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

Multivariate Data Analysis of Gas-Metal Arc Welding Process

Rajesh Ranjan * **Anurag Talati** ** **Megan Ho** ***
Hussain Bharmal **** **Vinay A. Bavdekar** ** **Vinay Prasad** **,1
Patricio Mendez **

* *Department of Polymer and Process Engineering, Indian Institute of Technology Roorkee, Roorkee, Uttarakhand, India, 247667*
** *Department of Chemical and Materials Engineering, University of Alberta, Edmonton, AB, T6G 2V4 Canada*
*** *Department of Materials Engineering, University of British Columbia, Vancouver, BC, V6T 1Z3 Canada*
**** *Department of Mechanical Engineering, Indian Institute of Technology Bombay, Mumbai, India, 400076*

Abstract: Gas-metal arc welding is a widely used welding process. The testing of such welds is done offline and in most cases after the welding operation is over. To monitor the progress of the welding run, it is essential to develop multivariate data analysis techniques that can classify the welds into good or bad runs and also be able to predict the quality variables. In this work, popular multivariate data analysis methods such as hierarchical clustering analysis, principal component analysis and partial least squares are used to develop classification and regression models to predict the weld quality based on various parameters. The results indicate that models obtained using these methods are effective in classification and prediction of weld quality and can be further developed for online and industrial uses in weld run monitoring.

Keywords: welding; data analysis; batch process; clustering; classification; multi-way principal component analysis; partial least squares; hierarchical clustering; soft sensors

1. INTRODUCTION

The gas-metal arc welding (GMAW) process is a commonly used welding process in industrial applications due to ease of operations and its versatility. In the GMAW process, an electric arc is formed between the consumable wire electrode and the workpiece metal. The arc formation causes the consumable wire and workpiece to melt and join. The area where the joining occurs is called a weld. To prevent contamination of the weld by the surrounding air during the welding process, an inert gas is fed along with the wire electrode to form a protective layer across the weld area during the weld process.

Conventionally, testing of the weld quality is performed off-line, with either destructive testing techniques (used on as few samples as possible) and non-destructive testing (NDT) techniques. The most common NDT is a visual inspection of the GMAW runs, which involves obtaining the penetration depth and the aspect ratio of the welds. All these testing techniques can only be used at the end of the welding runs and are mostly done on randomly selected samples. Univariate statistical analysis methods (Adolfsson et al., 1999; Siewert et al., 2008) have been previously used to monitor weld runs in various welding applications. Artificial neural network models have been developed to monitor the plasma radiation (García-Allende et al., 2009) in arc welds. These methods are univariate in nature, whilst the welding operations are multivariable operations. It is an established fact that univariate data analysis methods cannot accurately capture the effect of process variations in a multivariable process. This implies that multivariate techniques of modelling and analysis need to be used in order to ensure effective monitoring of the welding processes.

Data-driven multivariate statistical analysis methods such as hierarchical clustering analysis (HCA) (Jain et al., 1999), principal component analysis (PCA) and partial least squares (PLS) are widely used classification methods and are popular in process monitoring applications (Kresta et al., 1991). These methods are primarily used to analyse data sets with a large number of highly correlated variables. The HCA algorithms create a hierarchy of operations, in which the clusters are either broken into smaller parts in stages (divisive or top-down approach) or are agglomerated in stages (agglomerative or bottom-up approach) based on a metric of similarity. The principal assumption is that the process data conforms to a certain hierarchy of classification. Hence, the interpretations of the HCA results are more clear if *a priori* knowledge of the number of clusters or classes of the data is available. PCA is used when there is a single block of data, such as input data (Kresta et al., 1991). The method involves a linear coordinate transformation that spans the space of maximum variance, thereby capturing the maximum possible information in fewer dimensions. The PCA algo-

[1] Corresponding author; email: vprasad@ualberta.ca

rithm can be used only for continuous processes, where the autocorrelation of process variables is low (or ideally non-existent). Typical welding operations are batch processes, where the completion of each weld describes a batch operation. In batch operations, the aim is to minimise the batch-to-batch variations rather than the variation of process variables with time. This implies that the PCA algorithm cannot be used for classification and monitoring of batch processes as it is unable to distinguish process data from different batches. Nomikos and MacGregor (1994) developed the multiway PCA (MPCA) algorithm, which can be used for monitoring and classification of batch data. The MPCA algorithm unfolds the batch data along the time axis and reduces the dimension of this data. The conventional PCA is applied on the unfolded batch data. The batch-to-batch variability can be monitored using this approach. PLS is used to find the fundamental relations between the dependent variables and independent variables; it is a latent variable approach to modelling the covariance structures in these two spaces (Kresta et al., 1991). A PLS model will try to find the multidimensional direction in the space of independent variables that explains the maximum multidimensional variance direction in the space of dependent variables. Both PCA and PLS are dimension reduction methods that operate in such a way that the reduced dimension space still explains the majority of the variance in the data. The unsupervised nature of all these algorithms makes them popular candidates for separating available process data into different classes.

In this work, the methods described above are applied on GMAW data obtained from experiments done in the welding laboratory. The key contributions of this work are development of HCA and MPCA methods to classify the welding runs into different categories. Further, a PLS model is developed to predict the weld quality based on visual inspection data, such as weld aspect ratio and weld penetration depth. The results obtained indicate that the data analysis methods are able to classify the welds and predict weld quality variables better than the conventional practices used in the welding industry.

The paper is organised as follows. Section 2 gives the important details about the algorithms used in this work to classify the welds. A brief explanation of the welding experiments in given in Section 3, followed by a discussion of the results obtained. Conclusions drawn from the analysis of the results in presented in Section 4.

2. OVERVIEW OF HCA, MPCA AND MPLS

2.1 Hierarchical Clustering Analysis

Hierarchical clustering approaches classify the data into various classes, which are separated by the degree of similarity associated with them (Manning et al., 2008). Thus, the degree of similarity results in a structure that looks like an hierarchy. Such a structure is called a dendrogram. The degree of similarity is achieved by using an appropriate metric of distance between the different samples. In this work, the Euclidean distance is used as a metric of similarity. Once the similarity metric has been obtained, the data is grouped together through a linkage, which uses the distance metric generated to determine the proximity of the objects to each other. This groups the data into binary clusters, which are then linked to other binary clusters to form bigger clusters until all the data points have been linked with each other to form a hierarchical tree (Jain et al., 1999). The popular linkage functions include farthest distance, shortest distance and mean distance. In this work, the shortest distance function was used to created the linkages. Once all the data has been linked through binary clusters, a dendrogram can be obtained, which shows the clusters and their similarity. The number of clusters can, then be decided based on a cut-off of the similarity metric, which is drawn from user/operator experience.

2.2 Multiway Principal Component Analysis

Similar to the conventional PCA algorithm (Kresta et al., 1991), the MPCA algorithm is also a dimensionality reduction technique that can be applied to batch processes (Nomikos and MacGregor, 1994). Consider a batch process with $j = 1, 2, \ldots, J$ variables, available at time intervals $k = 1, 2, \ldots, K$. For all batches $i = 1, 2, \ldots, I$, the data can be arranged as $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$. The three-way data array $\underline{\mathbf{X}}$ is unfolded along the K dimension to result in a two dimensional matrix $\mathbf{X} \in \mathbb{R}^{I \times JK}$. This places the variables measured at one particular time interval across all batches one below the other, while all variables at different time intervals are placed one next to the other. Unfolding the data in this manner allows monitoring of batch-to-batch variations. The standard PCA algorithm is then applied on the matrix $\mathbf{X}$. The PCA algorithm transforms the data co-ordinates such that each successive direction explains the maximum residual variance in $\mathbf{X}$. For example, the first direction, $\mathbf{t_1}$, is obtained as the solution of the following optimization problem

$$\begin{aligned} \max_{\mathbf{p_1}} \left(\mathbf{t_1}^T \mathbf{t_1}\right) &= \mathbf{p_1}^T \mathbf{X}^T \mathbf{X} \mathbf{p_1} \\ \text{such that } \mathbf{p_1}^T \mathbf{p_1} &= 1 \end{aligned} \tag{1}$$

The solution of the above problem can be posed as an eigenvalue–eigenvector problem

$$\mathbf{X}^T \mathbf{X} \mathbf{p_1} = \lambda_1 \mathbf{p_1} \tag{2}$$

Thus, the PCA decomposes the covariance of $\mathbf{X}$ into a scores matrix $\mathbf{T}$ and an orthogonal loadings matrix $\mathbf{P}$. In order to minimise noise, loading vectors corresponding to the first R largest eigenvalues are retained. This results in a lower dimensional space that can still explain the largest possible variance in the data. The original three-way data array $\underline{\mathbf{X}}$ can be reconstructed as follows

$$\underline{\mathbf{X}} = \sum_{r=1}^{R} \mathbf{t}_r \bigotimes \mathbf{P} + \underline{\mathbf{E}} \tag{3}$$

where $\underline{\mathbf{E}}$ is the residual noise information that is not captured in the first R directions.

The Hotelling's T-squared distribution is used for online implementation of the MPCA algorithm. For the ith sample of process variables, the Hotelling's T-squared statistic is obtained as

$$T_i^2 = \mathbf{t}_i \mathbf{\Lambda}^{-1} \mathbf{t}_i \tag{4}$$

where $\mathbf{t}_i$ represents the scores of the ith sample and $\Lambda = diag\left[\lambda_1 \;\; \lambda_2 \;\; \ldots \;\; \lambda_R\right]$. An upper and lower threshold

of T_i are used to determine the normal batch runs. If the value of T_i crosses any of the thresholds, then the batch is said to deviate from normal behaviour.

2.3 Partial least squares

The partial least squares or projection onto latent structures (PLS) algorithm attempts to simultaneously reduce the dimensions of the spaces of independent variables ($\mathbf{X}$) and dependent variables ($\mathbf{Y}$) to obtain the latent vectors of the $\mathbf{X}$ and $\mathbf{Y}$ spaces that are highly correlated (Kresta et al., 1991). Thus, the PLS algorithm obtains a regression model between the transformed space of $\mathbf{X}$ and transformed space of $\mathbf{Y}$. A popular implementation of the PLS algorithm is briefly described here. Consider the following decomposition of the $\mathbf{X}$ and $\mathbf{Y}$ matrices

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + E_1 \tag{5}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^T + E_2 \tag{6}$$

where the matrices $\mathbf{T}$ and $\mathbf{U}$ are the scores matrices and $\mathbf{P}$ and $\mathbf{Q}$ are the loadings matrices of $\mathbf{X}$ and $\mathbf{Y}$ respectively. The regressor matrices are obtained as follows: Set $\mathbf{u}$ equal to one column of $\mathbf{Y}$. Obtain the corresponding scores as follows

$$\mathbf{w}^T = \mathbf{u}^T\mathbf{X}/\mathbf{u}^T\mathbf{u} \tag{7}$$

$$\mathbf{t} = \mathbf{X}\mathbf{w}/\mathbf{w}^T\mathbf{w} \tag{8}$$

Re-compute the scores of $\mathbf{Y}$ as follows

$$\mathbf{q}^T = \mathbf{t}^T\mathbf{Y}/\mathbf{t}^T\mathbf{t} \tag{9}$$

$$\mathbf{u} = \mathbf{Y}\mathbf{q}/\mathbf{q}^T\mathbf{q} \tag{10}$$

Solve equations 7–10 until the values of $\mathbf{t}$ and $\mathbf{u}$ converge. Obtain the loadings of $\mathbf{X}$ and regressor of $\mathbf{Y}$ as follows

$$\mathbf{p} = \mathbf{X}^T\mathbf{t}/\mathbf{t}^T\mathbf{t} \tag{11}$$

$$\mathbf{b} = \mathbf{u}^T\mathbf{t}/\mathbf{t}^T\mathbf{t} \tag{12}$$

Compute the residuals as

$$\mathbf{E}_1 = \mathbf{X} - \mathbf{t}\mathbf{p}^T \tag{13}$$

$$\mathbf{E}_2 = \mathbf{Y} - \mathbf{b}\mathbf{t}\mathbf{q}^T \tag{14}$$

To compute the next set of latent vectors, replace $\mathbf{X}$ and $\mathbf{Y}$ with $\mathbf{E}_1$ and $\mathbf{E}_2$. Repeat the steps till the desired number of latent vectors are obtained.

3. WELDING EXPERIMENTS

The welding experiments were performed using a GMAW power supply in CV (constant voltage) mode and 0.045 in solid steel wire. A mixture of 75% argon and 25% carbon dioxide was used as a shielding gas. Typical voltage and current changes in a batch run are shown in Fig. 1 and Fig. 2. Experience of welding operations indicate that the following parameters affect the weld quality– contact tip to workpiece distance (CTWD), wire feed speed (WFS) and voltage. These are usual "essential variables" in codes, such as the ASME boiler and pressure vessel code (BVPC) (ASME, 2010). The range of these parameters that result in an acceptable weld are typically known through experience. Seventeen experiments of GMAW were run for different values of CTWD, WFS and voltage, which are shown in Table 1. The order of the batch runs was deliberately selected randomly, so as to minimize the effect of contact tip wear on weld quality. For example, batch runs in which voltage was manipulated were not run in sequence. The determination of what constitutes a "good" or "bad" weld depends on the application, and can involve radically different metrics. In this work, the metrics of weld quality used were the aspect ratio (AR) and penetration depth (PD) of the weld in each batch run.

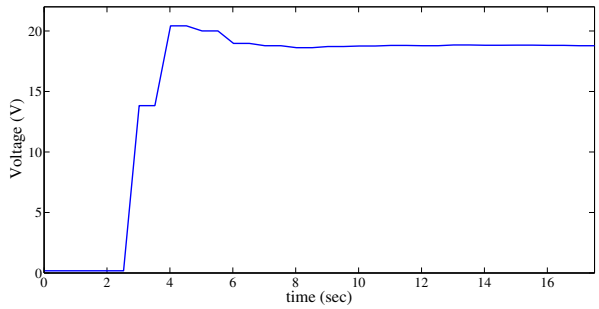

Fig. 1. Voltage trajectory for a typical run

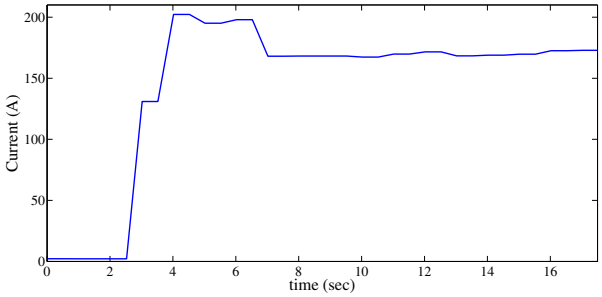

Fig. 2. Current trajectory for a typical run

Table 1. Operating conditions of different batch runs

S. No.	CTWD (mm) (mm)	WFS (m/min)	Voltage (V)	Travel speed (m/min)
1	9	3.8	18	0.3
2	12	3.8	18	0.3
3	15	3.8	18	0.3
4	19	3.8	18	0.3
5	22	3.8	18	0.3
6	15	2.8	18	0.3
7	15	3.2	18	0.3
8	15	5.1	18	0.3
9	15	5.5	18	0.3
10	15	3.8	16	0.3
11	15	3.8	17	0.3
12	15	3.8	19	0.3
13	15	3.8	20	0.3
14	15	3.8	18	0.1
15	15	3.8	18	0.2
16	15	3.8	18	0.4
17	15	3.8	18	0.5

3.1 Results

The different input conditions result in batches that have unequal run-time. In order to develop models using the methods mentioned in Section 2, it is essential to have batches of equal length. This was achieved by truncating the batches with longer run-times. The resulting loss in information is minimal since these parts involve winding down and shutting the welding process, which has very

little impact on the final weld quality. Initially, the standard deviations of the current and voltage values from an arbitrarily selected "base run" were evaluated to see if they could be used to classify the batch operations into groups of similar runs. This is standard industrial practice and also used as a univariate monitoring technique (Adolfsson et al., 1999). From operators' experience, it was determined that batch number 9 was a "good run" in terms of its weld quality, and hence this was selected as the base case. For the rest of the runs, the standard deviation of the voltage and current was obtained around the mean value of the voltage and current of run 9. The results of this exercise are reported in Table 2. From the table, it can be seen that there is not much difference in the standard deviation values of all runs and hence it is difficult to classify the runs into similar groups of welding operations based on the standard deviation alone.

Table 2. Standard deviation of each batch from batch 9

Batch No.	Avg. V	std. dev (V)	Avg. I	std. dev (I)
1	17.933	9.224	131.748	133.057
2	17.956	12.433	132.637	142.375
3	16.655	9.260	140.283	132.351
4	17.777	9.344	148.009	136.172
5	17.835	9.477	143.235	137.981
6	17.835	9.477	143.235	137.981
7	17.884	9.244	132.644	132.736
8	17.606	9.299	186.101	136.605
9	**17.823**	**0**	**142.874**	**0**
10	17.869	9.039	124.675	129.143
11	17.705	8.958	126.024	127.569
12	18.829	9.230	139.394	135.039
13	19.839	8.962	132.675	137.488
14	17.930	8.802	116.599	125.571
15	18.013	8.799	104.188	126.690
16	17.775	9.184	132.218	130.661
17	17.673	9.110	166.337	126.806

Since the standard deviations in voltage and current were not able to classify the runs, other classification methods were used to obtain the similar groups in the welding experiments. HCA was performed on the two sets of data: (a) input data of current (I) and voltage (V) and (b) visual inspection data, AR and PD. The results of the HCA performed on input data are shown in Fig. 3. From the figure, it can be seen that similar runs can be clubbed as follows– runs (4, 11 and 12), runs (1 and 9) and runs (2, 5, 6, 7, 15 and 17). Runs 3, 8, 10, 13 and 14 are different from the other groups and do not fall into any group. However, when the HCA analysis is performed using the visual inspection data, the results are different. From Fig. 4, it can be seen that the grouping of the runs is different from that obtained through HCA on input data. This indicates that the classification of the inputs is not sufficient to predict the visual inspection data, and a correlation needs to be developed between them.

The MPCA algorithm was applied on the input data of the batch runs, I and V, to group batch runs that are similar to each other. For the purpose of this analysis, the size of all batches were maintained constant by truncating the longer runs. From Fig. 5 it can be seen that four principal components cumulatively explain more than 99% variance in the data. Fig. 6 is a biplot of the scores of the first two principal components. The runs that are similar

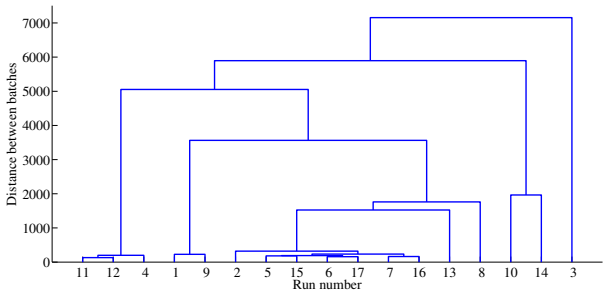

Fig. 3. HCA results on input data of the runs

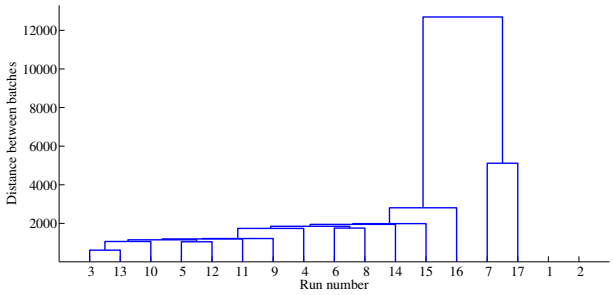

Fig. 4. HCA results on visual inspection data of the runs

get grouped close to each other, while the dissimilar ones are away from each other. The results indicate that runs (7, 16, 2, 15, 6, 5 and 17 were similar), (2, 4, and 11 were similar), (9 and 1 were similar) whereas the runs 3, 8, 13, 10 and 14 were different from the rest. It should be noted that the choice of the cut-off for clustering similar samples is completely based on user experience. The residuals between obtained data and predictions using the MPCA model are obtained and the sum of squared residuals (SSR) for each run is shown in Fig. 7. The results indicate that runs 8, 10 and 14 are not normal runs due to the higher scores, which is consistent with the analysis obtained through HCA and MPCA. This, however, still does not explain why runs 3 and 13 are different from the rest.

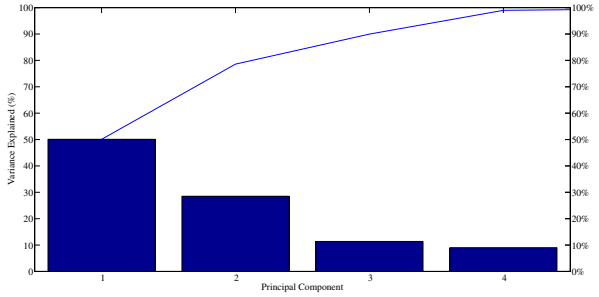

Fig. 5. MPCA: Variance explained by each principal component and cumulative variance

While HCA and MPCA are able to classify the batches according to their similarity, they are not able to predict the quality variables of the weld. PLS overcomes this shortcoming and helps in developing such models. Initially, PLS models were developed using average current (I_{avg}) and average voltage (V_{avg}) for each batch. The reason for

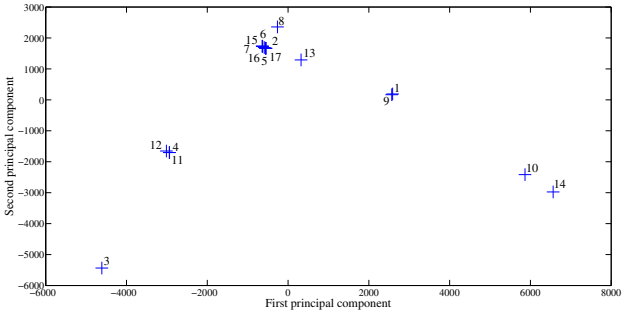

Fig. 6. MPCA: Scores plot of the batches indicating groups of similar runs

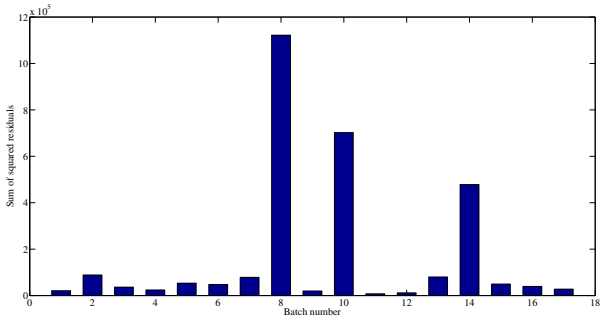

Fig. 7. Sum of squared residuals of MPCA predictions for all runs

using I_{avg} and V_{avg} is that operators use these values to gauge the progress of the weld run. The intention is to investigate if the PLS models using the average values of I and V for each batch are sufficient to provide good predictions of AR and PD. The following PLS models were developed: a) between (I_{avg}), (V_{avg}) and AR and b) between (I_{avg}), (V_{avg}) and PD. The model predictions vs. observed results are shown in Fig. 8 and Fig. 9. For a good prediction, the observations should lie along the diagonal. The results, however, indicate that this is not the case and the use of I_{avg} and V_{avg} is not a good indicator of the weld quality.

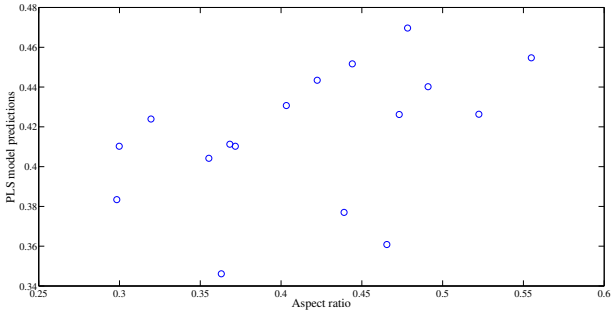

Fig. 8. PLS: model predictions using (I_{avg}), (V_{avg}) vs. observations of AR

Since I_{avg} and V_{avg} cannot be used to predict the AR and PD, the PLS models were obtained again by using the entire data of all runs. From Fig. 10 and Fig. 11 it can be seen that 4 components of the PLS model explain $> 96\%$ of the variance in AR and PD respectively. Plots of the PLS

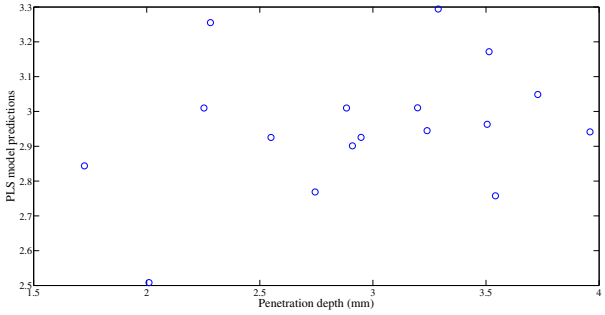

Fig. 9. PLS: model predictions using (I_{avg}), (V_{avg}) vs. observations of PD

model predictions vs obtained AR and PD measurements are shown in Fig. 12 and Fig. 13. From the figures, it can be seen that the PLS models are able to generate fairly accurate predictions of AR and PD.

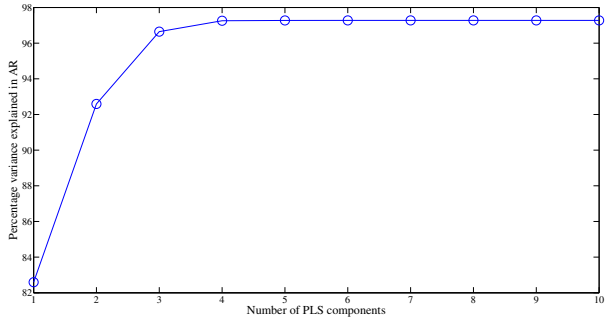

Fig. 10. PLS components vs. cumulative percentage variance explained for AR

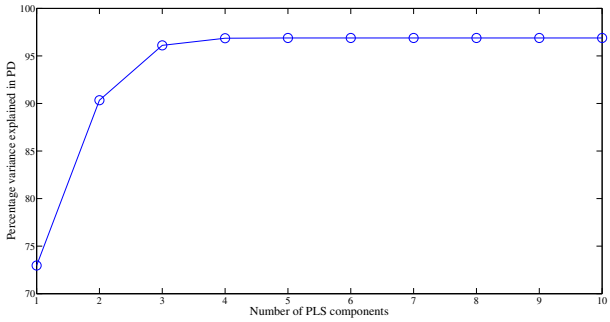

Fig. 11. PLS components vs. cumulative percentage variance explained for PD

4. CONCLUSIONS

In this work, data analysis methods such as MPCA and HCA were used to classify the GMAW runs in order to determine the "good quality" welds. PLS was used to build models that can predict the variables related to the quality of the welds. The use of such data analysis techniques make it possible for early detection of welding jobs that might not meet the expected quality parameters. Also, the use of such techniques can help reduce the need for performing metallography or using heuristic visual

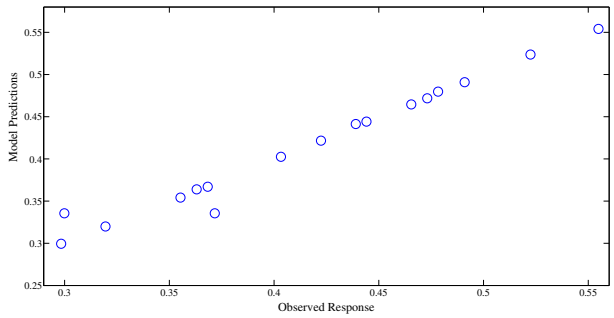

Fig. 12. PLS model for AR with 4 components: predictions vs. observed data

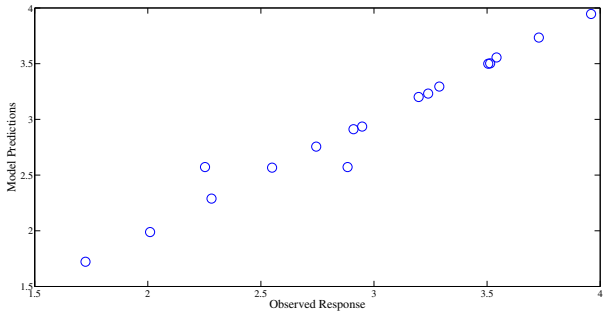

Fig. 13. PLS model for PD with 4 components: predictions vs. observed data

analysis techniques, which consume time and resources and can be done only at the end of the welding job.

The results indicate that the practice of using the standard deviation of the current and voltage values from a designated "good" run is not able to clearly differentiate between the welds. Further, the PLS models obtained using the average voltage and current values, which is a standard industrial practice to determine weld quality, do not give an accurate prediction of the weld quality. This implies that the trajectory of the inputs has an important role in the final outcome of the welding operation. Hence, HCA and MPCA models were built using the entire data of the welding runs. The HCA models developed using input data and visual inspection data do not result in the same classification, thereby indicating that the visual inspection data is not fully explained by the available input data. Based on the available input data, MPCA was applied to the welding batch runs in order to classify runs based on their similarity. The classification of the weld runs obtained from MPCA was similar to that obtained from HCA based on input data. The PLS models developed between entire input data and visual inspection data demonstrate that even with four PLS components, a reasonably good prediction of the visual inspection data can be obtained. This implies that the PLS models developed can be used in monitoring the welding runs and to predict the final outcome of the batch.

This work is an initial outcome of the data analysis work undertaken to develop models that can be used for monitoring and classification of welding runs on a solid steel wire. Further investigations and more experiments are also needed in determining the rest of the inputs that have an impact on the visual inspection data and improving the accuracy of the models, since the models obtained from these methods are reliable only when a fairly large amount of data is available. Models with greater details can be developed if metallography analysis of the welds is done in order to extract more parameters that are relevant to estimating the quality of the weld. It should also be noted that separate models need to be developed for different wire materials and wire types; however, the methodology developed here does not need to be modified.

ACKNOWLEDGEMENTS

The authors would like to gratefully acknowledge the financial support received from the Natural Sciences and Engineering Research council of Canada.

REFERENCES

S. Adolfsson, A. Bahrami, G. Bolmsjö, and I. Claesson. On-line quality monitoring in short-circuit gas metal arc welding. *Welding Journal*, 78(2):59S–73S, 1999.

ASME. *ASME boiler and pressure vessel code: an international code. I- Rules for construction of power boiler*. American Society of Mechanical Engineers, New York, 2010.

P.B. García-Allende, J. Mirapeix, O.M. Conde, A. Cobo, and J.M. López-Higuera. Spectral processing technique based on feature selection and artificial neural networks for arc-welding quality monitoring. *NDT and E International*, 42(1):56–63, 2009.

A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Comput. Surv.*, 31(3):264–323, 1999.

J. V. Kresta, J. F. MacGregor, and T. E. Marlin. Multivariate statistical monitoring of process operating performance. *Can. J. Chem. Eng.*, 69(1):35–47, 1991.

C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*. Cambridge University Press, New York, 2008.

P. Nomikos and J. F. MacGregor. Monitoring batch processes using multiway principal component analysis. *AIChE J.*, 40(8):1361–1375, 1994.

T. Siewert, I. Samardžić, Z. Kolumbić, and Š. Klarić. On-line monitoring system - an application for monitoring key welding parameters of different welding processes. *Tehnicki Vjesnik*, 15(2):9–18, 2008.