Preprints of the
9th International Symposium on Advanced Control of Chemical Processes
The International Federation of Automatic Control
June 7-10, 2015, Whistler, British Columbia, Canada

A Calibration Model Maintenance Roadmap

Barry M. Wise* and Robert T. Roginski*

**Eigenvector Research, Inc., Wenatchee, WA 98801
USA (Tel: 509-662-9213; e-mail: bmw@eigenvector.com).*

Abstract: Multivariate calibration, classification and fault detection models are ubiquitous in QbD (Quality by Design) and PAC and PAT (Process Analytical Chemistry and Technology, respectively) applications. They occur in the both the development of processes and their permissible operating limits, (*i.e.* models for relating the process design space to product quality), and in manufacturing (*i.e.* models used in monitoring and control). Model maintenance is the ongoing servicing of these multivariate models in order to preserve their predictive abilities. It is required because of changes to either the sample matrices or the instrument or response. The goal of model maintenance is to sustain or improve models over time and changing conditions with the least amount of cost and effort. This paper presents a roadmap for determining when model maintenance is required, the probable source of the response variations, and the appropriate approaches for achieving it. Methods for evaluating model robustness in order to identify models with lower ongoing maintenance costs are also discussed.

.

Keywords: Multivariate Calibration, Regression Modeling, Model Updating.

1. INTRODUCTION

Model maintenance can be roughly defined as the ongoing upkeep of (primarily) multivariate calibration and fault detection models in order to maintain their predictive abilities. The goal of model maintenance is to preserve or improve models over time and changing conditions with the least amount of cost and effort. This document outlines the reasons model maintenance is required and some approaches for dealing with it. It is written primarily from the perspective of spectroscopic instruments but much of it is applicable to soft sensor models in the process environment.

1.1 Why is Model Updating Necessary?

The reasons that calibration models need updating can be divided into two cases. The first is when the calibration set simply needs to be expanded. In this case, nothing has really changed with the response of the instrument to specific analytes. But the addition of new analytes or other previously unseen variations (such as changes in particle size distribution or the drift of a chemical process to a new steady state) make the old models biased. In these cases the calibration space must be expanded. In order for multivariate models to ignore irrelevant variation, they must have samples exhibiting this variation in the calibration data. Thus, when new variations are added, new samples must be added which exhibit the variation.

The second case is when the samples are the same but the measurement system response function has changed. This is often due to changes in the measurement hardware (new light source, clouding of optics, wavelength registration shift). This is really the instrument standardization problem. Changes in measurement conditions (temperature, pH, pressure) and changes in the sample matrix can have similar effects though they aren't actually changes in the hardware.

2. DETERMINING WHEN MODELS NEED UPDATING

Simply put, models need updating when their performance degrades. However, it is not always a simple matter to determine when that has occurred. Methods for identifying model degradation can be divided into those that rely on external validation samples and those that rely on internal model diagnostics.

2.1 External Validation Samples

The most obvious way to determine when a model is not predicting well is of course by checking the sample property predictions versus the reference method. Various SPC rules can be used on the deviations between the values predicted by the model and the reference values (such as the Western Electric rules). Procedures for doing confirmatory reference measurements are industry and applications specific and there is no "one size fits all" strategy. In many industries, especially pharmaceutical, a risk-based approach is preferred where limits are established based on potential system faults.

2.2 Model Diagnostic Measures (Q, T^2, etc.)

In the absence of external validation samples, or in between available validation samples, it is possible to use the multivariate model diagnostics. Though these go by somewhat different forms and names, they generally consist of some sort of model residual (that measures the orthogonal difference between a sample and the modelled data) and a leverage (that measures how far a sample is from the center of the data set, typically in some weighted fashion to account for some directions being more common than others). We

prefer the residual measure known as Q and Hotellings T^2 for leverage (Jackson 1991).

It is convenient to partition "uniqueness" between these two measures as they have different complementary interpretations. High T^2 samples have the same directions of variation as the calibration samples but are more extreme, while high residual samples exhibit new variations. In spectroscopic applications, high T^2 samples tend to be made up of the same analytes as the calibration data, but at extreme concentrations or unusual combinations of concentrations. Samples with high Q values have new analytes that make the samples unique, or other new variations (changes in the instrument or measurement conditions, etc.).

As with external validation, SPC rules can be used to monitor Q and T^2 values for each sample and detect when they are trending towards unacceptable values.

2.3 Setting Action Limits on Model Performance

While model prediction performance and internal diagnostics can be monitored, a key question is "At what point should action be taken?" With regard to agreement of predictions with external validation samples, answering this is fairly straightforward: choose the degree of deviation that pushes the instrument beyond the accuracy that is required for the application. Implicit in this assessment of the degree of deviation is the time frame over which the performance is monitored; simply put, we must avoid taking action based upon a statistically insufficient number of observations.

Setting limits on model internal diagnostic measures is somewhat more problematic. Just because new samples are producing high Q and T^2 values doesn't guarantee that the predictions are bad, it simply indicates that they are not to be trusted. Consistently high Q or T^2 values indicate that the current data does not fall within the range of the calibration data and certainly the prediction should be checked against the primary reference method. At this point, adjusting the model should be considered.

Our experience is that model predictions suffer less on samples with high T^2 values than high Q values (relative to their 95% or 99% limit based on the calibration data). This makes sense if the response of the measurement system is linear with respect to the property of interest (*e.g.* analyte concentration) and primary interferences. In these cases it would be expected that the model should be able to extrapolate significantly beyond the range of the calibration data with reasonable accuracy. On the other hand, high Q values indicate new variations that, even in relatively small amounts, may result in biased predictions.

2.4 Example

A synthetic data set was made up to illustrate the previous points. Consider a system based on NIR spectroscopy with one analyte of interest, (iso-octane), and a number of interferents, (initially heptane, toluene, decane and eventually also xylene). Pure component spectra were calculated from a real NIR data set (the pseudo-gasoline data from PLS_Toolbox 2014) using Classical Least Squares (CLS). These pure component spectra were then used along with a structured noise model created from the original data to create calibration data and "normal" test /validation data. In these data sets only iso-octane, heptane, toluene and decane were present and had approximately the same mean concentration values and covariance.

Two additional data sets were created to illustrate common problems with prediction data. The first of these had the same four chemicals and the calibration data, but they had a different covariance structure than the calibration data resulting in many samples with unusually high T^2 values. In the final data set xylene was introduced as a previously un-modelled interferent resulting in samples with unusually high Q values. Each data set has 30 samples and 401 spectral channels.

A Partial Least Squares (PLS) regression calibration model was built to predict the concentration of iso-octane using mean centering and 5 Latent Variables (LVs). Figure 1 shows the calibration model applied to the test set. As expected the test data (red diamonds) fall substantially within the calibration model limits. Also as expected prediction errors (RMSEP = .22) are slightly larger than the calibration error (RMSEC = .16) but close to the error of cross-validation (RMSECV = .21).

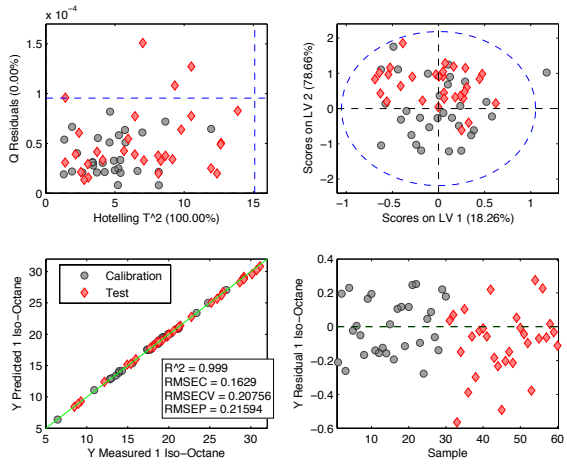

Figure 1. Calibration model applied to "normal" data including Q versus T^2 (upper left), scores on first two latent variables (upper right), predicted versus actual values for iso-octane (lower left) and prediction residual versus sample number (lower right). In all plots the calibration data is shown as grey circles while the test data is shown as red diamonds.

Figure 2 shows the calibration model applied to the high T^2 data set. The Q versus T^2 plot shows many test samples with very high T^2 values, some more than four times the 95% limit while Q values are only modestly higher. The scores plot shows that the test data largely fall outside the range of the calibration data. The predictions, however, have only moderately higher error despite the fact that they have gone completely outside the range of the original data. This is expected due to the linearity of the system.

Figure 3 shows the calibration model applied to the high Q data set. The Q versus T^2 plot shows many test samples with Q values 10-20 times the 95% limit indicating a new variation has been introduced into the data. This is of course the analyte xylene that was not present in the calibration data. The prediction error plot shows that the error gets progressively bigger versus sample number. This is because the concentration of xylene was ramped up over the data samples. But the concentration of xylene was not high relative to the other components, reaching a maximum of 6 (arbitrary units) while the other analytes had mean concentrations in the range of 14-38. The prediction error averages over 10 times the error on the original test set.

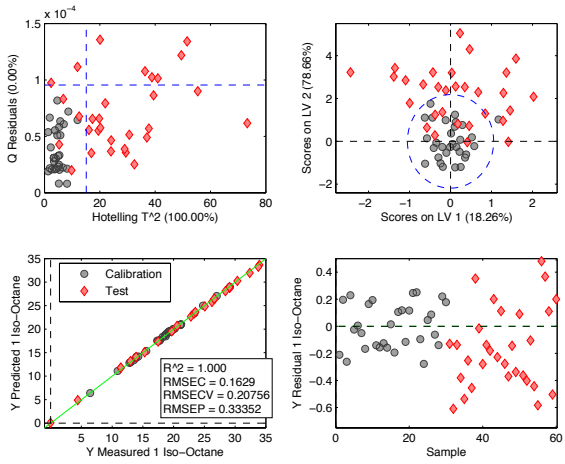

Figure 2. Calibration Model applied to high T^2 data. Individual plots as in Figure 1.

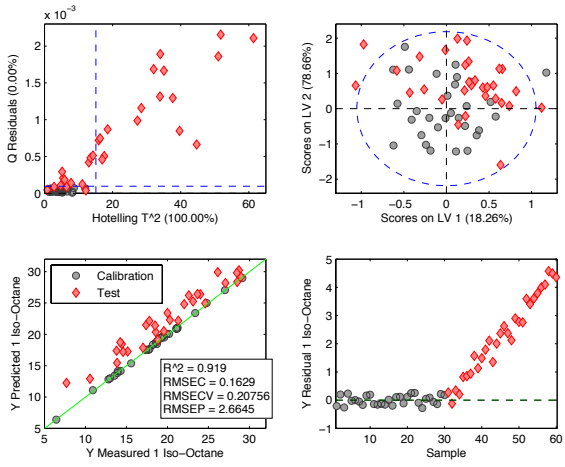

Figure 3. Calibration model applied to data with progressively high Q values. Individual plots as in Figure 1.

This example illustrates that in linear spectroscopic systems high T^2 samples might be tolerated if they do not also have high Q values. Samples with high Q values, however, are typically worse for prediction. The example also illustrates how some of the important prediction plots (Q versus T^2, scores plots) might look for this situation. Additional information can be obtained by investigating the Q residuals (Wise, 1989) now commonly known as contribution plots along with T^2 contribution plots (Wise, 1996).

3. MODEL UPDATING METHODS

There are many methods available for updating calibration models. The method of choice depends upon the nature of the change in the system. As a general principle, simple updating methods should be used first.

3.1 Slope and Bias Adjustments

Perhaps the simplest method for updating models is to post-process the predictions with a slope and bias adjustment (DiFoggio, 1995). The need for this would be apparent from comparison with reference validation samples. Slope/bias adjustments can be expected to work under a limited set of circumstances. In the event the samples have a new analyte in them at a fixed concentration, or a formerly fixed concentration analyte has moved to a new concentration, it would be expected that the model predictions would be in error by a constant value, *i.e.* bias. Other relatively simple effects might lead to gain changes, such as clouding of optics or changes in sample pathlength, which would require a change in the model slope.

Slope/bias adjustments, however, will not correct for any new variation in the data, such as a new analyte with a variable concentration.

3.2 Adding Samples to the Calibration Set

Under some circumstances, models can be updated by simply expanding the calibration set. For instance, in the event that a new analyte has been introduced into the samples, or a previously fixed analyte has begun to vary, then the model will produce incorrect predictions, and the model residual should indicate that the incoming samples are unusual. In this scenario, adding samples that exhibit the new variation to the calibration data may be sufficient. Multiple samples will likely be required in order for the model to capture the new variation. Methods for upweighting new samples, in order to minimize the number of samples required, exist (Stork 1999).

3.3 Use of Instrument Standardization/Calibration Transfer Procedures

Instrument standardization/calibration transfer methods are appropriate when the relationship between the samples and the data have changed due to changes in the instrument response, measurement conditions or sample matrix. Most of these methods attempt to map the response of the measurement system in its current state back to its response in its state during calibration. Methods include Shenk and Westerhuis, Direct Standardization (DS), Piecewise Direct Standardization (PDS, Wang 1991, 1992 and 1995), Spectral Space Transformation (SST Du 2011), Procrustes Analysis (PA, Anderson 1999), Artificial Neural Network (ANN) variants (Bouveresse 1996, Despagne 1998, Dreassi 1998), strategies based on modelling the model mismatch (Setarehdan 2002, Elizalde 2005), Wavelet Transforms (Greensill 2001), etc.

The vast majority of these methods require the use of transfer samples that are measured on the original instrument and the instrument to be standardized. Obviously, if using this method is contemplated when setting up a new measurement system, the transfer samples should be measured when the first calibration data is taken. Transfer samples ideally are similar to the calibration samples and exercise the instrument in similar ways. However, transfer samples can be more stable (non-perishable) samples such as rare earth oxide glasses. The key is that they have to vary in the areas of the spectrum upon which the calibration depends.

Other methods can be used to eliminate the differences between instruments while preserving the things they have in common. This includes Orthogonal Signal Correction (OSC, Wold 1998, Sjöblom 1998) Generalized Least Squares (GLS, Martens 2003) weighting and explicit drift correction (Gujral 2010). The down side of these methods is that they tend to reduce net analyte signal because they remove variation not common to both data sets. If this variation is similar to variation due to the analytes of interest, the subsequent calibration models may suffer. Finite Impulse Response (FIR, Blank 1996) filtering belongs in this group of methods and has the added feature that it does not require matched transfer samples.

The Extended Mixture Model (EMM, Martens 1991) and the very similar Prediction Augmented Classical Least Squares (PACLS, Haaland 2000) are modelling techniques that can be easily updated when new components are added to a calibration set. In practice, it is only important to model the subspace of the new components rather than the pure components themselves. Thus, the calibration set can be augmented with a basis for new variation, such as the loadings from a PCA model, or in some instances, polynomial baseline functions.

Of the methods listed above, we have had the best results with DS and PDS. Some recent work suggests SST as a viable alternative as well. The main advantage of PDS is that it is able to work with a very small number of transfer samples, as few as three in some instances. However, as more transfer samples become available (10 or more), other methods, such as DS and SST, may outperform it. Also, all of these methods preserve net analyte signal and have the ability to map features at one wavelength to a different wavelength.

The most commonly used standardization methods are compared in Table 1.

3.4 Automatic Model Updating

Methods for automatically updating multivariate linear regression models are well established. Recursive least squares (RLS, Hayes 1996) can be used to update MLR models as new data becomes available. It would be possible, for instance, to automatically update an MLR model every time a new reference value as available.

RLS can be formulated either with or without "forgetting." Without forgetting, models converge to the same result that would be obtained if the all the data had been used to develop the model in the first place. When forgetting is used, however, past data is progressively deemphasized, leading to a model that is based primarily on recent data. An adjustable parameter controls how fast past data is forgotten. The trade-off becomes that of adapting to new variations quickly versus basing the model on too little data resulting in a model with large prediction errors.

Table 1: Comparison of Properties of Common Standardization Methods.

Method	Number of meta-parameters	Y values *not* required ?	Use original calibration model?	Spectra un-modified?	Transfer sets *not* function of Y?	Retains net analyte signal?	Can use generic standards ?	Number transfer samples required
DS	1	✔	✔	✔	✔	✔	✔	High
PDS	2	✔	✔	✔	✘	✔	✔	Low
SST	1	✔	✔	✔	✘	✔	✔	Medium
GLS	1	✔	✘	✘	✔	✘	✔	Medium
OSC	2-3	✘	✘	✘	✘	✘	✘	Medium

Methods for automatically updating PLS and PCR type models are less well established (Helland 1992, Qin 1998). Various approaches have been attempted with some success. While these methods have shown great utility in some instances, it is hard to imagine that automatic updating methods would be allowed in a regulated environment.

3.5 Complete Recalibration

Complete recalibration should only be considered as a last resort. It is done in the event that the original model has essentially no relevant information, which is unlikely.

4. THE MODEL MAINTENANCE ROADMAP

All the pieces discussed above come together in the flowchart shown in Figure 4. The starting point of the roadmap is a preliminary study and initial calibrations. After this a maintenance plan should be developed. Once the model is online its performance is monitored via its own internal diagnostic measures and (hopefully) regular comparison to external reference methods.

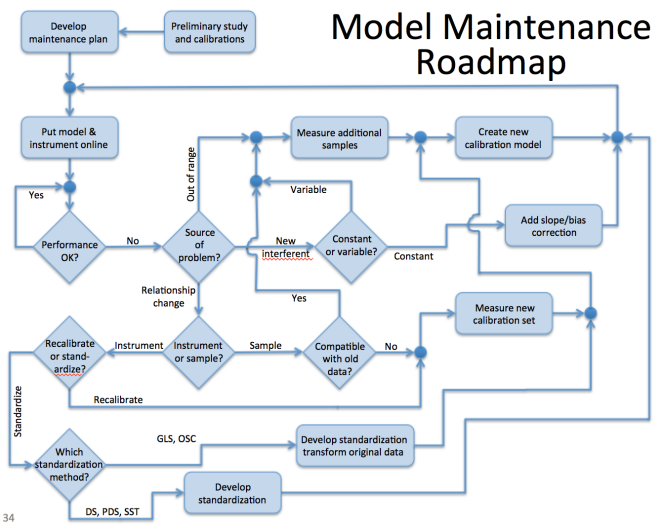

Figure 4. Model maintenance roadmap flowchart

When a performance problem is detected the first step is to establish the source of the problem. If the problem is simply out-of-range samples, additional samples can be measured

and the calibration model redeveloped. If a new interferent is detected, samples can also be added to the initial calibration set. Alternately, if the new interferent is constant, it may be possible to use a slope and bias adjustment on the existing model output.

If the source of the model performance degradation is a change in the relationship between the measurement variables and the output, the reason for this should be established. If it is a sample matrix problem it may or may not be possible to add new samples to the existing calibration set depending upon the degree of the mismatch. If it is not possible, a whole new calibration set may be required.

If instrument change is the source of the problem, there are many standardization methods available and a choice must be made. Complete recalibration is of course an option, but a very undesirable one. "Filtering" methods such as GLS and OSC are used on the original data and on any new samples and so require rebuilding the calibration model. "Mapping" techniques such as DS, PDS and SST can be applied to the new data and then the original calibration model can be applied directly.

5. AVOIDING MODEL UPDATING ALTOGETHER

Sometimes updating models is unavoidable due to significant changes in the system, either the samples or the hardware. However, care during the calibration process can help minimize the chances that updating will be needed. For instance, certain data pre-processing schemes affect the robustness of multivariate models. The use of spectral derivatives makes models very sensitive to changes in the wavelength registration. Even very small changes of a fraction of a channel can make model performance degrade significantly. An example of this is shown in Figure 5, where the prediction error for a series of models (for moisture in corn) with different preprocessing is shown as a function of registration shift in the spectrometer.

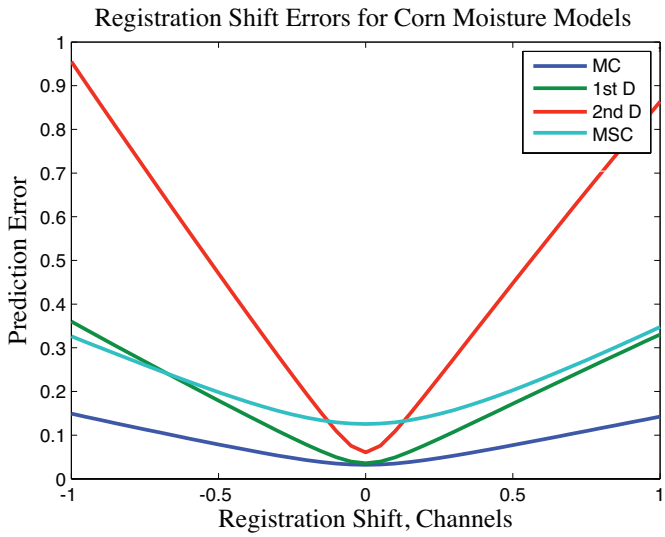

Figure 5. Model prediction error as a function of registration shift for models with mean centering, first derivative, second deriviative and multiplicative scatter correction.

On the other hand, the addition of data smoothing, which often does not improve a model under initial conditions, can improve robustness to registration shifts. Variable selection can also be a significant factor in the robustness of models (Swierenga 1998 and 1999).

The number of factors or LVs in models also affects their robustness over time. Models with too many LVs tend to be more brittle and predictions suffer considerably when even minor new constituents are added to the samples. Models with fewer LVs tend to not perform as well initially but not degrade as severely over time.

The robustness of models to new interferents may be tested by adding peaks of variable width to the data and translating them across the wavelength axis. This process can be repeated for models with different numbers of LVs as illustrated in Figure 6. In Figure 6 the segment for each number of model LVs shows the prediction error for very narrow to very wide interferents top to bottom and interferent center location side to side. It can been seen that while models with fewer LVs are never as good (as indicated by their lighter blue background) they also do not suffer as badly to interferents as models with more LVS (more intense areas of high prediction error). It is evident that the largest errors are indicated for the model with 8 LVs for narrow interferents near 1900 nm.

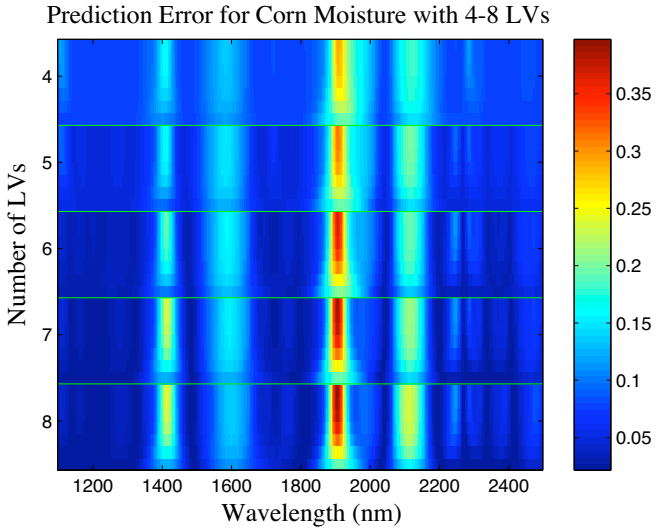

Figure 6. Model prediction error as a function interferent peak width, peak location, and number of model LVs.

As in the example with registration shift in Figure 5, different preprocessing schemes can affect the models ability to tolerate new interferents. The wavelength ranges considered can also impact model robustness and longevity.

We have seen where some modellers actually over-fit their models with the strategy of using them for a short period of time and subsequently update their models considerably more regularly. We don't consider this to be a conservative approach in either the amount of risk it involves or the level of effort needed to update the models much more frequently.

6. CONCLUSIONS

Many applications of PAC ultimately fail because it was assumed that the model-instrument system would operate forever as it did on day one. This is seldom the case and the installation plan should include a plan for updating the model going forward. There are many things to be considered in this plan, and no single procedure can address all possible changes in the system. The model maintenance roadmap, while likely not exhaustive, covers the most likely possibilities and can serve as a starting point for developing a plan for a specific application. Finally, modelling choices affect the robustness of the system to changes and should be considered during model development.

REFERENCES

Anderson, C.E. and Kalivas, J.H. (1999). Fundamentals of Calibration Transfer through Procrustes Analysis. *Applied Spectroscopy*, **53**(10), pps. 1268-1276.

Blank, T.B., Sum, S.T., Brown, S.D. and Monfre, S.L. (1996). Transfer of Near-Infrared Multivariate Calibrations without Standards, *Anal. Chem.*, **68**(17), pps. 2987-2995.

Bouveresse, E., Hartmann, C., Massart, D.L., Last, I.R. and Prebble, K.A. (1996). Standardization of Near-Infrared Spectrometric Instruments. *Anal.Chem.* **68** pps 982–990.

Despagne, F., Walczak, B. and Massart, D.L. (1998). Transfer of Calibrations of Near-Infrared Spectra Using Neural Networks. *Appl. Spec.*, **52**(5), pps 732-745.

DiFoggio, R. (1995). Examination of Some Misconceptions About Near-Infrared Analysis, *Appl. Spec.*, **49**, pps. 67-75.

Dreassi, E., Ceramelli, G., Perruccio, P.L. and Corti, P. (1998). Transfer of Calibration in Near-Infrared Reflectance Spectrometry. *Analyst*, **123**.

Du, W., Chen, Z.P., Zhong, L.J., Wang, S.X., Yu, R.Q., Nordon, A., Littlejohn, D. and Holden, M. (2011). Maintaining the predictive abilities of multivariate calibration models by spectral space transformation. *Analytica Chimica Acta*, **690** pps 64-70.

Elizalde, O., Asua J.M. and Leiza, J.R. (2005). Monitoring of Emulsion Polymerization Reactors by Raman Spectroscopy: Calibration Model Maintenance. *Applied Spectroscopy*, **59**(10), pps. 1280-1285.

Greensill, C.V., Wolfs, P.J., Spiegelman C.H. and Walsh, K.B. (2001). Calibration Transfer between PDA-Based NIR Spectrometers in the NIR assessment of Melon Soluble Solids Content. *Appl. Spec.*, **55**, pps. 647–653.

Gujral, P., Amrhein, M., Wise, B.M., and Bonvin, D. (2010). Framework for explicit drift correction in spectroscopic calibration models. J. Chemometrics, 24, pps. 534-543.

Haaland, D.M. and Melgaard, D.K., (2000). New Prediction-Augmented Classical Least-Squares (PACLS) Methods: Application to Unmodeled Interferents. *Applied Spec.* **54**(9), pps. 1303-1312.

Hayes, Monson H. (1996). *Statistical Digital Signal Processing and Modeling, section 9.4: Recursive Least Squares*. Wiley. p. 541.

Helland, K., Berntsen, H.E., Borgen, O.S. and Martens, H., (1992). Recursive algorithm for partial least squares regression. *Chemo. Intel. Lab. Sys.*, **14**, pps. 129–137.

Jackson, J.E., (1991), *A User's Guide to Principal Components*, Wiley.

Martens, H. and Næs, T. (1991). *Multivariate Calibration*, Section 3.6.3, Wiley.

Martens, H., Høy, M., Wise, B.M., Bro, R., and Brockhoff, P.B. (2003). Pre-whitening of data by covariance-weighted pre-processing. *J. Chemometrics*, **17**(3), pps 153-165.

Miller, C.E., (2010). Chemometrics in PAT in Katherine Bakeev, editor, "Process Analytical Technology: Spectroscopic Tools and Implementation Strategies for the Chemical and Pharmaceutical Industries," John Wiley and Sons.

PLS_Toolbox for use with MATLAB, Version 7.9, (2014). Eigenvector Research, Inc. Wenatchee, WA, USA.

Qin, S. Joe, (1998). Recursive PLS algorithms for adaptive data modelling. *Computers Chem Engng*, **22**(4/5) pps. 503-514.

Setarehdan, S.K., Soraghan, J.J., Littlejohn, D. and Sadler, D.A. (2002). Maintenance of a calibration model for near infrared spectrometry by a combined principal component analysis–partial least squares approach. *Analytica Chimica Acta,* **452** pps. 35–45.

Sjöblom, J., Svensson, O., Josefson, M., Kullberg, H., and Wold, S. (1998). An evaluation of orthogonal signal correction applied to calibration transfer of near infrared spectra. *Chemo. and Intell. Lab. Sys.*, **44**, pps. 229-244.

Stork, C.L. and Kowalski, B.R. (1999). Weighting Schemes for Updating Regression Models-a Theoretical Approach. *Chemometrics Intell. Lab. Syst.*, **48**(2), pps. 151-166.

Swierenga, H., de Weijer, A.P., Buydens, L.M.C., (1999). Robust calibration model for on-line and off-line prediction of poly(ethylene terephthalate) yarn shrinkage by Raman spectroscopy. *J. Chemom.* **13** pps. 237–249.

Swierenga, H., Groot, P.J., Weijer, A.P., Derker, M.W.J. and Buydens, L.M.C. (1998). Improvement of PLS Model Transferability by Robust Wavelength Selection. *Chemo. Lab Sys*, **41** pps. 237-248.

Wang, Y., Veltkamp, D.J. and Kowalski, B.R., (1991). Multivariate Instrument Standardization. *Anal. Chem.*, **63**(23), pps. 2750-2756.

Wang, Y., Lysaght, M.J. and Kowalski, B.R. (1992). Improvement of Multivariate Calibration through Instrument Standardization, *Anal. Chem.*, **64**(5), pps. 562-565.

Wang, Z., Dean T., and Kowalski, B.R., (1995). Additive Background Correction in Multivariate Instrument Standardization. *Anal. Chem.*, **67**(14), pps. 2379-2385.

Wise, B.M., Ricker, N.L. and Veltkamp, D.J., (1989). Upset and Sensor Failure Detection in Multivariate Processes. AIChE Annual Meeting.

Wise, B.M. and Gallagher, N.B., (1996). The Process Chemometrics Approach to Process Monitoring and Fault Detection, J. Process Control, **6**(6), 329-348.

Wold, S., H. Antti, F. Lindgren and J. Öhman, (1998). Orthogonal signal correction of near- infrared spectra, *Chemo. and Intell. Lab. Sys.*, **44**, pps. 175-185.

