Universidad Central “Marta Abreu” de Las Villas.

Facultad de Ingeniería Industrial y Turismo

Tesis en opción al título académico de Master en Ingeniería Industrial Mención Calidad.

Título: Contribución al Control Estadístico de Procesos Multivariado usando R

Autor: Ing. Edgar Santos Fernández

Tutor: Dr. Ing. Carlos Machado Osés

Curso 2012-2013

Pensamiento

“La estadística sin ciencia es incompleta,
la ciencia sin estadística es imperfecta” – K. V. Mardia”.

Agradecimientos

Agradecimientos

Estoy profundamente agradecido al equipo de R y a los contribuyentes por la creación y desarrollo del lenguaje y el entorno.

A los editores de Springer quienes hicieron posible este proyecto.

A mi amigo y mentor Michele Scagliarini por su colaboración y ayuda y por ser verdadero coautor de la función *mpci*.

Quero además agradecer a la Universidad Central de Las Villas y en especial a los profesores Dr. Carlos Machado Oses por ser tutor de esta obra y a Dr. Allán Aguilera Martínez.

Agradezco a la Empresa de Telecomunicaciones de Cuba S.A. (ETECSA) por todo el soporte durante este período.

De corazón agradezco a Carmen Fernández Ferrer por la inspiración. Además a Jessica Hernández por toda su ayuda y dedicación.

En fin a todos los que contribuyeron a este proyecto.

A mi padre Eugenio Santos Miyares (Cuqui) (1950-2011) quien está siempre presente en nuestros corazones.

Dedicatoria

Resumen

Resumen

En la actualidad el uso intensivo de la adquisición de datos y la utilización de computadoras on-line en el monitoreo de procesos ha conducido a un incremento de la ocurrencia de procesos industriales con dos o mas variables correlacionadas, en los cuales el control estadístico de procesos y los análisis de capacidad deben ser llevados a cabo usando metodologías multivariantes. Aún cuando las capacidades de cómputo se han incrementado exponencialmente la disponibilidad de aplicaciones para resolver problemas prácticos en el Control Estadístico de Procesos Multivariado (MSQC) son en la actualidad limitadas. Para resolver este problemática se realizó un análisis exhaustivo de la literatura y posteriormente se programaron y publicaron en forma de paquetes la mayoría de las técnicas encontradas en esta disciplina. Además se ilustraron detalladamente el uso de dichas herramientas mediante del uso de las funciones desarrolladas y la utilización de varios sets de datos didácticos. Finalmente se presentaron los casos de estudio de dos disciplinas en las cuales dichas técnicas contribuyen a establecer el control estadístico.

Índice

Índice

Introducción

En la actualidad el uso intensivo de la adquisición de datos y la utilización de computadoras on-line en el monitoreo de procesos ha conducido a un incremento de la ocurrencia de procesos industriales con dos o más variables correlacionadas, en los cuales el control estadístico de procesos y los análisis de capacidad deben ser llevados a cabo usando metodologías multivariantes.

Desafortunadamente, a pesar del incremento en las capacidades de cómputo, en el Control Estadístico de Procesos Multivariado (MSQC) las soluciones de software son limitadas o presentan restricciones de uso para manejar los problemas de la industria y promover la instrucción académica; resultando esta la *situación problémica* de esta investigación.

El *problema científico* resulta: ¿Cómo desarrollar varias funciones computacionales que posibiliten la disponibilidad on-line y de forma gratuita de las principales herramientas en el MSQC así como realizar una exposición descriptiva en el uso de las mismas?.

El objetivo general de la investigación es suplir el vacío relativo a la disposición de herramientas para llevar a cabo técnicas de MSQC y desarrollar una guía de utilización de las mismas.

Los *objetivos específicos* resultan los siguientes:

1- Realizar una revisión bibliográfica extensiva que permita conocer el estado del arte en esta disciplina.
2- Desarrollar aplicaciones en lenguaje R que permitan llevar a cabo las herramientas más importantes del MSQC y ponerlas a disposición de la comunidad de usuarios.
3- Realizar una descripción del uso de las herramientas desarrolladas en la medida en que se introducen los aspectos teóricos.

4- Exponer varios casos de estudio prácticos en el que se aplican la mayoría de las técnicas abordadas usando las herramientas desarrolladas.

Como hipótesis de la investigación tenemos que si se implementan las principales técnicas disponibles en la literatura, entonces la comunidad de usuarios tendrá a su disposición un conjunto de herramientas gratuitas que permita tanto el uso en la industria como en la academia. Lo anterior será validado entre otras vías si una vez publicados los paquetes se reporta el uso de los mismos y de acuerdo el nivel de descarga por los usuarios.

El valor práctico de la investigación subyace en que en la actualidad los problemas relativos a MSQC son abordados en la literatura fundamentalmente de forma teórica. Mientras que valor económico se sustenta sobre la disponibilidad de aplicaciones de forma gratuita y con facilidad de descargar.

Se pretende que el tipo de investigación a realizar sea descriptivo, es decir ir ilustrando la utilización de las técnicas desarrolladas en la medida en que se introducen los aspectos teóricos.

La opción del R como lenguaje y entorno para llevar a cabo la investigación se debe a que se ha convertido en la “lingua franca” o universal de análisis de datos, es fácil de usar, es de código abierto, gratuito, multiplataforma y un software muy flexible.

Además R cuenta con una gran comunidad de usuarios y está creciendo su uso empresarialmente y académicamente.

Los ejemplos son presentados de forma clara y sencilla y contienen los fragmentos de código utilizados.

En la investigación se usa el paquete de R llamado MSQC elaborado por el autor y disponible en :http://www.cran.r-project.org/package=MSQC en el CRAN (the

Comprehensive R Archive Network) de R. Contiene diez sets de datos didácticos los cuales son usados en los diferentes capítulos.

Además se utiliza el paquete MPCI disponible en http://www.cran.r-project.org/package=MPCI elaborado también por el autor en conjunto con el Profesor Michele Scagliarini, Universita di Bologna, Italia. La implementación de esta aplicación fue publicada en (Santos-Fernández,Scagliarini 2012).

Una versión extensiva de esta investigación fue publicada en (Santos-Fernández 2013). En esta fuente se pueden encontrar más detalles de las técnicas abordadas así como su utilización.

Introducción

Capítulo 1

Capítulo 1 Gráficos de Control Multivariados

Con las mejoras en los sistemas de adquisición de datos es usual tratar con procesos con más de una característica de calidad a ser monitoreada. Una práctica común es controlar la estabilidad del proceso utilizando gráficos de control. Esta práctica incrementa la probabilidad de falsas alarmas de fuentes de variación.

Por tanto el análisis debe ser llevado acabo con un enfoque multivariado, es decir, las variables deben ser analizadas conjuntamente, no de forma independiente.

Este capítulo presenta la distribución normal multivariante, la estructura de datos, la función mult.chart que posibilita la ejecución en R de las más usadas graficas de control multivariantes como:

- elipsoide de control o gráfico χ2.
- gráfico T^2 o gráfico de Hotelling.
- gráfico de Media Móvil con Pesos Exponenciales (Multivariate Exponentially Weighted Moving Average (MEWMA)
- gráfico de Suma Acumulada (Multivariate Cumulative Sum (MCUSUM))
- gráficos basados en Análisis de Componentes Principales (Principal Components Analysis (PCA))

1.2 Estructura de Datos

Con el objetivo de proveer una mejor comprensión de los aspectos teóricos, en esta sección se presenta un sumario de la estructura de datos utilizada en todos los métodos.

Como se muestra en la Fig. 1.1, casi todos los problemas que se abordan tratan con k muestras de tamaño n tomadas de p características de calidad.

$$
\begin{array}{cc|cccc}
 & & \multicolumn{4}{c}{Characteristic\ (j)} \\
 & & 1 & 2 & \ldots & p \\
 & & \multicolumn{4}{c}{sample\ size(n)} \\
 & 1 & x_{111}\, x_{211} \cdots x_{n11} & x_{121}\, x_{221} \cdots x_{n21} & \cdots & x_{1p1}\, x_{2p1} \cdots x_{np1} \\
Sample & 2 & x_{112}\, x_{212} \cdots x_{n12} & x_{122}\, x_{222} \cdots x_{n22} & \cdots & x_{1p2}\, x_{2p2} \cdots x_{np2} \\
(k) & \vdots & \vdots & \vdots & \ddots & \vdots \\
 & m & x_{11m}\, x_{21m} \cdots x_{n1m} & x_{12m}\, x_{22m} \cdots x_{n2m} & \cdots & x_{1pm}\, x_{2pm} \cdots x_{npm}
\end{array}
$$

Fig. 1.1 Representación gráfica de la estructura de datos

Donde x_{ijk} es la i^{th} observación de la j^{th} característica de calidad en la k^{th} muestra.

Frecuentemente los parámetros de la distribución (μ y σ) son desconocidos y deben ser estimados a través de $\bar{x}$ y S respectivamente siendo calculados como sigue:

$$\bar{\bar{x}}_j = \frac{\sum_{k=1}^{m} \bar{x}_{jk}}{m} \text{ (1.1) donde } \bar{x}_{jk} = \frac{\sum_{i=1}^{n} x_{ijk}}{n} \text{ (1.2)}$$

Existen caso en que las muestras están compuestas por una sola observación. Este caso es denominado de observaciones individuales y es presentado más adelante.

Por otro lado, S es estimado:

$$S = \begin{matrix} \bar{S}_1^2 & \bar{S}_{12} & \cdots & \bar{S}_{1p} \\ \bar{S}_{12} & \bar{S}_2^2 & \cdots & \bar{S}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{S}_{1p} & \bar{S}_{2p} & \cdots & \bar{S}_2^2 \end{matrix} \text{ (1.3)}$$

donde la diagonal de los elementos son las varianzas y el resto las covarianzas, siendo:

$$\bar{S}_j^2 = \frac{\sum_{k=1}^{m} S_{jk}^2}{m} \text{ (1.4) con } S_{jk}^2 = \frac{\sum_{i=1}^{n} \left(x_{ijk} - \bar{x}_{jk}\right)^2}{n-1} \text{ (1.5)}$$

y $\bar{S}_{jl} = \frac{\sum_{k=1}^{m} S_{jlk}}{m}$ (1.6) con $j \neq l$ siendo

$$S_{jlk} = \frac{\sum_{i=1}^{n} \left(x_{ijk} - \bar{x}_{jk}\right)\left(x_{ijk} - \bar{x}_{lk}\right)}{n-1} \text{ (1.7)}$$

El vector de medias (Xmv) es obtenido en R como

```
x.jk <- apply(x1, 1:2, mean)
```

primero calculando la media de cada muestra, y luego utilizando la función colMeans

```
Xmv <- colMeans(x.jk)
```

Con respecto a la matriz de covarianza muestral, esta puede ser lograda directamente usando la función covariance incluida en el paquete MSQC

S <- covariance(x)

1.3 La función mult.chart

La ejecución de los gráficos de control multivariantes puede ser llevada a cabo en R usando la función *mult.chart* la cual es una función general que permite desarrollar las propuestas más aceptadas como:

- gráfico T^2 o gráfico de Hotelling.
- gráfico de Media Móvil con Pesos Exponenciales (Multivariate Exponentially Weighted Moving Average (MEWMA)
- gráfico de Suma Acumulada (Multivariate Cumulative Sum (MCUSUM))
- gráfico T^2 o gráfico de Hotelling
- gráfico de Suma Acumulada (Multivariate Cumulative Sum (MCUSUM)) propuesto por (Crosier 1988)
- gráfico de Suma Acumulada (Multivariate Cumulative Sum (MCUSUM)) propuesto por (Pignatiello,Runger 1990)

La selección del tipo de ficha a usar es realizado especificando el argumento type= "t2", "mewma", "mcusum" o "mcusum2" en el mismo orden como fueron introducidos anteriormente

Usando la ayuda de la función

- help(package = "MSQC")

se obtiene una descripción bien detallada.

En la función x debe ser una matriz o un arreglo y conjuntamente con el argumento type son los únicos obligatorios

Otra importante funcionalidad es la fase que puede ser I o II(siendo I por defecto) y el nivel de significación (alpha) fijado en 0.01.

Finalmente la función devuelve:

- el estadístico T^2
- el Límite de Control Superior (UCL)

- la matriz de covarianza (S)
- el vector de medias (Xmv)
- y si algún punto cae fuera del UCL y su descomposición.

La ejecución de esta función toma solamente unas pocas centésimas de segundo lo cual puede ser verificado haciendo:

> system.time(mult.chart(dowel1, type = "chi", alpha = 0.05))

1.4 El Mapa de Contorno y el Gráfico de Control χ^2

En la distribución normal multivariante la densidad es descrita por un elipsoide con ejes en dirección a los vectores propios (e) de la matriz de covarianza, fijando μ como el origen y con longitud $\pm c\sqrt{\text{a}_j}e_j$ (1.8),

siendo

$$(x-«)'\Sigma^{-1}(x-«)=c^2 \quad (1.9)$$

Si x sigue $N_p(«,\Sigma)$ entonces

$(x-«)'(\Sigma)^{-1}(x-«)$ sigue ¡ $^2_{\sim,p}$.

Por tanto,

$$(x-«)'\Sigma^{-1}(x-«)\leq \text{¡}^2_{\sim,p} \quad (1.10)$$

Para ilustrar la construcción de un elipsoide de confianza considerar el dataset llamado dowel que contiene 40 muestras de dos características de calidad correlacionadas (diámetro y longitud) registradas en el proceso de manufactura de una espiga

> data("dowel1")

La construcción de la elipse de control es como sigue:

Fijando el nivel de significación.

> alpha <- 0.05 y haciendo

> p <- ncol(dowel1)

Entonces son estimados el vector de medias y la matriz de covarianza:

> Xmv <- colMeans(dowel1)

> S <- covariance(dowel1)

Entonces son obtenidos:

$\text{«}' = \begin{bmatrix} 0.50 & 1.00 \end{bmatrix}$ y $\Sigma = \begin{bmatrix} 3.70\text{e-}05 & 5.00\text{e-}05 \\ 5.00\text{e-}05 & 3.04\text{e-}04 \end{bmatrix}$

El cálculo de los valores y vectores propios es desarrollado usando la función eigen:

> DDe <- eigen(S)$values

> Ue <- eigen(S)$vectors

Entonces se obtienen:

$\text{a}' = \begin{bmatrix} 4.39\text{e-}04 & 3.02\text{e-}05 \end{bmatrix}$,

$e_1' = \begin{bmatrix} 0.22 & -0.98 \end{bmatrix}$ y

$e_2' = \begin{bmatrix} -0.98 & 0.22 \end{bmatrix}$

Entonces ploteando el origen dado por Xmv. (a 0.50, 1.00) con los ejes respectivos:

> plot(Xmv[1], Xmv[2], xlim = c(0.46,0.54), ylim = c(0.95,1.06), xlab = "diameter", ylab = "length",pch = 3)

La dirección de los ejes está dada por los vectores propios:

> inc <- atan ((Xmv[2] + Ue[2,1] - Xmv[2]) / (Xmv[1] + Ue[1,1] - Xmv[1]))

Calculando la longitud relativa a los ejes x y y :

> b <- (sqrt(DDe[1]) * sqrt(qchisq(1 - alpha,p))) * sin(inc)

> a <- (sqrt(DDe[1]) * sqrt(qchisq(1 - alpha,p))) * cos(inc)

> d <- (sqrt(DDe[2]) * sqrt(qchisq(1 - alpha,p))) * sin(inc)

> c <- (sqrt(DDe[2]) * sqrt(qchisq(1 - alpha,p))) * cos(inc)

Finalmente, son trazados los ejes usando:

> arrows(Xmv[1], Xmv[2], Xmv[1] + a, Xmv[2] + b)

> arrows(Xmv[1], Xmv[2], Xmv[1] - a, Xmv[2] - b)

> arrows(Xmv[1], Xmv[2], Xmv[1] - d, Xmv[2] + c)

> arrows(Xmv[1], Xmv[2], Xmv[1] + d, Xmv[2] - c)

Y la elipse resulta conectando los ejes por los extremos. Afortunadamente es relativamente fácil dibujar una elipse en R utilizando el algoritmo:

```
> angle <- seq(0, 2 * pi, length.out = 200)
> ch <- cbind(sqrt(qchisq(1 - alpha,2)) * cos(angle), sqrt(qchisq(1 - alpha,2)) * sin(angle))
> lines(t(Xmv - ((Ue %*% diag(sqrt(DDe))) %*% t(ch))),type = "l")
```

La Fig. 2.3 (a) muestra el resultado:

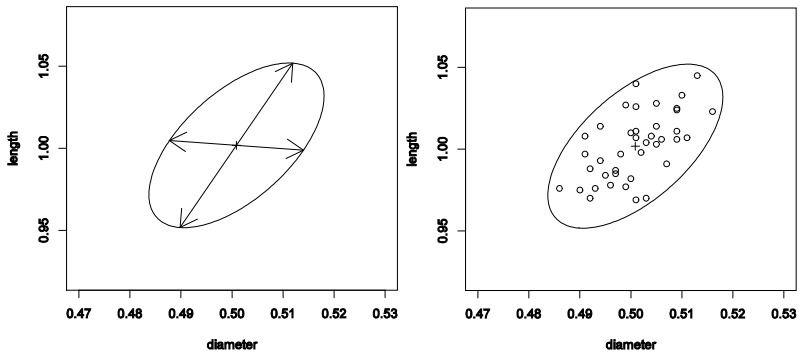

Fig. 1.2 (a) Elipse de control con los ejes para el dataset dowel1. (b) Scatter plot del dataset dowel1 con la elipse de control

Este procedimiento es conocido además como elipse de confianza. La Fig. 2.3 (b) muestra la adición de los puntos del arreglo dowel1:

```
> points(dowel1)
```

Como no se obtiene ningún punto fuera de la elipse, no existe evidencia de causas especiales, consecuentemente el proceso está en control. Notar que si se plotean los límites de los gráficos de control univariantes, la diferencia que existe entre ambas áreas. De hecho, cuatro puntos caen fuera de esta área.

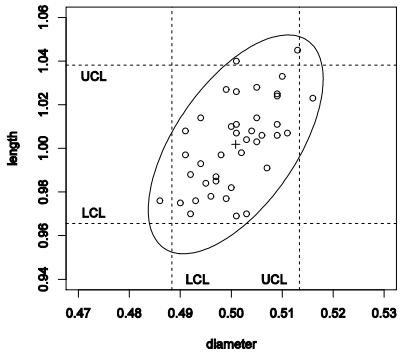

Fig. 1.3 Scatter plot del dataset dowel1 con la elipse de confianza y los limites univariantes.

La identificación de los puntos fuera del elipsoide es uno de las mayores limitaciones de esta herramienta, aunque esto puede ser resuelto insertando el número de la muestra en el gráfico cuando la cantidad de puntos no es excesivamente grande.

Otra desventaja es la complejidad para construir la elipse cuando p > 2 lo cual puede ser resuelto usando gráficos de control χ2, resultante de graficar el estadístico:

$$n(x-«)^{'}(\Sigma)^{-1}(x-«) = ¡^{2}_{\sim,p} \quad (1.11)$$

donde n es el tamaño de muestra y UCL el límite de control superior.

$$UCL = ¡^{2}_{\sim,p} \quad (1.12)$$

Cuando μ y Σ son estimados a través una muestra lo suficientemente grande, entonces puede ser utilizado este gráfico aunque los parámetros sean desconocidos.

Utilizando la función mult.chart

> mult.chart(dowel1, type = "chi", alpha = 0.05)

Entonces la función devuelve:

```
[1] "Chi-squared
Control Chart"
$ucl
[1] 6
$t2
[1,] 1.61
 [2,] 0.30
…
 [39,] 1.58
[40,] 1.64
$Xmv
[1] 0.50 1.00
$covariance
             [,1]         [,2]
[1,] 4.91e-05 8.59e-
05
[2,] 8.59e-05 4.20e-
04
```

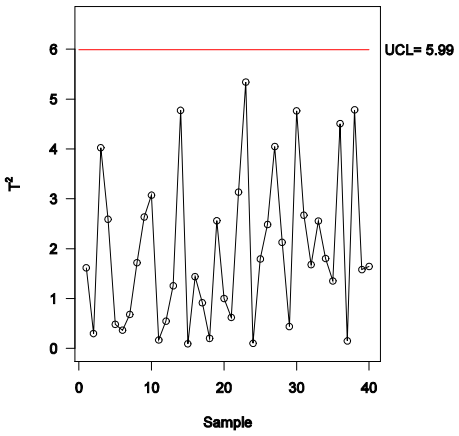

Fig. 1.4 χ^2 control chart for the dowel1 dataset.

Mostrando resultados similares al de la elipse de control. Una ventaja de esta gráfica es que permite la evolución de las muestras a través del tiempo.

A continuación se brinda una guía sobre el uso de las fases. Usualmente, estos estudios son divididos en dos fases bien diferentes una de la otra.

Fase I: en esta etapa es aplicado un análisis retrospectivo para valorar si el proceso está o no en control una vez que la primera muestra es tomada. Este tipo de estudio es usado cuando las fichas de control son establecidas por primera vez y con el objetivo de traer el proceso a control estadístico. Debe puntualizarse que se requiere un profundo análisis antes de establecer el estado de control.

Fase II: en esta fase las fichas son empleadas para verificar si el proceso permanece en control. En este caso la variabilidad es monitoreada usando la media y covarianza de la Fase I. Para más detalles ver (Woodall 2000)

Entonces, usando la media y la matriz de covarianza de la primera etapa es posible controlar producciones futuras (Fase II) almacenadas en el arreglo dowel2.

Empleando la elipse de control calculada en la Fase I, solo es necesario añadir los puntos de la Fase II

```
> data("dowel2")
> points(dowel2, pch = 4)
```

El argumento pch = 4 permite diferenciar los puntos. Un punto cae fuera del 95^{th} elipse indicando la presencia de causas especiales en el proceso.

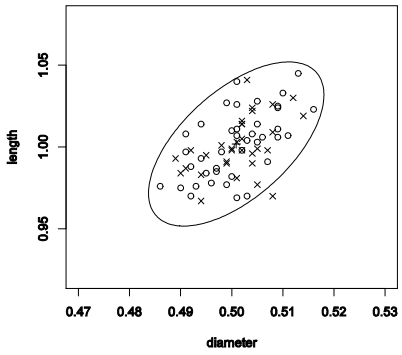

Fig. 1.5 Elipse de confianza en la Fase II para el dataset dowel2.

Por el contrario usando la gráfica de control χ^2 .

Como el vector de medias y la matriz de covarianzas son usados como parámetros de la distribución:

```
> vec <- (mult.chart(dowel1, type = "chi", alpha = 0.05)$Xmv)
> mat <- (mult.chart(dowel1, type = "chi", alpha = 0.05)$covariance)
```

Finalmente son pasados a la función mult.chart

```
> mult.chart(dowel2, type = "chi", Xmv = vec, S = mat, alpha = 0.05)
```

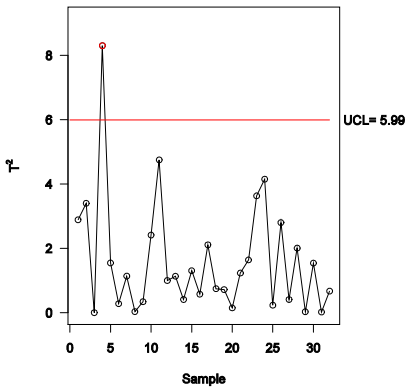

Fig. 1.6 Gráfico de control χ^2 para la Fase II del dataset dowel2.

La cuarta muestra cae fuera del UCL, como consecuencia, existen evidencias de causas especiales incidiendo, luego el proceso está fuera de control.

1.5 Gráfico de Control de Hotelling (T^2)

El origen de este gráfico de control data del estudio de Harold Hotelling quien aplicó este método al lanzamiento de las bombas en la Segunda Guerra Mundial.

El procedimiento de (Hotelling 1947) se ha convertido sin dudas en el más aplicado en el control de procesos mutivariante y resulta la extensión multivariada de la ficha de control de Shewhart.

Es frecuente en la práctica que los parámetros μ y Σ sean desconocidos y por consecuencia deben ser estimados a través de los estimadores insesgados $\bar{x}$ y S. Basado en la generalización del estadístico t de Student:

$t = \frac{\bar{x} - «}{S/\sqrt{n}}$ (1.13) haciendo $t^2 = \frac{(\bar{x} - «)^2}{S^2/n} = n(\bar{x} - «)\left(S^2\right)^{-1}(\bar{x} - «)$ (1.14)

Entonces la generalización resulta:

$T^2 = n(\overline{X} - \overline{X})'\left(S\right)^{-1}(\overline{X} - \overline{X})$ (1.15)

siendo $\overline{X}$ y S el vector de medias y la matriz de covarianza respectivamente.

El estadístico T^2 sigue una distribución F de Fisher con p y (mn-m-p+1) grados de libertad. Por tanto para establecer el control en la Fase I el límite de control superior resulta: $UCL = \frac{p(m-1)(n-1)}{mn-m-p+1} F_{\sim,p,mn-m-p+1}$ (1.16)

Mientras que para monitorear futuras observaciones (Fase II) el límite está dado por

$$UCL = \frac{p(m+1)(n-1)}{mn-m-p+1} F_{\sim,p,mn-m-p+1} \quad (1.17)$$

Acá en (2.25) el número de muestras (m) se refiere a las muestras preliminares tomadas para establecer en control en la Fase I. Notar que carece de límite de control inferior análogamente al gráfico χ^2.

Esta ficha es empleada en estudios introductorios y posee un buen rendimiento para largas variaciones de la media.

De acuerdo con (Lowry,Montgomery 1995) la aplicación de este gráfico requiere entre 2 y 10 características de calidad tomando más de 20 muestras de tamaño 2,3 o 10. Aunque estos valores son con frecuencia limitados por la naturaleza propia del problema.

A continuación se muestra la construcción de esta gráfica.

Usando el set de datos carbon, el cual contiene datos de tres características de calidad medidas en fibras de carbono (diámetro interior, grosor y longitud en pulgadas)

El cálculo en R resulta:

```
> data("carbon1")
> mult.chart(type = "t2", carbon1)
```

A continuación se muestra la salida:

Por ejemplo, si se desea obtener solamente el estadístico T^2 teclear solamente:

```
> mult.chart(type = "t2", carbon1)$t2
```

```
[1] "Hotelling Control Chart"
$ucl
[1] 11.35
$t2
      [,1]
 [1,] 4.99
 [2,] 4.66
...
[29,] 3.54
[30,] 1.40
$Xmv
[1]  0.99  1.04 49.98
$covariance
       [,1]   [,2]   [,3]
[1,] 0.0025 0.0036 0.0067
[2,] 0.0036 0.0140 0.0100
[3,] 0.0067 0.0100 0.0590
```

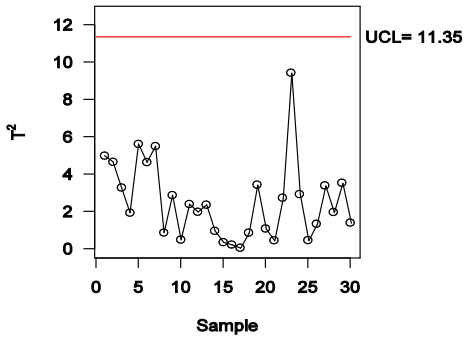

Fig. 1.7 Gráfico de Control Hotelling para el set carbon1.

Para ver los casos específicos de gráfico de control para observaciones individuales, sobre la interpretación de las respuestas y descomposición; así como para encontrar el uso de fichas de control usando Análisis de Componentes Principales ver (Santos-Fernández 2013). Además, en esa misma fuente se expone un capítulo al chequeo de los supuestos tanto de multinormalidad como aleatoriedad a los cuales se hace referencia en próximos capítulos.

1.6 El gráfico de Varianza Generalizada

De la misma forma que en los gráficos univariados la media es monitoreada simultáneamente con gráficos de dispersión, en escenarios multivariados resulta extremadamente útil monitorear la variabilidad del proceso. Esto se debe a que en las cartas multivariadas de Shewhart se asume que la dispersión permanece constante. Esta hipótesis debe ser chequeada en la práctica.

Hasta la fecha varios métodos han sido propuestos para monitorear la variabilidad pero sin dudas el gráfico de varianza generalizada es el más aceptado. Para más detalles ver por ejemplo (Alt 1985) o (Montgomery 2004)

Esta ficha resulta de plotear el determinante de la matriz de covarianza a través de límites naturales superior e inferior.

Cuando la matriz de covarianza Σ es conocida los parámetros resultan:

$$UCL = |\Sigma|\left(b_1 + 3b_2^{1/2}\right) \quad (1.18)$$

$$CL = b_1|\Sigma| \quad (1.19)$$

$$LCL = \max\begin{cases} |\Sigma|\left(b_1 - 3b_2^{1/2}\right) \\ 0 \end{cases} \quad (1.20)$$

donde: $b_1 = \dfrac{1}{(n-1)^p}\prod_{j=1}^{p}(n-j)$ (1.21) y

$$b_2 = \frac{1}{(n-1)^{2p}}\prod_{j=1}^{p}(n-j)\left[\prod_{i=1}^{p}(n-i+2) - \prod_{i=1}^{p}(n-i)\right] \quad (1.22)$$

Con frecuencia Σ es desconocida por lo que debe ser estimada a través de S basada en la relación:

$|S| = b_1|\Sigma|$ (1.23).

Por tanto los parámetros resultan:

$$UCL = \frac{|S|}{b_1}\left(b_1 + 3b_2^{1/2}\right) \quad (1.24)$$

$$CL = |S| \quad (1.25)$$

$$LCL = \max\left\{\frac{|S|}{b_1}\left(b_1 - 3b_2^{1/2}\right)\right. \quad (1.26)$$
$$\left.0\right\}$$

El paquete MSQC contiene una función específica llamada gen.var que permite desarrollar esta herramienta en R. La función solo requiere como argumento un arreglo de $p \times m \times n$ dimensiones.

Por ejemplo:

> gen.var(carbon1)

obteniéndose:

```
[1] "Generalized Variance
Control Chart"
$‘Upper Control Limit‘
[1] 4.3e-06
$‘Lower Control Limit‘
[1] 0
$stat
        [,1]
 [1,] 3.1e-07
```

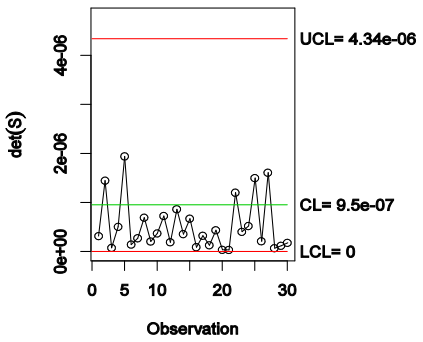

Fig. 1.8 Gráfico de la Varianza Generalizada.

[2,] 1.4e-06

...

[29,] 1.2e-07

[30,] 1.8e-07

1.7 Gráfico de Media Móvil con Pesos Exponenciales (MEWMA)

El Gráfico de Media Móvil con Pesos Exponenciales (MEWMA) es la extensión natural de la carta EWMA propuesta por (Roberts 1959). Esta técnica fue introducida por (Lowry et al. 1992) y es más sensible detectando cambios no aleatorios en el proceso.
El gráfico MEWMA posee el siguiente estadístico:

$T^2 = Z_i' \Sigma_{Z_i}^{-1} Z_i > h$ (1.27) donde: $Z_i = \lambda X_i + (1-\lambda) X_{i-1}$ (1.28)

siendo $Z_o = 0$, λ es la matriz $p \times p$ diagonal de la constante de alisamiento con

$0 < \lambda_i \leq 1$

aunque en la práctica no hay razón para emplear diferentes valor de λ en un mismo problema.

En el caso específico cuando hay presencia de subgrupos racionales, i.e. n>1; simplemente se reemplaza X_i por $\overline{X}_i$.

Lowry et al.(1992) propone dos alternativas para calcular $\Sigma_{z,}$

$\Sigma_{Z_i} = \frac{\lambda\left[1-(1-\lambda)^{2i}\right]}{2-\lambda}(\Sigma)$ (1.29) o $\Sigma_{Z_i} = \frac{\lambda}{2-\lambda}(\Sigma)$ (1.30)

Siendo la primera la de mejor desempeño.

Para ejecutar el gráfico MEWMA en R se utiliza la función mult.chart especificando type = "mewma"

Otro elemento que se le debe pasar a la función es el parámetro lambda, aunque en ausencia del mismo se toma por defecto valor igual a 0.1

En el gráfico MEWMA la matriz de covarianza puede ser estimada de tres formas distintas: para subgrupos racionales, para observaciones individuales usando el método de (Sullivan,Woodall 1996b) y el de (Holmes,Mergen 1993).

En la determinación del UCL, mult.chart usa el método sugerido por (Bodden,Rigdon 1999). Una limitación que presenta es la cantidad de opciones para seleccionar lambda, p y el valor de ARL limitados a los siguientes valores:

p para valores 2,3,...,10

lambda para 0.1, 0.2,...,0.9

y ARL solo para 200

Sin embargo el usuario puede entrar como argumento un valor deseado de UCL obtenido por ejemplo de (Prabhu,Runger 1997)o (Bodden,Rigdon 1999).

A través de la funcion mult.chart:

> mult.chart(type = "mewma", carbon1)

obteniéndose:

```
[1] "MEWMA Control Chart"
$ucl
[1] 10.81
$t2
      [,1]
 [1,] 0.62
 [2,] 0.30
 ...
[29,] 0.07
[30,] 0.12
$Xmv
[1]  0.99  1.04 49.98
```

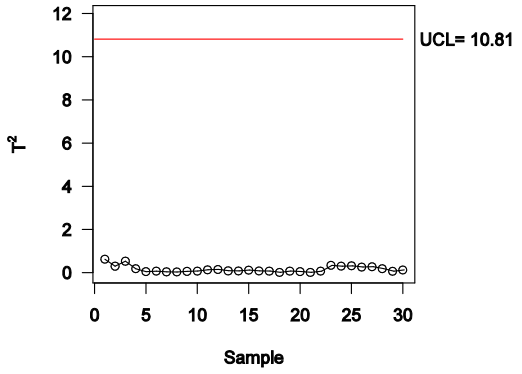

Fig. 1.9 Gráfico MEWMA

$covariance

[,1] [,2] [,3]

[1,] 0.0025 0.0036 0.0067

[2,] 0.0036 0.0140 0.0100

[3,] 0.0067 0.0100 0.0590

1.8 El Gráfico de Suma Acumulada (MCUSUM)

El Gráfico de Suma Acumulada (MCUSUM) aparece como una extensión del gráfico univariado CUSUM propuesto por (Page 1961). Este se enfoca en mejorar la sensitividad respecto a la variante T^2 para detectar pequeñas variaciones en el proceso y está basado en el principio de ir acumulando información de observaciones anteriores.

Existen fundamentalmente cuatro variantes de esta ficha de las cuales solamente dos serán expuestas a continuación

(Crosier 1988) presentó la siguiente variante:

$$T_i^{\,2} = \left[S_i^{'} \left(\frac{\Sigma}{n} \right)^{-1} S_i \right]^{1/2} > h \quad \text{(1.31) donde:}$$

$$S_i = \begin{cases} 0 & if \quad C_i \le k \\ \left(S_{i-1} + \overline{X}_i - «_o \right) \left(1 - \frac{k}{C_i} \right) & if \quad C_i > k \end{cases} \quad \text{(1.32) siendo } S_0=0,\ k>0 \text{ y}$$

$$C_i = \left[\left(S_{i-1} + \overline{X}_i - «_o \right)' \left(\frac{\Sigma}{n} \right)^{-1} \left(S_{i-1} + \overline{X}_i - «_o \right) \right]^{1/2} \quad \text{(1.33)}$$

Igualmente, el límite es fijado como $UCL = h$

La segunda variante presentada por (Pignatiello,Runger 1990) resulta la de mejor rendimiento.

$$T_i^2 = \max\left\{0, \left[S_i'\left(\frac{\Sigma}{n}\right)^{-1} S_i\right]^{1/2} - kn_i\right. \quad (1.34)$$ donde

$$S_i = \sum_{j=i-n_1+1}^{i}\left(\overline{X}_i - «_0\right) \quad (1.35)$$ y $$n_i = \begin{cases} n_{i-1}+1 & if \quad T_i^2 > 0 \\ 1 & otherwise \end{cases} \quad (1.36)$$

$$UCL = h$$

Ambos casos se ejecutan en R usando la función mult.chart especificando type = "mcusum" y "mcusum2" respectivamente.

Los argumentos k y h pueden ser entrados o la función por defecto utiliza los valores 0.5 y 5.5 respectivamente.

Por otro lado la función utiliza las mismas opciones que las variantes T^2 y MEWMA para estimar la matriz de covarianzas

Usando:

```
> data("carbon2")
> Xmv <- mult.chart(carbon1, type = "t2") $Xmv
> S <- mult.chart(carbon1, type = "t2") $covariance
> mult.chart(type = "mcusum", carbon2, Xmv = Xmv, S = S)
```

obtenemos:

"MCUSUM Control Chart by Crosier (1988)"

$ucl

[1] 5.5

$t2

[,1]

[1,] 1.86

[2,] 2.54

...

[24,] 8.69

[25,] 9.41

$Xmv

0.99 1.04 49.98

$covariance

[,1] [,2] [,3]

[1,] 0.0025 0.0036 0.0067

[2,] 0.0036 0.0140 0.0100

[3,] 0.0067 0.0100 0.0590

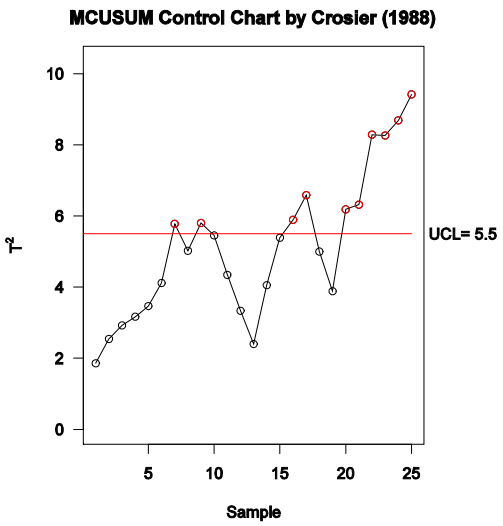

Fig. 1.10 MCUSUM usando el método de (Crosier 1988)

Mientras que especificando type = "mcusum2" R calcula la propuesta de (Pignatiello,Runger 1990).

"MCUSUM Control Chart by Pignatiello (1990)"

$ucl

[1] 5.5

$t2

[,1]

[1,] 1.86

[2,] 2.50

...

[24,] 7.16

[25,] 7.81

$Xmv

0.99 1.04 49.98

$covariance

[,1] [,2] [,3]

[1,] 0.0025 0.0036 0.0067

[2,] 0.0036 0.0140 0.0100

[3,] 0.0067 0.0100 0.0590

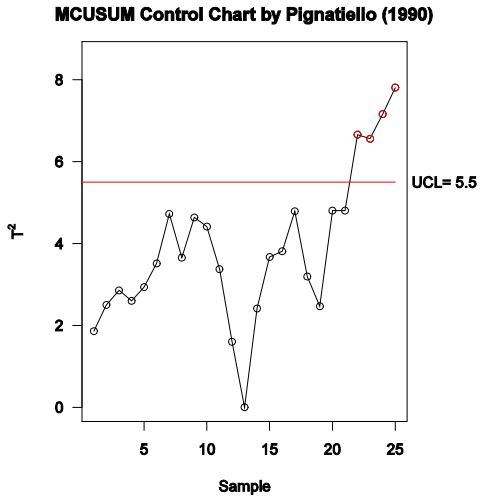

Fig. 1.11 MCUSUM usando el método de (Crosier 1988)

Capítulo 2

Capítulo 2 Índices de Capacidad de Procesos Multivariados

En secciones anteriores se expuso cómo la evaluación del rendimiento de un proceso compuesto por varias variables correlacionadas entre sí, debe ser llevado cabo a través de un enfoque multivariado. En esta sección se exponen las propuestas más importantes de índices de capacidad multivariantes.

Un índice de capacidad puede se descrito como el ratio entre las especificaciones y la dispersión del proceso y provee información acerca del cumplimiento de los requisitos.

Entre los primeros trabajos significativos en este campo encontramos los siguientes (Chan et al. 1991; Bothe 1992; Pearn et al. 1992).

Desde ese momento varios índices han sido propuestos entre los más reconocidos encontramos (Hubele et al. 1991), (Taam et al. 1993), (Shahriari et al. 1995) y (Chen 1994). Por su parte (Wang et al. 2000) llevó a cabo un estudio comparativo de las ultimas propuestas y discutió su utilidad específica.

(Wang,Chen 1998; Xekalaki,Perakis 2002; Wang 2005) sugieren índices basados en análisis de componentes principales como una extensión de las variantes univariadas C_p, C_{pm}, C_{pk} y C_{pmk} .

(Pearn,Kotz 2006) ofrecen una revisión de este campo actualizado hasta el 2004 y (Yum,Kim 2012) desarrollaron una exhaustiva revisión bibliográfica de los índices de capacidad en general.

Más recientemente (Pan,Lee 2010) proponen una modificación el índice de (Taam et al. 1993) con el objetivo de evitar sobreestimación, (Scagliarini 2011) estudia la presencia de errores en las mediciones en los índices basados en componentes principales y (Tano,Vännman 2011) desarrollaron una comparación entre los intervalos de confianza.

El número de propuestas se han incrementado significativamente en años recientes por tal motivo (Shinde,Khadse 2008) clasificaron los índices en cuatro categorías basadas en:

1- el ratio del volumen de la región de tolerancias respecto al ratio de la región del proceso e.g.: (Taam et al. 1993), (Shahriari et al. 1995), (Pan,Lee 2010), etc.

2- el uso de análisis de componentes principales (PCA) e.g: (Wang,Chen 1998; Xekalaki,Perakis 2002; Wang 2005), etc.

3- la probabilidad de productos no conformes como por (Wierda 1993), (Chen 1994), (Chen et al. 2003), (Castagliola,Castellanos 2005), etc

4- otros.

2.1 La función mpci

La medición de la capacidad de un proceso desde la perspectiva multivariada puede ser implementada con la función mpci que es una función genérica.

Esta función permite calcular los índices más aceptados como por ejemplo:

- el vector (Shahriari et al. 1995)
- el índice (Taam et al. 1993)
- la propuesta de (Pan,Lee 2010)
- los índices de (Wang,Chen 1998)
- la propuesta de (Xekalaki,Perakis 2002)
- los índices de (Wang 2005)

La selección del tipo de índice a utilizar es seleccionando especificando el argumento:
index = "shah", "taam", "pan", "wang", "xeke" o "wangw"

en el mismo orden en que han sido introducidos.

La función contiene tres argumentos obligatorios:

x, que es el la matriz o el frame de datos

LSL y USL que resultan los límites de especificación inferior y superior respectivamente.

La meta o Target del proceso puede ser especificada y en caso de omisión la función la calcula como el punto medio entre las tolerancias.

En el caso bivariado el argumento lógico “graphic” permite obtener una representación gráfica de los índices. Para el valor específico p = 3 características de calidad, en próximas secciones se ilustra el uso de gráficos tridimensionales usando el paquete rgl.

Para los tres primeros índices, alpha es la proporción de productos no conformes (fijado convencionalmente en 0.0027).

En el caso de los índices basados en PCA, alpha es el nivel de significación para los métodos descritos más adelante.

Para estos últimos índices el argumento “npc” permite especificar el número de componentes a retener.

La función es además capaz de desarrollar cinco métodos para seleccionar la cantidad de componentes introduciendo:

method = 1, 2, …o 5

ó utilizando el nombre de la rutina

method = "Percentage".

Luego de la ejecución de la función mpci se obtiene una lista que contiene un vector para el (Shahriari et al. 1995) , una lista de cuatro elementos para los índices que emplean PCA y una valor único para los índices de (Taam et al. 1993) y (Pan,Lee 2010).

La función cuenta con una ayuda la cual ofrece mayor detalle sobre su uso. Usar por ejemplo help(package = "MPCI"). Otros ejemplos de la función aparecen en (Santos-Fernández,Scagliarini 2012).

2.2 El Vector Multivariado de Capacidad de Procesos

El Vector Multivariado de Capacidad de Procesos fue introducido por (Shahriari et al. 1995) basado en el trabajo original de (Hubele et al. 1991). Este índice consiste en un vector de tres componentes el cual es definido como:

$$[CpM, PV, LI] \quad (2.1)$$

Y esta basado en el supuesto de que el proceso sigue una distribución normal multivariada.

El primer componente es CpM que es el ratio de las áreas o volúmenes entre las tolerancias y la región modificada del proceso.

$$CpM = \left[\frac{\prod_{i=1}^{p} (USL_i - LSL_i)}{\prod_{i=1}^{p} (UPL_i - LPL_i)} \right]^{1/p} \quad (2.2)$$

Siendo p el número de características de calidad.

Ambas áreas y volúmenes son rectángulos en procesos bivariados y prismas rectangulares en casos tridimensionales.

El área definida por las tolerancias es mostrada en Fig. 3.1 como el rectángulo externo.

Por otro lado, la región modificada del proceso es construida como el rectángulo más pequeño que circunscribe el elipsoide o contorno llamado región de proceso.

El elipsoide es el contorno de probabilidad centrado en la media del proceso la cual es construido por la descomposición espectral de la matriz de covarianzas centrada en el vector de medias

Los bordes de la región del proceso: el límite inferior del (LPL_i) y el superior (UPL_i) son calculados resolviendo el sistema de ecuaciones de la primera derivada de la forma cuadrática de acuerdo con (Nickerson 1994).

$$(X - «)'(\Sigma)^{-1}(X - «) = \text{i}^{2}_{\sim,p} \quad (2.3)$$

con una distribución χ^2 con p grados de libertad y nivel de significación α.

Las soluciones de la ecuación para cada dimensión esta dada por:

$$LPL_i = «_i + \sqrt{\frac{\text{i}^{2}_{\sim,p}\det\left(\Sigma_i^{-1}\right)}{\det\left(\Sigma^{-1}\right)}}$$

$$UPL_i = «_i + \sqrt{\frac{\text{i}^{2}_{\sim,p}\det\left(\Sigma_i^{-1}\right)}{\det\left(\Sigma^{-1}\right)}} \quad (2.4)$$

donde det() es el determinante y Σ_i^{-1} es la matriz que se obtiene eliminando la i^{th} columna y fila.

Valores de CpM mayores que 1, indican que la región modificada del proceso es más pequeña que la región de tolerancias.

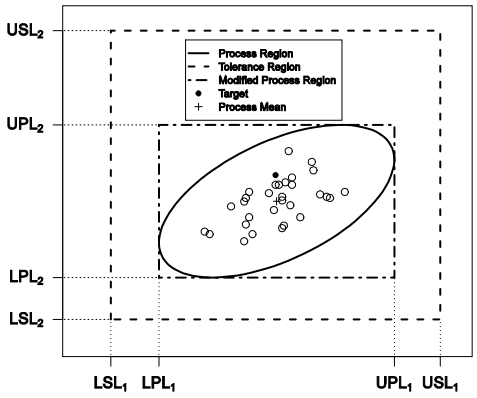

Fig. 2.1 Representación gráfica de la región modificada del proceso.

El segundo componente del vector (PV) está dada la cercanía entre la meta y la media del proceso, expresada como la hipótesis que

$$PV = P(T^2 > \frac{p(m-1)}{m-p} F_{p,m-p}) \quad (2.5)$$

donde: $T^2 = n(\overline{X} - «)^{'}(S)^{-1}(\overline{X} - «)$ (2.6)

y $F_{p,m-p}$ la distribución F con p y m-p grados de libertad respectivamente.

PV toma valores ente 0 y 1, y valores cercanos a cero indica que la media del proceso está distante de la meta o target.

Finalmente, el tercer componente (LI) compara las localizaciones de la región modificada del proceso con las tolerancias, mostrando cuándo alguna parte de la región del proceso cae fuera de la región de tolerancias.

Valores de LI = 0 implican que como mínimo en una dirección la región de tolerancias es superada.

$$LI = \begin{cases} 1 \\ 0 \end{cases}$$

En conclusión el índice de (Shahriari et al. 1995) provee una comparación entre los volúmenes de las regiones, la proximidad de los centros y la extensión de las regiones.

El uso de la función mpci para el cálculo del índice de (Shahriari et al. 1995) se muestra a continuación usando un vector bivariado

Usando el set llamado dowel conjuntamente con las siguientes tolerancias: LSL_1 = 0.47 y USL_1 = 0.53 así como LSL_2 = 0.90 y USL_2 = 1.10

El cómputo se realiza usando la función mpci usando el argumento "shah"

```
> library("MPCI")
> data("dowel1")
> LSL <- c(0.47, 0.90)
> USL <- c(0.53, 1.10)
> Target <- c(0.50, 1.00)

> mpci(index = "shah", dowel1, LSL, USL, Target, graph = TRUE)
```

El argumento graph provee en el caso bidimensional (p = 2) una representación grafica. La salida se muestra a continuación:

```
[1] "Shahriari et al.
(1995) Multivariate
Capability Vector"
$CpM
[1] 1.33
$PV
          [,1]
[1,] 0.74
$LI
[1] 1
```

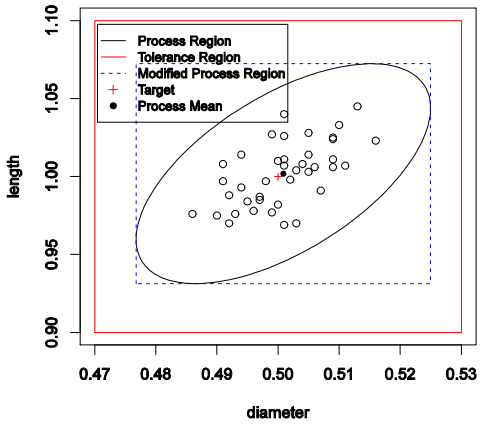

Fig. 2.2 Representación gráfica del método de Shahriari

Como CpM es mayor que uno, entonces el la región modificada del proceso es menor que la región de tolerancias. El valor PV nos es suficientemente cercano a cero como para aseverar que existen diferencias significativas entre los centros y finalmente el valor LI=1 indica que la región del proceso está contendida íntegramente en la región de tolerancias. Finalmente el proceso es encontrado capaz.

2.3 El Índice de Capacidad Multivariado

Otro índice multivariado ampliamente aceptado es el índice MCpm propuesto por (Taam et al. 1993). Éste es definido como el radio de los volúmenes de los elipsoides de las regiones de tolerancias y la región del proceso (ver Fig. 3.3).

A diferencia con el primer componente del vector de (Shahriari et al. 1995) el cual es calculado como el radio de los hipercubos, el índice MCpm es el ratio de los elipsoides.

La región de tolerancias modificadas es el mayor elipsoide construido en la región de tolerancias y centrado en el target.

El índice es calculado como:

$$MCpm = \frac{vol.(R_1)}{vol.(R_2)} \quad (2.7)$$

donde R_1 y R_2 son las regiones de tolerancias modificadas y el elipsoide de confianza respectivamente. Este radio puede ser estimado como:

$$MCpm = \frac{Cp}{D} \quad (2.8)$$

con

$$Cp = \frac{vol.(tolerance \quad region)}{vol.(estimated \quad process \quad region)} \quad (2.9)$$

el numerador es el hiperelipsoide con volumen determinado por:

$$vol.(tolerance \quad region) = \frac{2®^{p/2} \prod_{j=1}^{p} l_j}{p\Gamma(p/2)} \quad (2.10)$$ donde:

l_i son las longitudes de los semi-axes.

Por otro lado:

$$vol.(estimated \quad process \quad region) = |S|^{1/2} (®K)^{p/2} [\Gamma(p/2+1)]^{-1} \quad (2.11)$$

donde K es el percentil de la distribución χ^2 y

$$D = \left[1 + \frac{m}{m-1}(\overline{X} - «)' (S)^{-1} (\overline{X} - «)\right]^{1/2}$$ (2.12) por tanto

$$MCpm = \frac{vol.(R_1)}{\left\{|S|^{1/2} (®K)^{p/2} [\Gamma(p/2+1)]^{-1}\right\} * \left[1 + \frac{m}{m-1}(\overline{X} - «)' (S)^{-1} (\overline{X} - «)\right]^{1/2}}$$ (2.13)

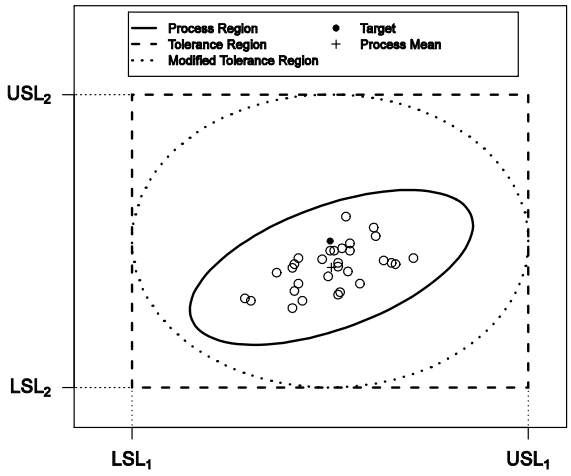

Fig. 2.3 Representación gráfica de la región modificada de tolerancia.

Cuando el valor del índice es mayor que 1 y el vector de medias es igual a la meta, implica que el volumen del proceso es menor que la región de tolerancias modificadas.

El cálculo del índice MCpm en R es usando el argumento index = "taam" y de la siguiente forma

> mpci(index = "taam", dowel1, LSL, USL, Target, graph = TRUE)

Entonces la función devuelve:

Finalmente el índice encuentra capaz el proceso puesto que se obtiene un valor CpM >1

[1] "Taam et al. (1993) Multivariate Capability Index (MCpm)"

$MCpm

[,1]

[1,] 2.19

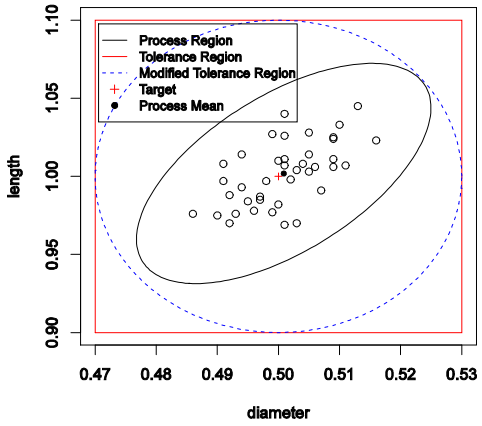

Fig. 2.4 Representación gráfica del método de Taam

2.4 Revisión del Índice de Capacidad Multivariado.

Una de las más recientes propuestas de índice se debe al trabajo de (Pan,Lee 2010), el cual resulta un caso especial de (Taam et al. 1993).

Los autores señalan que la propuesta de (Taam et al. 1993) puede sufrir sobreestimación si las características de calidad no son independientes. En este caso la región de tolerancias está dado por:

$$(X-T)'\left(A^*\right)^{-1}(X-T) = \text{¡}^{2}_{p,1-\sim} \quad (2.14) \text{ donde}$$

$$A^*_{ij} = r_{ij}\left(\frac{USL_i - LSL_i}{2\sqrt{\text{¡}^{2}_{p,1-\sim}}}\right)\left(\frac{USL_j - LSL_j}{2\sqrt{\text{¡}^{2}_{p,1-\sim}}}\right) \quad (2.15)$$

Y r_{ij} es el coeficiente de correlación entre i and j.

Finalmente el índice propuesto resulta:

$$NMCpm = \left(\frac{|A^*|}{|S|}\right)^{1/2} \quad (2.16)$$

La figura siguiente muestra el elipsoide propuesto con estilo de línea longdash (lty = 5).

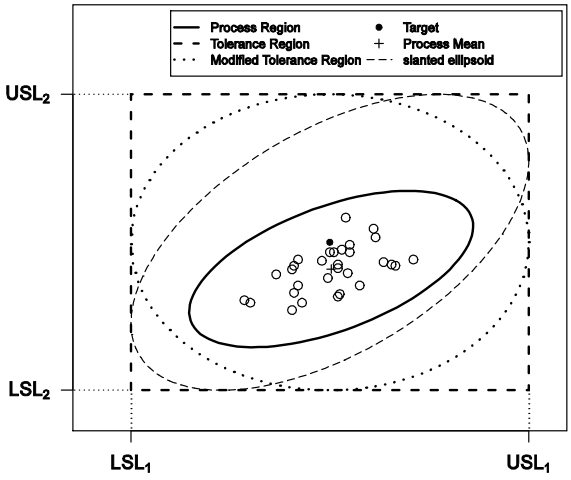

Fig. 2.5 Representación gráfica de la región revisada.

El cálculo en R es como sigue:

```
> mpci(index = "pan", dowel1, LSL = LSL, USL = USL, graph = TRUE)
```

```
[1] "Pan and Lee
(2010) Multivariate
Capability Index
(NMCpm)"
$NMCpm
        [,1]
[1,] 1.75
```

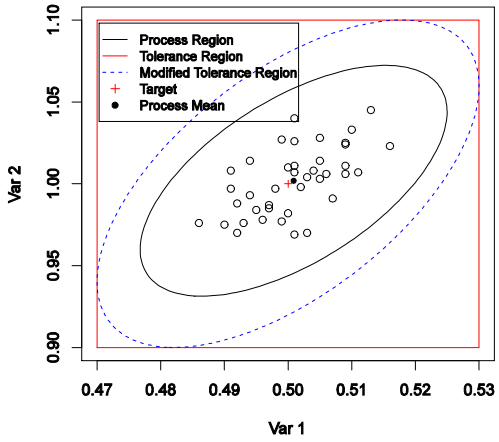

Fig. 2.6 Representación gráfica del método de Pan

2.5 Índices de Capacidad Multivariantes basados en Análisis de Componentes Principales (PCA)

Muchos índices basados en el uso de Componentes Principales han sido propuestos en los últimos años.

Algunos de los más aceptados fueron sugeridos por (Wang,Chen 1998), (Xekalaki,Perakis 2002) y (Wang 2005).

Como esta metodología comienza con un análisis PCA, entonces se obtienen variables no correlacionadas así como una reducción de la dimensionalidad.

Los índices se basan en la descomposición espectral de la matriz de covarianza:

$$\Sigma = UDU' \quad (2.17)$$

donde U es la matriz de vectores propios y D la matriz diagonal de valores propios.

$$D = diag(\mathsf{a}_1, \mathsf{a}_{2,\ldots,} \mathsf{a}_p) \quad (2.18)$$

El ith componente principal resulta como

$$PC_i = u_i' x$$

Las especificaciones de ingeniería (Especificación Superior, Inferior y la Meta) son transformadas como sigue:

$$LSL_{PC_i} = u_i' LSL$$

$$USL_{PC_i} = u_i' USL$$

$$T_{PC_i} = u_i' T \quad (2.19)$$

donde $i = 1,2,\ldots, p$ son los ith componentes principales.

Normalmente el primer componente es responsable de la mayoría de la variabilidad, por tanto la dimensionalidad puede ser reducida sin una pérdida significativa de información.

El problema consiste en cuántos componentes se deben retener.

En la próxima sección se introducen cinco métodos para definir este asunto

La propuesta de (Wang,Chen 1998) es la extensión multivariante de los índices univariados C_p, C_{pk}, C_{pm} y C_{pmk}.

$$MC_p = \left(\prod_{i=1}^{3} C_{p;PC_i} \right)^{1/3} \quad (2.20)$$

donde

$$C_{p;PC_i} = \frac{USL_{PC_i} - LSL_{PC_i}}{6\pm_{PC_i}} \quad (2.21)$$

y donde υ es el número de componentes principales y

$\pm_{PC_i} = \sqrt{a_i}$.

De la misma forma se obtienen MC_{pk}, MC_{pm} y MC_{pmk} reemplazando $C_{p;PC_i}$ por $C_{pk;PC_i}$, $C_{pm;PC_i}$ y $C_{pmk;PC_i}$ respectivamente, donde

$$C_{pk;PC_i} = \min\left\{ \frac{USL_{PC_i} - «}{3\pm_{PC_i}}, \frac{« - LSL_{PC_i}}{3\pm_{PC_i}} \right\} \quad (2.22),$$

$$C_{pm;PC_i} = \frac{USL_{PC_i} - LSL_{PC_i}}{6\sqrt{\pm^2_{PC_i} + (« - T)^2}} \quad (2.23)$$

y

$$C_{pkm;PC_i} = \frac{\left(USL_{PC_i} - LSL_{PC_i}\right)/2 - \left| « - \left[\left(USL_{PC_i} + LSL_{PC_i}\right)/2 \right] \right.}{3\sqrt{\pm^2_{PC_i} + (« - T)^2}} \quad (2.24)$$

Para ilustrar el uso en R usando el set bimetal1 y las siguientes especificaciones:

$$LSL = \begin{bmatrix} 19.0 & 39.0 & 13.0 & 20.2 & 24.5 \end{bmatrix}$$

$$USL = \begin{bmatrix} 23.0 & 41.0 & 17.0 & 23.8 & 27.5 \end{bmatrix}$$

$$\text{Target} = \begin{bmatrix} 21.0 & 40.0 & 15.0 & 22.0 & 26.0 \end{bmatrix}$$

Para estos datos los valores y vectores propios resultan:

$$U = \begin{bmatrix} 0.5968 & 0.5476 & 0.5108 & 0.2881 & 0.0044 \\ 0.2731 & 0.0408 & -0.0739 & -0.5246 & 0.8019 \\ 0.6413 & -0.7344 & -0.0752 & 0.2014 & -0.0561 \\ 0.2988 & 0.3948 & -0.8505 & 0.1404 & -0.1083 \\ 0.2620 & 0.0575 & 0.0673 & -0.7626 & -0.5848 \end{bmatrix}$$

$$D = \begin{bmatrix} 0.1700 & 0.0659 & 0.0396 & 0.0148 & 0.0023 \end{bmatrix}$$

Reteniendo los dos primeros componentes principales

```
>mpci(index = "wang",
 bimetal1,
 LSL,
 USL,
 Target,
 method = 1,
 perc = 0.80)
```

Obteniéndose:

"Wang and Chen (1998) Multivariate Process Capability Indices (PCI) based on PCA"	$MCp $MCpm
$‘number of principal components‘ [1] 2	[1] 1.34 [1] 1.24
	$MCpk $MCpmk

[1]	[1]
1.13	1.04

Como en la propuesta de (Wang,Chen 1998) los componentes son tomados con la misma importancia aun cuando el primer componente tiene mas peso que los restantes; (Xekalaki,Perakis 2002) propusieron corregir lo anterior de acuerdo con la varianza explicada por cada componente principal.

$$MXPC_p = \frac{\sum_{i=1}^{3} a_i C_{p;PC_i}}{\sum_{i=1}^{3} a_i} \quad (2.25)$$

$MXPC_{pk}$, $MXPC_{pm}$ y $MXPC_{pmk}$ .se obtienen de forma similar.

El cálculo del índice de (Xekalaki,Perakis 2002) se presenta a continuación:

> mpci(index = "xeke", bimetal1, LSL, USL, Target, method = 1, perc = 0.80)

obteniéndose la siguiente salida:

"Xekalaki and Perakis (2002) Multivariate Process Capability Indices (PCI) based on PCA"	$MCp	$MCpm
	[1]	[1] 2.17
	2.31	$MCpmk
$‘number of principal components‘	$MCpk	[1] 2.05
[1] 2	[1]	
	2.18	

Por otro lado, Wang (2005) sugiere otra variante usando la media geométrica

El índice resulta:

$$MWC_p = \left(\prod_{i=1}^{3} C_{p;PC_i}^{a_i} \right)^{1/\sum_{i=1}^{3} a_i} \quad (2.26)$$

De igual forma son calculados MWC_{pk}, MWC_{pm} y MWC_{pmk}

El uso de la función mpci en esta variante es especificando index = "wangw"

```
> mpci(index = "wangw", bimetal1, LSL, USL, Target, method = 1, perc = 0.80)
```

```
"Wang(2005) Multivariate Process Capability Indices(PCI) based on PCA"
$`number of principal components`
2
$MCp
[1]
1.91
$MCpk
[1]
1.70
$MCpm
[1] 1.77
$MCpmk
[1] 1.58
```

2.6 Metodología para seleccionar el número de Componentes Principales

En secciones anteriores se expuso como el Análisis Componentes Principales permiten la reducción de la dimensionalidad de los datos en los cuales $1 \leq a \leq p$ componentes principales pueden ser retenidos.

Existen varios métodos que posibilitan decidir cuántos componentes deben ser retenidos o usados con la finalidad de evitar la pérdida significativa de información.

A continuación se expone una metodología contenida en la función mpci la cual consta de cinco métodos encontrados en la literatura.

Método 1 o Porcentage:

Esta técnica garantiza que se retenga al menos el porcentaje especificado de proporción acumulada de varianza. Se fija normalmente en un 80%.

Usando el set de datos llamado bimetal1 tomado del proceso de elaboración de un bimetal

Usando summary(princomp(bimetal1)) R muestra una sumario de la proporción de la varianza, la proporción acumulada y los valores propios.

Tabla 2.1 Importancia de los componentes.

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	0.40486	0.25205	0.19551	0.11965	0.04673
Proportion of Variance	0.58091	0.22515	0.13547	0.05074	0.00774
Cumulative Proportion	0.58091	0.80606	0.94152	0.99226	1.00000

Si se usa el límite del 80% entonces los primeros dos componentes deben ser retenidos.

Método 2 o Average:

El segundo método está basado en la retención de los componentes principales cuyos valores propios sean mayores que la media de los mismos

$$\sum_{i=1}^{p} a_i / p$$

Los valores propios son calculados como:

```
eig <- eigen(cov(bimetal1))$values
print(eig)
[1] 0.169984728 0.065883347 0.039640343 0.014847291 0.002264529
```

Si mean(eig)=0.05852405, entonces solo los dos primeros componentes cumplen con la condición.

Método 3 o Scree: El gráfico scree es un procedimiento visual que plotea los valores propios. Permite determinar cuáles componentes son significativamente alejados de la línea que forman los últimos valores propios.

En el siguiente gráfico se muestra cómo el primer componente se separa de la línea recta, por tanto solamente el primer componente debe ser retenido.

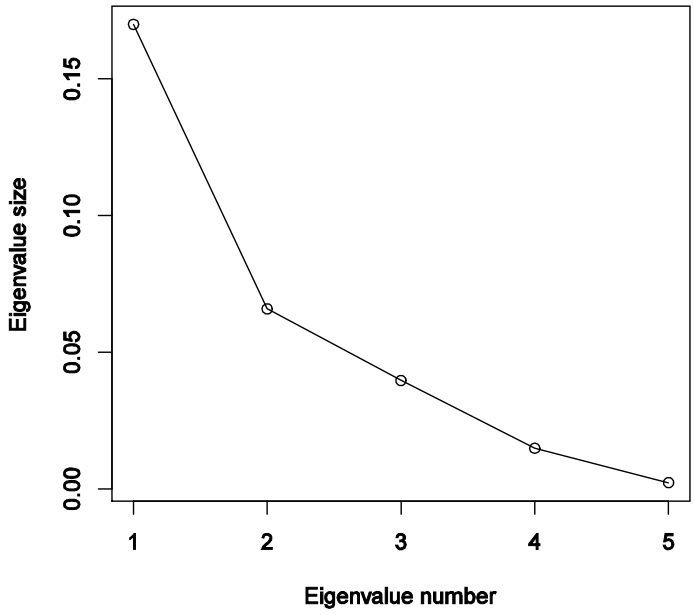

Fig. 2.7 Scree graph for the eigenvalues.

Método 4 o (Bartlett 1954) Test: Este método es un test estadístico diseñado para ignorar aquellos componentes no significativos del resto y asume multinormalidad. Usualmente este método produce un número mayor de componentes que los métodos descritos anteriormente. Ver por ejemplo (Rencher 2002)

$$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_p$$
$$H_1 : \alpha_i \neq \alpha_j \text{ for some } i \neq j$$

$$\bar{\alpha} = \sum_{i=p-k+1}^{p} \alpha_i / k \quad (2.27)$$

donde k es la secuencia p, p-1, p-2, …,1

$$\chi^2 = \left(n - \frac{2p+11}{6}\right)\left(k \ln \bar{a} - \sum_{i=p-k+1}^{p} \ln a_i\right) \quad (2.28)$$

$$\chi^2 \geq \chi^2_{\sim,1/2(k-1)(k+2)} \quad (2.29)$$

Los resultados teóricos y prácticos de χ^2 son:

Tabla 2.2 Valores del test estadístico.

Eigenvalue	k	χ^2	$\chi^2_{\sim,1/2(k-1)(k+2)}$
0.16998	5	93.80	33.20
0.06588	4	56.57	25.26
0.03964	3	39.82	18.21
0.01485	2	19.06	11.83
0.00226	1	0	0

Esto implica que los primeros cuatro son significativamente diferentes. Por tanto de acuerdo con el test de Bartlett se deben retener los cuatro primeros.

Método 5 o (Anderson 1963) Test:

Otra técnica ampliamente usada es el test de Anderson que diferencia además los componentes significativos de los que no los son.

$$H_0 : a_1 = a_2 = \ldots = a_p$$
$$H_1 : a_i \neq a_j \text{ for some } i \neq j$$

donde k = 1, 2, …, p

$$\chi^2 = -3 \sum_{i=k+1}^{p} \ln(a_i) + 3(p-k)\ln\left(\frac{\sum_{i=k+1}^{p} a_i}{p-k}\right) \quad (2.30)$$

$$i^2 \geq i^2_{\sim,1/2(p-k-1)(p-k+2)} \quad (2.31)$$

Los resultados se muestran en la siguiente tabla:

Tabla 2.3 Valores del test estadístico.

Eigenvaluek		i^2	$i^2_{\sim,1/2(k-1)(k+2)}$
0.16998	0	103.37	33.20
0.06588	1	62.34	25.26
0.03964	2	43.88	18.21
0.01485	3	21.01	11.83
0.00226	4	0	0

Este método encuentra los primeros cuatro significativamente diferentes.

Capítulo 3

Capítulo 3 Casos de Estudio y reportes sobre el uso

3.1 Caso de Estudio #1. Control del lanzamiento de los pitchers en el Béisbol

En este caso de estudio son aplicadas la mayoría de las técnicas presentadas, específicamente en la evaluación del rendimiento de los lanzadores en el béisbol.

De acuerdo con la Major League Baseball (MLB) la zona de strike es "aquella área sobre el home plate cuyo límite superior es la línea horizontal hasta la mitad entre la parte superior de los hombros y la parte superior de los pantalones; y con límite inferior la parte detrás de las rodillas…"

Resulta un prisma pentagonal con 20 pulgadas de ancho y altura determinada por la estatura del bateador en posición de batear la bola. Aunque esta dimensión varía de una bateador a otro; normalmente tiene dimensión desde 1.6 hasta 3.5 pies sobre el home.

El árbitro canta strike cuando el lanzamiento cae en esa área y el bateador no hace swing.

Aunque los lanzadores mueven la pelota estratégicamente en diferentes posiciones de la zona de strike tratando que el bateador no pueda hacer contacto con la pelota, frecuentemente el desempeño del pitcher es medido por la habilidad de poner la bola en la zona a cierta velocidad

En este caso de estudio fueron recolectados los datos de la bitácora de lanzamientos de la base de datos de la MLB (http://gd2.mlb.com/components/game/mlb/) del lanzador C.C. Sabathia de los Yankees de New York.

Fueron seleccionados dos set de datos de juegos contra Tampa Bay: el primero el día 10 de julio de 2011 y el segundo el 12 de Agosto del mismo año. Ambos están alojados en el paquete MSQC como sabathia1 y sabathia2 respectivamente.

Esta bitácora posee gran cantidad de información acerca del lanzamiento, pero en este estudio solamente son estudiados las variables de velocidad a la salida (medida en mph) y localización (en pies) al cruzar el home.

Este último punto es medido relativo a un sistema de coordenadas con origen en el home. El eje z está verticalmente orientado, mientras que el eje x se localiza de forma horizontal orientado a la derecha del catcher.

Solamente los lanzamientos de recta rápida son considerados y cada muestra corresponde a un bateador promediando las variables de cada lanzamiento. Notar que un bateador consume más de un turno en un juego.

Para realizar el procesamiento en R

```
> data("sabathia1")
```

usando:

```
> colMeans(sabathia1)
> covariance(sabathia1)
> cor(sabathia1)
```

se obtiene

$$\bar{x} = \begin{bmatrix} 0.1074 \\ 2.9430 \\ 94.4108 \end{bmatrix}$$

$$S = \begin{bmatrix} 0.22 & 0.09 & 0.05 \\ 0.09 & 0.27 & -0.25 \\ 0.05 & -0.25 & 1.50 \end{bmatrix}$$

y

$$r = \begin{bmatrix} 1 & 0.37 & 0.09 \\ 0.37 & 1 & -0.39 \\ 0.09 & -0.39 & 1 \end{bmatrix}$$

Notar la correlación directa entre las primeras dos variables, siendo negativa entre la posición vertical y la velocidad. El scatterplot así lo demuestra visualmente.

```
> pairs(sabathia1)
```

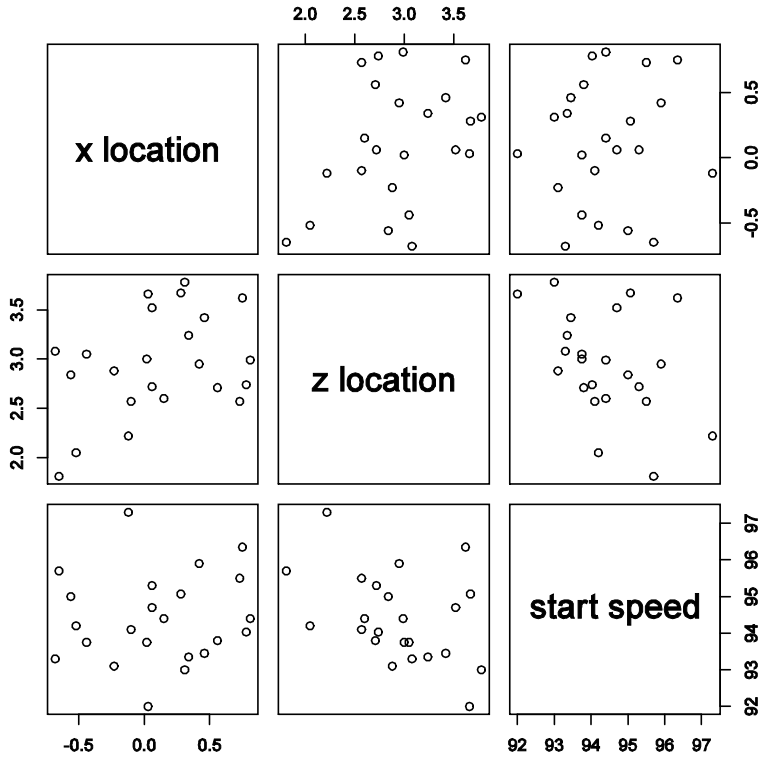

Fig. 3.1 Matriz de scatterplot de la posición vertical y horizontal y la velocidad.

Un útil análisis inicial puede ser realizado construyendo un scatterplot en 3D

>library(rgl)

> plot3d(ellipse3d(cov(sabathia1), centre = colMeans(sabathia1), level=0.99), xlab = "", ylab = "", zlab = "",type="wire", col="gray1", alpha=0.2)

> points3d(sabathia1, size = 4, cex = 2, add = TRUE)

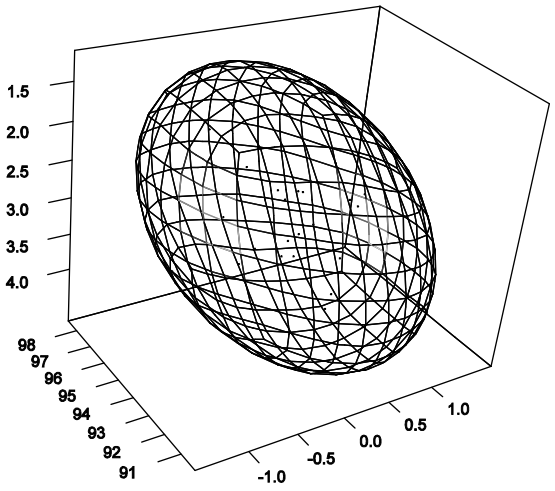

Fig. 3.2 Diagrama de dispersión tridimensional con elipsoide de confianza.

Moviéndose a través de las coordenas puede ser determinado cómo todos las observaciones caen dentro de estos límites. No se encuentran valores extremos.

Entonces llevando a cabo un gráfico de Hotelling

```
> mult.chart(type = "t2", sabathia1)
```

```
[1] "Hotelling Control
Chart"
$ucl
[1] 13.31
$t2
     [,1]
 [1,] 4.37
 [2,] 1.65
...
[22,] 6.95
[23,] 6.06
$Xmv
[1]  0.11  2.94 94.41
$covariance
     [,1]   [,2]   [,3]
[1,] 0.220  0.092  0.051
[2,] 0.092  0.270 -0.250
[3,] 0.051 -0.250  1.500
```

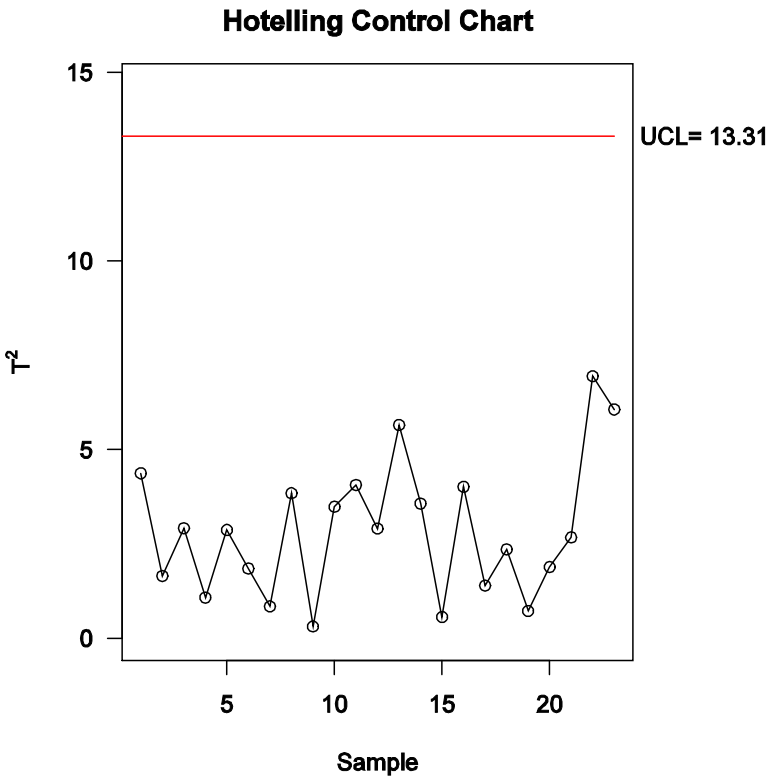

Fig. 3.3 Gráfico de control de Hotelling de los datos de sabathia1.

Ningún punto cae fuera del UCL, por tanto no existe evidencia para rechazar el control estadístico del proceso. EL resultado final así lo demuestra.

Entonces, fijando el primer juego para calcular los parámetros y analizar el segundo partido como Fase II:

```
> colm <- nrow(sabathia1)
> vec <- (mult.chart(sabathia1,type = "t2")$Xmv)
> mat <- (mult.chart(sabathia1,type = "t2")$covariance)
```

usando

```
> data("sabathia2")
> par(mfrow = c(1,2))
> mult.chart(type = "t2", sabathia2, Xmv = vec, S = mat, colm = colm)
> mult.chart(type = "mewma", sabathia2, Xmv = vec, S = mat)
```

Entonces R devuelve:

The following(s) point(s) fall outside the control limits[1] 16 20

$‘Decomposition of‘ [1] 16

	t2 decomp	ucl	p-value	1	2	3
[1,]	12.5255	8.0686	0.0016	1	0	0
[2,]	10.7037	8.0686	0.0031	2	0	0
[3,]	4.2001	8.0686	0.0511	3	0	0
[4,]	16.8950	12.1448	0.0000	1	2	0
[5,]	18.1565	12.1448	0.0000	1	3	0
[6,]	11.3942	12.1448	0.0003	2	3	0

$‘Decomposition of‘ [1] 20

	t2 decomp	ucl	p-value	1	2	3
[1,]	0.4091	8.0686	0.5282	1	0	0
[2,]	9.6004	8.0686	0.0048	2	0	0
[3,]	0.8664	8.0686	0.3609	3	0	0
[4,]	13.4175	12.1448	0.0001	1	2	0
[5,]	1.3922	12.1448	0.2671	1	3	0

0				[6,]	15.0562	12.1448	0.0001 2
[7,]	19.4116	16.1352	0.0000 1 2 3	0			
				[7,]	22.4067	16.1352	0.0000 1 2 3

El análisis muestra cómo los puntos 16 y 20 caen fuera del umbral, i.e.: el pitcher parece estar fuera de control. La descomposición del estadístico T^2 muestra cómo en la muestra 16 ambas localizaciones en la horizontal y en la vertical presentan variaciones no aleatorias. En contraste en bateador número 20 solo la localización en la vertical produce la alarma.

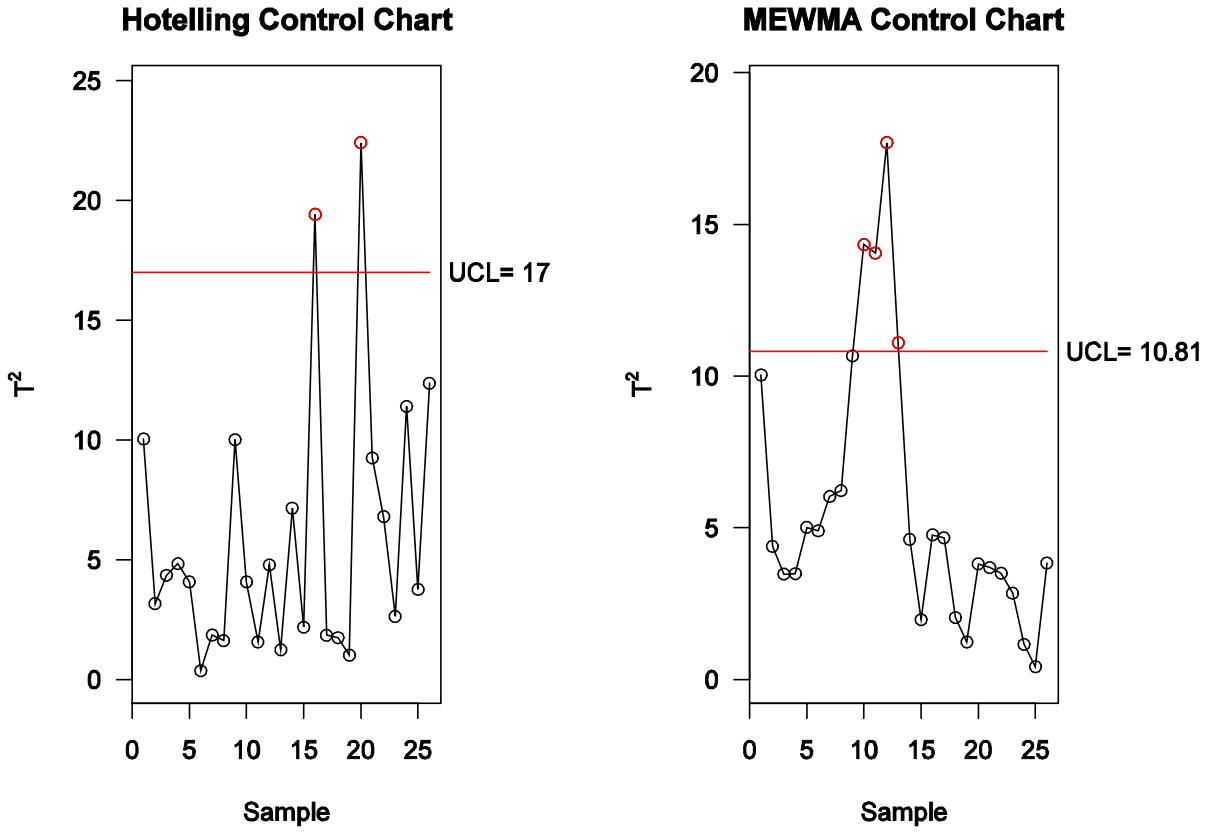

Fig. 3.4 Gráfico Hotelling (a) y (b) MEWMA de los datos guardados en set de datos sabathia2.

Con el objetivo de mejorar la detección de pequeños cambios en el proceso, las fichas MEWMA y MCUSUM son ejecutadas.

Por ejemplo la ficha MEWMA detecta variación no aleatoria en la media en el 10mo bateador ver Fig. 3.4 y MCUSUM de acuerdo a (Crosier 1988) y (Pignatiello,Runger 1990) en las muestras 9 y 10 respectivamente.

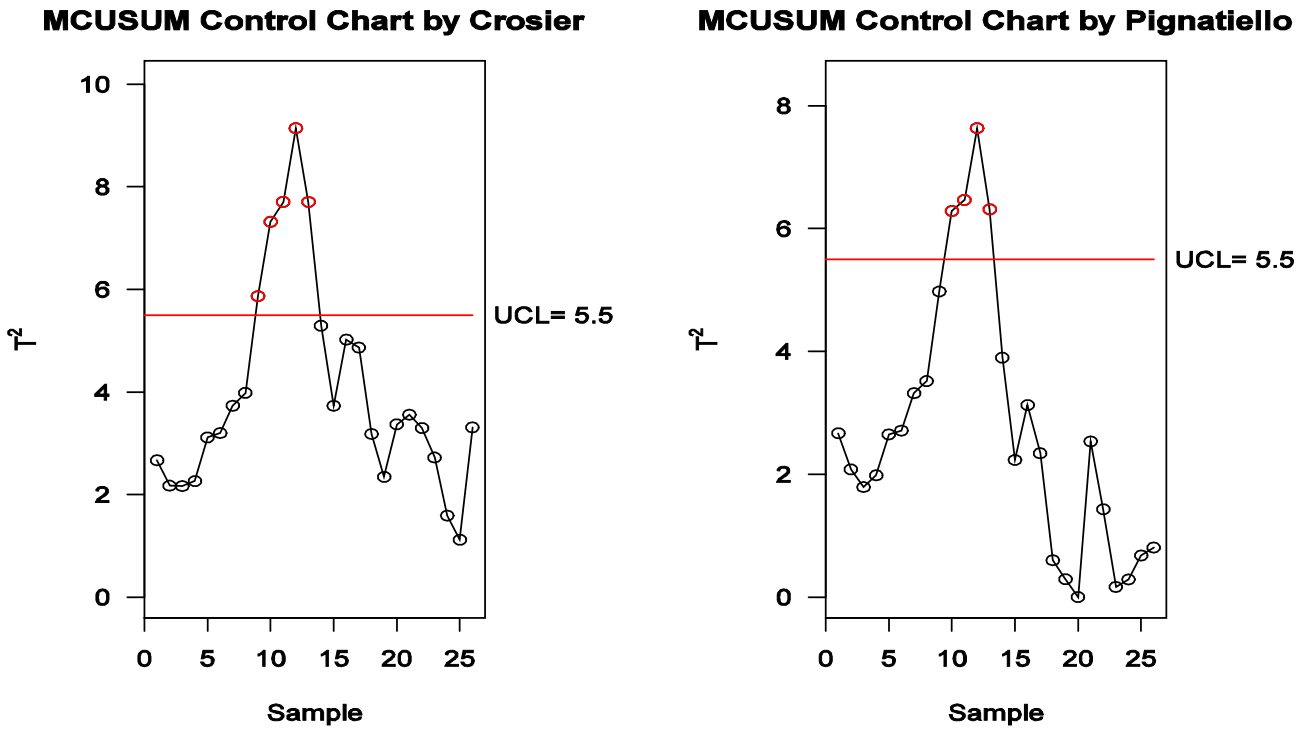

Fig. 3.5 Gráfica MCUSUM para el dataset sabathia2.

Notar que este estudio no pretende probar per se cuándo un lanzador está en control o no. Existen muchas otras variables que deben ser analizadas.

El objetivo es proponer una herramienta que permita el monitoreo del control sobre la zona de strike y la velocidad.

Otro aspecto importante a ser considerado es que aunque el lanzador presente control estadístico sobre las variables medidas, este puede ser bateado por lo que el score puede ser engañoso.

Luego puede ser desarrollado un estudio de capacidad para evaluar el cumplimiento con las especificaciones del la zona de strike especifica del arbitro.

El primer partido estudiado que fue usado como Fase I tuvo a Ron Kulpa como arbitro. La zona de strike fue construida como el rectángulo delimitador de la elipse de confianza dada por todos los lanzamientos cantados strike en este juego y almacenado en el dataset llamado kulpa. Por tanto usando la función proc.reg los límites son calculados

```
> data("kulpa")
> LSL <- as.vector(proc.reg(kulpa, alpha = 0.1)$LPL)
> USL <- as.vector(proc.reg(kulpa, alpha = 0.1)$UPL)
```

Notar que fue seleccionado alpha = 0.1 para evitar una área muy extensa

```
> data("sabathia.ind")
> par(mfrow = c(1,3))
```

```
> mpci(index = "shah", sabathia.ind, LSL = LSL ,USL = USL,alpha = 0.1, graph = TRUE)
```

```
> mpci(index = "taam", sabathia.ind, LSL = LSL ,USL = USL,alpha = 0.1, graph = TRUE)
```

```
> mpci(index = "pan", sabathia.ind, LSL = LSL ,USL = USL,alpha = 0.1, graph = TRUE)
```

[1] "Shahriari et al. (1995) Multivariate Capability Vector"	[1] "Taam et al. (1993) Multivariate Capability Index (MCpm)"	[1] "Pan and Lee (2010) Multivariate Capability Index (NMCpm)"

```
$CpM                $MCpm           $NMCpm
[1] 0.94                  [,1]            [,1]
$PV                 [1,] 0.73       [1,] 0.73
          [,1]
[1,] 6.72e-05
$LI
[1] 0
```

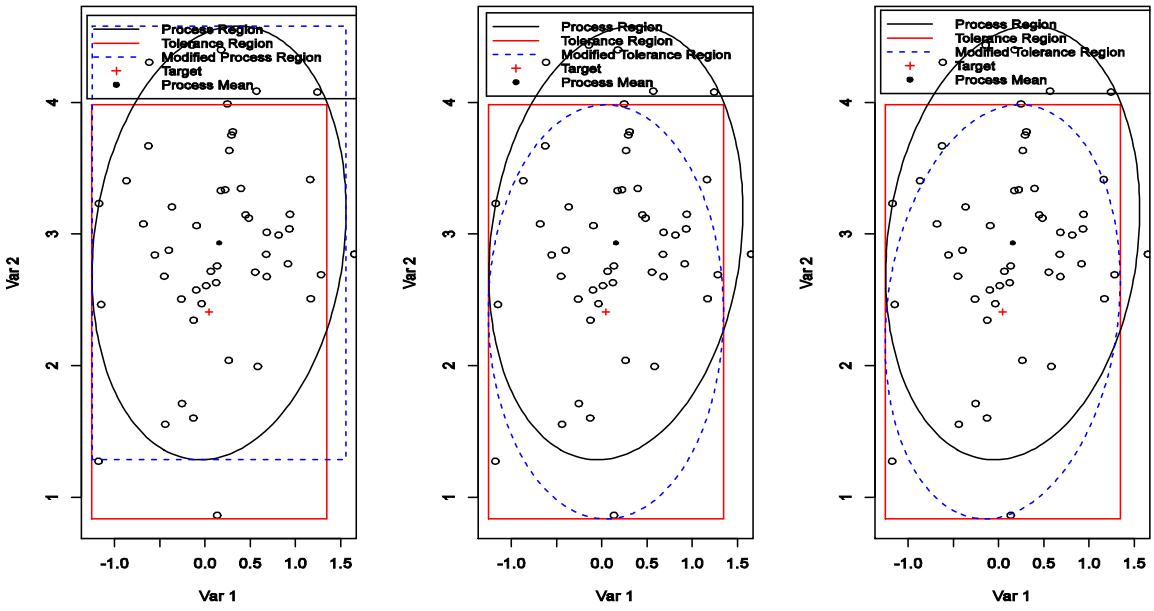

Fig. 3.6 Índices (Shahriari et al. 1995), (Taam et al. 1993) y (Pan,Lee 2010) calculados a partir de los datos de sabathia1.

La Fig. 3.6 muestra la salida de los tres índices calculados.

Notar la diferencia entre la meta y la media del proceso expresado en el extremadamente bajo valor de PV en el índice de (Shahriari et al. 1995).

La mayoría de los puntos están localizados sobre la parte superior de la zona de strike y la región del proceso no está contenida por la región de tolerancia, por tanto LI= 0.

Por otro lado, el radio de las áreas del índice de (Shahriari et al. 1995) produce

Un valor elevado mientras que los otros índices como (Taam et al. 1993) y (Pan,Lee 2010) logran valores más bajos (0.73).

Tener en cuanta que la llamada proporción de producción no conforme, en este caso la proporción de bolas que caen fuera de la zona de strike es como promedio de un tercio en las estadísticas de la MLB.

Estos índices pueden se de gran utilidad para llevar a cabo comparaciones entre pitchers y entre las diferentes zonas de strike de los árbitros las cuales varían en todos los partidos.

Finalmente chequeando el supuesto de multinormalidad (MVN) con la prueba de Henze-Zirkler y la de Royston. Estas pruebas están además incluidas en el paquete MSQC. Para mas detalles ver (Santos-Fernández 2013)

HZ.test(sabathia1)	HZ.test(sabathia2)
[1] 0.75 0.49	[1] 0.69 0.52
Royston.test(sabathia1)	Royston.test(sabathia2)
test.statistic	test.statistic
p.value	p.value
1.49 0.68	1.61 0.65

Por otro lado verificando la carencia de autocorrelación. Este tema es también abordado en detalle en (Santos-Fernández 2013)

```
> par(mfrow = c(2,3))
> for( i in 1 : ncol(sabathia1) ){par(mar = c(4.1,4.5,3,1))
>     acf(sabathia1[,i],lag = 7,las = 1, main= colnames(sabathia1)[i])}
> for( i in 1 : ncol(sabathia2) ){ par(mar = c(4.1,4.5,3,1))
>     acf(sabathia2[,i],lag = 7,las = 1, main= colnames(sabathia2)[i])}
```

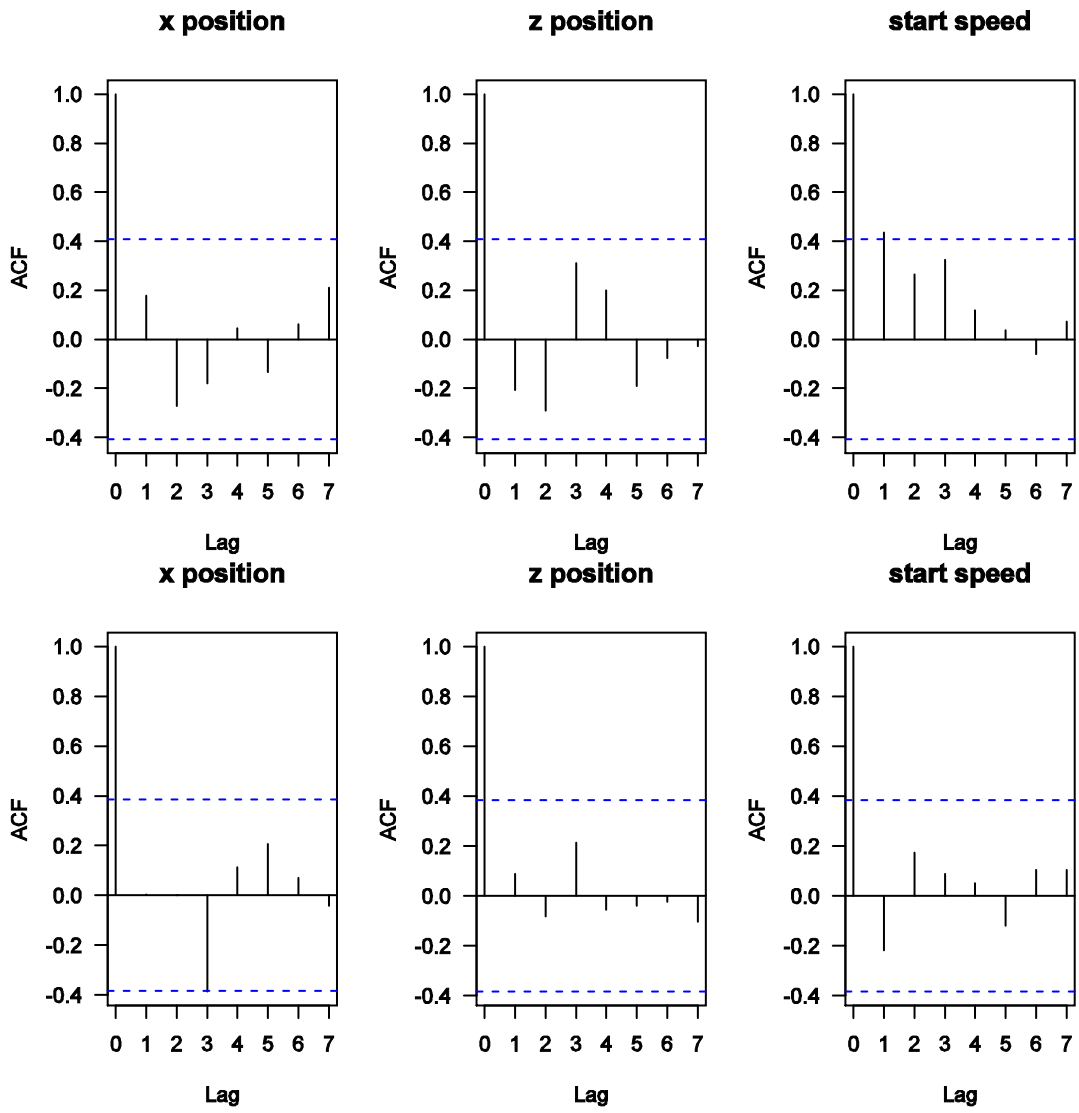

Fig. 3.7 Correlograma de ambos conjuntos de datos.

Notar que no existen evidencias de la carencia de normalidad y que no se obtiene autocorrelación.

Este estudio muestra el amplio espectro de aplicaciones del control estadístico de procesos multivariados, en el cual la combinación de las gráficas de control y los estudios de capacidad de procesos permiten evaluar el desempeño de los pitchers así como la habilidad para cumplir con las especificaciones de la zona de strike.

3.2 Caso de Estudio #2. Tiro con Arco

El tiro con arco es un deporte competitivo gobernado por la World Archery Federation (WA) en el cual los arqueros disparan a una diana a cierta distancia. El establecido en los Juegos Olímpicos posee una diana de 122 centímetros la cual se localiza a 70 metros.

La competencia individual cuenta con dos etapas. En la primera denominada etapa de ranking, cada arquero dispara 72 flechas en 12 grupos de seis flechas cada una. Luego comienza la segunda etapa con las rondas entre el primer rankeado contra el 64, el segundo contra el 63 y así sucesivamente, disparando 18 flechas en grupos de tres. Los ganadores avanzan hasta completar tres ciclos. Entonces los ocho restantes arqueros continúan la eliminación disparando 12 flechas en grupos de tres; siendo el campeón el que queda sin derrotar.

El conjunto de datos llamado archery1 consiste en 72 disparos en grupos de tres flechas de la etapa de ranking de cierto arquero y el conjunto archery2, 54 disparos de la etapa de eliminación con el mismo tamaño del subgrupo.

Notar que la información es dada en coordenadas x y y aunque en las competiciones de tiro el puntaje es basado en la localización de la flecha en aros concéntricos con una valor establecido.

Las Fig. 3.8 (a) y (b) muestran el scatterplot de los lanzamientos individuales sobre una diana de 122 centímetros.

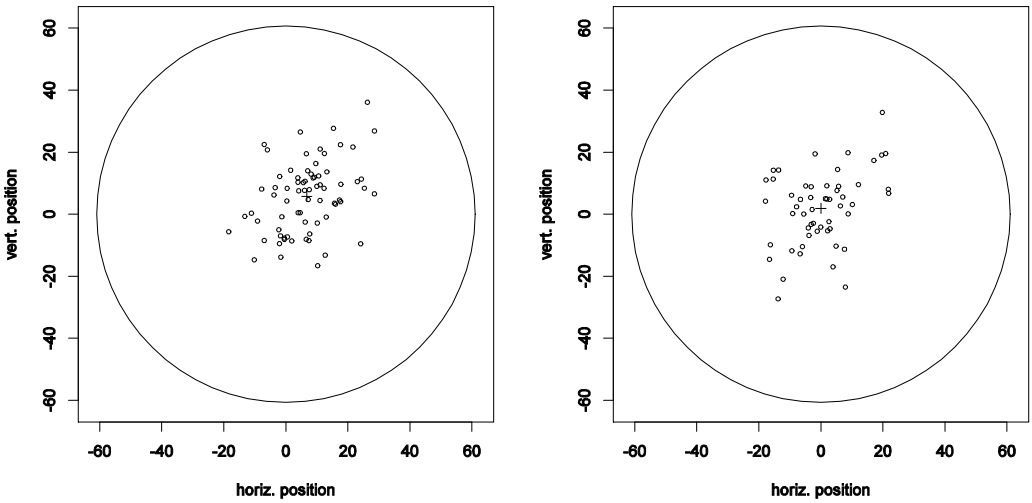

Fig. 3.8 Scatterplot de ambos conjuntos de datos.

```
> data("archery1")
> data("archery2")
```

Usando

```
> cor(cbind(c(archery1[,1,]),c(archery1[,2,])))
```

es calcualda la correlación

$$r = \begin{bmatrix} 1 & 0.37 \\ 0.37 & 1 \end{bmatrix}$$

Luego usando una ficha de control de Hotelling para la primera etapa

```
mult.chart(archery1, type = "t2")
```

obteniendose:

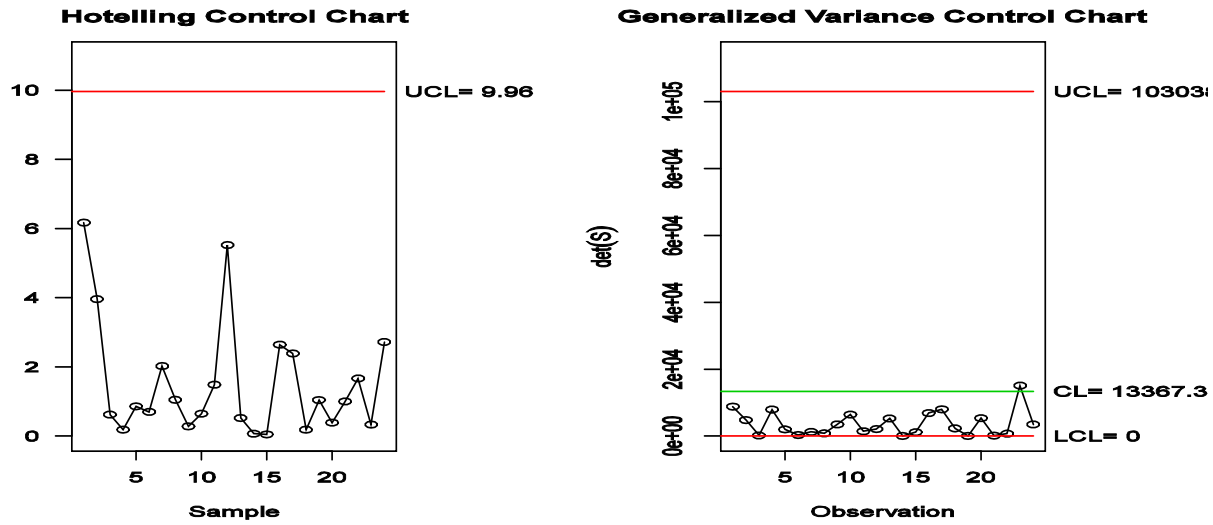

Fig. 3.9 Gráfico de Hotelling y de Varianza Generalizada para el dataset archery1.

De acuerdo con está ficha el proceso parece estar en control, puesto que no se encuentra evidencia de variaciones no aleatorias. Luego el análisis puede ser complementado con el gráfico de varianza generalizada. Este último tampoco reporta operación no aleatoria.

```
> gen.var(archery1)
```

Suponer que es deseado usar la ronda de ranqueo como Fase I y control de esta forma las futuras observaciones almacenadas en archery2:

```
> colm <- nrow(archery1)
> vec <- (mult.chart(archery1,type = "t2")$Xmv)
> mat <- (mult.chart(archery1,type = "t2")$covariance)

> par(mfrow = c(2,2))
> mult.chart(archery2,type = "t2", Xmv = vec, S = mat, colm = colm)
> mult.chart(archery2,type = "mewma", Xmv = vec, S = mat)
> mult.chart(archery2,type = "mcusum", Xmv = vec, S = mat)
> mult.chart(archery2,type = "mcusum2", Xmv = vec, S = mat)
```

Entonces R devuelve:

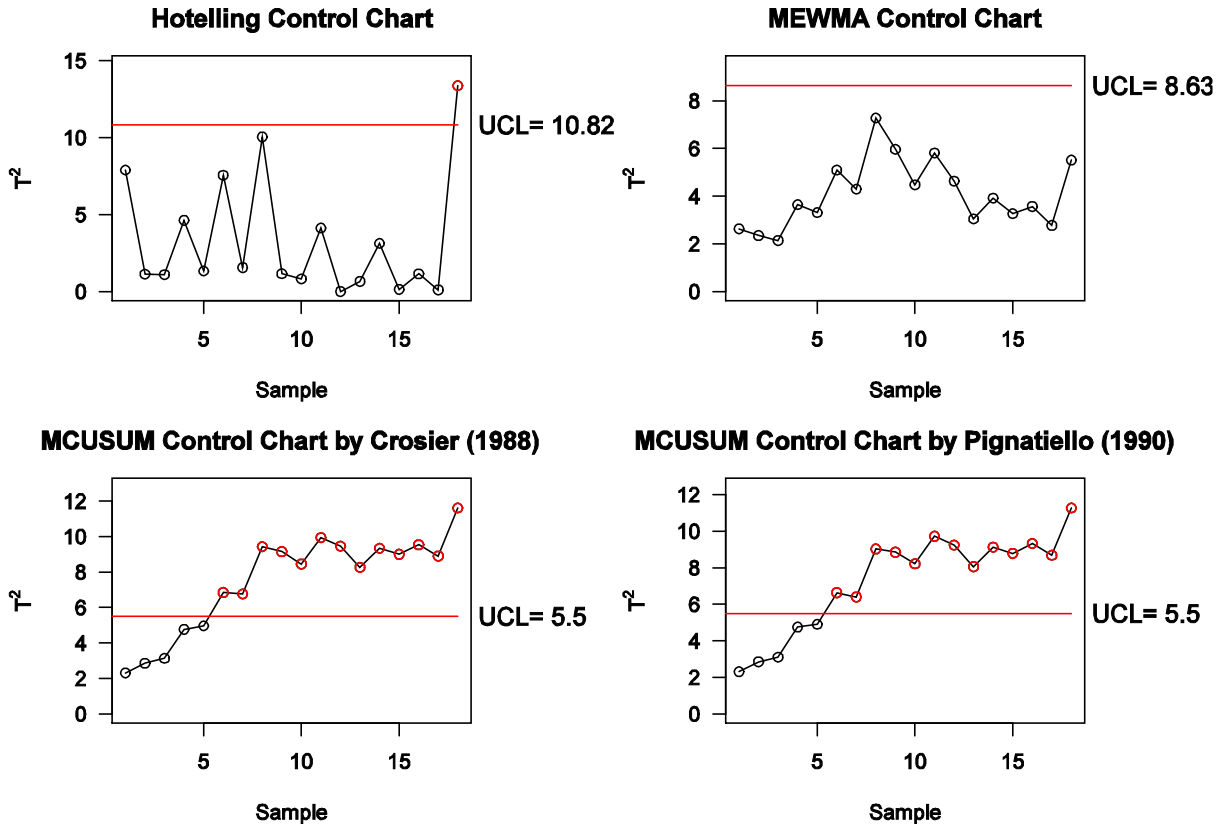

Fig. 3.10 Gráficos de Hotelling, MEWMA y MCUSUM del conjunto archery2 .

The following(s) point(s) fall outside the control limits[1] 18

$‘Decomposition of‘

[1] 18

	t2	decomp	ucl	p-value		
[1,]	11.4353	7.8065		0.0035	1	0
[2,]	0.0008	7.8065		0.9778	2	0
[3,]	13.3752	11.4390		0.0003	1	2

El gráfico de Hotelling detecta la muestra 18 fuera del UCL. La descomposición muestra que la causa se debe a una variación en la horizontal.

Por su parte la gráfica MEWMA no detectó causas especiales mientras que la ficha (Crosier 1988) y (Pignatiello,Runger 1990) presentan una detección temprana de la fuente de variación.

Ilustrando lo engañoso que resultan estas fichas cuando los requisitos no son satisfechos así como puede el mal uso causar ajustes en los procesos cuando no es necesario; se analizará el supuesto de normalidad.

> HZ.test(apply(archery1,1:2,mean))	> HZ.test(apply(archery2,1:2,mean))
0.07 0.73	0.43 0.40
> Royston.test(apply(archery1, 1:2, mean))	> Royston.test(apply(archery2, 1:2, mean))
test.statistic p.value	test.statistic p.value
7.02 0.03	3.49 0.18

Como resultado, fuerte evidencia apunta a rechazar el supuesto de multinormalidad en los primeros datos. Como resultado una transformación a normalidad es requerida.

Usando la transformación de Johnson:

```
> arch.mean1 <- apply(archery1,1:2,mean); arch.mean2 <- apply (archery2, 1:2, mean)
> arch.trans1 <- matrix(0, nrow(arch.mean1), ncol(arch.mean1))
> arch.trans2 <- matrix(0, nrow(arch.mean2), ncol(arch.mean2))

> library("Johnson")
> arch.trans1[,1] <- RE.Johnson(arch.mean1[,1])$transformed; arch.trans1[,2] <- RE.Johnson(arch.mean1[,2])$transformed
> arch.trans2[,1] <- RE.Johnson(arch.mean2[,1])$transformed; arch.trans2[,2] <- RE.Johnson(arch.mean2[,2])$transformed
```

Verificando la normalidad sobre los datos transformados

> HZ.test(arch.trans1)	> HZ.test(arch.trans2)
0.32 0.48	0.99 0.15
> Royston.test(arch.trans1)	> Royston.test(arch.trans2)
test.statistic p.value	test.statistic p.value
2.48 0.29	0.44 0.80

Notar cómo se obtienen p-valores satisfactorios luego de la transformación

Luego verificando la presnesia de autocorrelación.

```
> par(mfrow = c(2,2))
> for( i in 1 : ncol(arch.trans1) ){par(mar = c(4.1,4.5,3,1))
>     acf(arch.trans1[,i],lag = 7,las = 1, main= colnames(arch.trans1)[i])}
> for( i in 1 : ncol(arch.trans2) ){          par(mar = c(4.1,4.5,3,1))
>     acf(arch.trans2[,i],lag = 7,las = 1, main= colnames(arch.trans2)[i])}
```

Como resultado no se encontró dependencia del tiempo. Por tanto, no existe evidencia para rechazar el supuesto de aleatoriedad o independencia.

Luego realizando el mismo análisis con los gráficos de control se obtienen los siguientes resultados: en la etapa de ranking el arquero parece estar en control estadístico. Luego usando esta ronda para controlar la de eliminación (Fase II) no se obtienen causas especiales. Este resultado difiere enormemente del inicial y muestra cómo la falta de normalidad puede producir falsas alarmas en los procesos.

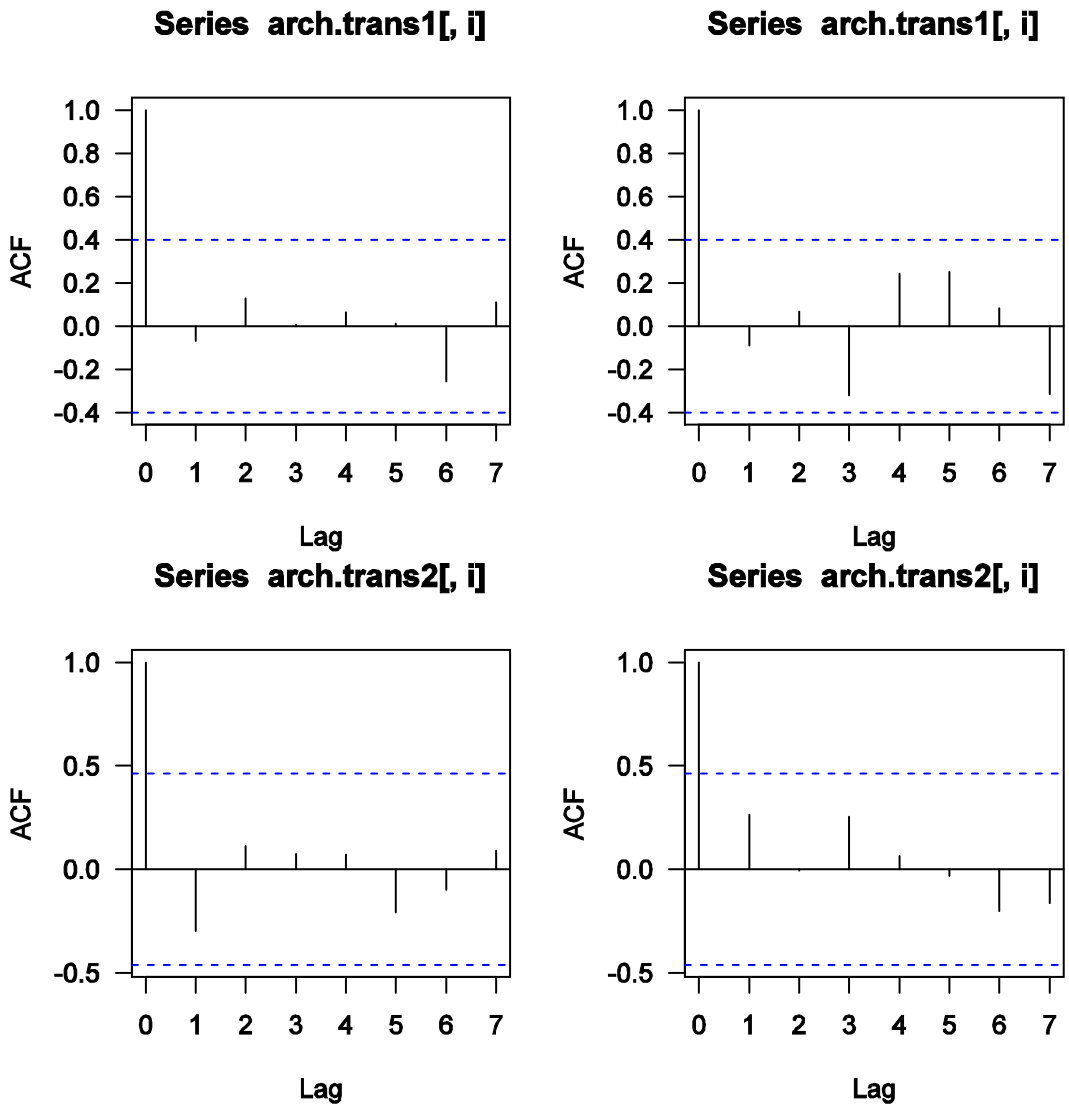

Fig. 3.11 Correlograma de ambos set de datos (archery1 y archery2) luego de la transformación.

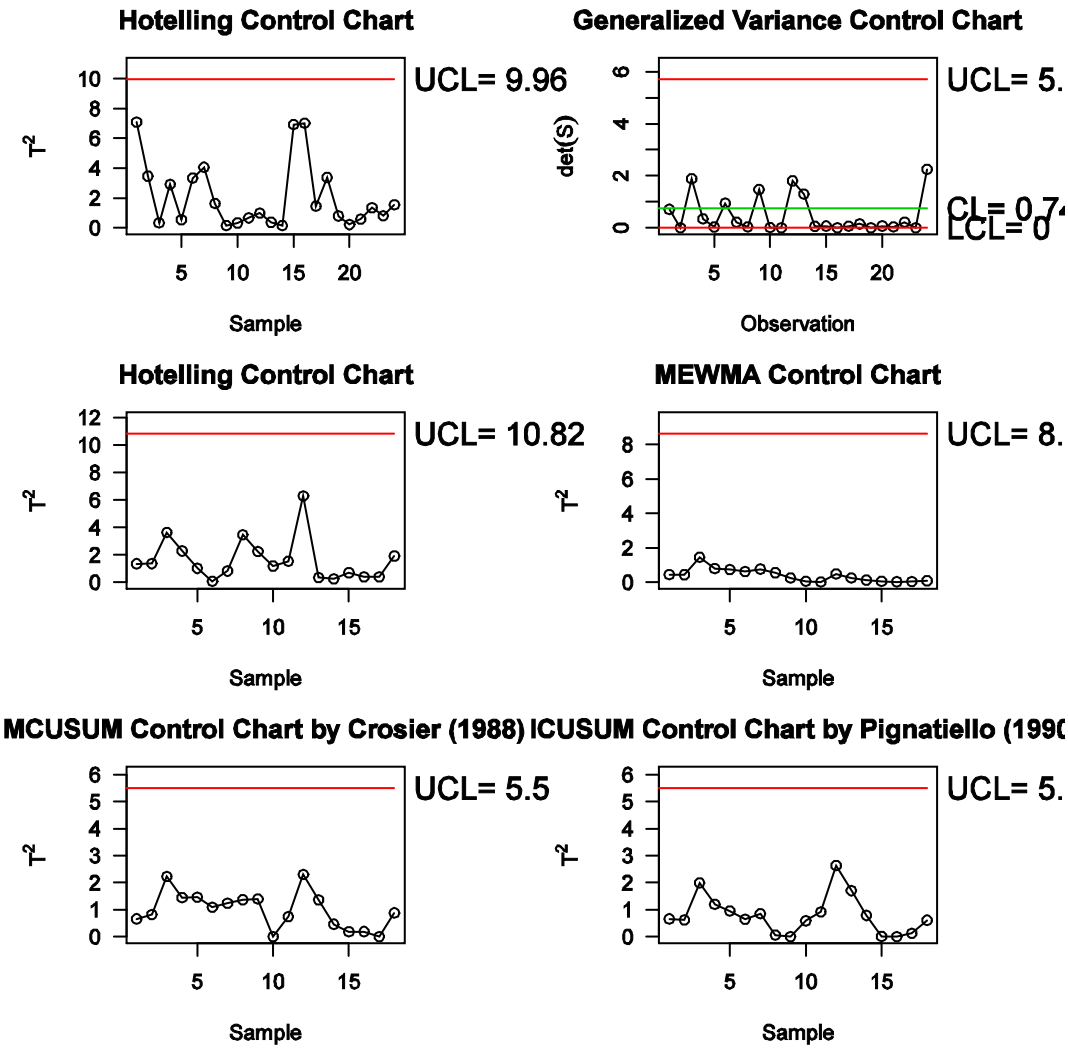

Fig. 3.12 Gráficos de control para ambos sets de datos.

3.3 Reportes de uso.

Usualmente los usuarios de las funciones y paquetes de R reportan vía email al autor o a quien brinda mantenimiento a las mismas. Lo anterior sucede mayoritariamente en los siguientes casos:

- Cuando se reportan “bugs”, errores o mal funcionamiento
- En casos en que el usuario desea que implemente alguna nueva funcionalidad o aplicación.
- Cuando el usuario desea agradecer por el aporte del autor a la comunicada.
- Etc.
-

A continuación mostramos algunos de los usuarios que han reportado uso de la paquetería expuesta en esta investigación:

Arnoldo Frigessi	University of Oslo Section of Biostatistics Institute of Basic Medical Research PO BOX 1122 Blindern N-0317 Oslo, Norway
Martin Boďa	Univerzita Mateja Bela v Banskej Bystrici > Ekonomická fakulta > Katedra kvantitatívnych metód a informačných systémov > Tajovského 10 > 975 90 Banská Bystrica
Didier Murillo	Colombia
Kiran Kumar G	Airbus.(Leading Aircraft Manufacturer) Europe.
Ivania Cerón Souza	Postdoctoral Fellow - Naos Smithsonian Tropical Research Institute Apartado 0843-03092 Panama, República de Panama Tel: +(507) 212 8838

	Fax: +(507) 212 8790
Michele Scagliarini	Dipartimento di Scienze Statistiche Università di Bologna Via Belle Arti 41, 40126 Bologna, Italy Phone: +39 051 2098253 Fax: +39 051 232153
Yi Yin	Department of Ecology College of Urban and Environmental Sciences Peking University, Beijing, P.R. China, 100871 Tel: +86-13811995112
Lukasz Komsta, PhD, DSc	Department of Medicinal Chemistry Medical University of Lublin Jaczewskiego 4, 20-090 Lublin, Poland Fax +48 81 742 36 91

En el caso específico de la publicación (Santos-Fernández,Scagliarini 2012) se puede apreciar en la Web de origen (http://www.jstatsoft.org/v47/i07) la cantidad de descargas:

Paper:	MPCI: An R Package for Computing Multivariate Process Capability Indices	[download] (1194)
Supplements:	MPCI_1.0.4.tar.gz: R source package (application/x-gzip, 12.4 KB)	[download] (159)
Code:	v47i07.R: R example code from the paper (application/octet-stream, 2.9 KB)	[download] (170)

Lo anterior muestra parcialmente el nivel de utilización del paquete MPCI.

Conclusiones y Recomendaciones

Conclusiones

Luego de cumplir con los objetivos de la investigación podemos arribar a las siguientes conclusiones:

1. Queda demostrado que en presencia de procesos con dos más variables correlacionadas se deben realizar análisis de capacidad así como gráficas de control multivariantes
2. El uso del lenguaje y entorno de programación R permite desarrollar herramientas potentes y fáciles de aplicar
3. En nuestro conocimiento los paquetes que acompañan esta investigación resultan las primeras aplicaciones tanto en R como en lenguajes de alto nivel como MATLAB y Octave para tratar con variantes multivariadas.
4. Las herramientas desarrolladas posibilitan solucionar los problemas más frecuentes relativos al MSQC.
5. El uso de estas técnicas contribuyen a implementar el control estadístico incluso en disciplinas como el Béisbol y el Tiro con Arco.

Recomendaciones

A modo de contribución a la utilidad de la investigación se procede a plantear las siguientes recomendaciones.

1. Como estas disciplinas se encuentran en pleno desarrollo, se prevee la implementación de nuevas propuestas tanto de índices como de fichas de control.
2. Implementar las herramientas correspondientes para el diseño experimental con varias variables de respuesta, otra de las subdisciplinas del Control Estadístico de la Calidad Multivariante.

Bibliografía

Bibliografía

1. Alt, F.B.: Multivariate Quality Control. In: Encyclopedia of Statistical Sciences, vol. 6. John Wiley & Sons, (1985)
2. Anderson, T.W.: Asymptotic theory for principal component analysis. The Annals of Mathematical Statistics **34**, 122–148 (1963)
3. Apley, D.W., Tsung, F.: The autoregressive T-squared chart for monitoring univariate autocorrelated processes. Journal of Quality Technology **34**, 80–96 (2002)
4. Bartlett, M.S.: A note on the multiplying factors for various X2 approximations. Journal of the Royal Statistical Society: Series B **16**, 296–298 (1954)
5. Bodden, K.M., Rigdon, S.E.: A Program for Approximating the In-Control ARL for the MEWMA Chart. Journal of Quality Technology **31**, 120-123 (1999)
6. Borror, C.M., Montgomery, D.C., Runger, G.C.: Robustness of the EWMA control chart to non-normality. Journal of Quality Technology **31**(3), 309-316 (1999)
7. Bothe, D.R.: A capability study for an entire product. ASQC Quality Congress Transactions, **46**, 172-178 (1992)
8. Box, G.E.P., Cox, D.R.: An Analysis of Transformations. Journal of the Royal Statistical Society. Series B (Methodological) **26**(2), 211-252 (1964)
9. Box, G.E.P., Jenkins, G.: Time Series Analysis: Forecasting and Control. Holden-Day, (1976)
10. Castagliola, P., Castellanos, J.-V.G.: Capability Indices Dedicated to the Two Quality Characteristics Case. Quality Technology & Quantitative Management **2**(2), 201-220 (2005)
11. Chan, L.K., Cheng, S.W., Spiring, F.A.: A Multivariate Measure of Process Capability. Journal of Modeling and Simulation **1**, 1-6 (1991)
12. Chatfield, C.: The Analysis of Time Series: An Introduction, 4 ed. Chapman & Hall, New York (1989)
13. Chen, H.: A Multivariate Process Capability Index Over a Rectangular Solid Zone. Statistica Sinica **4**, 749-758 (1994)
14. Chen, K.S., Pearn, W.L., Lin, P.C.: Capability Measures for Processes with Multiple Characteristics. Quality and Reliability Engineering International **19**, 101-110 (2003)
15. Chou, Y.-M., Polansky, A.M., L, M.R.: Transforming non-Normal Data to Normality in Statistical Process Control. Journal of Quality Technology **30, 2,** April, 133-141 (1998)
16. Crawley, M.J.: The R Book. Wiley-Blackwell, (2007)
17. Crosier, R.B.: Multivariate Generalizations of Cumulative Sum Quality-Control Schemes. Technometrics **30**(3), 291-303 (1988)
18. D'Agostino, R., Pearson, E.S.: Tests for Departure from Normality. Empirical Results for the Distributions of b2 and √ b1. Biometrika **60**(3), 613-622 (1973)
19. D'Agostino, R.B.: Transformation to normality of the null distribution of g1. Biometrika **57**(3), 679–681 (1970)
20. D'Agostino, R.B., Belanger, A., Jr, R.B.D.A.: A suggestion for using powerful and informative tests of normality. The American Statistician **44**(4), 316–321 (1990)

21. Doganaksoy, N., Faltin, F.W., Tucker, W.T.: Identification of out-of-control multivariate characteristic in a multivariable manufacturing environment. Communications in Statistics—Theory and Methods **20**, 2775–2790 (1991)
22. Healy, J.D.: A Note on Multivariate CUSUM Procedures. Technometrics **29**(4), 409-412 (1987)
23. Henze, N., Zirkler, B.: A Class of Invariant Consistent Tests for Multivariate Normality. Communications in Statistics - Theory and Methods **19**(10), 3595-3617 (1990)
24. Holmes, D.S., Mergen, A.E.: Improving the performance of T-square control chart. Quality Engineering **5**(4), 619-625 (1993)
25. Hotelling, H.: Multivariate Quality Control. McGraw-Hill, (1947)
26. Hubele, N.F., Shahriari, H., Cheng, C.S.: A Bivariate Process Capability Vector. 299-310 (1991)
27. Jackson, J.E.: A User Guide to Principal Components. John Wiley & Sons, New York (1991)
28. Jarque, C.M.: Jarque-Bera Test. In: Lovric, M. (ed.) International Encyclopedia of Statistical Science. Springer, (2010)
29. Jarque, C.M., Bera, A.K.: Efficient tests for normality, homoscedasticity and serial independence of regression residuals. Economics Letters **6**(3), 255–259 (1980)
30. Jarque, C.M., Bera, A.K.: A test for normality of observations and regression residuals. International Statistical Review **55**(2), 163–172 (1987)
31. Johnson, N.L.: Systems of Frequency Curves Generated by Methods of Translation. (1949)
32. Juran, J.M., Godfrey, A.B.: Juran's Quality Handbook. (1998)
33. Kalagonda, A.A., Kulkarni, S.R.: Multivariate quality control chart for autocorrelated processes. Journal of Applied Statistics **31**(3), 317–327 (2004)
34. Kotz, S., Lovelace, C.R.: Process Capability Indices in Theory and Practice. Hodder Education Publishers, (1998)
35. Lowry, C.A., Montgomery, D.C.: A review of multivariate control charts. IIE Transactions **27**(6), 800-810 (1995)
36. Lowry, C.A., Woodall, W.H., Champ, C.W., Rigdon, S.E.: A Multivariate Exponentially Weighted Moving Average Control Chart. Technometrics **34**(1), pp. 46-53 (1992)
37. Mardia, K.V.: Measures of multivariate skewness and kurtosis. Biometrika **57**, 519-530 (1970)
38. Mardia, K.V.: Applications of some measures of multivariate skewness and kurtosis for testing normality and robustness studies. Sankhya **36**, 115–128 (1974)
39. Mason, R., Tracy, N., Young, J.: Monitoring a multivariate step process. Journal of Quality Technology **28**, 39–50 (1996)
40. Mason, R.L., Tracy, N.D., Young, J.C.: Decomposition of T-square for multivariate control chart interpretation. Journal of Quality Technology **27**, 99–108 (1995)
41. Mason, R.L., Young, J.C.: Multivariate Statistical Process Control with Industrial Application, 1 ed. Society for Industrial and Applied Mathematics, (2001)
42. Mecklin, C.J., Mundfrom, D.J.: An Appraisal and Bibliography of Tests for Multivariate Normality. International Statistical Review **72**(1), 123–138 (2004)
43. MLB. http://gd2.mlb.com/components/game/mlb/.

44. MLB: The Strike Zone: A historical timeline http://mlb.mlb.com/mlb/official_info/umpires/strike_zone.jsp.
45. Montgomery, D.C.: Introduction to Statistical Quality Control, 5 ed. John Wiley & Sons, (2004)
46. Montgomery, D.C.: Introduction to Statistical Quality Control, 5 ed. John Wiley & Sons, (2005)
47. Murphy, B.J.: Selecting out-of-control variables with T-squared multivariate quality procedures. The Statistician **36**, 571–583 (1987)
48. Nickerson, D.M.: Construction of a Conservative Confidence Region from Projections of an Exact Confidence Region in Multiple Linear Regression. The American Statistician **48**(2), 120-124 (1994)
49. NIST / SEMATECH e-Handbook of Statistical Methods. http://www.itl.nist.gov/div898/handbook/
50. Page, E.S.: Cumulative Sum Charts. Technometrics 3(1), 1-9 (1961)
51. Pan, J.-N., Lee, C.-Y.: New capability indices for evaluating the performance of multivariate manufacturing processes. Quality and Reliability Engineering International **26**(1), 3–15 (2010)
52. Pearn, W., Kotz, L.S., Johnson, N.L.: Distributional and Inferential Properties of Process Capability Indices. Journal of Quality Technology **24**, 216–231 (1992)
53. Pearn, W.L., Kotz, S.: Encyclopedia And Handbook of Process Capability Indices: A Comprehensive Exposition of Quality Control Measures. World Scientific Pub Co Inc, (2006)
54. Pignatiello, J., Runger, G.: Comparisons of Multivariate CUSUM Charts. Journal of Quality Technology **22**(3), 173-186 (1990)
55. Prabhu, S.S., Runger, G.C.: Designing a multivariate EWMA control chart. Journal of Quality Technology **29**, 8-15 (1997)
56. Rencher, A.C.: Methods of Multivariate Analysis. John Wiley & Sons, (2002)
57. Roberts, S.W.: Control chart tests based on geometric moving averages. Technometrics **42**(1), 97-102 (1959)
58. Royston, J.P.: An Extension of Shapiro and Wilk's W Test for Normality to Large Samples. Applied Statistics **31**(2), 115–124 (1982)
59. Royston, J.P.: Some Techniques for Assessing Multivarate Normality Based on the Shapiro- Wilk W. Journal of the Royal Statistical Society. Series C (Applied Statistics) **32**(2), 121-133 (1983)
60. Royston, J.P.: Approximating the Shapiro-Wilk W-Test for non-normality. Statistics and Computing **2**(3), 117-119 (1992)
61. Royston, J.P.: Remark AS R94: A remark on Algorithm AS 181: The W test for normality. Journal of the Royal Statistical Society. Series C (Applied Statistics) **44**(4), 547-551 (1995)
62. Santos-Fernández, E.: Multivariate Statistical Quality Control Using R. Springer, (2013)
63. Santos-Fernández, E., Scagliarini, M.: MPCI: An R Package for Computing Multivariate Process Capability Indices. Journal of Statistical Software **47**(7), 1-15 (2012)
64. Scagliarini, M.: Multivariate process capability using principal component analysis in the presence of measurement errors. AStA Advances in Statistical Analysis **95**(2), 757-765 (2011)

65. Shahriari, H., Hubele, N.F., Lawrence, F.P.: A Multivariate Process Capability Vector. Proceedings of the 4th Industrial Engineering Research Conference **1**, 304-309 (1995)
66. Shapiro, S., Wilk, M.: An analysis of variance test for normality. Biometrika **52**, 591–611 (1965)
67. Shinde, R.L., Khadse, K.G.: Multivariate process capability using principal component analysis. Quality and Reliability Engineering International **25**(1), 69–77 (2008)
68. Slifker, J.F., Shapiro, S.S.: The Johnson system: selection and parameter estimation. Technometrics **22**(2), pp. 239-246 (1980)
69. Sullivan, J.H., Woodall, W.H.: A Comparison of Multivariate Control Charts for Individual Observations. Journal of Quality Technology **28**, 398-408 (1996a)
70. Sullivan, J.H., Woodall, W.H.: A Comparison of Multivariate Quality Control Charts for Individual Observations. Journal of Quality Technology **28(4)** (1996b)
71. Taam, W., Subbaiah, P., Liddy, W.J.: A Note on Multivariate Capability Indices. Journal of Applied Statistics **20**, 339-351 (1993)
72. Tano, I., Vännman, K.: Comparing Confidence Intervals for Multivariate Process Capability Indices. Quality and Reliability Engineering International **28**(4), 481–495 (2011)
73. Testik, M.C., Runger, G.C.: Mining Manufacturing Quality Data. In: Ye, N. (ed.) Handbook of Data Mining Lawrence Erlbaum Associates Publishers, New Jersey (2003)
74. Thode, H.C.: Testing For Normality. Marcel Dekker, (2002)
75. Thode, H.C.: Normality Tests. In: Lovric, M. (ed.) International Encyclopedia of Statistical Science Springer, (2010)
76. Tracy, N.D., Young, J.C., Mason, R.L.: Multivariate control charts for individual observations. Journal of Quality Technology **24**, 88–95 (1992)
77. Velilla, S.: A note on the multivariate Box-Cox transformation to normality. Statistics and Probability Letters **17**, 259-263 (1993)
78. Venables, W.N., Ripley, B.D.: Modern Applied Statistics with S. Fourth Edition. Springer, (2002)
79. Wang, C.H.: Constructing Multivariate Process Capability Indices for Short-Run Production. The International Journal of Advanced Manufacturing Technology **26**, 1306-1311 (2005)
80. Wang, F.K., Chen, J.C.: Capability Index Using Principal Components Analysis. Quality Engineering **11**, 21-27 (1998)
81. Wang, F.K., Hubele, N., Lawrence, F.P., Miskulin, J.D., Shahriari, H.: Comparison of Three Multivariate Process Capability Indices. Journal of Quality Technology **32**, 263-275 (2000)
82. Weisberg, S.: Applied Linear Regression, 3 ed. Wiley/Interscience, (2005)
83. Wierda, S.J.: A multivariate process capability index. In: ASQC Quality Congress Transactions. 342-348 (1993)
84. Wierda, S.J.: Multivariate statistical process control—recent results and directions for future research. Statistica Neerlandica **48**, 147–168 (1994)
85. Woodall, W.H., Ncube, M.M.: Multivariate CUSUM Quality-Control Procedures. Technometrics **3**(3), 285-292 (1985)

86. Xekalaki, E., Perakis, M.: The Use of Principal Component Analysis in the Assessment of Process Capability Indices. Proceedings of the Joint Statistical Meetings of the American Statistical Association, The Institute of Mathematical Statistics, The Canadian Statistical Society. New York. Marcel Dekker, New York (2002)
87. Yum, B.-J., Kim, K.-W.: A bibliography of the literature on process capability indices: 2000–2009. Quality and Reliability Engineering International **27**(3), 251–268 (2012)