minerals MDPI

1. Multivariate Polynomial Regression Analysis

A multivariate polynomial regression of Raman parameters against temperatures was performed using the Multivariate polynomial regression program in Matlab [1]. Multiple regression analysis is generally used to represent the relationships between a dependent variable against several indipendent variables [2]). Depending on the relation between the dependent and each independent variable, linear or non-linear regression can be adopted. Polynomial regression is a form of non-linear regression in which the dependent/independent variables relationship can be modelled as an nth degree polynomial equation. In this work, the regression fit is performed with the least square (LS) method that calculates the sum of the squared errors to find a set of estimators that minimizes the sum, as in Equation (S1).

A general polynomial regression can be expressed by:

$$y = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \cdots + \beta_k x_i^k + e_i \quad \text{for } i = 1, 2, \ldots, n \tag{S1}$$

where k is the degree of the polynomial, y is the value of the dependent variable Y, x_i is the value of the independent variable X for the i^{th} case, β_0 is the Y-intercept of the regression surface, each $\beta_{1,2 \ldots k}$ is the slope of the regression surface with respect to variable X, n is the number of predictors and e_i is the random error component for the i^{th} case. It should be emphasized that while polynomial regressions assure high-quality reproduction of the analytical data, caution must be used in extrapolating the results beyond the range of observations (T in this study).

The ability of the model to capture the analytical trends can be expressed by the R^2 and mean absolute error (MAE) values. R^2 gives the percentage of variance that all independent variables in the model explain collectively, whereas MAE measures the accuracy of forecast between different items or products [2] and is expressed by the equation:

$$MAE = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{S2}$$

Cross-validation testing was performed to estimate the accuracy of the predictive model. In general, cross-validation methods learn and test all possible ways to divide the original dataset into a training set and a validation set [3]. In the classical validation approach, the split of the dataset is usually made randomly into two. However, this can lead to high variability, depending on the selection of the training set. To overcome this problem, in this work we adopted a *Leave-one-out-cross*-validation (LOOCV). LOOCV generates estimates for n models where n is the length of dataset observations. For each model, one observation is left out to be used as test, whereas the remaining n-1 observations are used as a training set until the loop is complete. Repeated iteration allows for non-random sampling of the training set, thus reducing variability in the generation of the estimates.

1.1. Multivariate Polynomial Equation

A second-order polynomial equation was found to satisfactorily correlate Raman parameters against the dependent variable (temperature) with a R^2 of 0.96 and a MAE of 4% (Tabble 3). In general, a R^2 value of 0.9 or above denote a very good fit, while a MAE value of 4% (0.04) indicates accurate forecasting (MAE<10%).

The resulting equation is the following form:

$$
\begin{aligned}
T_{predicted}(^{\circ}C)& \\
= {} & 16683.09 \times wD/wG + 70.49 \times wG + 2.32 \times wG \times wD/wG \\
& - 1674.44 \times pD - 11.41 \times pD \times wD/wG - 0.052 \times pD \times wG \\
& - 1319.62 \times \Delta D - G - 7.96 \times \Delta D - G \times wD/wG - 0.029 \times \Delta D - G \times wG \\
& + 0.82 \times \Delta D - G \times pD + 1249295.60 + 0.50 \times \Delta D - G^2 + 0.55 \times pD^2 \\
& + 0.01 \times wG^2 + 105.83 \times (wD/wG)^2
\end{aligned} \tag{S3}
$$

Table S1. Error values from multiple polynomial regression. Acronysms: MAE: mean absolute error; MAESTD: mean absolute error standard deviation; CVR²: cross validation R-square; CVMAE: cross validation mean absolute error; CVMAESTD: cross validation mean absolute error standard devia-tion.

R^2	0.9583
MAE	0.0422
MAESTD	0.0405
CVR^2	0.9478
CVMAE	0.0451
CVMAESTD	0.0473

The quality of the regression is further confirmed by the high regression coefficient and low mean absolute error after LOOCV (respectively CVR^2 and CVMAE in Table S1).

References

1. Cecen, A. Multivariate Polynomial Regression 2020. Available online: https://www.github.com/ahmetcecen/MultiPolyRegress (accessed on 28 January 2022).
2. Ostertagová, E. Modelling using polynomial regression. *Procedia Eng.* **2012**, *48*, 500–506, doi:10.1016/j.proeng.2012.09.545.
3. Rennie, J.D.M. On the value of leave-one-out cross-validation bounds. In *Computers in Biology and Medicine*; Elsevier: Amsterdam, The Netherlands, 2003; pp. 123–129.