Proc. of the European Safety and Reliability Conference (ESREL 2026)
Edited by José C. Matos, Paulo B. Lourenço, Michael Beer, Daniel Oliveira, Edoardo Patelli, Hélder S. Sousa, Rui A. Silva and Monica Santamaria

doi: 10.3850/ESREL2026061419_esrel26-p27727-cd

Data-Driven Multivariate Monitoring of Gas Well Integrity

Shawn C. Liebenberg
Focus Area for Pure and Applied Analytics, North-West University, Potchefstroom, South Africa.
E-mail: Shawn.Liebenberg@nwu.ac.za

Roelof L. J. Coetzer
Focus Area for Pure and Applied Analytics, North-West University, Potchefstroom, South Africa.
E-mail: Roelof.Coetzer@nwu.ac.za

Ockert C. Koekemoer
School of Mechanical Engineering, North-West University, Potchefstroom, South Africa.
E-mail: Ockert.Koekemoer@nwu.ac.za

Gert J. Nel
School of Mechanical Engineering, North-West University, Potchefstroom, South Africa.
E-mail: Gert.Nel@nwu.ac.za

Maintaining the integrity of gas wells is crucial for safe and efficient production, as failures can result in significant environmental and operational consequences. This research presents a multivariate statistical process monitoring framework designed to identify deviations that indicate potential well integrity problems. The methodology uses Principal Component Analysis (PCA) on observed well data to capture the main sources of variation across various pressure and flow metrics. The derived PCA scores and residuals are statistically monitored using Hotelling's T^2 and Squared Prediction Error (SPE) statistics to pinpoint abnormalities. To enhance fault detection and interpretation, we investigate a combined monitoring index that merges the T^2 and SPE statistics, yielding a balanced measure of deviations in well integrity. We apply this framework to simulated oil well data and to observed gas well data. The suggested framework is shown to provide a data-driven solution for well monitoring, aiding in the identification of abnormal behavior in gas well operations.

Keywords: well integrity, process monitoring, principal components, multivariate statistics

1. Introduction

The integrity of gas wells is a key issue for both environmental protection and public safety. Failures in gas well components can allow hazardous gases, such as methane, to escape and migrate into shallow aquifers, reach the surface, or be released into the atmosphere. A major climate-related priority for regulators is to restrict and reduce these fugitive emissions from wells. As a result, continuous monitoring of gas wells is a critical aspect of the production process. Modern well production systems are equipped with instruments and sensors that collect a multitude of data directly at the well and along the downstream gas flow. Multivariate statistical process monitoring (MSPM) provides a set of tools to process data and produce meaningful results for detecting operational anomalies. Principal component analysis (PCA) is often used as a preprocessing step in MSPM to reduce dimensionality while preserving the covariance structure of the multivariate data. Furthermore, monitoring statistics computed in the reduced space and the residual space allow for more sensitive detection of fault signals. This paper evaluates a data-driven pipeline for multivariate monitoring of gas well integrity based on principal component analysis. Monitoring is performed using statistics defined for both reduced space and residual space, complemented by variable contribution analyses for diagnostic interpretation. In addition, a combined monitoring index is used to integrate information from both spaces. The objective is to deliver a practical monitoring framework and associated visualizations that operators can directly use under realistic field conditions. The remainder of the paper is summarized below. Section 2 provides background information on MSPM and related work. Section 3 details the methods that will be used to monitor gas well integrity. Section 4 will show the methods applied to simulated oil well data and observed gas well data, while Section 5 provides conclusions and future work.

2. Background

In this section, we provide the relevant background for the methods discussed in Section 3 and applied in Section 4.

2.1. *Principal Component Analysis*

PCA is widely used as a method to reduce the dimensionality in datasets with a large number of variables, while preserving as much information about the original set of data as possible. This is accomplished by creating new, uncorrelated variables (principal components) through linear combinations of the variables in order to maximize the variance explained by the latent variables. From an algebraic perspective, principal components are a linear combination of d variables $x_1, x_2, \ldots x_d$, and from a geometric standpoint, the linear combinations are treated as new coordinate systems, rotating the original system with respect to the variables $x_1, x_2, \ldots x_d$ as new coordinate axes. See Johnson et al. (2002) for more details. For the practical implementation of PCA, reduced variables are found through eigenvalue decomposition, also known as spectral decomposition. To establish the theoretical background, we adopt the notation of Liebenberg et al. (2024). If $\mathbf{X}$ is an $n \times d$ matrix with mean-zero columns, principal component analysis seeks the linear combination $\mathbf{Xm}$ that maximizes its variance $\mathrm{Var}(\mathbf{Xm}) = \mathbf{m}^\top \mathbf{S}\,\mathbf{m}$, subject to an appropriate normalization of $\mathbf{m}$, where $\mathbf{m} = (m_1, \ldots, m_d)^\top$ and $\mathbf{S}$ denotes the $d \times d$ sample covariance matrix Rodwell et al. (2021). Let $(\lambda_i, \mathbf{e}_i)$, $i = 1, \ldots, d$, be the ordered eigenvalue–eigenvector pairs of $\mathbf{S}$. The ith principal component score vector is then

$$Y_i = \mathbf{e}^\intercal \mathbf{X} = e_{i1}\mathbf{X}_1 + e_{i2}\mathbf{X}_2 + \cdots + e_{ip}\mathbf{X}_d,$$

for $i = 1, 2, \ldots, d$. Moreover, $\mathrm{Var}(Y_i) = \lambda_i$. Equivalently, PCA may be performed via the singular value decomposition (SVD). The SVD of $\mathbf{X}$ is

$$\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\intercal,$$

where $\boldsymbol{\Sigma}$ is a diagonal matrix of singular values and $\mathbf{V}$ contains the right singular vectors, which correspond to the PCA loadings. From this decomposition, $\mathbf{X}^\intercal\mathbf{X} = \mathbf{V}\boldsymbol{\Sigma}^2\mathbf{V}^\intercal$, and setting $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^2$ yields $\mathbf{X}^\intercal\mathbf{X} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\intercal$. The sample covariance matrix is therefore

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}^\intercal\mathbf{X} = \frac{1}{n-1}\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\intercal.$$

Equivalently, the scaled SVD

$$(n-1)^{-1/2}\mathbf{X} = \mathbf{U}\tilde{\boldsymbol{\Sigma}}\mathbf{V}^\intercal,$$

with $\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}/\sqrt{n-1}$, yields $\mathbf{S} = \mathbf{V}\tilde{\boldsymbol{\Sigma}}^2\mathbf{V}^\intercal$, showing that the principal directions obtained from the SVD coincide with those from the eigendecomposition of $\mathbf{S}$ (see Russell et al. (2000) for details). In reduced r dimensions, the PCA score matrix is $\mathbf{Z}^* = \mathbf{X}\mathbf{V}_r$, which provides the rank-r approximation $\mathbf{X} \approx \mathbf{Z}^*\mathbf{V}_r^\intercal$. Geometrically, the rows of $\mathbf{Z}^*$ give the coordinates of the samples in the r-dimensional reduced space, while each column corresponds to the scores of a single principal component across the samples. The variance of the ith column of $\mathbf{Z}^*$ is $\lambda_i = \sigma_i^2/(n-1)$, $i = 1, \ldots, r$, so the eigenvalues of $\mathbf{S}$ quantify the variance explained by the corresponding eigenvectors (principal components). If the first r principal components explain a sufficiently large fraction of the total variance, they provide an adequate low-dimensional representation of the data. The PCA loadings are given by the columns of $\mathbf{V}_r$.

2.2. *Multivariate statistical process monitoring*

In general, statistical process monitoring refers to the application of univariate and multivariate statistical techniques to ensure product quality in manufacturing environments. These techniques allow for the early identification and correction of deviations from the mean of the process or specified operating limits, thereby reducing the risk of producing nonconforming goods.

A variety of methodological approaches have been proposed and applied in the context of statistical process monitoring. In the multivariate setting, principal component analysis has been extensively used for process monitoring and control as demonstrated by Yu (2012), among others. Common monitoring statistics derived from PCA include Hotelling's T^2-statistic and the squared prediction error (SPE), also known as the Q-statistic, which are frequently employed to detect out-of-control conditions. Such applications are illustrated in Ferrer (2007). Control charts is a popular tool which provide a graphical representation of process behaviour over time. Examples of multivariate control chart applications in industrial settings can be found in Ueda et al. (2022), and the references therein. For a more recent example on MSPM see, Jin et al. (2024).

3. Method

Variable-reduction techniques such as PCA are particularly important in MSPM because multivariate datasets frequently contain correlated variables. High correlation among process variables can distort monitoring statistics and their visualizations, reducing the reliability of anomaly detection. The MSPM methods examined below depend on PCA to produce a lower-dimensional set of (approximately) uncorrelated scores. When PCA is integrated into MSPM, it serves three roles: dimensionality reduction, improved visualization, and outlier detection. The first r principal components and their score vectors are used to compute the T^2-statistic; the residual subspace, of dimension $r^* = d-r$, is used to compute the SPE, and a combined statistic merges information from both subspaces. Calculating monitoring statistics in this reduced/orthogonal space mitigates the adverse effects of highly correlated original variables. Yu (2012) provide an application of PCA-based monitoring in simulated examples. The following sections discuss in detail the T^2, SPE, and combined statistics.

3.1. *Hotelling's T^2-statistic*

The general approach to identifying multivariate outliers is to quantify the extent to which each observation deviates from the central tendency of the data. A commonly used measure for this purpose is Hotelling's T^2-statistic. For mean-centered observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, the T^2 statistic for the ith observation is defined as

$$T_i^2 = \mathbf{x}_i^\intercal \mathbf{S}^{-1} \mathbf{x}_i,$$

where $\mathbf{S}$ denotes the sample covariance matrix and $\mathbf{x}_i = [x_{i1} \ x_{i2} \ \ldots \ x_{id}]$, $i = 1, 2, \ldots, n$. For a new observation $\mathbf{x}_{new}$, the corresponding statistic is computed using the covariance matrix estimated from the original (Phase I) data, $T^2_{new} = \mathbf{x}_{new}^\intercal \mathbf{S}^{-1} \mathbf{x}_{new}$. Although this statistic has the same quadratic form as Hotelling's T^2, its interpretation in a process-monitoring context differs from that of the classical setting. In particular, classical results assume that the observation contributes to the estimation of $\mathbf{S}$, which yields an exact finite-sample F-distribution under multivariate normality. In statistical process monitoring, however, $\mathbf{S}$ is treated as fixed once estimated from Phase I data and the statistic is evaluated for independent new (Phase II) observations. In this case, the resulting distance measure, often referred to as the D-statistic following Westerhuis et al. (2000), is asymptotically χ^2-distributed. Consequently, control limits based on the upper $100(1-\alpha)\%$ quantile of a χ^2-distribution are commonly used for monitoring new observations.

In PCA-based monitoring, PCA is first applied to the sample covariance matrix $\mathbf{S} = \frac{1}{n-1}\mathbf{X}^\intercal\mathbf{X}$, where $\mathbf{X} = [\mathbf{x}_1 \ \mathbf{x}_2 \ \ldots \ \mathbf{x}_n]^\intercal$. A spectral decomposition yields

$$\mathbf{S} = \begin{bmatrix}\mathbf{V}_r \ \mathbf{V}_{r^*}\end{bmatrix} \begin{bmatrix}\mathbf{\Lambda}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{\Lambda}_{r^*}\end{bmatrix} \begin{bmatrix}\mathbf{V}_r \ \mathbf{V}_{r^*}\end{bmatrix}^\intercal,$$

where $\mathbf{V}$ contains the eigenvectors of $\mathbf{S}$ and $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues. Retaining r principal components gives the reduced covariance approximation $\hat{\mathbf{S}} = \mathbf{V}_r\mathbf{\Lambda}_r\mathbf{V}_r^\intercal$. The monitoring statistic for a new observation is then

$$T^2_{new} = \mathbf{x}_{new}^\intercal \mathbf{V}_r \mathbf{\Lambda}_r^{-1} \mathbf{V}_r^\intercal \mathbf{x}_{new},$$

which, under standard conditions, is asymptotically χ_r^2-distributed. An observation is identified as a fault when T_{new}^2 exceeds the corresponding χ_r^2 control limit. Denote this $100(1-\alpha)\%$ control limit, $\tau^2 = \chi_\alpha^2(r)$.

3.2. *Squared prediction error*

The squared prediction error (SPE) measures the residual variation between an observation and its projection into the subspace of the principal components used in the PCA model Slišković et al. (2012). Throughout, observations are assumed to be mean-centered using the Phase I sample mean. The SPE quantifies the magnitude of the component of an observation that lies in the residual subspace orthogonal to the retained principal components. Alcala and Qin (2011) define the SPE for a centered observation $\mathbf{x}$ as

$$SPE = \mathbf{x}^\intercal \mathbf{V}_{r^*} \mathbf{V}_{r^*}^\intercal \mathbf{x},$$

where $\mathbf{V}_{r^*}$ contains the eigenvectors associated with the discarded $r^* = d - r$ principal components. For a new centered observation $\mathbf{x}_{new}$, the corresponding statistic is given by

$$SPE_{new} = \mathbf{x}_{new}^\intercal \mathbf{V}_{r^*} \mathbf{V}_{r^*}^\intercal \mathbf{x}_{new}.$$

The SPE is monitored using a control limit of the form $\delta^2 = g^{SPE}\chi_\alpha^2(h^{SPE})$, corresponding to a confidence level of $100(1-\alpha)\%$. This control limit is based on a moment-matching approximation to the distribution of the SPE. Here,

$$g^{SPE} = \frac{\theta_2}{\theta_1}, \qquad h^{SPE} = \frac{\theta_1^2}{\theta_2},$$

with $\theta_1 = \sum_{i=r+1}^{d} \lambda_i$, and $\theta_2 = \sum_{i=r+1}^{d} \lambda_i^2$, where λ_i denotes the i^{th} eigenvalue of the sample covariance matrix $\mathbf{S}$.

3.3. *Combined monitoring statistic*

During process monitoring, it may be desirable to rely on a single monitoring statistic rather than using both the T^2 and SPE statistics separately. To address this, Yue and Qin (2001) proposed a combined monitoring statistic that integrates information from both the T^2 and SPE measures. The combined statistic is defined as

$$\varphi = \mathbf{x}^\intercal \mathbf{\Phi} \mathbf{x}, \text{ where } \mathbf{\Phi} = \frac{\tilde{\mathbf{C}}}{\delta^2} + \frac{\mathbf{D}}{\tau^2},$$

with $\tilde{\mathbf{C}} = \mathbf{V}_{r^*}\mathbf{V}_{r^*}^\intercal$, and $\mathbf{D} = \mathbf{V}_r \mathbf{\Lambda}_r^{-1} \mathbf{V}_r^\intercal$. The combined monitoring statistic is associated with a control limit $\zeta = g^\varphi \chi_\alpha^2(h^\varphi)$, corresponding to a confidence level of $100(1-\alpha)\%$, where

$$g^\varphi = \frac{r/\tau^4 + \theta_2/\delta^4}{r/\tau^2 + \theta_1/\delta^2}, \; h^\varphi \quad = \frac{\left(r/\tau^2 + \theta_1/\delta^2\right)^2}{r/\tau^4 + \theta_2/\delta^4}.$$

Within this framework, Alcala and Qin (2011) conducted a simulation study demonstrating the effectiveness of the three monitoring statistics discussed in this section for fault diagnosis. Although the combined index provides a good summary of overall process behaviour, it does not by itself reveal the underlying sources of deviation. To address this, contribution plots are examined in Section 3.5. Another potential concern with combined statistics is the individual contributions may mask one another, leading to missed detections. However, since the combined statistic is an additive combination of two nonnegative quantities, cancellation between terms is not possible

3.4. *Control charts*

The results of monitoring statistics are typically visualized using control charts, which are widely applied in practice for the implementation of MSPM. Control charts provide a graphical display of process behavior over time, allowing changes in process variability and violations of control limits to be easily identified. In such charts, the value of the monitoring statistic for each observation (such as the T^2, SPE, or combined statistic) is plotted alongside its corresponding control limit.

Santos-Fernández (2012) provide a detailed illustration of multivariate control chart construction and include accompanying code for generating control charts from sample data in an associated R package.

3.5. *Variable Contribution*

Once MSPM identifies observations deemed faulty, further investigation, such as computing variable contributions, can help pinpoint the variable responsible for the anomaly in that observation. Alcala and Qin (2011) demonstrated the use of the T^2, SPE, and combined statistics, together with their variable contributions, for fault diagnosis in a simulation study. They described several types of variable contributions that can be computed for each statistic; among these are complete decomposition contributions (CDC), partial decomposition contributions, and reconstruction-based contributions. Each type admits a general representation in terms of a matrix $\mathbf{M}$, whose form depends on the monitoring statistic under consideration. In this paper, we restrict attention to the CDC. The CDC decomposes a monitoring statistic into the sum of its variable contributions. Its general expression is

$$CDC_i^{\mathit{Index}} = \mathbf{x}^T \mathbf{M}^{(1/2)} \xi_i \xi_i^T \mathbf{M}^{(1/2)} \mathbf{x},$$

where ξ_i denotes the ith column $(i = 1, 2, ..., n)$ of the identity matrix. The term *Index* will later be replaced with the specific monitoring statistic of interest. For different choices of $\mathbf{M}$, the CDC can be expressed using the matrices employed to compute the corresponding monitoring statistic. Choosing $\mathbf{M} = \mathbf{D} = \mathbf{V}_r \mathbf{\Lambda}_r^{-1} \mathbf{V}_r^\intercal$, the CDC for the T^2 monitoring statistic becomes

$$CDC_i^{T^2} = \left(\xi_i^T \mathbf{V}_r \mathbf{\Lambda}_r^{-(1/2)} \mathbf{V}_r^T \mathbf{x} \right)^2.$$

To calculate the CDC for the SPE statistic, set $\mathbf{M} = \tilde{\mathbf{C}} = \mathbf{V}_{r^*} \mathbf{V}_{r^*}^\intercal$. Since $\tilde{\mathbf{C}}$ is idempotent, we have $\tilde{\mathbf{C}}^{1/2} = \tilde{\mathbf{C}}$, which implies $\tilde{\mathbf{C}}\mathbf{x} = \tilde{\mathbf{x}}$. Consequently, the CDC is given by

$$CDC_i^{SPE} = \left(\xi_i^T \tilde{\mathbf{x}} \right)^2 = \tilde{x}_i^2.$$

For the combined statistic, $\mathbf{M}$ is chosen as $\mathbf{M} = \mathbf{\Phi} = \frac{\tilde{\mathbf{C}}}{\delta^2} + \frac{\mathbf{D}}{\tau^2}$, yielding

$$CDC_i^{\mathit{Combined}} = \mathbf{x}^T \mathbf{\Phi}^{(1/2)} \xi_i \xi_i^T \mathbf{\Phi}^{(1/2)} \mathbf{x}.$$

4. Analysis

In this section, we apply the proposed MSPM techniques to both simulated oil well data and real observed gas well data. The simulated oil well data set provides a controlled environment, enabling an evaluation of the detection capabilities in known fault scenarios. The real gas well data complement this analysis by demonstrating the practical applicability of the methods under operational conditions.

4.1. *Monitoring of oil well data*

We used the simulated subset of the 3W data set introduced by Vargas et al. (2019), which was designed to closely mimic the dynamic behavior of offshore oil production wells while allowing controlled injection of fault events. The simulated data consist of multivariate time series generated from a realistic well model, capturing pressure, temperature, and flow-related measurements under normal operation, faulty steady-state conditions, and transient fault scenarios. Each time series instance is annotated by labels indicating normal operation (class = 0), faulty steady state (class = 1), or one of several transient fault events (class = 101–108), corresponding to different types of undesirable operational disturbances. The simulated data set maintains challenges encountered in real industrial systems, including strong class imbalance, heterogeneous variable scales, and frozen signals. This makes it particularly suitable for research on multivariate process monitoring, fault detection, and early warning systems.

The simulated 3W oil well dataset consists of high-frequency, time-stamped multivariate measurements. Recorded variables include pump discharge pressure (`P.PDG`), production tubing pressure (`P.TPT`) and temperature (`T.TPT`), pressure at the production choke manifold (`P.MON.CKP`), and temperature upstream of the choke (`T.JUS.CKP`). Each observation is associated with a time stamp and an operational state label (`class`) indicating normal operation, faulty steady state, or transient fault events. From the 8 possible fault options, we chose to illustrate the monitoring procedure on the *hydrate in the production line* scenario. Hydrate is a crystalline compound that is formed by water and natural gas and therefore resembles ice. Hydrate formation occurs most frequently in gas pipelines and gas-producing wells, however, hydrates also form in oil-producing wells Vargas et al. (2019). Hydrate can accumulate to the extent that it completely obstructs flow. Figure 1 shows the univeriate time series plots for the production variables. The green shading represents the period for normal operation, blue is the transient fault state, and red is the fault steady state. We see a sharp increase or decrease in all variables during the transient fault state. Figure 2 shows the control charts for the monitoring statistics calculated on the PCA. We see that all the monitoring statistics flags the fault early in the transient fault state. Variable-level contributions were quantified using the CDC associated with each monitoring statistic. For the first observation identified as out of control, CDC values were computed for each variable which were reported in Table 1. In addition, CDC values were aggregated over the first 50 out-of-control observations and represented as a proportion of the total contribution, as seen in Table 2. From Table 1, the T^2-statistic attributes the initial fault mainly to `T.JUS.CKP` and `P.MON.CKP`, whereas the SPE statistic shows more distributed contributions. The combined statistic identifies `T.TPT` as the dominant contributor at the point of detection. From Table 2, the T^2-contribution show that `T.JUS.CKP` and `P.MON.CKP` dominates, and the SPE chart shows `T.TPT`, while the combined statistic integrates both effects, yielding `T.TPT` as the dominant contribution.

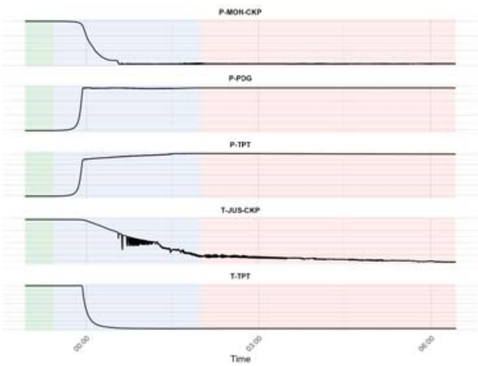

Fig. 1.: Time series for variables of simulated oil well.

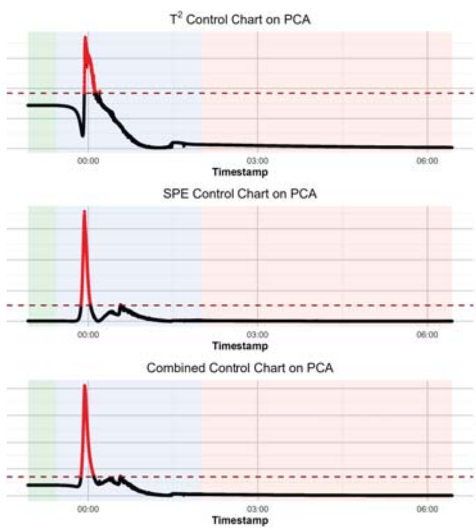

Fig. 2.: Multivariate monitoring statistics for simulated oil well.

Table 1.: Variable-level CDC values for the first out-of-control observation under different monitoring statistics

Variable	T^2	SPE	Combined
`P.PDG`	0.7673	0.0580	0.0086
`P.TPT`	0.8652	0.0750	0.0162
`T.TPT`	0.9095	0.1500	0.8606
`P.MON.CKP`	0.9874	0.0467	0.4486
`T.JUS.CKP`	1.2218	0.0115	0.0379

Table 2.: Mean relative CDC contributions across the first 50 out-of-control observations

Variable	T^2	SPE	Combined
`T.JUS.CKP`	0.3168	0.0342	0.0223
`P.MON.CKP`	0.2027	0.1425	0.3150
`T.TPT`	0.1779	0.4329	0.6098
`P.TPT`	0.1646	0.2177	0.0302
`P.PDG`	0.1381	0.1727	0.0228

4.2. *Monitoring of gas well data*

A typical natural gas well consists of a drilled borehole that is lined with several layers of steel casing, each cemented in place to support the well structure; isolating surrounding formations, and protecting groundwater (see Figure 1 in Cao et al. (2023)). Shallow conductor and surface casings stabilize near-surface formations, while the intermediate casing provides additional isolation at greater depths. The production casing extends down to the gas-bearing reservoir and forms a sealed pathway for gas to flow upward. Inside this casing, a narrower production tubing carries the gas to the surface. At ground level, the well is fitted with a wellhead assembly, including casing and tubing heads and a set of valves commonly referred to as the Christmas tree, which regulates flow, maintains pressure, and provides basic safety control during production Speight (2018).

Natural gas wells are typically monitored through measurements of pressures (wellhead, tubing, casing, and formation), temperatures, and flow rates of gas, condensate, and water, which provide insight into reservoir performance and well integrity. In this paper, the data originates from gas wells located in Southern Africa. The variables monitored in Tables 3 and 4 include tubing head pressure (`THP`), flow tubing Head pressure (`FTH`), line pressure (`PI`), annulus pressures A (`PI_A`), B (`PI_B`), line temperature (`TI`), and gas production rate (`RATE`). Many additional variables were included in the MSPM, however, only the top contributing variables for the first flagged abnormal (or faulty) observation, as identified by the monitoring statistics in Figure 3, are reported. The

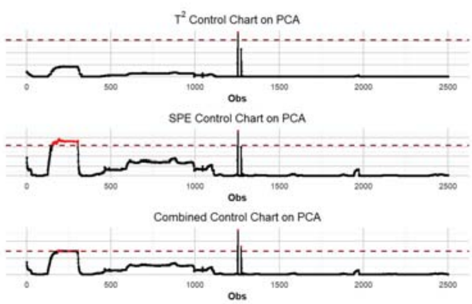

Fig. 3.: Multivariate monitoring statistics for gas well.

three control charts in Figure 3 show different but related behavior. The dashed red line marks the control limit, and the red points indicate when the limit is exceeded. The SPE chart rises early in the series and stays above the limit for a period, while this change is only weakly visible in the Combined chart and not detected by the T^2-chart until much later. The T^2-chart instead shows a later, sharp spike that is detected by all three measures. This suggests that SPE is more sensitive to early, local changes in the process, whereas T^2 responds mainly to larger changes affecting the system as a whole, with the Combined statistic lying between the two. From a practical point of view, early SPE signals can be used as an initial warning that prompts closer inspection of key variables, while Combined or T^2 signals provide stronger confirmation before major operational decisions are made. The variable contribution summaries in

Table 3.: Variable contributions SPE.

Variable	**Contribution**	**Prop.**
`PI_A`	8.6223	0.6643
`PI`	2.2845	0.1760
`TI`	0.7511	0.0579
`THP`	0.5827	0.0449
`PI_B`	0.4802	0.0370
`FTH`	0.2198	0.0169

Table 4.: Variable contributions Comb.

Variable	**Contribution**	**Prop.**
`PI`	0.7167	0.6727
`THP`	0.1497	0.1405
`FTH`	0.0815	0.0765
`RATE`	0.0561	0.0527
`TI`	0.0212	0.0199
`PI_B`	0.0139	0.0130
`PI_A`	0.0132	0.0123

Tables 3 (SPE) and 4 (Combined) show a different allocation of contribution across the monitoring variables. Both tables display the contribution and the proportion of total contribution. For the SPE statistic (Table 3), `PI_A` exhibits the largest contribution (8.6223, 66.43% of the total), followed by `PI` (2.2845, 17.60%); the remaining variables each contribute only a few percent. By contrast, the Combined statistic (Table 4) attributes most of its contribution to `PI` (0.7167, 67.2%), with `THP` and `FTH` together accounting for roughly 21% and the other variables individually contributing at the 1–5% level. The two monitoring statistics therefore highlight different drivers of the flagged

event. Importantly, the SPE statistic also reported an abnormal signal earlier in the monitoring sequence than the Combined statistic, indicating potentially greater sensitivity for early detection in this case. Together, these results suggest a complementary monitoring strategy; SPE provides an early alarm that appears driven by `PI_A`, whereas the Combined statistic offers a broader view that may be useful for confirmation.

5. Conclusion

In this paper, we presented a multivariate statistical process monitoring framework for gas wells based on principal component analysis, with monitoring statistics computed in both the principal component and residual subspaces. The methodology was evaluated on simulated data reflecting realistic oil well operating conditions and subsequently applied to real gas well data, demonstrating its ability to detect process faults. The results indicate that the SPE statistic is sensitive to early deviations, while the combined index remains responsive but exhibits more conservative behaviour. Since PCA is restricted to linear structure, extensions to nonlinear dimension reduction techniques may enhance detection capability. In addition, alternative strategies for combining monitoring statistics and a systematic investigation of their sensitivity constitute natural directions for further research.

Acknowledgement

The authors would like to express their sincere gratitude to Mr Derrick Piater, Senior Manager at Sasol Gas Sourcing Operations, for his valuable guidance and ongoing support in the development of the well integrity and risk assessment process. The authors would also like to acknowledge Jennifer L. Liebenberg for making her dissertation and accompanying code available which provided valuable reference material and technical insight.

References

Alcala, C. F. and S. J. Qin (2011). Analysis and generalization of fault diagnosis methods for process monitoring. *Journal of Process Control 21*(3), 322–330.

Cao, L., J. Sun, B. Zhang, N. Lu, and Y. Xu (2023). Analysis of multiple annular pressure in gas storage well and high-pressure gas well. *Energy Engineering: Journal of the Association of Energy Engineers 120*(1), 35.

Ferrer, A. (2007). Multivariate statistical process control based on principal component analysis (mspc-pca): some reflections and a case study in an autobody assembly process. *Quality Engineering 19*(4), 311–325.

Jin, W., W. Wang, Y. Wang, Z. Cao, and Q. Jiang (2024). Distributed monitoring of nonlinear plant-wide processes based on ga-regularized kernel canonical correlation analysis. *Reliability Engineering & System Safety 252*, 110421.

Johnson, R. A., D. W. Wichern, et al. (2002). Applied multivariate statistical analysis.

Liebenberg, J. L., R. Coetzer, M. Cockeran, and T. Nombebe (2024). Multivariate statistical process monitoring and diagnostic analysis of compositional data.

Rodwell, D. T., C. J. van der Merwe, and S. Gardner-Lubbe (2021). Categorical cva biplots. *Computational Statistics & Data Analysis 163*, 107299.

Russell, E., L. Chiang, and R. Braatz (2000, 01). *Data-Driven Methods for Fault Detection and Diagnosis in Chemical Processes*.

Santos-Fernández, E. (2012). *Multivariate statistical quality control using R*, Volume 14. Springer Science & Business Media.

Slišković, D., R. Grbić, and Ž. Hocenski (2012). Multivariate statistical process monitoring. *Tehnički vjesnik 19*(1), 33–41.

Speight, J. G. (2018). *Natural gas: a basic handbook*. Gulf Professional Publishing.

Ueda, R. M., Ì. R. Agostino, and A. M. Souza (2022). Analysis and perspectives on multivariate statistical process control charts used in the industrial sector: a systematic literature review. *Management and Production Engineering Review 13*.

Vargas, R. E. V., C. J. Munaro, P. M. Ciarelli, A. G. Medeiros, B. G. do Amaral, D. C. Barrionuevo, J. C. D. de Araújo, J. L. Ribeiro, and L. P. Magalhães (2019). A realistic and public dataset with rare undesirable real events in oil wells. *Journal of Petroleum Science and Engineering 181*, 106223.

Westerhuis, J. A., S. P. Gurden, and A. K. Smilde (2000). Generalized contribution plots in multivariate statistical process monitoring. *Chemometrics and intelligent laboratory systems 51*(1), 95–114.

Yu, J. (2012). Local and global principal component analysis for process monitoring. *Journal of Process Control 22*(7), 1358–1373.

Yue, H. H. and S. J. Qin (2001). Reconstruction-based fault identification using a combined index. *Industrial & engineering chemistry research 40*(20), 4403–4414.