Preprints, 10th IFAC International Symposium on Advanced Control of Chemical Processes Shenyang, Liaoning, China, July 25-27, 2018

A Systematic Framework for Assessing the Quality of Information in Data-Driven Applications for the Industry 4.0

Marco S. Reis *

** CIEPQPF - Department of Chemical Engineering, University of Coimbra Pólo II, Rua Sílvio Lima, 3030-790 Coimbra, Portugal (e-mail: marco@eq.uc.pt)*

Abstract: Managing and improving the quality of information generated in data-driven empirical studies is of central importance for Industry 4.0. A fundamental and necessary condition for conducting these activities is to be able to *measure* the quality of information – "*If you can not measure it, you can not improve it*" (Lord Kelvin). It is somewhat surprising that, with so many efforts devoted to take the most out of the available data resources, not much attention has been paid to this key aspect. Therefore, in this article we described and apply a framework, the InfoQ framework, for evaluating, analyzing and improving the quality of information generated in the variety of data-driven activities found in the Chemical Processing Industry (CPI). This systematic framework can be used by anyone involved in conducting these activities, irrespectively of the context and the specific goals to achieve. For instance, it can either be used to provide a preliminary assessment of the project risk, by analyzing the adequacy of the data set and analysis methods to achieve the intended goal, as well as to perform a SWOT analysis on an ongoing project, to improve it and increase the quality of information generated, i.e., increasing its InfoQ. The framework is applied to a real world case study in order to illustrate its implementation, utility and relevance. The author recommend its routine adoption, as part of the Definition stage in any data-driven task, such as in Lean Six Sigma projects, exploratory studies, on-line and off-line process monitoring, predictive modelling and diagnostic & troubleshooting activities.

Keywords: Quality of information; InfoQ; Industry 4.0; Big Data; Predictive analytics.

1. INTRODUCTION

More than ever, data abounds in the new era of Industry 4.0 and Big Data. Virginia Rometty's (CEO of IBM) well-known quote is a clear signal that the critical role of data is finally being acknowledged by all industry stakeholders: "*What steam was to the 18th century, electricity to the 19th and hydrocarbons to the 20th, data will be to the 21st century. That's why I call data a new natural resource*." The pressure is rapidly building up on enterprises around the world, to take the most out of this resource, in order to turn it into a source of competitive advantage, process/quality improvement and economic growth (Reis et al., 2016). The importance currently given to data is also being transferred to all connected elements – the communicating vessels principle. This means that technology for handling, storing and retrieving large amounts of data, as well as analytics to process them, are also under pressure to develop proper solutions and to keep the pace of the increasing demands imposed by Industry 4.0. Of course, this generates even more data, originating a virtuous cycle that is gaining momentum and outreaching all industrial sectors and activities (Reis et al., 2016).

With so much focus being given to data, it becomes critical to be able to measure the quality of information generated in data-centric activities. Examples abound where the mere use of data is not enough for achieving the analysis goals (Harford, 2014; Reis et al., 2016), and the reason lies on the low quality of information generated. Therefore, it is now the right time to develop and implement a systematic approach for assessing this fundamental aspect, to support the planning and implementation of data-driven activities, as well as their improvement.

In this article we present the concept of information quality, InfoQ, originally proposed by Kenett and Shmueli (Kenett and Shmueli, 2014, 2016), adapting it for the first time to the Chemical Processing Industry (CPI). InfoQ is defined as "*the potential of a data set to achieve a specific (scientific or practical) goal by using a given empirical analysis method*". InfoQ, depends upon a set of structuring aspects of any data-driven project, called the InfoQ-components, namely: the specific analysis goal, g; the available data set, X; the empirical analysis method, f; the utility measure, U. According to the definition of InfoQ, these elements are related with each other through the following analytical expression (InfoQ is the level of Utility, U, achieved by applying the analytical method f to the data set X, given the activity goal g):

$$InfoQ(f, X, g) = U\{f(X \mid g)\} \quad (1)$$

This article is organized as follows. Section 2 provides an overview the proposed InfoQ framework, together with its 8 assessment dimensions. Section 3 illustrates the application

Copyright © 2018 IFAC

of the framework to a real world case study from the Semiconductors industry. The paper is concluded with a summary of the contents of the article and prospects of future activities, in Section 4.

2. THE INFOQ FRAMEWORK

As mention in the previous section, the quality of information generated in an empirical study, InfoQ, depends on the quality of its 4 components:

- *Analysis goal, g*. The purpose of the analysis, in statistical or data science terms. A broad classification of goals include the following categories: descriptive/exploratory studies, predictive modelling and diagnosis/causal explanation activities.
- *Data set, X*. The data set used for accomplishing the goal. Data can arise from different sources, such as observational industrial data, data collected from planned experiments, laboratory data, computer simulations, etc., and with different structures.
- *Empirical analysis method, f*. The data analysis method adopted to process the data set *X*, in order to achieve the goal, *g*. Methods can be of different types, such as {parametric, semi-parametric, non-parametric}, {probabilistic, deterministic, algorithmic}, {linear, non-linear}, {single-block, multi-block}, etc.
- *Utility, U*. A measure of the extent to which the analysis goal, *g*, is achieved. It usually consists of suitable performance metrics such as *RMSEP* or R^2_{Pred} for predictive activities, measures of statistical power (e.g., p-values) for diagnosis, and goodness of fit and discrimination for descriptive goals.

The evaluation of InfoQ can be made directly upon the analysis of its components. Such unspecified multidimensional assessment process, raises however some questions of reproducibility and operationability, which will certainly affect the buy-in and adoption by industrial practitioners. Therefore, in order to make the assessment process well-defined and systematic, and to prevent overlooking important aspects to consider during the assessment of InfoQ, a set of 8 dimensions were proposed that should be explicitly addressed during the assessment process. They contemplate different aspects that are necessary, in general, to take into account for determining the value of information in a data-driven empirical study. These dimensions, $\theta = [D_1 \quad D_2 \quad \cdots \quad D_8]^T$, intervene in the quality of the four InfoQ components (*g, U, X, f*), in a way that may be different depending on the component under analysis. Therefore, instead of computing InfoQ by assessing directly the quality of the 4 components, one can do it indirectly, analysing the 8 underlying dimensions that structure their quality. The assessment should be made following the guidelines of the Delphi method, in order to avoid personal bias and converge to consensus decisions. These 8 dimensions are briefly described below (based on the initial proposal of Kenett and Shmueli, with some adaptations to make them applicable to the CPI context):

2.1 Data Resolution (D_1)

In the CPI context, resolution is usually connected to the aggregation level of data. One type of aggregation, regards data granularity. It often occurs that collected data may have different levels of granularity, meaning that their values regard the state of the process over different windows of time, during which measurements were collected and averaged, resulting in the end in a single aggregated value. This process results in recorded values representing averages of minutes, hours, days, weeks, shifts, production units (lots), etc. This is called multiresolution data. A distinct topic (but often confused with multiresolution), is multirate data. Multirate regards the existence of multiple acquisition rates, usually from instantaneous (high resolution) measurements (Rato and Reis, 2017; Reis and Saraiva, 2006a, 2006c; Willsky, 2002). In the scope of this InfoQ dimension, one considers the appropriateness of both data granularity and acquisition rate for the purposes of the analysis.

2.2 Data Structure (D_2)

Data structure refers to the type(s) of data and their characteristics, such as:

- Structured (arrays of numbers, cross-sectional, network data, time series) or unstructured (text, images, sound & vibration records);
- Tensor nature (0th-order, such as process sensors; 1st-order, such as spectra; 2nd-order such as images, etc.) (Reis and Saraiva, 2006b, 2012);
- Presence of noise, outliers, missing data, bad segments (plant shutdowns and transients) (Chiang et al., 2003; Reis et al., 2009; Walczak and Massart, 1997);
- Single-block or multi-block (i.e., when a single or multiple natural groups of variables exist and their integrity should be maintained) (Campos et al., 2017; Westerhuis et al., 1998);
- Static or time-delayed structure (meaning a lagged-correlation pattern) (Ku et al., 1995; Rato and Reis, 2013a, 2013b);
- Observational (i.e. “happenstance data”, using R.A. Fisher terminology) or Causal (namely collected following a DOE plan) (Box, 1957; Box et al., 2005).

2.3 Data Integration (D_3)

This dimension regards the existence of multiple sources of data that could convey relevant and complementary information for achieving the project goal, if properly integrated through *f*. They can arise from different points in the process (raw materials, operations, quality, customers, etc.) or form different measurement devices (process sensors, environmental data, laboratory analytical devices, etc.).

2.4. Timeliness or Temporal Relevance (D_4)

The extraction of knowledge from data happens in a workflow, roughly composed by the following stages: i) planning; ii) data collection; iii) data analysis; iv) deployment. Dimension D_4 regards the impact of the duration of each stage, and the gaps in between, on InfoQ.

2.5 Selection of Data and Chronology (D_5)

This dimension regards the variables selected and the temporal relationships between them, in the context of *g*. Much of the success of constructing models for process optimization and diagnosis goals, rely on having access to measurements of critical variability drivers. This is fundamental for developing input-output models for process control & optimization or to perform troubleshooting activities, but not so critical for process monitoring and soft sensor applications.

2.6 Generalizability (D_6)

This InfoQ dimension is relative to the potential to generalize the analysis outcomes to the desired universe targeted by the empirical study. Observational data allows inferences regarding similar operation conditions. On the other hand, the active collection of data (through DOE) enables the capability for exploring operation modes beyond those used before, generalizing inferences to other conditions. Therefore, this dimension assesses the ability of *X* and *f* to be extended to the circumstances of interest (established in *g*), as well as the adequacy of *U* to capture this performance.

2.7 Operationalisation (D_7)

This dimension addresses the complexity in operationalizing the empirical study within the existent capabilities of the company. It regards the difficulties involved in data collection, analysis and deployment of solutions. Timeliness (D_4) regarded the aspect of time, but here the emphasis is in the complexity in carrying out the several stages involved and the accessibility to the resources necessary to do it (other than time).

2.8 Communication (D_8)

This dimension comprises the rigour, completeness and clarity, with which the following aspects are established and communicated:

- The goals of the project – to the project team;
- The results obtained – to the project stake holders.

So, it regards both the quality of the project Definition stage, as well as the quality of the communication of the outcomes obtained to the relevant stakeholders, from which the project impact is, to a great extent, dependent.

3. OPERATIONALIZATION

The 8 dimensions described in the previous subsection (InfoQ-dimensions) should be properly combined in order to compute an InfoQ-score. A new InfoQ assessment strategy is proposed, that is based on the decomposition of InfoQ into its 4 components, and then onto the 8 dimensions that contribute to them: $\theta \rightarrow \gamma \rightarrow InfoQ$, where $\theta = [D_1 \quad D_2 \quad \cdots \quad D_8]^T$ represents the eight InfoQ-dimensions, and $\gamma = [g, U, X, f]^T$ stands for the 4-dimensional vector of InfoQ-components. This decomposition is depicted in Fig.1.

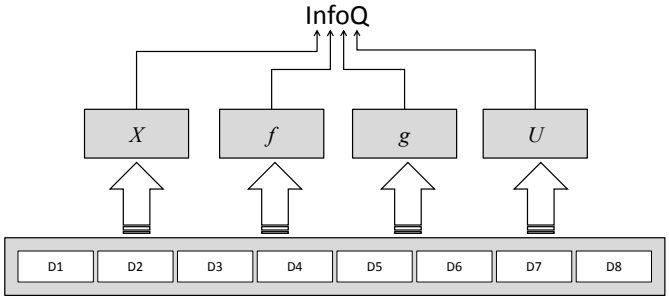

Fig. 1. The decomposition of InfoQ into its components (*X, f, g, U*) and the 8 dimensions that determine their quality.

Not all dimensions are relevant for assessing each component. Table 1 specifies which dimensions are actually considered in the assessment of each component.

Table 1. Summary table of the InfoQ-dimensions affecting the four InfoQ-components (*X,f,g,U*)

InfoQ-compon.(→) InfoQ-dimens.(↓)	***X***	***f***	***g***	***U***
Data Resolution (D_1)	✓	✓		
Data Structure (D_2)	✓	✓		
Data Integration (D_3)	✓	✓		
Timeliness (D_4)	✓	✓		
Selection of Data and Chronology (D_5)	✓	✓		
Generalizability (D_6)	✓	✓		✓
Operationalisation (D_7)	✓	✓		
Communication (D_8)		✓	✓	

The assessment is made in 3 stages, as detailed bellow.

3.1 Stage 1

For each component, C_j, the user assesses each dimension, D_i, connected to it (see Table 1) and computes the associated scores: $score-\mathbf{d}_i^j$. The assessment of each dimension, D_i, *w.r.t.* a given component, C_j, is made with resort to a Likert scale, with 5 levels, [1–5] with "1" indicating low achievement in that dimension and "5" indicating high achievement. These ratings, $\left\{\mathbf{d}_i^j\right\}_{i=1:8}$, are filled in by the user, and are then normalized using a desirability function with a scale [0–1], leading to the normalized assessment scores, represented by: $score-\mathbf{d}_i^j$.

3.2 Stage 2

This stage is of a computational nature, where the scores obtained from stage 1 for each component are combined to compute the scores for the quality of each component: $score-\mathbf{d}_i^j \rightarrow score-\mathbf{c}_j^{InfoQ}$. This data fusion operation is made through the weighted geometric mean of the individual desirabilities that are relevant to a given component. Contrary to the original approach, weights are now introduced, to reflect the different focus and priorities associated with the different analysis goals:

$$score-\mathbf{c}_j^{InfoQ} = \left\{\prod_{k\in\mathbf{I}^j}\left(score-\mathbf{d}_k^j\right)^{w_k^j}\right\}^{\frac{1}{\sum_{k\in\mathbf{I}^j} w_k^j}} \quad (2)$$

3.3 Stage 3

This is also a computational stage, where the component scores are combined to finally obtain the InfoQ: $score-\mathbf{c}_j^{InfoQ} \rightarrow InfoQ$. We do not consider different weights for the different components, which amounts to assume that they are all equally relevant for establishing InfoQ:

$$InfoQ = \left\{\prod_{j=1:4}\left(score-\mathbf{c}_j^{InfoQ}\right)\right\}^{\frac{1}{4}} \quad (3)$$

4. CASE STUDY

In this section, a case study is presented to illustrate the implementation of the proposed InfoQ framework in the context of CPI. The impact of the options followed at the level of the methods adopted, *f*, or regarding features present in the data set, *X*, are also brought to the analysis and discussed.

4.1 Description

This case study regards a semiconductor project (the name of the company cannot be disclosed), whose purpose was to derive an inferential model (virtual metrology) that could be used in the future for purposes of fast release of wafer batches or even for process control (run-to-run control). FDC data was provided by the semiconductor manufacturer (FDC means Fault Detection and Classification, and consists mostly of process operation variables, such as flows, pressures, temperatures, etc.), together with Metrology data for the key dimensions of the wafer. The FDC data regards almost 1000 wafer batches, but the Metrology data was collected for only approximately 50 batches, which furthermore do not always coincide with those in the FDC data set.

The team decided to fuse the two data sets (FDC and Metrology) using the wafer lot reference and developed inferential models using several predictive modelling approaches, such as least squares regression with variable selection (forward stepwise regression), penalized regression (LASSO) and partial least squares (PLS). The methods' performance was assessed using cross-validation. Good fitting and predictive scores were obtained for the least squares variable selection methodology.

4.2 InfoQ Assessment of the Initial Study

Implementing the workflow for InfoQ assessment (Stage 1), each component was evaluated using the dimensions that are relevant for its quality definition (see Table 1). The following paragraphs contain some observations of the ratings given to each dimension *w.r.t.* to a given component (*g, U, X, f*).

- *Assessing InfoQ-X*. Several datasets are available, namely FDC and Metrology data, but their integration is limited because the overlap of records for the same wafers is low. Therefore, the collection protocol could have been better designed from the standpoint of potentiating better integration capabilities (*D3*). The low superposition between datasets also causes many records to be discarded, leading to low resolution data (*D1*). On the other hand, the dataset took considerable time to be collected and made available to the analytics team, and the collection process was very complex (*D7*) – by the time it was analysed, the process may have suffered some changes, which may limit the deployment of results (*D4*). The data structure correspond to a 2-way table composed by observational or passively collected data (*D2*), and the main process variables were included in the analysis (*D5*), which are both positive aspects for developing a Virtual Metrology predictive model for this process (D6).
- *Assessing InfoQ-f*. The methods adopted are in general capable to deal with the features present in the dataset, such as multicollinearity, sparsity

and noise (*D1-D3*), and can be implemented in useful time and within the resources available in the team (*D4, D7*). The methods also have built-in features for selecting the relevant variables (*D5, D6*) and for generalization to the process of interest (namely parsimony and parameter estimation stability).

- *Assessing InfoQ-g*. It is not clear from the goal statement whether the objective is to develop a predictive model for Virtual Metrology or for Control/Optimization. A better goal definition is therefore needed (D8), as the nature of the models required for these two goals, differs.

- *Assessing InfoQ-U*. The performance of the predictive model was evaluated using cross-validation, which is a sound approach for assessing the predictive capabilities of the model, under situations where data is not so abundant (D6). However an independent test set would be a preferable solution in the future, especially if the purpose is to conduct process control.

The assessment of the initial study resulted in the scores for the components and for the InfoQ, presented in Fig 2. From the analysis of these results, one can verify that the overall quality of information is not very high (0,69), and the main concerns are in the InfoQ-components: data set and goal. Therefore these should be carefully analysed and solutions devised for their improvement, in order to increase the value of information generated in the study.

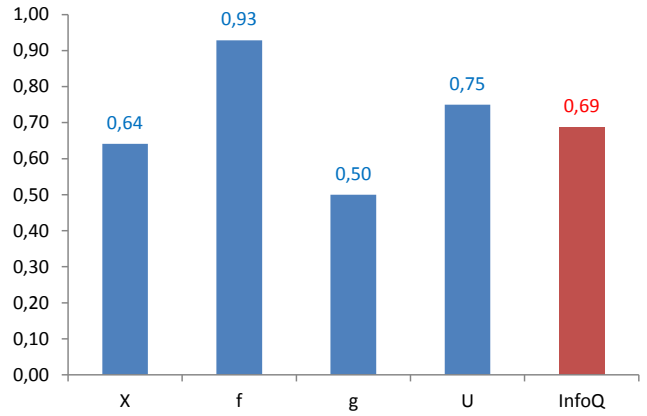

Fig. 2. Decomposed InfoQ assessment: initial study.

4.2 InfoQ Assessment of the Final Study

After closely analysing the elements of the initial study, and the InfoQ assessment performed, several improvement opportunities were detected, namely:

- The decision of data to be collected should result from a consensus analysis made by the process team and the analytics team and not only a decision of the process team. With this, better integration capabilities (D3) can be expected and the resolution of data will also be improved (D1).

- The goal definition must also be clearly defined, namely if it regards the development of a virtual metrology model, or if the purpose is to derive an input-output model for process control and optimization. This can make a significant difference on the type of models needed and the data structure required for analysis. For instance, input-output models for process control require the realization of system identification experiments, which were not contemplated in the original data collection plan.

- An independent test set should be collected, especially if the purpose is to conduct process control.

With this changes implemented in the future, the quality of information generated by the study can improve from the initial level of 0.69 to 0.92, indicating a significantly higher level of achievement of the project goals (Fig. 3).

As Fig. 3 depicts, there is an evolution in the assessment of each InfoQ-component from the initial to the final stage, with the implementation of the improvement initiatives, namely in *X, f* and *U*.

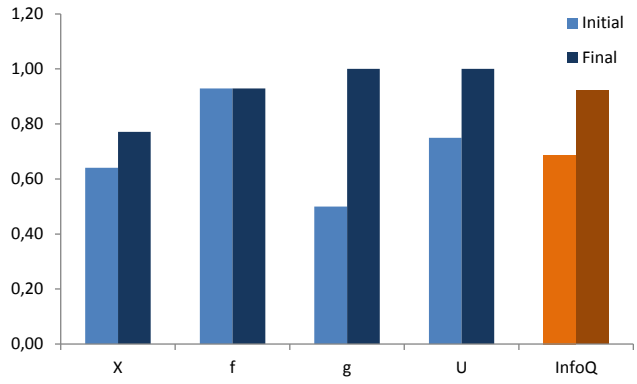

Fig. 3. Decomposed InfoQ assessment: final study.

6. CONCLUSIONS

In this paper, we presented a systematic framework for assessing the quality of information in data-driven empirical studies. This is a missing piece in most data analytics efforts, which we believe can bring further insights and contribute significantly to the improvement of their effectiveness.

The proposed framework can be used for:

- Planning and optimizing the implementation of data-driven activities in Industry 4.0

- Assessing the quality of information generated in data-driven empirical studies.
- A posteriori diagnosis and reporting of strengths and weaknesses of any data analysis activities (SWOT analysis).
- Tool for supporting decision making on how to improve the design or data-driven empirical studies, maximizing InfoQ.

Future work will contemplate the reporting and analysis of more applications of this methodology, with the purpose to support practitioners in developing their data-centric projects in the era of Industry 4.0

ACKNOWLEDGEMENTS

Special thanks to Ron Kenett for the challenging discussions around the topic of this paper, among many others. The author also acknowledges financial support through project 016658 (references PTDC/QEQ-EPS/1323/2014, POCI-01-0145-FEDER-016658) co-financed by the Portuguese FCT and European Union's FEDER through the program "COMPETE 2020".

REFERENCES

Box, G.E.P. (1957). Evolutionary operation: A method for increasing industrial productivity. *Applied Statistics - Series C,* 6 (2), 81-101.

Box, G.E.P., Hunter, J.S., and Hunter, W.G. (2005). *Statistics for Experimenters: Design, Innovation, and Discovery* (2nd ed.). Wiley, Hoboken, NJ (USA).

Campos, M.P., Sousa, R., Pereira, A.C., and Reis, M.S. (2017). Advanced predictive methods for wine age prediction: Part II - a comparison study of multiblock regression approaches. *Talanta,* 171, 121-142.

Chiang, L.H., Pell, R.J., and Seasholtz, M.B. (2003). Exploring process data with the use of robust outlier detection algorithms. *Journal of Process Control,* 13 (5), 437-449.

Harford, T. (2014). Big data: are we making a big mistake. *Significance,* December, 14-19.

Kenett, R.S., and Shmueli, G. (2014). On Information Quality. *Journal of the Royal Statistical Society A,* 177 (1), 3-38.

Kenett, R.S., and Shmueli, G. (2016). *Information Quality: The Potential of Data and Analytics to Generate Knowledge*. Wiley.

Ku, W., Storer, R.H., and Georgakis, C. (1995). Disturbance Detection and Isolation by Dynamic Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems,* 30, 179-196.

Rato, T.J., and Reis, M.S. (2013a). Advantage of Using Decorrelated Residuals in Dynamic Principal Component Analysis for Monitoring Large-Scale Systems. *Industrial & Engineering Chemistry Research,* 52 (38), 13685-13698.

Rato, T.J., and Reis, M.S. (2013b). Fault detection in the Tennessee Eastman process using dynamic principal components analysis with decorrelated residuals (DPCA-DR). *Chemometrics and Intelligent Laboratory Systems,* 125, 101-108.

Rato, T.J., and Reis, M.S. (2017). Multiresolution Soft Sensors (MR-SS): A New Class of Model Structures for Handling Multiresolution Data. *Industrial & Engineering Chemistry Research,* 56 (13), 3640-3654.

Reis, M.S., Bakshi, B.R., and Saraiva, P.M. (2009). Denoising and Signal to Noise Enhancement: Wavelet Transform and Fourier Transform. In S. Brown, R. Tauler & B. Walczak (Eds.), *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis* (Vol. 2, pp. 25-55). Elsevier, Oxford.

Reis, M.S., Braatz, R.D., and Chiang, L.H. (2016). Big Data - Challenges and Future Research Directions. *Chemical Engineering Progress,* Special Issue on Big Data (March), 46-50.

Reis, M.S., and Saraiva, P.M. (2006a). Generalized Multiresolution Decomposition Frameworks for the Analysis of Industrial Data with Uncertainty and Missing Values. *Industrial & Engineering Chemistry Research,* 45, 6330-6338.

Reis, M.S., and Saraiva, P.M. (2006b). Multiscale Statistical Process Control of Paper Surface Profiles. *Quality Technology and Quantitative Management,* 3 (3), 263-282.

Reis, M.S., and Saraiva, P.M. (2006c). Multiscale Statistical Process Control with Multiresolution Data. *AIChE Journal,* 52 (6), 2107-2119.

Reis, M.S., and Saraiva, P.M. (2012). Prediction of Profiles in the Process Industries. *Industrial & Engineering Chemistry Research,* 51, 4524-4266.

Walczak, B., and Massart, D.L. (1997). Noise Suppression and Signal Compression Using the Wavelet Packet Transform. *Chemometrics and Intelligent Laboratory Systems,* 36, 81-94.

Westerhuis, J.A., Kourti, T., and MacGregor, J.F. (1998). Analysis of Multiblock and Hierarchical PCA and PLS Models. *Journal of Chemometrics,* 12, 301-321.

Willsky, A.S. (2002). Multiresolution Markov Models for Signal and Image Processing. *Proceedings of the IEEE,* 90 (8), 1396-1458.