REVISTA DE MÉTODOS CUANTITATIVOS PARA LA ECONOMÍA Y LA EMPRESA (30). Páginas 279-296. Diciembre de 2020. ISSN: 1886-516X. D.L: SE-2927-06. www.upo.es/revistas/index.php/RevMetCuant/article/view/3509

Diseño de una carta de control basada en Análisis de Componentes Principales. Un caso de estudio

RAHMER, BRUNO DE JESÚS
Facultad de Ingeniería. Fundación Universitaria Tecnológico Comfenalco (Colombia)
Correo electrónico: brunodejesus.2509@gmail.com

SOLANA GARZÓN, JOSÉ
Fundación Universitaria Tecnológico Comfenalco (Colombia)
Correo electrónico: ingjosemsolanag@gmail.com

GARZÓN SAÉNZ, HERNANDO
Fundación Universitaria Tecnológico Comfenalco (Colombia)
Correo electrónico: hnando2001@gmail.com

ORTIZ PIEDRAHITA, GUSTAVO
Fundación Universitaria Tecnológico Comfenalco (Colombia)
Correo electrónico: gustavoaop@gmail.com

RESUMEN

El abanico de métodos y técnicas cuantitativas utilizadas para la detección de variaciones de fondo en procesos de manufactura se circunscriben en una disciplina matriz, catalogada como control estadístico de procesos. Tales métodos poseen una aptitud sobresaliente para diagnosticar el estado general de sistemas productivos y realizar una evaluación simultánea de diversas características de calidad interrelacionadas. En el marco de esta investigación, se propone el análisis y monitoreo de un proceso químico sustentado en los principios teóricos del análisis en componentes principales. De este modo, las variables originales seleccionadas son representadas en un espacio dimensional más compacto, bajo la hipótesis de normalidad multivariante. Se construye en la fase posterior, un gráfico de control basado en los cuadrados de los errores de predicción con el objeto de evaluar el comportamiento de los factores subyacentes. Los resultados indican que el proceso no es marginalmente estable y existe un margen de reducción de variabilidad significativo.

Palabras clave: industria química, reducción de datos, análisis estadístico, control de calidad multivariante, variabilidad.
Clasificación JEL: C19; C38.
MSC2010: 62H25.

Artículo recibido el 10 de agosto de 2018 y aceptado el 1 de julio de 2019.

Design of a control charter based on principal component analysis. A case study

ABSTRACT

The set of quantitative methods and techniques used to detect assignable variations in manufacturing processes are contained within a discipline classified as statistical process control. Such methods carry out precise evaluations on the general state of productive systems and carry out simultaneous monitoring of various interrelated quality characteristics. In the framework of this research, the analysis of a chemical process based on the theoretical principles of principal component analysis is proposed, which enables the representation of the original variables in a compact dimensional space. In the later phase, a control graph based on the squares of the prediction errors is constructed in order to evaluate the behavior of the composite variables found. The results indicate that the process is not marginally stable and it is necessary to reduce its variability margin.

Keywords: chemical industry, data reduction, statistical analysis, multivariate quality control, variability.
JEL classification: C19; C38.
MSC2010: 62H25.

1. Introducción.

Las técnicas de análisis multivariante han sido reconocidas en múltiples dominios (incluyendo el monitoreo industrial) por su capacidad superlativa para la detección de anomalías en procesos donde ocurren comovimientos entre las variables y donde existen ingentes cantidades de registros disponibles (Rahim, Siddiqui & Elshafei, 2014). Las técnicas utilizadas para la construcción de marcos de monitorización apuntan a estabilizar el proceso, reducir su nivel de variabilidad y prever la aparición de eventos de fondo que repercuten negativamente en el performance de los sistemas productivos (Camacho, Pérez-Villegas, García-Teodoro & Maciá-Fernández, 2016).

Aunque es innegable que tales técnicas no gozan de una generalizada popularidad, es preciso indicar que éstas permiten reducir cuotas de esfuerzo al momento de evaluar sincrónicamente, múltiples variables, sin asumir pérdidas de elevadas fracciones de información. En este sentido, los métodos multivariantes para el control de procesos se ocupan generalmente de medir la direccionalidad de las observaciones en un espacio multivariante, en contraste con los métodos ortodoxos, que sólo monitorean la magnitud y la variación de características individuales sin someter a escrutinio los cambios simultáneos entre ellas. De este modo, se superan ciertas restricciones operativas inamovibles que se imponen sobre los marcos de control convencionales.

Ciertas técnicas estadísticas multivariantes como el Análisis de Componentes Principales (ACP) han sido catalogadas como métodos de reducción de dimensionalidad y son muy utilizadas cuando hay una gran cantidad de variables de naturaleza cuantitativa. Su uso permite obtener un sistema transformado de coordenadas denominadas componentes principales, combinación lineal de las primitivas (Filho & Sant'Anna, 2016), amén de ser capaces de retener la variación global exhibida por los datos originales (Owen, 2014). Este tipo de métodos y sus extensiones, no sólo permiten la proyección de datos en un sub-espacio dimensional más parsimonioso, sino que posibilitan la implementación de metodologías alternativas para el diseño de cartas de control, como aquellas que se construyen para el análisis de variables latentes. Es sabido que los usos directos de las herramientas estándares para el control de operaciones son ineficientes para el abordaje de circunstancias anómalas. Así, por ejemplo, la existencia de variables colineales (lo que conduce a que la matriz de covarianza se convierta en una matriz casi singular y difícilmente invertible) y el incumplimiento de ciertos supuestos teóricos, produce un incremento de las tasas de falsas alarmas.

La metodología seguida en este caso de investigación tiene un énfasis puramente empírico, pues el objetivo ulterior es evaluar el estado actual de un proceso productivo en una unidad empresarial, localizada en el clúster petroquímico plástico cartagenero. La importancia de este análisis de caso queda refrendada, en tanto que llena algunas lagunas a nivel epistémico discernibles en estudios de caso similares, reportados en este espacio geográfico.

El uso de herramientas de monitorización que integran metodologías de síntesis de información como el exhibido en este artículo, ostentan propiedades teóricas deseables para ser utilizadas en entornos fabriles, a efectos de caracterización y evaluación sistemática de procesos en los que subyace una multiplicidad de características de calidad altamente correlacionadas y donde la cantidad y la calidad de los datos disponibles es aceptable.

2. Metodología.

La presente investigación se segmenta en tres apartes que serán explicados pormenorizadamente a continuación. En la fase inicial se procede a identificar y caracterizar el comportamiento de un conglomerado de variables del proceso productivo analizado. La definición y análisis de tales propiedades mensurables favorece a la explicación de un panorama general de la estabilidad general del proceso y las correlaciones entre cada una ellas. El conjunto de datos obtenidos vía selección muestral

aleatorizada se efectúa inicialmente y luego se realiza un somero análisis exploratorio de los mismos (Herrera, Herrera & Rahmer, 2017).

En la fase subsiguiente, se ejecuta el análisis de componentes principales utilizando los datos originarios. En términos generales, el análisis de componentes principales envuelve un conjunto de datos con observaciones sobre p variables numéricas para cada una de las n entidades o individuos. Estos valores definen vectores *n-dimensionales* $x_1, x_2 \dots x_n$ de modo que se busca una combinación lineal dada por: $\sum_{j=1}^{p} a_j \, x_j$ (Jolliffe & Cadima, 2016), cuyos coeficientes son obtenidos a partir de los vectores propios de la matriz de covarianza de los datos recabados.

Para la extracción secuencial de los componentes principales se dispone del procedimiento algorítmico catalogado como mínimos cuadrados parcialmente iterativos no lineales. La determinación de la cantidad óptima de las variables sustitutivas se realiza a través de la aplicación del método de validación cruzada de Krzanowski. Se subraya adicionalmente que, dentro de este análisis la extracción de componentes principales, se efectúa sobre variables tipificadas, puesto que las unidades de medida difieren.

Una vez concluido el análisis de componentes principales se construye una carta de control multivariante, siendo ésta la fase neurálgica del presente artículo, ya que permite detectar la ocurrencia de eventos anómalos y movilizaciones del proceso fuera del hiper-plano definido por el modelo de referencia. Bajo esta misma intencionalidad, se bosqueja un análisis auxiliar basado en la construcción de gráficos de monitorización para cada componente extraído. En todos los casos, los límites de especificación se computan a partir de la información histórica del proceso.

3. Resultados.

Para la adquisición de los datos se ha aplicado un plan de muestreo aleatorio simple sin reposición. En este procedimiento de selección muestral todos los sujetos tienen idéntica probabilidad de ser seleccionados y sin oportunidad para otra eventual selección.

Se cuenta inicialmente con un volumen bruto de datos en el que difícilmente se pueden identificar relaciones matemáticas. Por tanto, debemos realizar un tratamiento oportuno de la información recogida para su utilización en procedimientos estadísticos posteriores. En la Tabla 1 se listan 9 variables correspondientes a un proceso de fabricación de poliestireno expandido: Presión (P), Temperatura de vapor de agua (T), Resistencia a la difusión del vapor de agua (R), Densidad (D), Diámetro de pellets (Diam), Contenido de Pentano (PE), Concentración del polímero (PO), Concentración de oxígeno (O).

Se relacionan en la Tabla 1, los límites de especificación para cada variable, junto con algunos estadísticos descriptivos de dispersión y de tendencia central. Adicionalmente se muestran los respectivos factores de escala.

El hecho de que las métricas de las variables observadas difieren significativamente supondría a priori, una problemática patente a efectos de estimación de las componentes principales, pues estas características no son comparables entre sí. Además, aquellas características cuya varianza sea alta, irremediablemente dominarán las primeras sub-dimensiones. Esto es particularmente cierto para el caso de variables como Presión, Temperatura y Resistencia a la difusión, pues sus medias son significativamente superiores a las demás variables restantes. Para sortear esta circunstancia indeseable se procede a la estandarización de las variables originarias, de modo que ellas serán equiponderadas al instante de iniciar el análisis.

Tabla 1. Estadísticas descriptivas y escalamiento de las variables.

VARIABLE	Media	Desviación Estándar	Factor de Escala	Límite Inferior	Límite Superior
Presión (P)	19,510	9,659	9,659	20 Kpa	22 Kpa
Temperatura (T)	81,899	11,030	11,030	80ºC	100 ºC
Resistencia a la difusión (R)	22,366	5,620	5,620	20	40
Densidad Nominal (D)	9,086	5,332	5,332	9 Kg/m³	11 Kg/m³
Diámetro de pellets (Diam)	4,967	1,305	1,305	0,2 mm	1 mm
Contenido de Pentano (PE)	5,668	1,037	1,037	5%	7%
Concentración del polímero (PO)	0,233	0,050	0,050	0,03	0,07
Concentración de oxígeno (O)	0,967	0,328	0,328	0,92	0,98

Fuente: Elaboración propia.

3.1. Extracción de las Componentes Principales.

Intuitivamente, la finalidad esencial en este apartado es hallar un conjunto nuevo de direcciones ortogonales que definen la variabilidad máxima en términos de la estructura de varianza-covarianza de las variables originales. De esta manera, la información contenida en el conjunto completo de las componentes halladas es el equivalente exacto de la información original de los datos, conjeturándose que existen elementos redundantes y que sólo añaden dimensionalidad al problema estudiado (Montgomery, 2012). A través del ACP, los datos originales son proyectados en una representación dimensional mucho más compacta y parsimoniosa (Vanhatalo, Kulahci, & Bergquista, 2017), que posibilita el análisis y monitoreo de múltiples variables bajo un enfoque levemente simplificado.

Para el cálculo de la primera componente principal defínase:

$$y = e_1^{'}X, e_1^{'}e_1 = 1$$

De modo que:

$$Var(Y_1) \ = \ Var(l'\,X\,), \max_{l} Var(l'X\,) \ = \ Var(\,e_1^{'}\,X\,) \ = e_1^{'}\Sigma e_1$$

Para maximizar esta función multivariable sujeta a varias restricciones se dispone del método de Lagrange. Nótese que el vector desconocido e_1 proporciona la combinación lineal óptima. De este modo se tiene que:

$$\begin{Bmatrix} \max\{l'\Sigma l\} \\ l'l = 1 \end{Bmatrix} = \phi_1(l) = l'\Sigma l - \lambda(l'l)$$

Al derivar respecto a l y premultiplicar por 2 se obtiene que:

$$\Longrightarrow \frac{\partial \phi_1}{\partial l} = 2\Sigma l - 2\lambda l = 0 \Longrightarrow (\sigma - \lambda I)l = 0$$

Se supone que $\Sigma_{p\,x\,p}\ \ \lambda_1 \geq \lambda_2 \geq \cdots\ \lambda_p \geq 0$ con autovectores asociados $e_1, e_2, \ldots e_p$ y como $l'\Sigma l = \lambda l'l = 1, Var(l'\Sigma l) = \lambda$ y tomando $l = e_1$ que corresponde al mayor autovalor, se resuelve el problema de optimización propuesto, de modo que la primera componente principal queda expresada como $Y_1 = e_1'X$ y $Var(Y_1) = \lambda_1$.

La racionalidad del procedimiento anterior puede aplicarse para hallar la $r(+1)$ *-ésima* componente.

Para este caso se tiene que $Y_{r+1} = l'X$; $l'l$ =1; $l'\Sigma e_1 = 0, i = 1, \ldots r$

$$\phi_{r+1}(l) = l'\Sigma e_1 - (\lambda l'l - 1) - 2\sum_{i=1}^{r} v_i\, l'\Sigma e_1$$

Se demuestra que $\lambda_i \neq 0, i = 1, \ldots r$ el problema conduce a $v_i = 0, i = 1 \ldots r$. De esta manera el sistema que resuelve el problema de maximización viene dado por:

$$\{2\Sigma l - 2\lambda l = 0, \Sigma l - \lambda l = 0, \Sigma l - \lambda I = 0\}$$

Pueden ocurrir dos casos:

Si $\lambda_{r+1} \neq 0, \lambda = \lambda_{r+1}, l = e_{r+1}$y la $r(+1)$-*ésima* componente principal será:

$$Y_{r+1} = e_{r+1}'X$$

Y la varianza $Var(Y_{r+1}) = \lambda_{r+1}, \lambda$

En el otro caso posible cuando $\lambda_{r+1} = 0, \neq 0, i \neq r + 1$ se toma la combinación lineal de α_{r+1} y α_{r+1}y α_i para la que $\alpha_i \neq 0$ (De Ketelaere, Hubert & Schmitt, 2015; Severson, Molaro & Braatz, 2017).

3.2 Extracción de sucesivas componentes.

En lugar de ir obteniendo sucesivamente las componentes principales y resolver los sucesivos problemas de programación restringida y al final considerar globalmente todos, es preciso actuar globalmente desde un comienzo. Así, en lugar de ir resolviendo los sucesivos problemas de máximos condicionados, se parte de un resultado de maximización conocido.

Considérese el siguiente lema de maximización:

Sea A una matriz $p \times p$ definida positiva, con autovalores $\lambda_1, \lambda_2, \lambda_3 \ldots \lambda_p > 0$ y autovalores normalizados $e_1, e_2, e_3 \ldots e_p$ y sea x un vector $p \times 1$ arbitrario no nulo. Entonces se cumple que:

$$\max_x \frac{x'Ax}{x'x} = \lambda_1$$

Alcanzado en $x = e_1$

$$\max_x \frac{x'Ax}{x'x \neq 0} = \lambda_p$$

Alcanzado en $x = e_p$

$$\max_{x \perp e_1, e_2, e_3 \ldots e_p} \frac{x'Ax}{x'x \neq 0} = \lambda_{k+1}$$

Alcanzado en $x = e_{k+1}\ k = 1,2, \dots, p-1$

Sea $X = (X_1, \dots X_P)^{'}$ un vector aleatorio con matriz de covarianza conocida Σ definida positiva y real y sean (λ_i, e_i) los autovalores-autovectores Σ de $\lambda_1 \geq \lambda_2 \dots \geq \lambda_p > 0$ .La CP *i-ésima* Y_i antes definida viene dada por:

$$Y_i = e_i^{'}X + e_{i1}X_1 + \cdots + e_{ip}X_p, i = 1, \dots, p$$

Y se verifica:

$$Var(Y_i) = e_i^{'}\Sigma e_i = \lambda_i$$

$$Cov(Y_i, Y_j) = e_i^{'}\Sigma e_j = 0, i \neq j$$

Si hay autovalores iguales, pongamos λ_k, entonces los e_k asociados no son únicos, por lo que, en este caso las respectivas CP no son únicas.

Dada una matriz B, de dimensiones $p \times p$ con descomposición $\sum_{i=1}^{p} \lambda_i e_i^{'}$ sea la matriz $\mathbb{P} = (e_1, e_2, e_3 \dots e_p)$ conformada por columnas, con los autovectores normalizados e_i. Por consiguiente: $B = \mathbb{P}\Lambda\mathbb{P}^{'} = \sum_{i=1}^{p} \lambda_i e_i e_i^{'}$ siendo $\mathbb{P}^{'}\mathbb{P} = I$ y $\lambda = diag(\lambda_i, \lambda_2, \lambda_3 \dots \lambda_p)$. En el caso $\lambda_i > 0$ se puede utilizar esta descomposición para definir la matriz $B^{1/2} = \mathbb{P}\Lambda^{1/2}\mathbb{P}^{'} = \sum_{i=1}^{p} \sqrt{\lambda_i} e_i e_i^{'}$ dado que $B^{-1} = \mathbb{P}\Lambda^{-1}\mathbb{P}^{'} = \sum_{i=1}^{p} \frac{1}{\lambda_i} e_i e_i^{'}$

3.3 Validación cruzada de *Krzanowski*.

Para la extracción secuencial de los componentes principales se ha utilizado el procedimiento algorítmico denominando *mínimos cuadrados parcialmente iterativos no lineales*. En cada iteración ocurre un ajuste lineal de las columnas de *X* sobre un vector de puntuaciones t para la obtención de un vector de cargas p, seguida de una regresión lineal. Instantáneamente, se regresan las filas de *X* sobre el vector de cargas para la re-estimación de t hasta que se cumpla el criterio de convergencia predefinido. Este procedimiento algorítmico puede sintetizarse en los siguientes pasos (Quaglino & Vitelleschi, 2009):

1. Se selecciona una columna de la matriz de datos y se hace igual a un vector t_h

2. El vector t_h se utiliza para predecir la matriz *X* con el modelo de regresión:

$$X = t_h b_h^T + U$$

siendo $$\hat{b}_h^T = \frac{t_h^T X}{t_h^T t_h}$$

El estimador mínimo cuadrático de b_h^T y constituye la proyección de las columnas de *X* sobre la dirección de t_h, definida en el espacio de las *N* observaciones.

3. Se define el vector: $p_h = \hat{b}_h$

4. El vector p_h se normaliza, de manera que su longitud se hace igual a la unidad.

5. El vector p_h se utiliza para predecir la matriz X^T a partir del modelo visto en:

$$X^T = p_h b_h^T + F$$

6. Siendo el estimador mínimo cuadrático de b_h^T dado por:

$$\widehat{b}_h^T = \frac{t_h^T X^T}{p_h^T p_h}$$

7. La proyección de las filas de la matiz X sobre la dirección del vector p_h definida sobre el espacio de las K variables se obtiene de este modo.

8. Se define el vector $t_h = \hat{b}_h$

9. Se calcula la norma cuadrática de la diferencia entre los vectores obtenidos en el paso antecedente y en el inicial.

10. Se contrasta la norma cuadrática con el valor de tolerancia fijado en 10^{-6}. Si la diferencia es menor a este nivel se ha obtenido la h-ésima componente. En caso opuesto, se retorna al paso 2.

El procedimiento anteriormente descrito se aplica en reiteradas ocasiones para la obtención de *n-componentes*; sin embargo, para el retenimiento de una cantidad apropiada de sub-dimensiones y evitar el fenómeno de sobre-ajuste del modelo de PCA se aplica un método de validación cruzada como el de *Krzanowski*. Este esquema alternativo basado en el algoritmo de descomposición de valores singulares, tiene por objetivo determinar una cantidad óptima de componentes y facilitar la identificación de información redundante en la matriz de datos (Camacho & Ferrer, 2014). Para tal propósito, una prueba de significación estadística es aplicada sobre cada componente extraído a fin de determinar cuáles de ellos deben ser incorporados al modelo. El método propuesto asume que se pretende predecir los elementos x_{ij} de la matriz $\boldsymbol{X}$ a través del modelo:

$$x_{ij} = \sum_{t=i}^{m} v_{it}\, \gamma_t\, u_{tj} + e_{ij}$$

En este sentido, se predice el valor $\hat{x}_{ij}^{(m)}$ de x_{ij} $(i = 1, \dots g; j = 1 \dots e)$ para la selección de una cantidad de componentes m. La suma de los cuadrados de la diferencia entre los valores actuales y las estimaciones se cuantifica a través de la expresión:

$$PRESS(m) = \frac{1}{np} \sum_{i}^{n} \sum_{n}^{p} \left[\hat{x}_{ij}^{(m)} - x_{ij}\right]^2$$

Como se desea evitar sesgos predictivos, los x_{ij} no deben ser utilizados a efectos del cálculo de $\hat{x}_{ij}^{(m)}$ para cada i e j. El método de *Krzanowski* parte de la asunción de que la descomposición de valores singulares de la matriz X puede ser expresado como $X = UDV^T$.

Se define, subsiguientemente, el estadístico:

$$W_m = \frac{PRESS(m-1)PRESS(m)}{D_m} + \frac{PRESS(m)}{D_r}$$

siendo D_m los grados de libertad requeridos para ajustar los m-ésimos componentes y D_r los grados de libertad remanentes una vez ajustados los m-ésimos componentes W_m representa el incremento de la información predictiva en cada componente restante. El procedimiento estándar de validación cruzada consiste en subdividir X en varios grupos, eliminar cada grupo de los datos, evaluar los parámetros del predictor a partir de los datos restantes y predecir los valores eliminados (Dias & Krzanowski, 2006).

Siguiendo un criterio ortodoxo y sutilmente restrictivo, los componentes con mayor significancia estadística deberán poseer valores W_m mayores a la unidad. Incontestablemente, resultaría inapropiado detener la adición de componentes principales tan pronto como W_m, se localice por debajo de la unidad en la primera ocasión porque es una función de m no monotónica decreciente (Bógalo, 2012). Esto deja entrever que no existen cánones o criterios unificados para determinar qué proporción de variabilidad deba ser explicada por las componentes principales (Lazzarotto, Madalena, Chaves, & Texeira, 2016). Sin embargo, para propósitos de monitoreo, se seguirá esta pauta empírica.

En la Tabla 2 se compila un resumen de los resultados arrojados por el análisis de componentes principales. El parámetro R^2X cuantifica la bondad de ajuste del modelo al dar cuenta sobre la proporción de la variabilidad explicada por éste. Nótese que las tres componentes principales extraídas son capaces de explicar hasta aproximadamente el 93% de la variación total. Por otro lado, se muestra el valor de Q^2X, que es una expresión semejante al R^2X, excepto que evidencia un comportamiento menos inflacionario a medida que se acrecienta la complejidad del modelo. Ésta permite evaluar la capacidad predictiva del mismo, indicando que tal es aceptable, pues asume un valor próximo a la unidad.

El vector de eigenvalores, por su parte, puede ser conceptualizado como la magnitud de la varianza de las observaciones a lo largo de la dirección de su correspondiente autovector. Es un hecho fehaciente que el primer factor explica una fracción mayoritaria de la dispersión total de la nube y los sucesivos explican porciones cada vez más pequeñas de información de los datos originales.

Se postula, además, que la suma de las varianzas de las variables o inercia total de la nube de puntos es equivalente al sumatorio de las varianzas de las componentes principales. En este sentido, el porcentaje de inercia explicada por una componente i-ésima es:

$$\lambda_i / \sum_{i=1}^{p} \lambda_i = \lambda_i / \sum_{i=1}^{p} \mathrm{V}(\mathrm{x}_i)$$

siendo $\sum_{i=1}^{p} \mathrm{V}(\mathrm{x}_i) = traza(v)$, una expresión de la medida de variabilidad asociada a las variables originales (Pérez, 2004).

Tabla 2. Resumen de la extracción de las componentes.

COMPONENTE	R^2X	R^2X ACUM	AUTOVALOR	Q^2X	Q^2X ACUM	LÍMITE
1	0,7085	0,7085	5,6633	0,6051	0,6051	0,1347
2	0,1306	0,8391	1,0435	0,2059	0,6864	0,1525
3	0,0849	0,9240	0,6820	0,3362	0,7919	0,1763

Fuente: Elaboración propia.

A continuación, en la Tabla 3 se relacionan las saturaciones o cargas factoriales. En tanto que no son directamente observables, su determinación está parcialmente condicionada por criterios subjetivos. La carga de un factor cualquiera equivale a la correlación existente entre una variable original y un factor, obtenido por combinación lineal de las variables originales.

Ahora bien, sean las Componentes Principales Y_i asociadas al vector aleatorio X de matriz de covarianzas Σ conocida y sean (λ_i, e_i) sus autovalores-autovectores.

Para calcular ρ_{Y_i,X_k} ha de considerarse $h_k^{'} == (0,\ldots,0,1,0,\ldots,0)$, definido por $h_{ki} = \delta_{ki}$. Entonces:

$$Cov(Y_i, Y_j) = Cov\left(e_i^{'}X\ h_k^{'}X\right) = e_i^{'}\Sigma h_k = \lambda_i e_i$$

Con $Var(Y_i) = \lambda_i$, $Var(X_k) = \sigma_{kk}$, luego entonces:

$$\rho_{Y_i,X_k} = \frac{\lambda_i e_{ki}}{\sqrt{\lambda_i}\sqrt{\sigma_{kk}}} = \frac{e_{ki}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} = \frac{e_{ki}\sqrt{\lambda_i}}{\sigma_k} \quad i,k = 1 \ldots p$$

A partir de la expresión obtenida para ρ_{Y_i,X_k}, se infiere que la componente *k-ésima* del eigenvector e_i que proporciona la componente principal Y_i, cuantifica la relevancia absoluta que la variable original *k-ésima*, X_k sobre ésta. De manera que, cuanto mayor sea $|e_{ik}|$ mayor es la correlación entre X_k y la Y_i.

Se deriva de lo anterior que valores positivos próximos a 1 indican una fuerte correlación entre un componente y una variable cualquiera como sucede con las variables Presión y Contenido de Pentano en el primer factor; por el contrario, saturaciones negativas cercanas a la unidad revelan un acentuado grado de asociación negativa entre ambos.

Tabla 3. Matriz correlaciones entre componentes y variables.

	Componente 1	**Componente 2**	**Componente 3**
Presión (P)	**0,803**	0,587	0,066
Temperatura (T)	0,726	**0,743**	0,043
Resistencia a la difusión (R)	-0,895	0,186	**0,847**
Densidad Nominal (D)	-0,846	0,154	**0,764**
Diámetro de pellets (Diam)	**0,876**	-0,090	**0,825**
Contenido de Pentano (PE)	**0,883**	-0,211	0,321
Concentración del polímero (PO)	-0,783	**0,709**	-0,303
Concentración de oxígeno (O)	**0,877**	-0,087	-0,200

Fuente: Elaboración propia.

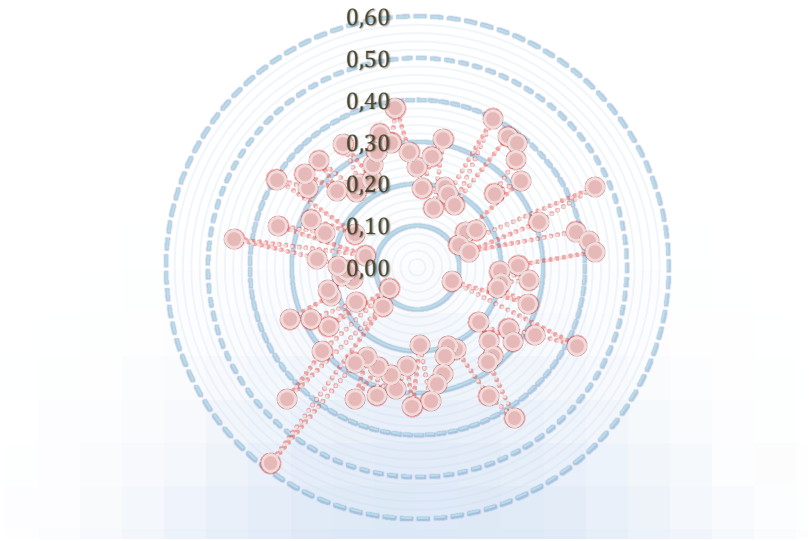
El parámetro estadístico utilizado para calcular la distancia euclidiana de las observaciones respecto al modelo se define por el ratio entre la distancia absoluta hacia al modelo S_i y la distancia normalizada respecto al mismo S_o tal y como se expresa en:

$$S_i = \sqrt{\sum_{k=1}^{K} e_{ik}^2 /(k - A)}$$

$$S_o = \sqrt{\sum_{i=1}^{N}\sum_{k=1}^{M} e_{ik}^2 /(k-P-P_0)(k-P)}$$

Como se observa en la Figura 1, las distancias normalizadas de las observaciones respecto al modelo oscilan entre 0.00 y 0.5, indicando que existen pocos outliers moderados.

Figura 1. Distancias normalizadas respecto al modelo.

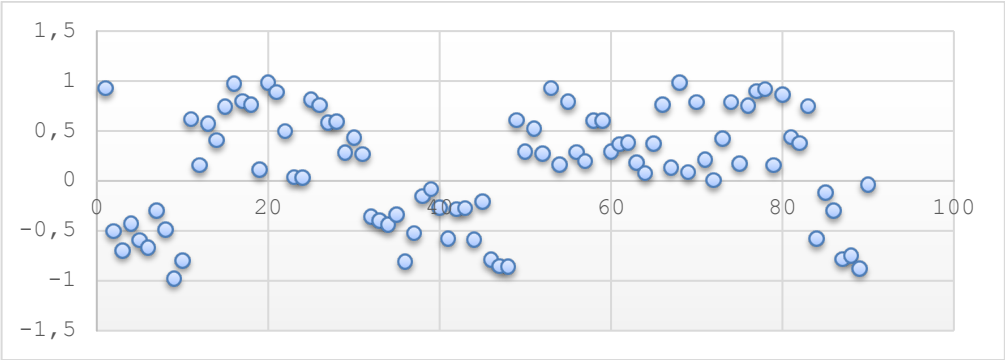


Fuente: Elaboración propia.

Al reducir la dimensionalidad y modelar los datos originales es que una vez extraída una cantidad óptima de variables sucedáneas es posible predecir, con cierto grado de exactitud los valores asumidos por el conglomerado de datos originales. El nivel de precisión predictiva incrementa en la medida que se incorporan más componentes al modelo, hasta cierto punto. Ahora bien, siempre existirá una discrepancia natural entre las observaciones originales y las predicciones generadas por el modelo de PCA. Tales, se denominan residuos y en condiciones ideales deberían exhibir valores bajos, sin desviaciones obvias de aleatoriedad.

Un somero examen visual de la Figura 2 permitirá concluir que existe un comportamiento no sistemático de los residuos, que además oscilan en un espectro reducido de valores, partiendo desde -1,0 hasta 1,0 sin incluir estas cantidades numéricas. Ello sugiere que el modelo de PCA se ajusta apropiadamente a los datos originales.

Figura 2. Gráfico de residuos.



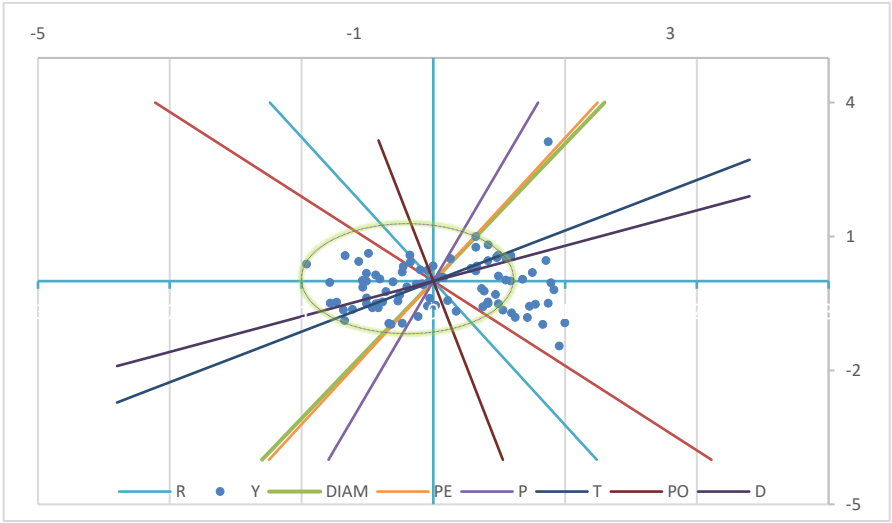
Fuente: Elaboración propia.

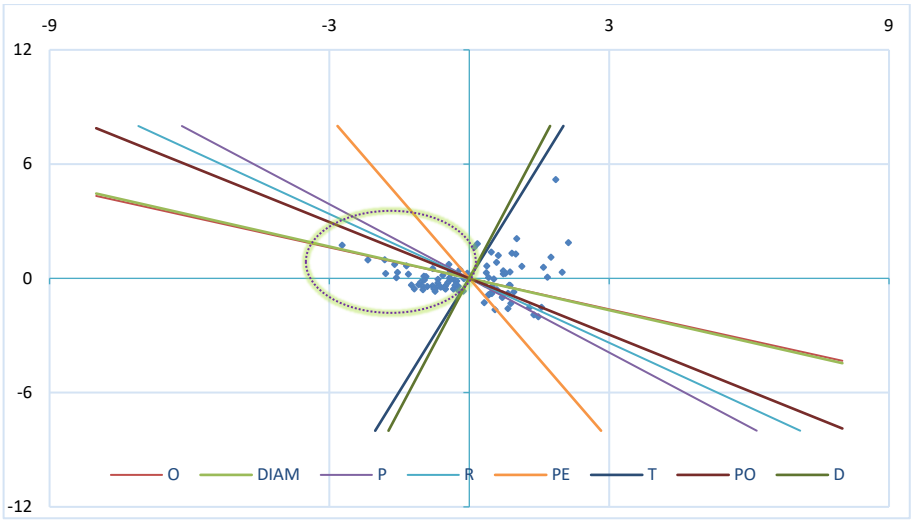
La Figura 3 no es nada más que una representación gráfica combinada de los *scores* estandarizados del componente principal 1 versus los componentes 2 y 3 junto con las proyecciones de las variables estandarizadas. La elipse que encierra la nube de puntos queda definida por:

$$e = \sqrt{(t_1/\sigma_1)^2 + (t_2/\sigma_2)^2}$$

Siendo σ_i la desviación estándar de los scores. Se observa que existen casos anómalos situados al exterior de los límites definidos, a pesar de que la nube de puntos se distribuye de forma homogénea.

Figura 3. Carta de control SPE.





Fuente: Elaboración propia.

3.2. Diseño de Carta de Control.

Ciertos diagramas de control multivariados están basados en la transformación de un vector $\mathbb{R}^p$ a un escalar a través de una función cuadrática que detecta la existencia de outliers o direcciones extremas.

Cuando se ha descartado la presencia de una estructura de correlación serial en las observaciones y el proceso verifica la condición de estacionariedad, los gráficos de control de esta naturaleza pueden ser implementados si la meta es el ajuste en la monitorización de procesos que exhiben alta dimensionalidad. A diferencia de los diagramas de control de T^2_{PCA} éstos son aptos para descubrir un cambio en la media del proceso si el desplazamiento es ortogonal a los primeros k vectores propios de la matriz de covarianza S (Phaladiganon, Bum, Chen & Jiang, 2013). La construcción de la carta de control que se reporta en la Figura 4, está basada en los cuadrados de los errores de predicción (SPE) dados por la expresión:

$$SPE_y = \sum_{k=1}^{i} (y_{nueva,i} - \hat{y}_{nueva,i})^2$$

El estadístico SPE representa la distancia ortogonal cuadrática de una nueva observación multivariada respecto al sub-espacio multivariante de las componentes principales. En otros términos, es una medida que cuantifica la falta de ajuste de la nueva muestra respecto al modelo que incluye los componentes retenidos, al detectar los datos proyectados que no son representados por éste. Para una muestra x_i los residuos denotados por r_i vienen dados por:

$$r_i = x_i - \hat{x}_i = x_i\left(I - P_l P_l^T\right)$$

Mientras que la magnitud de los residuos es equivalente a (Slišković, Grbić & Hocenski, 2012) la expresión dada por:

$$Q = \|r_i\| = r_i {r_i}^T = x_i\left(I - P_l P_l^T\right){x_i}^T$$

Siendo, r_i la i-ésima fila de la matriz residual que representa los errores de predicción del modelo, $P_l = [p_1, \ldots p_n]$ e I la matriz identidad o unitaria.

Los límites de control aproximados para Q pueden ser determinados siempre y cuando los datos puedan ser descritos por una distribución normal multivariante y los autovalores de la matriz covarianza sean conocidos. Así, el umbral Q_α responde a la siguiente expresión:

$$Q_\alpha = \theta_1 \left(\frac{z_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (1 - h_0)}{\theta_1^2} \right)^{\frac{1}{h_0}}$$

Siendo Q_α el límite superior de confianza para Q con un nivel de significancia α y z_α la estadística de la distribución normal estandarizada correspondiente al percentil superior $(1 - \alpha)$.

Entretanto,

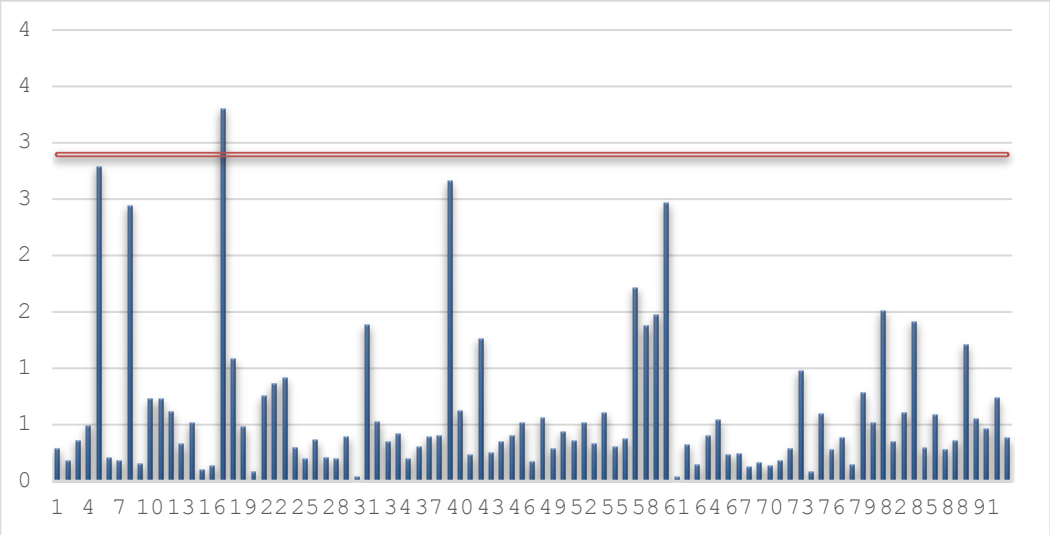
$$h_0 = 1 - \frac{2\theta_1 \theta_3}{3\theta_2^2}$$

y

$$\theta_1 = \sum_{j=k+1}^{p} \lambda_j^i$$

Siendo λ_j el valor propio de la matriz de covarianzas. Los errores entre los valores prospectados y los reales pueden medirse para determinar su significación estadística y caracterizar las actuales condiciones de operación. En caso tal que Q superase el umbral, es decir, que $Q > Q_\alpha$ entonces se sostendrá que existen indicios de la presencia de eventos que afectan la estructura de covarianza de X, tal y como se evidencia en la Figura 4, y por ello se concluye que en el proceso concurren fluctuaciones anómalas que no son explicadas por el modelo.

Figura 4. Carta de control SPE.



Fuente: Elaboración propia.

La presencia de muestras fuera de los límites del diagrama de control construidos a partir de los puntajes de los componentes o la existencia de tendencias y comportamientos sistemáticos en el gráfico

de scores muestra que el proceso está "fuera de control", es decir, que ha ocurrido una ruptura en la estructura de correlación del modelo estimado (Costa, Pedroza, Porto, Amorim & Lima, 2015).

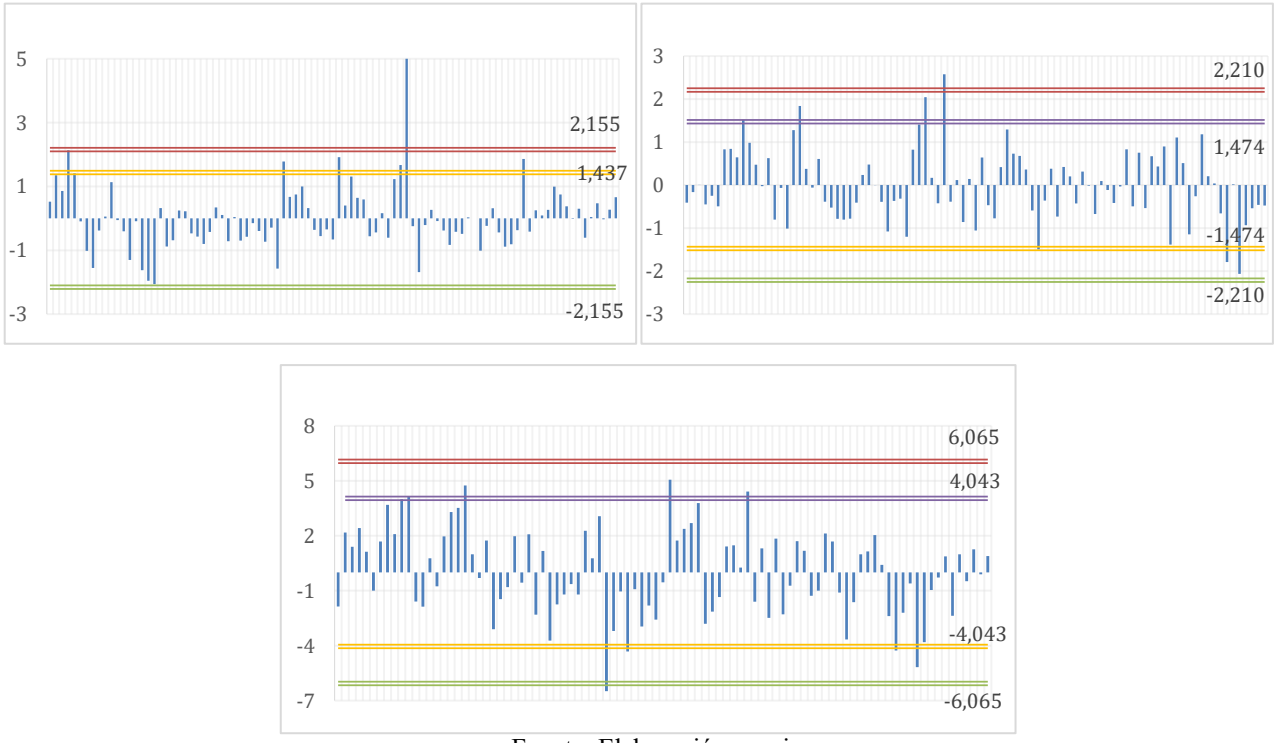
Ahora bien, puede esbozarse un análisis auxiliar fundamentado en los resultados provistos por las cartas de control univariadas construidas a partir de los *t-scores* de los componentes individuales, en las que se exhibe el comportamiento de las variables contribuyentes a la generación de señales alarmantes. Ello se percibe en la Figura 5. Bajo la asunción de normalidad, los límites de control para un nuevo $\vec{t}$-score en un intervalo de tiempo k y a un nivel de significancia α vienen dados por la expresión:

$$\pm t_{n-1,\frac{\alpha}{2}} S_{ref} (1 + \frac{1}{n})^{1/2}$$

Siendo S_{ref} la desviación estándar estimada para la muestra de $\vec{t}$-scores en un intervalo de tiempo denotado por k. La expresión $t_{n-1,\frac{\alpha}{2}}$ es el valor crítico de la variable estudentizada con $n-1$ grados de libertad y nivel de significancia $\frac{\alpha}{2}$ (Zude, 2008).

Es posible detectar en la Figura 5 que emergen señales fuera de control en las cartas de monitorización asociados a los componentes 1, 2 y 3. Por consiguiente, se afirma que el proceso está contaminado por la presencia de outliers, originando la ruptura temporal de la estabilidad del proceso. Se percibe que en ciertos intervalos de tiempo el proceso regresa a un estado donde sólo operan causas comunes de variabilidad. Empero, la existencia de comportamientos anómalos sigue siendo persistente. Esto permite concluir que los comovimientos de las variables contenidas en las sub-dimensiones halladas transgreden los límites naturales de variación para todos los casos evaluados. Por tal razón, convendría utilizar gráficos de control univariados que permitirían identificar plenamente las variables que contribuyen a la irrupción de anomalías en el proceso.

Figura 5. Carta de control para Scores.



Fuente: Elaboración propia.

4. Consideraciones finales.

En el presente caso de estudio se propuso un marco de monitorización para un proceso industrial de una empresa localizada en el clúster petroquímico plástico cartagenero. Se refrenda el hecho de que las cartas de control construidas a partir del análisis de componentes principales permiten evaluar el estado de sistemas productivos multidimensionales. Sin entrar en cargantes elucubraciones teóricas, se puede afirmar que el hallazgo de estas componentes permite reducir sustancialmente la dimensionalidad del problema estudiado, mediante la creación de estructuras de interdependencia que involucran las variables observadas.

Los resultados obtenidos corroboran que las tres componentes retenidas explican una fracción mayoritaria de la variabilidad original de la nube de datos y que la carta de control construida a partir de las sub-dimensiones registra la existencia de outliers o valores extremos. Se deriva entonces, la tesis de que el proceso no se halla en un estado de control estadístico, juicio revalidado al inspeccionar meticulosamente el comportamiento mostrado por los gráficos de control asociados a las componentes individuales. Por ello, es menester el direccionamiento de acciones destinadas a mejorar la consistencia del proceso productivo a fin de garantizar que las características de calidad exhiban un comportamiento relativamente homogéneo y sin traspasar los límites de variación natural. En un caso como éste sería provechosa la aplicación de herramientas de monitoreo univariadas, ya que una limitante de la metodología propuesta es su incapacidad para individualizar las variables que contribuyen sustancialmente a la generación de señales fuera de control en el corto plazo.

A la luz de los resultados hallados, puede anotarse que el abanico de posibilidades de aplicación de la metodología expuesta aquí es extenso. Se insta a abordar líneas de investigación focalizadas en la resolución de problemas investigativos no tan diferentes al aquí propuesto, en los que no se verifique íntegramente las hipótesis de partida que han de ser materializadas para garantizar una aplicabilidad óptima de las cartas de control multivariantes. Asimismo, se insta a diseñar herramientas de monitorización alternativas para el control de procesos fabriles en los que subyacen múltiples variables.

Agradecimientos

Se extiende especial agradecimiento al cuerpo científico del grupo C.I.P.T.E.C de la Fundación Universitaria Tecnológico Comfenalco por el apoyo técnico, financiero y académico dispensado para la consecución de los propósitos inherentes a esta investigación.

Referencias

Bógalo, J. (2012). Componentes subyacentes comunes en series temporales. Madrid: Universidad Nacional de Educación a Distancia. Recuperado de http://e-spacio.uned.es/fez/eserv/bibliuned:masterMatavanz-Jvbogalo/Documento.pdf

Camacho, J., & Ferrer, A. (2014). *Cross-validation in PCA models with the element-wise k-fold (ekf) algorithm: Practical Aspects. Chemometrics and Intelligent. Laboratory Systems*. Valencia: Elseiver.

Camacho, J., Pérez-Villegas, A., García-Teodoro, P., & Maciá-Fernández, G. (2016). PCA-based multivariate statistical network. *ScienceDirect, 59*, 118-137.

Costa, F., Pedroza, R., Porto, D., Amorim, M., & Lima, K. (2015). Multivariate Control Charts for Simultaneous Quality Monitoring of Isoniazid and Rifampicin in a Pharmaceutical Formulation Using a Portable Near Infrared Spectrometer. *Journal of the Brazilian Chemical Society, 26*(1), 64-73.

De Ketelaere, B., Hubert, M., & Schmitt, M. (2015). Overview of PCA-Based Statistical Process-Monitoring Methods for Time-Dependent, High-Dimensional Data. *Journal of Quality Technology, 47*(4), 318-335.

Dias, C., & Krzanowski, W. (2006). Choosing components in the additive main effect and multiplicative interaction (AMMI) models. *Scientia Agricola, 63*(2), 169-175.

Filho, D., & Sant'Anna, A. (2016). Principal component regression-based control charts for monitoring count data. *The International Journal of Advanced Manufacturing Technology, 85*, 1565-1574.

Herrera, J., Herrera, G., & Rahmer, B. (2017). Control Estadístico de Procesos para datos Correlados Serialmente. Un Caso de Estudio. *International Conference on Industrial Engineering and Operations Management* (pp. 891-904). Bogotá.

Jolliffe, I., & Cadima, J. (2016). Principal component analysis: a review and recent. *The Royal Society Publishing, 374*, 20150202. DOI: 10.1098/rsta.2015.0202.

Lazzarotto, E., Madalena, L., Chaves, A., & Texeira, L. (2016). Principal components in multivariate control charts applied to data instrumentation of dams. *Independent Journal of Management & Production, 7*(1), 17-37.

Montgomery, D. (2012). *Statistical Quality Control* (Séptima ed.). Jon Wiley & Sons.

Owen, J.A. (2014). Principal Component Analysis: Data Reduction and Simplification. *McNair Scholars Research Journal, 1*, 1-12.

Pérez, C. (2004). *Técnicas de Análisis Multivariante de Datos.* Madrid: Pearson Prentice Hall.

Phaladiganon, P., Bum, S., Chen, V., & Jiang, W. (2013). Principal component analysis-based control charts for multivariate nonnormal distributions. *Expert Systems with Applications: An International Journal, 40*(8), 3044-3054.

Quaglino, M., & Vitelleschi, M. (2009). Comparación de Métodos para el Tratamiento de Información Faltante en un Análisis de Componentes Principales sobre Datos Biológicos. *Revista de la Facultad de Bioquímica y Ciencias Biológicas, 13*, 115-124.

Rahim, M. A., Siddiqui, Y., & Elshafei, M. (2014). Integration of Multivariate Statistical Process Control and Engineering Process Control. *International Conference on Industrial Engineering and Operations Management* (pp. 470-480). Bali.

Severson, K., Molaro, M., & Braatz, R. (2017). Principal Component Analysis of Process Datasets with Missing Values. *Processes, 5*(3), 38.

Slišković, D., Grbić, R., & Hocenski, Ž. (2012). Multivariate statistical process monitoring. *Tehnicki Vjesnik-Technical Gazette, 19*(1), 33-41.

Vanhatalo, E., Kulahci, M., & Bergquista, B. (2017). On the structure of dynamic principal component analysis used in statistical process monitoring. *Chemometrics and Intelligent Laboratory Systems, 167*, 1-11.

Zude, M. (2008). *Optical Monitoring of Fresh and Processed Agricultural Crops.* New York: CRC Press Taylor & Francis Group.