Chemometrics and Intelligent Laboratory Systems 196 (2020) 103894

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

Variable contribution identification and visualization in multivariate statistical process monitoring

R.F. Rossouw [a,b,*], R.L.J. Coetzer [a], N.J. Le Roux [b]

[a] Data Science, Central Digital Office, Sasol, Private Bag 1, Sasolburg, 1947, South Africa
[b] Department of Statistics and Actuarial Science, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa

ARTICLE INFO

Keywords:
Biplots
Process monitoring
Variable contribution

ABSTRACT

Multivariate statistical process monitoring (MSPM) has received book-length treatments and wide spread application in industry. In MSPM, multivariate data analysis techniques such as principal component analysis (PCA) are commonly employed to project the (possibly many) process variables onto a lower dimensional space where they are jointly monitored given a historical or specified reference set that is within statistical control. In this paper, PCA and biplots are employed together in an innovative way to develop an efficient multivariate process monitoring methodology for variable contribution identification and visualization. The methodology is applied to a commercial coal gasification production facility with multiple parallel production processes. More specifically, it is shown how the methodology is used to specify the optimal principal component combinations and biplot axes for visualization and interpretation of process performance, and for the identification of the critical variables responsible for performance deviations, which yielded direct benefits for the commercial production facility.

1. Introduction

The Sasol Coal Gasification (SCG) plant in South Africa is a highly complex chemical facility. The chemical facility gasifies bituminous coal to synthesis gas, which is converted to fuels and chemicals via the company's suite of hydrocarbon processes. Specifically, the facility gasifies no less than 40 million tons of coal per year and has a combined production of approximately 4.6 × $10^6 m^3 nh^{-1}$ (normal cubic meters per hour) dry raw gas (RG). The continuous improvement of the coal gasification plant is of critical importance to the company to ensure a stable supply of high quality gas to the downstream units.

In order to maintain optimum product yields and to sustain throughput, an efficient multivariate process monitoring methodology is required for all the reactors. In this paper the reactors or gasifiers will be referred to as production processes. The coal gasification facility consists of two separate plants known as Gasification West and Gasification East. There are between 40 and 42 production processes on each plant which are organized into four production trains i.e., each train contains 10 or 11 production processes. A layout of the gasification plants is presented in Fig. 1. Each production process is equipped with a dedicated instrumentation which records online performance data on each process.

Since these processes are very complex, there may be many process variables of interest for monitoring the production performance. However, in this paper, and for the purpose of process monitoring of the production processes, process variables of interest include production volume, utility consumption and other stability measures on the reactors. Table 1 provides a list together with a high level description of the types of process variables considered. These are the variables that were identified together with the subject matter experts as the most relevant for monitoring the performance of the production processes. Note that the variables are indicated by neutral labels, due to confidentiality restrictions imposed by the company under consideration. In addition, the production volume variable was removed from this study due to measurement inconsistencies in the data. This posed no constraint as the main driver of this study was stability improvement. The variables were grouped into utility and stability types. Stability variables provide information on process stability for instance temperature, whereas utility variables measure some indirect input to the process for instance steam.

The commonly accepted approach of most of the multivariate process monitoring procedures is to first specify a historical reference set that is within statistical control. Since the variables may be many, multivariate data analysis techniques such as principal component analysis (PCA) are employed to project process variables onto a lower dimensional space where they can be jointly monitored given the in-control reference set.

[*] Corresponding author. Data Science, Central Digital Office, Sasol, Private Bag 1, Sasolburg, 1947, South Africa.
E-mail address: ruan.rossouw@sasol.com (R.F. Rossouw).

https://doi.org/10.1016/j.chemolab.2019.103894
Received 30 August 2019; Received in revised form 6 November 2019; Accepted 11 November 2019
Available online 15 November 2019

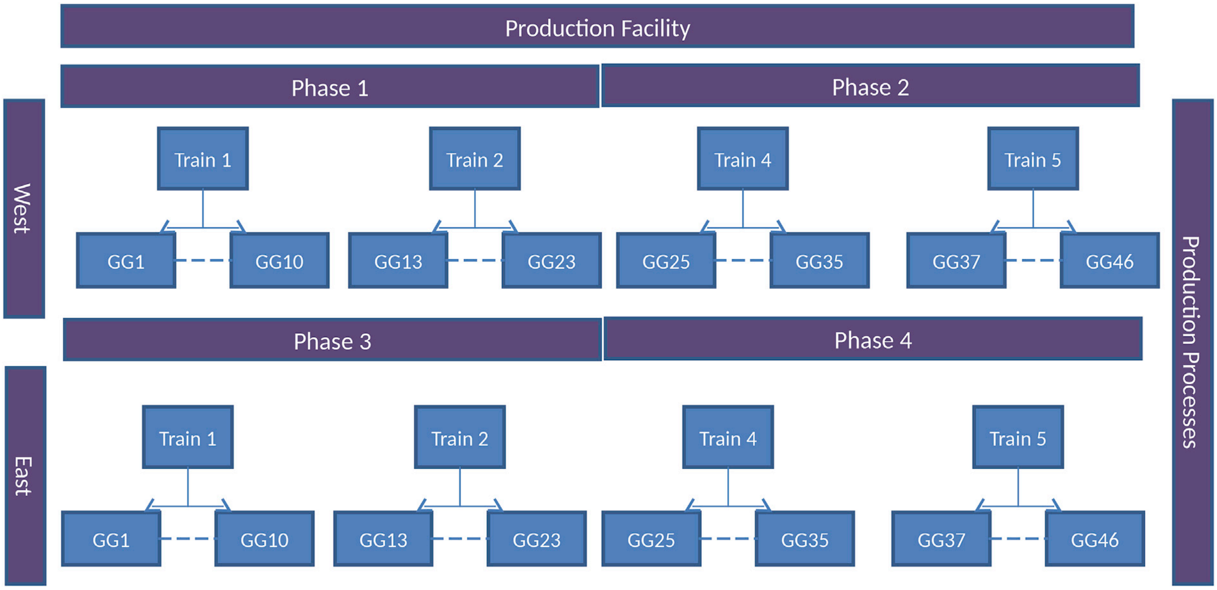


Fig. 1. Sasol coal gasification (SCG) overview.

Table 1
Variable types.

Variable	Variable Type
U1	Utility
U2	Utility
U3	Utility
S1	Stability
S2	Stability
S3	Stability
S4	Stability
S5	Stability
S6	Stability
S7	Stability
U4	Utility

The selection of an optimal reference set was discussed in detail in Coetzer et al. (2014)[10] for the coal gasification production facility. In this paper, the optimal reference set is considered as given, and the objective is to develop and present an efficient multivariate process monitoring methodology using the eleven variables specified for real time performance monitoring of the production processes. From Coetzer et al. (2014)[10], the optimal reference set specified for the Eastern Factory is utilized for the multivariate process monitoring of the production processes. The process variable data are captured in real time on a distributed control system (DCS), and may be captured at different time intervals for the different variables. For the current application, the real time monitoring is performed on 15 min aggregated data.

Given the optimal reference set, the PCA biplot methodology as described by Ref. [1] is applied to the multivariate statistical process monitoring of each production process. As all the production processes on one train receive the same coal feed, and are managed by the same operator, it is expected that the performance should be similar. Therefore, the use of one optimal reference set for all the production processes is sensible since the performance is monitored against the same reference set for all the processes. [2,3] discussed numerous advantages of biplots for process monitoring. Specifically, the biplot provides the ability to detect deviations from expected performance in real time, as well as which variables are mostly responsible for the deviation. Therefore, the implementation of monitoring biplots for real time performance monitoring provides the engineer with information on the important process variables for navigating the process back to expected or target performance quicker. In this paper, the common statistical methods of PCA and biplots are used together in an innovative way to develop and present an efficient multivariate process monitoring methodology for variable contribution identification and visualization, applied to a commercial chemical process with multiple parallel production processes.

Specifically, the methodology presented in this paper addresses the following key aspects:

- The selection of the optimal principal component combinations to utilise for the biplot scaffolding or visualization.
- The optimal axes combination to include for each principal component combination for purposes of visualization.
- The identification of critical variables responsible for process performance deviations for each principal component combination presented.

The paper is outlined as follows. In the next section an introduction to the theory of the PCA biplot as well as some measures of the predictive power of the biplots is presented. In Section 3 the application of the PCA biplot monitoring methodology to the coal gasification process is discussed. In Sections 4 and 5 the results and conclusions are presented respectively.

2. Introduction to PCA biplot theory

2.1. Preliminaries

The singular value decomposition (SVD) plays a central role in multivariate statistics [1,4–6]. The SVD of a $n \times p$ rank r matrix $\mathbf{X}$ with $n \geq p$ is defined as

$$\mathbf{X} = \mathbf{U}_{n\times r}\mathbf{\Sigma}_{r\times r}\mathbf{V}^T_{r\times p} \tag{1}$$

where $\mathbf{U}$ is a $n \times r$ orthonormal matrix and $\mathbf{V}$ is a $p \times r$ orthonormal matrix, $r \leq p$. The matrix $\mathbf{\Sigma}$ is an $r \times r$ diagonal matrix with diagonal elements the non-zero singular values of $\mathbf{X}$. The singular values are non-negative, and specified in decreasing order i.e., $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$.

In this paper the optimal approximation of an $n \times p$ data matrix $\mathbf{X}$ of rank r in fewer (say $r^*, r^* < r \leq p$) dimensions is of interest. Let $\widehat{\mathbf{X}}_{r^*}$ of rank r^* be the approximation of $\mathbf{X}$ in r^* dimensions. Then, $\widehat{\mathbf{X}}_{r^*}$ is obtained by minimizing of

$$\text{trace}\left[(\boldsymbol{X}-\widehat{\boldsymbol{X}}_{r^*})^T\left(\boldsymbol{X}-\widehat{\boldsymbol{X}}_{r^*}\right)\right] = \parallel \left(\boldsymbol{X}-\widehat{\boldsymbol{X}}_{r^*}\right)\parallel^2 \tag{2}$$

for given r^* [1].

The solution to this least-squares problem is given by the Eckart-Young theorem [7]. According to this theorem, the least squares problem is minimised in r^* dimensions by the r^* dimensional approximation for $\boldsymbol{X}$

$$\widehat{\boldsymbol{X}}_{r^*} = \boldsymbol{U}_{r^*}\boldsymbol{\Sigma}_{r^*}\boldsymbol{V}_{r^*}^T \tag{3}$$

where $\boldsymbol{U}_{r^*}$ is a $n \times r^*$ orthonormal matrix with r^* the first r^* columns of $\boldsymbol{U}$, $\boldsymbol{V}_{r^*}$ is a $p \times r^*$ orthonormal matrix (for $r_* < p$) with r^* the first r^* columns of $\boldsymbol{V}$, and $\boldsymbol{\Sigma}_{r^*}$ is a $r^* \times r^*$ diagonal matrix. The quality of the lower dimensional representation is defined as:

$$\frac{\sum_{i=1}^{r^*}\sigma_i^2}{\sum_{i=1}^{p}\sigma_i^2} \tag{4}$$

where σ_i is the ith singular value of $\boldsymbol{X}$. The rows of the r^* dimensional approximation $\widehat{\boldsymbol{X}}_{r^*}$ of $\boldsymbol{X}$ are given by $\boldsymbol{U}_{r^*}\boldsymbol{\Sigma}_{r^*} = \boldsymbol{X}\boldsymbol{V}_{r^*}$. In the PCA literature $\boldsymbol{X}\boldsymbol{V}_{r^*}$ is known as the scores of the PCA. The directions of the PCA biplot axes are given by the rows of $\boldsymbol{V}_{r^*}$, also known as the loadings. The r^* dimensional approximation of a new observational vector $\boldsymbol{x}_0$ is given by $\widehat{\boldsymbol{x}}_{r^*} = \boldsymbol{V}_{r^*}^T\boldsymbol{x}_0$.

[4] extended the biplot methodology to include calibration of the axes. Biplots are discussed in detail in Refs. [1,4,6].

2.2. PCA biplot predictivity

[8] proposed the axis predictivities as a measure of how well each biplot axis is representing the true variable in the biplot space. The axis predictivities are defined as the ratio of the sums of squares ($\widehat{\boldsymbol{X}}_{r^*}^T\widehat{\boldsymbol{X}}_{r^*}$) to the total sums of squares ($\boldsymbol{X}^T\boldsymbol{X}$):

$$\boldsymbol{\Pi} = \text{diag}\left(\widehat{\boldsymbol{X}}_{r^*}^T\widehat{\boldsymbol{X}}_{r^*}\right)\left[\text{diag}(\boldsymbol{X}^T\boldsymbol{X})\right]^{-1} \tag{5}$$

$$= \text{diag}\left(\boldsymbol{V}_{r^*}\boldsymbol{\Sigma}_{r^*}^2\boldsymbol{V}_{r^*}^T\right)\left[\text{diag}\left(\boldsymbol{V}\boldsymbol{\Sigma}^2\boldsymbol{V}^T\right)\right]^{-1}. \tag{6}$$

If $\boldsymbol{X}$ is standardized so that each variable has zero mean and unit variance then (5) and (6) simplify to give

$$\boldsymbol{\Pi} = \text{diag}\left(\widehat{\boldsymbol{X}}_{r^*}^T\widehat{\boldsymbol{X}}_{r^*}\right) \tag{7}$$

$$= \text{diag}\left(\boldsymbol{V}_{r^*}\boldsymbol{\Sigma}_{r^*}^2\boldsymbol{V}_{r^*}^T\right). \tag{8}$$

[9] proposed the mean standard prediction error (mspe) criterion as a measure of predictive power of the biplot axes, given by:

$$\text{mspe} = \frac{1^T\sum_{i=1}^{p}|\boldsymbol{x}_{(i)} - \widehat{\boldsymbol{x}}_{(i)}|}{n} \tag{9}$$

where 1^T is a $1 \times \boldsymbol{n}$ vector of 1's, and $\boldsymbol{x}_{(i)}$ is the ith column of $\boldsymbol{X}$ and similarly $\widehat{\boldsymbol{x}}_{(i)}$ is the ith column of $\widehat{\boldsymbol{X}}_{r^*}$. The mspe criterion assumes the initial $\boldsymbol{X}$ to be centered and scaled to mean 0 and unit variance for each column. The mspe value for each axis (or variable) is the mean absolute residual value for the actual against fitted values for each variable in $\boldsymbol{X}$. Therefore, a low mspe value represents good predictive power of the axis. Given the $\boldsymbol{X}$ is centered and scaled as above the maximum mspe value will be equal to 1.

Comparing the axis predictivity values with the mspe values the following observations can be made:

- The mspe value for each axis (or variable) is the mean absolute residual value of the actual minus the fitted values for each variable in $\boldsymbol{X}$.
- The axis predictivity value for a variable for a centered and scaled $\boldsymbol{X}$ is the error sums of squares value, or alternatively the corresponding diagonal element of $\boldsymbol{I} - \text{diag}((\boldsymbol{X} - \widehat{\boldsymbol{X}}_{r^*})^T(\boldsymbol{X} - \widehat{\boldsymbol{X}}_{r^*}))$.

Therefore, both the mspe and the axis predictivity are functions of the residual values $\boldsymbol{X} - \widehat{\boldsymbol{X}}_{r^*}$, and it is expected that comparable results will be obtained by applying these two criteria to PCA biplots.

2.3. PCA biplots for multivariate process monitoring

The PCA biplot can be extended to function as a multivariate process monitoring instrument [2,3,10]. In order to keep our notation simple in what follows we do not distinguish between random matrices or vectors and their values, but let the context dictate the meaning. If it can be assumed that the rows of the $n \times p$ reference set ($\mathbf{X}$) are approximately multivariate normally distributed with mean equal to $\overline{\mathbf{x}}$ and covariance matrix $\mathbf{S}$, it follows that the multivariate normal density is constant on surfaces where the Mahalanobis distance $(\mathbf{x} - \overline{\mathbf{x}})^T\mathbf{S}^{-1}(\mathbf{x} - \overline{\mathbf{x}})$ (with $\mathbf{x}^T$ a row of $\mathbf{X}$) is constant. That is, the rows of $\mathbf{X}$ lie on the surface of an ellipsoid centered at $\overline{\mathbf{x}}$.

It can then be shown that

$$(\mathbf{x} - \overline{\mathbf{x}})^T\mathbf{S}^{-1}(\mathbf{x} - \overline{\mathbf{x}})\tilde{}\chi_p^2, \tag{10}$$

where χ_p^2 denotes the chi-square distribution with p degrees of freedom, which is closely related to the T^2-statistic. The $(1-\alpha)$% concentration ellipse is defined by all observations that satisfy

$$(\mathbf{x} - \overline{\mathbf{x}})^T\mathbf{S}^{-1}(\mathbf{x} - \overline{\mathbf{x}}) \le \chi^2(\alpha) \tag{11}$$

[1]. Any observation that does not satisfy the above inequality can be flagged for further inspection. The T^2-statistic and its derivation is discussed in detail in Refs. [11,12].

2.4. Variable contribution identification

[11] p43 define an approach for the calculation of the contribution of each variable to the overall T^2-value:

1. Calculate the normalised scores for all r^* scores

$$\left(\frac{y_i}{\sigma_i}\right)^2 \tag{12}$$

where y_i is the i^{th} value of the projection $\boldsymbol{y} = \boldsymbol{V}_{r^*}^T\boldsymbol{x}$. Therefore, y_i is the score of the observation $\boldsymbol{x}$ projected on the i^{th} loading vector, and σ_i is the corresponding singular value.

Determine the a scores responsible for the out of control status by comparing each $\left(\frac{y_i}{\sigma_i}\right)^2$ to $(T_\alpha^2)^{\frac{1}{r^*}}$ and only retain the scores where

$$\left(\frac{y_i}{\sigma_i}\right)^2 > (T_\alpha^2)^{\frac{1}{r^*}}. \tag{13}$$

2. Calculate the variable contribution for each variable x_j to the out of control scores t_i

$$c_{ij} = \frac{y_i}{\sigma_i^2}v_{ij}\left(x_j - \mu_j\right) \tag{14}$$

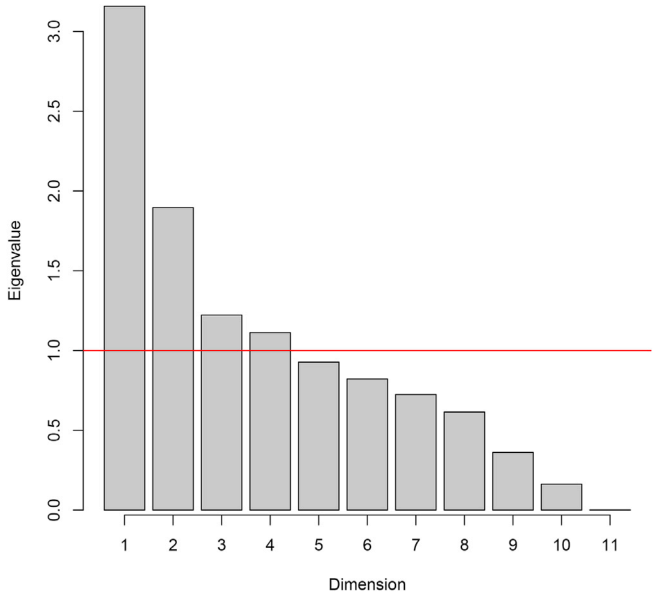


Fig. 2. Scree plot.

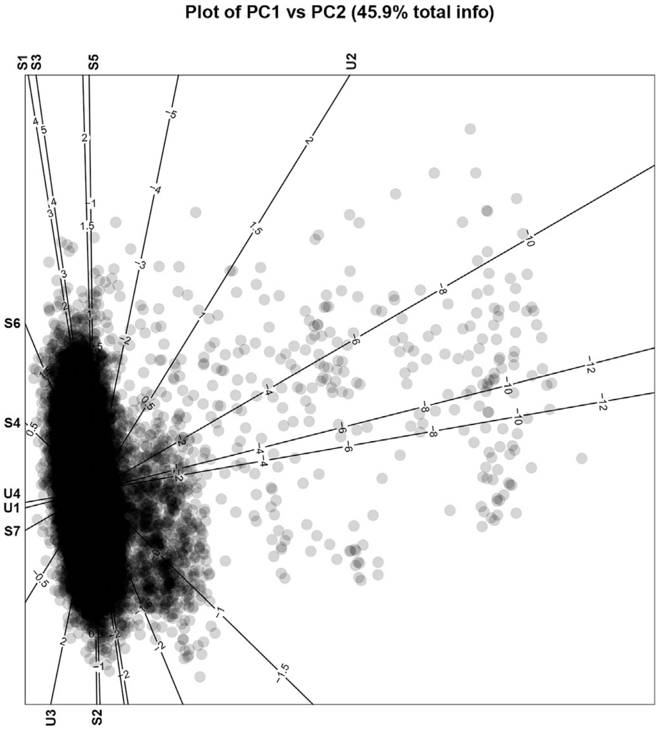


Fig. 3. PC1 vs PC2 - All Axes Included.

where v_{ij} is the (i,j)-th element of $\boldsymbol{V}_{r^*}$ and c_{ij} is the (i,j)-th element of $\boldsymbol{C}$ where $\boldsymbol{C}$ is an $r^* \times p$ matrix initialized with zero values. Therefore, the rows of $\boldsymbol{C}$ for in-control t_i will default to zero values.

3. Now, set any $c_{ij} < 0$ equal to 0.
4. Calculate the total variable contribution for the j-th variable x_j

$$t_j = \sum_{i=1}^{r^*} c_{ij} \tag{15}$$

where $\boldsymbol{t}$ is a vector of length p.

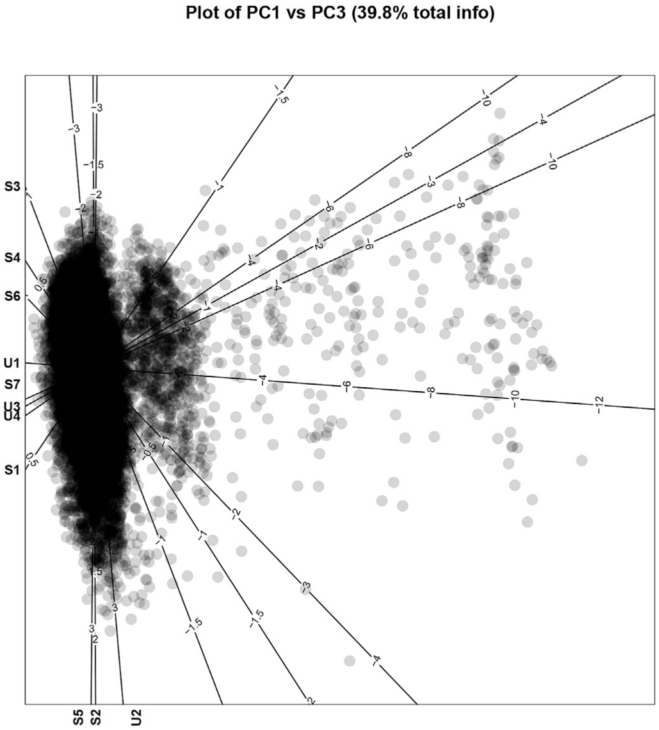


Fig. 4. PC1 vs PC3 - All Axes Included.

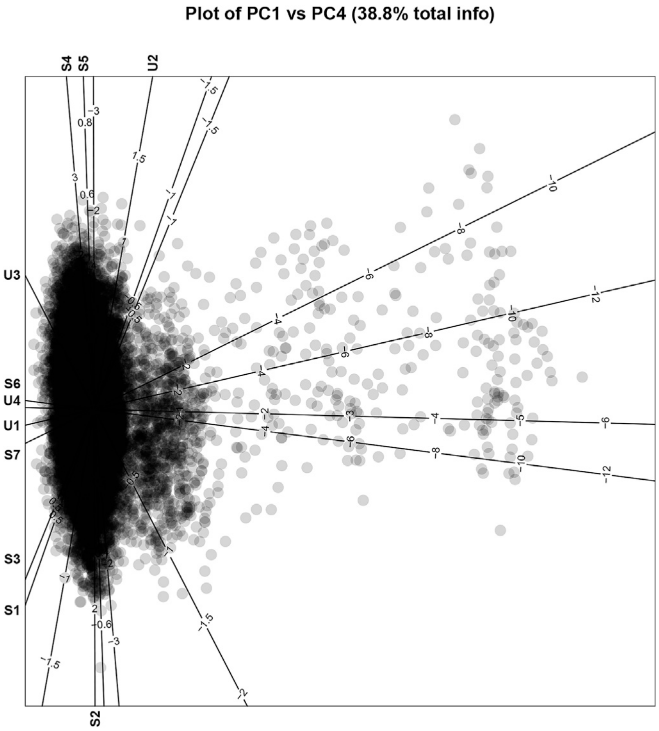


Fig. 5. PC1 vs PC4 - All Axes Included.

3. Application of PCA to the coal gasification process

3.1. Number of principal components to include

Various strategies are discussed in the literature to determine the number of principal components to retain [5,9]. In this paper the components were excluded for which the eigenvalues of the correlation

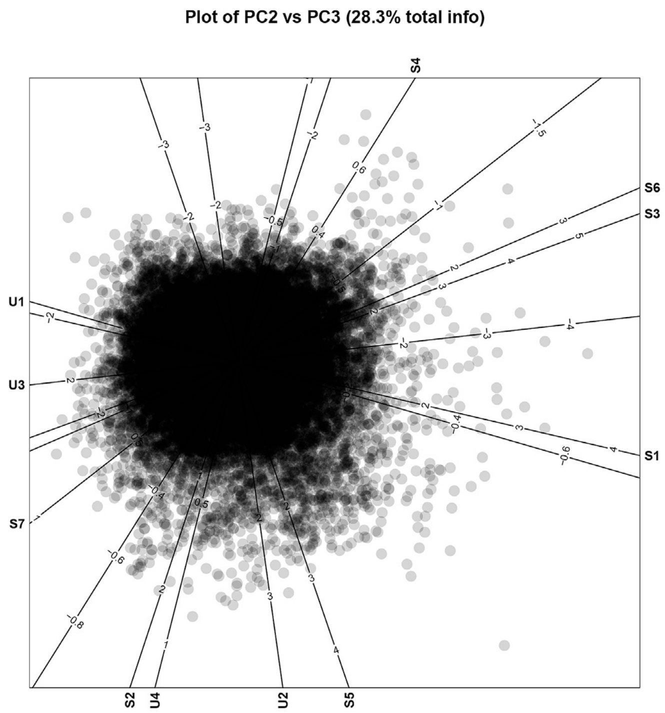


Fig. 6. PC2 vs PC3 - All Axes Included.

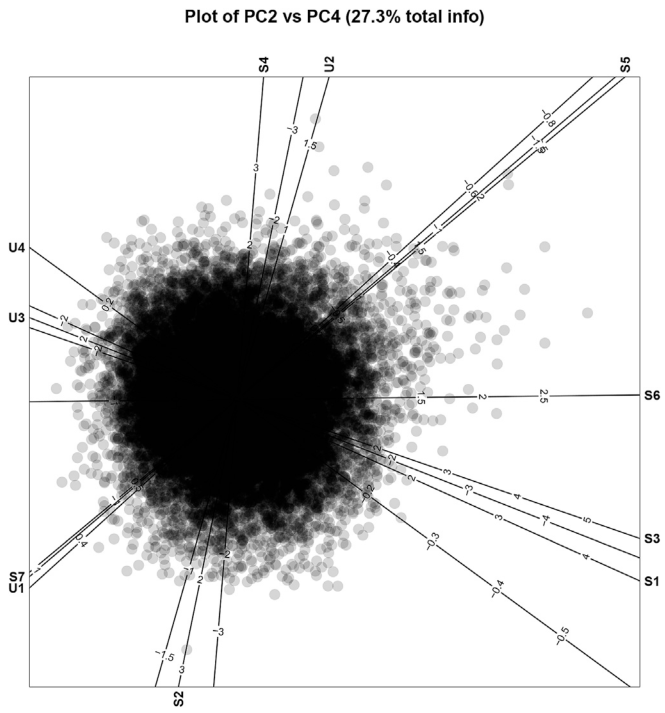


Fig. 7. PC2 vs PC4 - All Axes Included.

matrix are less than one. This method was first proposed by Ref. [13]; and is a special case of the more general methodology to exclude all the eigenvalues that are less than the average of all the eigenvalues i.e., if λ_i is defined as the i^{th} eigenvalue of $\boldsymbol{S}$ the average eigenvalue will be $\sum_i^p \frac{\lambda_i}{p}$. Note that the $\sum_i^p \lambda_i = \text{trace}(\boldsymbol{S})$ and therefore the average of the eigenvalues is equal to the average variance [14].

For the gasification facility discussed in Section 1, the eigenvalues of the correlation matrix of the reference set were calculated. The results are shown in Fig. 2. According to the specified criterion, four principal components should be retained.

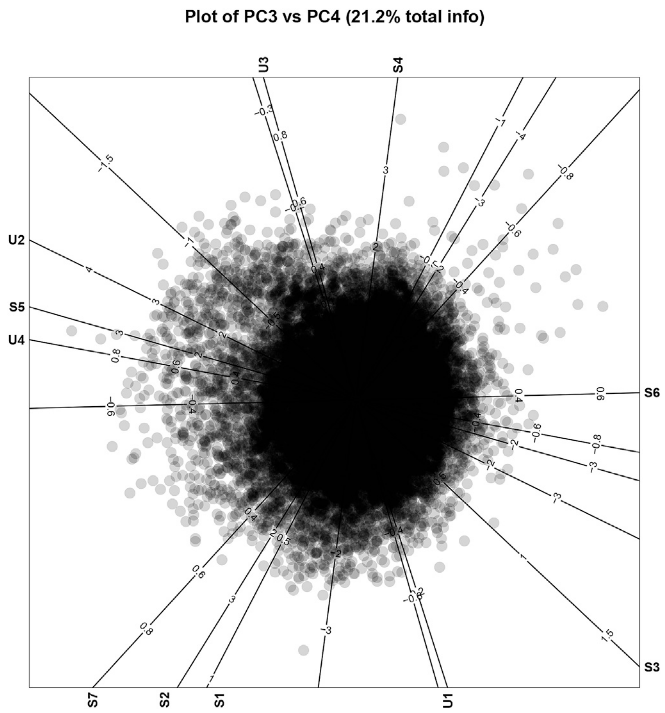


Fig. 8. PC3 vs PC4 - All Axes Included.

Table 2
PCA biplot quality for Figs. 3 to 8.

	PC (x-axis)	PC (y-axis)	Quality (%)
Fig. 3	1	2	45.94
Fig. 4	1	3	39.81
Fig. 5	1	4	38.81
Fig. 6	2	3	28.35
Fig. 7	2	4	27.34
Fig. 8	3	4	21.22

3.2. *Axis predictivity and interpretation*

Figs. 3–8 depict the PCA biplots of the reference set of the Eastern factory as discussed in Section 1. Note that for all the biplots generated in this section the smaller numbered principal component will be represented on the x axis and the higher numbered principal component on the y axis. For example, in Fig. 3 PC1 is represented along the x axis and PC2 along the y axis. The PCA biplot qualities (4) are provided in Table 2 for the different combinations. One observation from the figures is that there is one large group of data, which represents the majority of the operating conditions in the process variables considered, and a smaller group of data which was formed to the right of the major group, as well as a few additional operating conditions along the first principal component.

One explanation may be that these points are outliers which may be omitted from the reference set. However, from closer inspection, there is process information in this additional scatter. Specifically from Fig. 3, in the two dimensional biplot visualization, the operating conditions to the right are projected lower onto variables U1, U4, and S7, which indicates that the reactor was operated at lower load. The reactor load is usually reduced when the reactor becomes unstable due to either feed quality or mechanical problems. Therefore, from a process engineering perspective, these reactor conditions are sensible and expected during instability. This is one example where results from a purely data driven approach must be interpreted together with process knowledge in order to make the correct

Table 3
Axis predictivity.

	PC1	PC2	U1	U2	U3	S1	S2	S3	S4	S5	S6	S7	U4
1	1	2	0.95	0.07	0.59	0.41	0.04	0.67	0.08	0.11	0.45	0.77	0.91
2	1	3	0.94	0.53	0.12	0.06	0.18	0.11	0.08	0.44	0.25	0.74	0.94
3	1	4	0.94	0.14	0.14	0.08	0.36	0.09	0.61	0.03	0.23	0.74	0.91
4	2	3	0.01	0.50	0.48	0.36	0.22	0.64	0.02	0.55	0.24	0.05	0.04
5	2	4	0.01	0.11	0.50	0.38	0.40	0.62	0.56	0.14	0.22	0.05	0.01
6	3	4	0.00	0.57	0.03	0.03	0.54	0.06	0.56	0.47	0.02	0.02	0.03

Table 4
Axis mspe.

	PC1	PC2	U1	U2	U3	S1	S2	S3	S4	S5	S6	S7	U4
1	1	2	0.19	0.70	0.50	0.61	0.79	0.45	0.74	0.72	0.58	0.31	0.24
2	1	3	0.20	0.54	0.77	0.77	0.72	0.75	0.73	0.57	0.67	0.31	0.20
3	1	4	0.20	0.69	0.76	0.76	0.62	0.75	0.48	0.75	0.69	0.32	0.24
4	2	3	0.50	0.54	0.57	0.62	0.70	0.45	0.75	0.51	0.66	0.32	0.48
5	2	4	0.50	0.70	0.55	0.61	0.60	0.47	0.49	0.71	0.67	0.32	0.48
6	3	4	0.49	0.50	0.79	0.78	0.52	0.77	0.49	0.56	0.77	0.28	0.47

Table 5
PCA loadings for first four principal components.

	PC1	PC2	PC3	PC4
U1	-0.54	-0.07	0.02	-0.05
U2	0.12	0.11	-0.63	0.28
U3	-0.19	-0.50	-0.05	0.15
S1	-0.13	0.43	-0.08	-0.15
S2	0.00	-0.15	-0.38	-0.57
S3	-0.15	0.56	0.17	-0.15
S4	-0.14	0.07	0.10	0.70
S5	-0.01	0.25	-0.60	0.16
S6	-0.27	0.34	0.12	0.00
S7	-0.48	-0.15	-0.10	-0.10
U4	-0.54	-0.05	-0.16	0.03

deductions.

Although it is possible to interpret the biplots with all the axes included it will be demonstrated in the remainder of this section that:

- Using a predictivity measure to reduce the number of axes on each plot guides the user in interpreting the data.
- The predictivity measures contain valuable information on the variable effects and correlations.
- Interpreting additional principal component combinations instead of only using the traditional plot of first and second principal components, yields additional value to the interpretation of the data.

In this paper, both the axis predictivity (5) from Ref. [8] and the mspe criteria (9) from Ref. [9] were implemented. The results are summarized in Table 3 and Table 4. For a first approximation a criterion of axis predictivity above 0.5 and mspe below 0.5 was implemented. The results are highlighted in Tables 3 and 4.

First, the results from Tables 3 and 4 are compared. It is clear that although the axis predictivity and mspe results mostly agree there are some noticeable differences. These are most prominent for variables U1, S7 and U4 where for principal component combinations 2 × 3, 2× 4 and 3 × 4 the axis predictivities for U1, S7 and U4 are close to zero, while the mspe values are still in the acceptance range. The remaining differences are mostly on borderline cases, for example U2 for principal component combination 1 × 2 is accepted by the predictivity criterion, but rejected by the mspe criterion. The mspe value is however 0.54 which is very close to the (arbitrary) cutoff value of 0.5. To aid in the interpretation of the results the PCA loadings for the first four principal components (the first four columns of $\boldsymbol{V}_{r^*}$) are provided in Table 5. Loadings with an absolute value larger than 0.4 are highlighted. As the data were mean centered and unit scaled the relative sizes of the loadings can be interpreted. In addition, the origin of the PCA biplot will be at $x = 0, y = 0$, and therefore positive loadings for a variable will lead to predicted values on the biplot axes increasing towards the edges of the top right quadrant of the biplot for the specific variable. Closer investigation of Tables 3 and 5 leads to the following observations:

1. The axis predictivities for variables U1, S7 and U4 are higher in all combinations where principal component one is present. This observation is confirmed by the loadings for principal component one, where all three these variables have relatively large loadings compared to the other variables. U1, S7 and U4 are variables that are all directly related to reactor load. All the loadings are negative, and therefore in the same direction. It can therefore be concluded that principal component one predominantly explains the effect of the reactor load related variables on the performance of the reactor.
2. The axis predictivities for U3, S1 and S3 are higher in all combinations where principal component two is present. This observation is confirmed by the loadings for principal component two. Variable U3 has a relatively large positive loading, and variables S1 and S3 have relatively large negative loadings. Therefore, it can be concluded that variable U3 is negatively correlated with variables S1 and S3. Note that in PCA the signs of the components are arbitrary, but the relative signs of the loadings on a component can be interpreted. Variables S1 and S3 are two temperature variables inside the reactor. Variable U3 measures the reactor cooling agent level. Therefore, the negative correlation between the cooling agent level and the temperatures are sensible and expected from a process engineering perspective. From the statistical indices and the biplot, it can therefore be concluded that principal component two predominantly represents the relationship between the temperature inside the reactor and the amount of cooling agent applied.
3. The axis predictivities for U2 and S5 are higher in all combinations where principal component three is present. This observation is confirmed by the loadings for principal component three. Both

Table 6
Table with new data.

	U1	U2	U3	S1	S2	S3	S4	S5	S6	S7	U4
1	−0.30	1.71	0.64	−0.61	1.08	−2.08	−0.45	0.48	−0.88	0.10	0.21
2	−0.40	1.71	0.89	0.18	1.46	−2.06	−1.54	0.27	−0.83	0.08	0.11
3	−0.34	1.71	0.79	0.64	0.77	−2.01	−1.81	0.29	−0.88	−0.02	0.18
4	−0.22	1.71	0.71	−0.09	−0.04	−1.98	0.13	0.46	−0.95	0.03	0.29
5	−0.31	1.71	0.62	−0.97	−0.27	−1.89	−0.62	0.47	−1.04	0.13	0.20
6	−0.46	1.71	0.54	−1.11	−0.23	−1.95	−0.31	0.32	−1.03	0.22	0.03
7	−0.47	1.71	0.68	−1.07	−0.91	−2.10	−0.15	0.18	−1.12	0.07	0.03
8	−0.51	1.71	0.72	−1.25	−0.04	−2.12	−0.64	0.29	−1.11	0.26	−0.01
9	−0.58	1.71	0.46	−1.22	−0.62	−2.26	−0.66	0.21	−1.10	0.15	−0.08
10	−0.20	0.86	1.27	−1.21	−1.10	−2.91	−0.76	−0.03	−1.13	0.13	0.09
11	−0.09	0.70	1.23	−0.83	−1.01	−2.88	−0.31	−0.23	−1.12	0.19	0.14
12	−0.07	0.70	1.78	−1.48	−1.66	−2.62	−0.19	−0.38	−1.05	0.19	0.15
13	−0.12	0.70	0.74	−1.01	−1.57	−2.05	0.42	−0.19	−0.90	0.07	0.11
14	0.04	0.70	0.55	−0.36	−1.81	−1.81	0.97	0.15	−0.73	0.12	0.26
15	−0.28	0.70	0.32	−1.43	−1.35	−1.62	0.87	0.35	−1.03	0.13	0.14
16	−0.24	0.70	0.17	−0.94	−0.34	−1.35	−0.16	0.03	−1.09	0.12	0.13

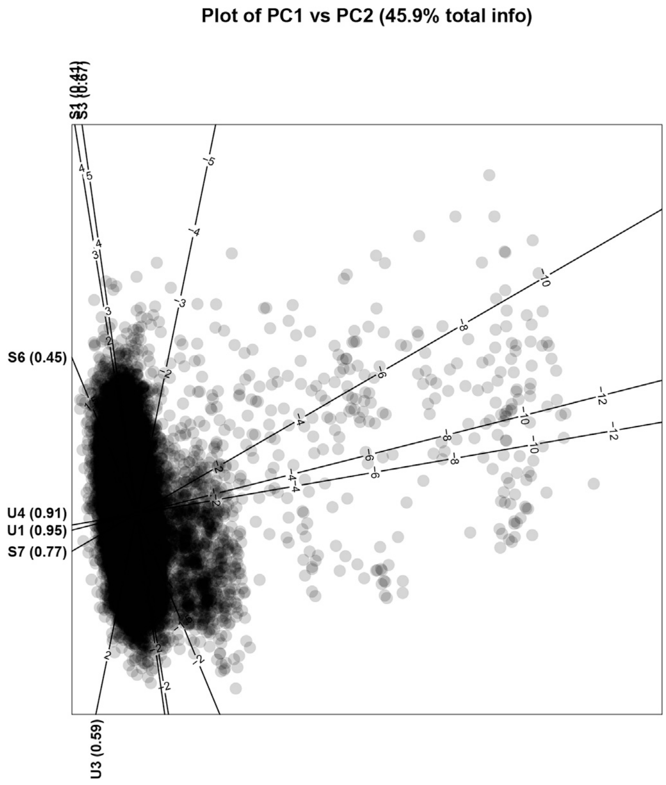


Fig. 9. PC1 vs PC2 - Axes included with Axis Predictivity Higher than 0.35.

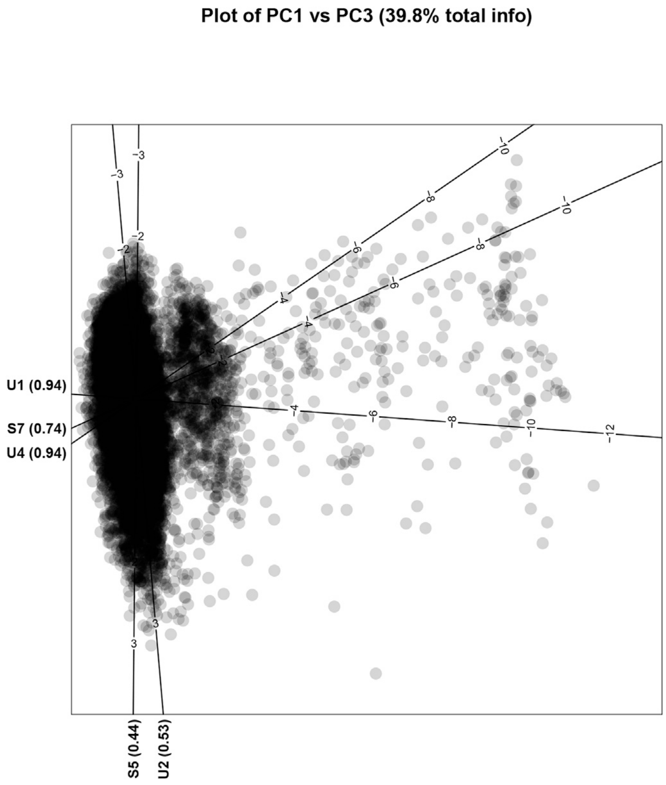


Fig. 10. PC1 vs PC3 - Axes included with Axis Predictivity Higher than 0.35.

variables U2 and S5 have relatively large negative loadings. Variable U2 is related to reactor load, and variable S5 is a reaction side product.

4. The axis predictivities for S2 and S4 are higher in all combinations where principal component four is present. This observation is confirmed by the loadings for principal component four. Variable S2 has a negative loading and S4 has a positive loading, and they are thus negatively correlated. Variable S2 is a temperature variable inside the reactor and variable S4 is a variable related to the difficulty of the ash removal from the reactor. Therefore, principal component four represents a temperature measurement inside in the reactor, with the accompanying changes in ash properties.

From above, it is clear that the axis predictivity values are linked to the underlying structure in the data, and are correlated with the PCA loadings (as follows from (8)). In addition, the results could be validated from a process engineering perspective.

It can therefore be concluded that the structure in the principal components successfully captures the underlying dependencies and effects of the variables in the gasification process. Therefore, the reference set selected is appropriate for monitoring the performance of the process. The axis predictivity values aided in highlighting the structure, and it is therefore suggested that predictivity is not only utilized to eliminate axes from the display, but also as an aid to analyze the underlying structure of the data. From the above analysis it was decided to utilise the axis predictivity values to choose the appropriate axes for the monitoring biplots. In addition, a cutoff value of 0.35 will be utilized to include all the relevant axes for each principal component combination in the study.

4. Results and discussion

To demonstrate the implementation of monitoring biplots an example

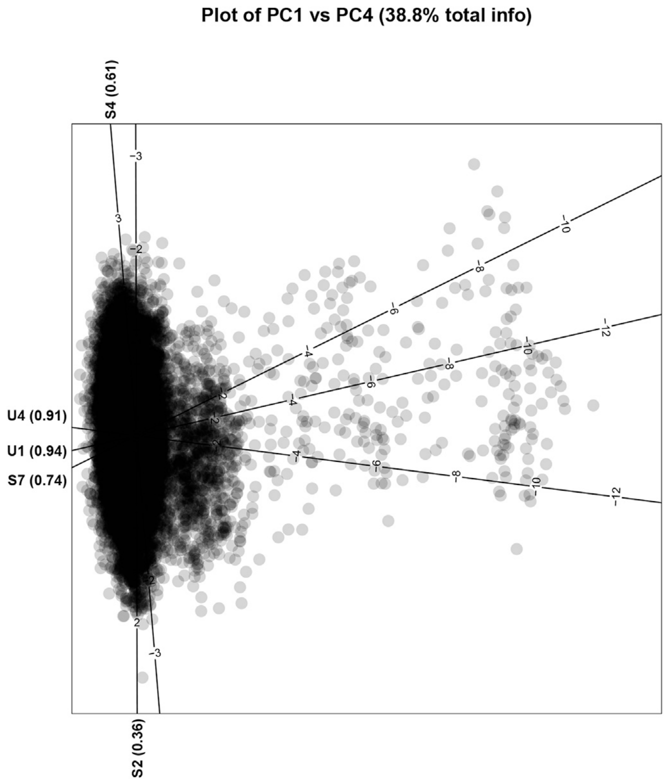


Fig. 11. PC1 vs PC4 - Axes included with Axis Predictivity Higher than 0.35.

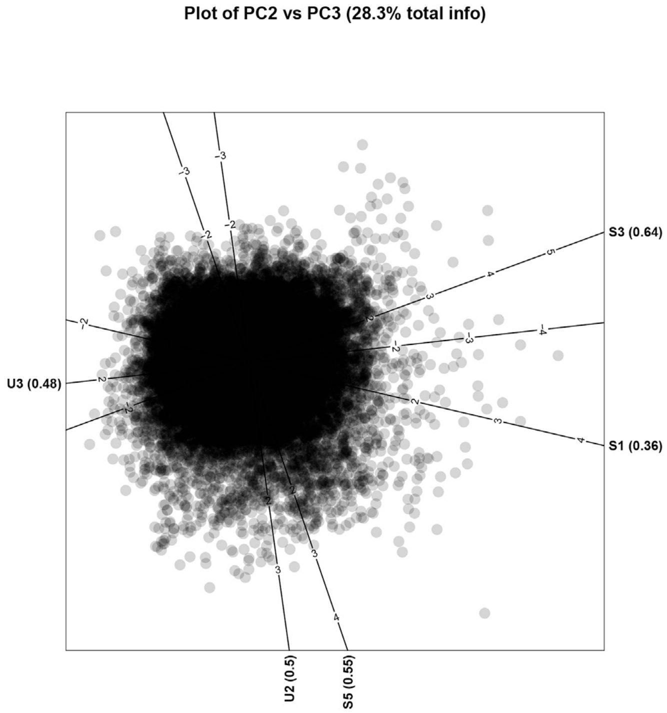


Fig. 12. PC2 vs PC3 - Axes included with Axis Predictivity Higher than 0.35.

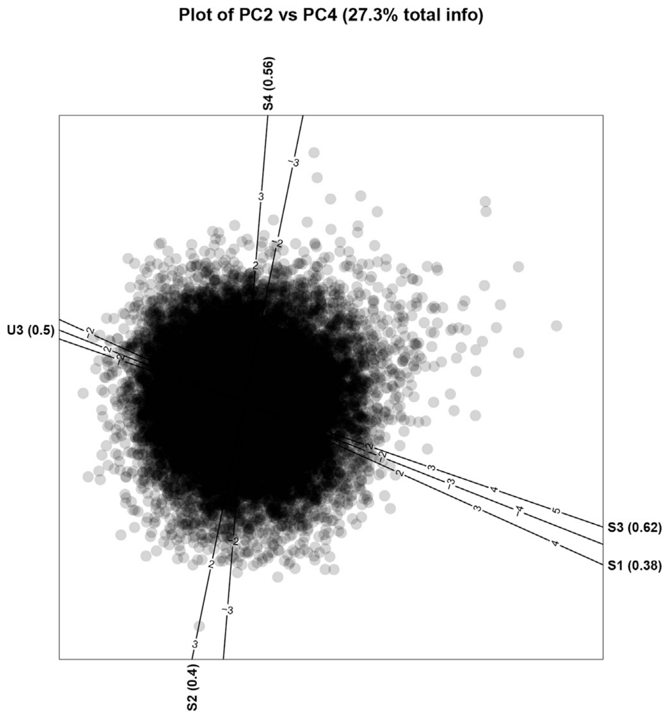


Fig. 13. PC2 vs PC4 - Axes included with Axis Predictivity Higher than 0.35.

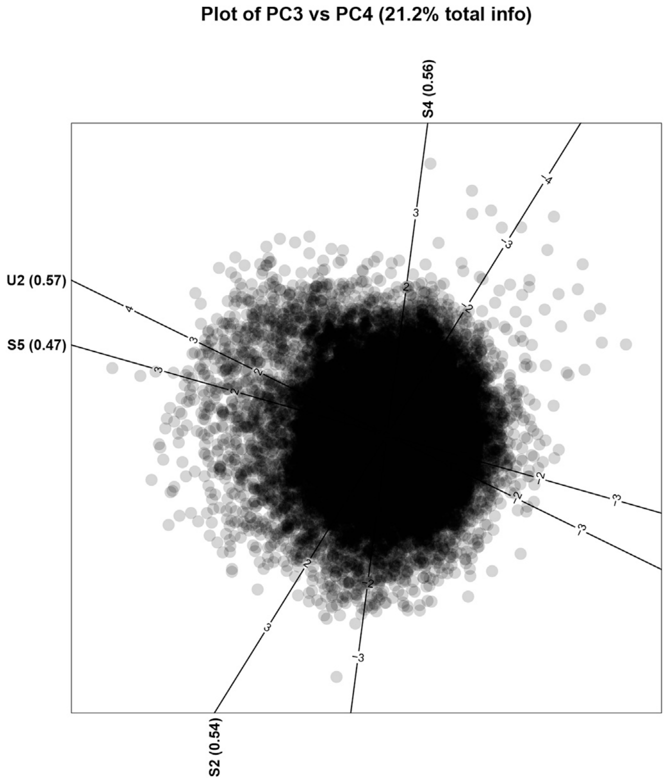


Fig. 14. PC3 vs PC4 - Axes included with Axis Predictivity Higher than 0.35.

of a four hour period in 15 min intervals depicted in Table 6 will be discussed. The row numbers indicate the sequence i.e., row one occurred first and row 16 occurred last. Due to confidentiality constraints the data set was centered and scaled using the column means and column standard deviations of the reference set. Note that the mean of the reference set will therefore be 0 for all the columns. It is therefore easy to compare the values in Table 6 to the reference set as any value above 0 is higher than the average value for the reference set, and any value below 0 is lower than the average value for the reference set. Additionally, the standard deviation of the reference set is scaled to 1, which can also be used to give context to the values in Table 6.

Fig. 19 to Fig. 24 depict the data in Table 6 projected on the

Table 7
Axis predictivity values with values above 0.35 highlighted.

	PC1	PC2	U1	U2	U3	S1	S2	S3	S4	S5	S6	S7	U4
1	1	2	0.95	0.07	0.59	0.41	0.04	0.67	0.08	0.11	0.45	0.77	0.91
2	1	3	0.94	0.53	0.12	0.06	0.18	0.11	0.08	0.44	0.25	0.74	0.94
3	1	4	0.94	0.14	0.14	0.08	0.36	0.09	0.61	0.03	0.23	0.74	0.91
4	2	3	0.01	0.50	0.48	0.36	0.22	0.64	0.02	0.55	0.24	0.05	0.04
5	2	4	0.01	0.11	0.50	0.38	0.40	0.62	0.56	0.14	0.22	0.05	0.01
6	3	4	0.00	0.57	0.03	0.03	0.54	0.06	0.56	0.47	0.02	0.02	0.03

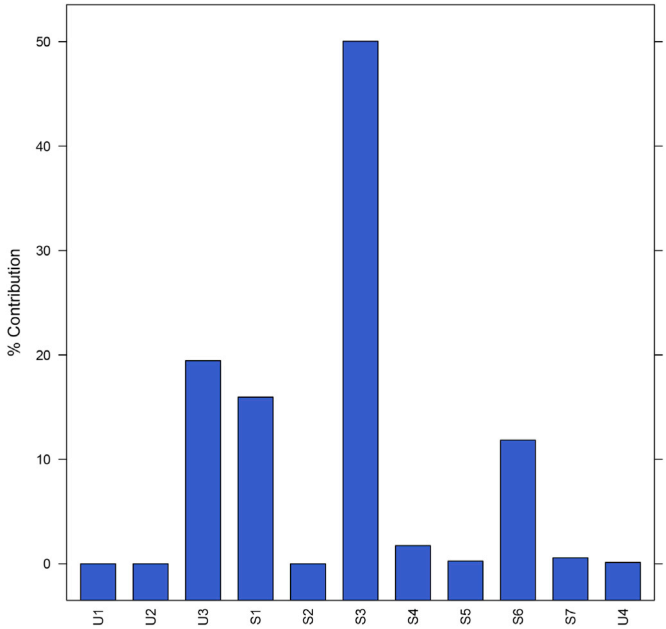


Fig. 15. Variable Contributions for point 10.

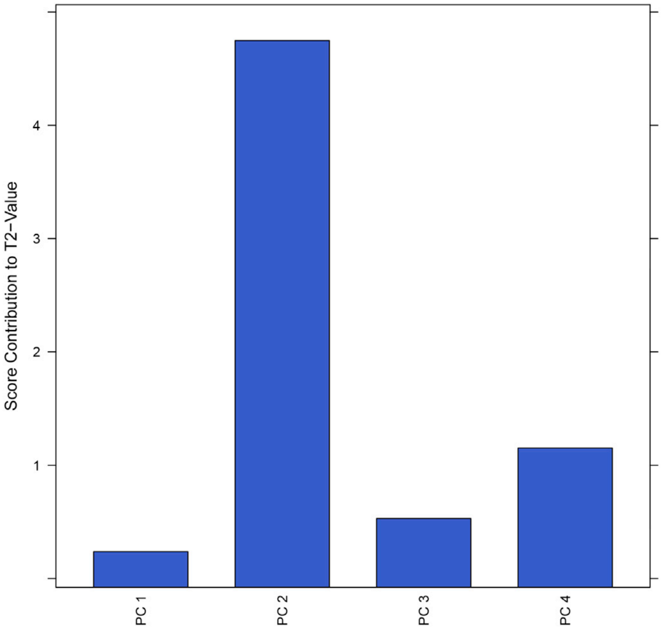


Fig. 16. Score Contributions for point 10.

monitoring biplots (based on the scaffolding (the relevant columns of $\boldsymbol{V}_{r^*}$) depicted in Fig. 9 to Fig. 14 for the different combinations of PC dimensions discussed in Section 3.2. The axes included for each dimension have an axis predictivity greater than 0.35 as indicated by the gray cells in Table 7.

As discussed in Section 1, PCA biplots provide for both the flagging of unexpected behaviour as well as the contributing variables on one graph. However, investigating all the principal component combinations can be a time consuming process, and the T^2 contribution value from (15) can be used in combination with the biplots to guide the user. In addition, the score contribution in (12) can be used to flag the principal component combinations that contributed most to the out of control T^2-value. Therefore, only the relevant principal component combinations need to be displayed on a specific biplot which will improve the interpretation of the results.

In the production facility these PCA biplots are displayed on a custom developed online web based interface. The biplots are updated every 15 min with the most recent process data available from the DCS systems. In the production facility the points outside the confidence ellipse are highlighted in red, and in combination with tool tips enables the easy identification of the relevant points. In addition, animation can provide a powerful tool to identify process trends. This allows the engineers and operators to monitor the production process continuously in real time. Note, due to printing constraints Figs. 19–24 depict the points outside the confidence ellipse by increased font size and the use of bold font, rather than the red coloring of the points used for the online implementation of the technology.

From Fig. 19 it can be concluded that points 10 to 12 are outside of the 90% confidence region. Points 10 to 12 project high on variable U3, and low on variables S1 and S3. These operating points indicate an

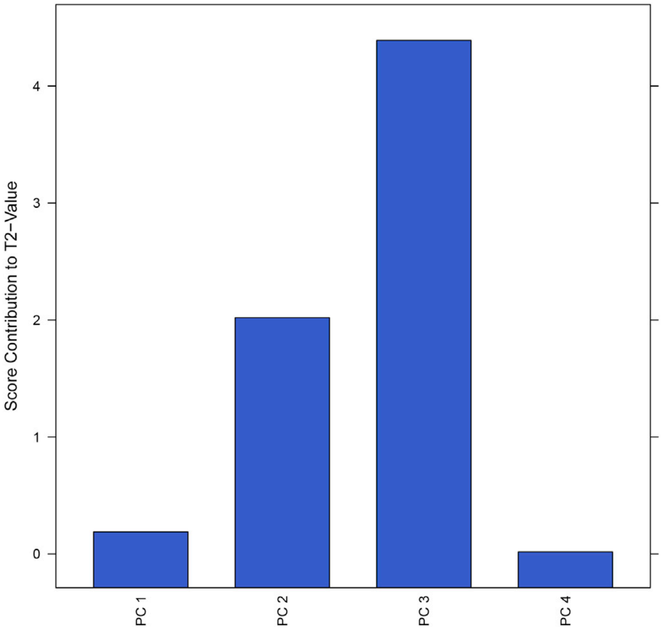


Fig. 17. Score Contributions for point 1.

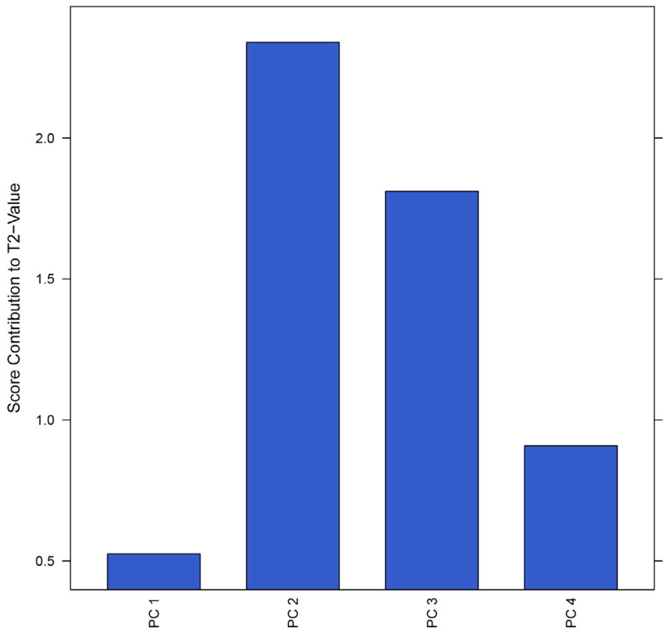


Fig. 18. Score Contributions for point 9.

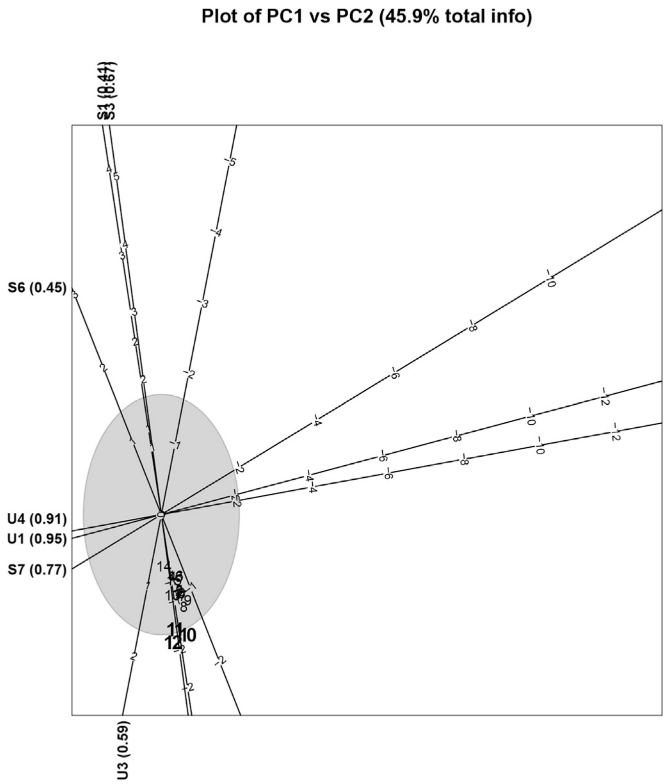


Fig. 19. PC1 vs PC2.

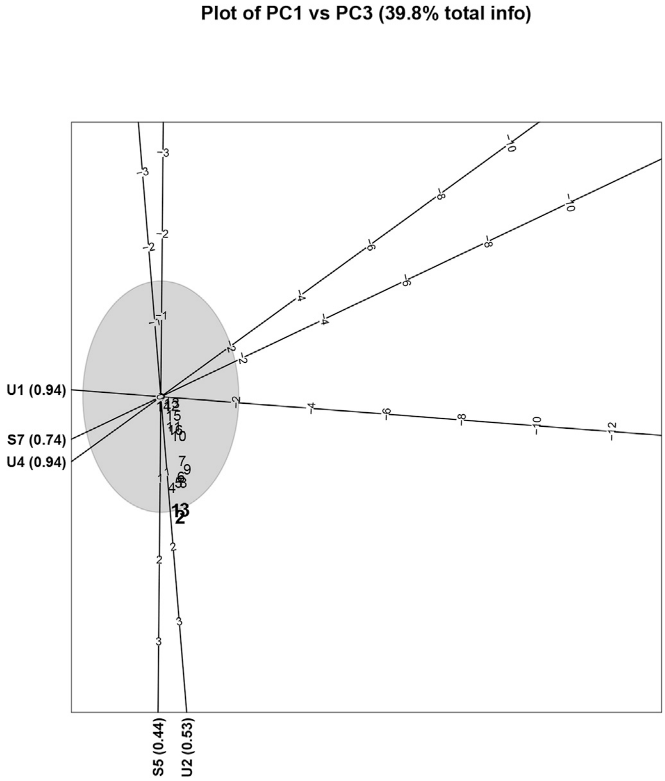


Fig. 20. PC1 vs PC3.

increase in reactor instability followed by an increase in the temperature readings, which necessitates an increase in the cooling agent. Considering Table 6 it is clear that variable U3 is higher for these points compared to the remaining points. Variable S3 is also lower for these points than for the remaining points, and variable S1 is also in general low for these points. As these points are almost exclusively separated on the PC2 dimension it is expected that they will be separated on all combinations including PC2. This can be confirmed by inspecting Figs. 22 and 23. Constructing the variable contribution plot discussed in Section 2.4 for point 10 (Fig. 15) confirms these results by highlighting variables U3, S1 and in particular S3. In addition, the score contribution plot for point 10 (Fig. 16) clearly highlights PC2.

Considering Fig. 20 it seems that the 15 min points moved from 1, 2 and 3 which were outside the concentration ellipse to a grouping of points 4 to 9, which up to the remaining points, were on target for these variables. This separation almost exclusively takes place on the PC3 dimension. From Table 6 it is noted that all these points (1–9) are high for variable U2 (and identical), which indicates the same operating point for the ratio. As expected, for variable S5 these points are somewhat higher in general. However, these two variables individually cannot explain the separation between points 1–3 and 4–9. Therefore, the remaining plots that contain PC3 are inspected. Fig. 22 indicates that the points are high on variable U3 and low on variables S3 and S1. Although this can again be confirmed from Table 6 these values still do not clearly differentiate between points 1 to 3 and 4 to 9. Fig. 24 accentuates the separation between these two groups of points. Specifically, points 1–3 project higher on variable S2 and lower on variable S4. This observation can be confirmed from Table 6. Points 1–3 are higher for variable S2 than the remaining points, and specifically points 2 and 3 are lower for variable S4 than the remaining points. These results highlight the importance of selecting the appropriate dimensions for out of control diagnostics. The score contribution plots for point 1 (Fig. 17) and point 9 (Fig. 18) clearly highlight the importance of PC2 and PC3 in understanding the behaviour of these points. The traditional score or biplot of the first two dimensions will not suffice to highlight all the relevant deviations, and would not have led to the correct diagnosis.

5. Conclusions

In this paper the PCA biplot and various measures of predictive power of the PCA biplot were investigated, and applied to a commercial industrial process.

Two measures of predictive power of the biplot axes were applied. Specifically, the mean standard predictive error (mspe) for PCA biplots

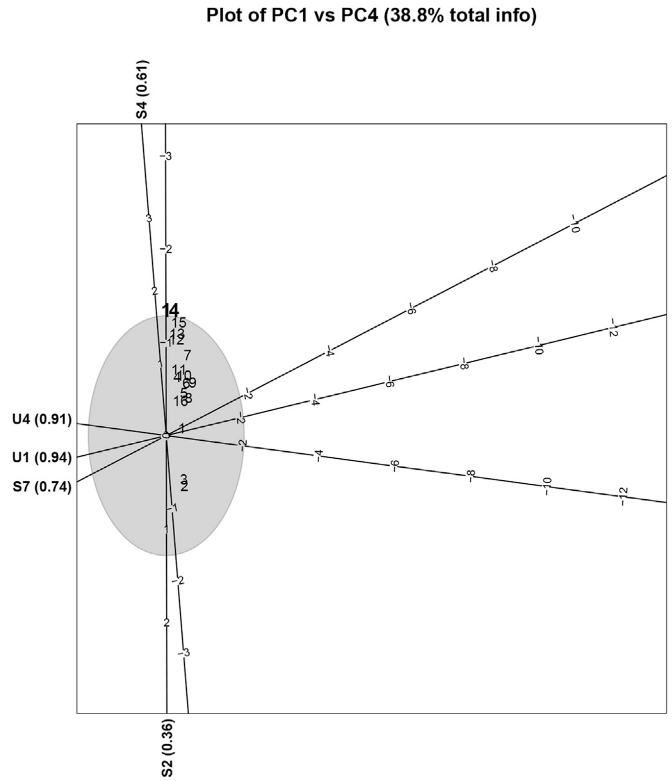


Fig. 21. PC1 vs PC4.

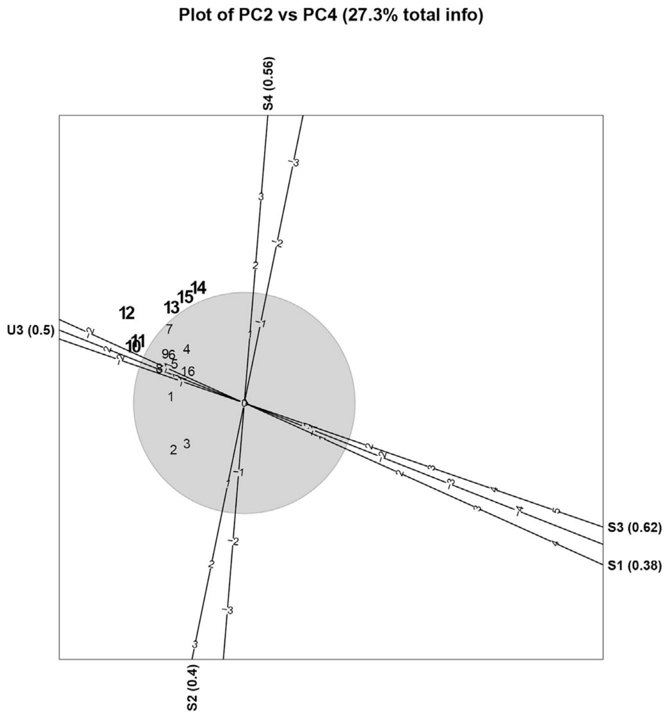


Fig. 23. PC2 vs PC4.

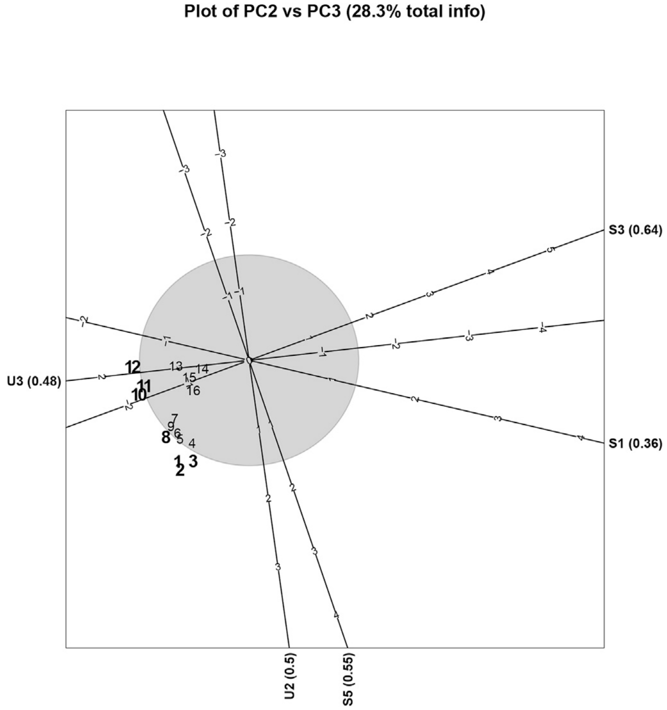


Fig. 22. PC2 vs PC3.

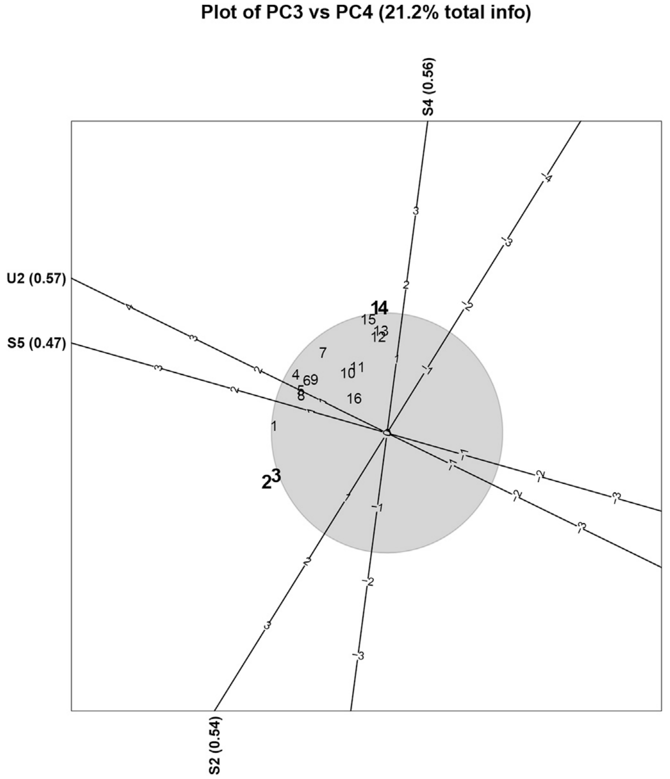
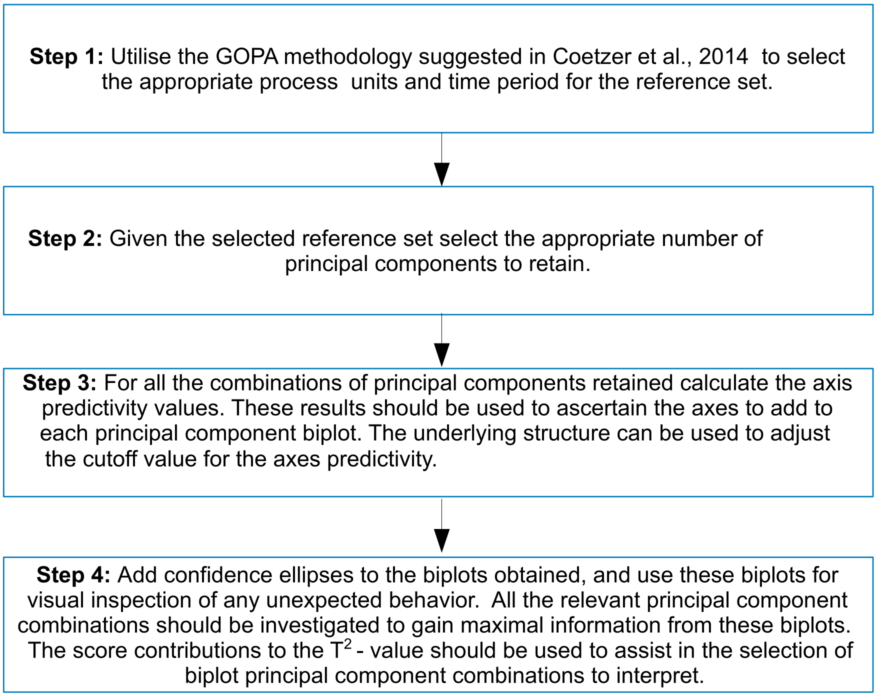


Fig. 24. PC3 vs PC4.

from Ref. [9] and the PCA axis predictivity from Ref. [8] were implemented and applied. In this case study, similar results were obtained using these two measures. However, the results from the PCA axis predictivities could be explained from a process perspective, and was investigated further.

From the axis predictivity results it was concluded that in addition to the intended axes allocation on the different biplots, the predictivity values on their own contains valuable information. Careful analysis of the axis predictivity highlights the variables captured by the different principal components. It could be illuminating to investigate the different groupings on the principal components for hidden interactions. The groupings contained in the case study were validated against the



Step 1: Utilise the GOPA methodology suggested in Coetzer et al., 2014 to select the appropriate process units and time period for the reference set.

Step 2: Given the selected reference set select the appropriate number of principal components to retain.

Step 3: For all the combinations of principal components retained calculate the axis predictivity values. These results should be used to ascertain the axes to add to each principal component biplot. The underlying structure can be used to adjust the cutoff value for the axes predictivity.

Step 4: Add confidence ellipses to the biplots obtained, and use these biplots for visual inspection of any unexpected behavior. All the relevant principal component combinations should be investigated to gain maximal information from these biplots. The score contributions to the T^2 - value should be used to assist in the selection of biplot principal component combinations to interpret.

Fig. 25. Steps for the implementation of a PCA biplot monitoring methodology.

underlying process, and found to be a true reflection of the underlying process dynamics.

Concentration ellipses were added to the selected principal component and axes combinations, and a new data set containing a four hour period of 15 min average data was projected onto the biplots. From this application it was concluded that more information is obtained by only including the axes with high predictive power, and by not only investigating the traditional PC1 × PC2 combination. It was also clear that a number of the PC combinations can be useful to extract the maximum information from the underlying structure of the data. The application of score contributions to the T^2- value was proposed to assist in the selection of biplot principal component combinations to display.

Therefore it is suggested that in any PCA biplot analysis the methodology presented in this paper should be employed. This is especially important for the real time on-line application of this case study. Information on the PCA biplot quality and the axis predictivities should be supplied on the plots to indicate to the user the confidence in the results. It is also prudent to compare the results obtained to the actual data to confirm the results, and to incorporate process understanding. It should however be noted that the actual process consists of 84 gasifiers, and it would be an overwhelming number of biplots to analyze if the user is not guided to the important gasifiers in a constructive manner. No part of the biplot analysis stands alone, and therefore the steps in Fig. 25 are suggested for the implementation of a PCA biplot monitoring methodology for multiple identical production processes.

In conclusion, PCA biplots are powerful visual aids in multivariate analysis. A significant amount of information can be obtained in one glance that would be virtually impossible to obtain from many tables of data. It is however necessary to exclude axes with low axis predictivity values as to not misguide the user, as the casual user will assign equal importance to all the axes as a direct consequence of the comparison of scatter plots to biplots often made in the literature. In addition, more eigenvalues than the first two should be investigated as there is possibly additional separation taking place on the remaining planes orthogonal to the first two principal components. The application of score contributions to the T^2- value and the selection of biplot principal component combinations to display is a novel approach to multivariate process monitoring. The R code developed to create the graphs in this paper will be discussed in detail in a future paper. The PCA biplot methodology presented in this paper is very effective for the real-time monitoring of unexpected behavior over short time periods for the coal to gas processing facility.

References

[1] J. Gower, S. Lubbe, N. Le Roux, Understanding Biplots, John Wiley & Sons, Chichester, 2011.

[2] C. Aldrich, S. Gardner, Le Roux, N. J, Monitoring of metallurgical process plant by using biplots, Am. Inst. Chem. Eng. J. 50 (2004) 2167–2186.

[3] R. Sparks, A. Adophson, A. Phatak, Multivariate process monitoring using the dynamic biplot, Int. Stat. Rev. 65 (3) (1997) 325–349.

[4] J.C. Gower, D.J. Hand, Biplots, Chapman & Hall, London, 1996.

[5] M. Greenacre, R. Primicerio, Multivariate analysis of ecological data. Fundaciòn BBVA, 2014.

[6] M. Greenacre, Biplots in practice. Fundaciòn BBVA, 2010.

[7] C. Eckart, G. Young, The approximation of one matrix by another of lower rank, Psychometrika 1 (3) (1936) 211–218.

[8] S. Gardner-Lubbe, N.J. Le Roux, J.C. Gower, Measures of fit in principal component and canonical variate analyses, J. Appl. Stat. 35 (9) (2008) 947–965.

[9] M.R. Alves, Evaluation of the predictive power of biplot axes to automate the construction and layout of biplots based on te accuracy of direct readings from common outputs of multivariate analyses: 1. application to principal component analysis, J. Chemom. 26 (2012) 180–190.

[10] R. Coetzer, R. Rossouw, N. Le Roux, Reference set selection with generalized orthogonal procrustes analysis for multivariate statistical process monitoring of multiple production processes, Chemometr. Intell. Lab. Syst. 132 (2014) 52–62.

[11] E. Russell, L. Chiang, R. Braatz, Data-driven Techniques for Fault Detection and Diagnosis in Chemical Processes. Springer, 2000 (London).

[12] J.F. MacGregor, T. Kourti, Statistical process control of multivariate processes, Contr. Eng. Pract. 3 (3) (1995) 403–414.

[13] H. Kaiser, The varimax criterion for analytic rotation in factor analysis, Psychometrika 23 (1958) 187–200.

[14] B. Everitt, T. Hothorn, An Introduction to Applied Multivariate Analysis with R. Springer, 2011 (New York).