Linking pulp variations to TMP operation by better selection and treatment of process data

ROBERT P. HARRISON AND PAUL R. STUART

ABSTRACT: Multivariate analysis (MVA) is widely used for troubleshooting, process monitoring, and advanced control. It has become easily accessible through desktop software packages. However, this least-squares statistical technique remains highly susceptible to the adage "garbage-in/garbage-out," notably with regard to process disturbances and other outliers. Using a thermomechanical pulp newsprint mill in Eastern Canada as a case study, we compared various ways of selecting and pre-treating raw process data to maximize the realism and usefulness of the black-box pulp quality models. We eliminated start-up and shutdown data to model steady-state conditions. A major conclusion of this work was that the partial least squares models were significantly improved by pre-treating the data, with respect to both statistical significance and physical interpretability. We therefore recommend an overall approach for applying MVA to industrial operating data. This involves a systematic method for removing dubious periods of operation, such as low production and aberrant process behavior, and filtering of all dependent and independent variables. Because no single model was able to cover all process scenarios, it seems that some kind of adaptive controller would be required to automate the TMP refining process.

Application: This study presents a straightforward method for selecting and pretreating TMP operating data, to improve statistical tracking of pulp quality variations.

Multivariate analysis (MVA) is increasingly being used for improving operations, whether through troubleshooting, process monitoring, or advanced process control. This statistical tool is widely available to plant personnel via user-friendly desktop computer packages, but it is highly susceptible to "garbage-in/garbage-out." This paper uses historical data from a real mill to explore ways to counter this challenge.

MVA reduces the number of variables within a dataset, to make it more manageable and understandable. The original variables are boiled down to a smaller number of new variables, or 'principal components,' that typically capture much of the variance of the initial dataset [1]. Each principal component is simply a linear combination of the original variables. Often, the principal components will correspond to hidden or latent variables, inherent to the system, that are not measured directly but have a physical interpretation. Modern computing power has made it possible to apply MVA to millions of data points, and it has been used in a variety of industrial sectors [2], including pulp and paper [3-6].

MVA is a statistical technique, entirely data-driven. It functions like a black box, finding relationships between input and output data only, with no physical modeling of any kind. As a least-squares method it inevitably overemphasizes extreme values such as large process fluctuations, at the expense of normal operating conditions. When applying MVA to raw data from an industrial facility such as a pulp and paper mill, it is therefore critical to select and pretreat the data adequately.

CASE STUDY: TMP NEWSPRINT MILL IN EASTERN CANADA

The case study for this project is a thermomechanical pulp (TMP) newsprint mill in eastern Canada. The mill experiences short- and long-term variations in final paper quality. Over a typical month, hourly averages for tear strength, burst, and tensile stiffness index (TSI) may vary ±10%. Within a calendar year, daily averages for these same parameters may vary ±12%-14%. The goal of the study was to use MVA to understand the correlations and trends that are inherent to the mill operation to determine which upstream parameters are most likely linked to pulp quality variations.

High-quality pulp requires good quality fibers that are well separated. Proper refiner operation is critical to this aim, with maximum separation, but minimum cutting of fibers. Changes in incoming chip quality are beyond the control of the TMP mill operator, and are difficult to measure [7]. To a certain degree, these external factors could theoretically be counteracted using controllable internal operating parameters such as chip feed rate, specific energy, dilution water flow, and plate gaps.

Typically, TMP operators use freeness as the main indicator of pulp quality, adjusting the set points for three variables: transfer screw speed, plate position (or hydraulic pressure), and dilution water flow rate [8]. The specific energy required to achieve a given freeness is usually related to the amount of long fiber in the pulp [9]. Modern control systems [3, 10] therefore use freeness and fiber length as indicators, generating a window within which the refiners should operate.

At the case study mill, key pulp quality parameters, including freeness, average fiber length, and fines content, are measured using an automated on-line Metso PulpExpert EXP sampler/analyzer. The mill operates on 100% wood chips, using a combination of black spruce and balsam fir, but no substantial data were available on the incoming wood chip quality or exact species mix. (Several on-line woodchip monitors are under development in Canada, though none is yet commercially available.)

The mill is equipped with Sunds Defibrator RGP-70-CD conical refiners, with independent control of the gap between the plates in both the flat and conical sections. There is no advanced process control on these refiners; key operating decisions are made by the operators. The main steps are:

1. Steam pre-heating of chips at 180 kPa for 3 min
2. Primary refining at 350 kPa and 40%-50% consistency
3. Secondary refining at 350 kPa and 40%-50% consistency

The latency chest has an approximate residence time of 45 min. The upstream equipment has a residence time of only a few minutes, so in this regard the latency chest dominates this section of the mill. The refiner plates have a normal lifespan of about 2000 hours.

Pulp refining is a highly complex process using a biological feed material, so fundamental models to date have been semi-empirical [11]. However, no models have been published specifically for conical refiners, hence the necessity of the black-box approach used in this study. The MVA models presented here are based on previous work that focused on identifying key variables, and their combinations, to best model the fundamentals of this system [12].

RELATING PULP VARIATIONS TO TMP OPERATION

A variant of MVA called *partial least squares* (PLS) was used to model latency chest pulp quality. As shown in Table I, the pulp quality parameters which serve as the Y variables were freeness, average fiber length, and percent fines. We selected a time increment of 1 hour, corresponding roughly to the overall residence time of the system (around 50 min). The three dependent variables are measured every 60-90 min on average, and so again a 1-hour period is appropriate.

A major challenge in this study was the regular starting and stopping of the TMP lines, due to built-in overcapacity relative to the papermaking section. Direct use of the raw data would yield meaningless MVA results, because the algorithm would attribute most of the correlation to the start-stop phenomenon, and not to actual changes in the process during normal operation.

As is the case with most MVA studies of existing operations, results are dependent on the natural variability preexisting in the dataset. Designed experiments can be used to counter this problem [13]. However, this study was limited to historical operating data which, though imperfect, are an important (and inexpensive) source of insight available to process engineers.

Variable	Unit
X variables	
1° and 2° motor loads	MW
Production rate	
(proportional to feed screw rotational speed)	tons/day
1° and 2° specific refining energy	kWh/ton
1° and 2° flat plate gap	mm
1° and 2° conical plate gap	mm
1° and 2° dilution flows (multiple)	L/min
1° and 2° steam pressure drop across refiner (îP)	kPa
1° and 2° blowline consistency (calculated)	% solids
1° and 2° plate age	h
Y variables	
Canadian Standard Freeness	mL
Average fiber length (length-weighted)	mm
Fines content, defined as small enough	
to pass through 200-mesh screen	
(76 m)	%(mass)

Table I. Variables used to generate partial least squares (PLS) model of single thermomechanical pulp line at case study mill.

Criterion	Parameter Used1	Type Of Output
Goodness of Fit	Q^2 for overall model Predicted vs. observed for each Y	Quantitative Scatter plot
Complexity of Model	Number of components required for a given Q^2	Quantitative
Realism of Model	Relative prominence of X variables Interpretability of components with regard to process fundamentals	Ranking Qualitative

[1]Q^2 is a measure of goodness of fit analogous to R^2, but it is specific to predictive power; it is the percentage of overall measured variance that is attributable to the model's predicted values. It is derived by separating the dataset into several segments, some used to create the model, others for testing it. Unlike R^2, which increases when a model is made more complex with additional principal components, Q^2 tends to plateau and then diminish when there is over-fitting.

Table II. Criteria used for comparing MVA models.

The various MVA models generated in this study were evaluated using two different, but parallel, approaches: *mathematical*, primarily based on goodness of fit; and *physical*, based on the interpretability of the components that were found. Table II lists the criteria used in this project.

We subjected the raw data to the step-by-step approach outlined in Fig. 1. We evaluated each of the preprocessing steps to compare and contrast the results obtained. For the most part, data pretreatment was done in Microsoft Excel before downloading the data into the MVA software: SIMCA-P (version 10) from Umetrics AB.

Minimum production, as opposed to mean production, refers to the smallest value recorded on a second-by-second basis within that 1-hour period. A threshold of 200 tons/day was selected because the TMP operators never intentionally operate the refiners below this rate.

We used *principal component analysis* (PCA)—the other main variant of MVA—to identify potential outliers. It differs from PLS in that it treats all variables the same, rather than dividing them into Xs and Ys. In fact, PLS is a combination of one PCA model for the Xs and another for the Ys, each one tweaked to match each other (hence the name "*partial* least squares").

PCA "score plots" show how each observation fits within the model space relative to all the others. The "distance-to-model plots" show how far each data point had to be projected to be included in the model. To eliminate subjectivity in interpreting these plots, only points falling outside Hotelling's T^2 (95-percentile) were considered outliers. The Simca-P software shows this threshold as an ellipse on the score plot, and as a horizontal line on the distance-to-model plot [14]. To identify which variables caused the outlier in the first place, we used the MVA "contribution plots." This ability is one of the main advantages of MVA over other black-box methods.

Exponentially weighted moving average (EWMA) is the most widely used form of filtering in the chemical process industries. Trimming and "winsorizing" automated features of the Simca P software [14] serve to remove extreme values. (Winsorizing is statistical jargon for replacing extreme values with the value at the cut-off point.) Since these features are routinely employed by MVA users, they were included in this study for the sake of comparison.

The periods under study were based on previous work in which the significance of summer and winter periods were examined [12]. Identical periods (March and August) were used within two separate years (2003 and 2004). The mill used normal refining plates during these periods.

REMOVAL OF LOW-PRODUCTION PERIODS

The main cause of production disruptions at the refiners is built-in overcapacity relative to the paper machines. When less pulp is needed, one of the four refining lines is temporarily shut down, such that the mill is operating with only three refiner lines about 70% of the time. All four lines are subjected to these shutdowns. Automatic shutdowns, called "feedguard" events, are triggered by excessive motor load and represent another cause of production disruption. Also, the refiners are stopped every 200 hours to change the direction of disk rotation, to prevent uneven wear of the refiner plates.

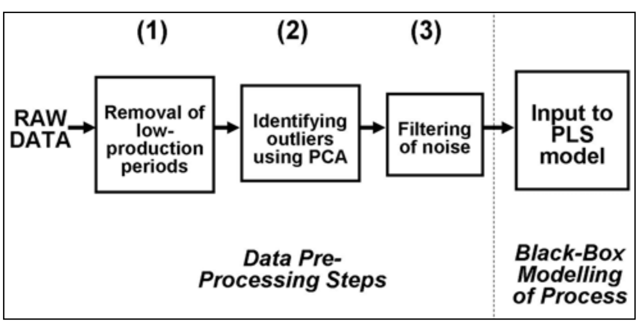

Figure 1. Overall data pretreatment strategy applied to TMP historical data.

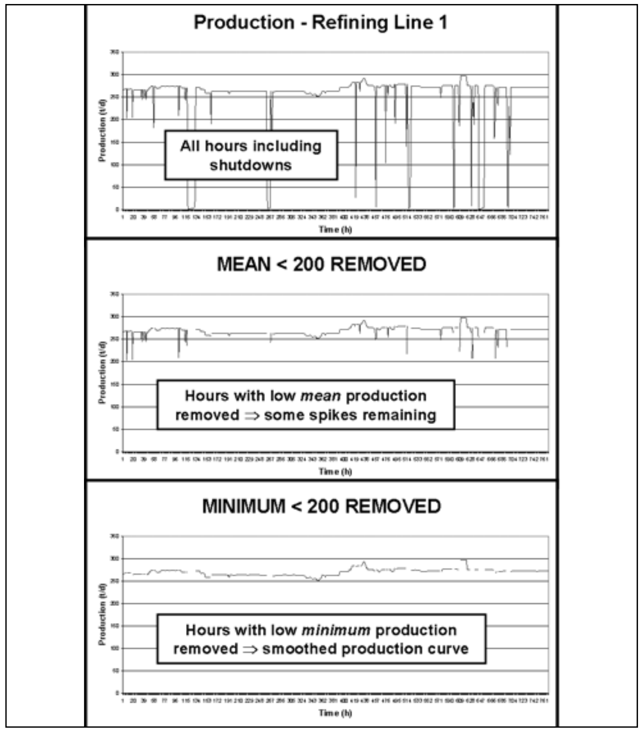

Figure 2. Hourly production rates for refining line 1 (expressed as metric tons per day) for the month of August 2003.

August 2003 was selected as the base case. Previous work at the mill had shown stronger MVA models for summer data, possibly due to poorer chip quality in winter. The two options for dealing with low production periods were compared. On the basis of the mean throughput, 58 hours were disqualified out of 768 for August 2003. Only removing periods where the mean production was below the threshold proved inadequate, however (Fig. 2). The mean in this case disguised short periods of low production. Using the more stringent minimum test, a further 32 hours were removed, which resulted in a much clearer picture of the overall periods suitable for further study.

To complete the comparison, PLS models were generated using different datasets (Table III). The Q^2 shown is a combined metric for all three dependent (Y) variables, namely latency chest freeness, fiber length, and fines content. In other words, these models were built for a group of dependent variables. The results obtained would be different if each dependent variable had been modeled separately.

Thus, the PLS results for August 2003 confirmed that the best option was to remove *a priori* the periods with minimum production below the threshold, and then take out the PCA outliers. However, the differences between some of the Q^2 values are slight. In the case of "mean < 200 removed" there was no appreciable difference when the outliers were removed. It was therefore necessary to investigate further by examining the actual nature of the data points that were removed in each case.

IDENTIFYING MAJOR OUTLIERS

For each of the four months under study, we created a PCA model using all variables and data points, including low-production periods, to test the effectiveness of MVA for detecting known outliers. The data points identified using Hotelling's T^2 ellipse (95-percentile) often corresponded to the periods of low production, but not always, and some low production periods were not detected at all.

Some of the outliers detected using the distance-to-model plot occurred during normal production periods, several hours after the last shutdown. Clearly, other variables were breaking with the model structure. Studying the contribution plots revealed that, for these points, plate gap and specific energy in the secondary refiner were higher than normal in some cases, and lower than normal in others. Several points showed higher or lower specific energy in the primary refiner. These results are difficult to interpret, because the residual low-production periods could be adversely affecting the model's overall validity.

Next, a PCA model was created that excluded periods of low *mean* production. The score plots revealed far fewer outliers, corresponding to periods where the minimum production was below the threshold. The distance-to-model plots also showed fewer outliers.

Finally, we created a PCA model excluding periods of low *minimum* production. This resulted in the model with the fewest outliers. For August 2003, fewer than 10 outliers were found on the score plot, with a further 32 found on the distance-to-model plot. This explains why there was no appreciable difference in when the outliers were removed.

Using the MVA contribution plots for these points, we could determine which variables had caused them to break with the overall correlational structure. The few dozen distance-to-model outliers were mainly due to changes in pulp parameters that were uncorrelated with the other variables, or to fluctuations in steam pressure drop across the refiner, possibly indicating temporary obstructions to counter-current steam flow between the refiner plates. Overall, this small number of outliers appeared to represent unusual operating conditions. Even though they had little effect on the Q^2, they were therefore removed before continuing to the next step of data pre-treatment.

Dataset	Overall Q^2	# of Comp.[2]	Comments
All hours	22%	5	Dominated by start-up and shut down
- minus outliers[1]	41%	5	Dominated by start-up and shutdown
Mean < 200 removed	39%	5	Start-up and shutdown still strongly evident
- minus outliers	39%	4	Start-up and shutdown still strongly evident
Min. < 200 removed	40%	4	Minimal evidence of start-up and shut down
- minus outliers	41%	4	Option retained

1Rows marked "- minus outliers" indicate that both types of PCA outliers were removed.
2The column identified as "# of Comp." refers to the number of significant components found, based on the best cumulative Q^2.

Table III. Results for partial least squares (PLS) models generated from refining line data, with different hours excluded.

Data Pretreatment Step	Number of Hours Removed
Low production periods, based on *minimum* throughput	90
PCA outliers, based on score plot	6
PCA outliers, based on distance-to-model plot	32
One-hour periods retained for partial least squares (PLS) model	641
TOTAL	768

Table IV. Hours retained and removed for August 2003.

Table IV summarizes the final tally of hours retained and hours removed for August 2003, out of the original 768 one-hour periods.

FILTERING NOISE

Figure 3 shows plots of the original primary and secondary motor loads for August 2003, overlaid by the filtered signals for two different values of alpha in EWMA. (Alpha is the weighting, between zero and one, given to the previous value in the sequence. The weighting of the current value is one-minus-alpha. An alpha of zero means there is no filtering.)

The lower value of 0.5 was the point at which smoothing started to become visibly apparent on the plots. The upper value of 0.8 was chosen because it yielded the smoothed curve that fit the original data the best, admittedly a somewhat subjective evaluation. At alpha values above 0.9, the curves became visibly overfitted. Note that with an extreme alpha (> 0.99), all signals begin to resemble a straight line.

At an alpha of 0.8 there is a noticeable shift of the raw data to the right, caused by the inertia of the moving average. Plotting the cross-correlation curves for filtered versus unfiltered data showed this shift to be about 2 hours. However, since this shift applies equally to all variables—both X and Y—the accuracy of the model should not be affected. Obviously this shift would have to be taken into account if filtered and unfiltered signals were used in combination.

Table V shows the results of the various PLS models for August 2003. The combined Q^2 value for the three dependent variables (freeness, fiber length, and fines) is shown are for a model with four components. In all cases, components numbered 5 and above contributed little or no incremental gain to the overall goodness of fit, and may therefore represent random noise in the data. The improvement in goodness of fit when EWMA is applied is striking, with a jump in Q^2 from 41% to 61% just by using an

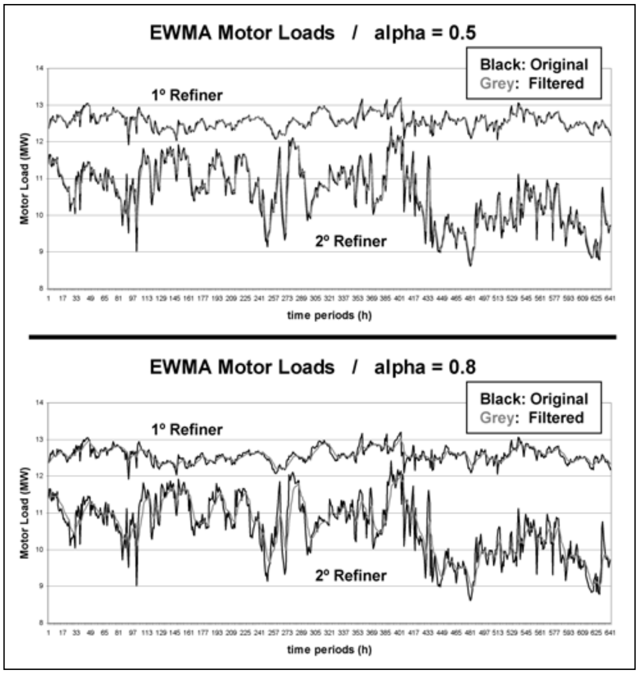

Figure 3. Filtered vs. original motor load signals for August 2003, showing smoothing effect of exponentially weighted moving average at different alpha values.

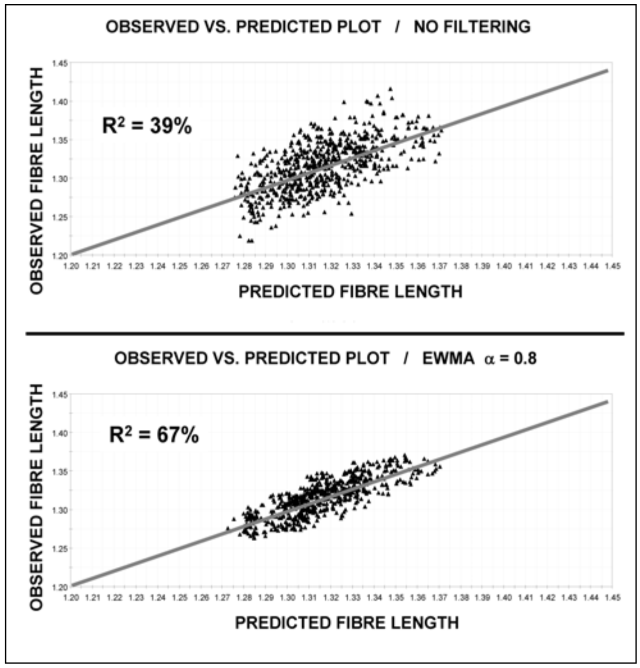

Figure 4. Observed vs. predicted plots for fiber length in the partial least squares (PLS) model, using different filtering methods, August 2003 (diagonal is 1:1 line). The R^2 value is that calculated by the PLS model.

alpha of 0.8.

The trimming/winsorizing step entailed the removal of the top and bottom 1% of values for each individual variable, and their replacement with the value at the cut-off point. This option had little effect on the results. This was not surprising. Intuitively this method does not seem suited to time series data, where the most extreme values might be due to process shifts and not aberrant measurements.

The improvement in accuracy with filtering is apparent in Fig. 4, which shows the observed vs. predicted plots generated for average pulp fiber length. The two plots show the data clouds with and without EWMA for August 2003.

Table VI indicates possible interpretations for the components for August 2003. These are based largely on which X and Y variables showed the largest PLS loadings, i.e., the weights assigned to each original variable by the PLS model. The larger the loading, the more that term contributed to the PLS model. Motor loads and dilution flow rates are often highly correlated with production rate. For PCA modeling, this could be a pitfall. This one dominant trend could eclipse other, more interesting correlations. However, when using a PLS model, it is acceptable to have highly correlated variables among the Xs, because a component in the X space can only contribute to the PLS model if it correlates with a trend among the Ys, regardless of the number of underlying variables.

Conditions at the primary refiner seem to mostly correlate with pulp freeness, whereas conditions at the secondary refiner more closely relate to fiber length and fines. This is consistent with discussions we have had with the TMP operators at the mill and with equipment suppliers. Similar components were found when the data were filtered, suggesting that the overall correlational structure

Filtering option	Overall Q^2	# of Comp.1
No filtering	40.63%	4
Trim / winsorize	40.56%	4
EWMA, α =0.5	50.13%	3
EWMA, α =0.8	60.75%	2

[1]The column titled "# of Comp." indicates the number of components required to achieve an overall cumulative Q^2 equivalent to that of the "no filtering" case.
EWMA = exponentially weighted moving average.

Table V. Comparison of partial least squares (PLS) results for different data filtering methods.

of the dataset was not compromised by using EWMA.

Note that freeness is logarithmic with almost any other variable. However, in a previous study [12], we replaced freeness with its logarithm and obtained exactly the same PLS models. The most likely explanation is that within our dataset the freeness only varied between 200 mL and 250 mL, which is too small a range for the nonlinearity to show itself. We therefore used the original freeness values.

Table VII compares results for the four different months studied. Compared to the August 2003 base case, March of that year showed a large number of start-ups and shutdowns, with over a third of the hours being identified for removal. The final PLS model was much weaker than for August, possibly related to greater variability in incoming chip quality (typically not well measured at Canadian newsprint mills). Again, the overall Q^2 was significantly improved by using EWMA. Furthermore, the prominent X variables for March were different for August, with the first component being dominated by low specific energy in the primary refiner, which the TMP operators often associate with lower quality chips.

Apart from the fact that the models were stronger, the 2004 findings were similar to those of 2003. However, once again, which X variables were prominent differed from one month to the other. The process clearly evolves over time, such that no two months have exactly the same coefficients. Using MVA for advanced control would therefore require some kind of adaptive controller, as reported by other authors [3].

RECOMMENDATIONS

The main focus of this paper was to examine, in a systematic way, various techniques for preselecting and pretreating TMP process data to increase the realism and usefulness of PLS black-box models for pulp quality. For the purpose of this study, pulp freeness, average fiber length, and fines content at the latency chest were combined into a single statistical model. The models were compared on the basis of pure statistics, notably goodness-of-fit, and on model interpretability. Significantly better PLS models were obtained when mill data were pretreated as follows:

1. Low-production periods were stringently removed so that no second-by-second data point during the entire hour fell below the threshold.
2. PCA outliers on both the score and distance-to-model plot were removed using Hotelling's T^2 (95-percentile).
3. Aggressive EWMA filtering was applied to all Xs and Ys.

While data pretreatment is clearly essential to successful application of MVA, these methods are generally compromises, with no one single "best" data pretreatment methodology. However, in our study the models using pretreated data were better, whether evaluated using statistical metrics or qualitative tests.

One possible explanation for the success of filtering is that MVA is normally blind to time series data, treating all data points the same regardless of how far apart they are in real time. Filtering with EWMA essentially overlays time-related information onto the dataset, introducing a dynamic element not present in the original unfiltered data. By providing the algorithm with more information about the system, filtering improves its ability to find correlations between the dependent and independent variables. EWMA did not appear to affect which X and Y variables were most prominent, suggesting that even with a high level of filtering, the model was still representative of the original data.

PLS outputs must be studied carefully to ensure that the results correspond to the known properties of the system. The PLS models from this study were consistent with expected results, such as lower specific energy being associated with less fiber development (higher freeness) and less fiber cutting (longer fiber length and lower fines). Another observed effect was plate age. However, caution must be exercised before ascribing cause-and-effect to what are purely statistical outputs, since both the Xs and the

#	Variables with Highest PLS Loadings[1]	Possible Interpretation of Component
1	X: 2° motor load (-) Y: freeness (+), fiber length (+), fines (-)	Impact of 2° refining energy on fiber length
2	X: 2° plate gaps (+), 2° steam P (-) Y: fiber length (+), fines (-)	Older plates requiring tighter gap in 2° refiner, yielding shorter fibers and more fines
3	X: 2° dilution (+), 1° plate gaps (+) 1° motor load (-) Y: freeness (+)	Lower primary refining energy, yield ing higher freeness
4	X: 2° steam P (+), 2° conical plate gap (-) Y: fiber length (-), fines (+)	Higher refining intensity in secondary refiner, cutting fibers, and generating fines

[1]The parentheses show whether the loading was positive or negative, which indicates either a positive or a negative correlation *vis-à-vis* the other variables. Caution must be exercised when assigning cause-and-effect relationships to purely statistical outputs.

Table VI. Interpretation of partial least squares (PLS) components for August 2003 mill data.

		Overall Q2		
Month	Hours Removed	No Filtering	EWMA ± = 0.8	Dominant X Variables
March 2003	273 out of 768 (36%)	23%	38%	1° energy, 1° dilution, 1° plate gaps
August 2003	128 out of 768 (17%)	41%	61%	1° and 2° energy, 2° plate gaps, 2° steam P
March 2004[1]	258 out of 768 (34%)	56%	67%	2° energy, 1° plate gaps, 1° and 2° consistency
August 2004	170 out of 768 (22%)	59%	68%	1° and 2° energy, 1° dilution, 1° and 2° consistency

[1]Maintenance on the PulpExpert unit during this month may have affected results.
EWMA = exponentially weighted moving average.

Table VII. Partial least squares (PLS) results for different months using identical variables and data pretreatment.

Ys might simply be responding together to an outside influence.

Different X variables were prominent from one month to the other. It would appear that the process evolves over time, such that no two months will yield exactly the same model, even if the list of key parameters remains constant. Unmeasured fluctuations in incoming chip quality is one plausible explanation, though without better data this remains conjecture. Using MVA for advanced control would therefore require an adaptive controller, in which the variables remain the same but the coefficients are continually updated.

Even our best models only explained 40% to 60% of the variance in the Y parameters, meaning that about half of the variance corresponds to unmeasured variables. The various on-line woodchip monitors under development should help to address this data gap. **TJ**

ACKNOWLEDGEMENTS

This work was completed with support from the Natural Sciences and Engineering Research Council of Canada (NSERC) Environmental Design Engineering Chair at École Polytechnique. We would also like to acknowledge Alain A. Roche of PAPRICAN and Martin Fairbank of Abitibi-Consolidated Inc. for their invaluable advice and inspiration.

LITERATURE CITED

1. Johnson, R.A. and Wichern, D.W., *Applied Multivariate Statistical Analysis*, Prentice Hall, New Jersey, 1992.
2. Mason, R.L. and Young, J.C., *Quality Progress*, 37(4): 89(2004).
3. Strand, W.C., Fralic, G., Moreira, A., et. al., *Proceedings of the 2001 International Mechanical Pulping Conference*, TAPPI Press, Baltimore, MD, Vol. 2, pp. 253-262.
4. Lupien, B., Lauzon, E., and Desrochers, C., *Pulp Paper Can.* 102(5): 19(2001).
5. Saltin, J.F. and Strand, B.C., *Pulp Paper Can.* 96(7): 48(1995).
6. Nobleza, G.C., 1997 *Proceedings of the Technical Section CPPA Annual Meeting*, PAPTAC Publications, Montreal, Canada, pp. B31-B36.
7. Ding, F., Benaouida, M., Bédard, P., et al., *Pulp Paper Can.* 106(2): T25(2005).
8. Cluett, W.R., Guan, J., and Duever, T.A., *Pulp Paper Can.* 96(5): 31(1995).
9. Miles, K.B. and Omholt, I., *Proceedings of the 2003 International Mechanical Pulping Conference*, TAPPI Press, Baltimore, MD, pp. 179-186.
10. Sidhu, M.S., Van Fleet, R., Dion, M.R., et al., *Proceedings from Control Systems 2004 Conference*, PAPTAC Publications, Montreal, Canada, pp. 107-112.
11. McDonald, D., Miles, K., and Amiri, R., *Pulp Paper Canada* 105(8): 27(2004).
12. Harrison, R.P., Leroux, R., and Stuart, P.R., *91st Annual PAPTAC Meeting Preprints*, PAPTAC Publications, Montreal, Canada, 2005, pp. D563-D568.
13. Elsinga, M., *Proceedings from IEEE Pulp & Paper Industry Technical Conference*, 2002, IEEE Publishing, Piscataway, NJ, pp. 10-15.
14. Umetrics, A.B., User Guide: Simca-P and Simca-P+ 10, Umea, Sweden, 2002.

Received: March 23, 2006
Accepted: March 30, 2006

This paper is also published on **www.tappi.org** and summarized in the June, 2006 Solutions! for *People, Processes and Paper* magazine (Vol. 89 No. 6).

INSIGHTS FROM THE AUTHORS

Many mills are faced with the problem of data overload. With this work, we're trying to help them extract useful information from these large databases to address product quality issues. Until now, we have focused on selecting which variables to use and on interpreting the multivariate analysis (MVA) results to have physical meaning. This was the first time we systematically applied different data pretreatment techniques to see which ones gave the best model of thermomechanical pulp (TMP) quality.

The most difficult aspect of this work is problems with the data. No matter how much we understand the process and focus on the right variables in the right combinations, there is no guarantee that this type of black box modeling will actually yield any useful results. We addressed this challenge by studying the data beforehand, with simple time plots for each variable, to make sure we were weren't just modeling instrument drift or calibration or other nonprocess trends.

The next step is to model all four TMP lines together, for overall pulp quality. What I discovered personally from this research is that sometimes it is only a small number of datapoints that "spoil the barrel." By eliminating these, we can greatly improve the accuracy and usefulness of the MVA models. The most surprising thing is that when things go well, you really can model a portion of the variability in pulp quality by using the TMP operating parameters.

Harrison

Stuart

All mills have product quality fluctuations they would like to be able to explain better. They have large amounts of operating data, but it is difficult to sift out any useful trends from the rest.

Stuart is professor, Chemical Engineering Department, Ecole Polytechnique-Montreal and chairholder-NSERC Design Engineering Chair, Process Integration in the Pulp and Paper Industry; email paul.stuart@polymtl.ca; Harrison, is doctoral student, NSERC Environmental Design Engineering Chair in Process Integration, Department of Chemical Engineering, École Polytechnique, Montreal, Quebec, Canada; email robert.harrison@polymtl.ca.