DATA PRE-PROCESSING TECHNIQUES FOR MULTIVARIATE ANALYSIS TO TREAT INDUSTRIAL OPERATING DATA FOR RETROFIT DESIGN

Robert P. Harrison, Paul R. Stuart

NSERC Environmental Design Engineering Chair in Process Integration, Department of Chemical Engineering, École Polytechnique, Montréal (QC)
Contact: paul.stuart@polymtl.ca

ABSTRACT

Multivariate Analysis (MVA), a statistical design tool for dealing with very large datasets, was applied to historical data from a Thermo-Mechanical Pulp (TMP) newsprint mill in Eastern Canada. Partial Least Squares (PLS) type MVA models were created to identify significant correlations between operating parameters in the woodchip refining section and variations in pulp quality. Understanding these relationships is of crucial importance to any eventual retrofit design for this process. This paper focusses on pre-selecting and pre-treating the raw process data, including infrequently measured variables, to maximize the realism and usefulness of the MVA black-box models. Key methods explored were ways of selecting low-production periods for removal, techniques for identifying and eliminating major outliers using MVA outputs, and noise filtering. A major conclusion of this work was that the PLS models were significantly improved by pre-treating the data. This paper recommends an overall design approach for applying MVA to industrial operating data, involving stringent removal of dubious periods of operation such as aberrant process behaviour, and an aggressive Exponentially Weighted Moving Average (EWMA) filtering of all dependent and independent variables.

INTRODUCTION

Multivariate Analysis (MVA) is a statistical design tool used for dealing with very large datasets. A small number of principal components is generated that capture much of the variance of the initial dataset, but at a much lower dimensionality. Each new component is simply a linear combination of the original variables, and essentially amounts to a rotated axis that better fits the data cloud. Often, these new principal components will correspond to hidden or “latent” variables inherent to the system that are not measured directly.

While this concept is quite straightforward, it is only the advent of modern computing power that has made it possible to apply it to millions of data points. Engineers and scientists can now draw useful conclusions for systems with dozens or even hundreds of variables, without sacrificing the information contained in the correlational structure. MVA has been used to understand and ultimately reduce process variation in a variety of industrial sectors [1] including pulp and paper [2, 3, 4, 5]. Every day, industrial plants generate vast numbers of real operating data, which can be studied using MVA to gain insights for better equipment or process design, yielding knowledge that was not available in the past and thus potentially creating opportunities for better retrofit design.

When applying MVA to raw operating data from an industrial facility such as a pulp and paper mill, it is critical to prepare the data adequately right from the start. MVA is a black-box statistical technique, entirely data-driven, and therefore highly susceptible to the problem of “garbage-in/garbage-out”. This paper explores these data pre-processing challenges, by incorporating work that was done previously along with new developments.

CASE STUDY: NEWSPRINT MILL

The case study for this project is a Thermo-Mechanical Pulp (TMP) newsprint mill in Eastern Canada, which experiences short- and long-term variations in final paper quality. The goal is to use MVA to understand the correlations and trends that are inherent to the mill operation, in order to determine which upstream parameters are most likely linked to these quality variations. The project is focussed on two different facets of the problem: first, a statistical component including pre-processing of data to enhance the results, and second, a chemical-physical interpretation of the results based on fundamental phenomena. This paper deals mainly with the former aspect, with the specific design objective of determining how best to use real operating data for possible retrofit improvements to the refining section, which is the heart of this type of newsprint mill. Here wood chips are converted into pulp, an aqueous suspension of cellulose fibres that is later screened, treated, formed and dried into the familiar large rolls of newsprint paper.

The TMP refining section is illustrated in Figure 1. The primary and secondary refiners are comprised of large rotating disks for disintegrating the wood chips into a fibrous pulp.

Fig. 1: Refining section of a Thermo-Mechanical Pulp newsprint mill. Red star indicates main pulp sampling location. Blue arrows indicate water dilution points.

This pulp is an intermediate product of crucial importance to final paper quality. Key pulp quality parameters include its permeability to free water flow (known as “freeness”), the average length of the cellulose fibres, and fines content, all measured using an automated in-line Metso PulpExpert EXP sampler/analyser. No substantial data were available on the incoming wood chip quality. The mill is equipped with Sunds Defibrator CD-70 conical refiners, in which the rotating disk has both vertical and conical sections, each with independent control of the distance between the plates. There are multiple dilution points, as indicated by the

blue arrows. There is no advanced process control on these refiners; all key operating decisions are made by the operators.

The latency chest has a 45-minute residence time to allow the wood fibres to disentangle themselves before proceeding to the papermaking section downstream. The rest of the flowsheet only has a residence time of a few minutes, so in this regard the latency chest dominates this section of the mill. Note that the refiner plates have a normal lifespan of about 2000 hours, during which time the bars on the plate surface are gradually worn down, affecting the process dynamics.

Pulp refining is a highly complex process using a biological feed material, so all fundamental models to date have been semi-empirical [6]. No equations have been published for conical refiners, however, hence the black-box approach taken in this study. The MVA models presented here are based on previous work that focussed on which variables to use, and in what combinations, to best model the fundamentals of this system [7].

CHALLENGES WHEN USING MVA TECHNIQUE

The MVA user is confronted with a series of challenges inherent to this statistical technique, and must often make decisions right from the start that may affect the final results. These challenges fall into three categories:

- Those requiring process knowledge or chemical engineering insight in order to be addressed
- Those requiring statistical, mathematical, or data-oriented solutions
- Those that are unavoidable, but must nevertheless be understood to ensure proper interpretation of the final results

The most obvious process-oriented decision is selecting which variables to use among the hundreds available in a modern industrial plant. A statistical analysis is only as good as the original data, and MVA tends to be blindly slanted towards those variables that are measured better or more frequently [3]. The key X (independent) and Y (dependent) variables used in this study are listed in Table I.

TABLE I: VARIABLES USED IN PLS MODEL

VARIABLE	UNIT
X-Variables	
1° & 2° motor loads	MW
Production rate (proportional to feed screw rotational speed)	t/d
1° & 2° specific refining energy	kWh/t
1° & 2° vertical plate gap	mm
1° & 2° conical plate gap	mm
1° & 2° dilution flows (multiple)	L/min
1° & 2° steam pressure drop across refiner (ΔP)	kPa
1° & 2° blowline consistency	% solids
1° & 2° plate age	h
Y-Variables	
Permeability to liquid water flow, using <i>Canadian Standard Freeness</i> test	mL
Average (length-weighted) fibre length	mm
Fines content, defined as small enough to pass through 200 mesh screen (76 μ m)	% (mass)

A major challenge in this study was the constant starting and stopping of the TMP lines, due to over-capacity relative to the papermaking section. Furthermore, the key operating variables at the refiners are measured every second (the data historian's lowest time increment), whereas pulp quality is only measured every few hours, creating the challenging situation illustrated in Figure 2. Direct use of the raw data would yield meaningless MVA results, since the algorithm would blindly attribute most of the correlation to the start/stop phenomenon and not to actual changes in the process. Once this issue has been addressed, however, it should be possible to use a variant of MVA called Partial Least Squares (PLS) to model both the frequent (X) and infrequent (Y) variables [8]. PLS models serve to maximize the covariance between a set of X variables, in this case refiner operating data, and a set of Y variables, namely the latency chest pulp quality. Note that these are mathematical, not physical, distinctions: the decision as to whether a given variable is included in the X set or in the Y set is entirely up to the PLS user.

Fig. 2: Contrasting nature of X and Y variables

Another decision the user must make is whether to use means, medians, maxima/minima, discrete values, interpolated values, moving averages, standard deviations, or other forms of the original data. In this case, we used the mean as the logical starting point, although we also explored the use of a moving average, and the minimum for production rate.

Determining which timescales are appropriate is also of key importance, and unfortunately this may often be based on the availability of data rather than on any process consideration. For this study, we selected a time increment of one hour, corresponding roughly to the overall residence time of the system (around 50 minutes). The pulp quality parameters which serve as the Y variables for the model are measured every 60-90 minutes on average, and so again a one-hour period is appropriate.

The overall question of which timescales to use will be further explored in an upcoming paper, in which various parts of the mill, namely multiple refining lines plus the papermaking section, will be combined to model final paper quality.

Among the mathematical challenges of using MVA are the fact that it is a least-squares technique, and thus highly sensitive to outliers, whether caused by sensor malfunctions, start-up and shutdown of individual pieces of equipment, time lags, or other problems. MVA is likewise highly sensitive to

instrument drift, since this can appear as a long-term trend to which the algorithm blindly ascribes statistical significance. Note that variables are usually normalised before use in the MVA model, so that even the slightest trend can take on major significance.

Noisy data present a different problem, because MVA is not a time-series technique *per se*. Each one-hour period is treated as a separate observation or trial, and thus the presence of spikes will result in a series of trials which bear little resemblance to each other. Some kind of smoothing is therefore required. Merely using a one-hour average is already a form of filtering, but for this study, we also included a moving average for the sake of comparison.

The components that make up MVA models are simply linear combinations of the original variables, nothing more. Fortunately, though, for a well controlled system the range of the process is relatively limited, and it is possible in many cases to model important phenomena as linear. However, in cases of extreme non-linearity, new variables can be created based on one or more of the original variables. For this study, one such term was always included in the model, namely specific energy, which is simply motor load divided by throughput. Another non-linear term to represent the unmeasured parameter of refining intensity was also added in some cases, as described below.

Compression is used to save data storage space, by only logging those values that differ significantly from the previously stored value. It has been shown that data compression can impact mean and standard deviation calculations [9]. Ideally, the data compression would be removed to gauge its real impact on this kind of MVA application, but as these were historical data, it was not possible in this case. Studying stored data for the key variables such as motor load revealed that the time between recorded values was as high as 30 to 40 seconds in some instances, amounting to a loss in resolution in the signal (the exception test limit was $\pm 1\%$ and the compression test limit $\pm 2\%$). However, for the most part, one-hour averages were used, so this should not have a major impact. Plotting each variable also served to ensure that there were no regular staggered patterns indicative of excessive compression. Note that no compression had been applied to the Y (pulp quality) data.

When performing MVA, time lags must be taken into account beforehand. For this study, in order to compare the X and Y variables it was necessary to insert a 45-minute lag to account for the latency chest. As a check, the liquid level in the latency was verified over the time periods studied, and it remained above 90% capacity at all times. Note that this is not a plug flow situation, since the latency chest is well mixed.

Another mathematical question is how to best represent process variability. Obviously, the recorded values for each variable do change with time, providing some variability information to the model, but again MVA is not a time-series technique, so the link between consecutive time periods is totally lost. An upcoming paper will explore ways of capturing the variability as an explicit input to the model.

An example of an unavoidable problem is that MVA is blind to variables that do not vary significantly, regardless of their actual physical importance. For instance, if the temperature in a chemical reactor were kept perfectly constant,

it could not correlate with any other variable and would therefore appear to be of no importance, even though it is fundamental. As is the case with most MVA studies of existing operations, results are dependent on the natural variability pre-existing in the dataset. Designed experiments can be used to counter this problem [10]. This study was entirely limited to historical operating data, which are nonetheless an important source of process insight available to the modern-day design engineer.

Care must be used when interpreting and assigning cause-and-effect relationships to MVA results [11]. This is especially true of variables affected by control loops, where the correlations found are often the exact opposite of reality, due to the action of the controller. MVA blindly finds correlations, nothing more, and the physical reality of the system plays no part in generating the statistical outputs. As such, MVA models created under closed-loop conditions cannot therefore predict open-loop behaviour, or vice versa [12, 13].

Many authors have emphasised that an MVA model only works within the domain where it was calibrated, i.e., the process must continue to behave in a similar fashion as the original dataset [12, 13]. Past process data may therefore not help in predicting future behaviour, a major pitfall. This is of course true of any black-box model in which no scientific principles of any kind are used to guide the algorithm.

Finally, MVA results can be difficult to interpret, since it is not always obvious how to assign meaning to the components that are derived mathematically. In all cases, interpretations of the MVA results must be based on an understanding of the process fundamentals, since the actual outputs from the MVA software are purely statistical [4]. The use of MVA for studying industrial process data is thus a task for design engineers, and not pure statisticians.

CRITERIA FOR EVALUATING MVA MODELS

The various MVA models that were generated 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. The criteria used in this project are listed in Table II.

TABLE II: CRITERIA FOR EVALUATING MVA MODELS		
CRITERION	PARAMETER USED	TYPE OF OUTPUT
GOODNESS OF FIT	Q^2 for overall model	Quantitative
	Q^2 for each Y	Quantitative
	Predicted vs. Observed for each Y	Scatter plot
COMPLEXITY OF MODEL	Number of components required for a given Q^2	Quantitative
REALISM OF MODEL	Relative prominence of X variables	Ranking
	Interpretability of components with regard to process fundamentals	Qualitative

Note that 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 components, and others for testing the model. Unlike R^2 , which always increases when a model is made more complex, Q^2 tends to plateau and then diminish when there is overfitting.

METHODOLOGY

The study followed the step-by-step approach outlined in Figure 3. The raw data were subjected to each of these pre-processing steps, in order to compare and contrast the results obtained.

Fig. 3: Overall data pre-treatment strategy

The data pre-treatment options used in the various trials are listed in Table III. For the most part, these were simply done manually in EXCEL prior to downloading the data into the MVA software, SIMCA-P (version 10) from Umetrics AB in Sweden.

TABLE III: DATA PRE-PROCESSING OPTIONS USED IN MVA TRIALS	
PRE-PROCESSING STEP	OPTION
REMOVAL OF LOW PRODUCTION PERIODS	1) Remove all periods where <i>mean</i> production < 200 t/d
	2) Remove all periods where <i>minimum</i> production < 200 t/d
	3) Use all time periods
IDENTIFYING MAJOR OUTLIERS	1) Identify outliers using PCA score plots only
	2) Identify outliers using <i>both</i> PCA score plots <i>and</i> distance-to-model plots
	3) Do not remove outliers
FILTERING NOISE	1) Smooth motor load data using EWMA
	2) Remove peaks by trimming and winsoring
	3) No filtering
NON-LINEAR TERM	1) Include non-linear energy term
	2) Do not include non-linear energy term

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

Principal Component Analysis (PCA), the other main variant of MVA, differs from PLS in that it treats all variables the same, rather than dividing them into X's and Y's. In other words, in PCA all variables are considered X's. In fact, PLS is just a combination of one PCA model for the X's and another for the Y's, adjusted to match each other (hence the name *Partial Least Squares*).

PCA score plots, showing how each observation fits within the model space relative to all the others, can be used to identify potential outliers. The other plot used in this study, distance-to-model, shows how close each data point is to the multi-dimensional "hyper-plane" defined by the components, i.e., how far it had to be projected to be included in the model. The distance-to-model plot can indicate additional, more subtle outliers. To eliminate subjectivity in interpreting these plots, only points falling outside Hotelling's 95-percentile were considered outliers. In Simca-P, this threshold appears as an ellipse on the score plot, and as a horizontal line on the distance-to-model plot [14]. Those variables that caused the outliers in the first place, by breaking with the established correlational structure, can be easily identified using MVA contribution plots, one of the main advantages of this technique over other black-box methods.

The reasoning for using exponentially weighted moving average (EWMA) is that it is the most widely used form of filtering in the chemical process industries [15]. Trimming and winsoring, which are automated features of the Simca-P software [14], serve to remove extreme values, for instance the top and bottom 1% ("winsoring" 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 "non-linear energy term" referred to in Table III is the multiplication of two existing process values: the solids content (or "consistency") at the refiner outlet, as calculated by simple mass/energy balance; and the steam pressure delta between the refiner inlet and outlet. Together, these are indicators of low "refining intensity", which is the specific energy per impact experienced by the wood fibres within the refiner, which cannot be measured directly. This non-linear term was used in a previous study at the same plant [7] and was included here to determine whether data pre-processing would change its effect on the model.

The periods under study were based on previous work in which the significance of summer and winter periods were examined [7]. Identical periods within two separate years were included, as shown in Table IV. Normal refiner plates, as opposed to experimental plates, were used during all of these periods.

TABLE IV: TIME PERIODS STUDIED	
DATE	INCREMENT
March 1-31, 2003	1 hour
August 1-31, 2003	1 hour
March 1-31, 2004	1 hour
August 1-31, 2004	1 hour

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; about 70% of the time the mill is operating with only three refiner lines. Feedguard events, automatic shutdowns triggered by excessive motor load, represent another important cause of production disruption. Also, the refiners are stopped every 200 h to change the direction of disk rotation, to prevent uneven wear of the plates. All four lines are operated exactly the same way.

August 2003 was selected as the base case, since previous work had shown stronger MVA models for summer data, possibly due to poorer chip quality in winter. Both options for dealing with low-production periods were compared. Based on the *mean* throughput, 58 hours were disqualified out of 768. Merely removing periods where the mean production was below the threshold proved inadequate, however, as shown in Figure 4 for August 2003. Clearly the mean in this case disguised short periods of low production. Using the more stringent *minimum* test, only a further 32 hours required removal, but this gave a much clearer picture of overall periods suitable for further study.

Fig. 4: Production plots for the month of August 2003

To complete the comparison, PLS models were generated using the different datasets as shown in Table V.

Rows marked “- outliers” give the results when both types of PCA outliers were removed. The column identified as “# OF COMP.” refers to the number of significant components found, based on the best cumulative Q^2 .

TABLE V: PLS RESULTS WITH DIFFERENT EXCLUDED HOURS			
DATASET	OVERALL Q^2	# OF COMP.	COMMENTS
All hours	22%	5	Dominated by start-up & shutdown
- outliers	41%	5	Dominated by start-up & shutdown
Mean < 200 removed	39%	5	Start-up & shutdown still strongly evident
- outliers	39%	4	Start-up & shutdown still strongly evident
Min. < 200 removed	40%	4	Minimal evidence of start-up & shutdown
- outliers	41%	4	Option retained

Thus, the PLS results for August 2003 confirmed that the best option was to remove the periods with minimum production below the threshold, as well as all the PCA outliers. Similar results were found for the other three months under study.

IDENTIFYING MAJOR OUTLIERS

For the four months under study, a PCA model was created 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 95-percentile often corresponded to the periods of low production, but not always. Some low-production periods were not detected at all, making this an unreliable method for determining periods of normal operation.

Furthermore, 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 from the correlation 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, indicating variability in the process. Several points showed higher or lower specific energy in the primary refiner, often a sign of density fluctuations in the incoming chips and thus another source of process variability. These interpretations are suspect, however, since the 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, indicating that the models were more self-consistent.

Finally, a PCA model was created excluding periods of low *minimum* production. This resulted in the most self-consistent PCA model, with the fewest outliers. For August 2003, fewer than 10 outliers were found in the score plots, all during a 12-h period when fibre length was unusually low, and

finest content was unusually high, possibly corresponding to a temporary drop in incoming chip quality. The few dozen distance-to-model outliers were mainly due to pulp parameters that were also possibly related to incoming chip quality, or to fluctuations in steam pressure drop across the refiner, possibly related to obstructed counter-current steam flow between the refiner plates. Overall, this small number of outliers appeared to represent unusual operating conditions, and were therefore removed before continuing to the next step of data pre-treatment.

The final tally of hours retained and hours removed for August 2003, out of the original 768 one-hour periods, is given in Table VI.

TABLE VI: HOURS RETAINED AND REMOVED FOR AUGUST 2003	
DATA PRE-TREATMENT STEP	NUMBER OF HOURS REMOVED
Low production periods, based on <i>minimum</i> throughput	90
PCA outliers, based on Score plot	6
PCA outliers, based on distance-to-model plot	32
One-hour periods retained for PLS model	641
TOTAL	768

FILTERING NOISE

Figure 5 shows plots of the original primary and secondary motor loads (blue and yellow) for August 2003, overlaid by filtered signals (red and green) for two different values of alpha in EWMA.

Fig. 5: Filtered vs. original motor load signals

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 seemed to fit the original data the best, admittedly a somewhat subjective evaluation. At alpha values above 0.9, the curves became visibly overfiltered. Note that with an extreme alpha (> 0.99), any signal will eventually resemble a straight line.

At an alpha of 0.8 there is a noticeable shift 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 to all variables, both X and Y, the accuracy of the model should not be affected. Obviously shift would have to be taken into account if filtered and unfiltered signals were used in combination. Note that the filtering also smoothed out the transitions between individual high-production periods, an unavoidable outcome which may not be optimal.

The results of the various PLS models for August 2003 are shown in Table VII. In all cases, the Q^2 values shown are for a model with 4 components. Higher components contributed little or no incremental gain to the overall goodness of fit, and hence probably represented random noise in the data. The column entitled “# of Comp.” indicates the number of components required to achieve an overall cumulative Q^2 equivalent to that of the “No filtering” case. The rows identified as “non-lin.” give the results when blowline consistency and steam pressure delta are replaced by the non-linear term.

TABLE VII: COMPARISON OF PLS RESULTS FOR DIFFERENT FILTERING METHODS					
Filtering option	Overall Q^2	Freeness Q^2	Length Q^2	Fines Q^2	# of Comp.
No filtering	40.63%	34.04%	39.05%	48.71%	4
non-lin.	40.83%	34.08%	39.74%	48.62%	4
Trim / winsor	40.56%	33.99%	38.72%	48.92%	4
non-lin.	40.92%	33.94%	39.74%	49.05%	4
EWMA, $\alpha = 0.5$	50.13%	40.42%	52.49%	57.44%	3
non-lin.	49.51%	39.09%	53.20%	56.20%	3
EWMA, $\alpha = 0.8$	60.75%	46.97%	66.80%	68.47%	2
non-lin.	59.02%	43.59%	67.93%	65.54%	3

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 alpha of 0.8.

The trimming/winsoring option had little impact on the results. This was not surprising, since intuitively this method does not seem at all suited to time series data, where the most extreme values might be due to process shifts and not aberrant measurements.

Replacement of two key X variables by a non-linear term (their product) resulted in a marginal improvement in the model using unfiltered data. Basically, a slightly higher overall cumulative Q^2 was achieved using fewer variables. However this benefit disappeared when filtering was applied.

This would suggest that data quality, and not MVA model structure, is paramount in generating a strong model.

Figure 6 shows the Observed vs. Predicted plots generated for average fibre length, with and without EWMA for August 2003. The improvement in accuracy when filtering is applied is clearly obvious from these two plots, which were generated automatically by the Simca-P software.

Fig. 6: Observed vs. Predicted plots for fibre length, using different filtering methods (red line is 1:1 diagonal)

Table VIII indicates possible interpretations for the components for August 2003, based largely on which X and Y variables showed the largest PLS “loadings”.

TABLE VIII: POSSIBLE INTERPRETATION OF PLS COMPONENTS		
#	VARIABLES WITH HIGHEST PLS LOADINGS	POSSIBLE INTERPRETATION OF COMPONENT
1	X: 2° motor load (-) Y: freeness (+), fibre length (+), fines (-)	Lower 2° refining energy, yielding longer fibres and fewer fines
2	X: 2° plate gaps (+) 2° steam ΔP (-) Y: fibre length (+) fines (-)	Less aggressive conditions in 2° refiner, yielding longer fibres and fewer fines
3	X: 2° dilution (+) 1° plate gaps (+) 1° motor load (-) Y: freeness (+)	Low density chips in the incoming feed, affecting primary refiner operation and yielding higher freeness
4	X: 2° steam ΔP (+) 2° conical plate gap (-) Y: fibre length (-) fines (+)	Overly aggressive conditions in conical section of secondary refiner cutting fibres and generating fines

Loadings are simply the weight assigned to each original variable by the PLS model. The larger the loading, the more that term contributed to the PLS model. 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.

It would appear that conditions at the primary refiner are mostly correlated with pulp freeness, whereas conditions at the secondary refiner are more closely related to fibre length and fines. Pure cause-and-effect cannot be attributed to statistical results without designed experiments, of course, but these trends are consistent with TMP process fundamentals. Similar components were found when the data were filtered, suggesting that the overall correlational structure of the dataset was not compromised by using EWMA.

Table IX gives a comparison of the results for the four different months that were studied. Compared to the August 2003 base case, March of that year showed a large number of start-up 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. Again, the overall Q^2 was significantly improved by using EWMA. Furthermore, the structure of the model for March was different to August, with the first component being dominated by low specific energy in the primary refiner, which is often associated with lower quality chips.

TABLE IX: PLS RESULTS FOR DIFFERENT MONTHS			
MONTH	HOURS REMOVED	OVERALL Q^2	
		NO FILTERING	EWMA $\alpha = 0.8$
March 2003	273 out of 768 (36%)	23%	38%
August 2003	128 out of 768 (17%)	41%	61%
March 2004	258 out of 768 (34%)	56%	67%
August 2004	170 out of 768 (22%)	59%	68%

Apart from the fact that the models were stronger, the 2004 findings were very similar to those of 2003. Some differences were found in the structure of the individual components, but the overall trends were similar.

DISCUSSION

Clearly, much better PLS models were obtained when the data are pre-treated. 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 original data.

The best results were achieved when:

- 1) the most stringent test was applied to throughput, namely that not a single 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 95-percentile

- 3) Aggressive EWMA filtering was applied to all X's and Y's

The models using pre-treated data were better in all respects, whether evaluated using statistical metrics or qualitative tests. One possible reason is that MVA is normally blind to time series data, treating all time increments as equivalent 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.

It is important to note that PLS outputs must be studied carefully, to ensure that the results correspond to the known properties of the system at hand. The PLS models from this study were consistent with expected results, such as lower specific energy being associated with less fibre development (higher freeness) and less fibre cutting (longer fibre length and lower fines). However, caution must be exercised before ascribing cause-and-effect to what are purely statistical outputs, since both the X's and the Y's might simply be responding together to an outside influence (such as incoming chip quality).

CONCLUSIONS

Multivariate Analysis (MVA), a statistical design tool for dealing with very large datasets, was applied to historical data from a Thermo-Mechanical (TMP) newsprint mill in Eastern Canada, in order to determine if it is possible to represent the process more accurately by pre-treating the data.

A variant of MVA called Partial Least Squares (PLS) was used to model correlations between operating parameters in the woodchip refining section and the observed variations in pulp quality, which are of crucial importance to any eventual retrofit design for this process.

Several methods were explored for pre-treating the data. Low-production periods were identified using two different selection methods, one based on mean throughput and the other on minimum throughput. Two techniques were explored for identifying and eliminating major outliers using MVA outputs, one based on score plots, the other on distance-to-model plots. Finally, a basic trim-and-winsor approach for filtering extreme data points was compared to Exponentially Weighted Moving Average (EWMA) filtering.

A major conclusion of this work was that the PLS models were vastly improved by pre-treating the data. The paper recommends an overall design approach for applying MVA to industrial operating data, involving stringent removal of dubious periods of operation such as low production and aberrant process behaviour, and an aggressive EWMA filtering of all dependent and independent variables prior to creation of the PLS model.

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.

REFERENCES

1. MASON, R.L., YOUNG, J.C. "Multivariate Thinking", Quality Progress, April 2004, 89-91 (2004).
2. STRAND, W.C., FRALIC, G., MOREIRA, A., MOSSAFFARI, S., FLYNN, G., "Mill-Wide Advanced Quality Control for the Production of Newsprint", IMPC Conference, Helsinki, Finland. (2001).
3. LUPIEN, B., LAUZON, E., DESROCHERS, C., "PLS Modelling of Strength and Optical Properties of Newsprint at Papier Masson Ltée", *Pulp and Paper Canada* 102(5): 19-21 (2001).
4. SALTIN, J. F., STRAND, B. C., "Analysis and Control of Newsprint Quality and Paper Machine Operation Using Integrated Factor Networks", *Pulp and Paper Canada* 96(7): 48-51 (1995).
5. NOBLEZA, G.C., "Multivariate Analysis of TMP Mill Operation Data", Technical Section CPPA Annual Meeting: B31-B36 (1997).
6. M^CDONALD, D., MILES, K., AMIRI, R. "The Nature of the Mechanical Pulping Process", *Pulp and Paper Canada* 105(8): 27-32 (2004).
7. HARRISON, R.P., LEROUX, R., STUART, P.R. "Representing TMP Process Fundamentals by Creating Non-Linear Terms in Multivariate Analysis", 91st Annual PAPTAC Meeting Preprints, D563-D568 (2005).
8. BURNHAM, A.J., MACGREGOR, J.F., VIVEROS, R. "Latent Variable Multivariate Regression Modelling", *Chemometrics and Intelligent Laboratory Systems* 48(2): 167-180. (1999)
9. THORNHILL, N.F., SHOUKAT CHOUDHURY, M.A.A., SHAH, S.L. "The Impact of Compression on Data-Driven Process Analyses", *Journal of Process Control* 14(4): 389-398 (2004).
10. ELSINGA, M. "TMP Optimization Using Multivariate Analysis", Proceedings from IEEE Pulp & Paper Industry Technical Conference, 10-15. (2002).
11. BROWNE, T., MILES, K., M^CDONALD, D., WOOD, J., "Multivariate Analysis of Seasonal Pulp Quality Variations in a TMP Mill", *Pulp and Paper Canada* 105(10): 35-39 (2004).
12. HODOUIN, D., MACGREGOR, J.F., HOU, M., FRANKLIN, M. "Multivariate Statistical Analysis of Mineral Processing Plant Data", *CIM Bulletin* 86(975): 230-34 (1993)
13. KRESTA, J.V., MARLIN, T.E., MACGREGOR, J.F. "Development of Inferential Process Models Using PLS", *Computers and Chemical Engineering* 18(7):597-611. (1994)
14. UMETRICS AB, *User Guide: Simca-P and Simca-P+* 10, Umea, Sweden, (2002).
15. BELITSKY, E. "Maintenance by Numbers", Plant Engineering and Maintenance Magazine, November (2001).