RR-96-09

Multivariate Process Monitoring with Partial Least Squares Regression: A Case Study

N. Yatawara
University of Waterloo

IIQP Research Report
RR-96-09
December 1996

MULTIVARIATE PROCESS MONITORING
with
PARTIAL LEAST SQUARES REGRESSION:
A CASE STUDY

Nihal Yatawara [1]
Department of Statistics
and Actuarial Science
University of Waterloo
Waterloo, ON N2L 3G1
Canada

SUMMARY

Latent variable regression methods are becoming increasingly popular in process industries due to their dimensionality reduction capabilities. In particular, when there are multivariate input - output relationships present, the partial least squares regression (PLSR) seems to offer a viable alternative to extensions of univariate control charts such as the Shewhart $\overline{X}$, CUSUM or the EWMA for monitoring purposes. PLSR tends to result in bias predictions, similar to ordinary least squares. However, its power as a diagnostic tool well compensates for this drawback. This paper gives an overview of the PLSR method and its use in monitoring the operating performance of a crusher used in a mineral processing plant.

Key words and phrases: Predictions, latent variables, Non-linear-iterative-Partial Least Squares.

1 Introduction

Most mineral processing industries now use computers in their operations and hence large amounts of data are collected routinely. This data contains a wealth of information that could be used on-line or off-line to monitor the performance of the operations and to improve upon it. The data also provide a measure of the current state of a process, which can be useful in implementing process control procedures. Often, such procedures seek combined engineering and statistical data based knowledge of the processes. Another use of data would be to build suitable models giving direct or indirect input-output relationships to better understand the process behaviour.

The classical approach to handling process control through univariate-control charting techniques such as the Shewhart, CUSUM, EWMA, suffer from several drawbacks. In the

[1] On leave from: Curtin University of Technology, G.P.O. Box U1987, Perth, W. A. 6001, Australia

presence of complex relationships among process variables, as in mineral processing industries, these simple graphical methods offer very little to be of any use as performance indicators. This is partly because of the difficulties involved in identifying the true state of a process from a display of several variables.

In addition the existence of complex correlations among variables prohibit the use of univariate control charts based upon independent assumptions. Most variables in mineral processing industries are correlated as they vary jointly with one another. Usually, the ore properties such as the hardness and the initial particle size affect the feed rate, the pressure exerted by the crushers and the amount that goes through the bypass etc. are correlated. Missing data and excessive noise also create further problems, making it difficult to extract accurate information. In Engineering process control, plant engineers use phenomenological assumptions to find solutions to these problems.

In this paper the complementary approach based on a statistical technique known as the Partial Least Squares Regression is discussed. This method is a multivariate statistical method which projects information on to low dimensional subspaces containing only the relevant information about the process.

2 Method of Partial Least Squares Regression (PLSR)

In mineral processing applications, there are input as well as output variables present and we need to find models which are capable of not only expressing the variability within input or output variables but which are also most predictive of the output.

Ideally multiple linear regression seems to be the simplest method which satisfies these requirements. However, it does not offer a meaningful solution in the presence of noisy correlated data. Even the recent extensions of multiple linear regression methods such as ridge or regularization do not answer the question of dimensionality reduction and the PLSR method seems to be most promising under these circumstances. The PLSR method creates a set of orthogonal latent vectors from a block of input variables usually associated with process parameters (X block) that maximises the covariance between those vectors and a block of output variables associated with product quality (Y block) (Hoskuldson (1988)).

In brief, the partial least squares regression seeks decompositions of the X and Y matrices as

$$X = t_1p_1' + t_2p_2' + \cdots\cdots + t_Ap_A' + E_A$$

$$\text{and} \quad Y = t_1q_1 + t_2q_2 + \cdots\cdots + t_Aq_A + F_A$$

where $X(n \times k)$ and $Y(n \times m)$ are matrices containing the explanatory and response variables, t_a's are the N-vectors of latent variables; p_a's are K-vector loadings, E_A and F_A are residual matrics and q_a's are scalars.

The problem of finding the scores and the loadings is of primary interest and in PLSR, these are determined by choosing subspaces of the column space of X sequentially and then projecting y onto these subspaces (Helland (1988)). These subspaces are constructed by linear combinations of columns of X, represented by Xw_i, where w_i is a vector of weights. To initialize the procedure the first weight w_1 is computed to yield the maximum sample covariance between Xw_1 and Y i.e., we maximize the quantity $w_1'X'XYY'Xw_1$, subject to the usual constraint that

$$w_1'w_1 = 1.$$

The maximum is obtained when w_1 is the largest eigenvector of the matrix $X'YY'X$, and the first PLSR latent variable t_1, is given by

$$t_1 = Xw_1.$$

This choice of w_1 would give the maximal reduction in the residual covariance matrix of Y:

$$Y'Y - \frac{Y'Xw_1w_1'X'Y}{w_1'X'Xw_1}.$$

To obtain further weights and latent variables the procedure is continued by replacing X and Y by their residual matrices

$$X_{i+1} = \left(I - \frac{t_it_i'}{t_i't_i}\right) X_i \quad \text{and} \quad Y_{i+1} = \left(I - \frac{t_it_i'}{t_i't_i}\right) Y_i$$

for $i = 1, 2, \ldots, A-1$ where $X_1 = X$ and $Y_1 = Y$ (Burnham et al (1996)).

The optimality results for PLSR predictions have not been discussed in the literature. However, it has been shown that (Goutis (1996), Frank and Friedman (1993)) the overall length of the vector of PLSR coefficients is less than those of the ordinary least squares.

Several applications of PLSR have appeared in a number of recent articles, (S. Wold (1987), Hoskuldson (1988), MacGregor (1991)) which also provide some mathematical details concerning the procedure. In practice, one can use the NIPALS (non-linear iterative partial least squares) algorithm to perform (Hoskuldson (1988)) the above computations, as described below.

1. Set u equal to a column of Y.

2. Obtain w by $w' = u'X/u'u$ (i.e. regress columns of X on u).

3. Normalize w to unit length.

4. Calculate the score vector by $t = Xw/w'w$.

5. Obtain $q' = t'Y/t't$ by regressing columns of Y on t.

6. Calculate a new vector u by $u = Yq/q'q$.

7. Check for convergence, if yes, to 8 if no to 2.

8. X loadings: $p = X't/t't$.

9. Regression: $b = u't/t't$.

10. Calculate the residual matrices:

$$\begin{aligned} E &= X - tp' \text{ and} \\ F &= Y - btq' \end{aligned}$$

11. Repeat steps after replacing X and Y by E and F.

3 PLSR monitoring methods and diagnostics

Fundamentally, the PLSR monitoring method uses Shewhart charting techniques but plots latent variables against each other rather than a quality measure against time. To be practically useful, the process variability should be described with three or less latent variables to allow simple charting. Also, a data sample when the process is "in-control" is needed as a reference data set, to determine control regions. If these data are reasonably accurately measured then the latent variables may be considered to follow a multivariate normal distribution and confidence intervals can be formed. (Kresta et al, 1991).

As we gather new observations, then the corresponding latent variables t_1, t_2, etc. can be calculated and they can be plotted in the in-control plots. Points that fall outside the in-control regions indicate a departure from the "normal" operating mode. Also the weight plots can be used in conjunction with the latent variable plots to understand out of control directions. Furthermore, squared prediction errors defined by

$$SPE_y = \sum_{i=1}^{m} \left(Y_i - \hat{Y}_i\right)^2,$$

where m is the number of Y variables, can be plotted against the latent variables. An observed increase in the value of the SPE_y indicates a change in the X variables, not explained by the reference set. As discussed in MacGregor et al., 1994, these simple diagnostic capabilities are some of the attractive features of the PLSR method that have captured the attention of the practitioners.

4 The Processing Plant and Data

The mineral processing plant under investigation is in Western Australia (Yatawara et al (1996)). The study here was done mainly to develop an efficient monitoring mechanism for the rolls crusher of the operation. As shown in Figure 1a, Ore is conveyed from a storage bin into a surge bin above two cylindrical crusher rollers that are powered by individual electric motors. The Ore is drawn into the crushing gap by the opposing rotation of the rollers. As the Ore is fed through the rollers they are forced apart and an opposing hydraulic pressure is applied to maintain the gap width between them. Instruments are located to measure the important parameters of the operation (Figure 1b)). The HPRC 1A and 1B differ only in that the powered rollers for 1B have variable speed drives to allow for variation in total Ore feed through the crushing plant area, and this instrument is represented by the variable VSPD1B.

The other variables are: Feed 1X - the (inferred feed rate calculated from conveyor speed (tonnes per hour)), Gap 1X - the gap width between the rollers (millimeters) Pres 1X1 and Pres 1X2 - opposing hydraulic pressures for each roller (Bar), PWR1X and PWR1XF - power draw for each roller motor (kilowatts). Note: X = A or B.

There is a conveyor that splits some of the Ore from the stream to the HPRC feed bins and is directed around the HPRC plant and onto conveyor CV53. This is the bypass Ore and is preset to a value depending on the total plant feed target and the current operating conditions of the HPRC and is called BYPASS (tonnes/hr). The screening plant divides the Ore stream into 3 sizes > 18 mm (cv5); between 3 mm and 18 mm (cv3); and < 3 mm. The < 3 mm material is called "degrit" and is considered as the critical Ore size for downstream plant efficiencies.

The data for the study were taken from 4 days, consisting of 1440 observations taken at 1 minute intervals on each day. Since the measurements are taken from 3 locations a lag correction was introduced, by treating CV53 as the reference point. Day 4 data set appeared to be the best candidate for a reference set and all other days were suitably calibrated (see Figure 2). The annotations for each days data are given in the following table.

1/8	1A in poor condition due to it being nearly fully worn. Segment change imminent. 1B operating moderately well.
3/8	1A off line for segment change 1B operating relatively poorly until later in day, variable ore dayshift
12/9	The ore was relatively soft and the pressures were reduced on 1B to try a new mode of operation
13/9	1A hardly operated during the day because of a segment change but 1B appeared to be operating well

***Table* 1 Annotations for Data Sets**

5 Application of PLSR for HPRC Data

The application of PLSR to process monitoring requires a reference data set to establish "in-control" boundaries. The data provided came from four different days, with differing operating conditions, none of which showed stability. However by comparison, set four was found to be the most suitable as a reference data set. The analysis was conducted in two parts to demonstrate that a control chart could be produced when either both crushers or only one crusher is operating. These are referred to as dual and single crusher studies respectively. The dual crusher study was not useful for examining the diagnostic capability of the PLSR method because crusher 1A was often faulty. Hence the results presented are mainly from the single crusher study with one (degrit) or two (degrit and crushed) response variables.

The F values associated with the PLSR algorithm for both studies indicate either 2 or 3 latent variables as significant (see Figures 3 and 4 in the Appendix). The t-value plots and the weight plots (see Figures 5 and 6) indicate that the single response variable model with three latent variables produced a better reference area. Also these plots show some evidence of two distinct clusters indicating two modes of operation. The SPE charts seem to show that some observations in the reference data are not predicted by the model. Both plots show significant "drifts" of the observations into the third quadrant. Thus using this model as a reference, the PLSR algorithm was run on the data sets 2 and 3 and the corresponding plots (see Figures 7 and 8) were compared with those of the reference set. A summary of the results for the data set 2 is given below.

Plot	$t_1 \times t_3$	$WGT_1 \times WGT_3$	
Observations	two small clusters, one above and one below the reference area	GAP1B	increase or decrease
		FEED1B	increase or decrease
		VSPD1B	increase or decrease
Plot	$t_1 \times t_2$	$WGT_1 \times WGT_2$	
Observations	large group of observations along the diagonal projecting well into the third quadrant	PRES1B1 and PRES1B2	increase
		FEED1B	decrease
		PWR1B and PWR1BF	decrease
		GAP1B	decrease
Plot	$t_2 \times t_3$	$WGT_2 \times WGT_3$	
Observations	same group and observations as before projecting in the negative t_2 direction	PRES1B1 and PRES1B2	increase
		FEED1B	decrease
		BYPASS	decrease

The results indicate an increase in PRES1B1 and PRES1B2, when compared with the reference data set. A check of the summary statistics for the Reference Set and for Day 2 data shows that the mean pressure during Day 2 is higher than for the reference data and the coefficient of variation is lower, indicating a consistently higher pressure for Day 2. By checking the mean value of the other variables, and placing more importance on those variables that also have low coefficients of variation (because they are *consistently* higher or lower than the reference values) it can be seen that FEED1B and GAP1B are also significantly and consistently lower for Day 2 data as predicted by the weight plots.

Reference Data

Variable	Minimum	Maximum	Range	Mean	Std Dev	CV
FEED1B	201.63	954.54	752.91	800.77	98.51	12.30
BYPASS	0.00	774.10	774.10	269.79	164.28	60.89
GAP1B	21.31	37.78	16.48	29.30	4.27	14.59
PRES1B1	28.88	41.10	12.22	34.56	2.70	7.83
PRES1B2	30.30	40.21	9.91	34.89	2.27	6.51
PWR1B	411.69	962.55	550.85	851.46	44.57	5.23
PWR1BF	133.59	929.04	795.46	820.60	49.36	6.01
VSPD1B	87.08	94.99	7.92	94.97	0.37	0.39

Day 2 Data

Variable	Minimum	Maximum	Range	Mean	Std Dev	CV
FEED1B	119.84	659.30	539.46	541.20	39.58	7.31
BYPASS	0.00	753.27	753.27	273.12	129.34	47.35
GAP1B	16.92	20.74	3.83	18.21	0.91	4.97
PRES1B1	34.10	37.79	3.69	35.39	0.82	2.32
PRES1B2	35.23	39.70	4.47	36.45	0.88	2.42
PWR1B	645.01	1043.99	398.98	890.15	93.27	10.48
PWR1BF	620.79	982.48	361.69	844.99	88.48	10.47
VSPD1B	87.08	94.99	7.92	94.97	0.40	0.42

***Table* 2: Summary Statistics for the Non Standardised Data for the Single Crusher Study**

6 Conclusion

The dual crusher study has shown that the PLSR method can be successfully applied to HPRC data, for performance monitoring. However, the diagnostics were much better when the weight charts were plotted from a single crusher analysis. The diagnostic tools provided by PLSR are not able to specify a cause for an "out-of-control" signal in the process, but they do indicate those process variables that contribute most weight to a particular shift in the observations on the latent variable planes. This information can assist an experienced plant operator to identify actual operating faults. Clearly the PLSR control charting model will be effective only if the reference data set is obtained from a known "in-control" period of operation. A real reference data set might produce a model different to what was developed in this study, where "constructed" reference sets were used. One obvious characteristic seen in the latent variable plots for reference data is the two clusterings. As pointed out by Kresta et al. (1991), this might indicate different modes of operation. Further investigation would also be required into the effects of different Ore types (soft or hard) on the PLSR model.

Although the PLSR model developed here considered past process data, the model can be adapted to an on-line monitoring system. In this way one can identify the time point at which the process begins to deviate from the normal operating mode. An obvious problem associated with the display of the control chart with the PLSR seems to be its complexity as compared to univariate shewhart plotting. However, for all practical purposes it would be sufficient to display only the previous p-observations, where p is determined as the period of interest (see Kresta et al. (1991)).

7 References

1. A.J. Burnham, R. Viveros and J.F. MacGregor. Frameworks for Latent Variable Multivariate Regression, *Journal of Chemometrics*, 10, 31-45, 1996.

2. I.E. Frank and J.H. Friedman. A Statistical Review of some Chemometrics Regression Tools, *Technometrics*, 35, 109-148, 1993.

3. P. Geladi and B.R. Kowalski. Partial Least Squares: A Tutorial, *Analytical Chemica Acta*, 185, 1-17, 1986.

4. C. Goutis. Partial Least Squares Algorithm Yields Shrinkage Estimators, *The Annals of Statistics*, 24, 816-824, 1996.

5. I.S. Helland. On the Structure of Partial Least Squares Regression, *Comm. Stat., B—Simulation and Computation*, 17, 581-607, 1988.

6. D. Hodovin, J.F. MacGregor, M. Hou and M. Franklin, Multivariate Statistical Analysis of Mineral Processing Plant Data, *CIM Bulletin Mineral Processes*, 86, 23-34, 1993.

7. A. Hoskuldsson, PLS Regression Methods, *Journal of Chemometrics*, 2, 211-228, 1988.

8. M. Kresta, J.F. MacGregor and T.E. Marlin, Multivariate Statistical Monitoring of Process Operating Performance, *Canadian Journal of Chemical Engineering*, 69, 35-47, 1991.

9. J.F. MacGregor, C. Jaecle, C. Kiparissides and M. Koutoudi, Process Monitoring and Diagnosis by Multiblock PLS Methods, *AiChE Journal*, 40, 826-838, 1994.

10. J.F. McGregor and T. Kourti, Statistical Process Control of Multivariate Processing, *Control Engineering Practice*, 3, 408-414, 1995.

11. N. Yatawara and J. Harrison. Multivariate Statistical Process Control of a Mineral Processing Industry, Proceedings of the QISM Conference, Cochin, India, Dec. 28-31, 1996

Acknowledgments

This work was partly written during the author's study leave at Waterloo and wishes to thank Professor Bovas Abraham former director of the IIQP for the financial support given. Also, the author thanks Jeff Harrison for carrying out the necessary computations.

Appendix

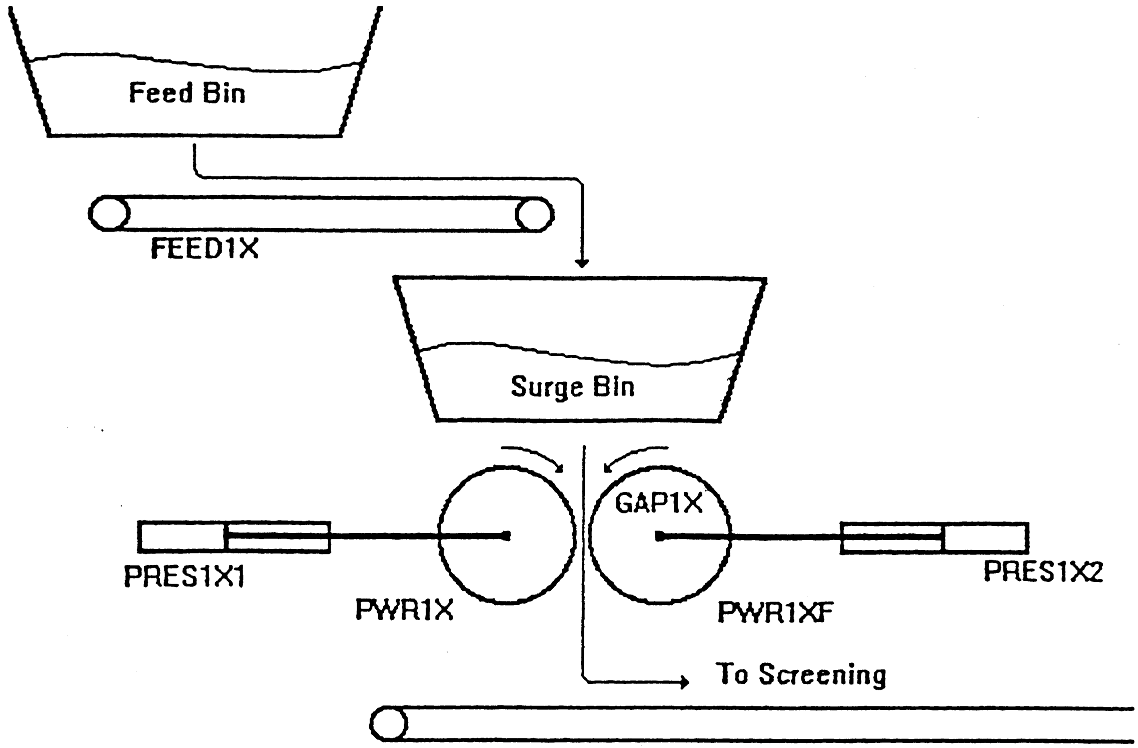

Figure 1a: Schematic of the HPRC Plant Showing Process Measurement Instruments

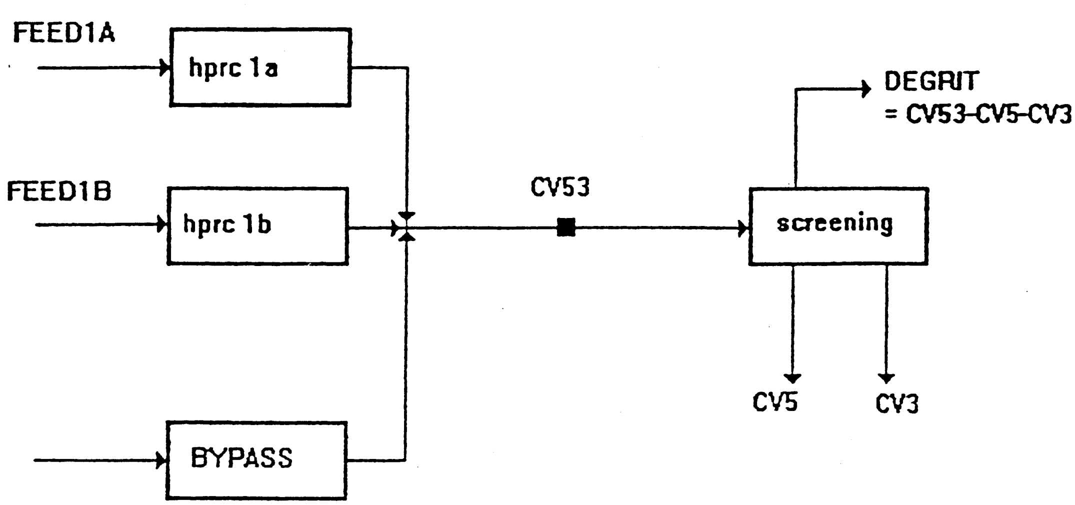

Figure 1b: Schematic of the Ore Crushing Process

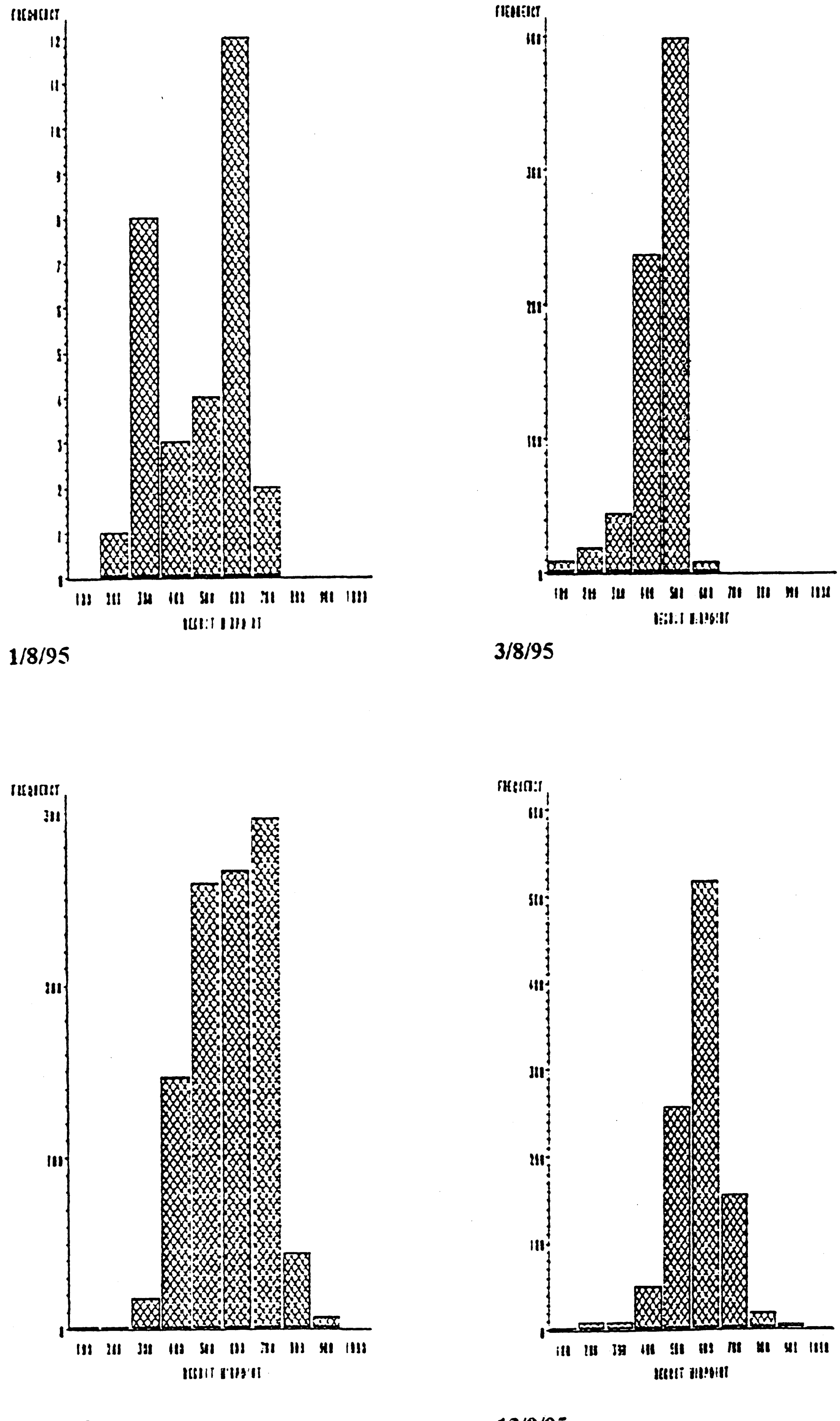

Figure 2: Histogram of DEGRIT for Crusher 1B for Each Data Set

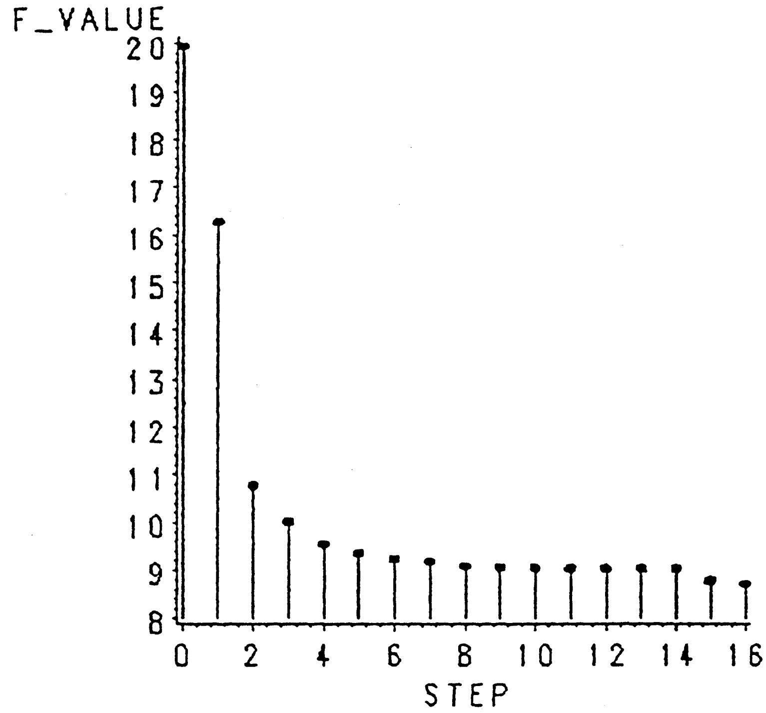

Figure 3: **F-Value by Latent Variables Extracted for Dual Crushers Reference Data with 1 Response Variable and BYPASS as an X Variable**

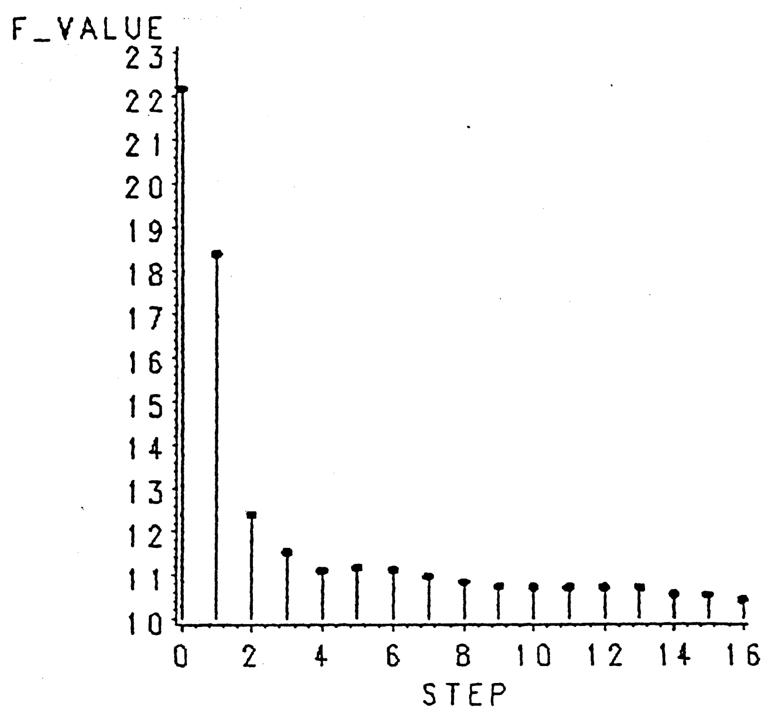

Figure 4: **F-Value by Latent Variables Extracted for Dual Crushers Reference Data with 2 Response Variables including a BYPASS Factor**

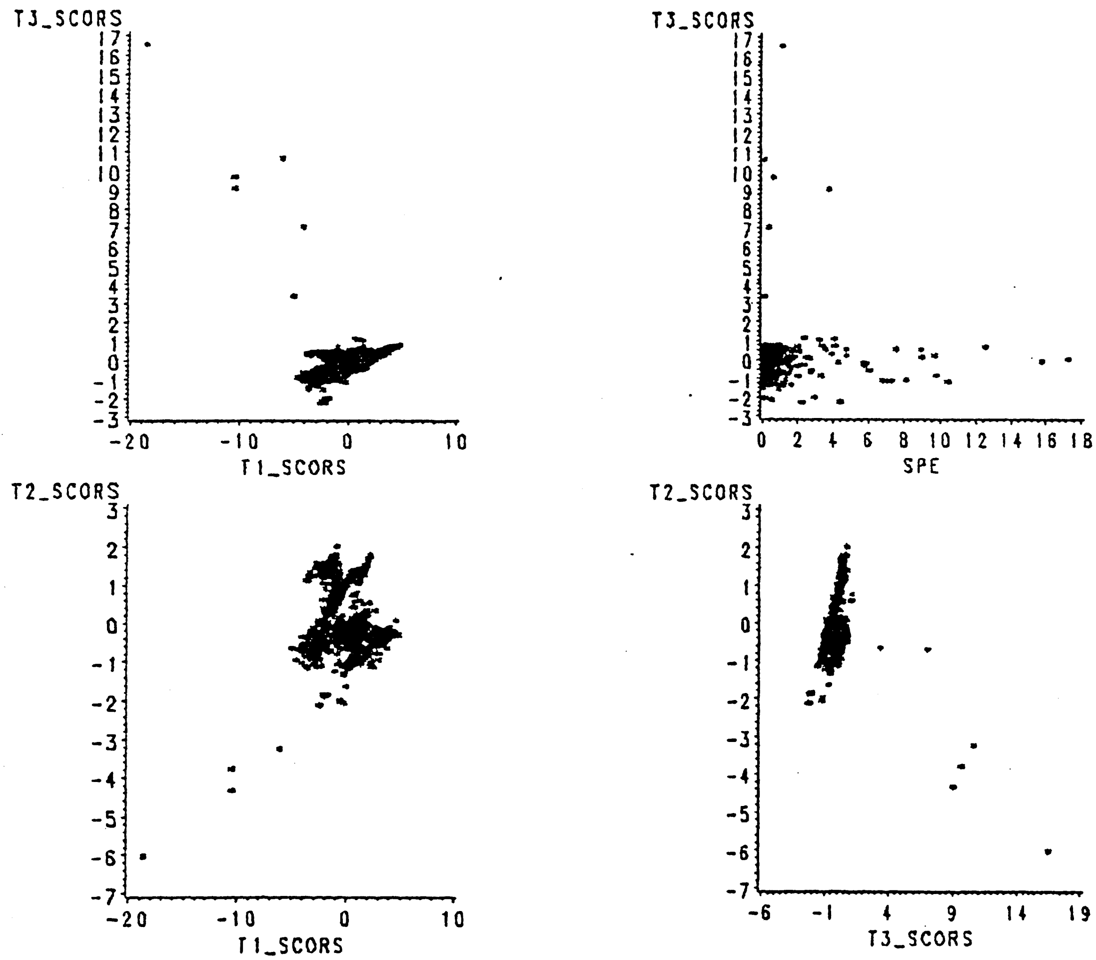

Figure 5: Monitoring and SPE Chart for Reference Set Crusher 1B (1 Response Variable)

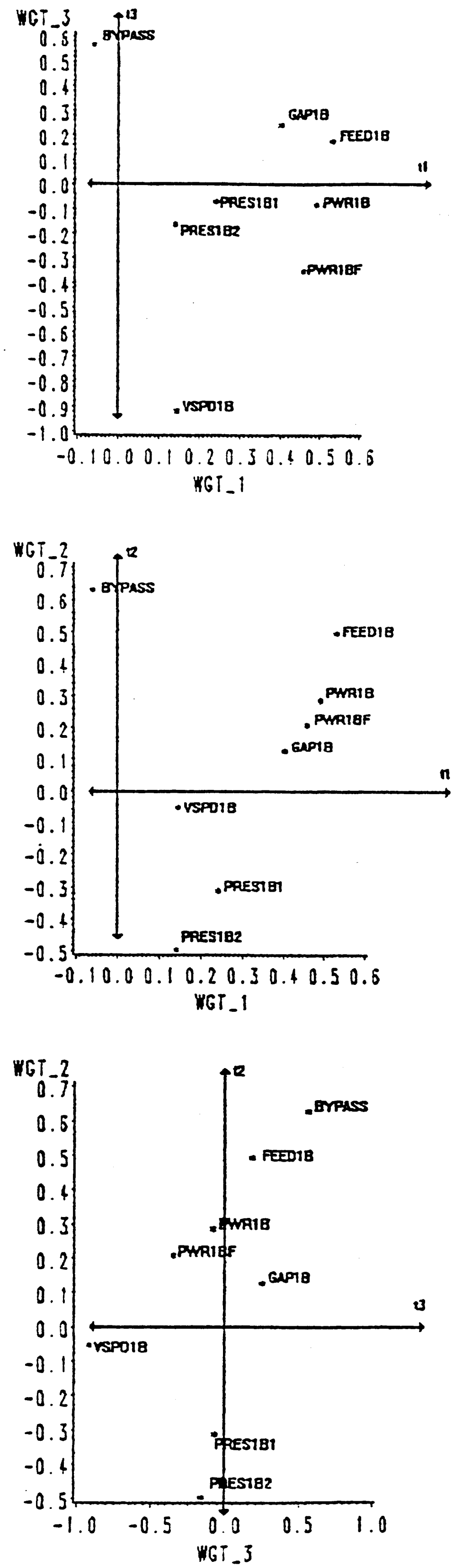

Figure 6: Plots of Latent Variable Weights for Single Crusher 1B (1 Response Variable)

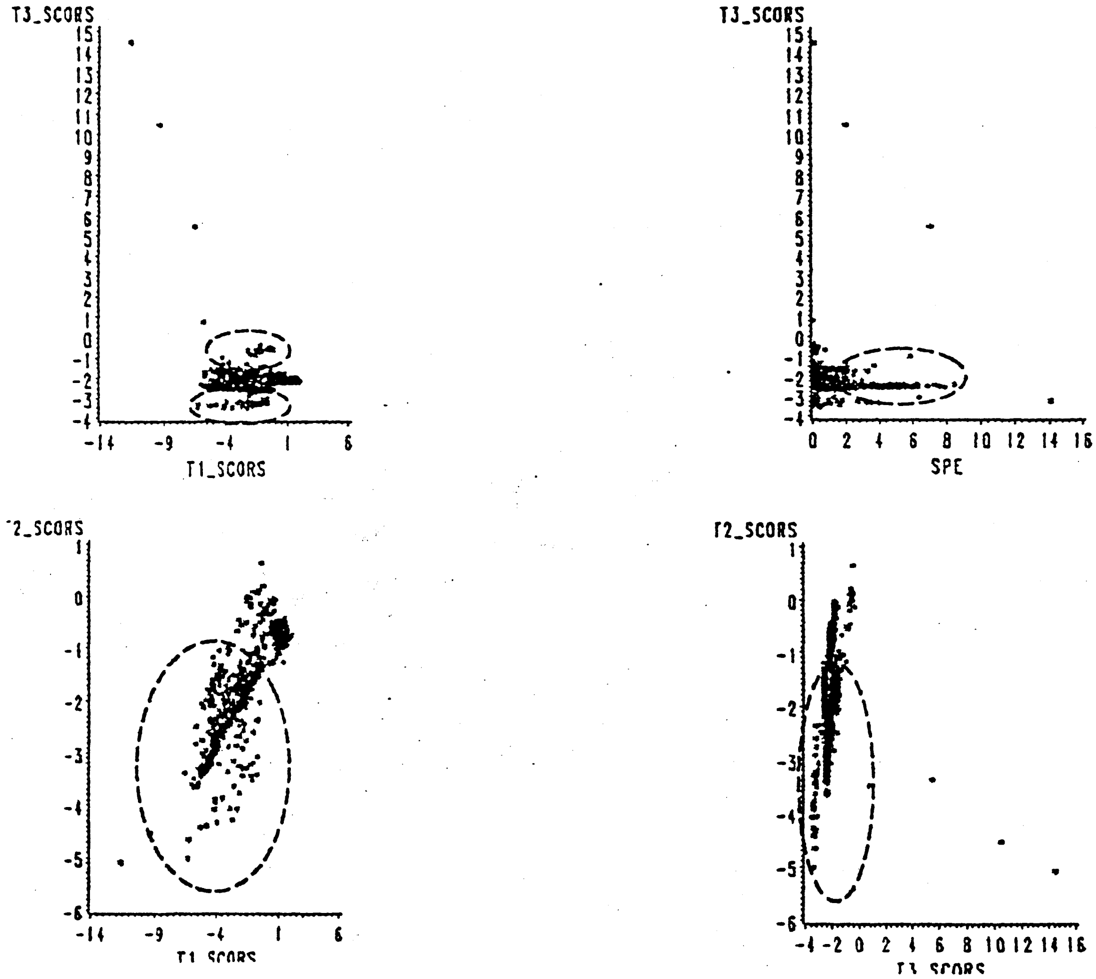

Figure 7: Monitoring and SPE Chart for 3/8/95 Data - Crusher 1B (1 Response Variable)

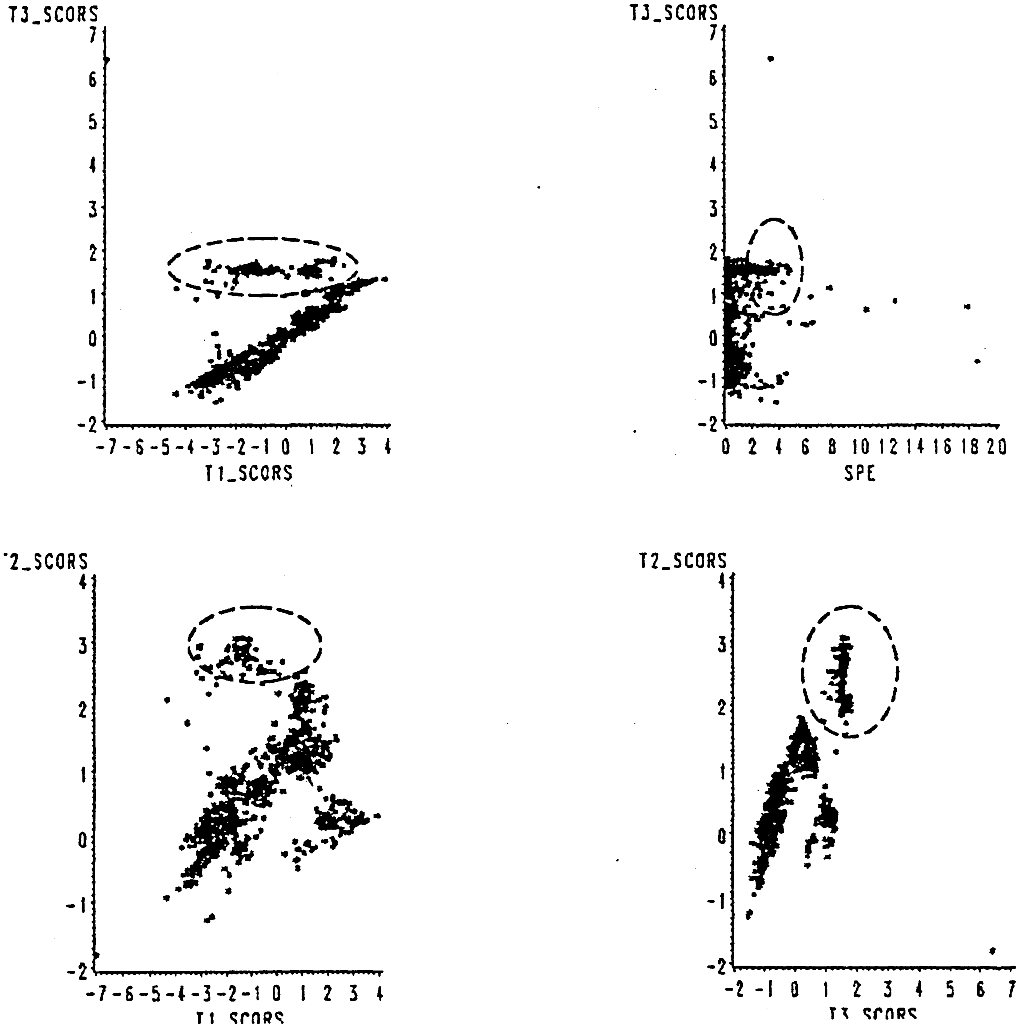

Figure 8: Monitoring and SPE Chart for 12/9/95 Data - Crusher 1B (1 Response Variable)