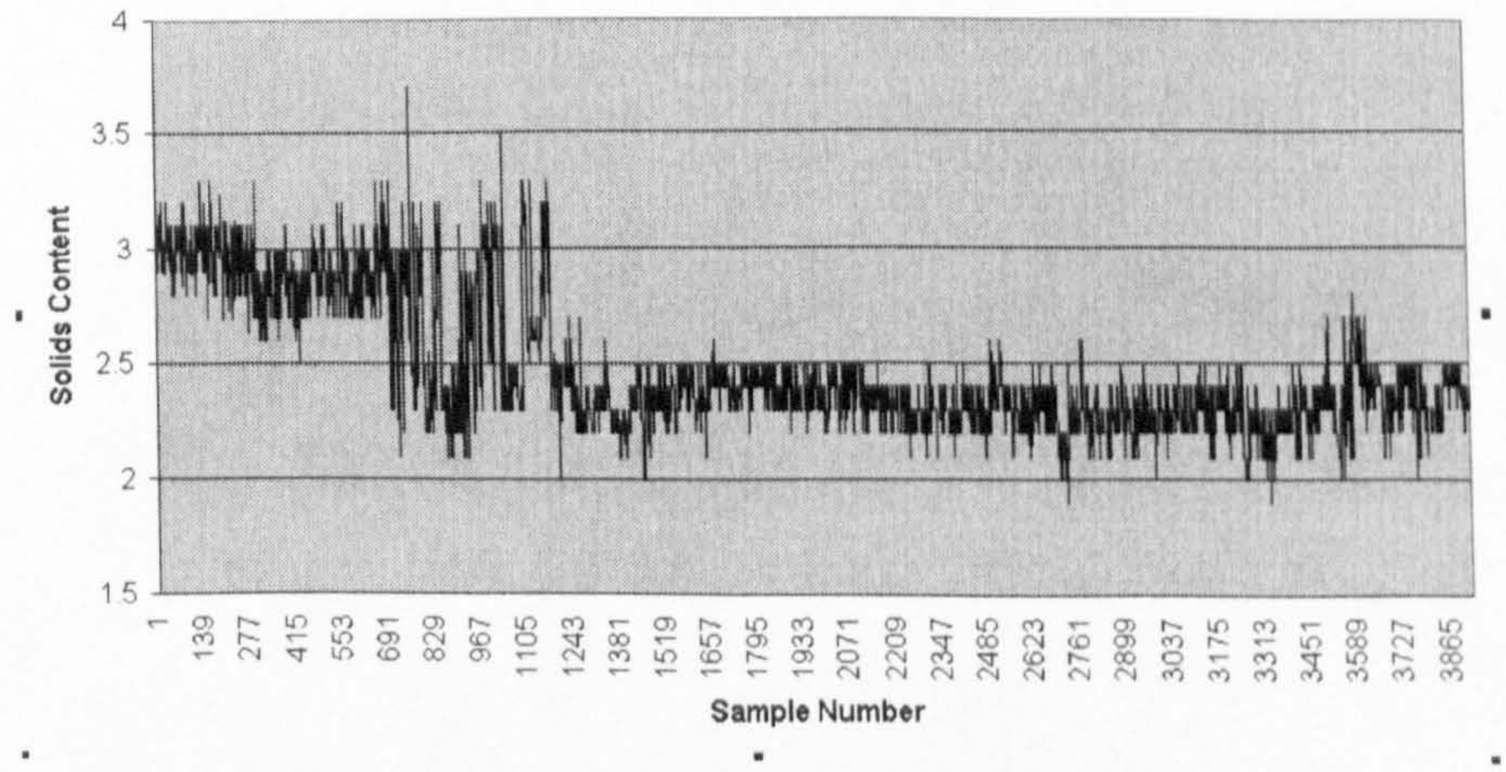


Figure 3: Plot of Solids Content Values, full data set

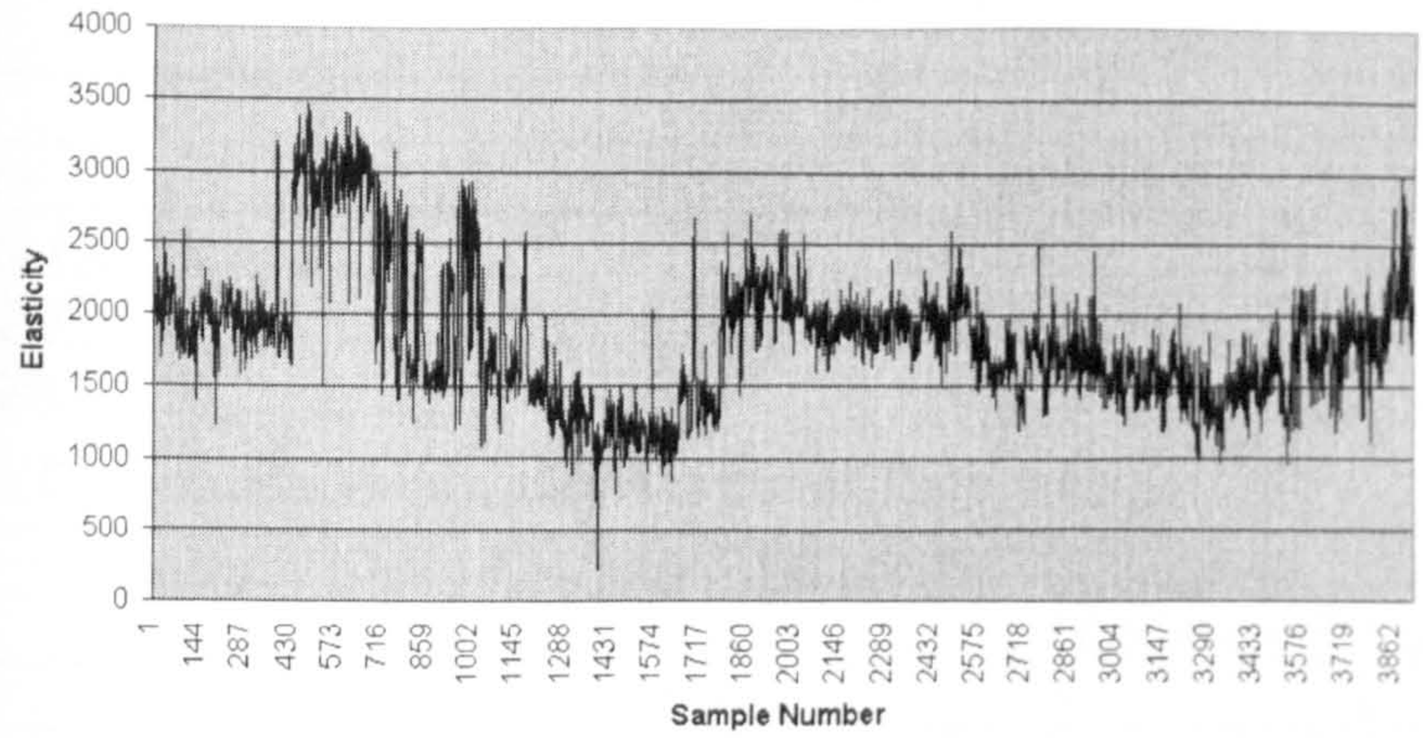


Figure 4: Plot of Elasticity Values, full data set

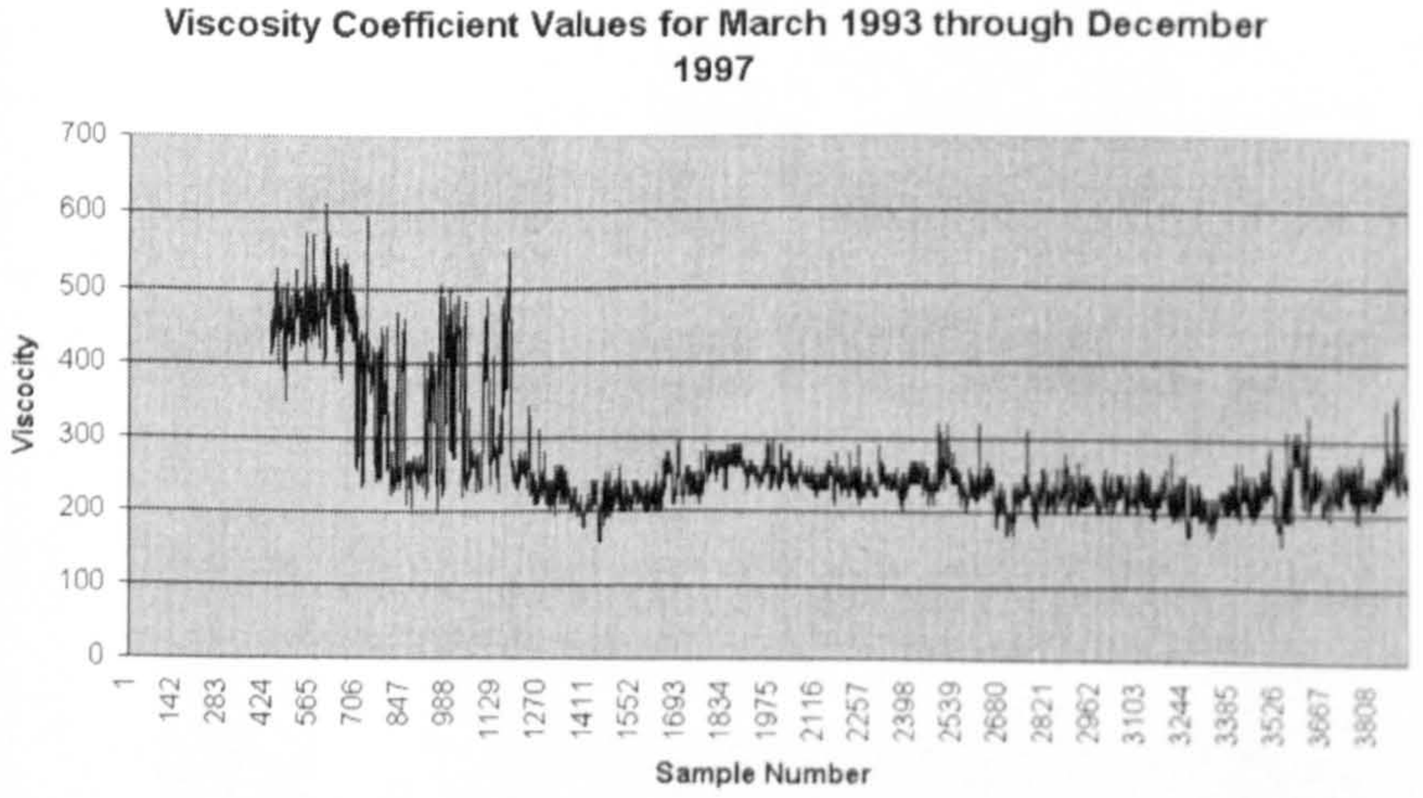


Figure 5: Plot of Viscosity Coefficients, full data set

	pH	Elasticity	Viscosity Coefficient	Solids Content	Fluid Absorption
pH	1.00				
Elasticity	-0.08	1.00			
Viscosity Coefficient	-0.39	0.84	1.00		
Solids Content	-0.42	0.70	0.90	1.00	
Fluid Absorption	-0.55	0.47	0.76	0.79	1.00

Table 1 : Correlation Values of full Intrasite Data set

	pH	Elasticity	Viscosity Coefficient	Solids Content	Fluid Absorption
pH	1.00				
Elasticity	-0.48	1.00			
Viscosity Coefficient	-0.30	0.88	1.00		
Solids Content	-0.15	0.72	0.85	1.00	
Fluid Absorption	0.14	0.26	0.46	0.50	1.00

Table 2: Correlation Values for Intrasite Data set, 1997

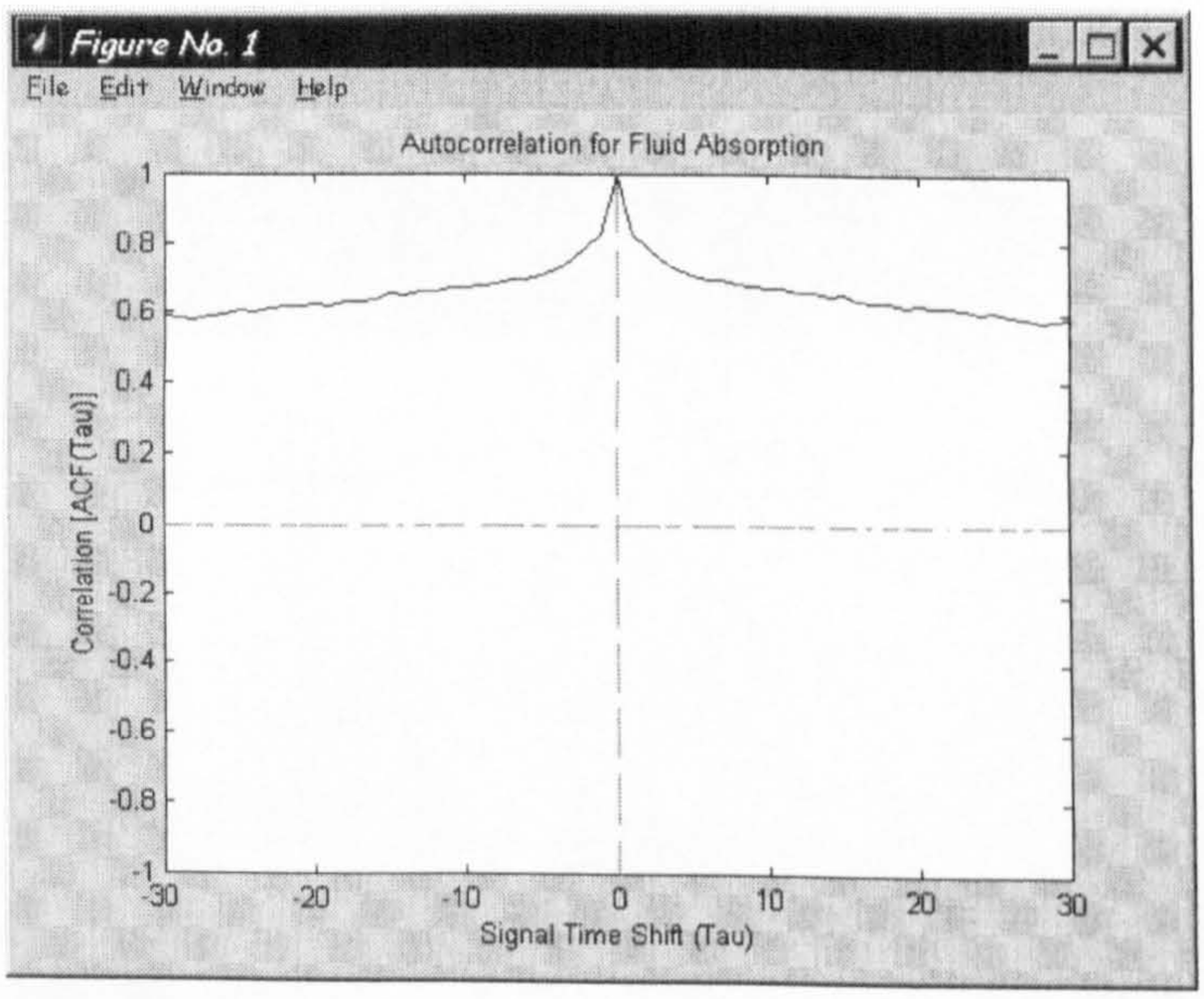


Figure 6 : Autocorrelation for Fluid

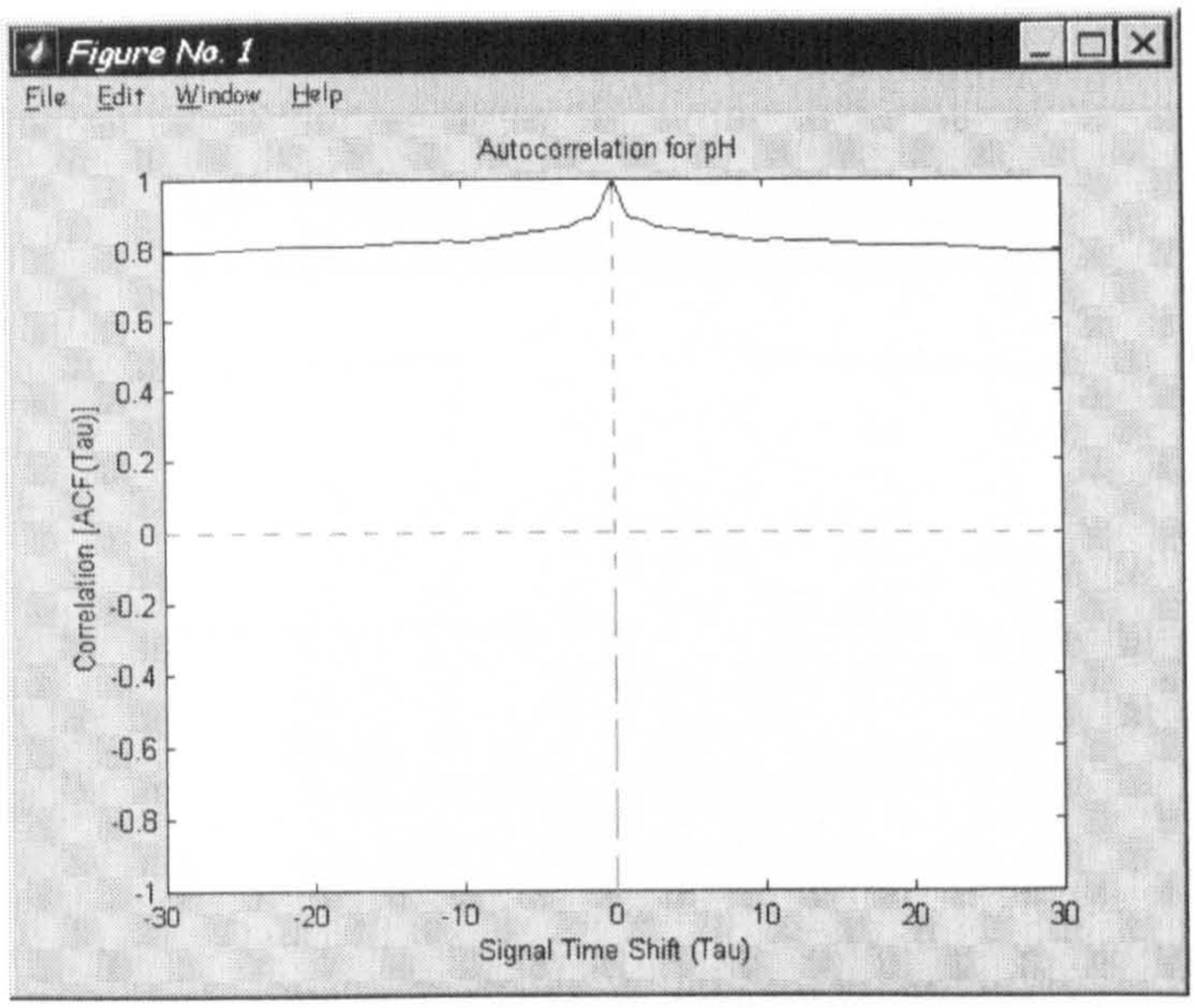


Figure 7: Autocorrelation for pH

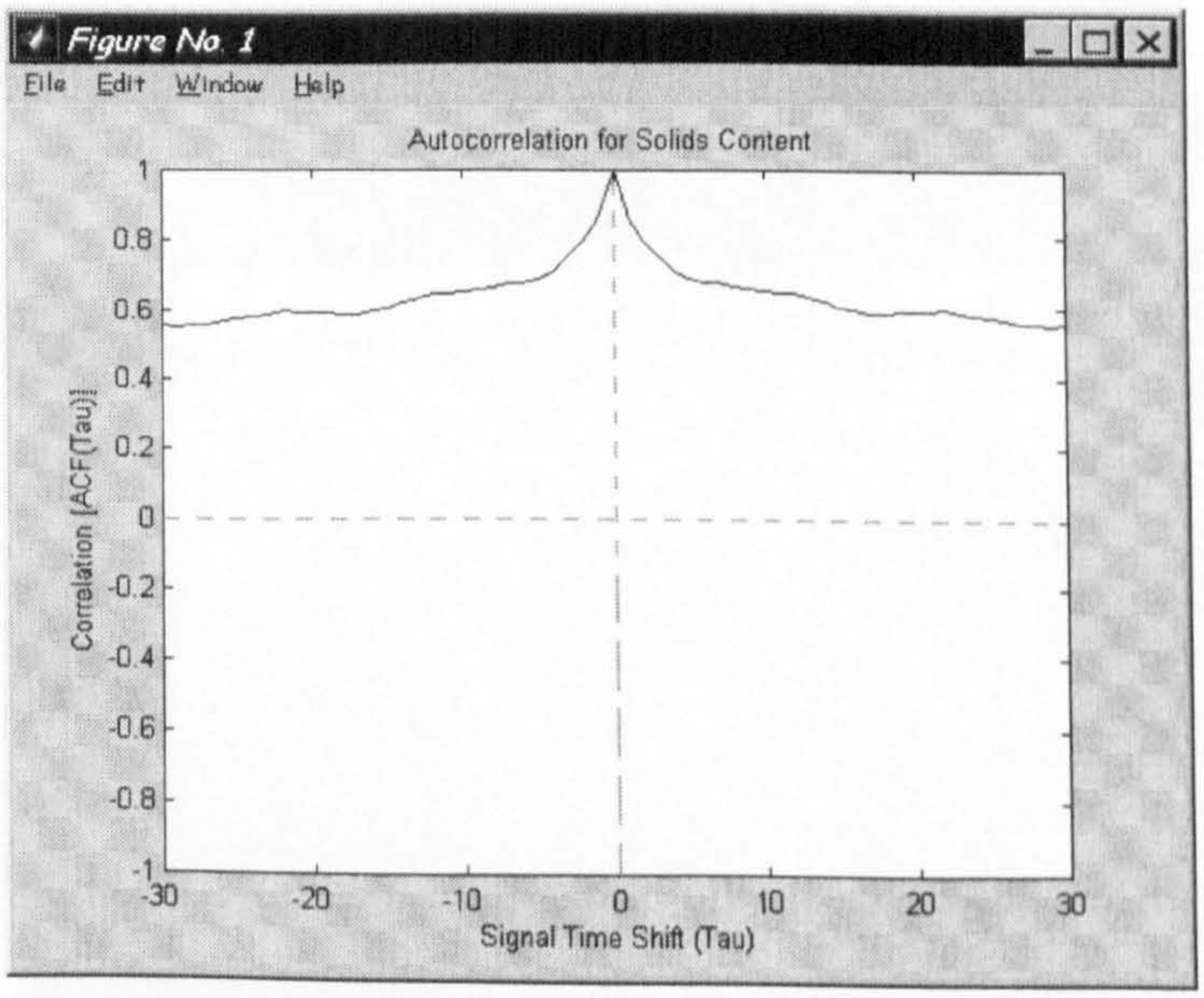


Figure 8: Autocorrelation for Solids Content

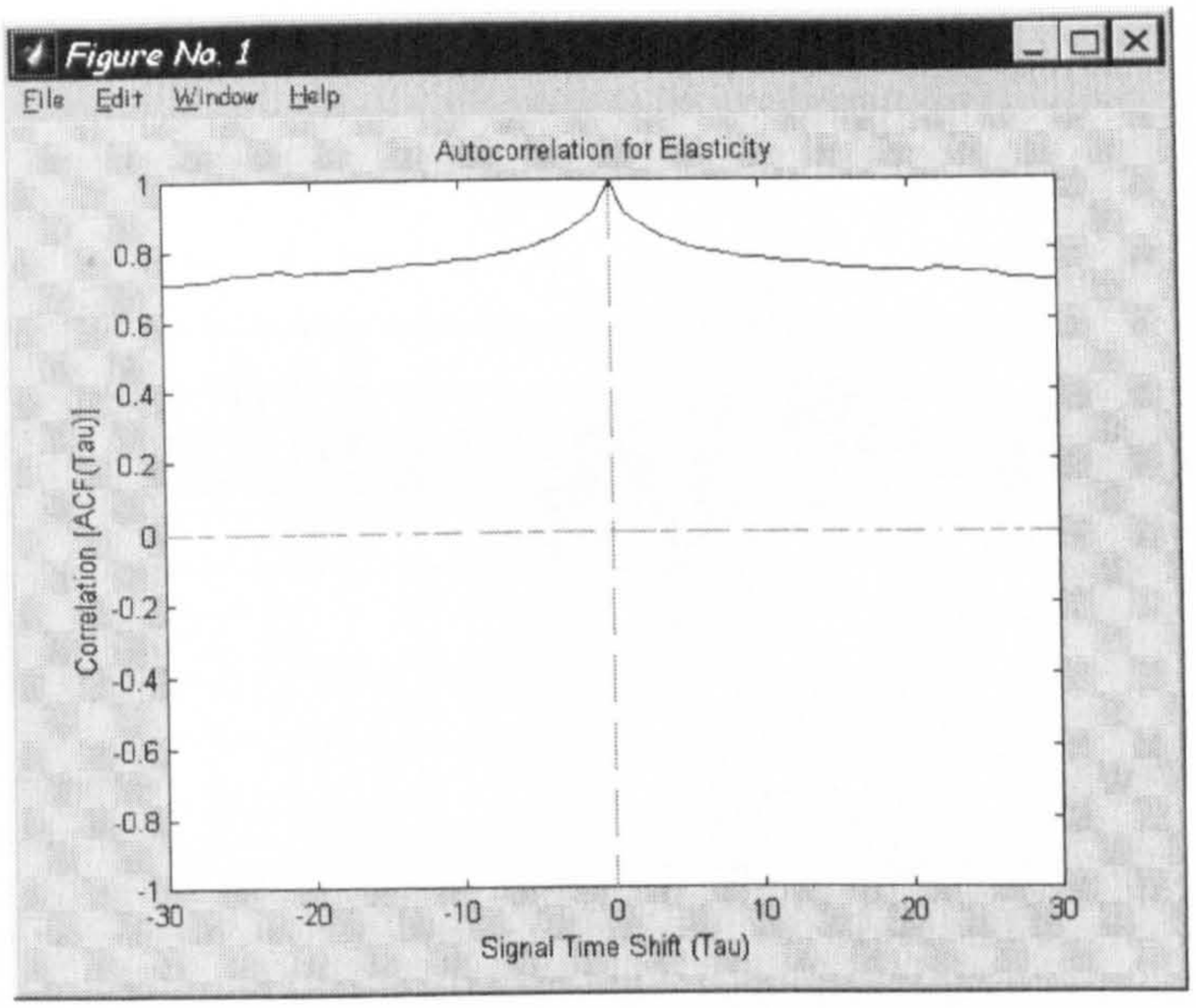


Figure 9: Autocorrelation for Elasticity

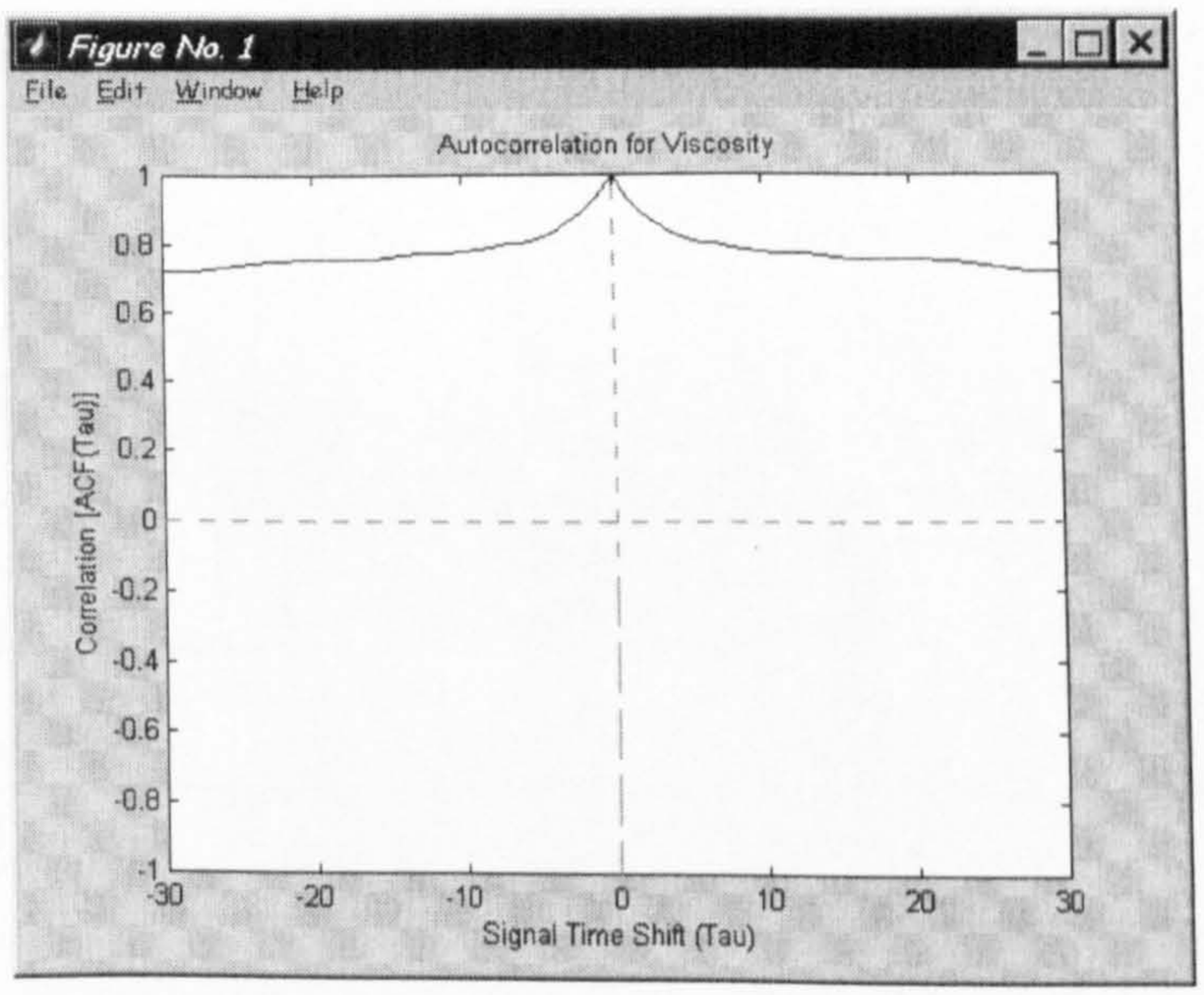


Figure 10: Autocorrelation for Viscosity

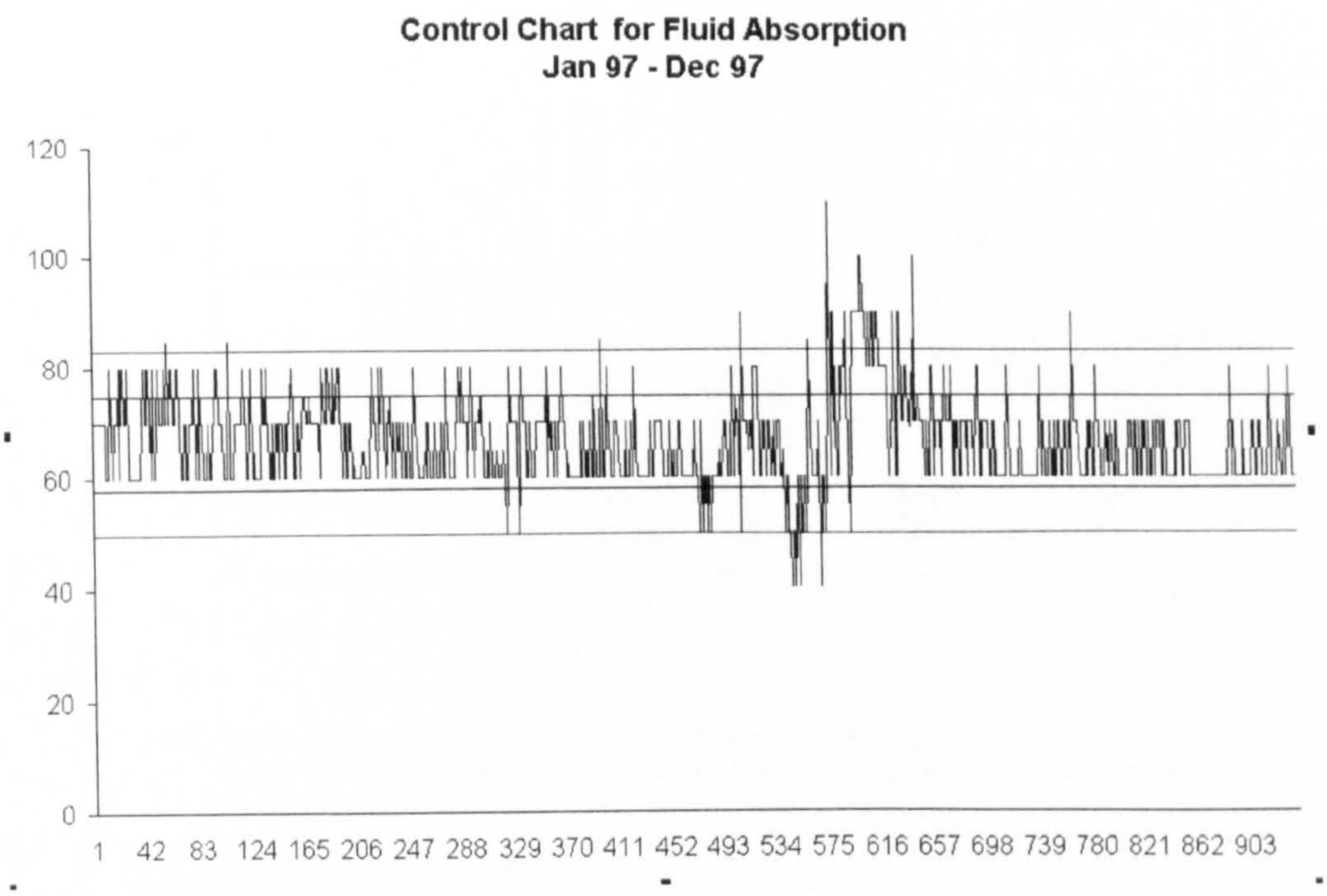


Figure 11: Control Chart for Fluid Absorption, Jan 97 - Dec 97

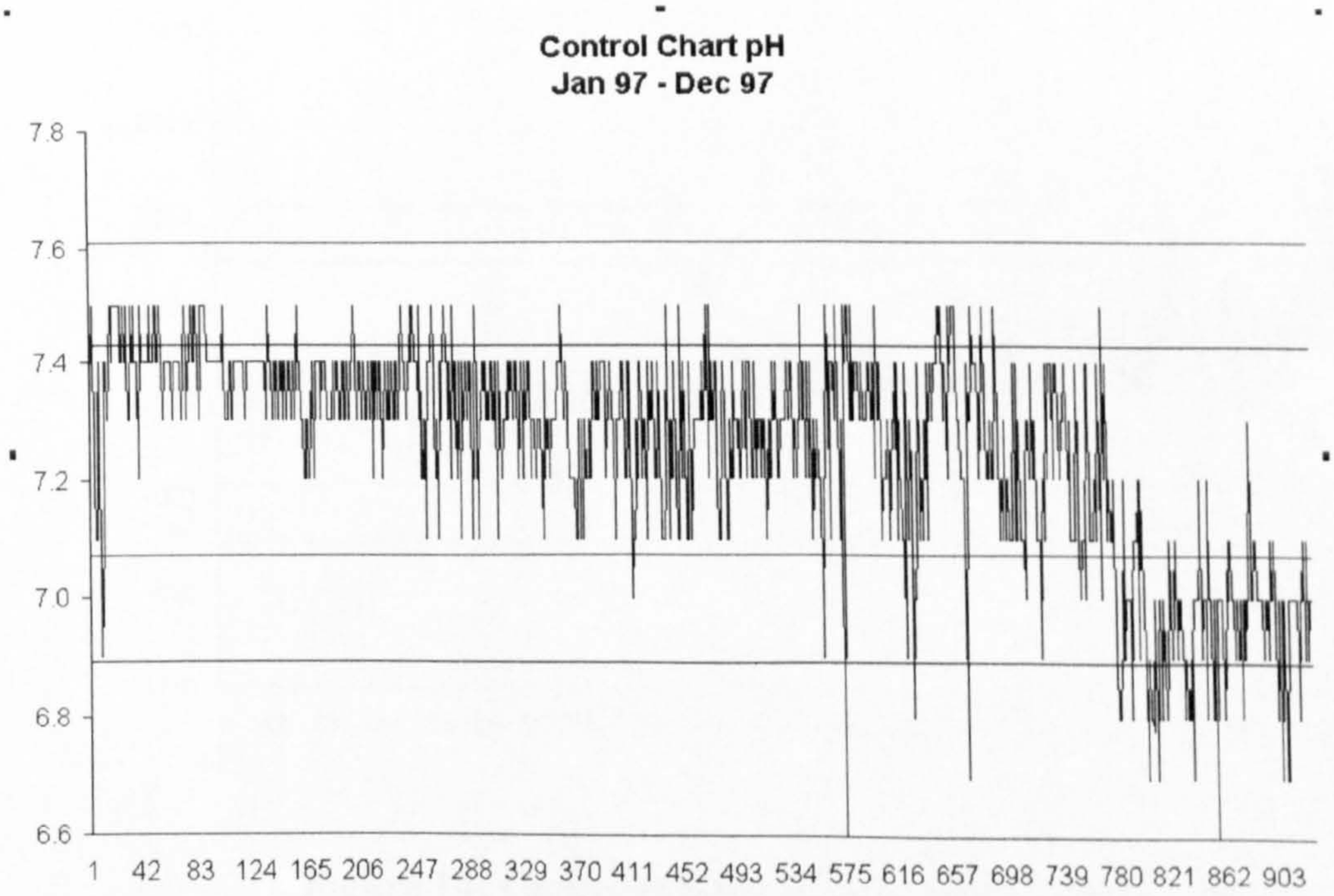


Figure 12: Control Chart for pH, Jan 97 - Dec 97

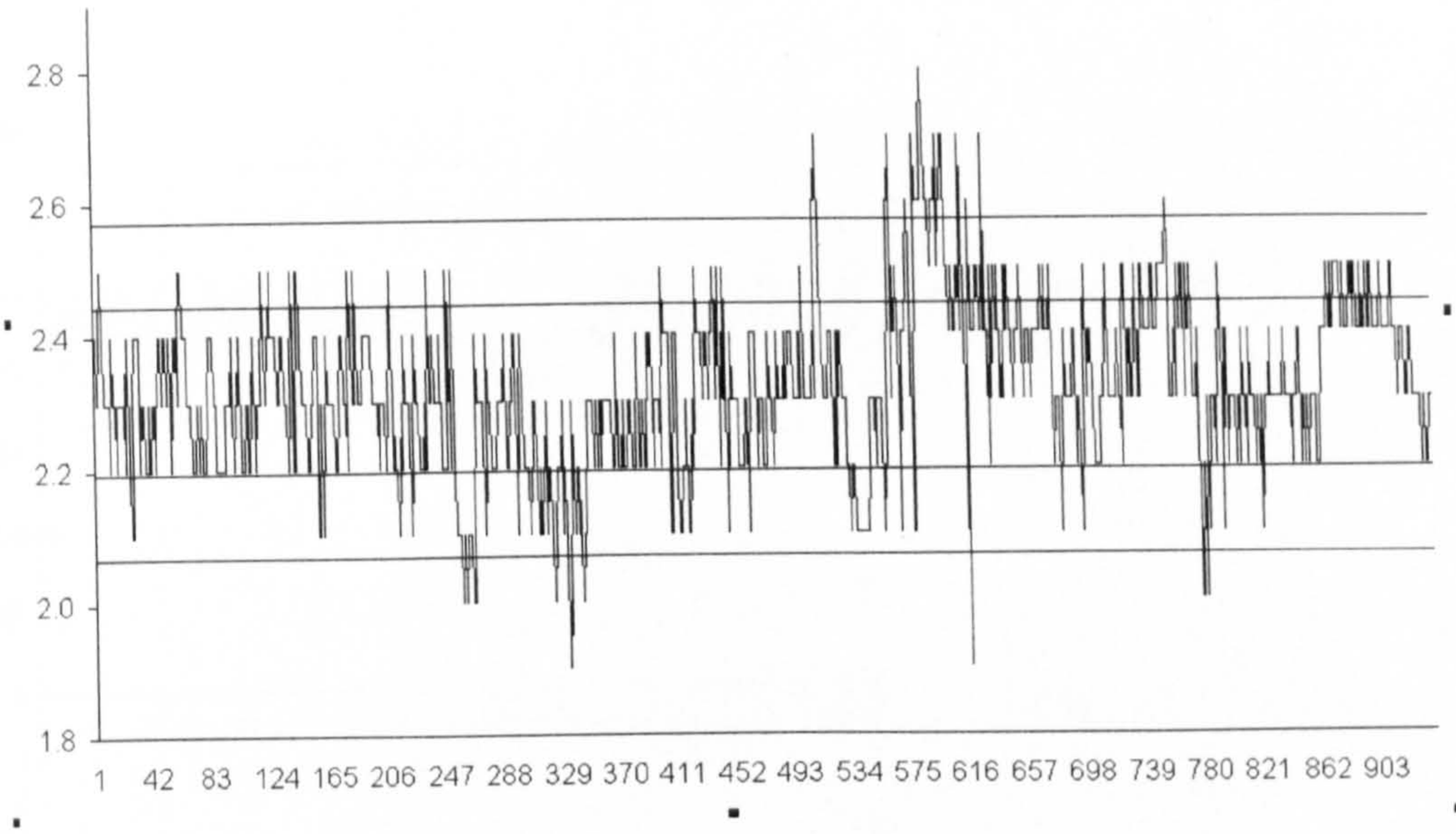


Figure 13: Control Chart for Solids Content, Jan 97 - Dec 97

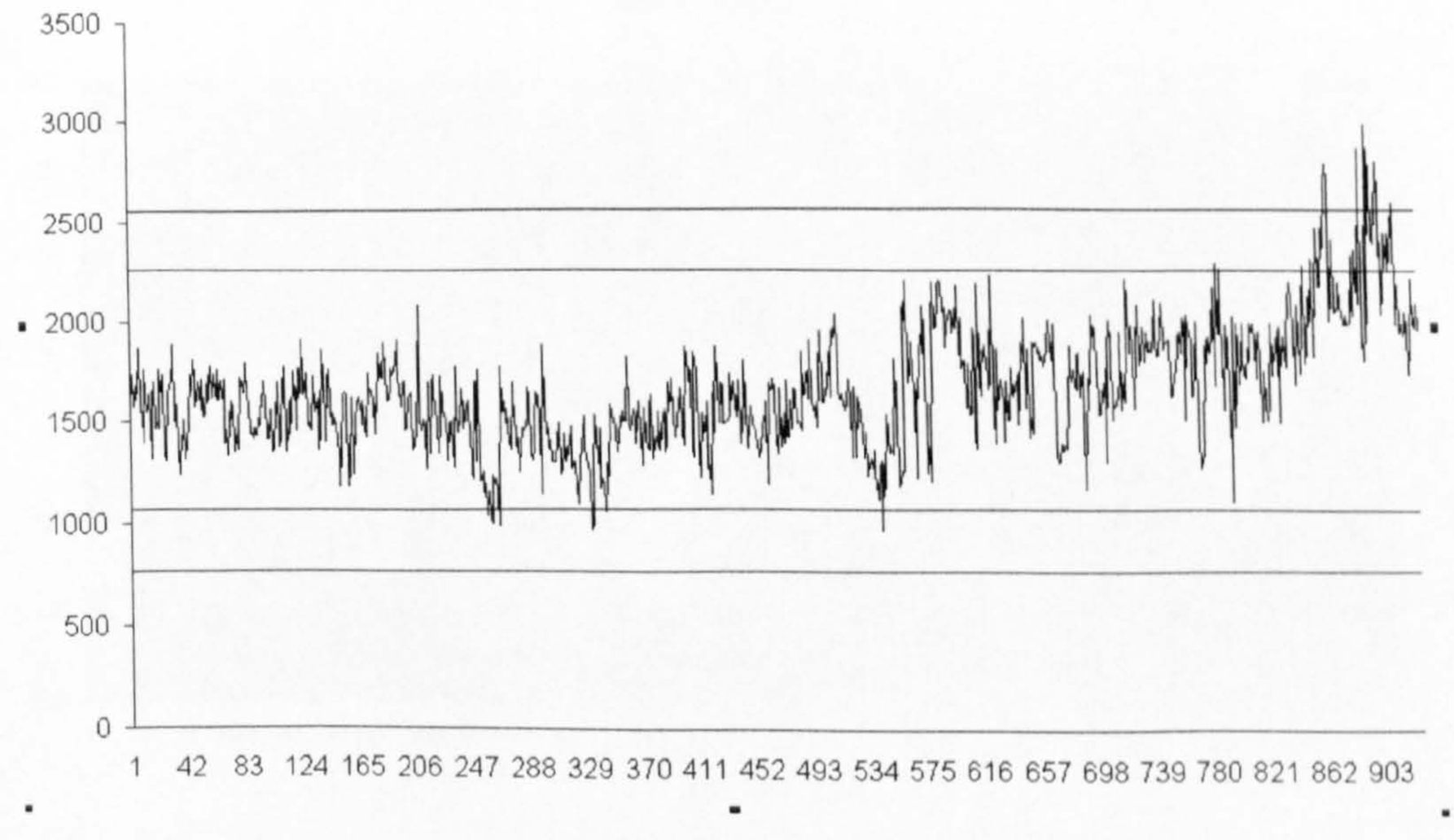


Figure 14: Control Chart for Elasticity, Jan 97 - Dec 97

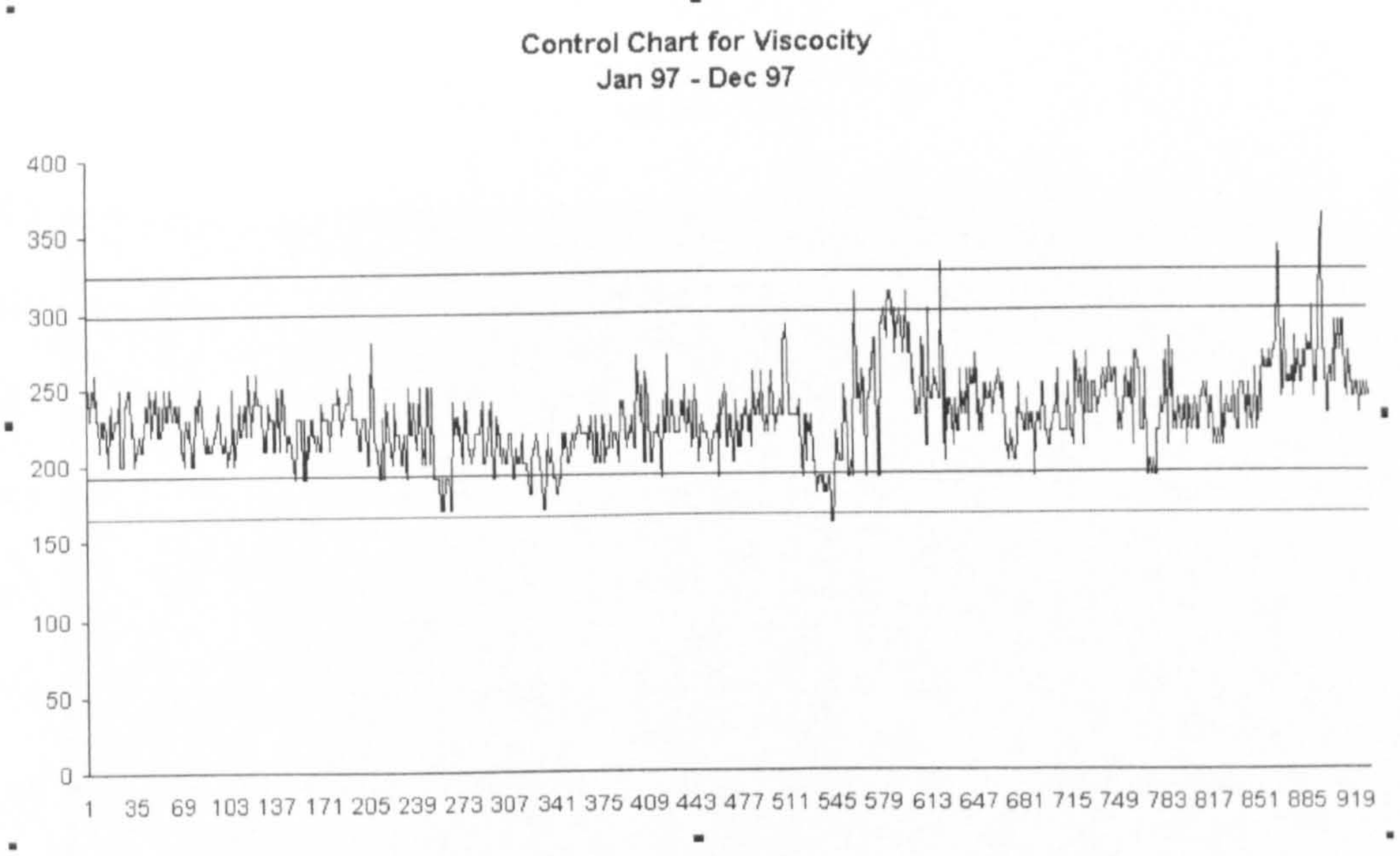


Figure 15: Control Chart for Viscosity, Jan 97 - Dec 97

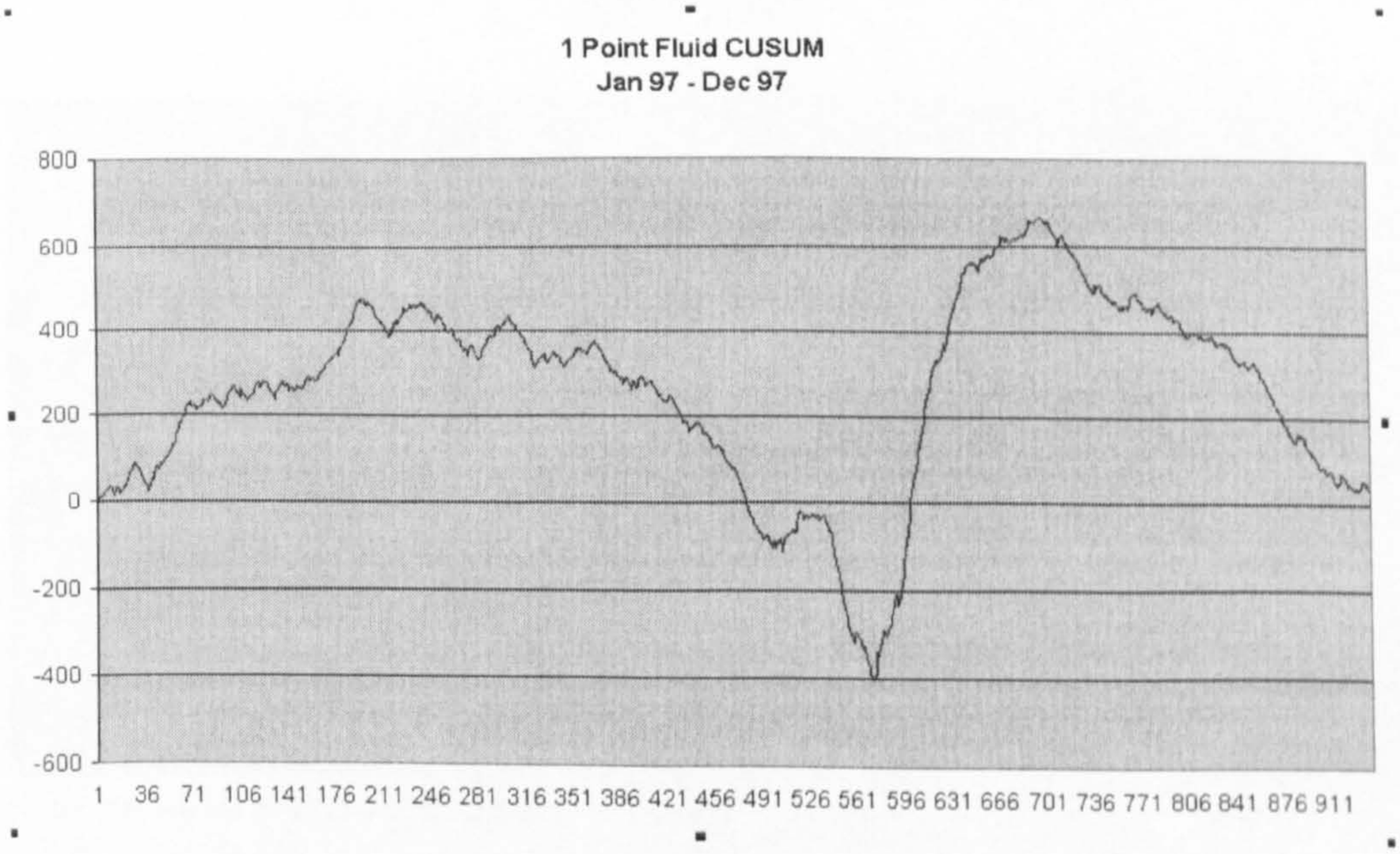


Figure 16: Cusum for Fluid Absorption, Jan 97 - Dec 97

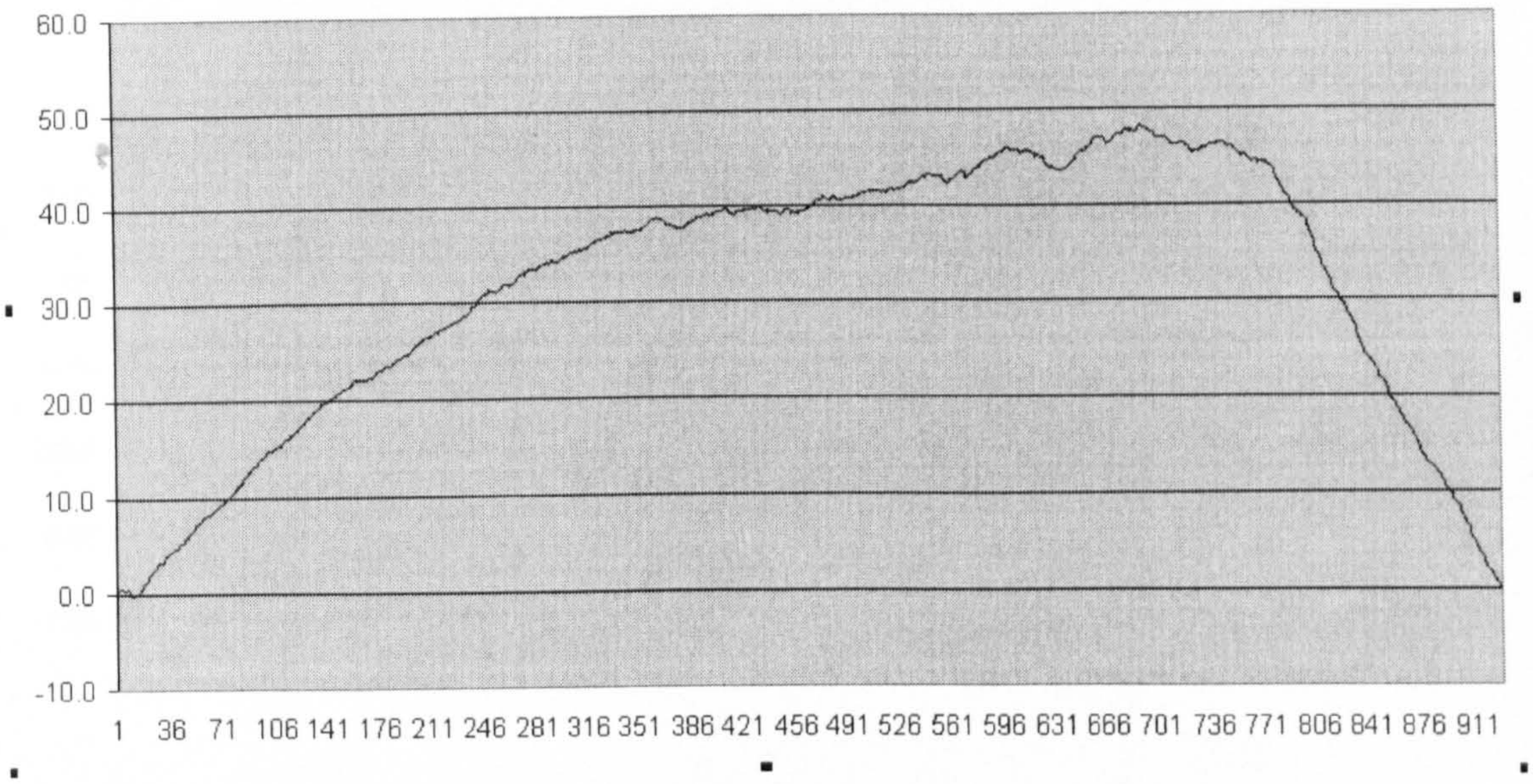


Figure 17: Cusum for pH, Jan 97 - Dec 97

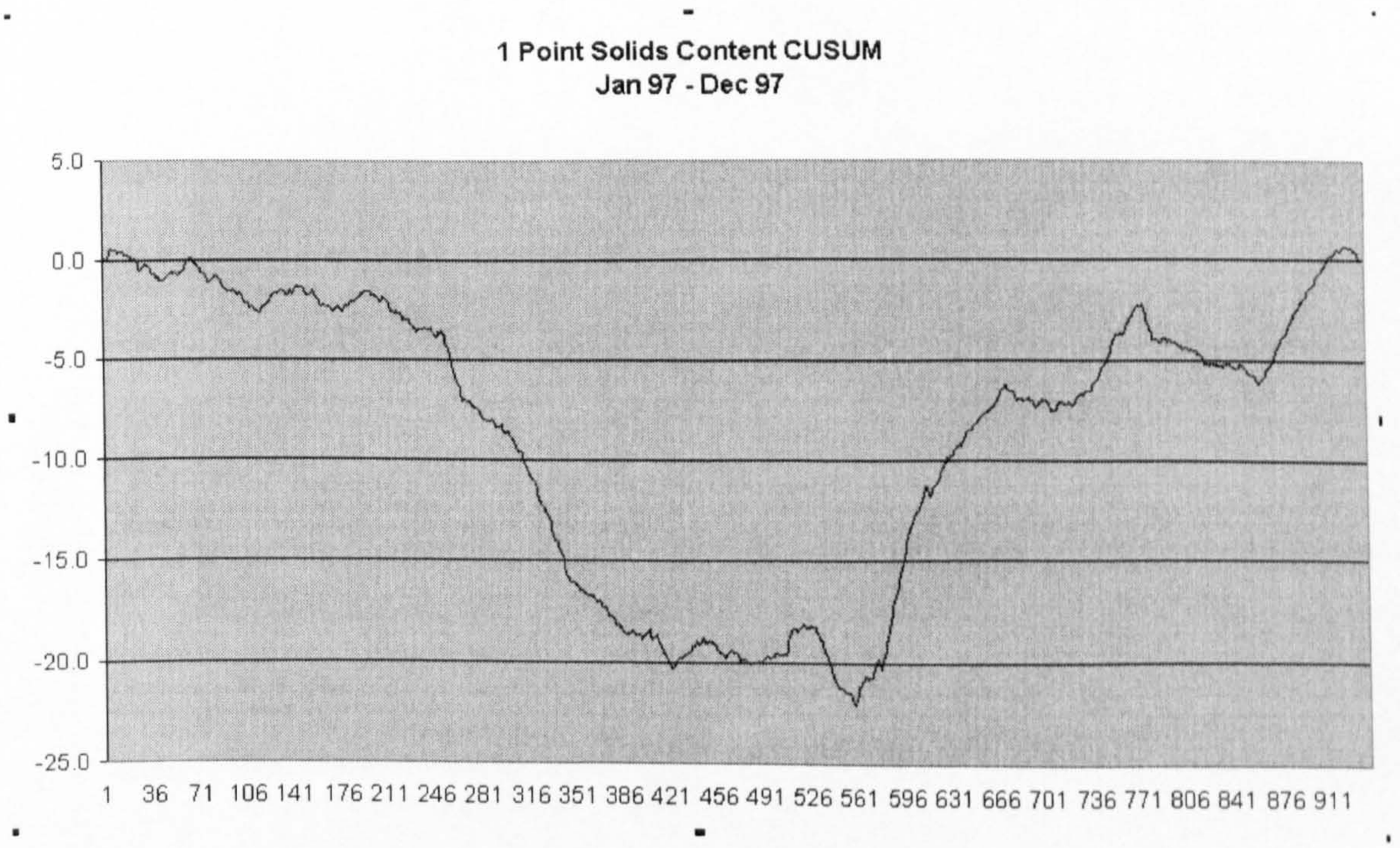


Figure 18: Cusum for Solids Content, Jan 97 - Dec 97

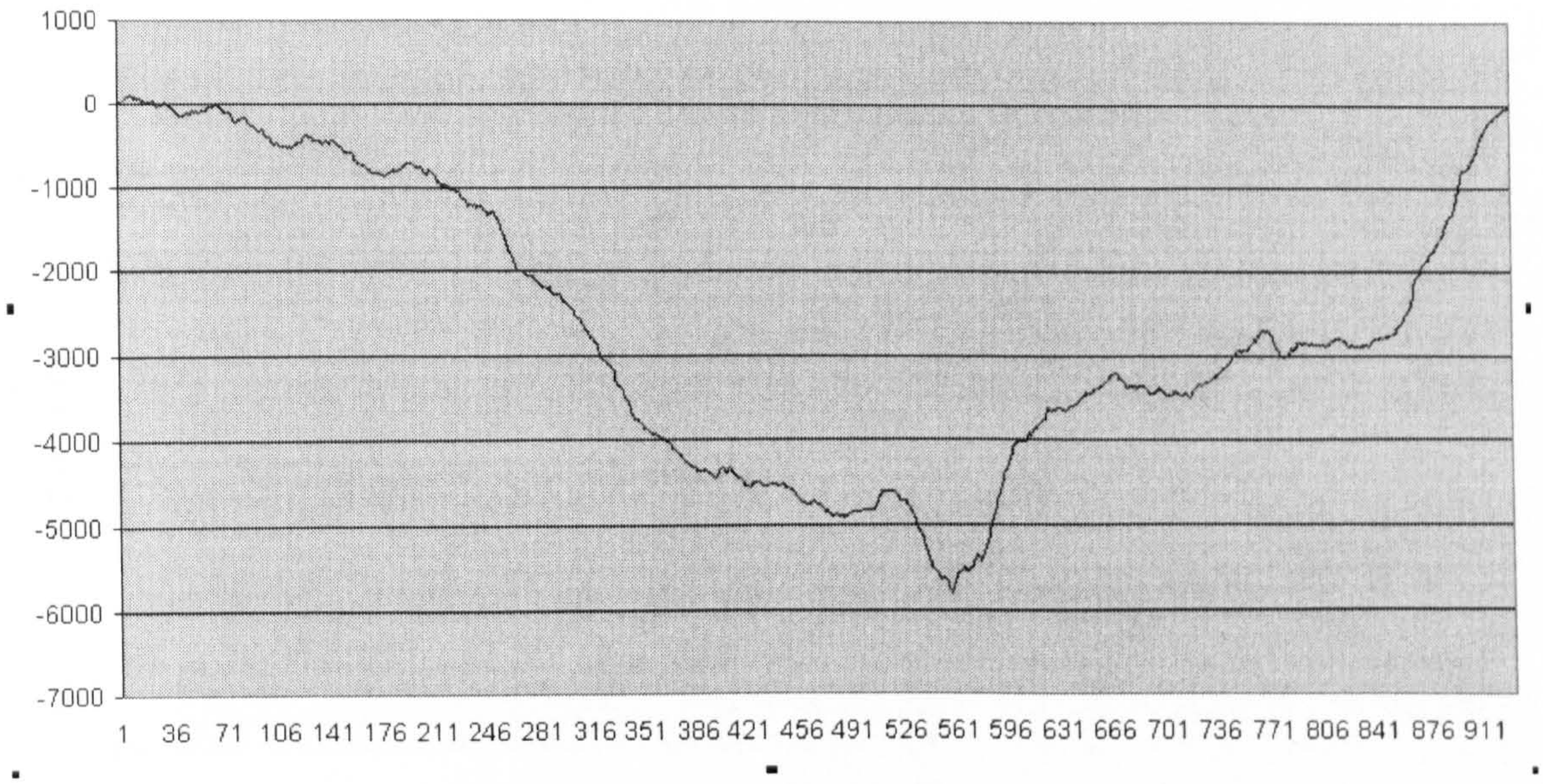


Figure 19: Cusum for Viscosity, Jan 97 - Dec 97

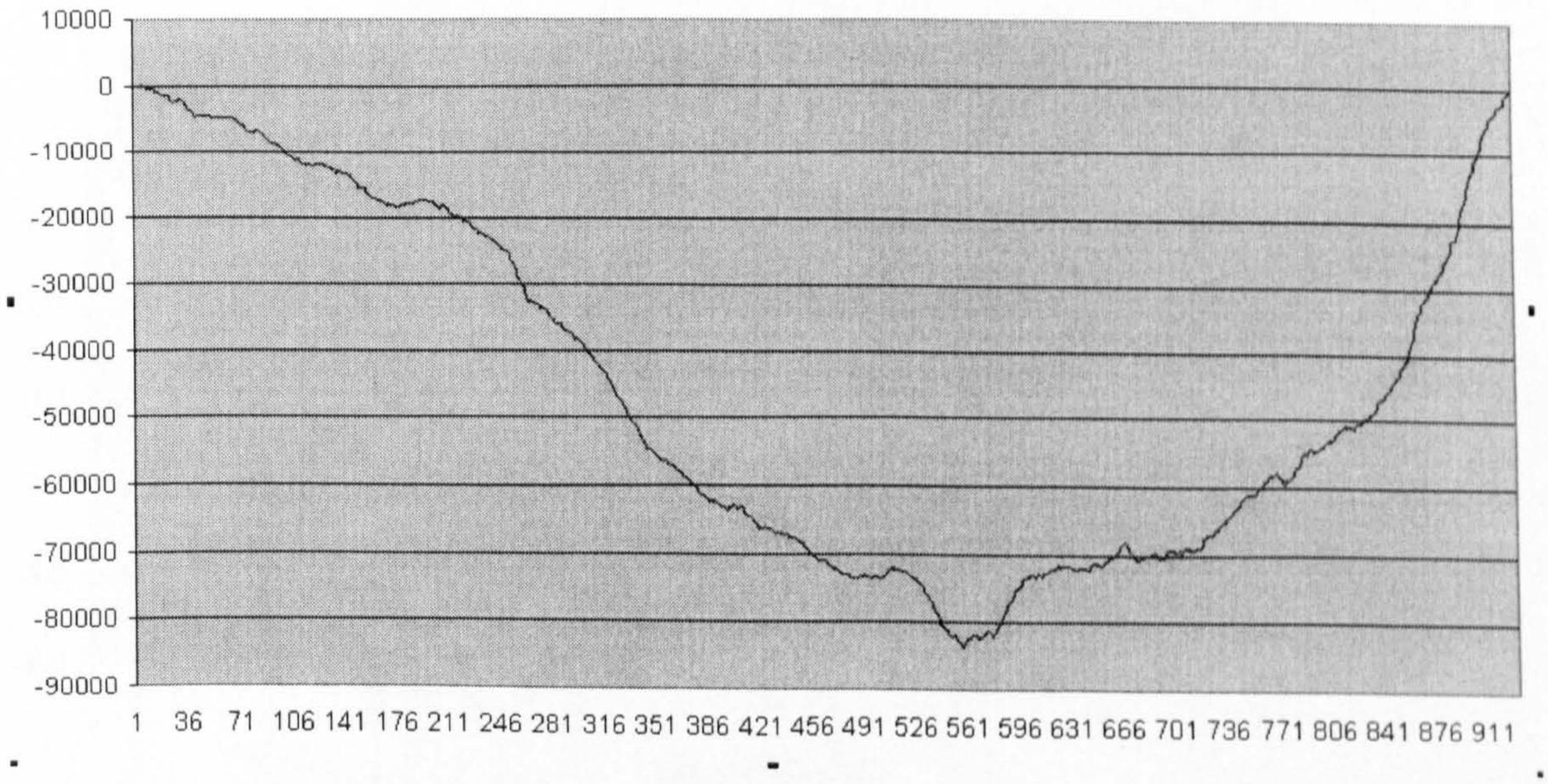


Figure 20: Cusum for Elasticity, Jan 97 - Dec 97

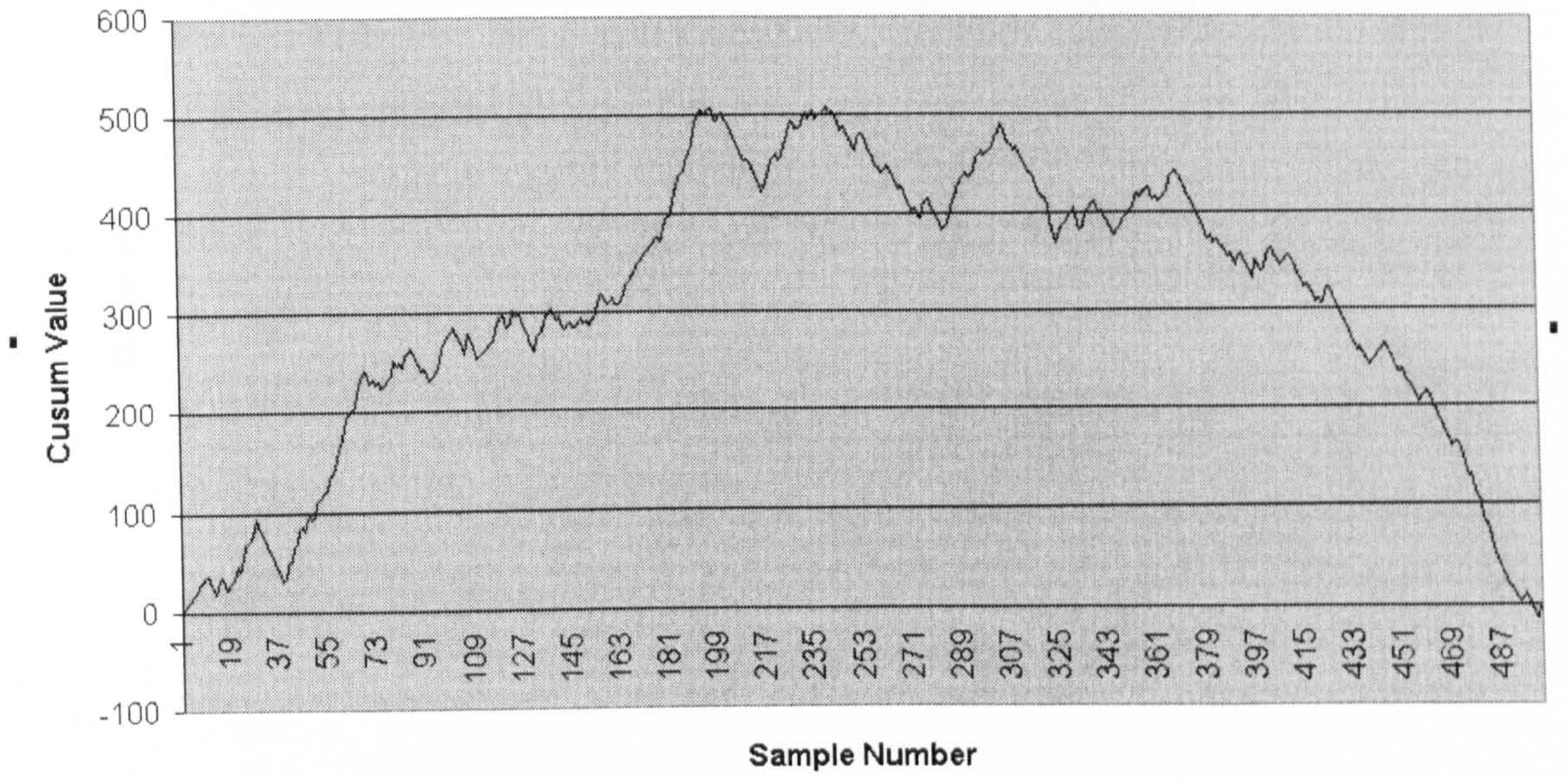


Figure 21: Cusum for Fluid Absorption, 1st half 1997

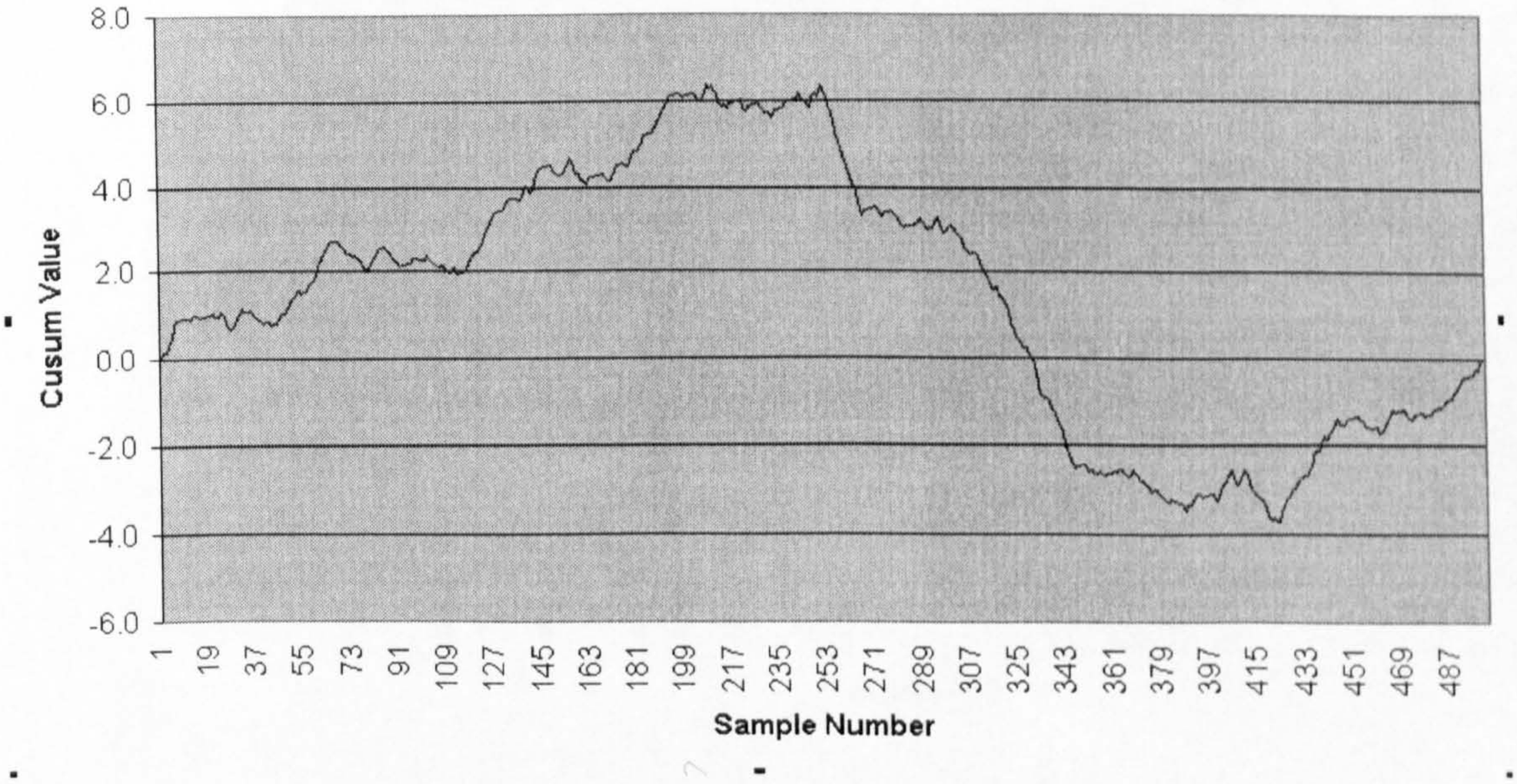


Figure 22: Cusum for Solids Content, 1st half 1997

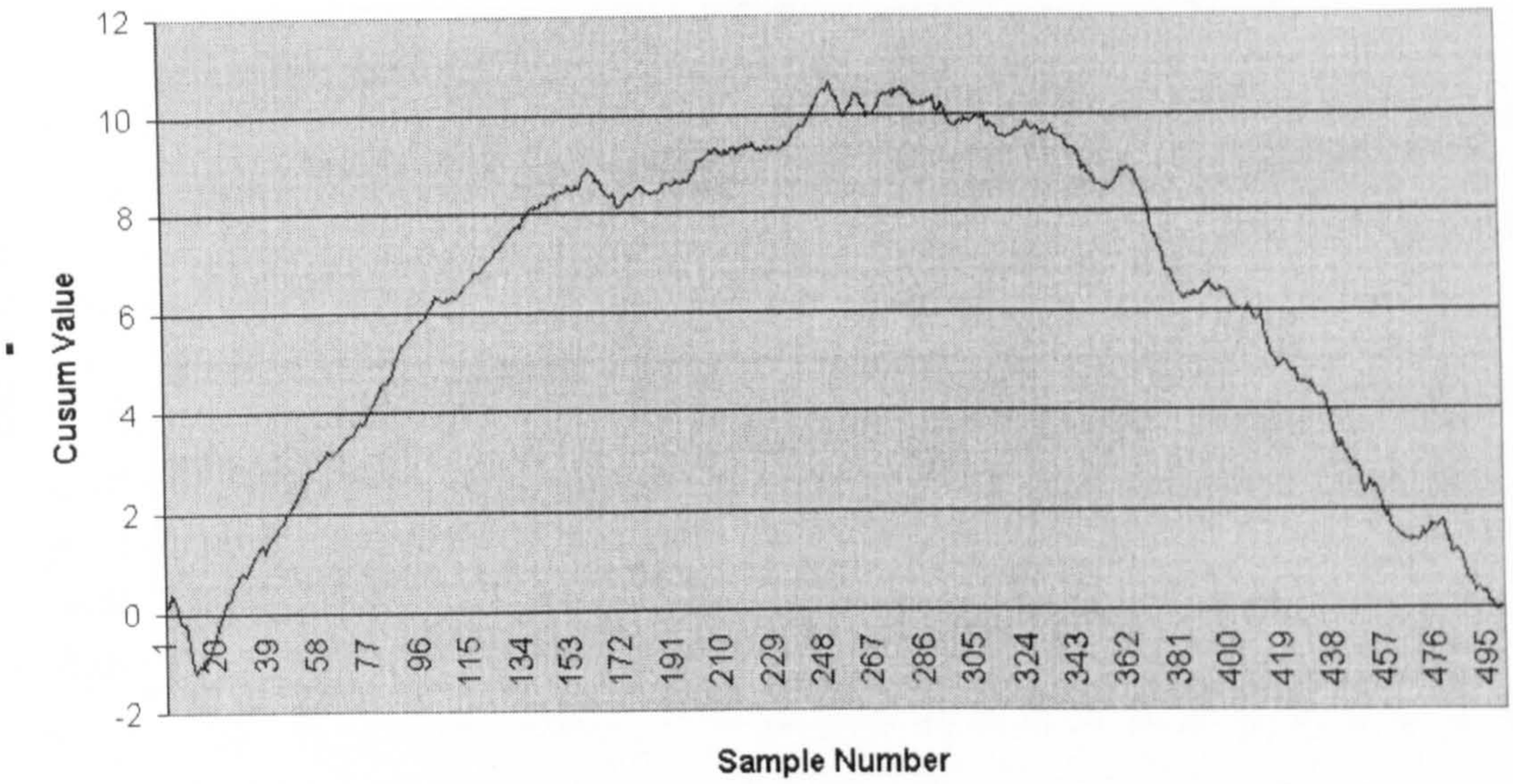


Figure 23: Cusum for pH, 1st half 1997

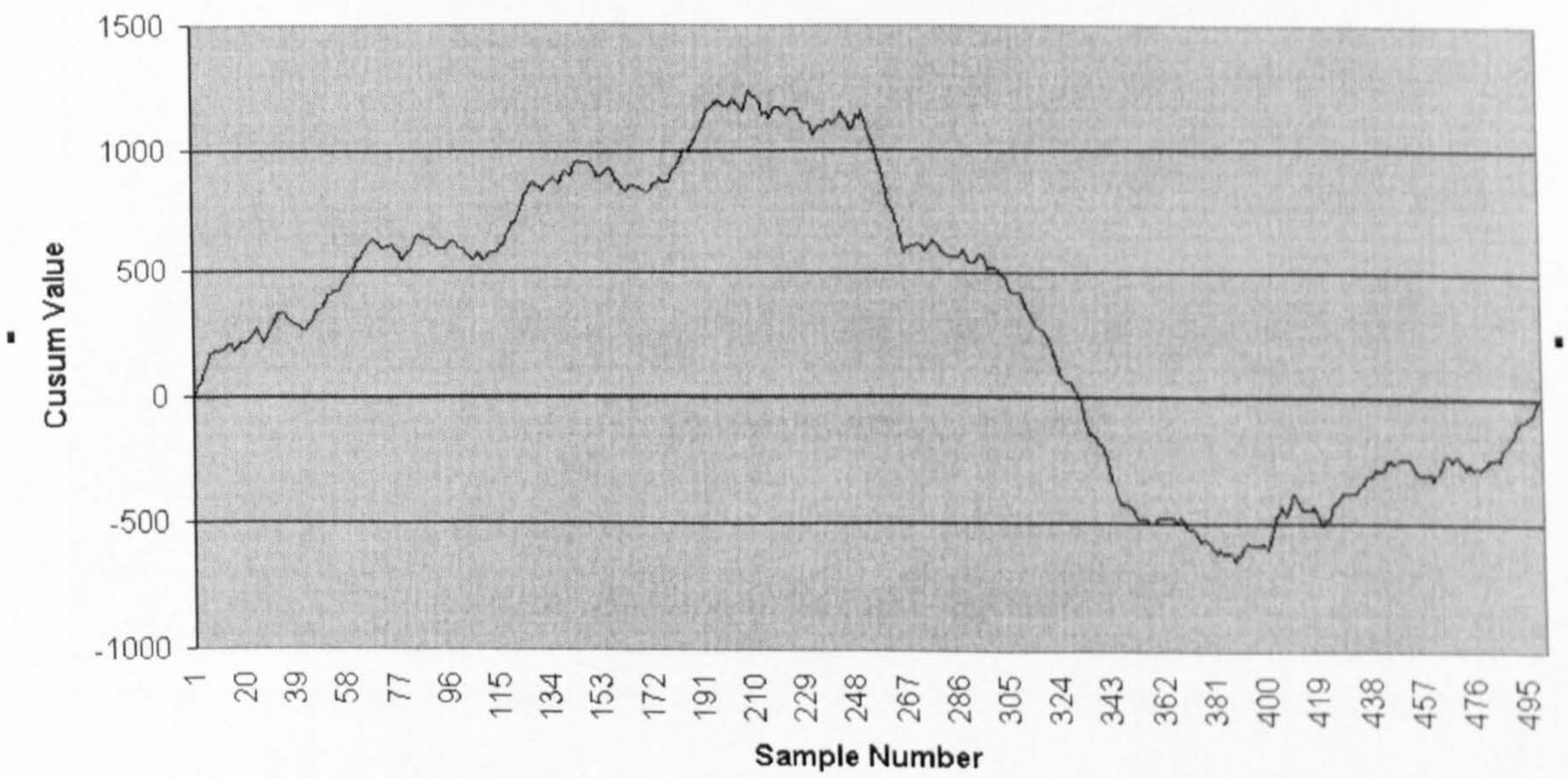


Figure 24: Cusum for Viscosity, 1st half 1997

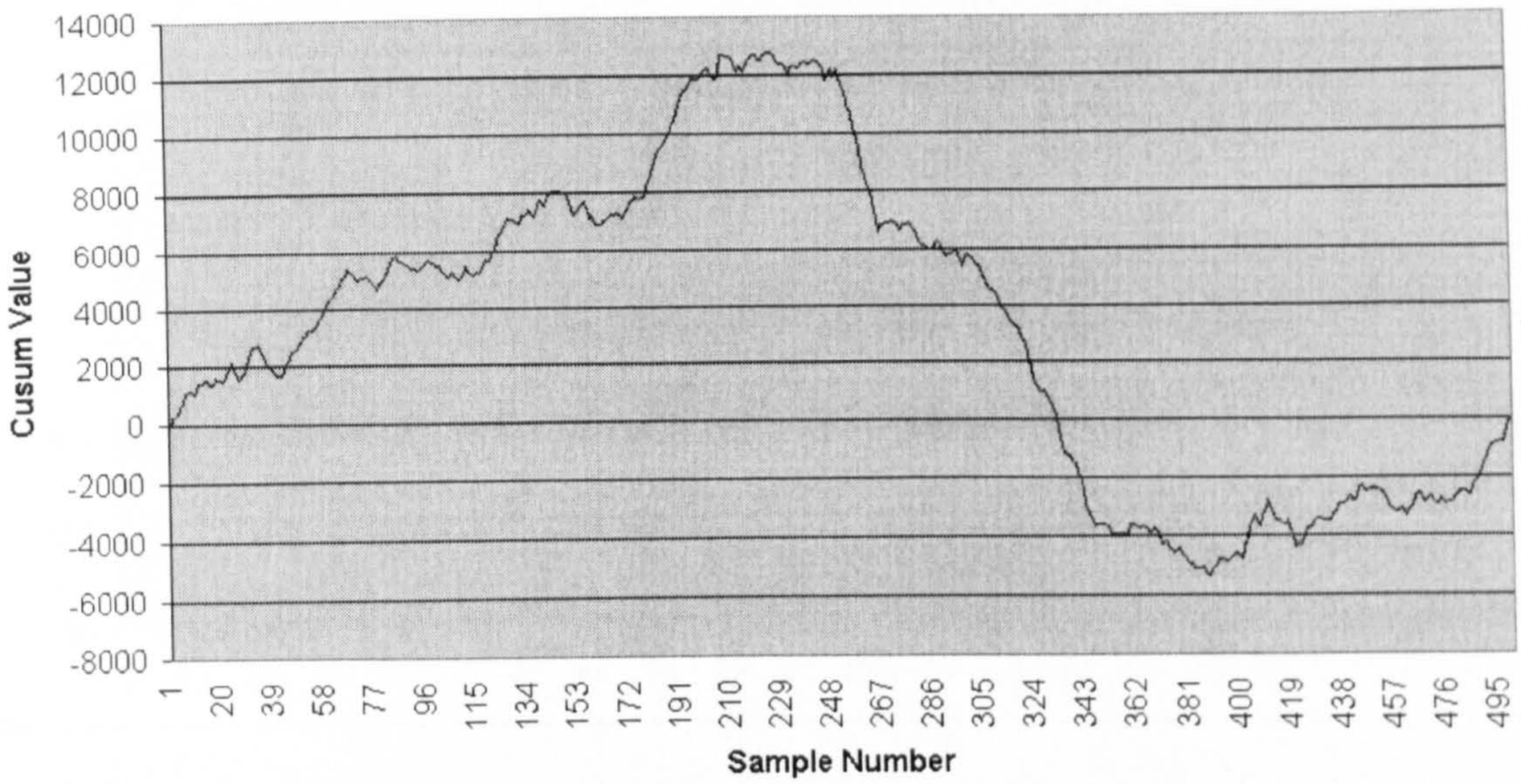


Figure 25: Cusum for Elasticity, 1st half 1997

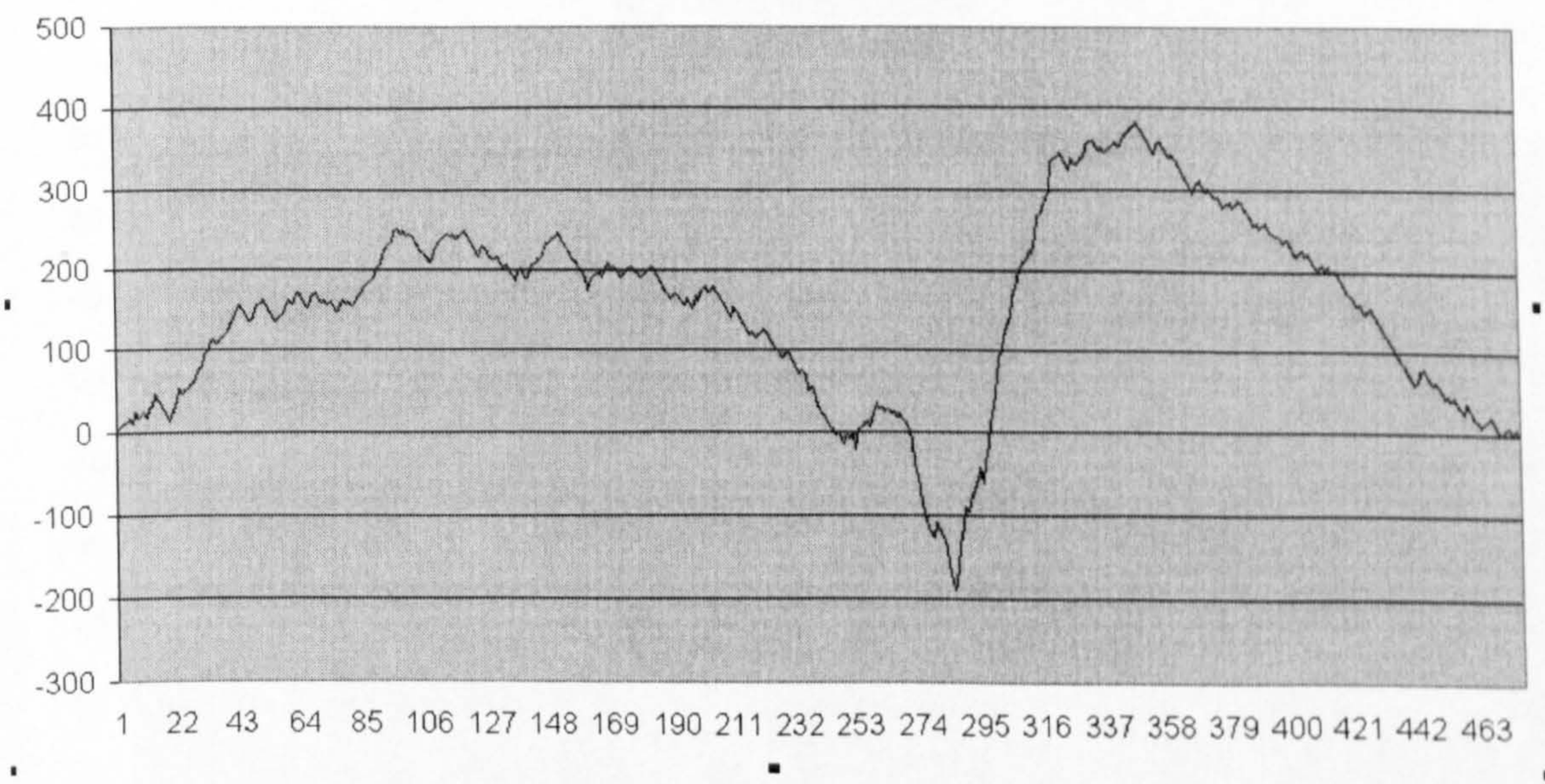


Figure 26: 2 Point Cusum for Fluid Absorption, Jan 97 - Dec 97

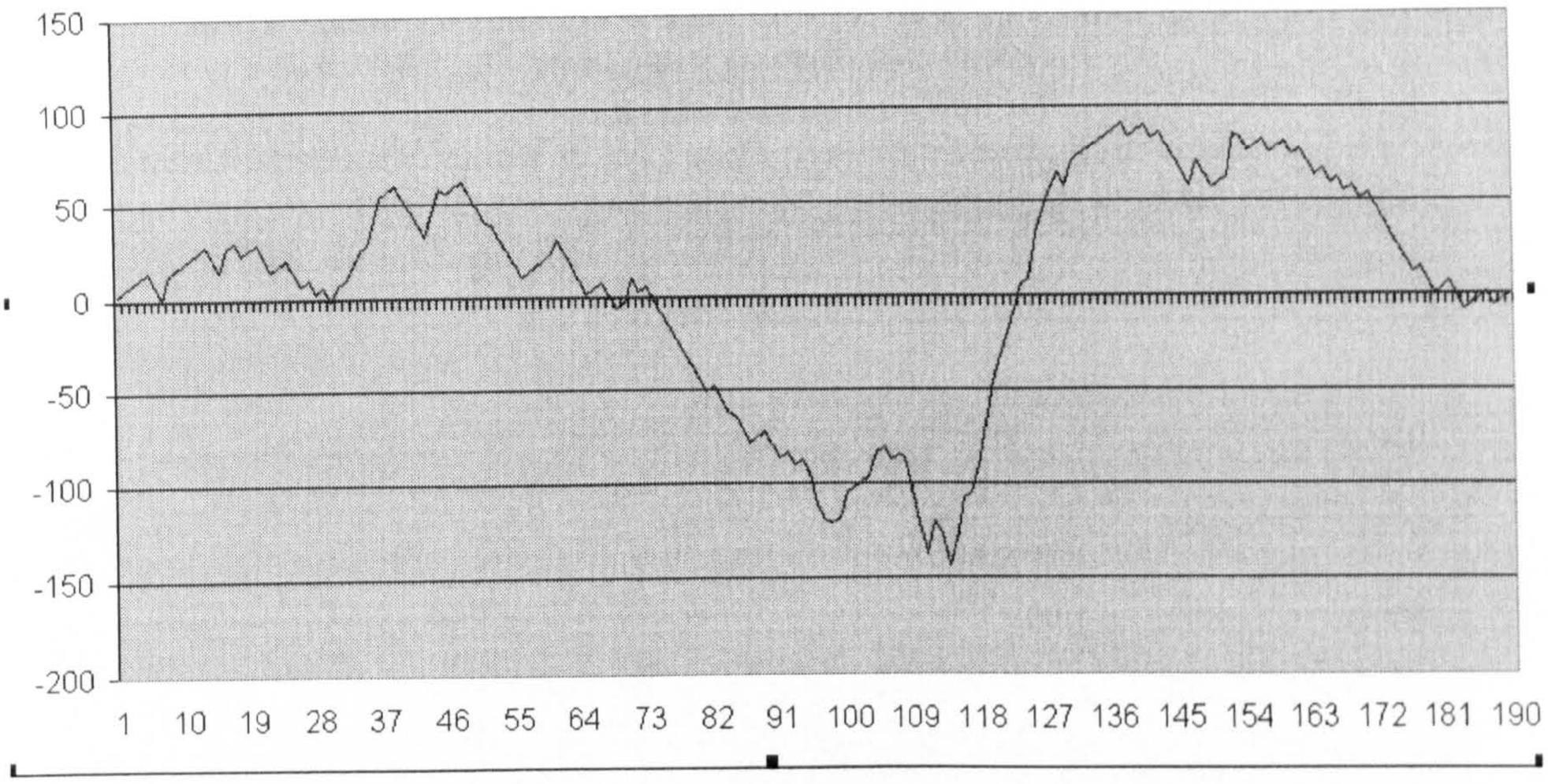


Figure 27: 5 Point Cusum for Fluid Absorption, Jan 97 - Dec 97

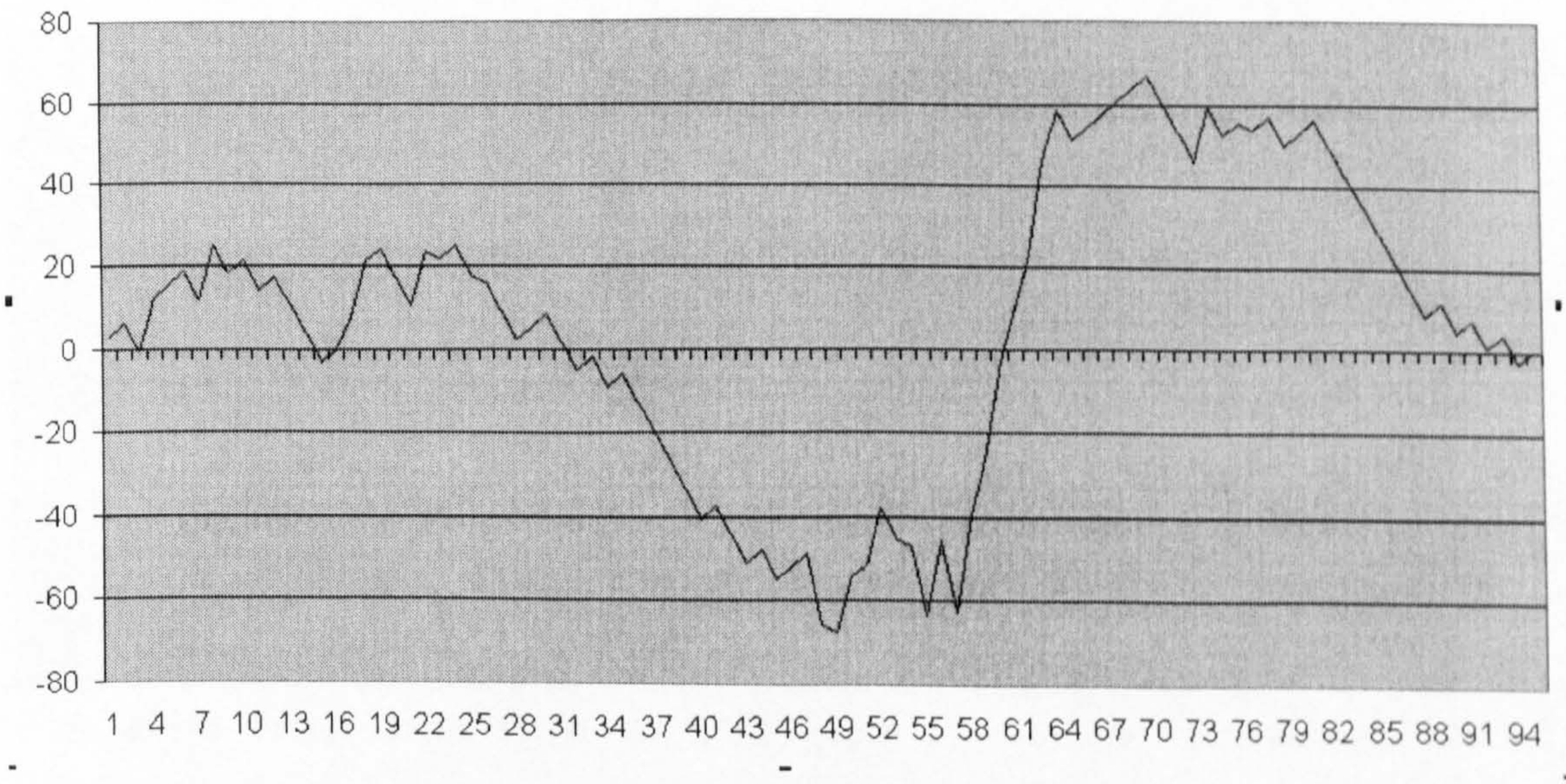


Figure 28: 10 Point Cusum for Fluid Absorption, Jan 97 - Dec 97

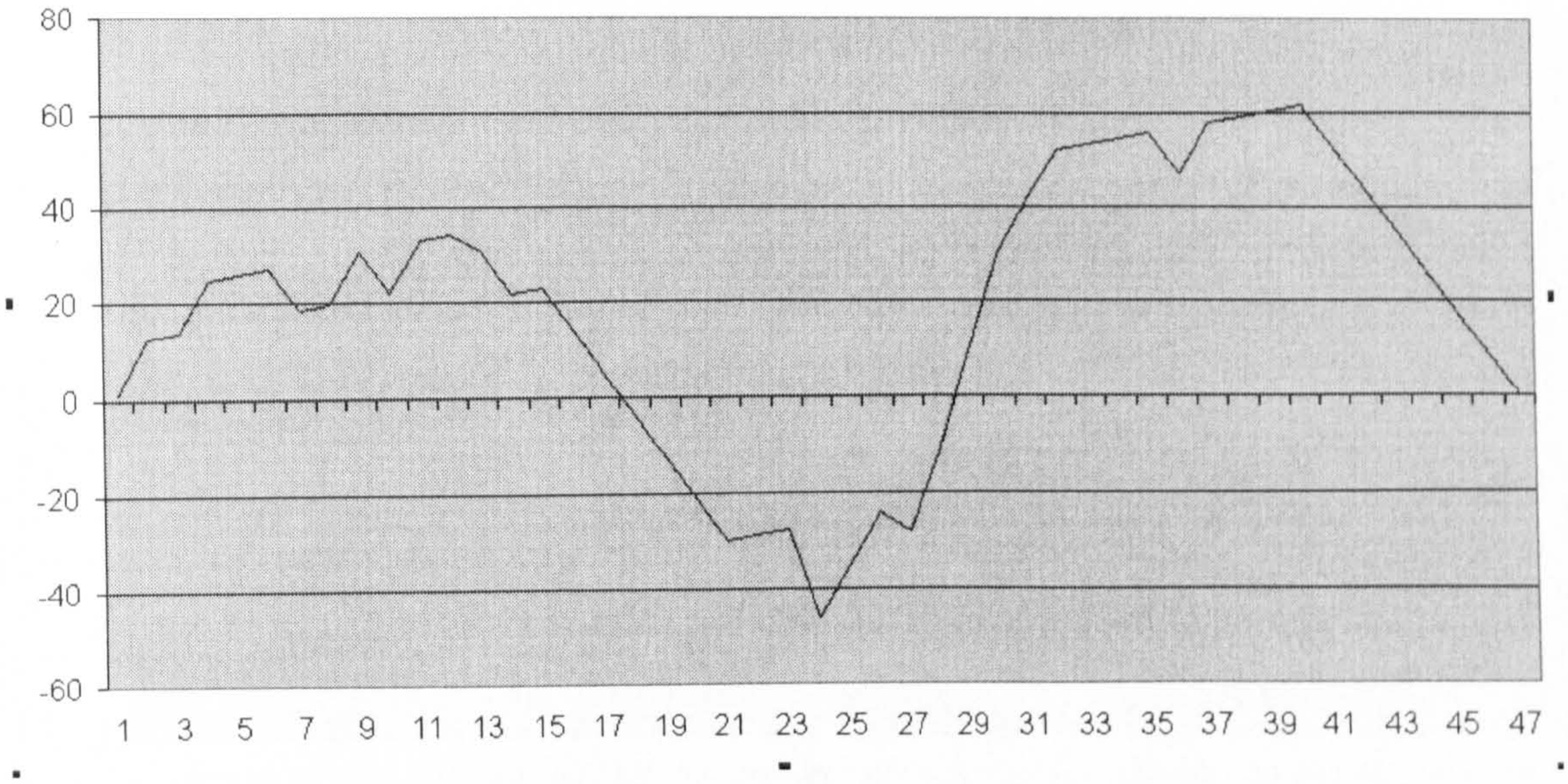


Figure 29: 20 Point Cusum for Fluid Absorption, Jan 97 - Dec 97

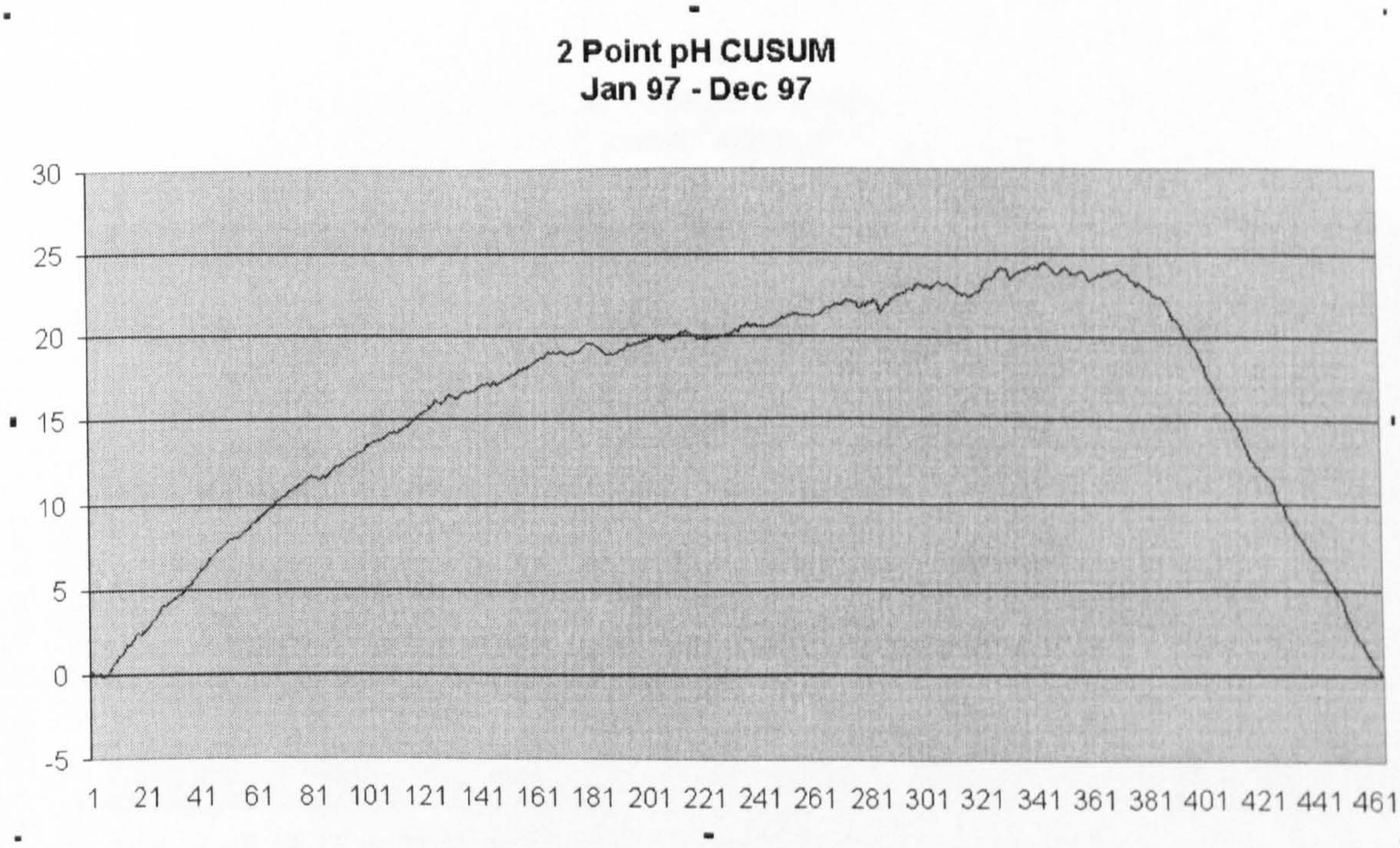


Figure 30: 2 Point Cusum for pH, Jan 97 - Dec 97

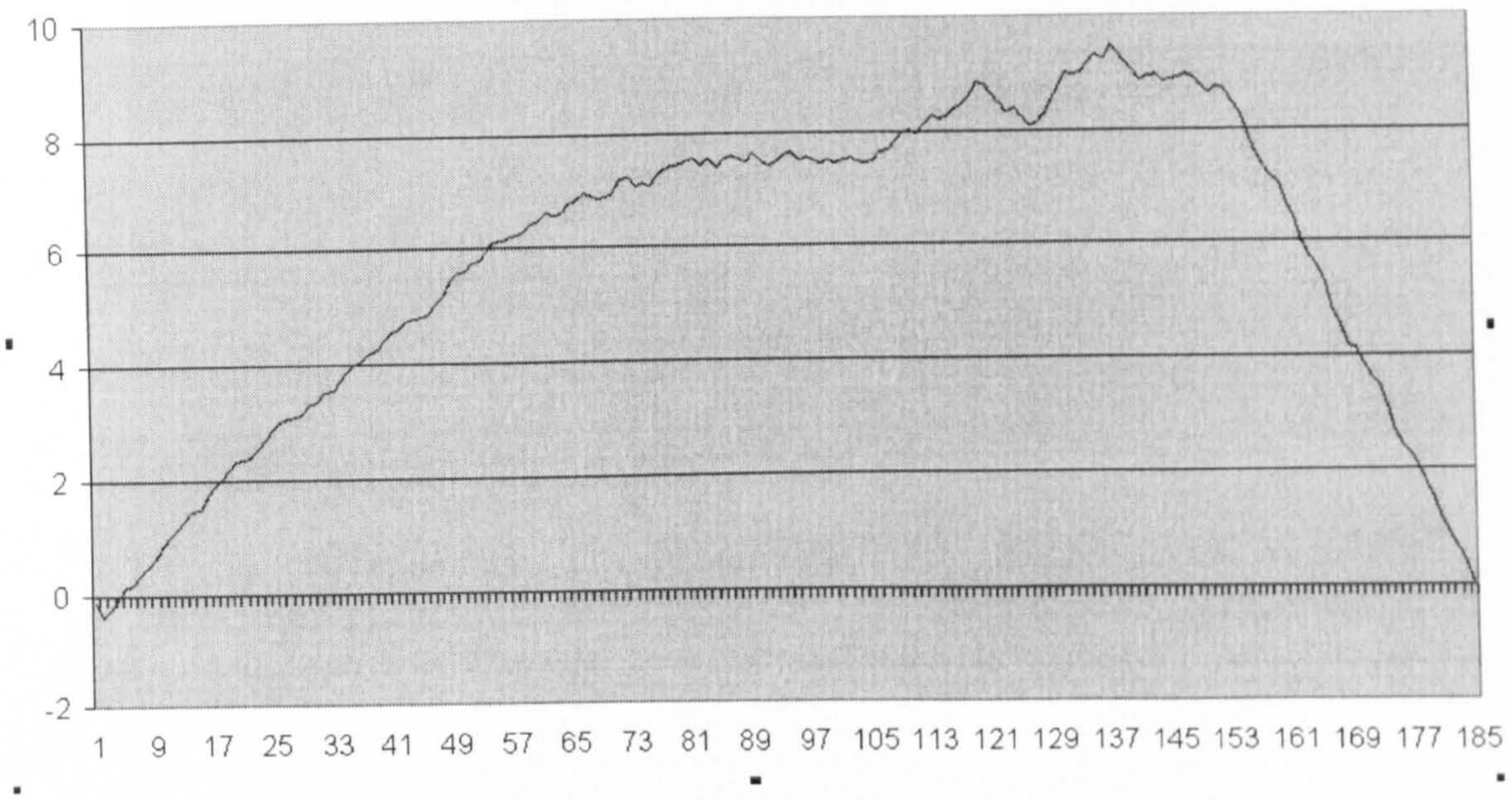


Figure 31: 5 Point Cusum for pH, Jan 97 - Dec 97

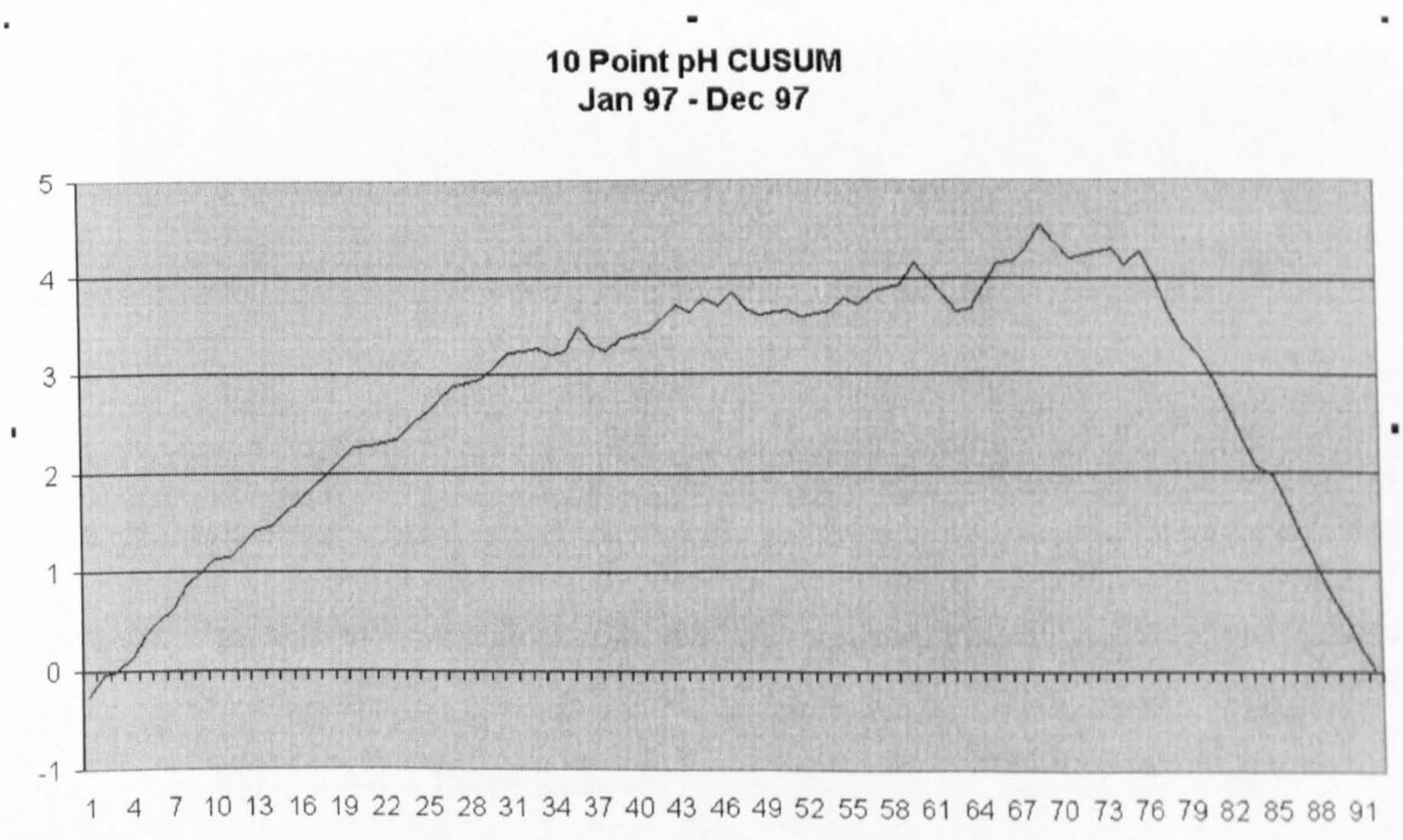


Figure 32: 10 Point Cusum for pH, Jan 97 - Dec 97

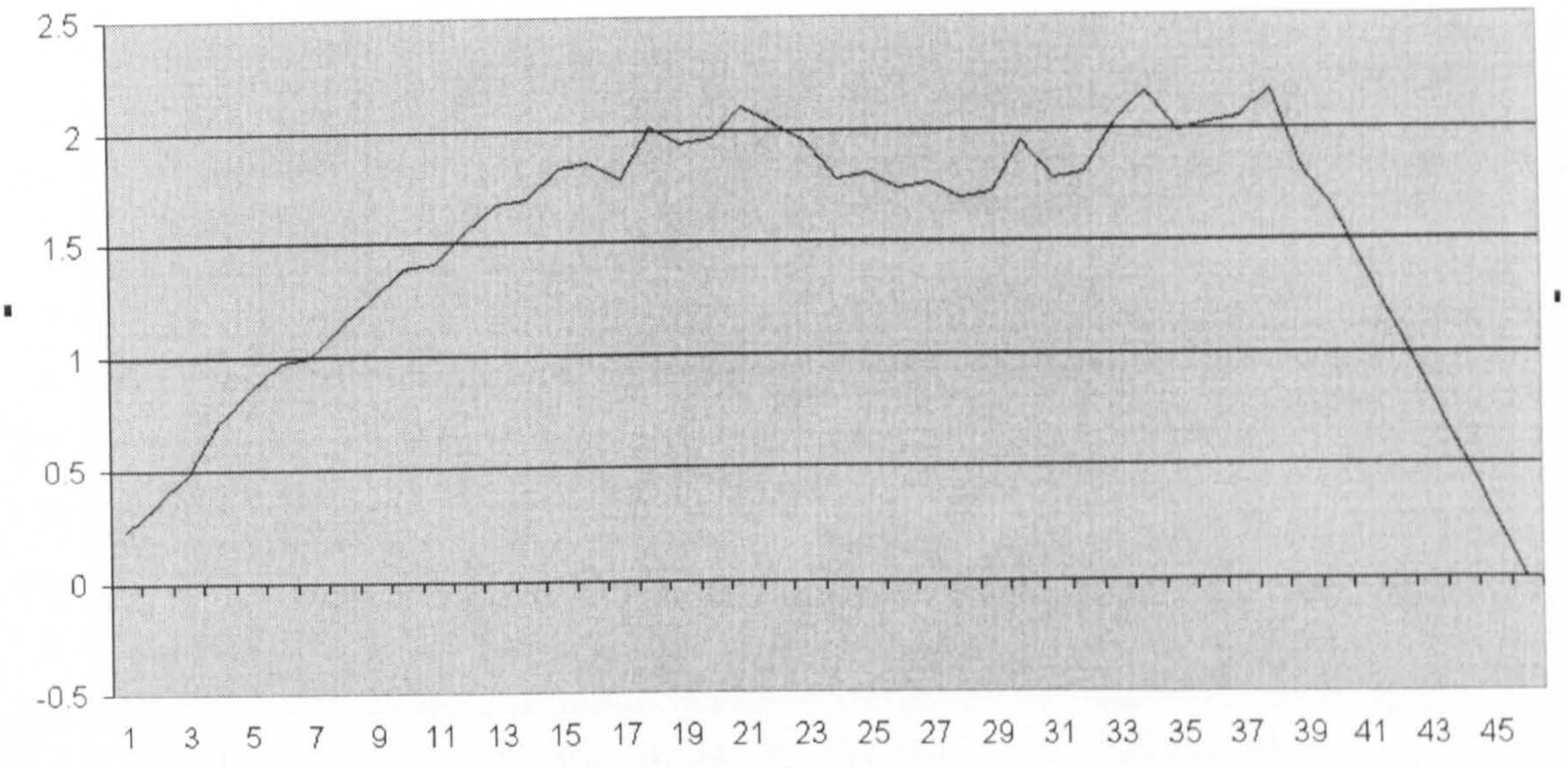


Figure 33: 20 Point Cusum for pH, Jan 97 - Dec 97

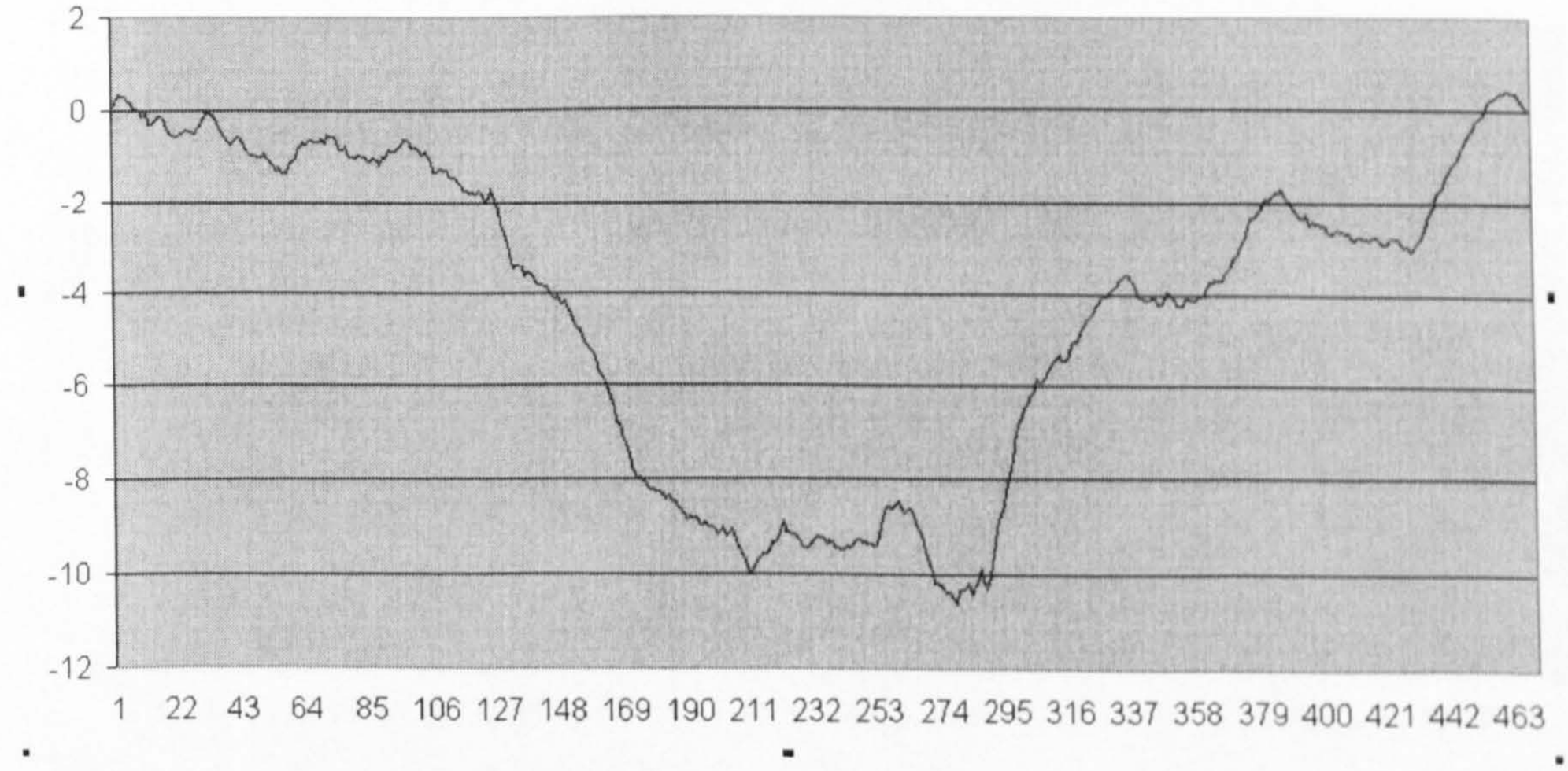


Figure 34: 2 Point Cusum for Solids Content, Jan 97 - Dec 97

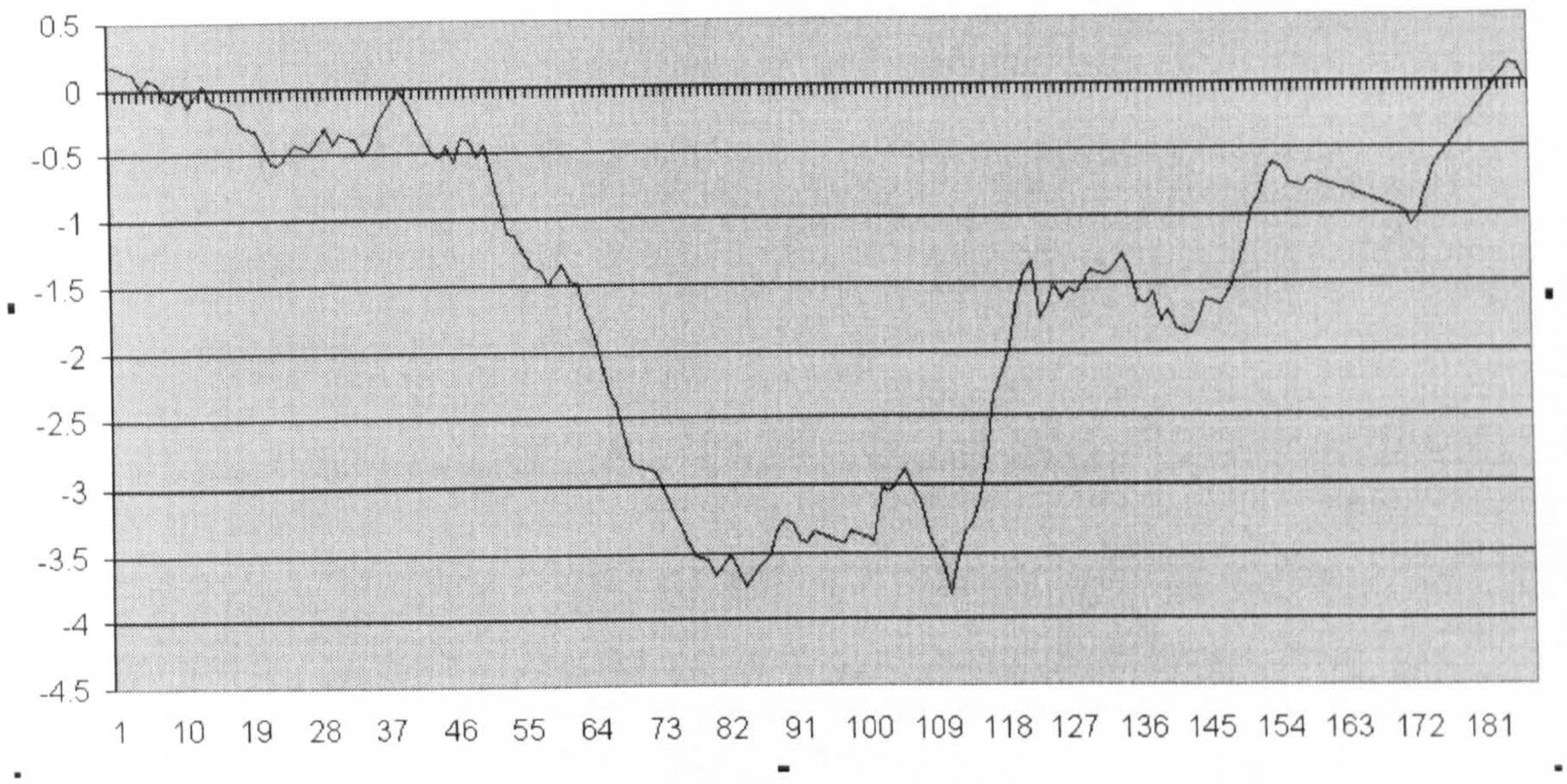


Figure 35: 5 Point Cusum for Solids Content, Jan 97 - Dec 97

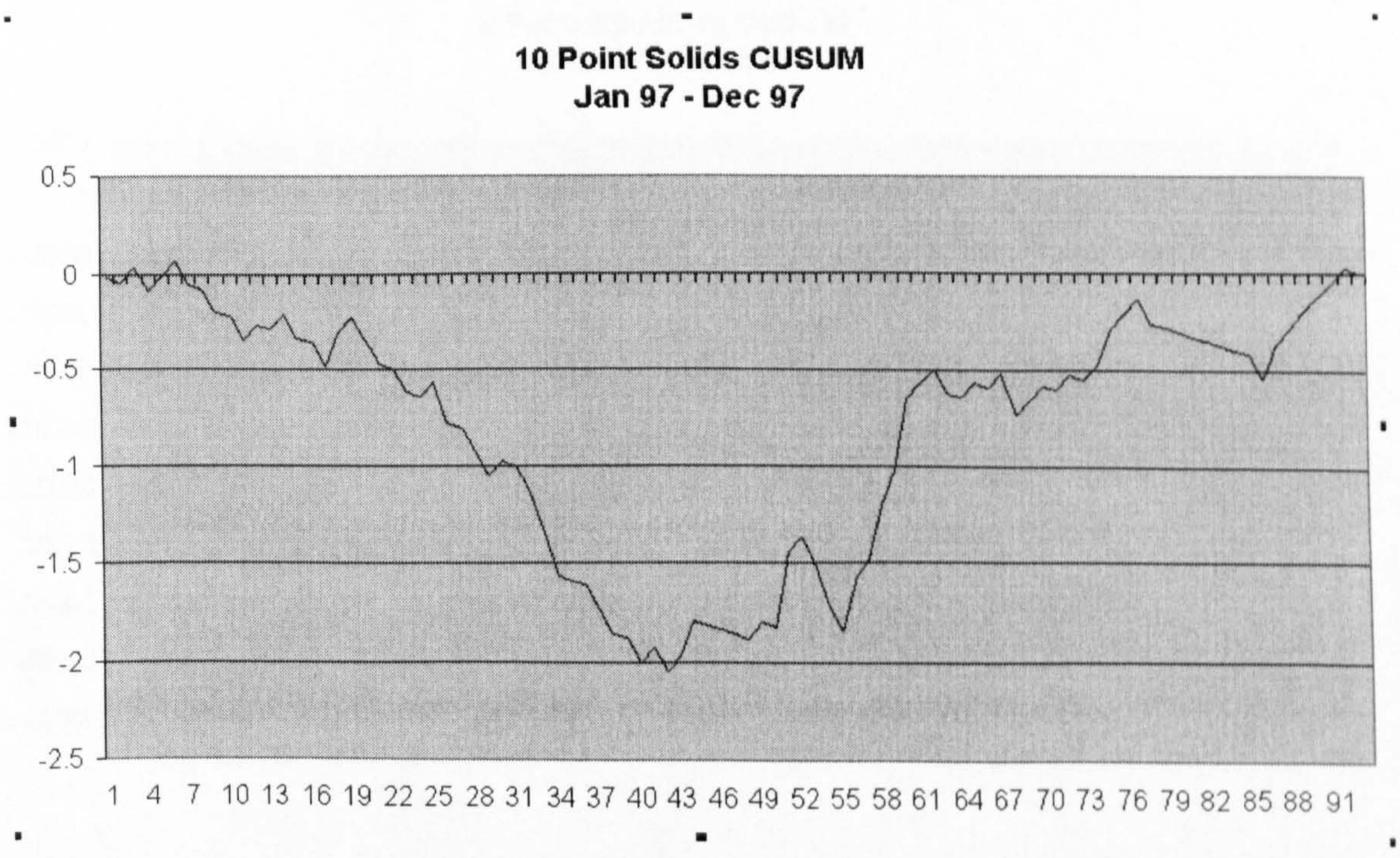


Figure 36: 10 Point Cusum for Solids Content, Jan 97 - Dec 97

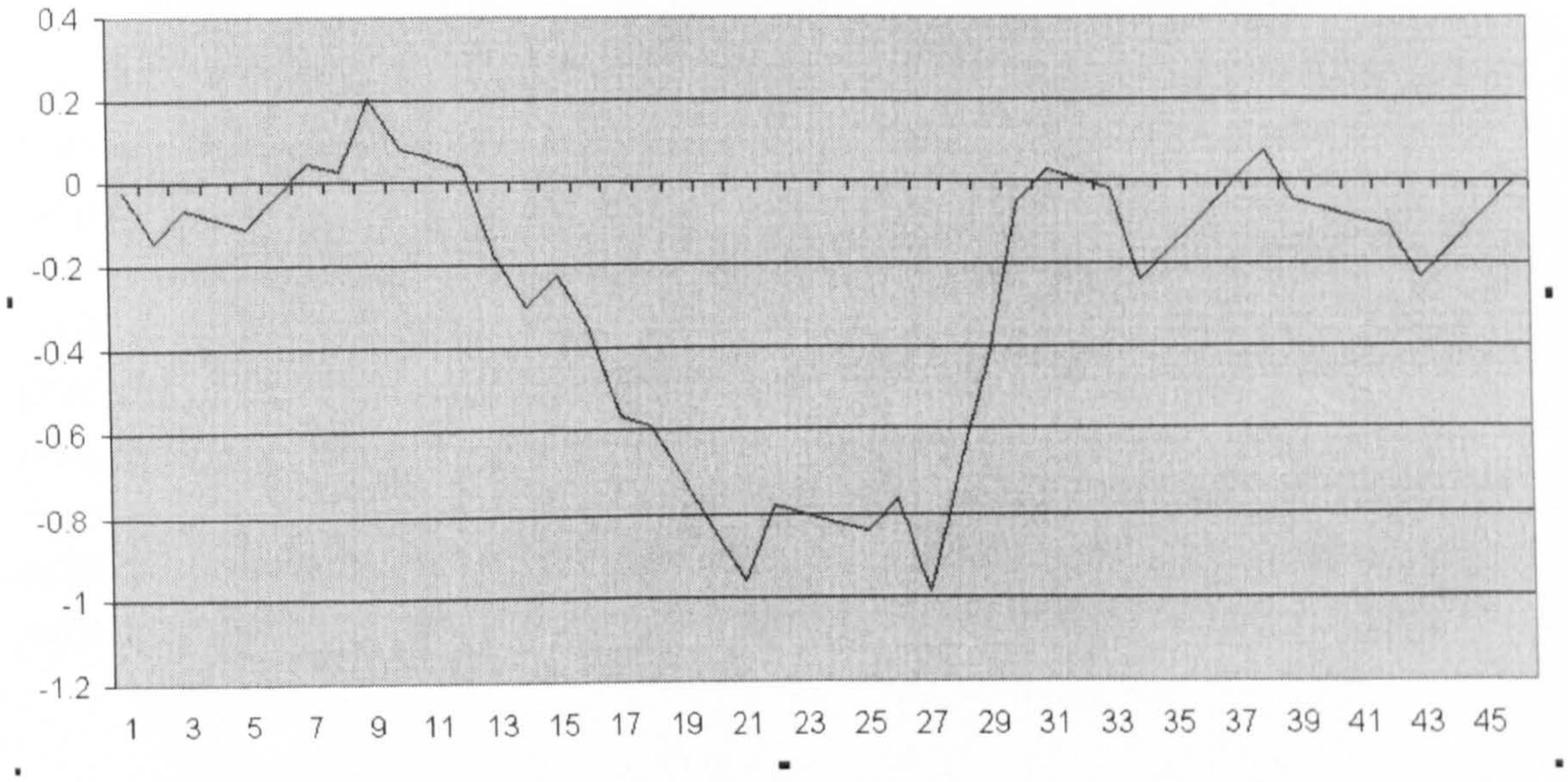


Figure 37: 20 Point Cusum for Solids Content, Jan 97 – Dec 97

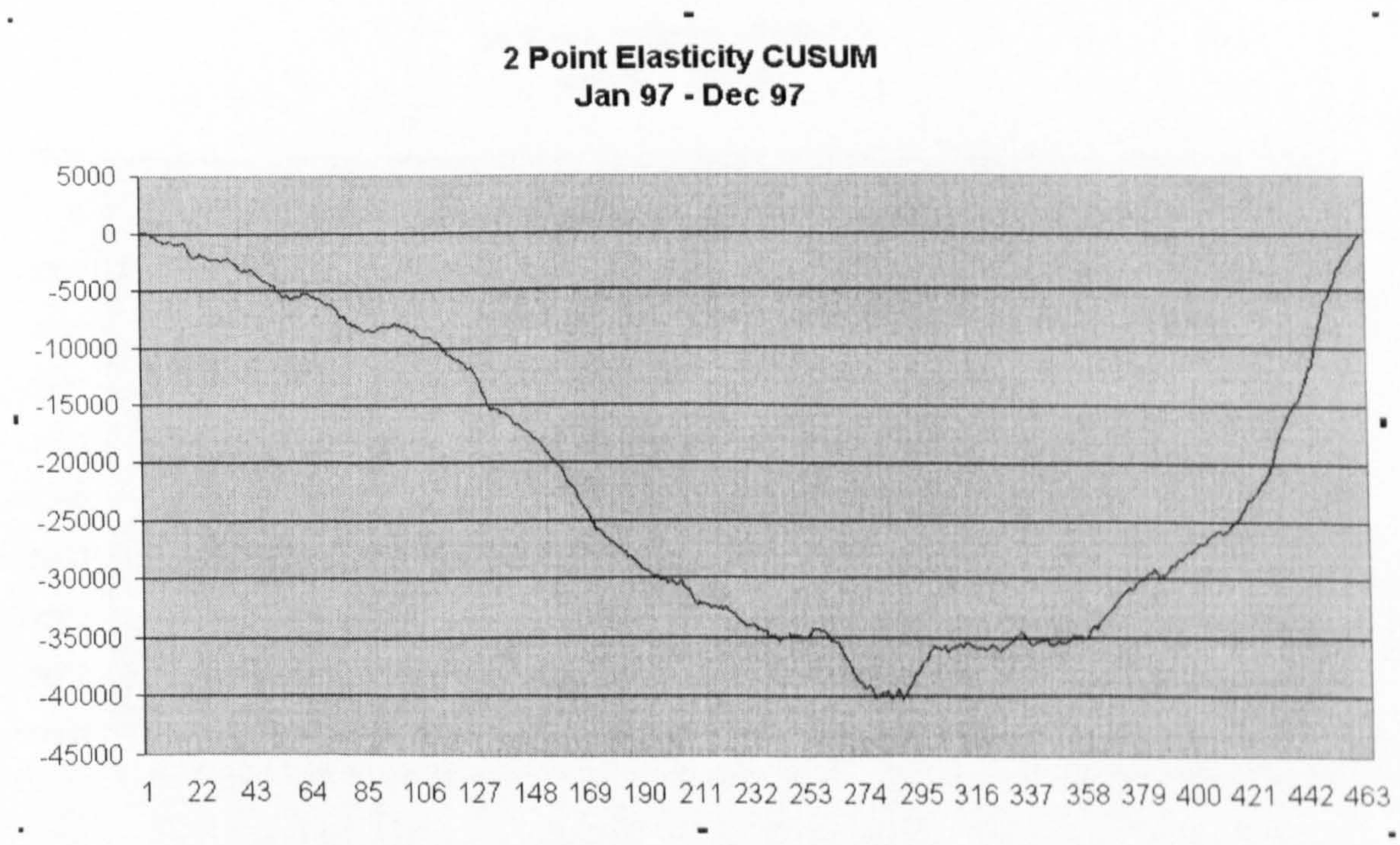


Figure 38: 2 Point Cusum for Elasticity, Jan 97 - Dec 97

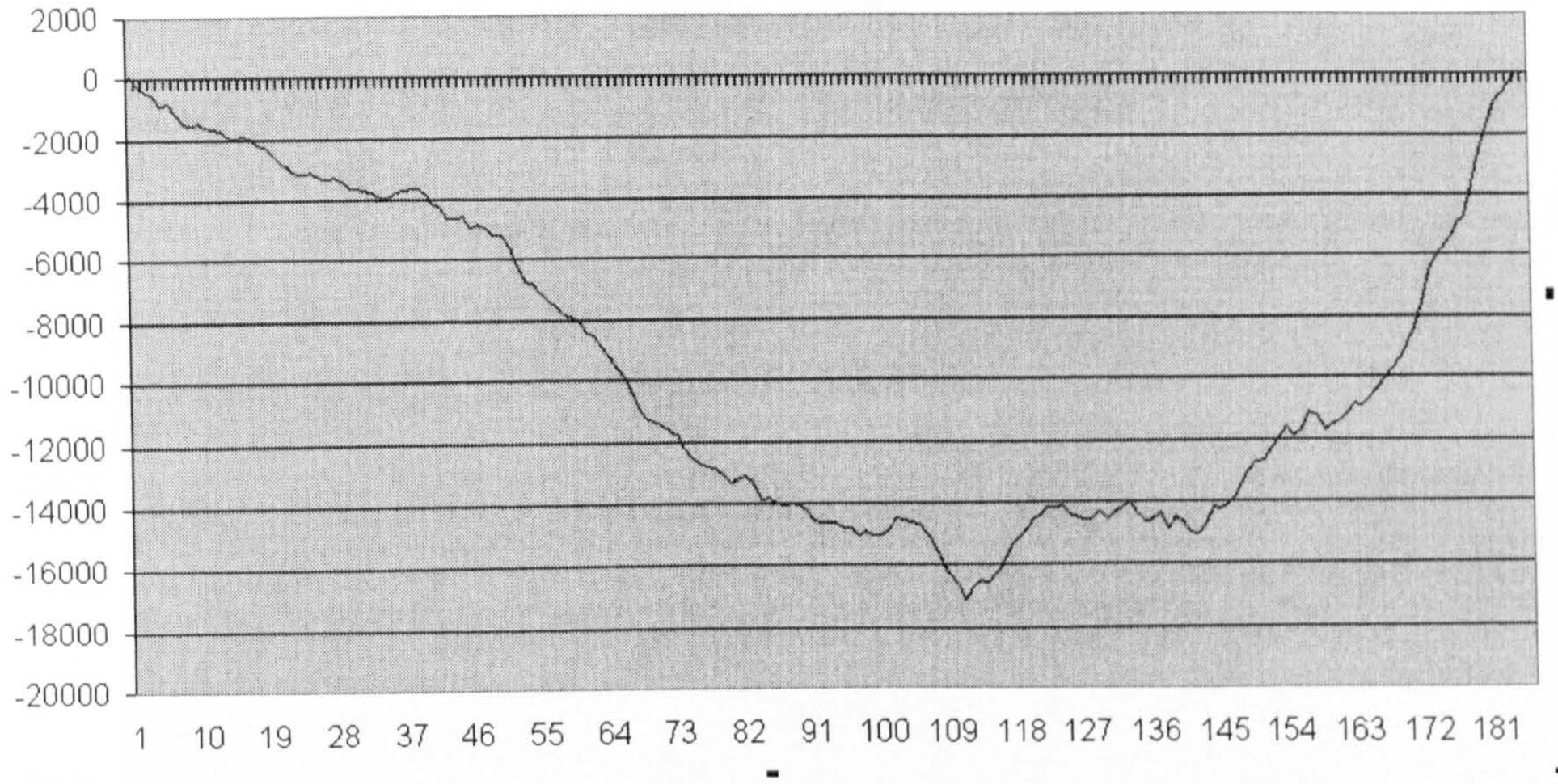


Figure 39: 5 Point Cusum for Elasticity, Jan 97 - Dec 97

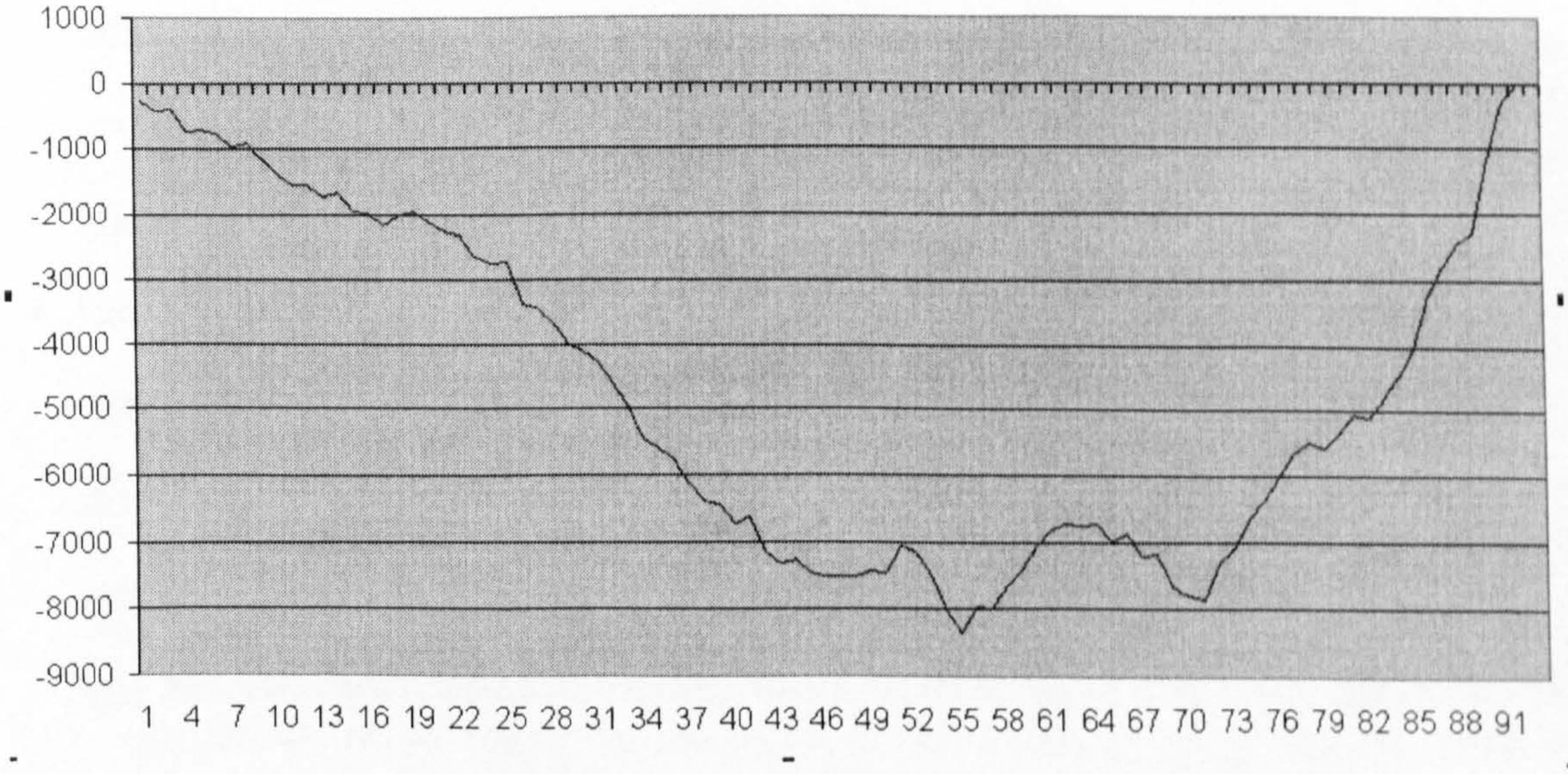


Figure 40: 10 Point Cusum for Elasticity, Jan 97 - Dec 97

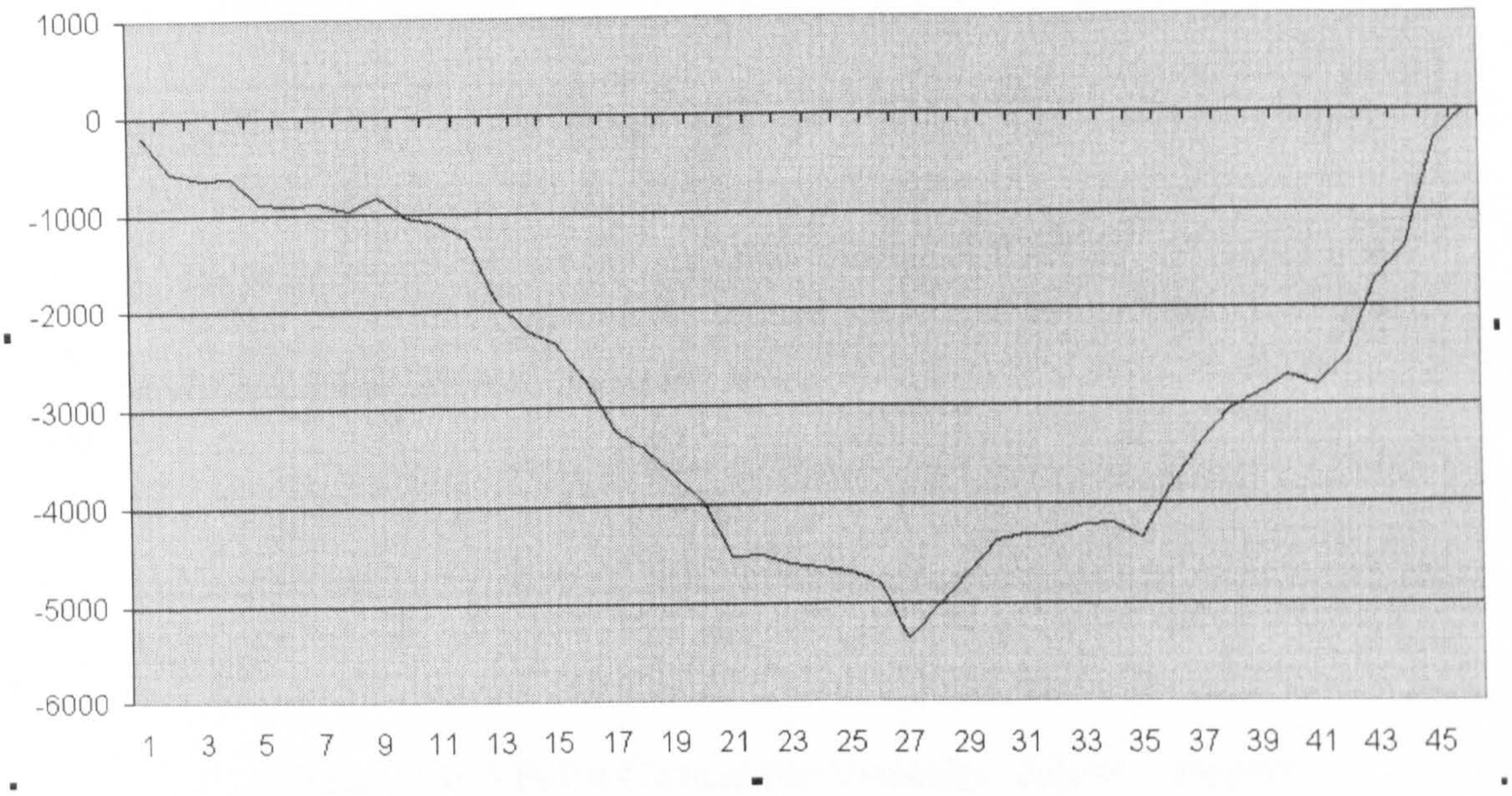


Figure 41: 20 Point Cusum for Elasticity, Jan 97 - Dec 97

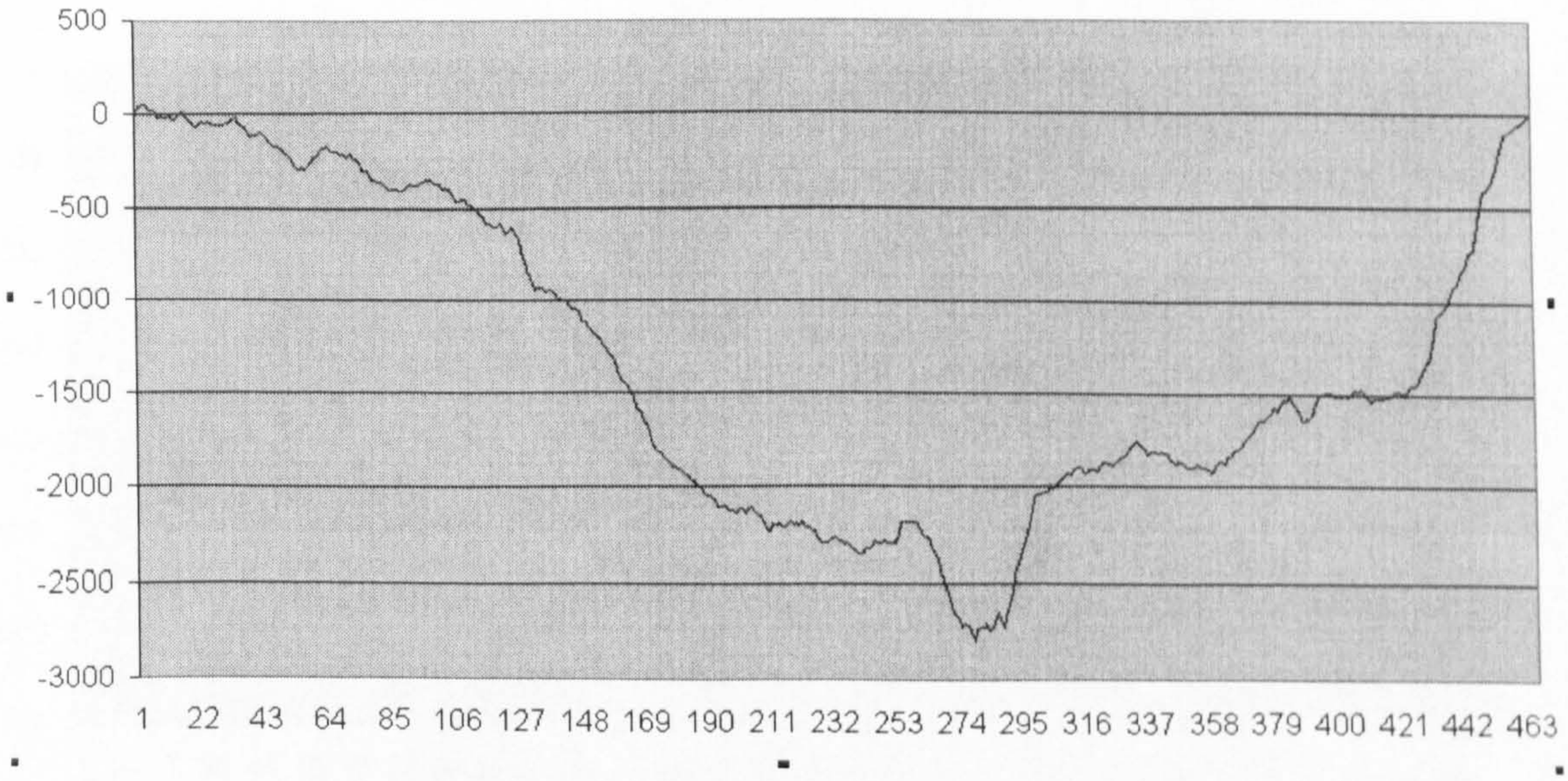


Figure 42: 2 Point Cusum for Viscosity, Jan 97 - Dec 97

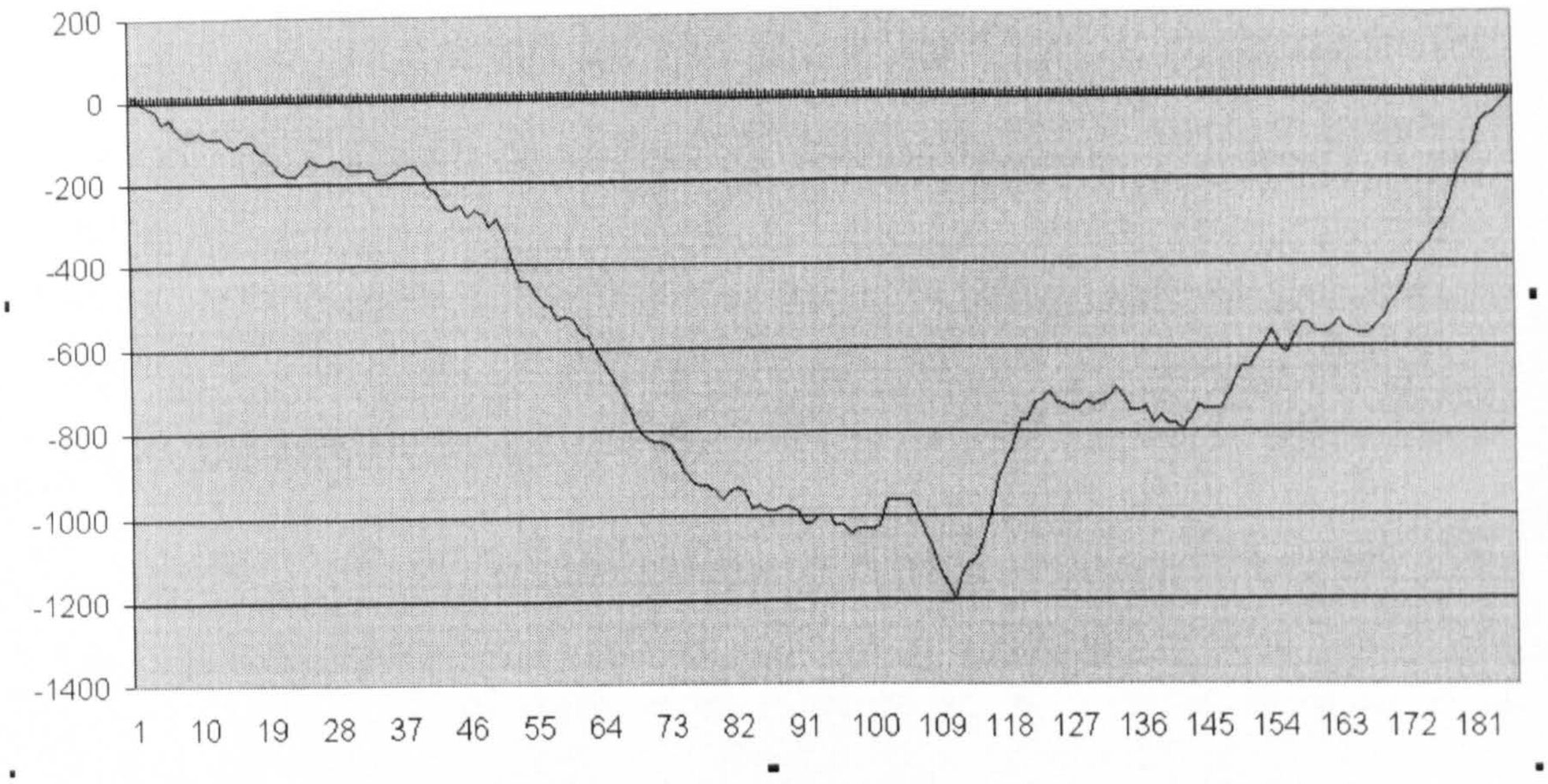


Figure 43: 5 Point Cusum for Viscosity, Jan 97 - Dec 97

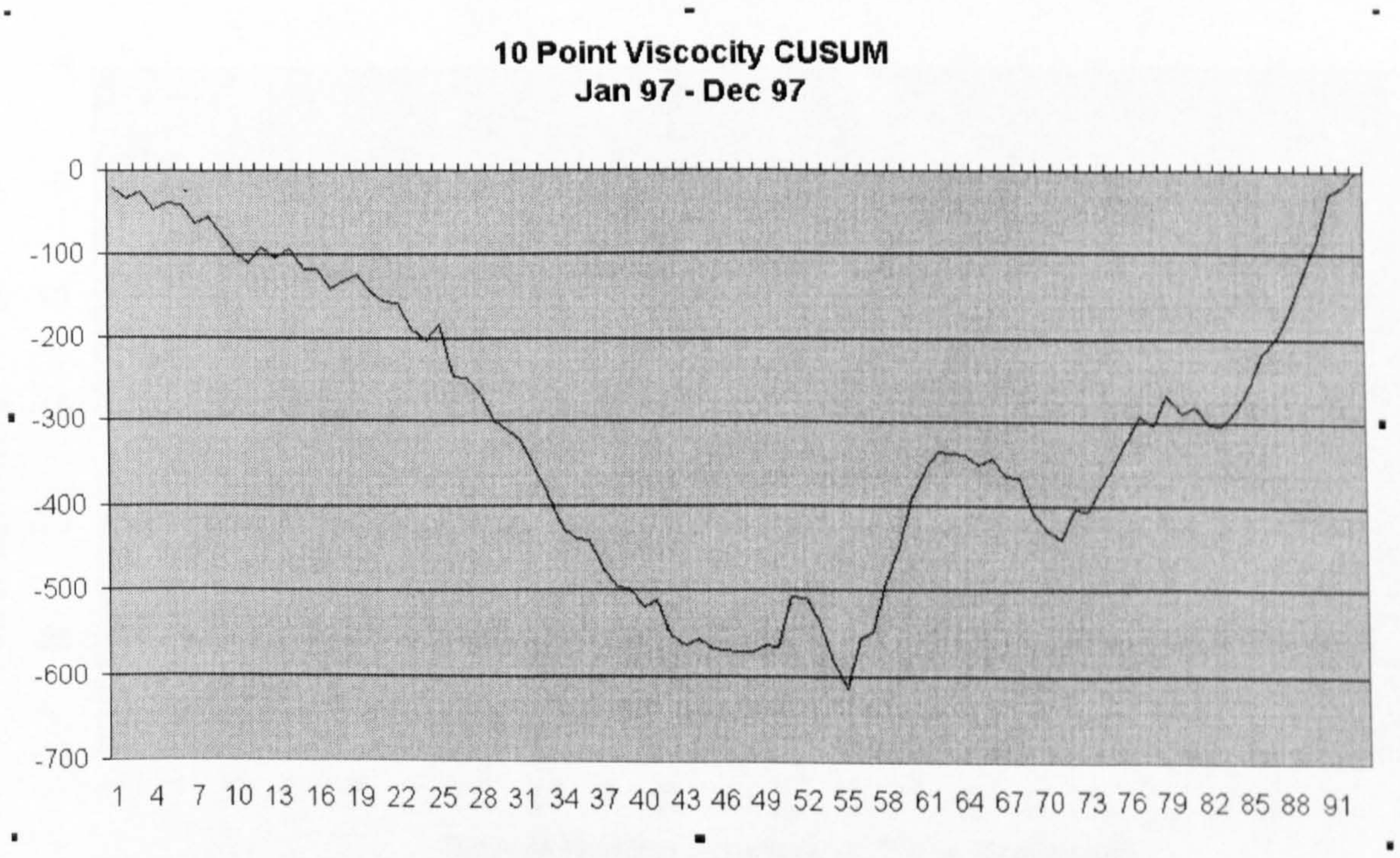


Figure 44: 10 Point Cusum for Viscosity, Jan 97 - Dec 97

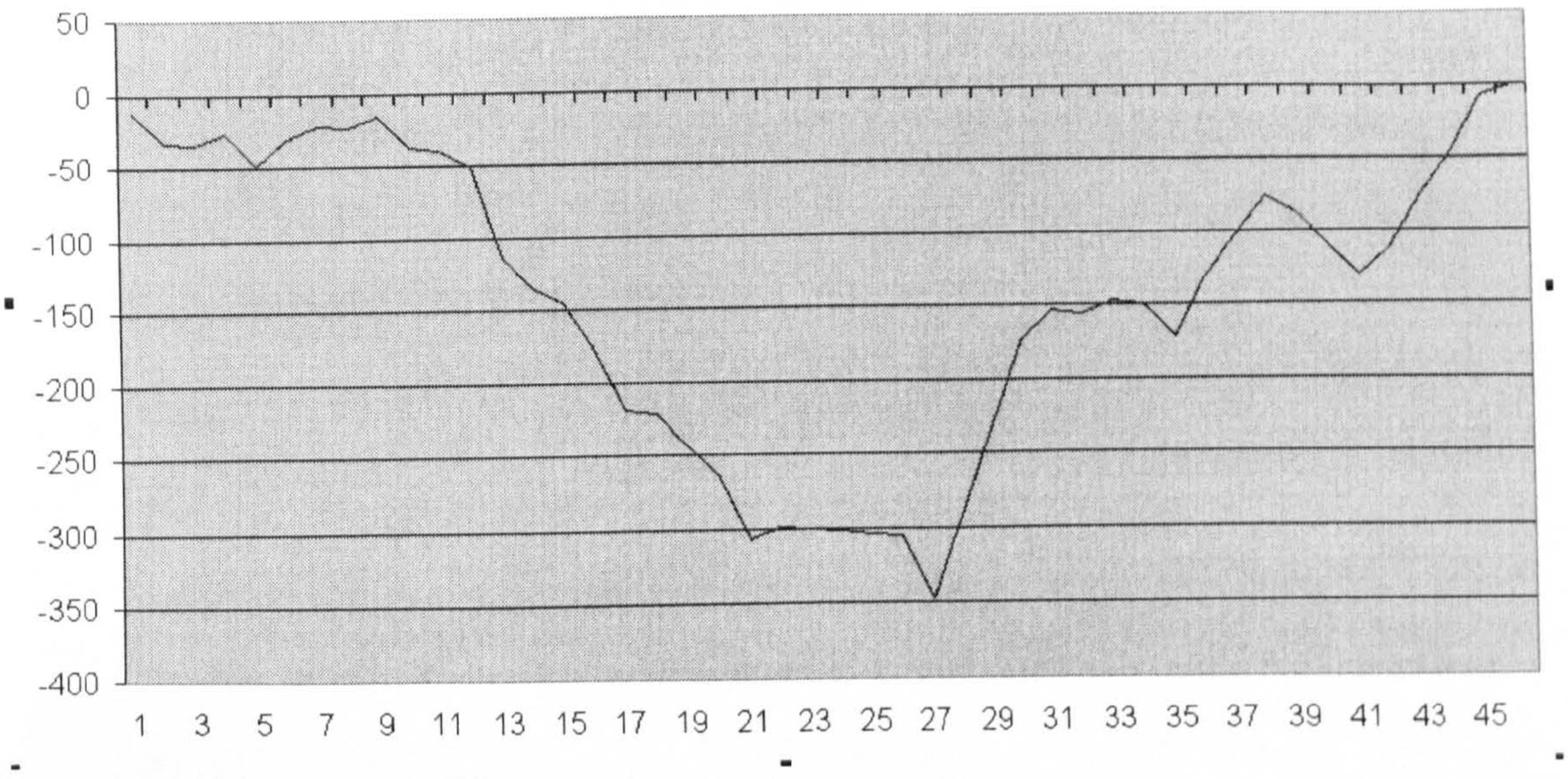


Figure 45: 20 Point Cusum for Viscosity, Jan 97 - Dec 97

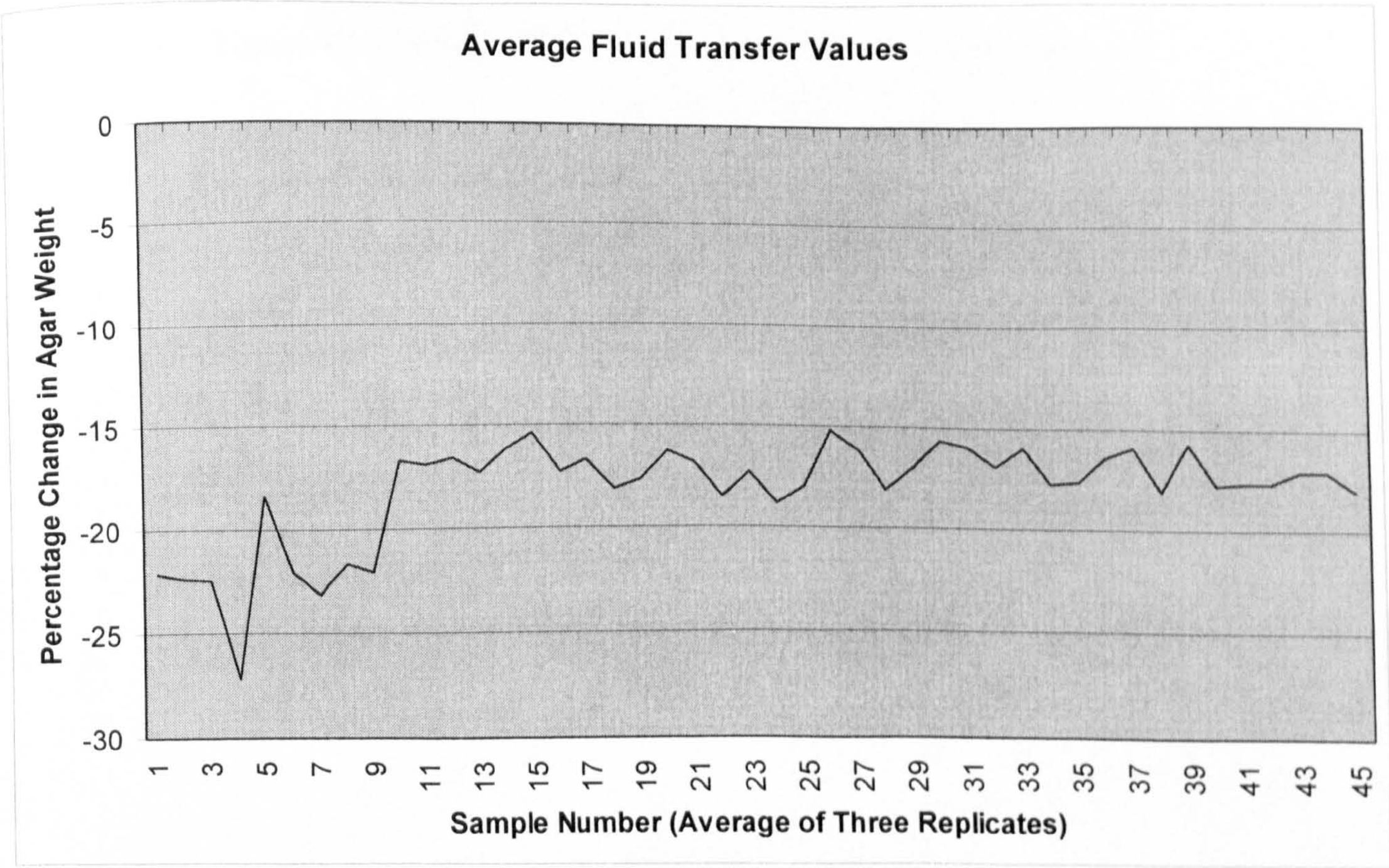


Figure 46: Fluid Transfer Test Results

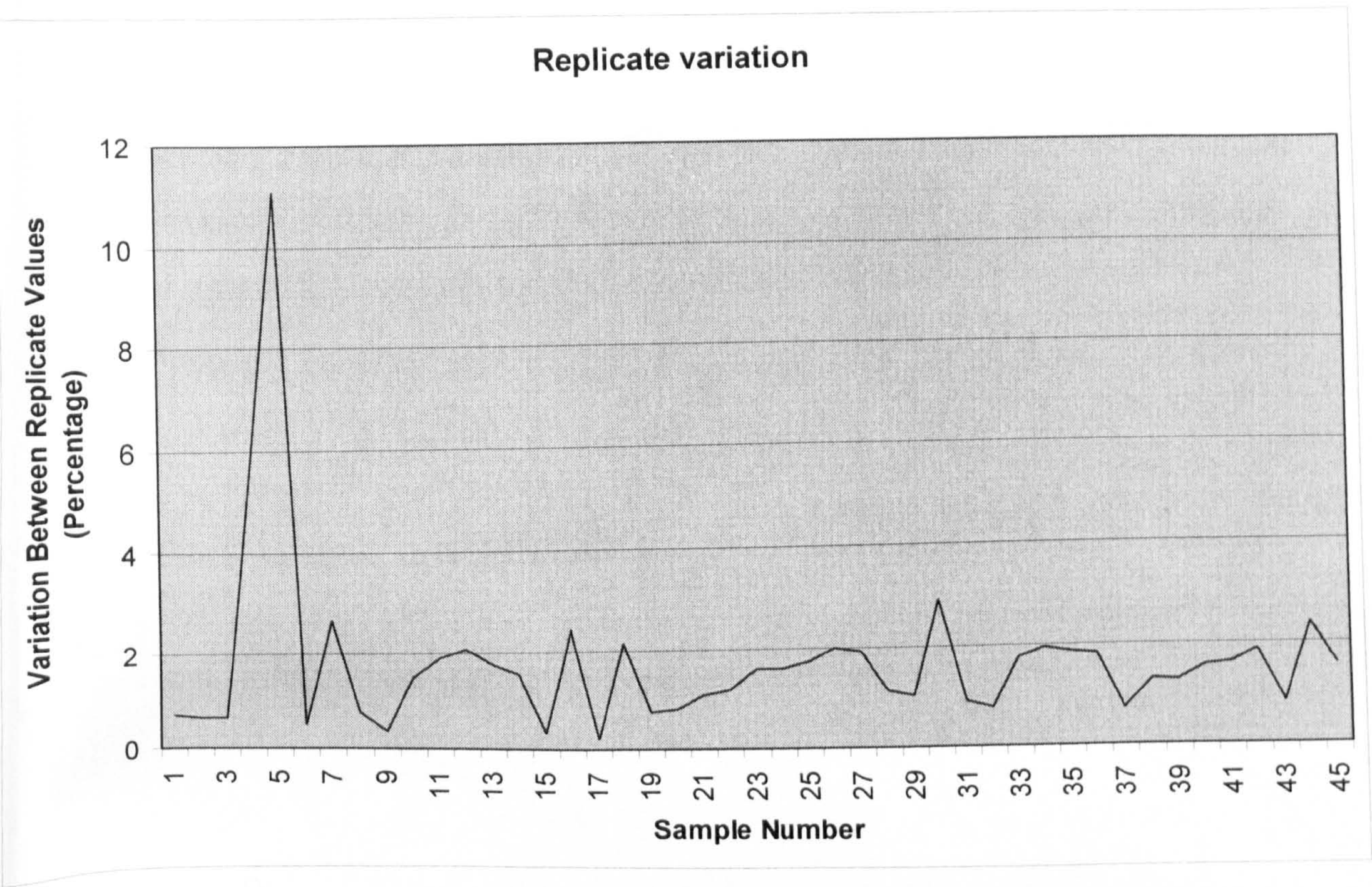


Figure 47: Variation between fluid transfer test replicates