A computerised algorithm to select automatically appropriate wavelength (or 20) variables for subsequent multivariate calibration modelling was applied to the determination of iron, manganese, potassium, calcium, titanium, silicon, aluminium, magnesium and sodium in a range of certified geological materials by XRF spectrometry. The application of partial least-squares (PLS) regression is shown to be superior in terms of predictive performance to univariate linear regression modelling and multiple linear regression analysis. The combined process of automated variable selection and PLS modelling is amenable to providing an automated XRF quantitative analysis software system.

Keywords: Geological materials; X-ray fluorescence spectrometry; partial least-squares regression; multivariate calibration modelling

XRF spectrometry is a routinely applied technique for both qualitative and quantitative multi-elemental analysis, particularly for solid samples which often require minimal pretreatment prior to analysis. The technique exhibits wide elemental coverage, good detection limits (typically mg kg\(^{-1}\)) and an extensive working range (to high percentage or pure materials). Compared with optical emission spectra, XRF spectra are relatively simple and well defined rules or heuristics can be developed to assist in the qualitative interpretation of the spectra. To this end, we have developed and reported a computerised expert system (AXIS) for automated XRF qualitative analysis.\(^1\)\(^\text{−}3\) From a digitised spectrum the emission peaks are extracted and, in decreasing order of intensity, are identified according to the elements present in the sample. Fuzzy logic and fuzzy operators are employed to account for uncertainty in the data and to mimic human interpretation behaviour. Results from the expert system on a wide range of sample types of widely different chemical composition have demonstrated the robust and accurate operation of the system.\(^3\) A useful and valuable extension to this software would be an automated quantitative analyser, providing the system with the ability to report not only the presence of an element but also its concentration in the sample under investigation.

In the simplest case, using the so-called univariate model, a linear regression analysis is performed between analyte concentration and fluorescence emission intensity recorded at a single wavelength. Where interference effects (spectral or matrix effects) are present, however, such a simple model may not be valid and other, additional, wavelengths can be introduced into the model in order to improve its predictive performance. A wide variety of these multivariate models are readily used in XRF analysis\(^4\)\(^\text{−}6\) and selection of appropriate ‘best’ model relies on the experience of the user. In many branches of spectroscopy in recent years, considerable attention has been devoted to developing and applying orthogonal multivariate regression models in which the independent variables are uncorrelated. These uncorrelated variables are formed from linear combinations of original, recorded variables. Many reports and discussions in the analytical and chemometrics literature have reviewed the relative merits of multivariate calibration models for quantitative analysis, and the advantages of orthogonal models such as principal components regression and partial least-squares (PLS) regression in terms of stability and robustness are well documented.\(^5\)\(^,\)\(^6\) Swerts et al.\(^7\) employed PLS regression for the determination of sulfur in sulfur–graphite mixtures using energy-dispersive XRF data. Although traditional least squares models failed to correct adequately for interferences and abnormal scattering of the excitation radiation, the PLS model was able to predict sulfur concentration with an accuracy of better than 5% in the concentration range 2–60% sulfur. Swerts et al.\(^8\) also successfully applied a similar PLS technique to the determination of 15 elements in Chinese porcelain, again using energy-dispersive XRF spectral data. PLS calibration models have also been studied by Urbanski and Kowalska\(^9\) for quantitative energy-dispersive XRF analysis of a range of sample types. They reported that the use of the PLS model not only gave greater robustness to interference effects but also provided new insights into aspects of the measurement methodology. Wang et al.\(^10\) compared wavelength-dispersive XRF fluorescence calibration using PLS regression with a conventional explicit calibration model employing multiple linear regression (MLR). With spectral data known to suffer from both absorption and enhancement interference effects, obtained from a set of nickel alloy samples, Wang et al. demonstrated the superiority of the PLS implicit model. One area of concern noted by Wang et al. referred to the need for further work devoted to the study of the selection of appropriate spectral regions for subsequent calibration model development for each analyte.

Hence an automated quantitative analysis software package should be capable of performing wavelength selection in addition to the application of a suitable calibration model, and in addition provide a chemically sound basis for its use and application. We have recently reported the preliminary results from a study of automated variable selection methods for wavelength-dispersive XRF data and subsequent application of a PLS regression algorithm.\(^11\) The results presented here demonstrate the application of the technique employed to a study of a series of geological samples analysed for nine elements using three dispersion crystals in the spectrometer.

Experimental
Twenty-five geological samples representing a wide range of such materials were examined, including basalt, granite, feldspar, soils and sediments (Table 1). Obtained from a variety of sources, including the US Geological Survey, the Geological Survey of Japan and the Canadian Reference Materials Project, no attempt was made to group samples into similar classes or sub-samples. The only criterion for selection and inclusion in the study was that each sample should have a certified value for each of the nine elements determined. Supplied as powders,
Results and discussion

The averages of the 25 sample spectra using each spectrometer dispersing crystal are illustrated in Fig. 1. Both the LiF (200) and the PET spectra are characteristic sharp-band spectra superimposed on a non-linear background typical of that obtained from a relatively low density matrix sample such as the geological materials examined here. The AX06 spectra appear broad-band owing to the greater digitisation frequency and the narrow range of angles scanned with this crystal. No pre-treatment, filtering or smoothing was performed on the data prior to numerical analysis. All calibration models were developed and evaluated using mean-centred data, i.e., the average spectrum using each crystal was subtracted from each sample spectrum recorded using that crystal.

Some of the problems associated with independent variable selection for subsequent calibration modelling have been discussed elsewhere, and an efficient and effective technique for XRF analysis is the application of a modified algorithm due to Brown et al. A linear relationship between emission intensity and analyte concentration is assumed and for each mean-centred digitised variable the following equation is applied:

\[ \hat{y} = a_1 x_1 + e_k \]  

(1)

where \( \hat{y} \) is the vector of predicted analyte concentrations, \( x_1 \) the vector of emission intensities at \( 2q \) angle \( k = 1, \ldots, K) \), \( a_1 \) the sensitivity (i.e., the slope of the intensity versus concentration regression line) using this variable for the analyte and \( e_k \) the error vector associated with assuming and fitting the linear model, i.e., the residuals. For iron as the analyte in the geological samples, a plot of \( a_1 \) versus angle is shown in Fig. 2(a). The spectral regions of high linear sensitivity to iron concentration are evident and are located about the major \( K\alpha \) and \( K\beta \) fluorescence lines of iron. In addition, the \( K\alpha \) lines from calcium and potassium can also be identified, indicating a linear correlation (positive for calcium and negative for potassium) between the concentrations of these elements and the concentration of iron. In order to enhance the selectivity of the 26 variable selection algorithm, the \( a_i \) values can be scaled by dividing each value by the variance, \( \sigma_k^2 \), of the residuals, \( e_k \). This serves to lessen the importance and influence of spectral regions with non-linear response and regions of spurious correlation due to high noise levels. Brown proposed this wavelength selection method using near-infrared absorbance spectra. For emission techniques, such as XRF, the dynamic range of the recorded response (i.e., emission intensity) can be several orders of magnitude greater. By performing a logarithmic transformation of the absolute sensitivity values, \( a_i \), the dynamic range characteristic of XRF emission intensities is suitably compressed, and the selection process is not overly dominated by any single most intense line. Hence our selection term, \( s \), can be expressed as:

\[ s_k = \log_{10} \frac{a_k^2}{\sigma_k^2} \]  

(2)
A plot of $s_2$ versus $2\theta$ angle is presented in Fig. 2(b). With spurious correlations now removed, the prominent peaks in this spectrum are due to the Kα and Kβ (first and second-order) lines of iron. It now remains to identify and select the regions about these peaks for inclusion in the calibration models. Brown\textsuperscript{13} has proposed a novel means for the selection of the optimum number and identity of the spectral regions for calibration. The selection process is achieved using the equation

$$Z = \frac{C_{(\eta)}}{\sum_{k=1}^{K} S_k}$$  \hspace{1cm} (3)

where $1-\eta$ is some selected confidence level and

$$C_{(\eta)} = \sqrt{\chi^2_{1,1} - \eta [k]}$$ \hspace{1cm} (4)

is the square root of tabulated chi-squared values on $k$ degrees of freedom and $K'$ is the reduced number of variables, $K' < < K$. To choose the value for $K'$, and the corresponding spectral regions, the computed $s$ values [eqn. (2)] are ranked from largest to smallest and, by eqn. (3), the function $Z$ is minimized. From the geological samples and for iron, the form of eqn. (3) is shown in Fig. 3. The number of $2\theta$ variables to be selected and used is given by the location of the minimum for the curve and the identity of the corresponding variables obtained from the ranked list. For any spectral data set, the location of the minimum from eqn. (3) is dependent on the chosen significance level, $1-\eta$. Hence, for iron and $\eta$ values of 10, 1 and 0.1% the sets ($A$) of spectral $2\theta$ variables from the geological sample spectra are as follows:

$\eta = 10\%: A_{0.9} = \{57.5, 57.6, 57.7, 57.8\}$
$\eta = 1\%: A_{0.99} = \{51.8, 51.9, 52.0, 57.5, 57.6, 57.7, 57.8, 57.9, 148.8, 148.9\}$
$\eta = 0.1\%: A_{0.999} = \{51.8, 51.9, 52.0, 57.5, 57.6, 57.7, 57.8, 57.9, 58.0, 148.6, 148.7, 148.8, 148.9\}$

(5)

Decreasing the significance level increases the number of $2\theta$ variables selected, and each larger set contains the elements of a previous set, i.e.,

$$A_{0.9} \subseteq A_{0.99} \subseteq A_{0.999}$$ \hspace{1cm} (6)

Reference to look-up XRF reference tables for the LiF\textsubscript{200} dispersing crystal identifies the four variables in set $A_{0.9}$ as being on the Fe Kα line. The $A_{0.99}$ set has added the Fe Kβ line and the second-order Fe Kα line. The final set, $A_{0.999}$, merely adds more points from these three lines.

This process of selecting $2\theta$ variables was undertaken for each of the nine analytes using all 25 reference samples, at each of the three significance levels indicated. In every case the technique resulted in a dramatic reduction of the number of variables and a chosen set of variables that can be readily
checked automatically for chemical consistency. Table 2 summarises the results of variable selection for the nine analytes. In all cases the technique has selected the appropriate Kα line for each element and, with the exception of iron, the algorithm indicates that only this line is required for subsequent calibration. Compared with simple univariate regression analysis, however, the technique identifies a range of wavelengths (20 values) about each line, thus providing an averaging effect and reducing the influence of spurious noise at one particular wavelength.

For the three sets of 20 variables (η = 0.1, 0.01, 0.001) selected for each analyte element, calibration models were developed using PLS regression and MLR. In addition to these multivariate models, univariate ordinary least-squares (OLS) regression was also undertaken using the characteristic Kα line of each element. The PLS algorithm employed was based on that reported by de Jong and referred to as SIMPLS. The algorithm avoids deflation of the full data matrix, and for single component analysis it has been shown to yield equivalent results to the more common NIPALS algorithm described by Martens and Næs.

Calibration model performance was compared using a full miss-one-out strategy on all 25 samples. For each model evaluated all samples less one were used to derive the model coefficients with the odd sample used for prediction or validation. The process was repeated 25 times, with each sample in turn used for prediction purposes. Numerical comparison of models was achieved by calculating the root-mean-square error of prediction (RMSEP):

\[ RMSEP = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2} \]  

where \( \hat{y}_i \) represents the residuals between predicted and known concentrations and \( n \) is the number of samples examined. Table 3 presents a summary of the results obtained. The RMSEP values indicated for PLS and MLR are those obtained using the set of variables giving the lowest RMSEP values with the PLS model. RMSEP values provide a useful quantitative summary of the performance of calibration models and it is evident from Table 2 that in every case the PLS model is as good as or superior to either of the more classical approaches. A more detailed picture of model performance can be obtained by visual inspection of residuals plots. Fig. 4 presents such a plot obtained from the determination of calcium. One feature of Fig. 4 is particularly worthy of note, viz., the calcium concentration values of the samples are not evenly distributed across the range encountered. This is not surprising given that the samples were not selected for any single, particular element but merely as representatives of the class of materials. For calcium, one sample has a considerably higher concentration than the others, hence with a miss-one-out validation strategy this sample will be determined by extrapolation from the calibration model developed using the remaining samples. It is not surprising in these circumstances that OLS and

### Table 2 Results of the variable selection algorithm at three significance levels for the nine elements studied

<table>
<thead>
<tr>
<th>Element</th>
<th>OLS</th>
<th>MLR</th>
<th>PLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fe</td>
<td>0.798</td>
<td>1.060</td>
<td>0.521 (4)</td>
</tr>
<tr>
<td>Mn</td>
<td>0.017</td>
<td>0.013</td>
<td>0.013 (4)</td>
</tr>
<tr>
<td>K</td>
<td>0.177</td>
<td>0.187</td>
<td>0.178 (1)</td>
</tr>
<tr>
<td>Ca</td>
<td>0.024</td>
<td>1.446</td>
<td>0.553 (1)</td>
</tr>
<tr>
<td>Ti</td>
<td>0.045</td>
<td>0.049</td>
<td>0.039 (1)</td>
</tr>
<tr>
<td>Si</td>
<td>1.471</td>
<td>2.016</td>
<td>1.482 (1)</td>
</tr>
<tr>
<td>Al</td>
<td>0.495</td>
<td>0.506</td>
<td>0.499 (2)</td>
</tr>
<tr>
<td>Mg</td>
<td>0.543</td>
<td>2.685</td>
<td>0.530 (3)</td>
</tr>
<tr>
<td>Na</td>
<td>0.446</td>
<td>0.968</td>
<td>0.465 (1)</td>
</tr>
</tbody>
</table>

For the three sets of 20 variables (η = 0.1, 0.01, 0.001) selected for each analyte element, calibration models were developed using PLS regression and MLR. In addition to these multivariate models, univariate ordinary least-squares (OLS) regression was also undertaken using the characteristic Kα line of each element. The PLS algorithm employed was based on that reported by de Jong and referred to as SIMPLS. The algorithm avoids deflation of the full data matrix, and for single component analysis it has been shown to yield equivalent results to the more common NIPALS algorithm described by Martens and Næs.

Calibration model performance was compared using a full miss-one-out strategy on all 25 samples. For each model evaluated all samples less one were used to derive the model coefficients with the odd sample used for prediction or validation. The process was repeated 25 times, with each sample in turn used for prediction purposes. Numerical comparison of models was achieved by calculating the root-mean-square error of prediction (RMSEP):

\[ RMSEP = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2} \]  

where \( \hat{y}_i \) represents the residuals between predicted and known concentrations and \( n \) is the number of samples examined. Table 3 presents a summary of the results obtained. The RMSEP values indicated for PLS and MLR are those obtained using the set of variables giving the lowest RMSEP values with the PLS model. RMSEP values provide a useful quantitative summary of the performance of calibration models and it is evident from Table 2 that in every case the PLS model is as good as or superior to either of the more classical approaches. A more detailed picture of model performance can be obtained by visual inspection of residuals plots. Fig. 4 presents such a plot obtained from the determination of calcium. One feature of Fig. 4 is particularly worthy of note, viz., the calcium concentration values of the samples are not evenly distributed across the range encountered. This is not surprising given that the samples were not selected for any single, particular element but merely as representatives of the class of materials. For calcium, one sample has a considerably higher concentration than the others, hence with a miss-one-out validation strategy this sample will be determined by extrapolation from the calibration model developed using the remaining samples. It is not surprising in these circumstances that OLS and

### Table 3 RMSEP values for validation using univariate ordinary least-squares (OLS), multiple linear regression (MLR) and partial least-squares (PLS) regression. The figures in parentheses indicate the number of factors used in the PLS model

<table>
<thead>
<tr>
<th>Element</th>
<th>OLS</th>
<th>MLR</th>
<th>PLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fe</td>
<td>0.798</td>
<td>1.060</td>
<td>0.521 (4)</td>
</tr>
<tr>
<td>Mn</td>
<td>0.017</td>
<td>0.013</td>
<td>0.013 (4)</td>
</tr>
<tr>
<td>K</td>
<td>0.177</td>
<td>0.187</td>
<td>0.178 (1)</td>
</tr>
<tr>
<td>Ca</td>
<td>0.024</td>
<td>1.446</td>
<td>0.553 (1)</td>
</tr>
<tr>
<td>Ti</td>
<td>0.045</td>
<td>0.049</td>
<td>0.039 (1)</td>
</tr>
<tr>
<td>Si</td>
<td>1.471</td>
<td>2.016</td>
<td>1.482 (1)</td>
</tr>
<tr>
<td>Al</td>
<td>0.495</td>
<td>0.506</td>
<td>0.499 (2)</td>
</tr>
<tr>
<td>Mg</td>
<td>0.543</td>
<td>2.685</td>
<td>0.530 (3)</td>
</tr>
<tr>
<td>Na</td>
<td>0.446</td>
<td>0.968</td>
<td>0.465 (1)</td>
</tr>
</tbody>
</table>

For the three sets of 20 variables (η = 0.1, 0.01, 0.001) selected for each analyte element, calibration models were developed using PLS regression and MLR. In addition to these multivariate models, univariate ordinary least-squares (OLS) regression was also undertaken using the characteristic Kα line of each element. The PLS algorithm employed was based on that reported by de Jong and referred to as SIMPLS. The algorithm avoids deflation of the full data matrix, and for single component analysis it has been shown to yield equivalent results to the more common NIPALS algorithm described by Martens and Næs.

Calibration model performance was compared using a full miss-one-out strategy on all 25 samples. For each model evaluated all samples less one were used to derive the model coefficients with the odd sample used for prediction or validation. The process was repeated 25 times, with each sample in turn used for prediction purposes. Numerical comparison of models was achieved by calculating the root-mean-square error of prediction (RMSEP):

\[ RMSEP = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2} \]  

where \( \hat{y}_i \) represents the residuals between predicted and known concentrations and \( n \) is the number of samples examined. Table 3 presents a summary of the results obtained. The RMSEP values indicated for PLS and MLR are those obtained using the set of variables giving the lowest RMSEP values with the PLS model. RMSEP values provide a useful quantitative summary of the performance of calibration models and it is evident from Table 2 that in every case the PLS model is as good as or superior to either of the more classical approaches. A more detailed picture of model performance can be obtained by visual inspection of residuals plots. Fig. 4 presents such a plot obtained from the determination of calcium. One feature of Fig. 4 is particularly worthy of note, viz., the calcium concentration values of the samples are not evenly distributed across the range encountered. This is not surprising given that the samples were not selected for any single, particular element but merely as representatives of the class of materials. For calcium, one sample has a considerably higher concentration than the others, hence with a miss-one-out validation strategy this sample will be determined by extrapolation from the calibration model developed using the remaining samples. It is not surprising in these circumstances that OLS and
MLR fail to predict accurately the concentration of this sample. It is noteworthy that PLS regression performs well and is better for extrapolation. Other elements in these samples exhibiting a similar heterogeneous distribution of concentrations with a single outlying sample are manganese and aluminium. For manganese, both the PLS and MLR models provide similar, good prediction values for the outlying sample, and for aluminium PLS and OLS perform equally well.

**Conclusion**

The results presented here, obtained from a diverse range of geological samples, demonstrate the potential for an automated variable selection and calibration procedure for quantitative XRF analysis. By providing an automatic means of selecting the spectral regions suitable for calibration, and which can be demonstrated as chemically relevant, and then applying PLS regression, a general calibration scheme is afforded that requires minimal input from the user and does not rely on the user’s prior experience or knowledge of the samples. Results for PLS regression are as good as or better than those obtained using more conventional methods, indicating that the technique may be used as a general model for XRF analysis. It remains to evaluate the performance of the procedure on a wider range of sample types and more analytes, including its ability to correct for matrix, physical and spectral interference effects, and to integrate the software with the qualitative expert system analyser.

**References**


*Received September 30, 1997*

*Accepted December 15, 1997*