O. Haavisto and H. Hyötyniemi. 2008. Partial least squares estimation of mineral flotation slurry contents using optical reflectance spectra. In: Proceedings of the 11th Conference on Chemometrics in Analytical Chemistry (CAC 2008). Montpellier, France. 30 June - 4 July 2008, volume 2 (Process Analysis), pages 81-85.

Partial least squares estimation of mineral flotation slurry contents using optical reflectance spectra

O. Haavisto H. Hyötyniemi
Helsinki University of Technology (TKK)
Department of Automation and Systems Technology
P.O.Box 5500, FI-02015 TKK, Finland
olli.haavisto@tkk.fi, heikki.hyotyniemi@tkk.fi

Keywords: Partial least squares, orthogonal signal correction, locally weighted regression, mineral flotation, reflectance spectroscopy.

1 Introduction

Froth flotation is one of the most important concentration methods in mineral processing (see e.g. [8]). Before flotation the ore is mixed with water and ground to small particle size in mills. The formed slurry is then processed in flotation cells, where the valuable mineral particles are lifted by air bubbles to the froth layer on the surface of the cell. The froth is allowed to overflow or skimmed to obtain the high-grade concentrate, whereas the low-grade slurry is removed from the bottom of the cell.

In order to effectively monitor and control a flotation process consisting of several flotation cells and slurry lines connecting them, it is essential to be able to assay the main slurry lines in real time and on-line. It was shown in [3] that visual and near-infrared (VNIR, $400-1000$ nm) diffuse reflectance spectroscopy can be successfully utilized as a supplemental method for determining the elemental contents of zinc concentrate slurry along with the traditional X-ray fluorescence (XRF) analyzer. The advantages of the spectral approach with respect to the XRF analysis are—as in many other applications—speed and simplicity. Spectral measurements combined with recursive partial least squares (rPLS) calibration (see e.g. [7]) improve the originally sparse XRF measurements providing a practically continuous estimate of the slurry contents.

Due to variations in slurry properties the calibration model requires recurrent updating. Fortunately, this is possible since new calibration samples are received regularly from the XRF analyzer. The main problem addressed in this study is how one should utilize these new data to optimally model the current relationship between the VNIR spectrum and the slurry element contents without losing information for wider calibration model relevance range from rare large concentration disturbances in the past. In [3] the learning was achieved by recursively updating the PLS model with nonlinear data preprocessing and rather fast exponential forgetting. The approach works well for the elements whose concentrations vary continuously, so that despite the forgetting the calibration model remains robust. However, in the (rare) case of process failures there occur sudden and large concentration changes which the current model cannot predict well, especially if a long time has passed since the previous such failure.

The aim of this work is to achieve a more robust modeling approach for the slurry element contents by using advanced data preprocessing and PLS calibration methods. To remove the effects of varying solids content and particle size from the measured spectra, orthogonal signal correction (OSC) [9] preprocessing method is applied. Additionally, the originally used recursive PLS is compared with a modified locally weighted regression (LWR) PLS algorithm that is suitable for continuous adaptation with new data.

2 Material and methods

The data for this study were obtained from the final zinc concentrate slurry flow in the concentration plant of Inmet Mining Corporation's Pyhäsalmi mine (located in Finland). The spectra were measured using an imaging spectrograph having the wavelength range of 400–1000 nm and nominal spectral resolution 5 nm. On-line measuring was performed by leading a constant slurry flow through a jet flow cell with a sapphire window. The window was illuminated with a regular halogen light source and the spectrum of the light reflected from the slurry was measured. The XRF analysis data were collected from the XRF analyzer of the concentrator plant. For more information on the equipment used in data collection, please refer to [3].

In Pyhäsalmi, the solids in the zinc concentrate typically contain about 55% zinc, 33% sulfur, 8% iron and 0.5% copper, which are the elements continuously measured by the XRF analyzer. The solids content of the flow is around $30 - 40\%$. However, during a process failure the zinc content decreases, whereas the iron and copper contents may significantly increase. In this study, the main focus is on modeling the copper contents where the relative changes are largest.

2.1 Collected Data

In total about 4500 valid sample pairs (matching spectrum and XRF measurements) were collected during the measurement period of 50 days. The data collection was performed continuously using the sampling interval of the XRF analyzer (about 16 minutes), excluding a small number of samples that were omitted due to the disturbances or maintenance breaks either in the XRF analyzer or spectrum measurement equipment. To reduce the amount of data and speed up the computations, the measured wavelength range was averaged to 20 values reducing the dimension of the spectrum from 960 wavelength values to 20. This was not detected to decrease the modeling performance. Additionally, a constant regressor variable was added to all spectra so that no data centering was required.

The collected flotation slurry data set was divided into two parts: the first 1000 samples were reserved for initializing the models, whereas the rest of the samples (3500) were utilized in model performance validation. The beginning of the initialization data contained two short but major process failures, where the copper concentration reached as high values as $4.5 - 5\%$ (Fig. 1). Near the end of the validation data set a third process failure with a very sharp increase in the copper content was detected.

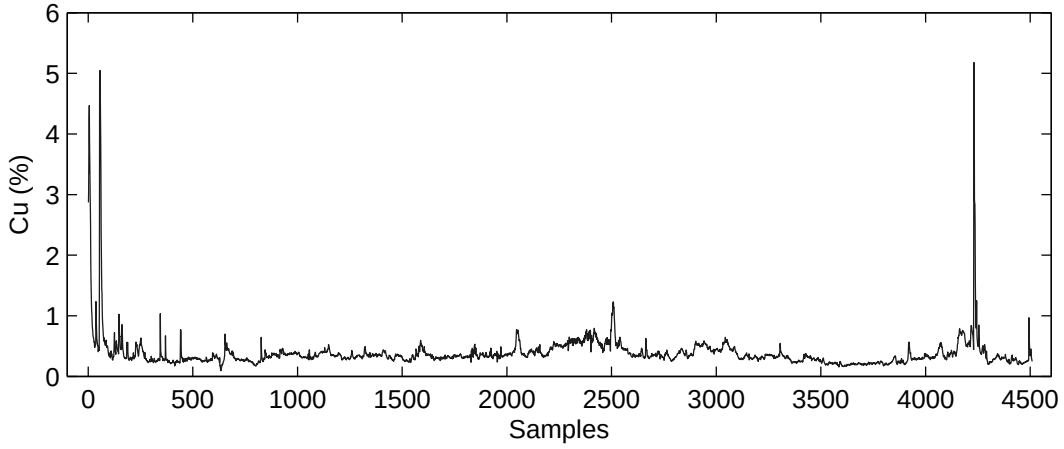

Figure 1: Copper content variation in the used data set.

All modeling was performed recursively and maintaining the temporal ordering of the samples to simulate the application of the models in on-line process monitoring. For validating the different calibration methods,

the root mean square error (RMSE) for all the validation samples was calculated using always the prediction given by the "old" model, i.e. the model updated with all the samples preceding the current sample. Thus, RMSE describes well the general prediction ability that can be achieved in real applications. However, since the main interest of the spectral analysis is to detect sudden major process changes as soon as possible, also the RMSE during the copper increase related to the process failure in the validation data was used to compare the modeling performances. This error (RMSE2) was calculated as the root mean square prediction error for the eight validation samples where the copper content was greater than 1.5% due to the process failure.

2.2 Continuous Locally Weighted Regression (CLWR)

Locally weighted regression (LWR) introduced in [1] is a method to extend the standard global linear regression into smooth nonlinear cases. The initial part of LWR is to collect a representative calibration data set. Then, for each new estimation sample $\mathbf{x}_{\text{est}}$, a new local linear model is calculated from the N_{est} nearest calibration data samples around $\mathbf{x}_{\text{est}}$. This model is then used to obtain the corresponding prediction $\hat{\mathbf{y}}_{\text{est}}$. Although originally presented only for least squares regression, LWR can also be combined with latent variable based methods [5] like principal component regression (PCR) and PLS.

In real life it is seldom enough to rely on a fixed calibration data set. Especially if the calibration model is to be used continuously for longer periods in a real process environment, the calibration samples tend to become less representative due to changes in the process. If new calibration samples are continuously available, better results are obtained when the data set is updated with them. As stated in Section 1, the use of exponential forgetting is not optimal in this situation, because one wants the model to remember the rare process failures. In the literature, some approaches are presented for how to utilize variable forgetting factor in order to maintain the required information (see e.g. [4]).

In this study, however, a simple method for discarding old calibration samples is utilized. Assuming that the current calibration regressors (spectra) are collected in the N rows of matrix $\mathbf{X}(k)$ and the corresponding responses (content values) in the rows of $\mathbf{Y}(k)$, the updating of the data set with the next calibration sample pair $\{\mathbf{x}(k+1), \mathbf{y}(k+1)\}$ is performed as follows: If the calibration set is still incomplete ($N < N_{\text{max}}$), the new sample is appended to the end of the data matrices to form the new calibration data sets $\mathbf{X}(k+1)$ and $\mathbf{Y}(k+1)$. If the maximum number of allowed calibration data is reached ($N = N_{\text{max}}$), one of the two most similar (measured by euclidian distance) old regressors is first searched:

$$i_{\text{min}} = \arg\min_i(|\mathbf{X}_i(k) - \mathbf{X}_j(k)|), \quad j = 1, \ldots, i-1, i+1, \ldots, N, \tag{1}$$

where $\mathbf{X}_i(k)$ is the ith row of $\mathbf{X}(k)$. The found sample is then replaced with the new calibration sample:

$$\mathbf{X}_i(k+1) = \begin{cases} \mathbf{x}(k+1), & \text{when } i = i_{\text{min}} \\ \mathbf{X}_i(k), & \text{when } i \neq i_{\text{min}} \end{cases} \text{ and } \mathbf{Y}_i(k+1) = \begin{cases} \mathbf{y}(k+1), & \text{when } i = i_{\text{min}} \\ \mathbf{Y}_i(k), & \text{when } i \neq i_{\text{min}}. \end{cases} \tag{2}$$

The purpose of the presented data discarding method is to maintain as wide calibration data set as possible that still contains the newest data available.

When a new (preprocessed) spectrum $\mathbf{x}_{\text{est}}$ is obtained, the prediction of the corresponding response $\hat{\mathbf{y}}_{\text{est}}$ is now based on the calibration data set and local regression. To start with, a fixed number (N_{est}) of samples in the data set $\mathbf{X}(k)$ that are closest (euclidian distance) to $\mathbf{x}_{\text{est}}$ are sought and used to form the estimation data set $\mathbf{X}_{\text{est}}$. The corresponding responses $\mathbf{Y}_{\text{est}}$ are selected from $\mathbf{Y}(k)$. Next, a local PLS model $\mathbf{B}$ is calculated using the estimation data $\{\mathbf{X}_{\text{est}}, \mathbf{Y}_{\text{est}}\}$ and the predicted value is obtained as

$$\hat{\mathbf{y}}_{\text{est}} = \mathbf{B}\mathbf{x}_{\text{est}}. \tag{3}$$

For the PLS estimation, the fast kernel-based algorithm presented in [2] was applied.

In addition to CLWR modeling, it was detected that bias correction (BC) improves the final predictions. Thus, after every new calibration sample, the current bias was calculated and then removed from the next prediction.

3 Results and discussion

Different PLS-based modeling approaches for the zinc concentration data were compared in this study. The starting point was the algorithm used in [3], where the recursive PLS algorithm (9 latent variables and forgetting factor 0.96) was applied to the reflectance data R that were preprocessed by rising the values to the power of $3/2$. There the nonlinear preprocessing was detected to give better results than when using the data as such. A more standard method to preprocess the data is to transform the spectra into absorbance values ($log(1/R)$) [6]. In addition to that, the regressor variations that are uncorrelated with the predicted values can be diminished using orthogonal signal correction (OSC). These two preprocessing methods with rPLS and CLWR were compared as shown in Table 1.

The OSC preprocessing was performed so that OSC was calculated for the initialization data (1000 samples), and the first latent variable was then removed from the whole data. Removing also the second latent variable was detected to generally decrease the modeling performance. The data set sizes $N_{\max}$ and N_{est} for the CLWR algorithm were selected by minimizing the RMSE, and the used values were $N_{\max} = 600$ and $N_{\text{est}} = 300$.

Preprocessing	Model	# LV	Q^2	RMSE	RMSE2
$R^{3/2}$	rPLS	9	0.90	0.064	0.99
$\log(1/R)$	rPLS	12	0.92	0.055	0.67
OSC	rPLS	11	0.91	0.058	0.73
$\log(1/R)$	CLWR + BC	8	0.89	0.070	0.83
OSC	CLWR + BC	6	0.92	0.060	0.43

Table 1: Comparison of all the modeling performances.

Table 1 lists the coefficient of determination Q^2, RMSE and RMSE2 for all the methods as well as the number of latent variables (#LV) used in the models. For each modeling approach, the optimal number of latent variables was selected by minimizing the RMSE. However, if increasing the model complexity did not improve the RMSE notably, the simpler model was chosen. Usually the same model was also optimal or close to optimal with respect to RMSE2.

In general the rPLS models required more latent variables than when CLWR was used. This is natural because each model in CLWR only describes simpler local dependencies, whereas rPLS applies for the whole data range. When absorbance data or OSC data preprocessing was utilized, the rPLS models performed better than the original model. As expected, with OSC preprocessing one latent variable could be omitted from the optimal rPLS model. However, the modeling performance was also slightly decreased.

Correspondingly, the effect of OSC preprocessing was significant in the CLWR models: Only 6 latent variables were necessary when OSC was used, and clearly better predictions were obtained with than without the preprocessing. This is most probably due to the fact that after the OSC preprocessing the search of similar regressors for the estimation data gives more uniform data sets. When CLWR is combined with OSC preprocessing, almost as good general modeling performance is obtained as with the best rPLS model.

The most interesting performance measure, however, is the ability of the models to predict the sudden increase in the copper content. This was evaluated by RMSE2, which represents the error of the model predictions during the process failure. Even though the data preprocessing and additional latent variables in rPLS models improved RMSE2, a significantly better value was obtained with the last modeling approach (OSC + CLWR + BC). The good performance is mainly due to the accurate prediction for the first test sample (Fig. 2), i.e. in the very beginning of the process failure. Even though the two rPLS models in general gave slightly more accurate predictions, they failed to predict this first test sample. That is because the previous large copper values had been forgotten in the exponential rPLS updating. On the contrary, the CLWR estimation data still contained

these values and thus the predictions given by CLWR were more accurate.

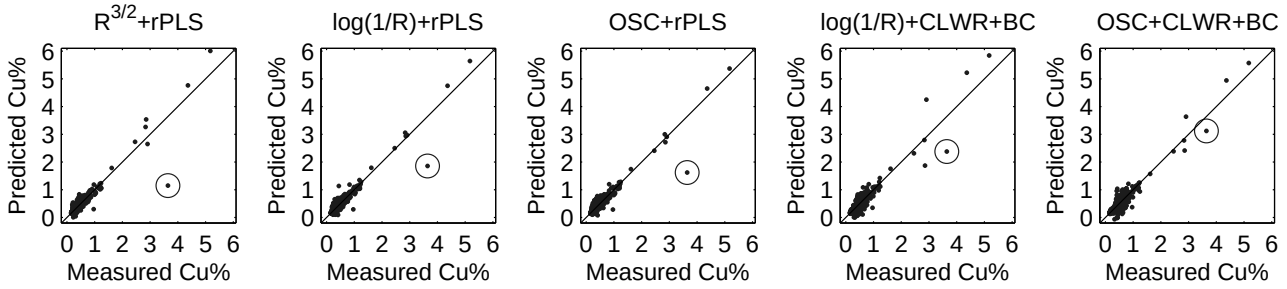

Figure 2: Validation data scatter plots for all the models. The first sample related to the process failure is circled.

The downside of the CLWR modeling is the slightly worse performance on the small response values. Additionally, the computational load of the algorithm is higher than in rPLS: the large calibration data set has to be stored in memory and for every prediction a new PLS model has to be calculated. These issues are to be considered in the future work.

4 Conclusion

Recursive partial least squares and continuous locally weighted regression approaches with and without orthogonal signal correction preprocessing were compared in this study. The aim was to find a robust calibration model for predicting mineral flotation slurry contents from VNIR reflectance data especially in the case where an unexpected major change in the process takes place after a long steady period. In general, this kind of changes are usually the most important ones to be predicted, since the faster for example a process failure is detected and corrective control actions started, the sooner the process can be restored to the normal state. It was shown that the otherwise accurate rPLS approaches did not perform well in the beginning of a process failure because of the exponential forgetting. On the other hand, the proposed CLWR model based on a continuously updated and OSC preprocessed calibration data set was able to predict the change significantly better.

References

[1] W. S. Cleveland, S. J. Devlin : Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403) : 596-610, 1988

[2] B. S. Dayal, J. F. MacGregor : Improved PLS algorithms, *Journal of Chemometrics*, 11 : 73-85, 1997.

[3] O. Haavisto, J. Kaartinen, H. Hyötyniemi : Optical spectrum based measurement of flotation slurry contents. *International Journal of Mineral Processing*, submitted in August 2007.

[4] H. Hyötyniemi : Information Content Weighting of Algorithms. *Preprints of the 10th IFAC Symposium on System Identification (SYSID'94), 1994*, Vol 3, pp. 417-422.

[5] T. Naes, T. Isaksson, B. Kowalski : Locally weighted regression and scatter correction for near-infrared reflectance data. *Analytical Chemistry*, 62 : 664-673, 1990.

[6] B. G. Osborne, T. Fearn, P. H. Hindle : *Practical NIR spectroscopy with applications in food and beverage analysis*, Longman, 2nd edition, 1993

[7] S. J. Qin : Recursive PLS algorithms for adaptive data modeling, *Computers & Chemical Engineering*, 22(4/5) : 503-514, 1998.

[8] B. A. Wills, T. J. Napier-Munn : *Wills' Mineral Processing Technology*, Butterworth-Heinemann, 7th edition, 2006.

[9] S. Wold, H. Antti, F. Lindgren, J. Öhman : Orthogonal signal correction of near-infrared spectra. *Chemometrics and Intelligent Laboratory Systems*, 44 : 175-185, 1998.