



**CIMAT**

---

---

Centro de Investigación en Matemáticas, A.C.

# **MONITOREO DE DATOS FUNCIONALES**

**Tesis**

Que para obtener el Grado de:

**Maestro en Ciencias con Orientación en Probabilidad y  
Estadística**

**P R E S E N T A:**

Rodrigo Plazola Ortiz

Directores:

Dr. Rogelio Ramos Quiroga

Dr. Enrique Raúl Villa Diharce

Guanajuato, Guanajuato, México

Diciembre del 2013

# Índice general

Índice general . . . . .	I
Índice de figuras . . . . .	III
Índice de tablas . . . . .	V
Agradecimientos . . . . .	V
Prefacio . . . . .	VI
<b>1. Cartas de Control</b>	<b>1</b>
1.1. Introducción . . . . .	1
1.2. Cartas de control . . . . .	2
1.3. Fases en la construcción de cartas de control. . . . .	4
1.4. Riesgo del muestreo . . . . .	5
1.5. Tipos de cartas de control . . . . .	6
1.6. Cartas de control multivariadas . . . . .	7
1.7. Cartas de control multivariadas $T^2$ de Hotelling . . . . .	9
1.7.1. Límite de control . . . . .	10
1.8. Carta de control multivariada $T^2$ de Hotelling para observaciones individuales	13
1.9. Longitud promedio de corrida (ARL) . . . . .	16
<b>2. Análisis de datos funcionales</b>	<b>18</b>
2.1. Introducción . . . . .	18
2.2. Datos funcionales . . . . .	19
2.3. Bases . . . . .	20
2.4. Splines . . . . .	22
2.4.1. B-Spline . . . . .	23
2.5. Splines suavizados . . . . .	26
2.6. Construcción de datos funcionales . . . . .	28
2.7. Construcción de datos funcionales con spline suavizados . . . . .	29
2.8. Estadísticas descriptivas para datos funcionales . . . . .	32

---

<b>3. Carta de control para datos funcionales</b>	<b>34</b>
3.1. Introducción . . . . .	34
3.2. Descripción del problema . . . . .	35
3.3. ARL para carta de control . . . . .	42
3.4. Carta de control para datos funcionales . . . . .	48
3.5. ARL de carta de control para datos funcionales . . . . .	53
3.6. Áreas Importantes . . . . .	57
3.7. ¿Qué sucede cuando no se tiene normalidad? . . . . .	59
3.8. Ejemplo . . . . .	62
<b>4. Conclusiones</b>	<b>70</b>
<b>Bibliografía</b>	<b>72</b>
<b>A. Estimaciones para el modelo de efectos mixtos.</b>	<b>74</b>
<b>B. Construcción de las funciones bases B-spline.</b>	<b>78</b>
<b>C. Cálculo de ARL con perturbación.</b>	<b>82</b>

# Índice de figuras

1.1. Elementos de una carta de control . . . . .	3
1.2. En figura a) Carta de control que muestra evidencia de proceso que está en control, mientras que en la figura b) da señal fuera de control. . . . .	4
1.3. Idea inicial de cartas multivariadas . . . . .	8
1.4. Distancia de Mahalanobis . . . . .	10
1.5. Carta de control $T^2$ de Hotelling . . . . .	14
2.1. Ejemplo de mediciones sobre una tabla. . . . .	20
2.2. Mediciones de 24 tablas, vistas como datos funcionales . . . . .	21
2.3. B-spline . . . . .	26
3.1. Datos obtenidos de <i>Amiri[1]</i> . . . . .	36
3.2. Ajuste de modelo cuadrático . . . . .	38
3.3. Ajuste de modelo de efectos mixtos . . . . .	41
3.4. Carta de control, fase I. . . . .	42
3.5. 20 simulaciones . . . . .	44
3.6. Ajuste B-spline . . . . .	50
3.7. Gráfico Q-Q . . . . .	50
3.8. Carta de control fase I . . . . .	51
3.9. Histograma de 10000 estadísticos $T^2$ . . . . .	52
3.10. Histograma de 10000 RLs . . . . .	54
3.11. Ajuste considerando nodosIM . . . . .	58
3.12. Ajuste considerando nodosN . . . . .	59
3.13. Método Bootstrap . . . . .	61
3.14. Curvas de 129 muestras de carne . . . . .	63
3.15. Ajuste B-Spline . . . . .	64
3.16. Curvas de las muestras con grasa $< 12\%$ . . . . .	64

3.17. Ajuste B-spline de las muestras con grasa <12% . . . . .	65
3.18. Carta de control Fase I . . . . .	66

# Índice de tablas

3.1. Estimaciones del modelo cuadrático . . . . .	37
3.2. Estimaciones del modelo de efectos mixtos . . . . .	40
3.3. ARL bajo control estadístico . . . . .	45
3.4. Resultados ARL . . . . .	46
3.5. Resultados ARL . . . . .	47
3.6. Coeficientes del ajuste <i>B – spline</i> . . . . .	49
3.7. ARL bajo control . . . . .	53
3.8. Resultados ARL, sólo considerando zona de interés. . . . .	55
3.9. Resultados ARL . . . . .	56
3.10. Resultados ARL . . . . .	60
3.11. Resultados de carta de control . . . . .	67
3.12. Mejor clasificación . . . . .	67
B.1. valores de RMP . . . . .	78
B.2. Matriz B-spline . . . . .	81

# Agradecimientos

Es para mi un verdadero placer utilizar este espacio para expresar mis más sinceros agradecimientos por la participación directa e indirecta de personas e instituciones, que han facilitado las cosas para que este trabajo llegue a un feliz término y más allá que me han permitido terminar mis estudios de maestría en una institución tan prestigiosa como lo es CIMAT, de la cual, estoy muy agradecido por estos años de enseñanza.

Primero deseo agradecer a mis padres por ser el pilar fundamental en todo lo que soy, en toda mi educación, tanto académica, como en la vida, por su incondicional apoyo perfectamente mantenido a través del tiempo, gracias por su inmenso amor y comprensión. Todo este trabajo ha sido posible gracias a ellos.

A mi esposa Perla y mi hija Azul por acompañarme en este largo camino lleno de dificultades, gracias por su paciencia, comprensión y su inmensurable amor.

Sinceras gracias a todos mis maestros en particular al Dr. Rogelio Ramos Quiroga, al Dr. Enrique Raúl Villa Diharce y al Dr. Yehua Li, por la paciencia, por los consejos, el apoyo y por la motivación para la culminación de mis estudios de maestría y para la elaboración de esta tesis.

Por último aunque no menos importante agradezco al Consejo de Ciencia y Tecnología (CONACYT), al Centro de Investigación en Matemáticas (CIMAT), y a la Universidad Estatal de Iowa (ISU) por su apoyo y patrocinio para la realización de mis estudios de maestría así como para la realización de este proyecto de tesis.

Infinitas gracias a todos.

# Prefacio

El objetivo principal de esta tesis es el estudio de las cartas de control para datos funcionales, las cuales se utilizan cuando la calidad de los artículos de un proceso se puede expresar como datos funcionales, por ejemplo, cuando la calidad de un motor depende del torque producido en diferentes valores de Revoluciones por minuto (RPM). En estos casos un artículo clasificado como defectuoso es aquel en que su correspondiente dato funcional, no tenga la misma “forma” que otros artículos que son clasificados como no defectuosos, es decir, es una comparación entre datos funcionales (curvas).

La carta de control funcional propuesta en esta tesis se basa en los coeficientes de un ajuste B-spline a los datos que se tienen del proceso, algo interesante es que esto ayuda en el análisis de grandes conjuntos de datos, en donde las herramientas clásicas pierden efectividad, debido a “la maldición de la dimensionalidad”, la causa común de estos problemas es que cuando aumenta la dimensionalidad, el volumen del espacio aumenta exponencialmente haciendo que los datos disponibles se vuelvan dispersos. Esta dispersión es problemática para cualquier método que requiera significación estadística.

Al trabajar con coeficientes de B-spline reducimos notablemente la dimensionalidad, entonces, la idea central es caracterizar a los datos funcionales mediante coeficientes de un B-spline y utilizar dichos coeficientes para comparar y monitorear las unidades que genera el proceso. De esta manera podemos identificar si el proceso está bajo control estadístico, si no lo está se investigan las causas y se corrigen para mejorar o mantener el proceso.



# Capítulo 1

## Cartas de Control

### 1.1. Introducción

La calidad se está convirtiendo en factor fundamental en la decisión del consumidor de muchos productos y servicios, independientemente de si el consumidor es un individuo, una corporación industrial o una tienda de autoservicio. Por consiguiente, la calidad es un factor clave que conduce al éxito empresarial, el crecimiento y la mejora de la posición competitiva.

Por todo lo anterior las empresas han recurrido a la estadística como parte de su estrategia para mejorar sus procesos, en esta área existe un conjunto de procedimientos que coadyuvan al mejoramiento y control de los procesos industriales, llamado Control Estadístico de Procesos (CEP). Dentro de este conjunto de procedimientos podemos encontrar herramientas gráficas de gran utilidad usadas para monitorear los procesos, las cartas de control. En este capítulo , daremos una pequeña introducción a la teoría de cartas de control para monitorear procesos industriales, enfocándonos en las cartas de control multivariadas  $T^2$  de Hotelling.

## 1.2. Cartas de control

Una Carta de Control (**CC**) es una herramienta estadística usada principalmente para el estudio y control de procesos repetitivos, de esta manera la carta de control provee información importante para mejorar los procesos. Recordemos que uno de los objetivos del control estadístico de procesos es eliminar la variabilidad en el proceso, aunque en muchas ocasiones no se puede eliminar por completo. Las CC son una herramienta efectiva para reducir la variabilidad lo más que se pueda.

La base de la teoría de las CC es que la variabilidad de un proceso se clasifica en dos grandes grupos:

- **Variaciones originadas por causas aleatorias:** Esta variabilidad esta presente en el proceso independientemente de lo bien diseñado o de lo cuidadosamente mantenido, en otras palabras, es la variación inherente de la calidad o del proceso y sobre esta variabilidad nada se puede hacer.
- **Variaciones originadas por causas especiales o asignables:** Esta variabilidad proviene generalmente de tres fuentes: las maquinas mal ajustadas, errores del operador, o materias primas defectuosas, y por lo general contribuyen a un mal rendimiento del proceso por ello es importante identificarlo y eliminarlo.

En la Figura 1.1, se muestra una CC típica, en la cual se grafica una estadística  $y$  que mide la característica de interés en el proceso, obtenida de una muestra, contra el número de la muestra o el tiempo. Es habitual conectar los puntos  $y$  en la carta de control mediante segmentos de linea recta, de este modo es más fácil visualizar como la secuencia de puntos se comporta en el tiempo.

Una carta de control contiene:

- **Línea Central (LC):** La cual representa el promedio de los valores de la estadística  $y$ , correspondientes a un estado bajo control del proceso.
- **Línea de control Superior (LCS) y Línea de control Inferior (LCI):** Estos límites representan el rango de variabilidad del proceso y estos límites de control se escogen de tal manera de que si el proceso esta bajo control, exista una alta probabilidad de que los puntos  $(y_i, i)$  queden dentro de estos límites.

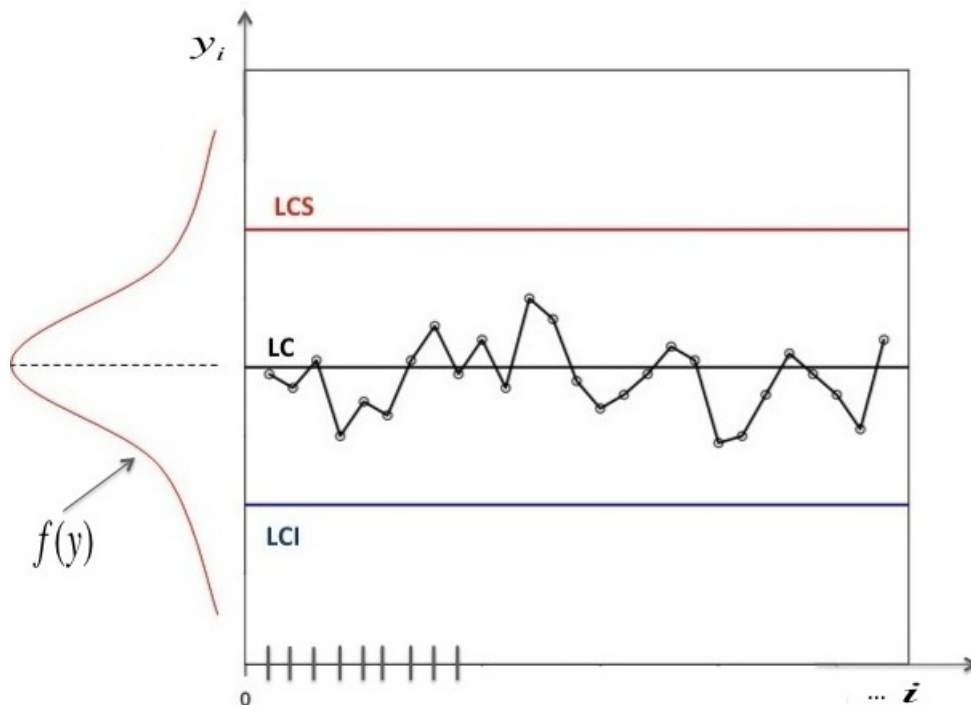


Figura 1.1: Elementos de una carta de control típica.

Notemos que estos límites sólo dependen del proceso y no de lo que se desea obtener del proceso, es decir, no se deben confundir con aquellos límites que definen la frontera de conformidad para una unidad individual generada por el proceso, llamados límites de especificación.

Las cartas de control permiten considerar al proceso como:

- **Proceso en control:** Si los puntos graficados caen dentro de los límites de control, entonces, el proceso se considera como proceso en control y ninguna acción es necesaria.
- **Proceso fuera de control:** Si un punto cae fuera de los límites de control o si se observan variaciones no aleatorias dentro de los límites de control, el proceso se considera fuera de control con respecto a su tendencia central. Entonces, se debe iniciar una búsqueda de las causas asignables de esta variación.

Algunas veces a las cartas de control se agregan dos rectas paralelas a la línea central usualmente localizadas a una distancia de  $2\sigma$ , llamadas límites de aviso. Si 2 de 3 puntos consecutivos aparecen entre los límites de aviso y de control, el proceso debe supervisarse con mayor cuidado o detenerse. En la Figura 1.2 se presentan dos cartas de control que muestran evidencia de procesos dentro y fuera de control.

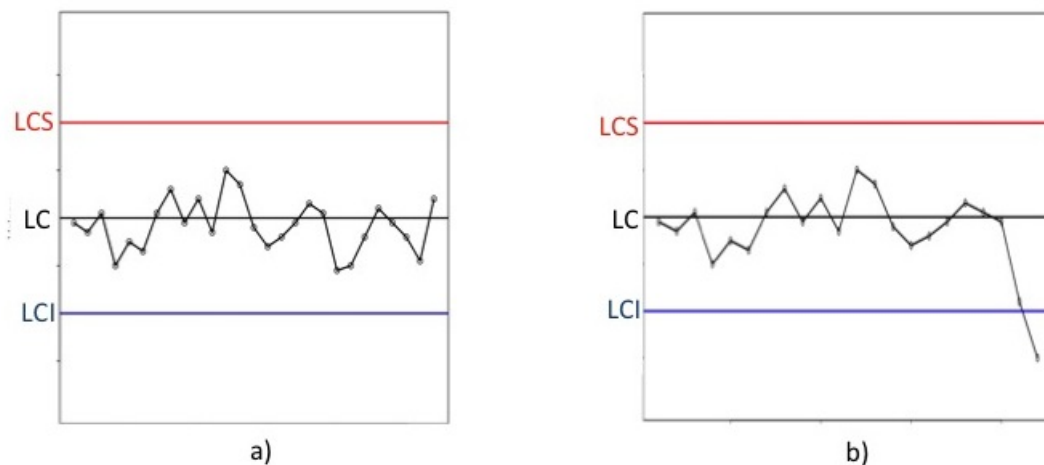


Figura 1.2: En figura a) Carta de control que muestra evidencia de proceso que está en control, mientras que en la figura b) da señal fuera de control.

### 1.3. Fases en la construcción de cartas de control.

En la construcción usual de cartas de control se utilizan dos Fases, cada una con

objetivos diferentes. En la Fase I, un conjunto de observaciones del proceso se recopilan y analizan, suponiendo que dicho proceso está en control, a estas observaciones se les conoce como datos históricos del proceso y una vez analizados nos servirán para estimar los parámetros, si es que se desconocen. En esta Fase I, se construye una carta de control, escogiendo los límites de control de manera que nos ayuden a determinar si el proceso ha estado en control, durante el tiempo en que se escogieron estos datos históricos y para determinar si es fiable utilizar estos datos para controlar la producción futura.

La Fase II inicia una vez que tengamos datos históricos “limpios” del proceso, en estables condiciones y representativos de un proceso en control. En la Fase II, utilizamos una carta de control para monitorear el proceso, comparando los nuevos conjuntos de observaciones con límites de control, los cuales, fueron estimados utilizando los datos históricos.

En la Fase I obtenemos  $n$  subgrupos de observaciones, los cuales analizamos construyendo límites de control con estos mismos subgrupos. Usualmente, se utiliza  $n = 20$  ó  $25$  subgrupos. En la Fase I se supone que el proceso está en control, el objetivo es llevar el proceso a un control estadístico, este conjunto de datos históricos se analizan y si resulta que el proceso está fuera de control, entonces, se analizan las causas y se corrigen en el proceso, luego se vuelven a obtener muestras y esto se repite hasta obtener un conjunto de datos históricos bajo control estadístico del proceso.

## 1.4. Riesgo del muestreo

Existe una conexión entre las cartas de control y las pruebas de hipótesis. Esencialmente una carta de control es una prueba de hipótesis, en donde las hipótesis nula e hipótesis alternativa son:

$$H_0 : \text{El proceso está bajo control} \quad \text{y} \quad H_1 : \text{El proceso está fuera de control},$$

un punto graficado dentro de los límites de control es equivalente a no rechazar la hipótesis nula, y un punto graficado fuera de los límites de control es equivalente a rechazar la hipótesis nula.

En una carta de control, después de calcular la estadística utilizada en la construcción y de acuerdo a un criterio utilizado para juzgar si el proceso está en control, concluimos que el proceso está en control o fuera de control. Por la aleatoriedad de los procesos e igual que en las pruebas de hipótesis, corremos el riesgo de equivocarnos:

- **Riesgo tipo I:** Es el riesgo de que una muestra extrema conduzca a tomar una acción, cuando no ha ocurrido cambio en el proceso.
- **Riesgo tipo II:** Es el riesgo de que una muestra caiga dentro de los límites de control, aunque haya ocurrido un cambio en el proceso.

## 1.5. Tipos de cartas de control

Las cartas de control se pueden clasificar en dos grupos generales: para variables y para atributos. Si la característica de calidad se puede medir o puede ser expresado por un número sobre una escala continua, es usualmente llamada para variable, las cuales requieren de un instrumento de medición (Volumen, Voltaje, Peso etc.).

Las cartas de control para variables más usadas son:

- $\bar{X}$  (Promedios).
- $R$  (Rangos).
- $S$  (Desviación estándar).
- $X$  (Medidas individuales).

- $T^2$  de Hotelling (Multivariadas).

Algunas características de calidad no son medidas sobre una escala continua, en este caso juzgamos cada unidad generada por el proceso como conforme o no conforme, en base a que presente o no cierto atributo o se cuenta el número de defectos que aparecen en cada unidad. Las cartas de control utilizando este tipo de características de calidad son llamadas cartas de control para atributos.

Las cartas de control para atributos mas usadas son:

- $p$  (Proporción o fracción de artículos defectuosos).
- $np$  (Número de unidades defectuosas).
- $c$  (Número de defectos).
- $u$  (Número de defectos por unidad).

Además de estas cartas de control existen una amplia gama de cartas de control. En el presente trabajo se hablará principalmente de cartas de control multivariadas  $T^2$  de Hotelling, para mayor referencia de las otras cartas de control se puede consultar el libro *Montgomery* (1991)[8]

## 1.6. Cartas de control multivariadas

En todo el análisis anterior consideramos cartas univariadas, es decir, sólo interesaba monitorear una característica de calidad. Pero en algunos casos, es necesario monitorear varias características de calidad simultáneamente, donde la respuesta esta dada por un

vector que mide  $p$  características de calidad  $x^T = (x_1, x_2, \dots, x_p)$ . En estos casos decimos que tenemos un problema de control estadístico de procesos multivariados.

En el monitoreo de  $p$  características de calidad, la idea inicial fue tomar una carta de control para cada una de las características de calidad, como lo muestra la Figura 1.6 para dos características de calidad  $x^T = (x_1, x_2)$ , en este tipo de cartas de control se tenían dos criterios de salida de control:

- Un punto  $x = (x_1, \dots, x_p)$  es una señal fuera de control si al menos una coordenada se sale de control.
- Un punto es una señal fuera de control si todas sus coordenadas están fuera de control.

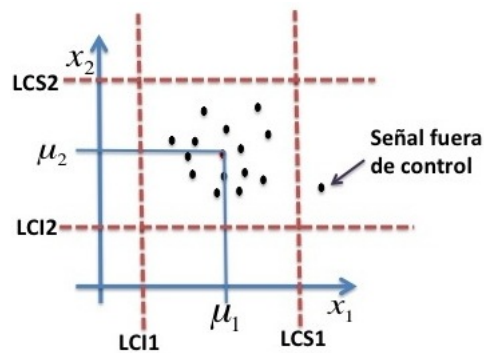


Figura 1.3: Idea inicial de cartas multivariadas



El problema con este enfoque es que no se toma en cuenta que existe una correlación entre las características de calidad, donde es claro que en la mayoría de los casos se tiene una cierta correlación debido a que las características de calidad se están midiendo del mismo objeto. Después se introdujeron las cartas  $T^2$  de Hotelling para el monitoreo de procesos multivariados, estas cartas sí consideran la correlación entre las diferentes variables

## 1.7. Cartas de control multivariadas $T^2$ de Hotelling

Cuando se desea controlar varias características simultáneas, la herramienta más usada es la carta de control multivariada propuesta por Hotelling basada en el estadístico  $T^2$  de Hotelling. Aquí la suposición fuerte es que las mediciones sucesivas de las características siguen una distribución normal multivariada. Sin embargo, la normalidad de las observaciones no siempre puede suponerse.

Las cartas de control multivariadas están diseñadas para monitorear el proceso de producción de un producto, el cual interesa controlar, con  $p$  variables de calidad posiblemente correlacionadas. Consideremos el caso en que se miden 2 características  $X_1$  y  $X_2$  a un cierto producto, supongamos que las dos características de calidad están distribuidas conjuntamente como una normal bivariada, una muestra de  $n$  artículos se seleccionan del proceso en intervalos regulares y se obtienen medidas de ambas variables.

Cuando interesa controlar la media del proceso, un método consiste en obtener la distancia del vector bivariado medio de cada muestra al vector de la media del proceso, para esto tomamos la distancia de Mahalanobis, la Figura 1.4 muestra ésta idea, la distancia de Mahalanobis se define como:

**Definición 1.6.1. Distancia de Mahalanobis.** La distancia de Mahalanobis entre dos vectores aleatorios  $\vec{X}$  y  $\vec{Y}$  con la misma distribución de probabilidad y con matriz de

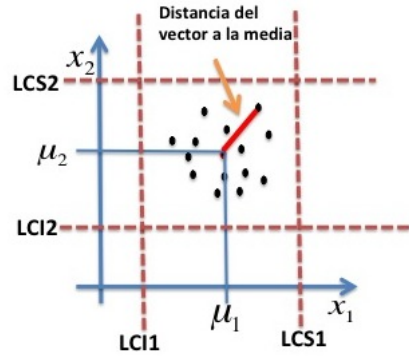


Figura 1.4: Distancia de Mahalanobis

varianzas y covarianzas  $\Sigma$  es:  $d_p(\vec{X}, \vec{Y}) = \sqrt{(\vec{X} - \vec{Y})^T \Sigma^{-1} (\vec{X} - \vec{Y})}$ .

### 1.7.1. Límite de control

En esta subsección, explicaremos como se elige el límite de control para una carta de control multivariada  $T^2$  de Hotelling, lo que tenemos en este tipo de cartas de control son observaciones  $x_1, x_2, \dots, x_n$   $p$ -variadas donde  $x_1, x_2, \dots, x_n \sim N_p(\mu, \Sigma)$ , donde  $N_p(\mu, \Sigma)$  denota una distribución Normal  $p$ -variada con vector de medias  $\mu$  y matriz de varianzas  $\Sigma$ .

La interpretación de  $x_1, x_2, \dots, x_n$  ( $p$ -variadas) es que cada  $x_i$  indica la medición de  $p$  características de calidad al artículo  $i$ , y el conjunto  $x_1, x_2, \dots, x_n$  es la muestra del proceso que se toma cada cierto intervalo de tiempo. Sean  $x_1, x_2, \dots, x_n$  variables aleatorias

independientes con distribución  $x_1, x_2, \dots, x_n \sim N_p(\mu, \Sigma)$ . Definimos,

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

como la media muestral. Demostraremos que

$$n(\bar{x} - \mu)^T \Sigma^{-1} (\bar{x} - \mu) \sim \chi_p^2,$$

donde  $\chi_p^2$  es una distribución chi-cuadrada con  $p$  grados de libertad, como suponemos que  $x_1, x_2, \dots, x_n \sim N_p(\mu, \Sigma)$ , entonces,  $\bar{x} \sim N_p(\mu, \Sigma/n)$ . Luego para demostrar que  $T^2 = n(\bar{x} - \mu)^T \Sigma^{-1} (\bar{x} - \mu) \sim \chi_p^2$ , usaremos la función generadora de momentos:

$$\begin{aligned} M_{T^2}(t) &= E(e^{tT^2}) \\ &= E\left(e^{tn(\bar{x}-\mu)^T \Sigma^{-1} (\bar{x}-\mu)}\right) \\ &= \int e^{tn(\bar{x}-\mu)^T \Sigma^{-1} (\bar{x}-\mu)} (2\pi)^{-\frac{p}{2}} |\Sigma/n|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}n(\bar{x}-\mu)^T \Sigma^{-1} (\bar{x}-\mu)\right\} dx_1 \dots dx_p \\ &= \int (2\pi)^{-\frac{p}{2}} |\Sigma/n|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}n(\bar{x}-\mu)^T (\Sigma^{-1} - 2t\Sigma^{-1})(\bar{x}-\mu)\right\} dx_1 \dots dx_p \\ &= |\Sigma^{-1} - 2t\Sigma^{-1}|^{-1}/n^{\frac{1}{2}} |\Sigma/n|^{-\frac{1}{2}} \int (2\pi)^{-\frac{p}{2}} |(\Sigma^{-1} - 2t\Sigma^{-1})^{-1}/n|^{-\frac{1}{2}} \\ &\quad * \exp\left\{-\frac{1}{2}n(\bar{x}-\mu)^T (\Sigma^{-1} - 2t\Sigma^{-1})(\bar{x}-\mu)\right\} dx_1 \dots dx_p \\ &= |(\Sigma^{-1} - 2t\Sigma^{-1})^{-1}/n^{\frac{1}{2}} |\Sigma/n|^{-\frac{1}{2}} \\ &= |(\Sigma^{-1} - 2t\Sigma^{-1})/n^{-1}|^{-\frac{1}{2}} |\Sigma/n|^{-\frac{1}{2}} \\ &= |I_p - 2tI_p|^{-\frac{1}{2}} \\ &= (1 - 2t)^{-\frac{p}{2}}. \end{aligned}$$

Con lo que vemos que la función generadora de momentos de  $T^2$  es  $M_{T^2}(t) = (1 - 2t)^{-\frac{p}{2}}$ , por lo tanto:

$$T^2 = n(\bar{x} - \mu)^T \Sigma^{-1} (\bar{x} - \mu) \sim \chi_p^2 \quad (1.1)$$

Con esto vemos que el límite de control superior que se debe elegir es una cuantíl de ésta distribución, esto es:

$$LCS = \chi_{\alpha,p}^2$$

$$LCI = 0$$

donde  $\chi_{\alpha,p}^2$  es el  $100 \times (1 - \alpha)$  cuantíl de la distribución Ji-cuadrada con  $p$  grados de libertad.

**Cuando se desconocen  $\mu$  y  $\Sigma$  del proceso.** En la practica, por lo general se desconocen  $\Sigma$  y  $\mu$  del proceso y lo que se hace es estimarlos de  $m$  conjuntos de observaciones, observaciones que se toman cuando se supone que el proceso esta en control, a este conjunto de observaciones se le conoce como datos históricos del proceso, supongamos que cada uno de los  $m$  conjuntos tienen  $n$  observaciones, entonces, las estimaciones para cada conjunto son:

$$\bar{x}_k = \frac{1}{n} \sum_{i=1}^n x_{ik} \quad k = 1, \dots, m,$$

$$s_k^2 = \frac{1}{n-1} \sum_{i=1}^n (x_{ik} - \bar{x}_k)(x_{ik} - \bar{x}_k)^T \quad k = 1, \dots, m.$$

Luego tomamos la media considerando todos los grupos,

$$\bar{\bar{x}} = \frac{1}{m} \sum_{k=1}^m \bar{x}_k,$$

$$S = \frac{1}{m} \sum_{k=1}^m s_k^2.$$

Ahora sustituyendo las estimaciones en la ecuación (1.1) obtenemos:

$$T^2 = n (\bar{x} - \bar{\bar{x}})^T S^{-1} (\bar{x} - \bar{\bar{x}}). \quad (1.2)$$

Recordemos que existen dos distintas fases en la construcción de cartas de control. La Fase I utiliza cartas de control para asegurar que los datos históricos, utilizados para estimar a  $\bar{\bar{x}}$  y  $S$ , fueron obtenidos de un proceso en control estadístico.

De acuerdo a *Montgomery [8]*, en la Fase I los límites de control están dados por:

$$\begin{aligned} LCS &= \frac{p(m-1)(n-1)}{mn-m-p+1} F_{\alpha,p,mn-m-p+1} \\ LCI &= 0, \end{aligned} \tag{1.3}$$

y en la Fase II cuando la carta de control es utilizada para monitorear la producción futura, los límites de control son:

$$\begin{aligned} LCS &= \frac{p(m+1)(n-1)}{mn-m-p+1} F_{\alpha,p,mn-m-p+1} \\ LCI &= 0. \end{aligned} \tag{1.4}$$

Notemos que el  $LCS$  en la ecuación (1.3) es igual que el  $LCS$  en la ecuación (1.4) multiplicada por  $(m+1)/(m-1)$ .

Cuando  $\mu$  y  $\Sigma$  son estimados con un gran número de datos históricos, es usual utilizar como límite de control superior  $LCS = \chi_{\alpha,p}^2$ , tanto en la Fase I como en la Fase II, esto debido a que si usamos  $m > 20$  ó  $25$  subgrupos de datos históricos los límites de control para la Fase I y Fase II coinciden.

*Lowry y Montgomery [12]* muestran que en muchas situaciones con un gran número de datos históricos, utilizar el límite de control ji-cuadrada en ambas fases es apropiado. Estos autores muestran tablas sobre esto, indicando el número mínimo de subgrupos  $m$  para muestras de  $n = 3, 5, y 10$  y  $p = 2, 3, 4, 5, 10 y 20$  características de calidad. La Figura 1.5 muestra una carta típica  $T^2$  de Hotelling.

## 1.8. Carta de control multivariada $T^2$ de Hotelling para observaciones individuales

En los análisis anteriores consideramos que para la construcción de cartas de control, necesitamos obtener un conjunto de piezas en determinados tiempos. Las cartas de control multivariadas para observaciones individuales, se utilizan para monitorear características de calidad en el caso en que se trabaje con procesos lentos y/o costosos, en los cuales para obtener una muestra de la producción se requiere de periodos relativamente largos y/o muy costosos.

Las estimaciones no difieren mucho del caso en que se obtienen conjuntos de muestras, lo

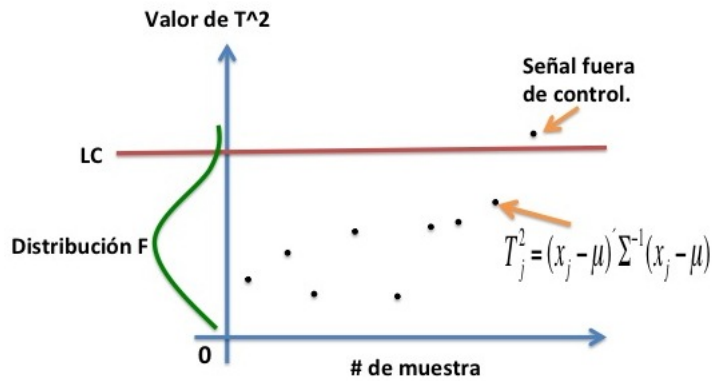


Figura 1.5: Carta de control  $T^2$  de Hotelling

único es que una observación individual se toma como un conjunto.

Supongamos que tenemos  $m$  muestras de datos históricos, cada una de tamaño  $n = 1$ , con  $p$  características de calidad observadas en cada muestra. Sea  $\bar{x}$  y  $S$  el vector de medias y matriz de varianzas, respectivamente, de estas observaciones. Entonces, el estadístico de la ecuación 1.2 se convierte en:

$$T^2 = (x - \bar{x})^T S^{-1} (x - \bar{x}). \tag{1.5}$$

Para definir los límites de control para la fase II utilizaremos el siguiente teorema,

**Teorema 1.8.** Sea  $T^2 = my'W^{-1}y$ , donde  $y \sim N_p(0, \Sigma)$ ,  $W \sim W_p(m, \Sigma)$ , y  $y$  y  $W$  son independientes. Usamos  $N_p$  y  $W_p$  para denotar una distribución Normal y Wishart  $p$ -dimensional, respectivamente. Luego,

$$\frac{m - p + 1}{p} \frac{T^2}{m} \sim F(p, m - p + 1).$$

Ahora consideremos un conjunto de datos históricos multivariados  $x_1, \dots, x_m$  y una

futura observación  $x_f$ , donde cada  $x_i$  es un vector de observaciones  $p$ -dimensionales. Si

$$x_i \sim N_p(\mu, \Sigma),$$

entonces,

$$\bar{x}_m \sim N_p(\mu, \Sigma/m)$$

$$(m-1)S_m \sim W_p(m-1, \Sigma).$$

Ahora supongamos que  $x_f$ ,  $\bar{x}_m$  y  $S_m$  son independientes, esto se debe a que  $\bar{x}_m$  y  $S_m$  son calculados mediante un conjunto de datos históricos y  $x_f$  es una futura observación. Entonces,

$$x_f - \bar{x}_m \sim N_p\left(0, \left(\frac{m+1}{m}\right)\Sigma\right)$$

y

$$\sqrt{\frac{m}{m+1}}(x_f - \bar{x}_m) \sim N_p(0, \Sigma).$$

Definimos el estadístico:

$$T^2 = \left(\frac{m}{m+1}\right)(x_f - \bar{x}_m)^T S_m^{-1}(x_f - \bar{x}_m),$$

entonces,

$$\frac{(m-p)}{p} \frac{T^2}{(m-1)} \sim F(p, m-1-p+1),$$

lo cual nos conduce a:

$$\frac{m-p}{p(m-1)} \frac{m}{(m+1)} (x_f - \bar{x}_m)^T S_m^{-1}(x_f - \bar{x}_m) \sim F(p, m-p),$$

con esto vemos que:

$$(x_f - \bar{x}_m)^T S_m^{-1}(x_f - \bar{x}_m) \sim \frac{p(m-1)(m+1)}{m(m-p)} F(p, m-p).$$

Con todo lo anterior se demuestra que en la Fase II, los límites de control para este estadístico son:

$$\begin{aligned} LCS &= \frac{p(m+1)(m-1)}{m^2 - mp} F_{\alpha, p, m-p} \\ LCI &= 0 \end{aligned} \tag{1.6}$$

Cuando el número de datos históricos  $m$  es grande digamos,  $m > 100$  algunos utilizan una aproximación al límite de control,

$$LCS = \frac{p(m-1)}{(m-p)} F_{\alpha, p, m-p} \quad (1.7)$$

o

$$LCS = \chi_{\alpha, p}^2 \quad (1.8)$$

Para  $m > 100$  la ecuación 1.7 es una aproximación razonable. *Lowry y Montgomery [12]* muestran que el límite de control Ji-cuadrada debe usarse con cuidado. Sólo si  $p$  es grande digamos,  $p > 10$ , luego si  $m > 250$ .

*Tracy et. al. [13]* mostraron que cuando se utiliza la misma muestra para estimar los parámetros y ser analizada, esto es cuando en la ecuación 1.5  $x, \bar{x}$  y  $S^{-1}$  son dependientes ya que la observación  $x$  se utilizó para calcular  $\bar{x}$  y  $S^{-1}$ , como en la Fase I, el límite de control debe basarse en la distribución beta. Entonces, los límites de control para la Fase I son:

$$\begin{aligned} LCS &= \frac{(m-1)^2}{m} \beta_{\alpha, p/2, (m-p-1)/2} \\ LCI &= 0, \end{aligned} \quad (1.9)$$

donde  $\beta_{\alpha, p/2, (m-p-1)/2}$  es el  $100 \times (1 - \alpha)$  cuantíl de una distribución beta con parámetros  $p/2$  y  $(m-p-1)/2$ . Aproximaciones a los límites de control de la Fase I con una distribución  $F$  o una distribución ji-cuadrada no son adecuados.

## 1.9. Longitud promedio de corrida (ARL)

Cuando el proceso está siendo monitoreado en línea, es decir, se encuentra en la Fase II, el número de puntos graficados en la carta hasta que aparece una señal fuera de



control, se le conoce como longitud de corrida y se denota por RL (run length).

El valor de RL cambia aleatoriamente de ensayo a ensayo, si repetimos varias veces bajo las mismas condiciones el experimento de observar un proceso desde el inicio hasta que encontramos un punto fuera de los límites de control, observaremos que RL varía de ensayo a ensayo. El valor esperado de RL es denotado como ARL (Average run length), éste puede calcularse teóricamente y por simulación. La cantidad ARL se utiliza para comparar diferentes cartas de control.

La longitud de corrida RL es una variable aleatoria con distribución geométrica de parámetro  $p$ :

$$\mathbb{P}(RL = y) = (1 - p)^{y-1}p, \quad y = 1, 2, \dots, \quad (1.10)$$

donde  $p$  es la probabilidad de caer fuera de los límites de control, esto es:

$$p = \mathbb{P}(Y > LCS),$$

con esto vemos que el valor esperado de RL, es  $\mathbb{E}(RL) = \frac{1}{p}$ , y  $Var(RL) = \frac{(1-p)}{p^2}$ . Por lo cual se obtiene que bajo condiciones de control estadístico:

$$ARL = \frac{1}{p} \quad (1.11)$$

Por ejemplo, en las cartas de control univariadas tipo Shewhart bajo distribución normal se tiene que  $p = 0.0027$  y  $ARL = 370.4$ , es decir, en promedio una carta de control dará una señal fuera de control cada 370 puntos a pesar de que el proceso este bajo control. De esta manera podemos comparar cartas de control calculando su respectivo ARL, es deseable que cuando el proceso está en control obtengamos valores de ARL grandes, pero si el proceso esta fuera de control, obtengamos valores pequeños.

# Capítulo 2

## Análisis de datos funcionales

### 2.1. Introducción

El análisis de datos funcionales es una rama de las matemáticas y más concretamente de la estadística, que estudia y analiza la información contenida en curvas, superficies o cualquier elemento que varíe sobre un espacio continuo. En el presente trabajo nos restringimos únicamente al estudio de la información contenida en curvas y no en otro tipo de elementos.

Los métodos y modelos empleados en el análisis de datos funcionales guarda un cierto parecido con los empleados en el análisis de datos multivariados, pudiendo complementarse con otras técnicas, como por ejemplo los modelos de regresión lineales y no lineales.

Los objetivos de este tipo de análisis son esencialmente los mismos que en cualquier otra área de la estadística, entre los que puede mencionarse son: la obtención de una representación de los datos de tal manera que facilite su entendimiento y resalte las distintas características, el estudio de posibles patrones de variabilidad de los datos, la explicación del porqué de las variaciones de un resultado, o bien la comparación de conjuntos de datos con respecto a cierto tipo de variaciones. Para lograr estos objetivos, en el presente trabajo utilizamos los programas estadísticos R y SAS.

## 2.2. Datos funcionales

El análisis de datos funcionales, a pesar de ser una disciplina temprana, inició sus pasos en la década de los 60 del siglo XX, la cual cuenta con una multitud de publicaciones y estudios. De la literatura referente a datos funcionales, es necesario destacar como referencia básica los libros de *Ramsay y Silverman*[2 – 4], donde tratan muchos de los problemas básicos de la estadística funcional.

En el análisis de datos funcionales, la unidad básica de información es el dato funcional, en general cualquier observación que varíe en un continuo se puede considerar como un dato funcional. Por ejemplo, un conjunto de imágenes de alta resolución es un ejemplo de datos funcionales en un dominio de dos dimensiones. En la práctica, estos sucesos son recogidos por máquinas que toman muestras de una determinada variable en distintos puntos del continuo que se considere.

En el contexto multivariado los datos provienen de la observación de la familia aleatoria  $\{X(t_j)\}_{j=1,\dots,J}$ . En análisis funcional se asume que las muestras son observaciones de una familia continua  $\chi = \{X(t); t \in T\}$ . El caso más sencillo es el de una curva unidimensional,  $T = \mathbb{R}$  pero también puede ser  $T = \mathbb{R}^2$  en imágenes o expresiones para casos más complejos.

Del libro *Ferraty y Vieu*[9] extraemos la definición de *dato funcional*:

**Definición 2.2.1** Una variable aleatoria  $\chi$  se llama variable funcional si toma valores en un espacio infinito dimensional (espacio funcional). Una observación  $X$  de  $\chi$  se llama dato funcional.

A manera de ejemplo, consideremos los datos provenientes de *Walker y Wright* [5], el cual presenta un ejemplo en la fabricación de un cierto tipo de tablas de surfear, donde es importante estudiar el patrón de distribución de la densidad de las piezas, la densidad es medida usando un perfilómetro el cual usa un láser para tomar una serie de mediciones a través de la tabla. Una medición típica de una tabla arroja una gráfica como la que aparece en la Figura 2.1

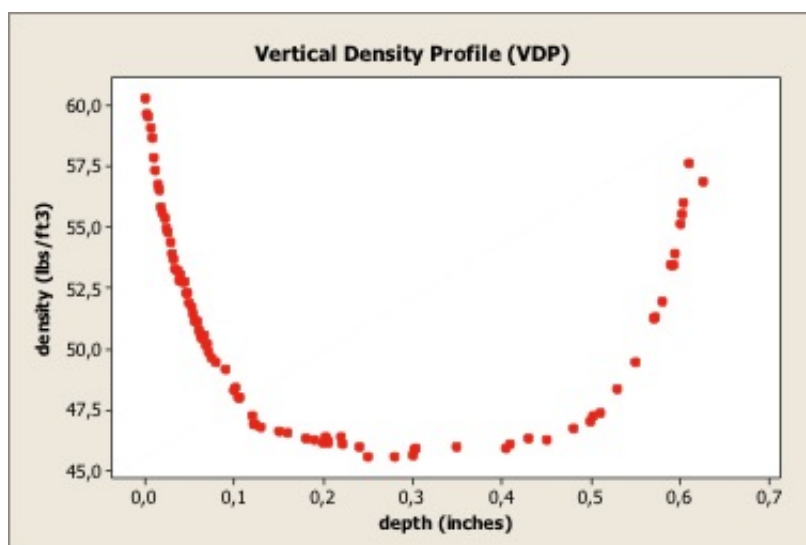


Figura 2.1: Ejemplo de mediciones sobre una tabla.

En la industria de fabricación de tablas de surfear, lo que se quiere es que una tabla tenga una alta densidad en los extremos para que sea resistente, pero se quiere que la densidad dentro de la tabla sea pequeña para que al mismo tiempo la tabla sea liviana.

En este estudio se monitoreó un proceso de producción de tablas, notemos que los datos son discretos por lo que el primer paso fue transformarlos a datos funcionales, el resultado del monitoreo de 24 tablas se muestra en la Figura 2.2.

### 2.3. Bases

Puesto que en la práctica, los datos funcionales son obtenidos y registrados de un modo discreto, en muchos casos es necesario convertir los datos disponibles en bruto a funciones sobre las cuales podemos trabajar, para lo cual pueden emplearse distintos tipos de ajustes. Tomando en cuenta lo anterior, la representación de estos datos como datos funcionales mediante una base ortonormal brinda una ventaja tanto teórica como práctica, esta base la utilizamos para transformar datos discretos en datos funcionales.

**Definición 2.3.1 (Base)** Una base es un conjunto de funciones conocidas  $\{\phi_k\}_{k \in \mathbb{N}}$  tales que cualquier función puede ser aproximada, tan bien como se requiera, mediante

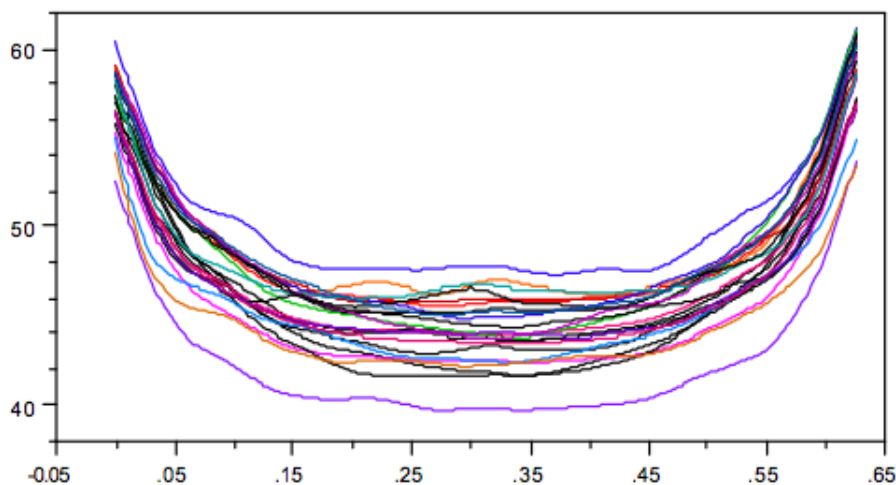


Figura 2.2: Mediciones de 24 tablas, vistas como datos funcionales

una combinación lineal de  $K$  de ellas con  $K$  suficientemente grande. De esta manera, la observación funcional puede aproximarse como  $X(t) \approx \sum_{k=1}^K c_k \phi_k(t)$ .

$$X(t) = c^T \Phi, \quad (2.1)$$

donde  $c$  es un vector de dimensión  $K$  que contiene los coeficientes  $c_k$  y  $\Phi$  es una matriz que contiene los valores de  $\phi_k(t)$ .

En este enfoque es crucial tanto la elección de la base como del valor de  $K$ , para esto no existe una ley que nos permita hacer una elección óptima. Esto dependerá del problema en estudio.

En ocasiones puede ser suficiente la utilización de una interpolación simple, que consiste en la unión de los datos adyacentes mediante segmentos de línea recta, pero esta aproximación puede ser inadecuada, si posteriormente deseamos obtener información a partir de las derivadas de la función. Una alternativa es el ajuste mediante una serie de Fourier, que permite mantener la periodicidad de los datos y generalmente suele proporcionar resultados razonables. Otra opción es el ajuste basado en splines polinomiales.

El uso de funciones tipo Spline es uno de los sistemas de aproximación más común en el estudio de datos funcionales no periódicos, y ha sido el empleado en este trabajo

para el ajuste de los datos. Su éxito radica en la combinación de la rapidez del cálculo junto con una gran flexibilidad.

## 2.4. Splines

En este trabajo se utilizan funciones Spline, por lo que dedicaremos unas líneas a la descripción de ellas.

Las funciones spline son las más utilizadas cuando se desean aproximar series de datos no periódicos. Estas funciones son polinomios ajustados en un espacio o periodo de tiempo  $T$ . EL primer paso para definir un spline es dividir el intervalo sobre el cual se desea aproximar la función de interés en  $L$  subintervalos separados por valores  $\tau_l$ ,  $l = 1, \dots, L - 1$ . Estos valores son conocidos como nodos. Sobre cada intervalo, un spline es un polinomio de orden  $m$ .

De manera general una función spline es determinada por el orden de los segmentos polinomiales y la secuencia de los nodos  $\tau$ , (*Ramsay y Silverman*[4]). Para construir una función spline se debe especificar una base de funciones  $\phi_k$  que tienen las siguientes características:

- Cada función  $\phi_k$  es una función spline de orden  $m$  definida por una secuencia de nodos  $\tau$ .
- Cualquier función spline definida por  $m$  y  $\tau$  puede ser expresada como una combinación lineal de  $\phi_k$ .

La base más utilizada para aproximar funciones spline es la *B-splines* desarrollada por *de Boor*[6]. Este tipo de bases están disponibles en una gran cantidad de software in-

cluyendo  $R$ . Otras bases de funciones splines son las funciones potencias, de base truncada, M-spline y spline naturales.

### 2.4.1. B-Spline

La idea central de este trabajo se basa en B-splines por lo que a continuación describimos como se construyen, de acuerdo al método recursivo propuesto por Carl de Boor.

Empezamos con una partición o secuencia de nodos, esto es, una secuencia no decreciente  $t = (t_i)$  de  $N+2$  valores reales llamados “nodos” ( $N \geq 0$  son llamados “nodos interiores” y siempre existen dos puntos finales,  $t_0$  y  $t_{N+1}$ ). tal que

$$t_0 \leq t_1 \leq \dots \leq t_{N+1}.$$

Definimos el conjunto de nodos aumentados

$$t_{-k} = \dots = t_0 \leq t_1 \leq \dots \leq t_N \leq t_{N+1} = \dots = t_{(N+1)+k},$$

lo anterior indica que los nodos límites  $t_0$  y  $t_{N+1}$  se repiten  $k$  veces, donde  $k$  indica el grado de la base B-Spline (Esto es necesario para formar recursivamente B-splines). Podemos reajustar el índice  $i$  de los nodos, empezando del último  $t_{-k}$ , entonces, los  $N + 2(k + 1)$  nodos aumentados  $t_i$  serán identificados por  $i = 0, \dots, N + 2k + 1$ .

Para cada uno de los nodos  $t_i$ ,  $i = 0, 1, \dots, N + 2k + 1$ , definimos de manera recursiva un conjunto de valores reales de las funciones  $B_{i,j}(t)$  ( para  $j = 0, 1, \dots, k$ , donde  $k$  es el grado de la base  $B - spline$ ) como se describe a continuación:

El B-spline de grado 0 para esta secuencia de nodos son las funciones:

$$B_{i0}(t) = X_i(t) = \begin{cases} 1, & \text{si } t_i \leq t < t_{i+1} \\ 0 & \text{otro caso.} \end{cases} \quad (2.2)$$

Estas funciones son continuas por la derecha y además cumplen que:

$$\sum_i B_{i0}(t) = 1, \text{ para todo } t. \quad (2.3)$$

En particular,

$$t_i = t_{i+1} \text{ implica } B_{i0}(t) = X_i(t) = 0. \quad (2.4)$$

De este B-spline de grado 0, podemos obtener el B-spline de mayor grado mediante recurrencia:

$$B_{ik}(t) = w_{ik}(t)B_{i,k-1}(t) + (1 - w_{i+1,k}(t))B_{i+1,k-1}(t), \quad (2.5)$$

donde,

$$w_{ik}(t) = \begin{cases} \frac{t-t_i}{t_{i+k}-t_i}, & \text{si } t_i \neq t_{i+k} \\ 0 & \text{otro caso,} \end{cases} \quad (2.6)$$

entonces, el B-spline de primer orden esta dado por:

$$B_{i1}(t) = w_{i1}(t)X_i(t) + (1 - w_{i+1,1}(t))X_{i+1}(t), \quad (2.7)$$

notemos que consiste en dos piezas lineales unidas para formar una pieza continua en el intervalo  $[t_i, t_{i+2})$ , por esta razón a  $B_{i1}(t)$  se le conoce como B-spline lineal. Si  $t_i = t_{i+1}$  (tenemos  $X_i = 0$ ), pero sigue siendo  $t_{i+1} < t_{i+2}$ , luego  $B_{i1}(t)$  consiste en sólo una pieza.



El B-spline de grado 2 se obtiene mediante:

$$\begin{aligned}
 B_{i2} &= w_{i2}(t)B_{i1}(t) + (1 - w_{i+1,2}(t))B_{i+1,1}(t) \\
 &= w_{i2}(t)w_{i1}(t)X_i(t) + (w_{i2}(t)(1 - w_{i+1,1}(t)) + (1 - w_{i+1,2}(t))w_{i+1,1}(t))X_{i+1}(t) \\
 &\quad + (1 - w_{i+1,2}(t))(1 - w_{i+2,1}(t))X_{i+2}(t).
 \end{aligned} \tag{2.8}$$

Esto muestra que en general  $B_{i2}(t)$  consiste en 3 piezas cuadráticas, que se unen suavemente en los nodos, para formar una sola función cuadrática la cual se vuelve cero fuera del intervalo  $[t_i, t_{i+3})$ . Coincidencias en los nodos  $t_i, \dots, t_{i+3}$  cambiarían esto.

Después de  $k - 1$  pasos de recurrencia, obtenemos que la forma de  $B_{ik}(t)$  es

$$B_{ik}(t) = \sum_{j=i}^{i+k} b_{jk}(t)X_j(t), \tag{2.9}$$

donde cada  $b_{jk}(t)$  es un polinomio de grado  $k$ . entonces, el B-spline de grado  $k$  consiste en la unión en los nodos de polinomios de grado  $k$ , el cual es cero fuera del intervalo  $[t_i, t_{i+k})$ .

En particular,  $B_{ik}(t)$  es una función cero en el caso en que  $t_i = t_{i+k}$  ya que si el primer nodo es igual al último, entonces todos son iguales, debido a que suponemos una secuencia no decreciente de nodos, luego por la ecuación 2.4 vemos que  $X_j(t) = 0$ , por lo que en la ecuación 2.9 se ve que  $B_{ik}(t)$  sería en este caso la función cero.

También tenemos que  $B_{ik}(t)$  es positiva en el intervalo  $(t_i, t_{i+k})$ , debido a que  $w_{ik}(t)$  y  $1 - w_{i+j,k}(t)$  son positivos en ese intervalo, esto último surge por la elección de los nodos (sucesión no decreciente) y de la definición de  $w_{ik}$  y  $1 - w_{i+j,k}$ .

**La función B-spline.** Una función B-spline de grado  $k$  es una curva paramétrica, compuesta de una combinación lineal de bases *B-spline*  $B_{i,k}(t)$  de grado  $k$ , obtenida mediante:

$$B(t) = \sum_{i=0}^{N+k} c_i B_{i,k}(t), \quad t \in [t_0, t_{N+1}],$$

los  $c_i$  son llamados “puntos control” o “puntos de Boor”. Para un *B-spline* de grado  $k$  con  $N$  nodos interiores existen  $M = N + k + 1$  puntos control.

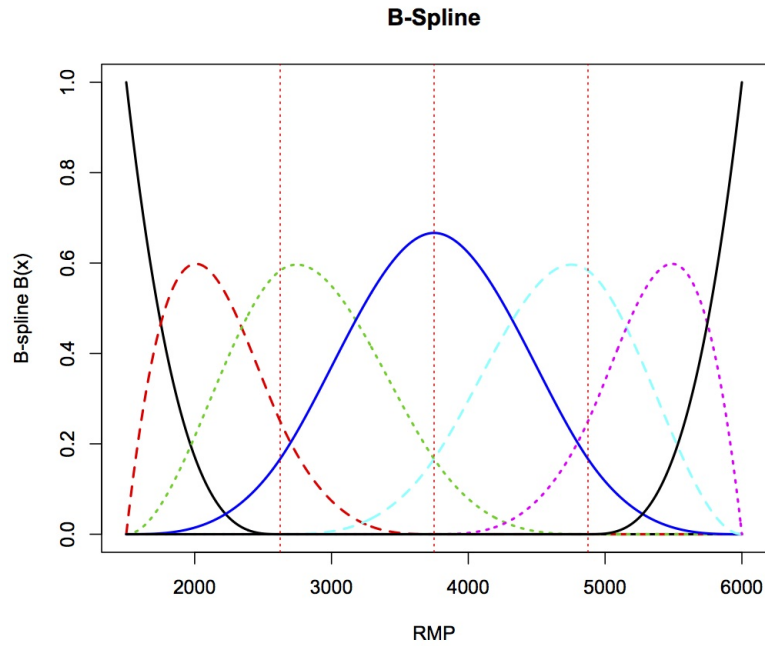


Figura 2.3: B-spline

La Figura 2.3 muestra siete funciones bases B-spline para una función B-spline de orden tres con intercepto, definidas por tres nodos interiores equiespaciados ( 2625, 3750 y 4875), y los puntos limites inferior y superior son (1500, 6000), respectivamente. Ver apéndice B para más detalles de la construcción.

## 2.5. Splines suavizados

El método de Spline Suavizado estima una curva  $x$  de las observaciones

$$y_j = x(t_j) + \epsilon_j. \quad (2.10)$$

Haciendo explícitos dos objetivos posibles en la curva, por un lado, se desea asegurar que la estimación de la curva dé un buen ajuste de los datos, por ejemplo, en términos de la suma al cuadrado de los errores  $\sum_j \{y_j - x(t_j)\}^2$ . Por otro lado, no queremos que el ajuste

sea demasiado bueno porque si esto pasa la curva  $x$  es excesivamente rugosa o localmente variable.

Una medida popular para la rugosidad de una función es

$$PEN_2(x) = \int \{D^2x(s)\}^2 ds = \|D^2x\|^2, \quad (2.11)$$

esto evalúa la curvatura total en  $x$ , o alternativamente, el grado con el que  $x$  se aparta de la línea recta, por lo que para funciones altamente variables se puede esperar grandes valores de  $PEN_2(x)$ , porque su segunda derivada es grande en la mayor parte de rango de interés.

Podemos definir la suma de cuadrados del error penalizado como:

$$PENSSSE_\lambda(x | y) = \sum_j \{y_j - x(t_j)\}^2 + \lambda \times PEN_2(x), \quad (2.12)$$

las estimaciones son obtenidas encontrando la función  $x$  tal que minimice a  $PENSSSE_\lambda(x)$  sobre un espacio de funciones  $x$ , para las cuales  $PEN_2(x)$  es definido.

El parámetro  $\lambda$  es un parámetro de suavizamiento, si  $\lambda \rightarrow \infty$  la curva ajustada  $x$  tiende a una aproximación de regresión lineal estándar de los datos observados. Por otro lado cuando  $\lambda \rightarrow 0$  la curva ajustada aproxima a una interpolación de los datos, que satisface  $x(t_j) = y_j, \forall j$ .

En muchos casos de análisis de datos funcionales es de interés la estimación de las derivadas. En estos casos la penalidad (2.11) puede no ser adecuada, ya que ésta controla sólo la curvatura de  $x$ .

Si la derivada de orden  $m$  es requerida, uno probablemente debería penalizar dos derivadas adelante. Por ejemplo, la estimación de la aceleración es mejor si usamos:

$$PEN_4(x) = \int \{D^4x(s)\}^2 ds = \|D^4x\|^2. \quad (2.13)$$

## 2.6. Construcción de datos funcionales

De la expresión (2.1), se observa que el problema consiste en estimar el vector de constantes  $c$ , el cual contiene el peso (o ponderación) que tiene cada función de la base sobre la función de interés  $x(t)$ . *Ramsay y Silverman*[4] realizaron la estimación del vector  $c$  utilizando la metodología de mínimos cuadrados ordinarios y ponderados de la siguiente manera:

Suponga que  $y_j$  es el valor observado de la variable  $y$  en el tiempo  $t_j$  (para  $j = 1, \dots, n$ ), tal que  $y_j$  puede ser expresado como:

$$y_j = x(t_j) + \epsilon_j = \phi^T(t_j)c + \epsilon_j, \quad (2.14)$$

donde  $x(t_j)$  es el valor de la función ajustada mediante (2.1) en el tiempo  $t_j$  y  $\epsilon_j$  es el error de estimación en el momento  $t_j$ . Se supone que estos errores son estadísticamente independientes y siguen una distribución normal con media cero y varianza constante. En la última igualdad de la expresión (2.14),  $\phi_j^T(t)$  es un vector fila que contiene las  $K$  funciones de la base evaluadas en el momento  $t_j$ .

Siguiendo la metodología de estimación por mínimos cuadrados ordinarios, la expresión a minimizar es:

$$SSE(c) = \sum_j^n \left[ y_j - \sum_k^K c_k \phi_k(t_j) \right]^2 = \sum_j^n \left[ y_j - \phi^T(t_j)c \right]^2, \quad (2.15)$$

sea  $Y$  un vector conformado por todos los  $y_j$  (valores a ser ajustados),  $\Phi^T$  una matriz de tamaño  $n \times K$  conformada por los valores de las  $K$  funciones base  $\phi_k(t_j)$  y  $\epsilon$  el vector de residuales de tamaño  $n \times 1$ , entonces la expresión (2.14) puede ser escrita en forma matricial como

$$Y = \Phi^T c + \epsilon, \quad (2.16)$$

de esta forma, la expresión (2.15) se puede expresar como:

$$SSE(c) = (Y - \Phi^T c)^T (Y - \Phi^T c), \quad (2.17)$$

resolviendo para  $c$  obtenemos el estimador  $\hat{c}$  que minimiza la suma de cuadrados de (2.17), y esta dado por:

$$\hat{c} = (\Phi\Phi^T)^{-1}\Phi Y, \quad (2.18)$$

con esto podemos ver que el vector de valores ajustados  $\hat{y}$  es

$$\hat{Y} = \Phi^T (\Phi\Phi^T)^{-1} \Phi Y. \quad (2.19)$$

Si se sospecha de errores autorrelacionados o no se puede suponer varianza constante para los mismos, se extienden los resultados anteriores al minimizar ahora la expresión:

$$SMSSE(y | c) = (Y - \Phi^T c)^T W (Y - \Phi^T c), \quad (2.20)$$

donde  $W$  es una matriz simétrica definida positiva que permite realizar diferentes ponderaciones a las observaciones y productos residuales. Si la matriz de varianzas y covarianzas para los errores  $\Sigma_\epsilon$  es conocida se puede utilizar  $W = \Sigma_\epsilon^{-1}$  (*Ramsay y Silverman*[4]). El estimador de  $c$  en la expresión (2.20), vía mínimos cuadrados ponderados es:

$$\hat{c} = (\Phi W \Phi^T)^{-1} \Phi W Y \quad (2.21)$$

## 2.7. Construcción de datos funcionales con spline suavizados

Teniendo en cuenta que cuando se suaviza la serie original mediante una gran can-

tividad de funciones de una base se puede incurrir en sobreestimación ( en un caso extremo se podría tener una función base por observación, lo que generaría un mal suavizamiento y por consiguiente un mal ajuste), en *Ramsay y Silverman*[4] se propone un método de suavizado que tiene en cuenta un factor de penalización por mal suavizamiento. En este método la estimación del vector  $c$  se obtiene al minimizar la expresión:

$$PENSSSE_{\lambda}(x | y) = \sum_j \{y_j - x(t_j)\}^2 + \lambda \int \{D^m x(s)\}^2 ds, \quad (2.22)$$

donde  $\lambda$  representa el parámetro que penaliza la curvatura relacionada a la bondad de ajuste, cuantificada en la suma de cuadrados del error en el primer término de (2.22) y  $D^m$  representa un operador diferencial de orden  $m$  aplicado a  $x(t)$ . Cuando  $m = 2$  este operador diferencial es denominado la curvatura de  $x(t)$ .

El uso de un operador diferencial  $D^m$  para definir el mal suavizamiento, implica que cuando  $\lambda \rightarrow \infty$  se estará forzando el ajuste a ser cada vez más cerrado a la solución de la ecuación diferencial  $D^m x(t) = 0$ . Si  $D^m$  es de orden 2, ésta solución será de orden 2 o una línea recta. Por otro lado, si  $\lambda \rightarrow 0$  no se estaría penalizando por mal ajuste o mal suavizamiento.

Remplazando  $x(t)$  por  $\Phi^T c$  en la expresión (2.22), se obtiene su forma matricial:

$$PENSSSE_{\lambda}(y | c) = (Y - \Phi^T c)^T W (Y - \Phi^T c) + \lambda \times PEN_m(x), \quad (2.23)$$

podemos escribir a  $PEN_m(x)$  de forma matricial como:

$$\begin{aligned}
 PEN_m(x) &= \int \{D^m x(s)\}^2 ds \\
 &= \int [D^m \Phi^T(s)c]^2 ds \\
 &= \int c^T D^m \Phi(s) D^m \Phi^T(s) c ds \\
 &= c^T \left[ \int D^m \Phi(s) D^m \Phi^T(s) ds \right] c \\
 &= c^T R c.
 \end{aligned}$$

donde

$$R = \int D^m \Phi(s) D^m \Phi^T(s) ds.$$

Sustituyendo esto en la expresión (2.23) obtenemos

$$PENSSSE(y | c) = (Y - \Phi^T c)^T W (Y - \Phi^T c) + \lambda c^T R c, \quad (2.24)$$

luego derivando (2.24) respecto a  $c$  e igualando a cero obtenemos:

$$-2\Phi W y + 2\Phi W \Phi^T c + 2\lambda R c = 0, \quad (2.25)$$

y resolviendo la ecuación anterior encontramos un estimador para  $c$ , dado por:

$$\hat{c} = (\Phi W \Phi^T + \lambda R)^{-1} \Phi W y. \quad (2.26)$$

De esta manera vemos que los valores ajustados  $\hat{y}$  son:

$$\hat{y} = \Phi^T (\Phi W \Phi^T + \lambda R)^{-1} \Phi W y. \quad (2.27)$$

De esta forma, la matriz  $H = \Phi^T(\Phi\Phi^T + \lambda R)^{-1}\Phi$  y por consiguiente:

$$\hat{y} = Hy. \quad (2.28)$$

Una de las metodologías utilizadas para determinar un valor adecuado del parámetro de suavizamiento  $\lambda$ , es la validación cruzada generalizada (GCV), el criterio se basa en determinar el valor de  $\lambda$  que minimice la expresión:

$$GCV = \left( \frac{n}{n - gl(\lambda)} \right) \left( \frac{PENSSSE}{n - gl(\lambda)} \right), \quad (2.29)$$

donde  $PENSSSE$  es la suma de cuadrados del error que depende del valor de  $\lambda$ , y  $gl(\lambda) = Tr(H)$ .

## 2.8. Estadísticas descriptivas para datos funcionales

Teniendo en cuenta que un primer análisis estadístico de un conjunto de datos debe empezar por un análisis descriptivo y exploratorio, se presentan a continuación algunas características descriptivas funcionales propuestas por *Ramsay y Silverman*[4] para resumir datos funcionales:

Sea  $x_i(t)$   $i = 1, 2, \dots, N$  una muestra de curvas o datos funcionales, la media, varianza, covarianza y correlación funcional están dadas por:

- Media funcional:

$$\bar{x}(t) = N^{-1} \sum_i x_i(t), \quad (2.30)$$

- Varianza funcional:

$$Var_x(t) = (N - 1)^{-1} \sum_i (x_i(t) - \bar{x}(t))^2, \quad (2.31)$$



- Función de covarianzas:

$$Cov(t_1, t_2) = (N - 1)^{-1} \sum_i [(x_i(t_1) - \bar{x}(t_1))][x_i(t_2) - \bar{x}(t_2)], \quad (2.32)$$

- Función de correlación:

$$Corr_x(t_1, t_2) = \frac{Cov_x(t_1, t_2)}{\sqrt{Var_x(t_1)Var_x(t_2)}}, \quad (2.33)$$

- Función de correlación cruzada:

$$Corr_{x,y}(t_1, t_2) = \frac{Cov_{x,y}(t_1, t_2)}{\sqrt{Var_x(t_1)Var_y(t_2)}} \quad (2.34)$$

*Ramsay y Silverman*[4] también plantean otros métodos o herramientas para el análisis descriptivo de datos funcionales como el análisis de componentes principales funcionales.

# Capítulo 3

## Carta de control para datos funcionales

### 3.1. Introducción

Los datos funcionales han tenido un gran auge en tiempos recientes, sus aplicaciones abarcan una gran cantidad de áreas, muchas técnicas estadísticas clásicas han evolucionado tomando en cuenta la formación de datos funcionales y las cartas de control no son la excepción. En la industria es común encontrarse con problemas en donde la característica de calidad es medida en el tiempo o variando un factor a cada uno de los artículos, entonces, lo que se obtiene es una curva por cada artículo de la muestra, a lo cual llamamos dato funcional. Este tipo de problemas se pueden atacar con cartas de control para datos funcionales.

En este tema de cartas de control para datos funcionales se tiene un antecedente inmediato en la institución CIMAT, en dicho trabajo de tesis se propone una carta de control para datos funcionales, basada en una base B-spline y con límites de control funcionales obtenidos mediante Bootstrap. *Rivera, D. (2011)*.

En el presente trabajo se hace una propuesta de carta de control para datos funcionales. comparamos esta carta de control, con la que se propone en *Amiri*[1] basada en los coeficientes aleatorios de un modelo de efectos mixtos cuadrático. En la comparación se analizan los datos de 26 motores de automóvil, que exhiben la relación entre velocidad

del motor, en revoluciones por minuto y el torque producido. Además que desarrollamos un ejemplo práctico tomando los datos de la base *teicator*, en esta base de datos, se presentan 240 muestras de carne molida a las cuales se les midieron 125 atributos: 100 son absorbencias, luego 22 coeficientes de componentes principales, después humedad, grasa y proteína. El objetivo principal era predecir el porcentaje de grasa en una muestra de carne molida con los 100 primeros atributos de absorbencia.

### 3.2. Descripción del problema

Algunas veces la calidad de un proceso o un producto se caracteriza mediante la relación entre una variable respuesta y una o más variables explicativas, conocida como Perfil. En *Amiri*[1] se muestran perfiles de la industria automotriz, en donde, se considera como una de las características más importantes de calidad del motor de un automóvil, la relación entre el Torque producido por el motor y la velocidad del motor medida en revoluciones por minuto (RPM). Para un tipo de motor (TU3, ensamblado en un automóvil Peugeot) se hacen mediciones de los valores del torque para 14 diferentes valores de RPM. Los datos que contiene *Amiri*[1] corresponden a 26 motores, cada uno probado a los siguientes valores de RPM:

1500	2000	2500	2660	2800	2940	3500
4000	4500	5000	5225	5500	5774	6000

y su correspondiente valor de Torque es medido. La Figura 3.1 muestra los datos obtenidos.

En este artículo reparametrizaron los valores de RPM, construyendo las RPM reparametrizadas como  $RMP_i - \overline{RMP}$ , para disminuir el efecto de colinealidad, luego ajustaron un modelo sencillo con una  $R^2$  ajustada grande, para cada uno de los 26 motores, utilizando

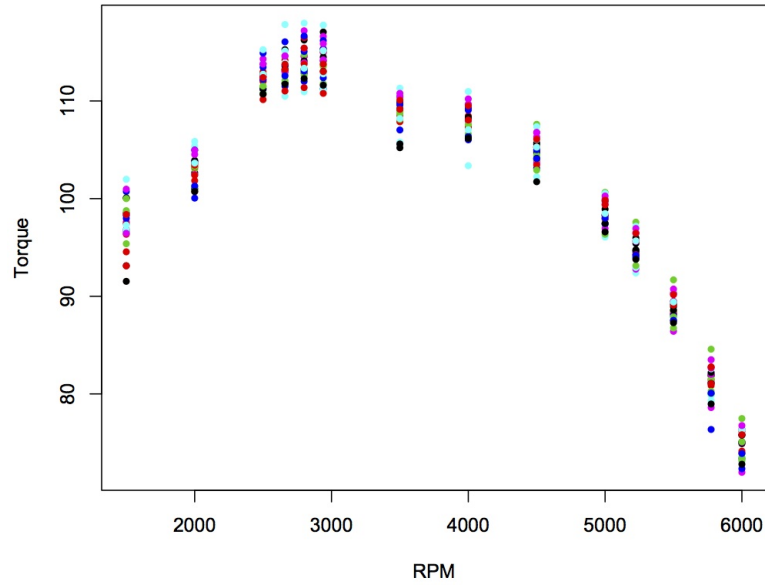


Figura 3.1: Datos obtenidos de *Amiri[1]*

un modelo cuadrático (el cual resultó con una  $R_{ajustada}^2 = 96.3$ ),

$$y_j = \beta_0 + \beta_1 x_j + \beta_2 x_j^2 + \epsilon_j, \quad (3.1)$$

donde el vector  $y_j$  contiene los valores de Torque, y  $x_j$  contiene los valores de RPM,  $(\beta_0, \beta_1, \beta_2)$  son los parámetros del modelo y  $\epsilon_j$  es el error de ajuste. En la Tabla 3.1, se muestran los parámetros encontrados para cada uno de los 26 motores.

Observamos el ajuste del modelo para cada uno de los 26 motores en la Figura 3.2 y notamos que existe una región donde el modelo cuadrático no se ajusta a los datos.

Para explicar y considerar la autocorrelación dentro de cada perfil se propuso un modelo de efectos mixtos de grado 2, el cual se expresa como:

$$y_j = X\beta + Zb_j + \epsilon_j, \quad (3.2)$$

Estimaciones del modelo cuadrático			
Motor	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
329	112.99	-0.00615	-0.00000521
449	110.95	-0.0057	-0.00000504
529	112.12	-0.00516	0.0000047
642	108.94	-0.0055	-0.0000045
724	107.81	-0.0060	-0.0000045
803	110.97	-0.0066	-0.0000052
930	111.14	-0.00609	-0.0000046
1148	112.45	-0.00609	-0.0000054
1171	111.09	-0.00577	-0.00000504
1516	111.66	-0.0067	-0.0000049
1791	111.97	-0.00707	-0.0000049
2600	113.27	-0.00605	-0.0000053
3100	110.96	-0.0061	-0.0000045
3720	109.89	-0.0051	-0.0000045
4025	110.26	-0.0063	-0.0000047
4068	112.65	-0.0071	-0.0000054
4926	114.69	-0.0061	-0.0000053
5155	111.95	-0.0058	-0.0000046
6143	110.03	-0.00532	-0.0000049
6844	111.46	-0.0057	-0.0000053
7811	109.74	-0.0059	-0.0000049
8007	109.87	-0.00608	-0.0000047
8623	110.88	-0.00566	-0.0000047
9388	113.31	-0.0057	-0.0000052
9404	109.42	-0.0056	-0.0000052
10430	112.14	-0.0056	-0.0000049

Tabla 3.1: Estimaciones que se obtuvieron con el modelo cuadrático  $y = \beta_0 + \beta_1x + \beta_2x^2$ , ajustado a las observaciones del Torque por cada uno de los 26 motores.

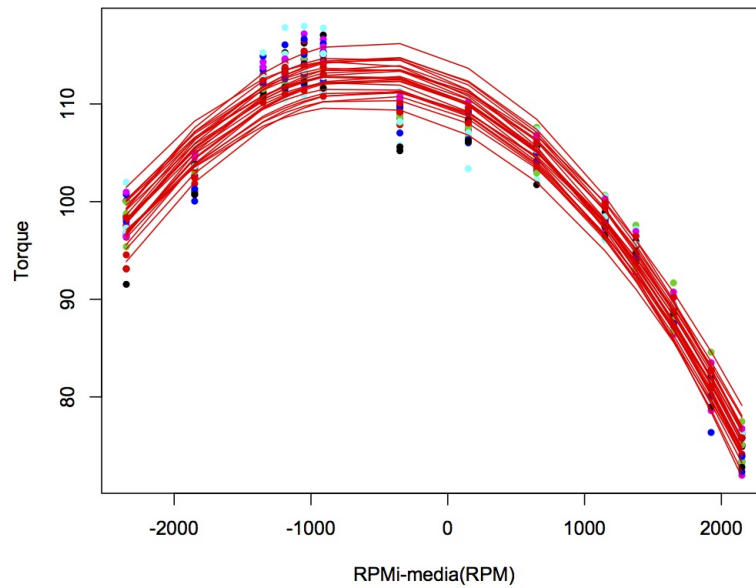


Figura 3.2: Ajuste de modelo cuadrático

donde  $y_j$  es la respuesta para el perfil  $j$  –ésimo,  $X$  es la matriz de observaciones,  $\beta$  es el vector de parámetros fijos,  $Z$  es la matriz asociada a los efectos aleatorios,  $b_j$  es el vector de efectos aleatorios y  $\epsilon_j$  el error. Se asume que  $b_j \sim N(0, D)$  donde  $D$  es la matriz de covarianzas para los efectos aleatorios y  $\epsilon_j \sim N(0, \Sigma)$ , donde  $\Sigma$  es una matriz de covarianzas de los errores, se asume que los efectos aleatorios y errores son independientes.

Notemos que las matrices de varianzas y covarianzas ( $D$  y  $\Sigma$ ) son parámetros a estimar, estimamos estos parámetros mediante máxima verosimilitud restringida, para esto se usa la rutina PROC MIXED en SAS, en este caso se supuso una estructura AR(1) en la matriz  $D$  y una estructura diagonal para  $\Sigma$ . Una vez obtenidas las estimaciones  $\hat{D}$  y  $\hat{\Sigma}$ , podemos obtener  $\hat{\beta}$ , este vector de efectos fijos estimado esta dado por:

$$\hat{\beta} = \left( X^T \hat{V}^{-1} X \right)^{-1} \left( X^T \hat{V}^{-1} \bar{y} \right), \quad (3.3)$$

donde  $\hat{V} = Z \hat{D} Z^T + \hat{\Sigma}$ , la cual contiene las estimaciones de las matrices de varianza y covarianza, y  $\bar{y}$  es la media de las observaciones. Las estimaciones para los efectos

aleatorios son obtenidos mediante:

$$\hat{b}_j = \hat{D}Z^T\hat{V}^{-1}(y_j - X\hat{\beta}). \quad (3.4)$$

En el Apéndice I se muestra la deducción de las ecuaciones (3.3) y (3.4) mediante máxima verosimilitud.

Una vez obtenidos los parámetros hacen una carta de control  $T^2$  de *Hotelling* para ver si no existen outliers en los 26 perfiles que se tienen o si existe alguno que este fuera de control respecto al resto. En el artículo Vargas[15] se muestra que un método efectivo para encontrar outliers es utilizar una carta de control  $T^2$  basada en diferencias sucesivas:

$$S_{SD} = \frac{1}{2(m-1)} \sum_{i=1}^{m-1} (\hat{b}_{j+1} - \hat{b}_j)(\hat{b}_{j+1} - \hat{b}_j)^T,$$

entonces, los puntos a graficar en esta carta de control  $T^2$  son los obtenidos mediante:

$$T_j^2 = (\hat{b}_j)^T [S_{SD}]^{-1} (\hat{b}_j), \quad (3.5)$$

donde el límite de control superior esta dado por el  $100 \times (1 - \alpha)$  cuantíl de la distribución ji-cuadrada de  $p = 3$  grados de libertad ( $\chi_{\alpha,p}^2$ ).

La Tabla 3.2 muestra las estimaciones de  $\hat{b}_j$ , para el conjunto de 26 motores que se presentan en el artículo en estudio, utilizando PROC MIXED de SAS, las cuales son muy parecidas a las presentadas en *Amiri*[1]. Enseguida en la Figura 3.3 se muestra el ajuste del modelo de efectos mixtos de grado 2 utilizando las estimaciones de  $\hat{b}_j$ .

Ahora en la fase I se investiga la existencia de outilers en los datos históricos, para ello se utiliza una carta de control  $T^2$  de *Hotelling* con los efectos aleatorios obtenidos para cada uno de los 26 motores, en la ecuación 3.5 se muestra la forma de obtener el

Estimaciones del modelo de efectos mixtos			
Motor	$\hat{b}_0$	$\hat{b}_1$	$\hat{b}_2$
329	0.957	-0.482	0.2022
449	-0.372	0.282	-0.132
529	0.998	0.059	0.0148
642	-1.179	0.716	-0.292
724	-1.928	0.768	-0.325
803	-0.612	-0.145	0.0261
930	0.377	-0.220	0.122
1148	0.107	-0.094	0.00107
1171	-0.272	0.236	-0.111
1516	0.386	-0.569	0.247
1791	0.711	-0.912	0.401
2600	1.016	-0.446	0.1809
3100	0.374	-0.245	0.1400
3720	-0.414	0.651	-0.2464
4025	-0.347	-0.083	0.045
4068	0.313	-0.809	0.303
4926	2.001	-0.894	0.382
5155	1.105	-0.373	0.2077
6143	-0.979	0.786	-0.349
6844	-0.409	0.325	-0.172
7811	-0.120	0.501	-0.236
8007	-0.708	0.230	-0.093
8623	0.110	0.134	-0.033
9388	1.153	-0.307	0.132
9404	-1.842	0.976	-0.466
10430	0.659	-0.077	0.046

Tabla 3.2: Estimaciones que se obtuvieron con el modelo de efectos mixtos  $y_j = X\beta + Zb_j + \epsilon_j$ , ajustado a las observaciones del Torque por cada uno de los 26 motores.



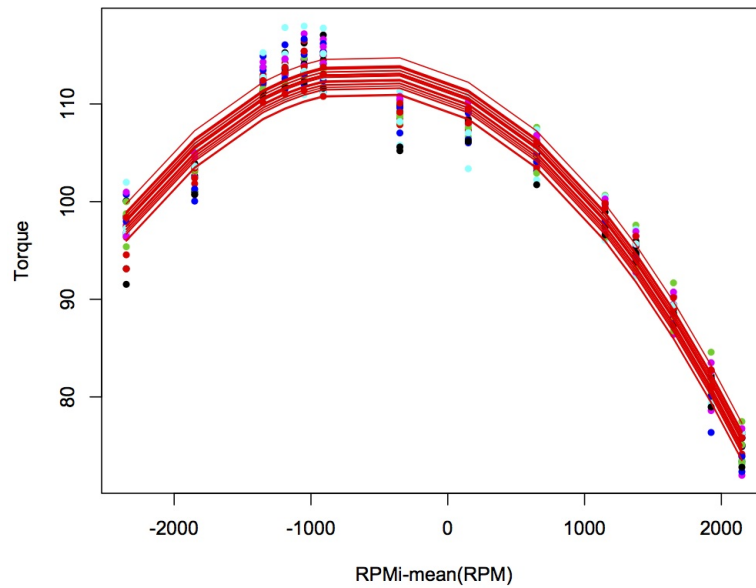


Figura 3.3: Ajuste de modelo de efectos mixtos

estadístico  $T_j^2$ , la carta de control obtenida se muestra en la figura 3.4 , donde notamos que todos los 26 puntos están por debajo del límite de control superior, lo cual nos indica que no existen outliers en los datos y con estos 26 motores podemos estimar los parámetros para construir la carta de control en la fase II.

En la fase II, construimos la carta de control utilizando el siguiente estadístico,

$$T_j^2 = (\hat{b}_j - \mu)^T \Sigma^{-1} (\hat{b}_j - \mu), \quad (3.6)$$

donde  $\mu$  y  $\Sigma$  (matriz de covarianzas muestral) están dados por:

$$\mu = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \quad y \quad \Sigma = \begin{pmatrix} 0.939 & -0.434 & 0.195 \\ -0.434 & 0.283 & -0.123 \\ 0.195 & -0.123 & 0.054 \end{pmatrix}, \quad (3.7)$$

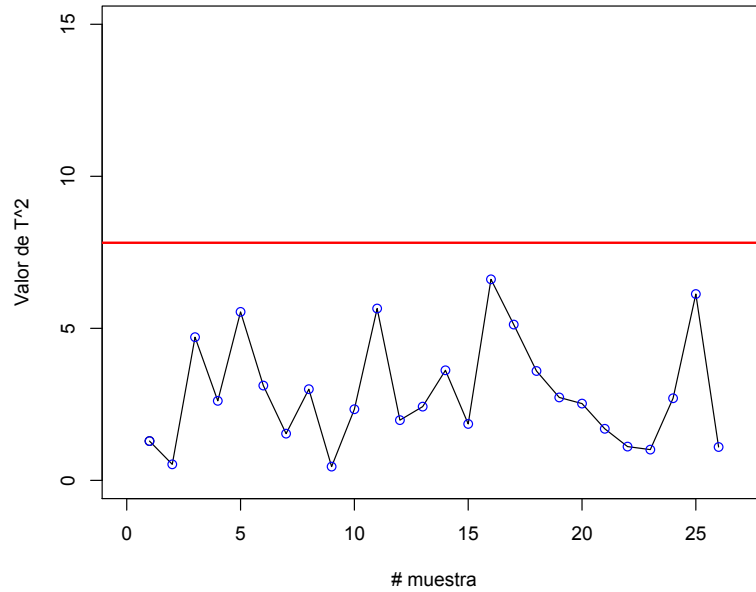


Figura 3.4: Carta de control, fase I.

tomando el límite de control superior como  $\chi_{(0.05,3)}^2$ , el cual es el  $100 \times (1 - 0.05)$  cuantíl de la distribución ji-cuadrada de 3 grados de libertad. Con esta carta de control podemos monitorear el proceso y decidir cuales motores están o no bajo control estadístico.

### 3.3. ARL para carta de control

Para encontrar el ARL que le corresponde a la carta de control descrita en *Amiri* [1], necesitamos simular nuevos perfiles de motores que correspondan a un proceso bajo control. Para hacer la simulación, suponemos que los perfiles de los motores tienen una distribución *Normal 14 – variada*,  $N_{14}(\mu, \Sigma)$ , esto debido a que con la prueba Mardia no rechazamos normalidad, los parámetros  $\mu$  y  $\Sigma$  los estimamos con los 26 perfiles de motores que presenta el artículo, esto es,

$$\mu = \frac{1}{26} \sum_{i=1}^{26} y_i \quad \text{y} \quad \Sigma = \frac{1}{25} \sum_{i=1}^{26} (y_i - \mu)(y_i - \mu)^T, \quad (3.8)$$

recordemos que  $y_i$  es un vector 14 – *variado* que representa el Torque medido en 14 diferentes revoluciones por minuto.

Algo importante de remarcar es que si simulamos de una distribución normal, como se mencionó anteriormente, ya estamos suponiendo que el proceso tiene media  $\mu$  y matriz de varianzas  $\Sigma$ , lo cual en un caso real no sería correcto, ya que sólo con 26 observaciones no podríamos aproximar de una manera adecuada los verdaderos parámetros del proceso, pero para efecto de comparación entre las dos cartas de control, la propuesta en esta tesis y la propuesta en el artículo *Amiri* [1], se procederá de esta manera. Algunas implicaciones de suponer que sabemos los verdaderos parámetros del proceso, radican en los límites de control que utilizamos en las cartas de control.

La figura 3.5 muestra la simulación de 20 perfiles mediante la distribución  $N_{14}(\mu, \Sigma)$ , las líneas rojas son las 20 simulaciones y las líneas negras son los 26 perfiles originales. Notemos que las 20 simulaciones presentan la misma forma y estructura de las 26 originales, por lo que concluimos que la simulación mediante esta distribución normal es adecuada para el estudio.

Hasta este momento tenemos un procedimiento para generar nuevos perfiles de motores, que nos servirán para calcular el ARL de esta carta de control. Así, simulamos un perfil y calculamos sus respectivos efectos aleatorios con la ecuación 3.4, después con la ecuación 3.6 calculamos el valor  $T_j^2$  que le corresponde, para luego graficarlo en la carta de control, la cual tiene límite de control  $\chi_{(\alpha,3)}^2$ . Repetimos este procedimiento hasta que la carta de control dé una señal fuera de control y registramos el número de repeticiones  $j$  hasta que ocurrió la señal, este valor de  $j$  es el  $RL$  definido en la sección (1.8). Repetimos esto 10000 veces y obtenemos 10000 valores de la  $RL$ , luego obtenemos la media de estos valores  $RL$  y esta media es exactamente el  $ARL$ .

Una vez obtenido el ARL bajo control estadístico del proceso para  $\alpha = 0.05$  y  $\alpha = 0.01$ , los cuales se muestran en la tabla 3.3, procedemos a introducirle una perturbación

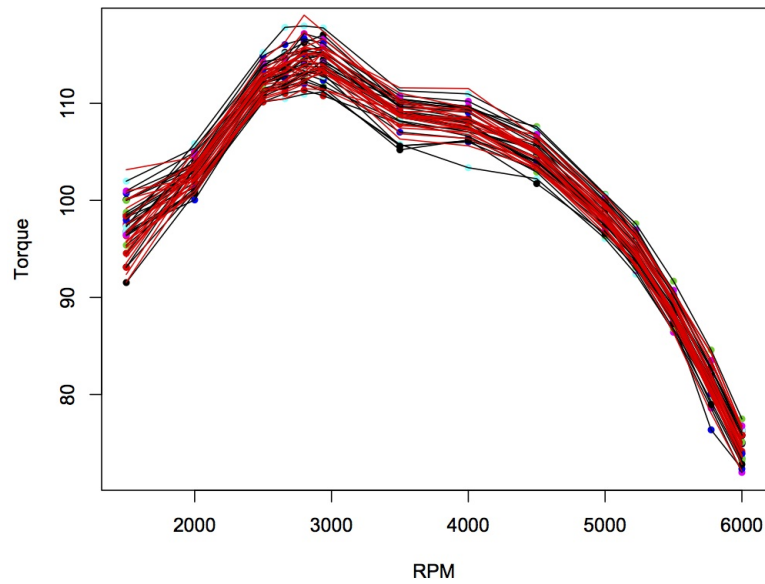


Figura 3.5: 20 simulaciones

a la media del proceso, para lo cual le sumamos un vector no nulo  $P$  a  $\mu$ , donde  $\mu$  es la media de la distribución  $N_{14}(\mu, \Sigma)$ , es decir, ahora simularemos nuevos perfiles suponiendo que siguen la distribución siguiente:

$$y_i \sim N_{14}(\mu + P, \Sigma).$$

La forma usual de  $P$  es  $P = aS$  donde  $a$  es un número real y  $S$  es la desviación estándar.

El vector desviación estándar  $S$  se muestra a continuación:

$$(2.52, 1.63, 1.40, 1.62, 1.80, 1.89, 1.70, 1.67, 1.52, 1.28, 1.38, 1.3, 1.78, 1.48),$$

lo que muestra este vector es la desviación estándar de 26 motores en cada uno de los 14 RPM.

Lo primero que observamos en el ajuste del modelo cuadrático es el pobre ajuste en los valores RPM: 2500,2660,2800,2940,3500. Por lo que en el tabla 3.4 se muestra el ARL

valor $\alpha$	ARL
0.05	19.9
0.01	99.8

Tabla 3.3: ARL bajo control estadístico

con una perturbación de  $P_i = aS_i$ , para cada uno de los 5 valores RMP mencionados, es decir, tendremos 5 vectores donde cada vector  $P_i$  contiene ceros en todas sus entradas excepto en la entrada  $i$ , donde  $i = 3, 4, 5, 6, 7$ , este orden de  $i$  corresponde a los valores 2500, 2660, 2800, 2940, 3500 RPM, respectivamente. En la entrada  $i$  del vector  $P_i$  se tiene la cantidad  $P = aS_i$ , donde  $S_i$  es la desviación estándar de 26 motores en el correspondiente valor de RPM.

En la Tabla 3.4 notamos que la carta de control propuesta en el artículo en estudio, no detecta de manera efectiva cambios en la media del proceso cuando este cambio se presenta en un sólo valor de RMP, es decir, cuando el vector  $P$  sólo tiene una entrada diferente de cero. En la tabla antes mencionada sólo se presentan los resultados para los valores de RMP: 2500, 2660, 2800, 2940, 3500 pero en el Apéndice C se muestran para todos los valores restantes de RPM, y en todos estos resultados se muestra la misma conclusión. La falta de sensibilidad antes mencionada de la carta propuesta en el artículo en estudio puede ocasionar problemas graves en la producción, ya que si el proceso esta produciendo motores con un mal desempeño, medido en Torque, en cierto valor de RPM esta carta detectará que el motor está en buenas condiciones cuando en realidad no lo está.

Ahora consideremos una perturbación en la media del proceso como  $P = aS$ , donde  $a$  es un valor real y  $S$  es un vector 14 – *variado* que contiene las desviaciones estándar en cada uno de los 14 valores de RMP. En la Tabla 3.5 se muestran los resultados de ARL para la carta de control propuesta en el artículo en estudio, considerando dos valores  $\alpha = 0.05$  y  $\alpha = 0.01$ .

$P_3$		$P_3$		$P_4$		$P_4$	
a	ARL	a	ARL	a	ARL	a	ARL
-0.1	20	0.1	20	-0.1	20	0.1	19.9
-0.5	19.9	0.5	19.9	-0.5	19.7	0.5	19.45
-0.8	19.8	0.8	19.4	-0.8	19.5	0.8	19.38
-1	19.5	1	19.3	-1	19.4	1	19.16
-2	18.2	2	18.3	-2	18.1	2	18.26
-3	16.3	3	16.2	-3	15.9	3	16.03

$P_5$		$P_5$		$P_6$		$P_6$	
a	ARL	a	ARL	a	ARL	a	ARL
-0.1	20	0.1	20	-0.1	20	0.1	20
-0.5	19.9	0.5	19.9	-0.5	19.8	0.5	19.8
-0.8	19.7	0.8	19.6	-0.8	19.5	0.8	19.4
-1	19.2	1	19.4	-1	19.2	1	19.2
-2	17.2	2	17.4	-2	16.3	2	16.3
-3	15.0	3	14.8	-3	13.4	3	13.2

$P_7$		$P_7$	
a	ARL	a	ARL
-0.1	20	0.1	20
-0.5	19.6	0.5	19.6
-0.8	18.7	0.8	18.6
-1	18.2	1	18.1
-2	14.0	2	14.2
-3	10.1	3	10.0

Tabla 3.4: Resultados ARL

Para $\alpha = 0.05$		Para $\alpha = 0.05$	
Valor de $a$	ARL	Valor de $a$	ARL
0.01	19.9	-0.01	19.9
0.1	19.4	-0.1	19.5
0.5	11.9	-0.5	11.8
0.6	9.9	-0.6	9.9
0.8	6.9	-0.8	6.8
1	4.8	-1	4.8
1.5	2.3	-1.5	2.3
2	1.4	-2	1.4

Para $\alpha = 0.01$		Para $\alpha = 0.01$	
Valor de $a$	ARL	Valor de $a$	ARL
0.01	99.1	-0.01	98.1
0.1	94.2	-0.1	95.3
0.5	46.8	-0.5	46.9
0.6	36.8	-0.6	36.0
0.8	21.8	-0.8	22.0
1	13.3	-1	13.5
1.5	4.5	-1.5	4.6
2	2.1	-2	2.1

Tabla 3.5: Resultados ARL

Analizando la Tabla 3.5 notamos que esta carta de control si detecta los cambios en la media del proceso cuando la perturbación afecta a las mediciones en cada uno de los 14 valores RMP, de hecho vemos que con una perturbación de  $a=0.5$  y  $a=-0.5$  con  $\alpha = 0.5$  el ARL tienen un cambio sustancial, hasta con una perturbación de  $a=2$  vemos que la carta detecta el cambio inmediatamente, es decir, con una perturbación de  $a=0.5$  la carta dejará pasar 12 motores defectuosos antes de registrar un motor fuera de control, mientras que con  $a=2$  ningún motor defectuoso pasará desapercibido. considerando  $\alpha = 0.01$  vemos las mismas conclusiones, se nota un cambio sustancial con una perturbación de  $a = 0.5$  y  $a = -0.5$ , hasta que con una perturbación de  $a = 2$  y  $a = -2$  la carta dejará pasar dos motores defectuosos antes de detectar el cambio en la media del proceso.

### 3.4. Carta de control para datos funcionales

En esta sección describimos la carta de control funcional propuesta en esta tesis, ilustramos el procedimiento utilizando los datos del artículo *Amiri*[1], los cuales se presentan en la Figura 3.1 para después comparar la carta de control funcional aquí propuesta contra la carta de control funcional propuesta en dicho artículo.

Los datos que se presentan en *Amiri*[1] los usaremos como datos históricos que nos servirán para encontrar los parámetros de nuestra carta de control, el primer paso del procedimiento consiste en un ajuste  $B - Splines$  de grado 3 con nodos interiores en 2625, 3750 y 4875 a cada uno de los 26 motores, cabe aclarar que los nodos son equidistantes y se eligió la mínima cantidad de ellos que conserven un buen ajuste a los datos, una de las características de esta carta es la reducción de dimensionalidad que con el ajuste mencionado reducimos de 14 a 7 la dimensión de los datos. La Tabla 3.6 muestra los coeficientes del ajuste  $B - spline$  para estos 26 motores y en la Figura 3.6 se muestra el ajuste  $B - spline$ , en esta última figura mostramos que efectivamente el ajuste es excelente ya que las curvas pasan justo por los puntos que se desean ajustar.

Uno de los supuestos fuertes de una carta de control  $T^2$  de Hotelling es que los coeficientes en este caso, tengan una distribución normal multivariada. En la Figura 3.7 se muestran gráficos Q-Q para cada uno de los 7 coeficientes de los 26 motores, por lo que obtenemos 7 gráficos. Notemos que estos gráficos muestran que los coeficientes provienen de una distribución Normal. Además la prueba Mardia no rechaza normalidad multivariada de los vectores  $7 - variados$ , por lo que tenemos que los coeficientes provienen de una distribución normal multivariada. Con este supuesto aclarado, podemos seguir utilizando la carta de control  $T^2$ .

Una vez encontrados los coeficientes del ajuste B-spline, procedemos a la fase I de la carta de control, en donde se elabora una carta de control  $T^2$  de Hotelling con los 26



Tabla 3.6: Coeficientes del ajuste  $B - spline$ 

motor	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$
329	98.484	95.744	126.156	105.508	108.404	86.060	75.090
449	96.361	93.568	123.527	104.410	105.648	85.277	75.159
529	98.702	98.793	119.764	107.497	108.695	88.695	77.731
642	96.569	94.727	121.073	100.957	106.689	84.167	75.276
724	96.698	95.732	119.883	100.156	104.921	82.791	73.284
803	97.580	96.841	122.495	105.563	104.992	82.821	72.065
930	99.984	98.937	121.348	105.328	107.525	85.709	75.288
1148	94.534	98.394	123.742	105.792	106.518	84.823	74.144
1171	96.491	93.685	123.685	104.546	105.784	85.384	75.249
1516	100.666	98.469	122.731	106.528	106.334	84.613	73.426
1791	101.905	100.959	122.375	107.598	106.874	82.954	73.817
2600	96.797	98.333	125.1319	106.340	108.380	85.408	75.985
3100	100.014	99.326	121.087	106.123	105.983	85.701	76.013
3720	97.362	98.186	117.437	104.776	107.209	87.090	76.294
4025	99.963	98.620	119.602	106.045	105.190	84.797	73.264
4068	97.925	100.110	125.597	104.766	108.401	81.486	71.791
4926	97.262	101.150	126.380	107.404	109.369	87.123	76.635
5155	100.927	101.214	119.491	108.266	107.514	87.871	76.591
6143	93.083	96.127	122.168	100.774	108.001	83.894	75.878
6844	93.077	100.66	119.984	106.583	103.992	86.761	74.128
7811	95.273	97.220	119.857	104.599	103.839	83.983	73.706
8007	98.169	95.607	122.539	101.451	108.222	83.134	74.068
8623	96.701	100.135	121.278	103.645	107.661	85.789	76.529
9388	96.409	100.698	122.463	107.705	108.461	86.673	76.754
9404	91.455	97.028	120.311	101.361	105.182	83.953	72.700
10430	98.282	97.292	121.944	106.570	108.008	87.013	75.982

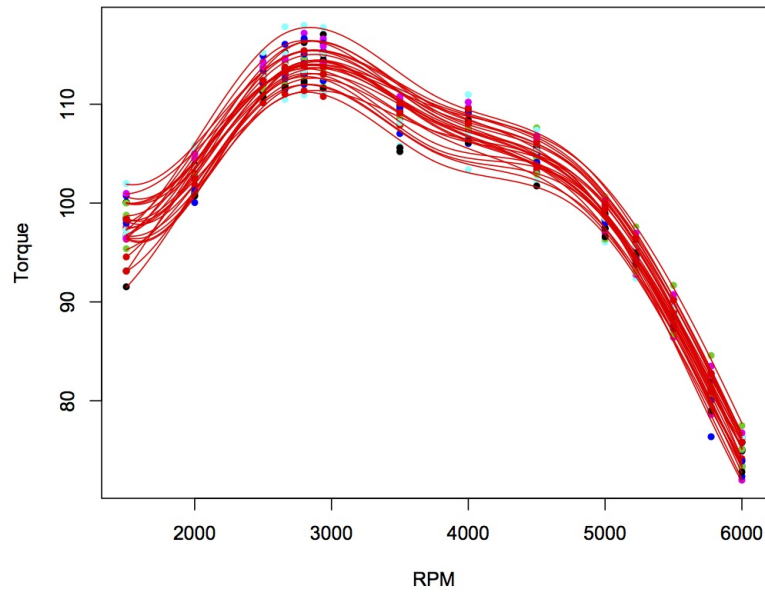


Figura 3.6: Ajuste B-spline

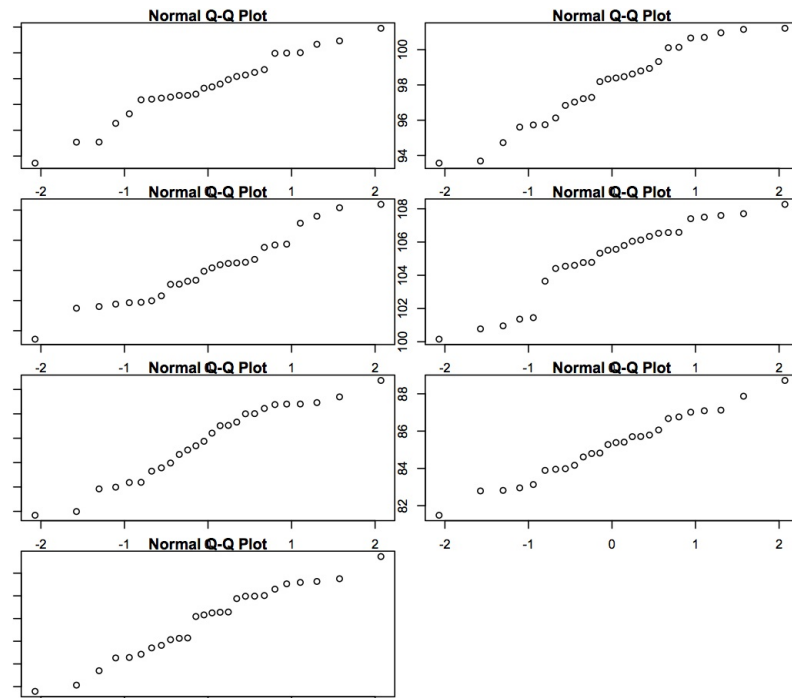


Figura 3.7: Gráficos Q-Q para cada uno de los 7 coeficientes que se obtiene en el ajuste *B – spline*

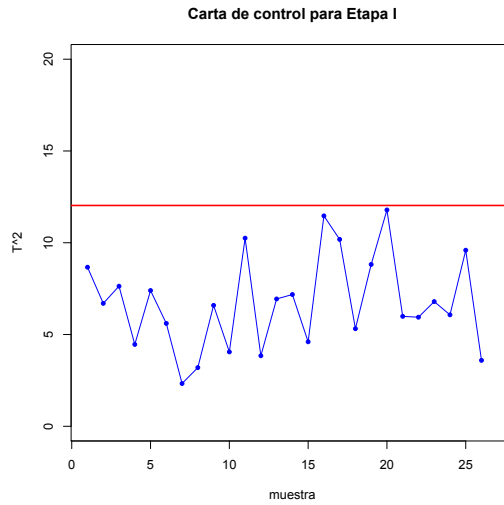


Figura 3.8: Carta de control fase I

motores, con el fin de identificar si en este conjunto de datos se tiene un outlier, para esto el estadístico es:

$$T_j^2 = (c_j - \bar{c})^T S^{-1} (c_j - \bar{c}) \quad (3.9)$$

donde,  $c_j$  es un vector *7-variado* que corresponde a los coeficientes resultantes del motor  $j$ -ésimo, luego  $\bar{c}$  es la media de los coeficientes de los 26 motores, y  $S^{-1}$  es la matriz de varianzas y covarianzas estimada con los coeficientes de los 26 motores.

Como se comentó en la sección 1.7, el límite de control en fase I tiene distribución beta debido a la dependencia entre  $c_j$ ,  $\mu$  y  $S^{-1}$ , por lo que se tiene que el límite de control superior es:

$$LCS = 12.03,$$

La carta de control resultante se muestra en la Figura 3.8.

Notemos que en la carta de control de la Figura 3.8 ningún punto esta por encima del límite de control, por lo que el conjunto de 26 motores es un grupo homogéneo y pueden ser utilizados en la fase II para estimar los parámetros de la carta de control. Las estimaciones para la fase II son:  $\bar{c} = \mu$  y  $S^{-1} = \Sigma$ , las cuales se muestran en la ecuación

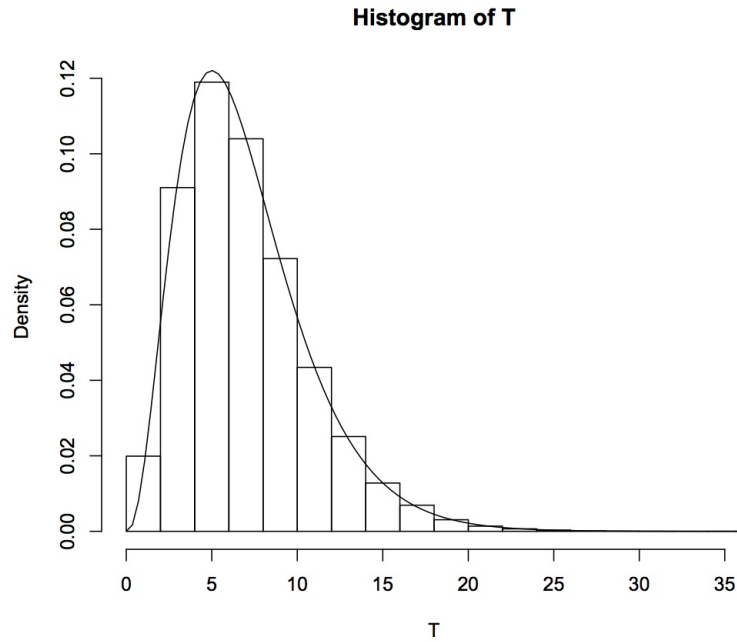


Figura 3.9: Histograma de 10000 estadísticos  $T^2$

3.8, los estadísticos a graficar en esta fase II están dados por:

$$T_j^2 = (c_j - \mu)^T \Sigma^{-1} (c_j - \mu) \quad (3.10)$$

En esta carta de control el límite de control superior esta dado por:

$$LCS = \chi_{0.05,p}^2 = 14.0671, \quad (3.11)$$

este límite de control superior corresponde a un cuantíl de la distribución *ji – cuadrada* de 7 grados de libertad, donde  $\alpha = 0.05$  representa la probabilidad de que un estadístico  $T_j^2$  sea mayor al  $LCS$ . Este límite se eligió ya que el estadístico que se muestra en la ecuación 3.10 tiene ésta distribución, para mostrar esto veamos la Figura 3.9 en la cual se muestran 10000 simulaciones del estadístico  $T_j^2$  y su respectivo histograma es obtenido, en el cual colocamos la densidad de una distribución  $\chi_7^2$ , notemos que el ajuste es excelente por lo que deducimos que  $T_j^2$  tiene una distribución  $\chi_7^2$ .

valor $\alpha$	ARL
0.05	19.96
0.01	99.8

Tabla 3.7: ARL bajo control

### 3.5. ARL de carta de control para datos funcionales

Para encontrar el ARL de esta carta de control necesitamos generar nuevos perfiles de motores, para generar nuevos perfiles los simulamos. En la sección 3.3 describimos un método para simular nuevos perfiles de motores, el cual se basa en suponer una distribución  $y_i \sim N_{14}(\mu, \Sigma)$ , donde estimamos los parámetros con los 26 perfiles de motores que se presentan en el artículo *Amiri*[1].

La manera de encontrar el ARL se describió en la sección 3.3, en este caso usaremos la misma idea, usando de igual manera la media de 10000 *RLs* para el cálculo de *ARL*. En la Figura 3.10 se muestra el histograma para 10000 *RLs* calculados mediante simulaciones, lo cual se describió en la sección 3.3, En este histograma se montó una distribución geométrica de parámetro  $p = 0.05$ , con esto demostramos que la variable aleatoria *RL* tiene distribución geométrica de parámetro  $p = 0.05$ .

En la Tabla 3.7 se muestra el ARL obtenido tomando  $\alpha = 0.05$  y  $\alpha = 0.01$  cuando el proceso esta en control.

En la Tabla 3.8 se muestran los resultados de ARL, para lo cual se utilizó una perturbación a la media del proceso de la forma  $P = aS_i$  para  $i = 3, 4, 5, 6, 7$ , estos valores de  $i$  corresponden a los valores de RPM 2500, 2660, 2800, 2940, 3500, respectivamente. Esto nos indica que el vector  $P$  tiene todas sus entradas iguales a cero excepto la entrada  $i$ , con  $a$  un valor real y  $S_i$  la desviación estándar de 26 motores en uno de los valores de RPM 2500, 2660, 2800, 2940, 3500 , respectivamente.

Lo que observamos en la Tabla 3.8 es que la carta de control propuesta en esta tesis es más efectiva que la de *Amiri* [1] para detectar cambios en la media del proceso, cuando la perturbación se presenta en una sola entrada del vector de medias, esta característica es

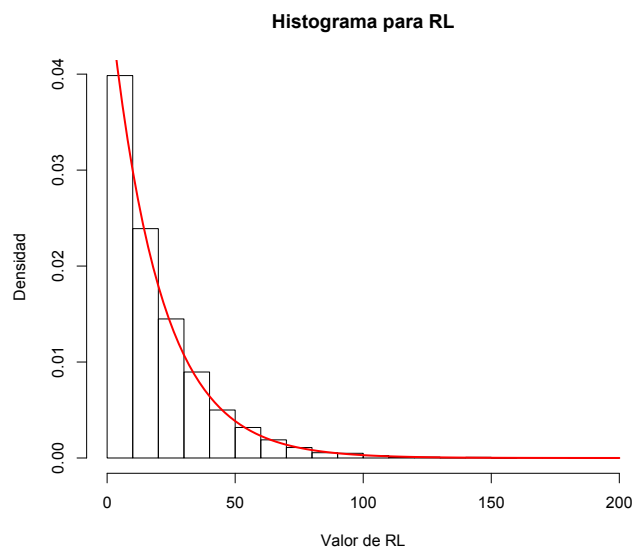


Figura 3.10: Histograma de 10000 RLs

muy importante ya que este cambio es muy pequeño, el cual en muchas ocasiones no puede ser percibido por personas alrededor del proceso. Para llegar a estas conclusiones basta comparar los resultados de la carta de control del artículo *Amiri[1]* que se presentan en la Tabla 3.4 contra los resultados obtenidos para la carta de control propuesta en esta tesis que se presentan en la Tabla 3.8. Aunque lo anterior se hizo para un área determinada en la cual se vio un pobre ajuste del modelo de efectos mixtos, en el Apéndice C se muestran para todos los demás RPM y se llegan a las mismas conclusiones.

Ahora consideremos perturbaciones a la media del proceso que no sólo afecten en un valor de RPM, sino que ahora el vector de perturbación sea  $P = aS$ , con  $a$  un valor real y  $S$  un vector 14 – *variado* que contiene la desviación estándar de 26 motores en cada uno de los 14 valores de RPM. Los resultados considerando dos valores  $\alpha = 0.05$  y  $\alpha = 0.01$  se muestran en la Tabla 3.9.

Para hacer una comparación amplia entre la carta de control propuesta en el artículo *Amiri [1]* y la carta de control propuesta en esta tesis, primero analizamos las tablas 3.4 y 3.8, de estas dos notamos que la carta de control propuesta en *Amiri [1]* no es

$P_3$		$P_3$		$P_4$		$P_4$	
a	ARL	a	ARL	a	ARL	a	ARL
-0.1	19.9	0.1	19.99	-0.1	19.9	0.1	19.92
-0.5	18.34	0.5	18.30	-0.5	18.21	0.5	18.17
-0.8	16.34	0.8	16.32	-0.8	15.87	0.8	15.93
-1	14.6	1	14.67	-1	14.25	1	14.23
-2	7.3	2	7.38	-2	6.78	2	6.77
-3	3.6	3	3.65	-3	3.22	3	3.22

$P_5$		$P_5$		$P_6$		$P_6$	
a	ARL	a	ARL	a	ARL	a	ARL
-0.1	20	0.1	19.94	-0.1	19.81	0.1	20
-0.5	18.33	0.5	17.94	-0.5	18.01	0.5	17.91
-0.8	15.53	0.8	15.61	-0.8	15.26	0.8	15.05
-1	13.61	1	13.49	-1	13.52	1	13.42
-2	6.15	2	6.13	-2	5.88	2	5.99
-3	2.89	3	2.87	-3	2.72	3	2.72

$P_7$		$P_7$	
a	ARL	a	ARL
-0.1	20	0.1	19.74
-0.5	14.24	0.5	14.54
-0.8	9.7	0.8	9.52
-1	7.12	1	7.03
-2	1.9	2	1.92
-3	1.11	3	1.1

Tabla 3.8: Resultados ARL, sólo considerando zona de interés.

Para $\alpha = 0.05$		Para $\alpha = 0.05$	
Valor de $a$	ARL	Valor de $a$	ARL
0.1	19.14	-0.1	19.39
0.5	10.62	-0.5	10.54
0.6	8.3	-0.6	8.52
0.8	5.26	-0.8	5.35
1	3.58	-1	3.48
1.5	1.59	-1.5	1.58
2	1.11	-2	1.11

Para $\alpha = 0.01$		Para $\alpha = 0.01$	
Valor de $a$	ARL	Valor de $a$	ARL
0.1	95.6	-0.1	96.4
0.5	39.8	-0.5	38.1
0.6	29.9	-0.6	29.3
0.8	15.7	-0.8	15.6
1	8.6	-1	8.6
1.5	2.6	-1.5	2.5
2	1.3	-2	1.3

Tabla 3.9: Resultados ARL

adecuada para detectar cambios pequeños en la media del proceso, en donde un cambio pequeño lo definimos como una perturbación en un sólo valor de RPM, de hecho aún con una perturbación de  $a = 2$  vemos que en la carta del artículo pasarían desapercibidos en promedio 16 motores en malas condiciones antes de detectar una salida de control, mientras que en la carta de control propuesta en esta tesis pasarían desapercibidos en promedio 6 motores en malas condiciones antes de que la carta detecte una salida de control.

Ahora analizando las tablas 3.5 y 3.9 vemos que la carta de control propuesta en esta tesis detecta de manera más eficaz cambios en la media del proceso cuando esta perturbación se presenta en todo el vector de medias, aunque la diferencia entre los ARLs de las dos cartas de control no es muy grande, la carta de control propuesta en esta tesis es mejor, por ejemplo con  $\alpha = 0.05$  la diferencia es de aproximadamente 2, si tomamos  $a = 0.8$  vemos que la carta de control de esta tesis dejará pasar 5 motores defectuosos antes de detectar una salida de control mientras que la carta del Amiri[1] dejará pasar 7



motores antes de detectar. Vemos que con  $\alpha = 0.01$  la diferencia es mayor pero siempre resulta mejor la carta de control propuesta en esta tesis.

En conclusión con estos resultados podemos decir que la carta de control propuesta en esta tesis detecta más rápido cambios pequeños y grandes del proceso que la carta de control propuesta en *Amiri*[1], por lo que en un proceso industrial es preferible utilizar la carta de control propuesta en esta tesis.

### 3.6. Áreas Importantes

En esta sección se mostrará como podemos dar más importancia a ciertas áreas de un perfil, es decir, si en un estudio se sabe que una cierta área del análisis es más importante que otra, entonces, se colocan más nodos en esa área y de esa manera si existen perturbaciones del proceso en esa importante área, la carta de control detectara más rápido el cambio que si la perturbación se presenta en otra área.

Para probar lo anterior utilizaremos la base de datos de motores que se encuentra en el artículo *Amiri*[1] y que se estudió en secciones anteriores. Para esto elegimos como área importante el área en el intervalo [2500, 2940] esto debido a su forma tan peculiar, luego colocamos 5 nodos de dos maneras distintas, en la primera la cual llamaremos *nodosIM* se colocan como sigue:

$$\text{nodosIM} = 2500, 2700, 2940, 3750, 4875$$

luego la segunda la cual llamaremos *nodosN* se colocan los 5 nodos de manera equiespaciada en el intervalo [1500, 6000], como sigue:

$$\text{nodosN} = 2250, 3000, 3750, 4500, 5250$$

los ajustes B-spline considerando *nodosIM* se muestra en la Figura 3.11 y considerando

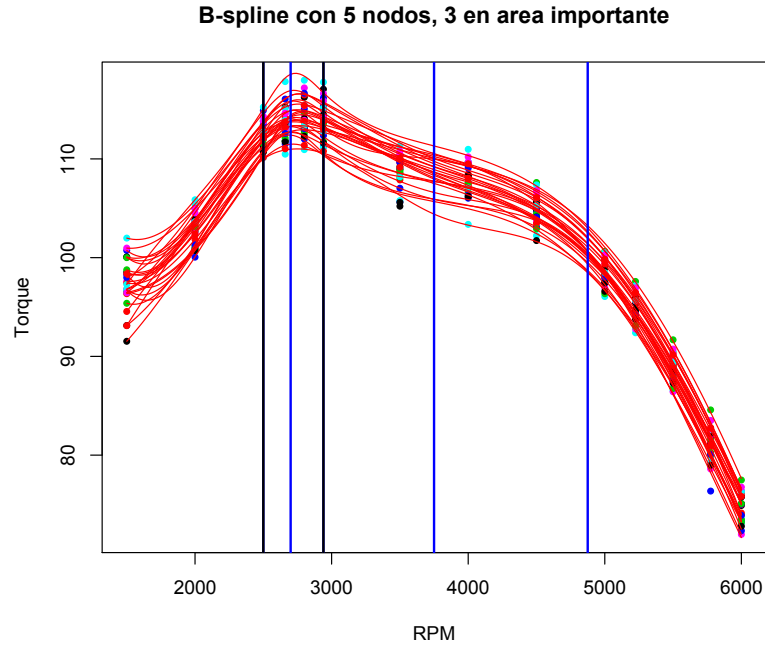


Figura 3.11: Ajuste considerando nodosIM

nodosN se muestra en la Figura 3.12, en ambas gráficas las líneas azules indican donde están colocados los nodos y las líneas negras delimitan el área importante.

Ahora calculamos los ARL de la carta de control funcional para cada manera de colocar los nodos, para esto consideramos un vector  $P_i = aS_i$ , donde  $a$  es un valor real y  $S_i$  es la desviación estándar de 26 motores en un valor RPM, como lo hicimos antes  $P_i$  es un vector con todas sus entradas iguales a cero excepto en la entrada  $i$ , para estos cálculos se simula considerando una distribución Normal 14 – *variada* tomando como  $\mu$  la media muestral de los 26 motores y como  $\Sigma$  la matriz muestral de varianzas y covarianzas de los 26 motores, se hicieron 10,000 repeticiones de RL y el límite de control para la fase 2 es:

$$\begin{aligned} LC &= \chi_{(0.05,9)}^2 \\ &= 16.9 \end{aligned}$$

En la Tabla 3.10 se muestran los resultados obtenidos, notemos que los primeros 4

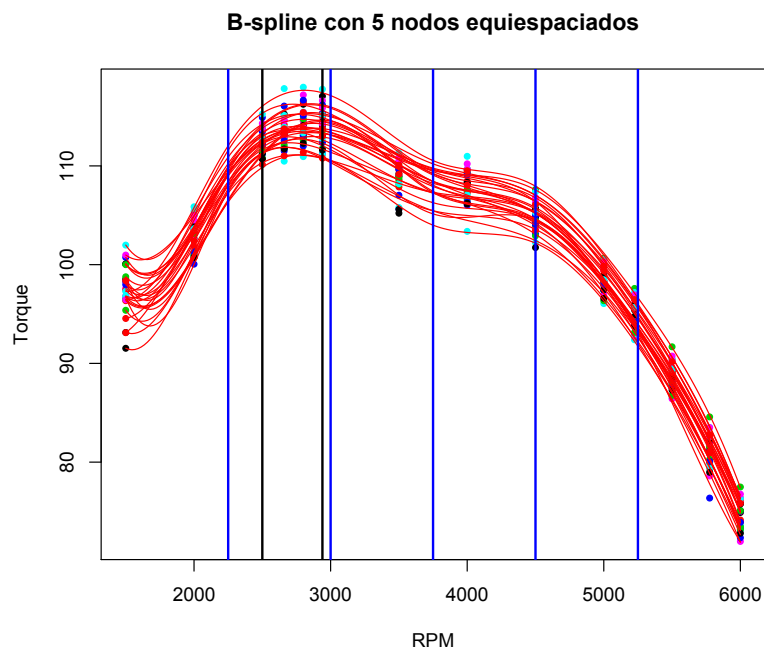


Figura 3.12: Ajuste considerando nodosN

tablas están dentro del área importante y se muestra que el ARL considerando nodosIM es menor que considerando nodosN, por otro lado vemos que los 2 últimos cuadros no están en el área importante en estos se muestra que se obtiene el mismo ARL considerando nodosIM que considerando nodosN.

Con lo anterior descrito podemos decir que utilizando la carta de control propuesta en esta tesis se pueden elegir regiones importantes en los datos y darle un mayor peso a éstas, si se tiene una región importante, entonces la carta de control detectará más rápido cambios en esa área, mientras que si el cambio se realiza fuera de la región importante la carta de control detectará el cambio igual de rápido que si colocamos los nodos equiespaciados. Éste último enfoque, se demostró que es más eficiente para detectar cambios en el proceso que la carta de control propuesta en *Amiri* [1]

### 3.7. ¿Qué sucede cuando no se tiene normalidad?

Como se mencionó en la sección 3.4 uno de los supuestos para la construcción de la

Perturbación en RPM=2500		
a	nodosIM	nodosN
0.1	18.9	19.8
0.5	6.1	13.1
0.8	2.49	7.7
1	1.6	5.3
2	1	1.4
3	1	1

Perturbación en RPM=2660		
a	nodosIM	nodosN
0.1	19.3	19.8
0.5	12.9	16.6
0.8	7.8	13.3
1	5.2	10.8
2	1.4	3.7
3	1	1.6

Perturbación en RPM=2800		
a	nodos IM	nodosN
0.1	19.4	19.8
0.5	10.7	17.6
0.8	5.7	14.9
1	3.8	12.5
2	1.3	5.22
3	1	2.3

Perturbación en RPM=2940		
a	nodosIM	nodosN
0.1	19.02	19.7
0.5	11.4	13.39
0.8	6.2	8.1
1	4.08	5.6
2	1.1	1.49
3	1	1.02

Perturbación en RPM=2000		
a	nodosIM	nodosN
0.1	19.3	19.4
0.5	10.9	10.9
0.8	5.7	5.6
1	3.7	3.7
2	1.1	1.1
3	1	1

Perturbación en RPM=3500		
a	nodosIM	nodosN
0.1	18.9	19.33
0.5	9.6	9.5
0.8	4.6	4.6
1	2.9	2.9
2	1.04	1.05
3	1	1

Tabla 3.10: Resultados ARL

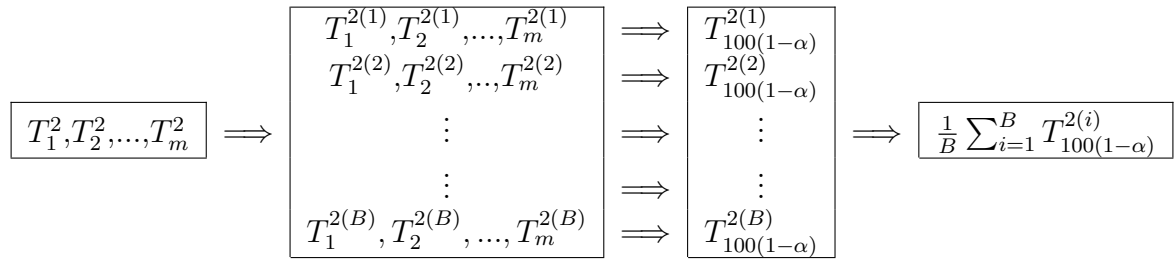


Figura 3.13: Método Bootstrap

carta de control  $T^2$  de Hotelling es que los coeficientes  $B$ -spline tengan una distribución normal multivariada, lo cual en la mayoría de los casos no se tiene, es decir, en la ecuación 3.12  $c_i$  debe tener una distribución normal multivariada.

$$T_i^2 = (c_i - \bar{c})^T S^{-1} (c_i - \bar{c}) \quad (3.12)$$

La principal dificultad con la ausencia de normalidad es que no se sabe exactamente la distribución del estadístico  $T^2$ , y por tanto no es claro cuales deberían ser los límites de control, dicha distribución en el caso normal es una distribución  $F$  o una distribución Ji-cuadrada, lo cual se mostró en la sección 1.8. Un enfoque muy útil para lidiar con la ausencia de normalidad se puede consultar en *Phaladiganon et al.* [14], en donde se propone un método Bootstrap para estimar los límites de control de la carta  $T^2$  de Hotelling, basado en las observaciones  $T_i^2$  con  $i = 1, 2, 3, \dots, m$  que se tienen de un proceso que esta en control, como en la fase I, recordemos que en esta fase I se tiene datos históricos del proceso, esto nos ayudará a determinar los límites de control. El esquema del procedimiento se muestra en la Figura 3.13.

Un dato importante es que en *Phaladiganon et al.* (2011), se muestra que resulta más eficiente utilizar este procedimiento comparándolo con un estimador Kernel, para determinar los límites de control. El número de muestras Bootstrap que se recomienda es de por lo menos  $B = 1,000$ .

### 3.8. Ejemplo

Para ilustrar el uso de la carta de control funcional propuesta en este trabajo, haremos una aplicación a una base de datos llamado *tecolor*. En esta base de datos se presentan 240 muestras de carne molida a las cuales se les midieron 125 atributos: 100 son absorbencias, luego 22 coeficientes de componentes principales, luego humedad, grasa y proteína. El objetivo principal era predecir el porcentaje de grasa en una muestra de carne molida con los 100 primeros atributos de absorbencia.

La medida de absorbencia es utilizada en espectrofotometría la cual se define como:

$$A_\lambda = -\log_{10}\left(\frac{I}{I_0}\right), \quad (3.13)$$

donde,  $I$  es la intensidad de la luz con una longitud de onda específica  $\lambda$ , que pasa por una muestra que analizamos (intensidad de la luz transmitida),  $I_0$  es la intensidad de la luz antes de que entre a la muestra (intensidad de la luz incidente). Las medidas de absorbencia son frecuentemente usadas en química analítica, ya que la absorbencia es proporcional al grosor de una muestra y la concentración de la sustancia en ésta. Se tienen 100 atributos de absorbencia porque se midió la absorbencia a 100 valores  $\lambda$  diferentes que representa las longitudes de onda.

Aplicaremos la carta de control funcional descrita en la Sección 3.4 para discernir entre muestras de carne que tienen mayor del 12% de grasa, esto se puede utilizar en el caso de que una empresa desee monitorear un proceso de producción de carne, suponiendo que 12% es un valor crítico, es decir, que la carne no se puede vender si presenta más del 12% de grasa por ejemplo. Para mostrar la efectividad de la carta de control, usaremos las primeras 129 muestras de carne para estimar los parámetros de la carta de control multivariada  $T^2$  de Hotelling, luego aplicaremos el procedimiento a las 111 muestras restantes, para de ésta manera medir la efectividad. La Figura 3.14 muestra las curvas de 129 muestras de carne molida, donde graficamos los diferentes valores de longitud de onda

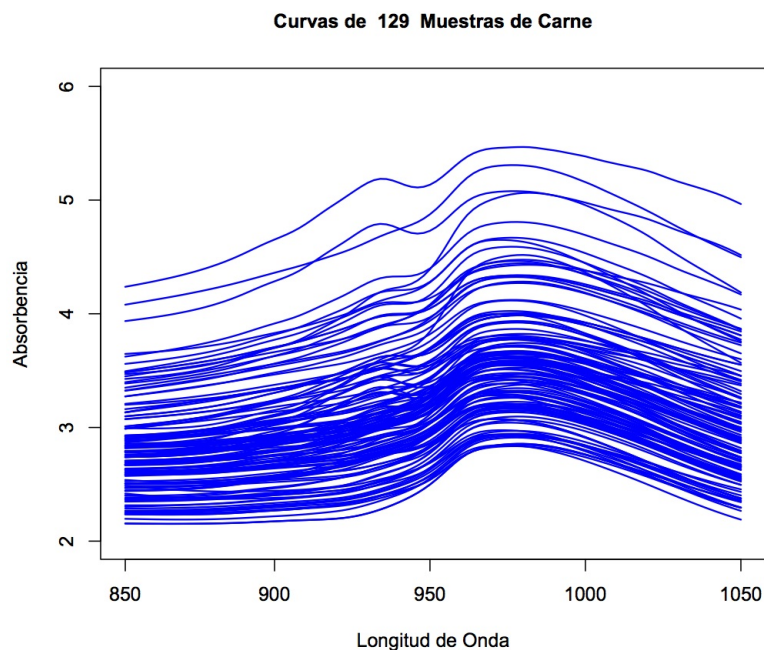


Figura 3.14: Curvas de 129 muestras de carne

contra la absorbencia de cada una de las 129 muestras.

De acuerdo a la carta de control funcional que se describe en este trabajo, lo siguiente es encontrar los coeficientes de un ajuste B-Spline de tercer grado, en este caso se decidió usar 14 nodos para el ajuste del cual resultaron 18 coeficientes. Los nodos y el grado se eligieron de tal manera que se obtuvieran pocos coeficientes en el ajuste pero con un ajuste razonable a los datos. La Figura 3.15 muestra el ajuste que se obtuvo con este número de nodos. Notemos que el ajuste es muy bueno, ya que las curvas de las Figuras 3.14 y 3.15 muestran gran semejanza.

Ahora de estas 129 muestras de carne, tomamos sólo las que tengan grasa menor a un 12%, las cuales son 58. En la Figura 3.16 se presentan las curvas de 58 muestras de las 129 que tienen grasa menor al 12%.

Después en la Figura 3.17 se muestra el ajuste B-spline con 14 nodos a estas 58 muestras de carne, lo cual resulta ser muy buen ajuste. Para calcular los parámetros de la carta de control  $T^2$  de Hotelling primero necesitamos determinar si las 58 muestras de

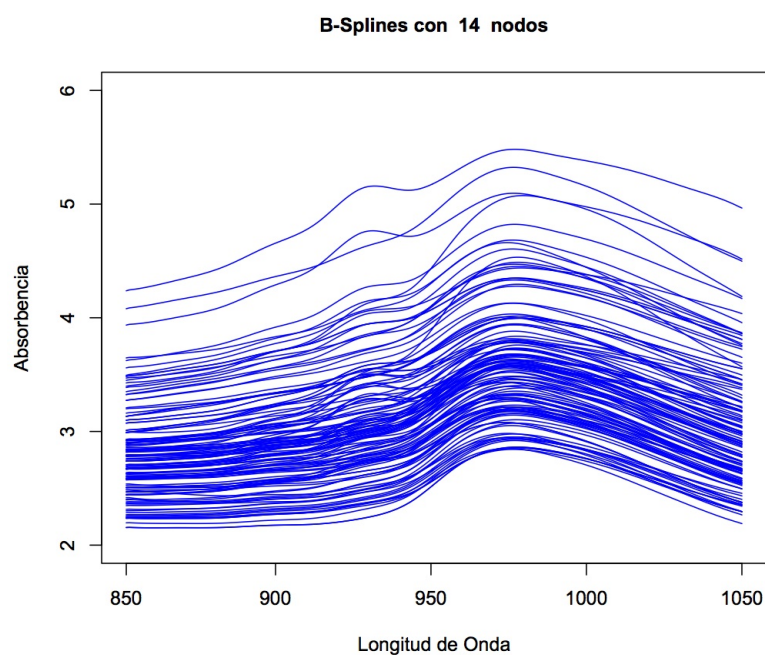
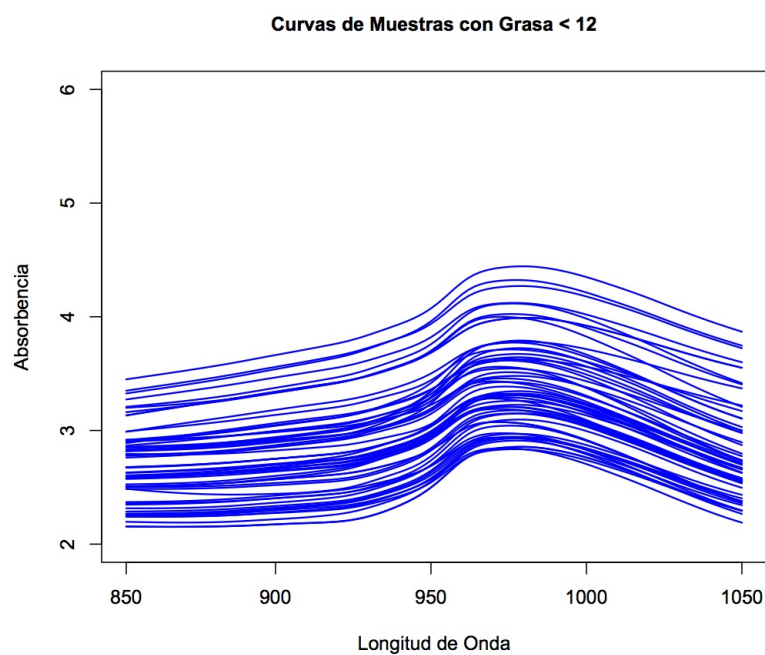
Figura 3.15: Ajuste  $B - spline$  a las 129 muestras de carne.

Figura 3.16: Curvas de las muestras con grasa &lt; 12%



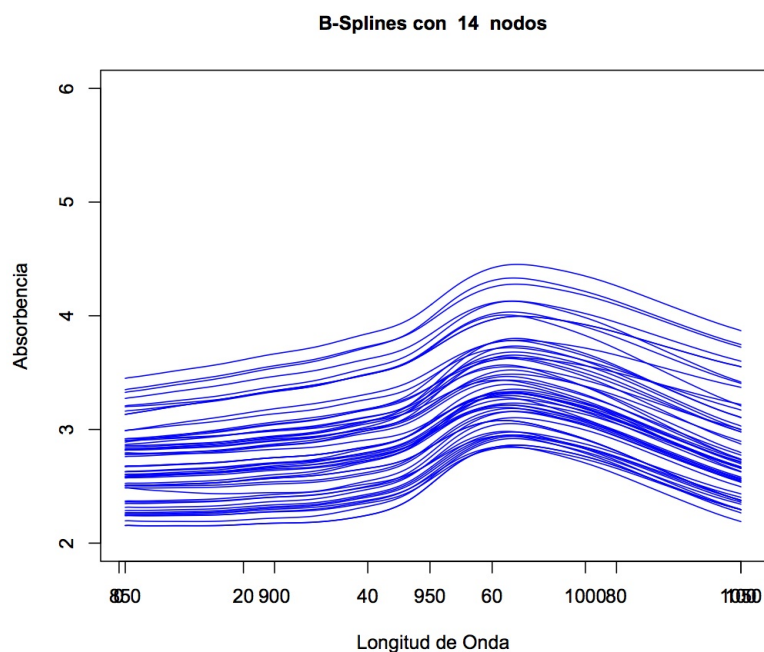


Figura 3.17: Ajuste B-spline de las muestras con grasa  $< 12\%$

carne con grasa menor a  $12\%$  están bajo control estadístico, para ello en la Figura 3.18 se muestra la carta de control para la Fase I, la cual tiene límite de control:

$$LC = 25.98$$

En la Figura 3.18 notamos que 8 puntos caen fuera del límite de control, entonces, lo que se hace en estos casos es parar e inspeccionar las muestras de carne que dan señal fuera de control para identificar por qué dan esta señal, en este caso al inspeccionar las 8 muestras de carne no se puede determinar por qué dan señal fuera de control, por lo que no se eliminan, al contrario, se siguen utilizando para las estimaciones; si se hubiera encontrado la causa de la salida de control, entonces, se remueve del conjunto. Otra razón por la que no se remueve de las estimaciones es que en este ejemplo no se tiene normalidad en los coeficientes, esto se muestra porque la prueba de Mardia (*Mardia [16]*) rechaza normalidad multivariada obteniendo los p-valores  $p.value.skew = 0.0000$  y  $p.value.kurt = 0.0000$ , donde claramente se rechaza normalidad.

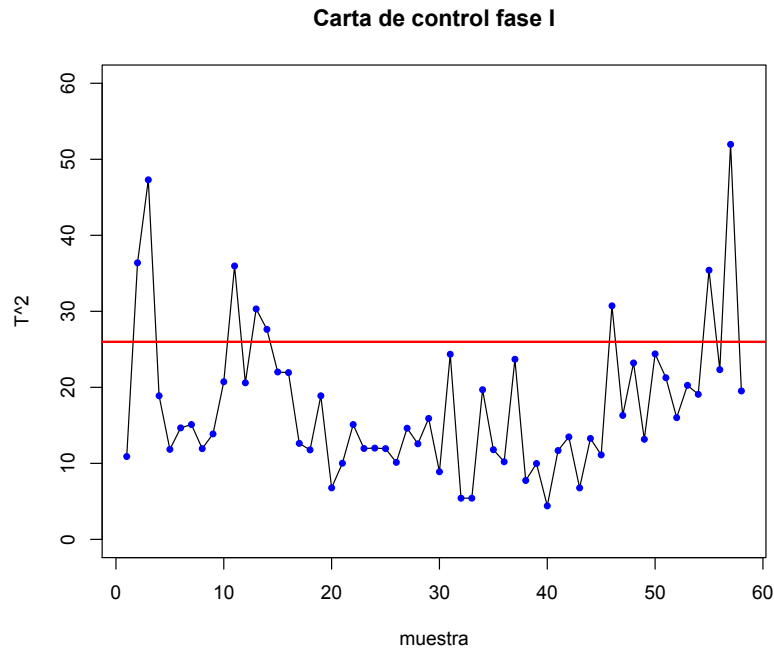


Figura 3.18: Carta de control Fase I

Ahora lo que hacemos es calcular los parámetros de la carta de control  $T^2$  de Hotelling con los coeficientes del ajuste  $B - spline$  de este conjunto de 58 muestras, los dos parámetros que estimamos son  $\bar{x}$  y  $S^{-1}$ . Con esto construimos en la fase II los estadísticos  $T_i^2$  con el resto de muestras de carne (111):

$$T_i^2 = (c_i - \mu)^T S^{-1} (c_i - \mu)$$

El límite de control superior para la carta de control  $T^2$  de Hotelling de acuerdo a la sección 1.8:

$$\begin{aligned} LCS &= \frac{p(m+1)(m-1)}{m^2 - mp} F_{1-\alpha, p, m-p} \\ &= 28.86 \end{aligned}$$

donde  $m = 58$ ,  $p = 18$ , esto se debe a que resultaron 18 coeficientes del ajuste  $B - Spline$ , y en este caso tomamos un  $\alpha = 0.05$ , es decir, la probabilidad de que una muestra esté por encima del límite de control es 0.05. En la Tabla 3.11 se muestran los resultados que arrojó

		Estimado	
		>12 %	<12 %
Real	>12 %	53	2
	<12 %	11	45

Tabla 3.11: Resultados de carta de control

		Estimado	
		>12 %	<12 %
Real	>12 %	50	5
	<12 %	5	51

Tabla 3.12: Mejor clasificación

la carta de control, en donde vemos que la carta de control dio conclusiones erróneas 13 veces, obteniendo un 88.2 % de decisiones correctas.

Moviendo  $\alpha$  y el número de nodos en el ajuste se encontró que la mejor clasificación se obtiene con  $\alpha = 0.003$  y  $n = 14$  nodos, obteniendo un 91 % de decisiones correctas, donde se obtuvo sólo 10 errores de clasificación como lo muestra la Tabla 3.12.

Enseguida utilizamos el procedimiento Bootstrap propuesto en *Phaladiganon et al. [14]* en el cual se utiliza la misma base de datos *tecolor* pero sólo se utilizan las 215 primeras muestras. Comparamos la carta de control propuesta en esta tesis utilizando el procedimiento Bootstrap con los métodos propuestos en *Wu y Liu [17]*, en este último se proponen dos métodos de clasificación basados en componentes principales funcionales llamados RSVM (Robust support vector machine) lineal funcional y RSVM no-lineal funcional, estos procedimientos los comparan con el método de Discriminante Lineal Funcional. En el ejemplo presentado en dicho artículo se desea predecir cuales muestras de carne tienen nivel de grasa mayor y menor a 20 %, se toman 155 muestras al azar de las 215 muestras, de este conjunto se extraen las muestras que tienen menor a 20 % de grasa y con estas se realiza las estimaciones y las 60 restantes se utilizan para realizar la clasificación. Los resultados que obtuvieron son:

- Análisis Discriminante Lineal-  $\frac{52}{60} = 86.6\%$  *Decisiones correctas*
- RSVM lineal Funcional-  $\frac{59}{60} = 98.3\%$  *Decisiones correctas*
- RSVM no-lineal Funcional-  $\frac{58}{60} = 96.6\%$  *Decisiones correctas*

Para aplicar el método de Bootstrap usaremos esta misma base de datos *teclator* con las primeras 215 muestras de carne, aplicaremos la carta de control funcional para clasificar las muestras con menor y mayor cantidad de 20% de grasa, el procedimiento paso a paso se muestra a continuación:

1. Del total de 215 muestras se eligen al azar 155 muestras las cuales serán usadas para estimar los parámetros de la carta de control funcional, y el resto 60 muestras serán usadas para predecir.
2. Elegimos las muestras que tengan grasa menor a 20% del conjunto de 155 muestras elegidas en el paso 1.
3. Hacemos el ajuste B-spline de grado 3 con 14 nodos equiespaciados y extraemos los coeficientes para cada una de las muestras elegidas en el paso 2.
4. Con los coeficientes que se obtienen en el paso 3, estimamos los parámetros de la carta de control  $\bar{x}$  y  $S^{-1}$ .
5. Aplicamos el procedimiento Bootstrap que se muestra en la Figura 3.13 con  $B = 5000$  y se obtiene el límite de control  $LC = \frac{1}{B} \sum_{i=1}^B T_{100(1-0.05)}^{2(i)}$ .
6. Para el resto, 60 muestras, obtenemos los coeficientes del mismo ajuste B-spline, luego aplicamos  $T_j^2 = (c_j - \bar{x})^T S^{-1} (c_j - \bar{x})$  y las clasificamos utilizando el límite de control  $LC$  obtenido en paso 5, y con esto obtenemos el porcentaje de correctas clasificaciones.

7. Repetimos todos los pasos anteriores 1,000 y obtenemos la media de todos los porcentajes de correctas clasificaciones obtenidas en paso 6.

Una vez aplicado el procedimiento anterior obtuvimos el siguiente resultado:

$$91.59\% \approx \frac{55}{60}$$

Con los resultados anteriores vemos que la clasificación obtenida utilizando el límite de control obtenido por Bootstrap resulta mejor que los resultados que obtenemos con el Análisis de Discriminante Lineal, y vemos que las clasificaciones correctas acumulan un buen número en comparación de los métodos desarrollados en *Phaladiganon et al.*[14] , aunque no resulta mejor, se obtienen muy buenos resultados debido a la sencilla aplicación de la carta de control funcional.

La carta de control funcional propuesta en esta tesis tiene una amplia gama de aplicaciones, este ejemplo muestra una de ellas. En este ejemplo utilizamos la carta de control funcional propuesta en esta tesis para detectar y discernir entre muestras de carne que tengan un cierto porcentaje de grasa, en este ejemplo tomamos un 12% y 20% de grasa como valor crítico. Este ejemplo sería útil en empresas que producen carne, cuando el porcentaje de grasa es una característica de calidad, supongamos que cuando la carne supere el 12% de grasa no es factible comercializarla, en este caso la empresa debe supervisar su producción y este tipo de cartas de control ayudan en esta tarea. Cuando en el proceso se detecta un punto fuera de control ( es decir, una muestra de carne con una cantidad mayor al 12% de grasa) el proceso se detiene y se investiga la causa para evitar pérdidas.

# Capítulo 4

## Conclusiones

A lo largo de este trabajo desarrollamos teoría de cartas de control y de Análisis de Datos Funcionales, la combinación de estas teorías dan como resultado cartas de control para datos funcionales, este tipo de cartas de control han tenido una gran importancia en tiempos recientes, ya que en la actualidad la obtención de datos es más sencilla, rápida y económica, gracias al avance de la tecnología. Por esta razón en muchos problemas de monitoreo de procesos industriales intervienen una gran dimensionalidad de los datos así como gran cantidad.

Las cartas de control son una herramienta muy usada en la industria, aunque es una herramienta sencilla tiende a resolver problemas de gran interés en un proceso industrial, además de que esta herramienta permite monitorear y detectar procesos que están fuera de control, de una manera muy visual y sencilla. Cuando una carta de control da señales fuera de control, se prosigue a investigar la causa para después corregirla, de esta manera las empresas evitan pérdidas económicas mejorando y manteniendo un proceso industrial.

En la actualidad la tecnología permite obtener información completa de un proceso, esto nos lleva a lidiar con dimensiones altas. Herramientas como las cartas de control típicas pierden eficacia en estas situaciones, ya que no se construyeron tomando en cuenta dimensiones altas, por ello el trabajar con datos funcionales nos permite reducir la dimensionalidad y de esta manera mejorar las herramientas clásicas.

En el artículo *Amiri*[1] se propone una carta de control para datos funcionales, la

idea central se describe en la sección 3.2. La carta de control propuesta en esta tesis se compara mediante ARL con la carta de control propuesta en el artículo mencionado. Resultando siempre mejor la carta de control propuesta en esta tesis.

En este trabajo describimos una nueva carta de control para datos funcionales basada en los coeficientes de un ajuste B-spline, con un determinado número de nodos. La elección de nodos sólo interviene en un buen ajuste de los datos, la ubicación de los nodos se hace de manera equidistante para que la carta sea más flexible y sencilla. También se mostró que si se tiene una cierta área importante en el estudio, entonces, es preferible la elección de cierto patrón de ubicación de los nodos, este patrón se basa en colocar más nodos en el área importante.

Comparando la carta de control propuesta en esta tesis con la carta de control del artículo *Amiri*[1], se observó que la carta de control propuesta en esta tesis es mejor para detectar cambios pequeños y grandes en el proceso, mientras que la carta de control propuesta en el artículo *Amiri*[1] se tardaría mucho en detectar estos cambios. En la industria es muy importante detectar los cambios pequeños del proceso, ya que regularmente los empleados y supervisores no se dan cuenta y pasan desapercibidos, mientras que los cambios grandes en el proceso regularmente se pueden detectar de manera más fácil, pero de igual manera los cambios grandes son sumamente importantes.

Por último, se consideró un ejemplo de aplicación de la carta de control para datos funcionales descrita en este trabajo, la cual resultó eficiente para resolver un problema de clasificación de muestras de carne, de acuerdo al contenido de grasa. Esta aplicación sería útil en una empresa que produce carne y que le interese monitorear el contenido de grasa en su producción como una característica de calidad.

Consideramos que esta carta de control funcional es una buena alternativa, se tenga o no normalidad, cuando la calidad que se quiere monitorear se puede representar de forma funcional, lo cual mejora algunas herramientas clásicas, ya que provee una manera eficiente para el análisis de grandes conjuntos de datos.

# Bibliografía

- [1] Amiri, A., Jensen, W.A., and Kazemzadeh, R.B., (2010). “A Case Study on Monitoring Polynomial Profiles in the Automotive Industry,” *Quality and Reliability Engineering International*, 26(5), 509-520.
- [2] Ramsay, J. O. and Silverman, B. W. (1997). *Functional Data Analysis*. New York: Springer-Verlag.
- [3] Ramsay, J. O. and Silverman, B. W. (2002). *Applied Functional Data Analysis, methods and case studies*. New York: Springer-Verlag.
- [4] Ramsay, J. O. and Silverman, B. W. (2009). *Functional Data Analysis*. New York: Springer-Verlag.
- [5] Walker, E. and Wright, P., (2000). Comparing curves using additive models. *Journal of Quality Technology*, 21, 118-129.
- [6] de Boor, C. (2001). *A practical guide to Splines*. New York: Springer-Verlag.
- [7] Pinheiro, C. J. and Bates, M. D. (2000). *Mixed-Effects Models in S and S-plus*. New York: Springer-Verlag.
- [8] Montgomery, D. C. (1991). *Introduction to statistical Quality control*. John Wiley & Sons.
- [9] Ferraty, F. and Vieu P. (2006). *Nonparametric Functional Data Analysis*. New York: Springer-Verlag
- [10] Ramsay, J. O., Hooker, G. and Graves, S. (2009). *Functional data analysis with R and Matlab*. New York: Springer-Verlag.
- [11] Gutiérrez, P. H. y De la Vara, S. R. (2004). *Control estadístico de calidad y seis sigma*. McGraw-Hill Interamericana, México.



- 
- [12] Lowry, C.A and Montgomery, D.C. (1995). A review of multivariate control charts. IIE transactions, 27, 800-810
- [13] Tracy, N.D, Young, J.C. and Mason, R.L. (1992). Multivariate control charts for individual observations. Journals of Quality Technology, 24,88-95
- [14] Phaladiganon, P., Bum, S.B, Chen,V.C.P. (2011). Bootstrap-Based  $T^2$  Multivariate Control Chart. Communications in statistics-Simulation and Computation, 40(5), p.645
- [15] Vargas J.A. (2003) Robust estimation in multivariate control charts for individual observations. Journal of Quality Technology ; 35(3):367-376
- [16] Mardia, K.V. (1974). Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. Sankhya, Series B, 36, 115-128.
- [17] Wu, Y. and Liu Y. (2013). Functional robust support vector machines for sparse and irregular longitudinal data. Journal of computational and Graphical Statistics, 22, 2, 379-395.

# Apéndice A

## Estimaciones para el modelo de efectos mixtos.

A continuación demostraremos que los estimadores para los efectos fijos y efectos aleatorios en un modelo de efectos mixtos utilizados en la carta de control funcional propuesta en el artículo *Amiri*[1], están dados efectivamente por las ecuaciones (3.3) y (3.4).

Sea  $y_{ij} = \text{Torque observado}$ ,  $i = 1, 2, \dots, M$  con ( $M=26$  motores) y  $j = 1, 2, \dots, R$  con ( $R=14$  Revoluciones por minuto observado). El modelo de efectos mixtos se expresa como:

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 r_j + \dots + \beta_k r_j^k + \\ &\quad b_{i0} + b_{i1} r_j + \dots + b_{ik} r_j^k + e_{ij} \\ &= (1, r_j, \dots, r_j^k) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + (1, r_j, \dots, r_j^k) \begin{pmatrix} b_{i0} \\ b_{i1} \\ \vdots \\ b_{ik} \end{pmatrix} + e_{ij} \\ &= X\beta + X\delta + e_i, \end{aligned}$$

$$\text{donde } X = \begin{pmatrix} 1 & r_1 & \cdots & r_1^k \\ 1 & r_2 & \cdots & r_2^k \\ \vdots & \vdots & \ddots & \vdots \\ 1 & r_R & \cdots & r_R^k \end{pmatrix}_{R \times (k+1)}, \quad \delta_i \sim N_{(k+1)}(0, D) \text{ y } e_i \sim N_R(0, \sigma^2 I).$$

El modelo se puede expresar como

$$y_i = X\beta + \nu_i \text{ con } \nu_i = X\delta_i + e_i \sim N_R(0, XDX^T + \sigma^2 I), \quad (1)$$

$$\text{donde } D = \begin{pmatrix} \sigma_0^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_k^2 \end{pmatrix}_{(k+1) \times (k+1)}, \quad \text{es decir, suponemos que los efectos aleatorios}$$

son independientes.

La verosimilitud considerando  $\delta_i \perp \delta_j$ ,  $e_i \perp e_j$  y  $\delta_i \perp e_i$  para  $i, j = 1, 2, \dots, (k+1)$ , se puede expresar como:

$$L(\beta, D, \sigma^2) = \prod_{i=1}^M \frac{1}{(2\pi)^{R/2} |XDX^T + \sigma^2 I|^{1/2}} \times \exp\left\{-\frac{1}{2}(y_i - X\beta)^T (XDX^T + \sigma^2 I)^{-1} (y_i - X\beta)\right\},$$

luego la log-verosimilitud esta dada por

$$l(\beta, D, \sigma^2) = -\frac{1}{2} \sum_{i=1}^M (y_i - X\beta)^T (XDX^T + \sigma^2 I)^{-1} (y_i - X\beta) - \frac{MR}{2} \log(2\pi) - \frac{M}{2} \log(|XDX^T + \sigma^2 I|),$$

el término que contiene una suma se puede ver como

$$\begin{aligned}
& \sum_{i=1}^M (y_i - X\beta)^T (XDX^T + \sigma^2 I)^{-1} (y_i - X\beta) \\
&= \sum_{i=1}^M y_i^T V^{-1} y_i + \mu^T V^{-1} \mu - 2\mu^T V^{-1} y_i \\
&= M\mu^T V^{-1} \mu - 2\mu^T V^{-1} M\bar{y} + \sum_{i=1}^M y_i^T V^{-1} y_i,
\end{aligned}$$

donde  $\mu = X\beta$ , y  $V = XDX^T + \sigma^2 I$ , luego derivando respecto a  $\beta$  obtenemos:

$$\frac{\partial l(\beta, D, \sigma^2)}{\partial \beta} = 2MX^T V^{-1} X\beta - 2MX^T V^{-1} \bar{y} = 0,$$

despejando  $\beta$  obtenemos el estimador de los efectos fijos, dado por

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} \bar{y}, \text{ con } V = XDX^T + \sigma^2 I.$$

La log-verosimilitud perfil para  $\sigma^2$  y  $D$  esta dada por

$$\begin{aligned}
l(\hat{\beta}, D, \sigma^2) &= -\frac{1}{2} \sum_{i=1}^M (y_i - X\hat{\beta})(X^T DX + \sigma^2 I)^{-1} (y_i - X\hat{\beta}) \\
&\quad - \frac{MR}{2} \log(2\pi) - \frac{M}{2} \log(|XDX^T + \sigma^2 I|),
\end{aligned}$$

para obtener los estimadores de  $\sigma^2$  y  $D$  minimizar la siguiente función respecto a  $\sigma^2$  y  $D$

$$l(\hat{\beta}, D, \sigma^2) = \frac{M}{2} \log(|XDX^T + \sigma^2 I|) + \frac{1}{2} \sum_{i=1}^M (y_i - X\hat{\beta})(X^T DX + \sigma^2 I)^{-1} (y_i - X\hat{\beta}).$$

Para estimar el vector de efectos aleatorios, hacemos lo siguiente:

De (1) vemos que  $y_i \sim N_R(XB, V)$  y  $\delta_i \sim N_{(k+1)}(0, D)$  luego,

$$\begin{aligned} Cov(y_i, \delta_i) &= Cov(X\beta + X\delta_i + \epsilon_i, \delta_i) \\ &= \underbrace{Cov(X\beta, \delta_i)}_{=0} + X \underbrace{Var(\delta_i, \delta_i)}_{=D} + \underbrace{Cov(\epsilon_i, \delta_i)}_{=0} = XD. \end{aligned}$$

Entonces, 
$$\begin{pmatrix} y_i \\ \delta_i \end{pmatrix} \sim N \left( \begin{pmatrix} XB \\ 0 \end{pmatrix}, \begin{pmatrix} V & XD \\ DX^T & D \end{pmatrix} \right)$$

Recordemos que si tenemos 
$$\begin{pmatrix} Y \\ Z \end{pmatrix} \sim N_p \left( \begin{pmatrix} \mu_Y \\ \mu_Z \end{pmatrix}, \begin{pmatrix} \Sigma_Y & \Sigma_{YZ} \\ \Sigma_{ZY} & \Sigma_Z \end{pmatrix} \right)$$

Entonces,  $Z | Y \sim N(\mu_{Z|Y}, \Sigma_{Y|Z})$  con  $\mu_{Z|Y} = \mu_Z + \Sigma_{ZY}\Sigma_Y^{-1}(Y - \mu_Y)$ ,  $\Sigma_{Z|Y} = \Sigma_Z - \Sigma_{ZY}\Sigma_Y^{-1}\Sigma_{YZ}$ .

Por lo que

$$E(\delta_i | y_i) = 0 + DX^tV^{-1}(y_i - X\beta) = DX^tV^{-1}(y_i - X\beta)$$

Este último es llamado “best linear unbiased predictor (BLUP)” de  $\gamma$ , y

$$\hat{\delta}_i = DX^tV^{-1}(y_i - X\hat{\beta})$$

es llamado “empirical BLUP”.

Por lo que con esto hemos deducido las ecuaciones:

$$\hat{\beta} = (X^TV^{-1}X)^{-1}X^TV^{-1}\bar{y}, \text{ con } V = XDX^T + \sigma^2I \quad (3.3)$$

$$\hat{\delta}_i = DX^tV^{-1}(y_i - X\hat{\beta}) \quad (3.4)$$

Las cuales son utilizadas en el artículo, para formar la carta de control.

# Apéndice B

## Construcción de las funciones bases B-spline.

A continuación detallaremos la construcción de las 7 funciones bases B-spline definidas en la sección 2.4.1 y utilizadas en la sección 3.4. para la construcción de la carta de control funcional propuesta en esta tesis, los datos que utilizamos son los datos de la sección 3.2 sobre Motores de la compañía Peugeot, los datos se muestran en la Figura 3.1, debido a que es una parte esencial de dicha carta de control dedicaremos este apéndice para detallar su construcción. Para su construcción tenemos 3 nodos interiores (2625 3750 y 4875), y los valores RMP donde se hacen las mediciones del torque son:

En este ejemplo utilizamos una función B-spline de grado 3 con 3 nodos interiores, por lo que considerando el intercepto, tenemos en total 7 funciones bases B-spline.

Primero comenzamos con un conjunto de nodos:

$$t = (1500, 2626, 3750, 4875, 6000)$$

1500	2000	2500	2660	2800	2940	3500
4000	4500	5000	5225	5500	5774	6000

Tabla B.1: valores de RMP

luego los nodos aumentados, dado que tenemos grado 3, son,

$$(1500, 1500, 1500, 1500, 2626, 3750, 4875, 6000, 6000, 6000, 6000)$$

donde,

$$t_0 = t_1 = t_2 = t_3 = 1500, t_4 = 2626, t_5 = 3750, t_6 = 4875, t_7 = t_8 = t_9 = t_{10} = 6000.$$

De la ecuación 2.5 vemos que,

$$B_{i,3} = w_{i,3}B_{i,2} + (1 - w_{i+1,3})B_{i+1,2}, \quad (\text{B.1})$$

sustituyendo la ecuación 2.8 obtenemos,

$$\begin{aligned} B_{i,3} = & w_{i,3}w_{i,2}w_{i,1}X_i + \left[ w_{i,3}w_{i,2}(1 - w_{i+2,2}) + (1 - w_{i+1,2})w_{i+1,1} \right. \\ & \left. + (1 - w_{i+1,3})w_{i+1,2}w_{i+1,1} \right] X_{i+1} + \left[ w_{i,3}(1 - w_{i+1,2})(1 - w_{i+2,1}) \right. \\ & \left. + w_{i+1,2}(1 - w_{i+2,2})(1 - w_{i+1,3}) + (1 - w_{i+2,2})w_{i+2,1}(1 - w_{i+1,3}) \right] X_{i+2} \\ & + (1 - w_{i+1,3})(1 - w_{i+2,2})(1 - w_{i+3,1})X_{i+3}. \end{aligned}$$

Donde

$$w_{i,k} = w_{i,k}(x) = \begin{cases} \frac{x-t_i}{t_{i+k}-t_i} & t_i \neq t_{i+k} \\ 0 & \text{otro caso} \end{cases}, \quad X_i = X_i(x) = \begin{cases} 1 & \text{si } t_i \leq x < t_{i+1} \\ 0 & \text{otro caso} \end{cases}$$

Lo que sigue es obtener los 14 valores de las funciones bases  $B_{0,3}(x), B_{1,3}(x), \dots, B_{7,3}(x)$ , donde los valores posibles para  $x$  son los 14 valores del cuadro B.1, Entonces, hacemos

algunas evaluaciones para  $B_{0,3}(x)$ :

$$\begin{aligned}
 B_{0,3} = & w_{0,3}w_{0,2}w_{0,1}X_0 + \left[ w_{0,3}w_{0,2}(1 - w_{2,2}) + (1 - w_{1,2})w_{1,1} \right. \\
 & \left. + (1 - w_{1,3})w_{1,2}w_{1,1} \right] X_1 + \left[ w_{0,3}(1 - w_{1,2})(1 - w_{2,1}) \right. \\
 & \left. + w_{1,2}(1 - w_{2,2})(1 - w_{1,3}) + (1 - w_{2,2})w_{2,1}(1 - w_{1,3}) \right] X_2 \\
 & + (1 - w_{1,3})(1 - w_{2,2})(1 - w_{3,1})X_3.
 \end{aligned}$$

de lo cual vemos que los únicos valores  $x$  para los que la función  $B_{0,3}$  es no cero es para los valores 1500, 2000 y 2500, y solo el término

$$(1 - w_{1,3})(1 - w_{2,2})(1 - w_{3,1})X_3$$

es diferente de cero en estos tres valores, obteniendo como resultado

$$B_{0,3}(1500) = 1, B_{0,3}(2000) = 0.1714, B_{0,3}(2500) = 0.0013$$

Siguiendo el procedimiento de esta manera, obtenemos la matriz  $(14 \times 7)$  mostrada en el cuadro B.2, esta matriz se utiliza para estimar los coeficientes de la función  $B - spline$  de la sección 3.4.



$x$	$B_{0,3}(x)$	$B_{1,3}(x)$	$B_{2,3}(x)$	$B_{3,3}(x)$	$B_{4,3}(x)$	$B_{5,3}(x)$	$B_{6,3}(x)$
1500	1	0	0	0	0	0	0
2000	0.1714	0.5980	0.2158	0.014	0	0	0
2500	0.013	0.3401	0.5413	0.117	0	0	0
2660	0	0.2273	0.5899	0.182	0.000005	0	0
2800	0	0.1505	0.5941	0.254	0.0006	0	0
2940	0	0.0933	0.5681	0.3348	0.0036	0	0
3500	0	0.0027	0.2960	0.62277	0.0784	0	0
4000	0	0	0.0784	0.62277	0.296	0.0027	0
4500	0	0	0.0061	0.3703	0.549	0.074	0
5000	0	0	0	0.117	0.541	0.3401	0.0013
5225	0	0	0	0.0544	0.412	0.5032	0.0301
5500	0	0	0	0.0146	0.215	0.598	0.1714
5774	0	0	0	0.00133	0.052	0.434	0.512
6000	0	0	0	0	0	0	1

Tabla B.2: Matriz B-spline

# Apéndice C

## Cálculo de ARL con perturbación.

En el presente Apéndice se presentarán los resultados de ARL tanto para la carta de control propuesta en esta tesis como para la carta de control propuesta en el artículo *Amiri*[1], estos ARL son considerando una perturbación en la media como un vector  $P_i = aS_i$ , con  $i = 1, 2, 8, 9, 10, 11, 12, 13, 14$ ,  $a$  es un número real y  $S_i$  es la desviación estándar de 26 motores en los valores de RMP

(1500, 2000, 4000, 4500, 5000, 5225, 5500, 5775, 6000)

que corresponden de izquierda a derecha a los valores  $i = 1, 2, 8, 9, 10, 11, 12, 13, 14$ . Notemos que el vector  $P_i$  tiene todas sus entradas iguales a cero excepto la entrada  $i$ .

Primero presentaremos los Resultados de ARL para la carta de control propuesta en el artículo en estudio:

$P_1$	
a	ARL
-0.1	19.5
-0.5	14.8
-0.8	10.2
-1	7.8
-2	2.4
-3	1.2

$P_1$	
a	ARL
0.1	19.6
0.5	14.9
0.8	10.2
1	7.8
2	2.4
3	1.2

$P_2$	
a	ARL
-0.1	19.9
-0.5	19.1
-0.8	18.1
-1	17.1
-2	11.9
-3	7.5

$P_2$	
a	ARL
0.1	20
0.5	19.2
0.8	18.0
1	17.2
2	11.9
3	7.5

$P_8$	
a	ARL
-0.1	19.8
-0.5	19.1
-0.8	18.4
-1	17.8
-2	13.1
-3	8.9

$P_8$	
a	ARL
0.1	19.9
0.5	19.2
0.8	18.6
1	17.8
2	13.2
3	9.1

$P_9$	
a	ARL
-0.1	20
-0.5	19.5
-0.8	18.8
-1	18.4
-2	14.9
-3	11.1

$P_9$	
a	ARL
0.1	19.9
0.5	19.6
0.8	18.7
1	18.1
2	14.4
3	11.0

$P_{10}$	
a	ARL
-0.1	20
-0.5	19.6
-0.8	19.4
-1	19.3
-2	17.5
-3	15.1

$P_{10}$	
a	ARL
0.1	20
0.5	19.8
0.8	19.5
1	19.3
2	17.6
3	15.2

$P_{11}$	
a	ARL
-0.1	20
-0.5	19.8
-0.8	19.4
-1	19.2
-2	17.7
-3	15.3

$P_{11}$	
a	ARL
0.1	20
0.5	19.8
0.8	19.5
1	19.1
2	17.8
3	15.1

$P_{12}$	
a	ARL
-0.1	20
-0.5	19.6
-0.8	19.1
-1	18.8
-2	17.0
-3	14.5

$P_{12}$	
a	ARL
0.1	20
0.5	19.7
0.8	19.4
1	18.8
2	17.2
3	14.4

$P_{13}$	
a	ARL
-0.1	19.8
-0.5	19.3
-0.8	18.1
-1	17.3
-2	12.6
-3	8.2

$P_{13}$	
a	ARL
0.1	20
0.5	19.4
0.8	18.3
1	17.7
2	12.5
3	8.3

$P_{14}$	
a	ARL
-0.1	20
-0.5	19.2
-0.8	18.0
-1	16.9
-2	11.1
-3	7.2

$P_{14}$	
a	ARL
0.1	20
0.5	19.1
0.8	18.1
1	16.8
2	11.3
3	7.0

A continuación se presentan los resultados ARL para la carta de control propuesta en esta Tesis:

$P_1$	
a	ARL
-0.1	19.8
-0.5	13.9
-0.8	9.4
-1	6.8
-2	1.8
-3	1

$P_1$	
a	ARL
0.1	19.9
0.5	14.2
0.8	9.2
1	6.7
2	1.8
3	1

$P_2$	
a	ARL
-0.1	19.6
-0.5	13.2
-0.8	8.1
-1	5.6
-2	1.5
-3	1

$P_2$	
a	ARL
0.1	19.5
0.5	12.9
0.8	8.0
1	5.7
2	1.5
3	1

$P_8$	
a	ARL
-0.1	19.6
-0.5	12.2
-0.8	6.8
-1	4.6
-2	1.2
-3	1

$P_8$	
a	ARL
0.1	19.5
0.5	12.1
0.8	6.9
1	4.6
2	1.3
3	1

$P_9$	
a	ARL
-0.1	19.3
-0.5	8.3
-0.8	3.8
-1	2.4
-2	1.1
-3	1

$P_9$	
a	ARL
0.1	19.4
0.5	8.3
0.8	3.8
1	2.4
2	1.2
3	1

$P_{10}$	
a	ARL
-0.1	19.8
-0.5	14.8
-0.8	10.8
-1	7.8
-2	2.2
-3	1.1

$P_{10}$	
a	ARL
0.1	19.6
0.5	15.2
0.8	10.4
1	7.7
2	2.1
3	1.2

$P_{11}$	
a	ARL
-0.1	19.6
-0.5	15.1
-0.8	10.1
-1	7.6
-2	2.1
-3	1.3

$P_{11}$	
a	ARL
0.1	19.7
0.5	14.8
0.8	10.2
1	7.5
2	2.1
3	1.1

$P_{12}$	
a	ARL
-0.1	19.5
-0.5	9.9
-0.8	4.9
-1	3.2
-2	1.1
-3	1

$P_{12}$	
a	ARL
0.1	19.7
0.5	10.1
0.8	4.9
1	3.1
2	1.1
3	1

$P_{13}$	
a	ARL
-0.1	19.6
-0.5	11.5
-0.8	6.2
-1	4.3
-2	1.1
-3	1

$P_{13}$	
a	ARL
0.1	19.6
0.5	11.6
0.8	6.4
1	4.3
2	1.2
3	1

$P_{14}$	
a	ARL
-0.1	19.2
-0.5	7.2
-0.8	3.1
-1	1.9
-2	1
-3	1

$P_{14}$	
a	ARL
0.1	19.1
0.5	7.1
0.8	3.0
1	1.9
2	1
3	1