CIMAT

Centro de Investigación en Matemáticas, A.C.

CARTAS DE CONTROL PARA DATOS FUNCIONALES

T E S I S

que para obtener el grado de

Maestro en Ciencias con Especialidad en Probabilidad y Estadística

p r e s e n t a

Diego Rivera García

Dr. Enrique Raúl Villa Diharce
Director de Tesis

Dr. Joaquín Ortega Sánchez
Codirector de Tesis

Guanajuato, Gto. Agosto de 2011

Índice general

Índice de figuras

Índice de cuadros

Agradecimientos

Un agradecimiento muy especial a CONACYT por el apoyo recibido para poder cursar esta maestría. De igual forma quisiera agradecer a CIMAT AC, y a los profesores que me acompañaron y orientaron durante esta travesia.

A mi familia, gracias por su apoyo incondicional, si él no hubiera podido estar donde estoy ahora. Mil gracias por todo, jamás podré pagar todo lo que han hecho por mi.

También quiero agradecer al Dr. Enrique Villa Diharce y al Dr. Joaquín Ortega Sánchez por brindarme el apoyo para la realización de este trabajo.

Al Dr. Miguel Nakamura Savoy gracias por sus valiosos consejos y enseñanzas.

A mis amigos, en especial a Joel Iglesias, Mario Santana, Gustavo Cano, Harol Moreno, Abelardo Motesinos, Norma Selomit Ramírez, Lina Vargas gracias por su apoyo y amistad.

Un agradecimiento muy especial a Leticia Escobar por ser uno de los motores que me impulsaron a crecer como estudiante y sobre todo como persona.

A Dios por darme paciencia y fuerza para la realización de este trabajo.

Prefacio

La creciente competitividad de los mercados ha conducido al mundo empresarial a buscar nuevas y mejores formas de ofrecer productos de mejor calidad, que les permitan convertirse en líderes dentro del sector que manejen.

El monitoreo de los procesos industriales ha sido de vital importancia en la mejora de productos, para ello diversas estrategias como el control estadístico de procesos han desarrollado herramientas que permitan hacer el monitoreo de una mejor manera. Las cartas de control son una herramienta gráfica que permite detectar anomalías ocurridas durante el proceso de producción. Sin embargo, dado que la característica de calidad puede estar en función de una o más covariables es necesario el desarrollo de nuevas técnicas que permitan capturar estas nuevas fuentes de variación.

El desarrollo de la estadística de datos funcionales ha logrado un cambio en el análisis de datos, ya que permite el análisis conjunto de curvas y no simplemente datos como se hace usualmente.

En el presente trabajo se hace una propuesta de cartas de control que permite monitorear características de calidad cuando estas estén en función de una o más covariables, las cartas de control para datos funcionales. Este trabajo se desarrolla de la siguiente forma: En el primer capítulo se hace una introducción a las cartas de control, en el capítulo 2 se realiza una breve introducción al análisis de datos funcionales, en el capítulo 3 se realiza un caso de estudio para un conjunto de datos dado y finalmente el el capítulo 4 se muestran las conclusiones obtenidas a lo largo del trabajo.

Capítulo 1

Cartas de control

1.1. Introducción

Ante la creciente competitividad en los mercados y la exigencias por parte de los consumidores, las industrias han buscado nuevas formas de monitorear sus procesos de producción, a fin de mejorar la calidad de sus productos.

En el área de estadística existe un conjunto de técnicas para el mejoramiento y control de las lineas de producción industrial, conocida como control estadístico de procesos (CEP). Dentro del CEP existe una herramienta gráfica de gran utilidad usada para el monitoreo de procesos, las cartas de control. En el presente capítulo se dará una breve introducción a las cartas de control para monitorear procesos industriales.

1.2. Cartas de control

Una carta de control es una herramienta estadística empleada para el estudio y control de procesos a través del tiempo. El objetivo de las cartas de control es el observar y analizar mediante el uso de datos estadísticos la variabilidad del proceso de interés a través del tiempo, (Gutiérrez y De la Vara, 2004 [9]).

Mediante el uso de las cartas de control se pretende identificar las principales fuentes de variación del proceso, las cuales se identifican como:

- **Variabilidad debida a causas comunes:** Variabilidad que aparece de manera

natural en el proceso debida al azar e inherente a la calidad. Nada se puede hacer sobre este tipo de variabilidad.

- **Variabilidad debida a causas especiales:** Variabilidad originada por circunstancias o situaciones especiales ajenas al proceso. Este tipo de variabilidad a menudo puede ser identificada y eliminada del proceso

La idea básica de una carta de control es que, mediante el cálculo de limites de control, podamos observar dónde varía el proceso a través del tiempo, graficando un estadístico, denotado por W, el cual mide la característica de interés en el proceso. Los elementos para construir una carta de control son:

- **Linea central (LC):** Esta linea representa el promedio de los valores de W.
- **Los límites de control inferior (LCI) y superior (LCS):** Estos límites definen el rango de variación del proceso, de tal manera que al estar el proceso bajo control estadístico, haya una alta probabilidad de que los valores de W se encuentren dentro de los límites de control.

Cabe mencionar que estos límites de control no corresponden a los límites de especificación, tolerancias o deseos del proceso. Estos son calculados a partir de la variación de los datos que se representan en la carta. La idea de su cálculo está en establecer los límites de forma que sea cubierto el mayor porcentaje de la variabilidad del proceso. Sin embargo, la elección de estos límites debe ser realizada con cuidado, ya que si se desea cubrir un alto porcentaje de variabilidad, los límites serán muy amplios. Esto dificultaría la detección de cambios; en cambio si el porcentaje es pequeño, los límites serán muy estrechos causando muchas señales en falso.

Así, si algún valor de W cae fuera de los límites de control esto indicará un evento inusual en el proceso. En la figura 1.2.1 se presentan los elementos de la carta de control.

La distinción entre los tipos de variación ayudará a caracterizar el funcionamiento del proceso, con el fin de decidir las acciones de control y mejora, para mantener el proceso bajo control estadístico.

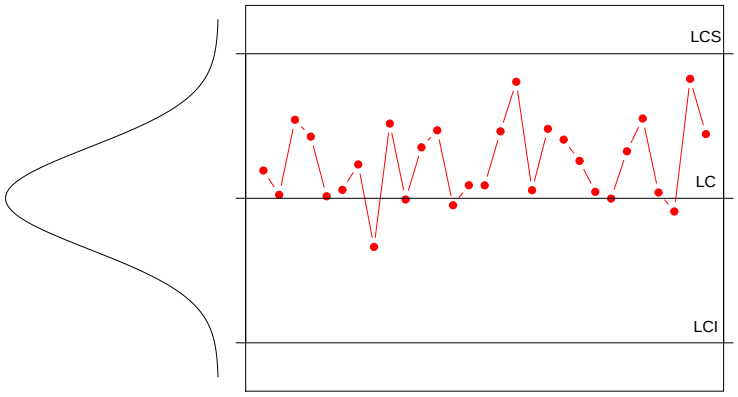
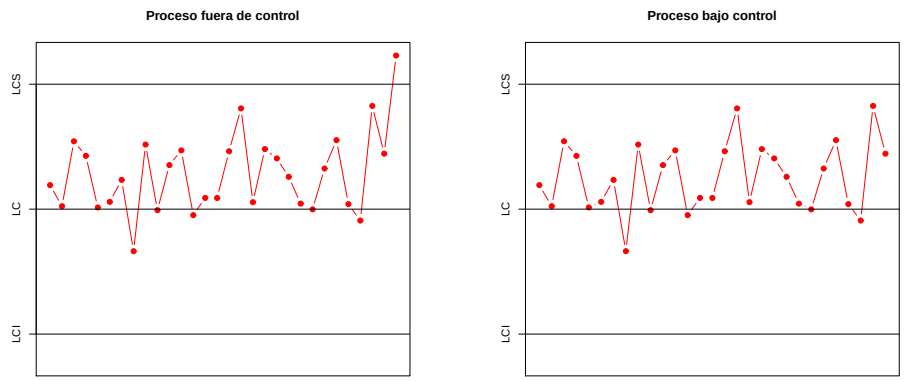


Figura 1.2.1: Elementos de una carta de control

Decimos que un proceso está bajo control estadístico si la variación del mismo se mantiene dentro de un rango preestablecido y su distribución no cambia. De manera análoga decimos que el proceso está fuera de control estadístico si su variación se debe a una o varias causas específicas y por lo tanto su distribución cambia. Esto se puede observar en la gráfica 1.2.2.

La manera más frecuente de encontrar estos límites es a partir de la relación entre la media y la desviación estándar de W. En el caso de que W siga una distribución normal con media μ_w y desviación estándar σ_w, se tiene que los límites están dados por $\mu_w - 3\sigma_w$ y $\mu_w + 3\sigma_w$, donde bajo control estadístico se ubica el 99.73 % de los posibles valores de W.

En el caso de tener una distribución diferente a la normal, y se tenga una distribución unimodal y con forma parecida a la normal, entonces, se aplica la regla empírica o la



a) b)

Figura 1.2.2: Control estadístico: a)Proceso fuera de control, b) Proceso bajo control

extensión del teorema de Chebyshev, la cual esta dada por

Teorema 1.2.1 (Desigualdad de Chebyshev) *Sea* $\mu = E(X)$ *y* $\sigma^2 = Var(X)$, *entonces* $P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$.

A partir de este resultado se obtiene una relación entre $\bar{X}$ y S que establece el porcentaje mínimo de datos que caen entre los límites $\bar{X} - kS$ y $\bar{X} + kS$.

Las cartas de control que siguen esta metodología fueron propuestas por el doctor Walter A. Shewhart (1924), y sus límites de control están dados por

$$LCI = \mu_w - 3\sigma_w, \quad LC = \mu_w, \quad LCS = \mu_w + 3\sigma_w. \tag{1.2.1}$$

Este tipo de cartas son conocidas como cartas tipo Shewhart. Con ellas y bajo condiciones de control estadístico se tendrá alta probabilidad de que los valores de W caigan dentro de los límites definidos en (1.2.1). En el caso de que se tenga una distribución normal esta probabilidad será 0.9973.

A menudo los parámetros de la distribución de W son desconocidos, por los cual deben ser estimados. Para ello se toma un conjunto de datos históricos, con los cuales se estiman

los parámetros de la distribución de W. Una vez estimados estos valores se realiza un proceso de depuración de la información hasta considerar el proceso bajo control estadístico. Este proceso se conoce como **Fase I** de la construcción de la carta de control.

Durante la **Fase II** el proceso se muestrea en línea y los valores de W obtenidos son graficados en la carta de control. Generalmente se muestrea un número m de subgrupos, que oscilan entre 20 y 50 de tamaño 5.

1.3. Riesgos del muestreo

Una vez construida la carta de control, mediante algún criterio se juzga si el proceso se encuentra o no bajo control. Por aleatoriedad del proceso se corre el riesgo de equivocarse y de que el proceso envíe señales falsas de estar fuera de control.

Dado lo anterior se tienen dos tipos de riesgo que se pueden encontrar:

- **Riesgo tipo I:** Es el riesgo de que una muestra conduzca a tomar una decisión cuando no haya ocurrido un cambio en el proceso.
  Riesgo tipo II: El riesgo de que una muestra se encuentre dentro de los límites de control a pesar de que haya ocurrido un cambio en el proceso.

De acuerdo con esto, nótese que a partir de las cartas de control es posible generar el siguiente juego de hipótesis:

$$H_0 : \text{El proceso está bajo control v.s. } H_1 : \text{El proceso está fuera de control} \tag{1.3.1}$$

con lo cual se tiene una relación entre las cartas de control y las pruebas de hipótesis.

1.4. Tipos de cartas de control

Existen dos grupos generales de cartas de control: para variables y para atributos. Las cartas de control para variables, se aplican al monitoreo de características de calidad del tipo continuo, las cuales requieren de un instrumento de medición (Peso, volumen, voltaje, etc.).

Las cartas de control para variables más usuales son:

- $\bar{X}$(Promedios)
- R (Rangos)
- S (Desviación estándar)
- X (Medidas individuales)
- T^2 (Multivariadas)

Las cartas de control para atributos se aplican cuando el producto o el proceso no es medido y simplemente es juzgado como conforme o no conforme, dependiendo del número de defectos o no conformidades que tiene. Las principales cartas de control para atributos 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 las cartas mencionadas, existe una gran variedad de cartas de control, con las cuales se pretende detectar más rápido un cambio en el proceso y reducir la frecuencia de falsas alarmas. Entre estas cartas las más conocidas son las cartas EWMA y CUSUM.

En el presente trabajo se hablará principalmente de las cartas para medidas individuales, para mayor referencia acerca de las cartas de control consultar del libro de Gutiérrez y De la Vara et al, (2004).

1.4.1. Carta de medidas individuales

Este tipo de cartas se usan para monitorear variables del tipo continuo en el caso que se trabaje con procesos lentos o costosos, en los cuales para obtener una muestra de la producción se requiere de periodos relativamente largos.

La determinación de los límites de control en este caso no difiere del caso de las otras cartas de control, esto es, mediante la estimación de la media y desviación estándar del estadístico W que se esté usando. En este caso es directamente la observación individual obtenida del proceso, X. Entonces los límites quedan determinados por

$$LCI = \mu_X - 3\sigma_X, \quad LC = \mu_X, \quad LCS = \mu_X + 3\sigma_X, \tag{1.4.1}$$

donde μ_X y σ_X son la media y desviación estándar del proceso, respectivamente.
La estimación de estos parámetros procede de la siguiente manera:

$$\mu_X = \bar{X} \;\; y \;\; \sigma_X = \frac{\bar{R}}{d_2}, \tag{1.4.2}$$

donde $\bar{X}$ es la media de las observaciones y $\bar{R}$ corresponde a la media de los rangos móviles de orden dos, esto es, el rango entre dos observaciones sucesivas del proceso. Mientras la constante d_2 está dada por

$$d_2 = E(R \mid X_i \sim N(0,1)) \tag{1.4.3}$$

donde $R = \text{máx}\{X_i\} - \text{mín}\{X_i\}$. Estas constantes usualmente vienen tabulados en los libros de control estadístico de procesos. La constante d_2 depende del tamaño del subgrupo y define la media del rango relativo $q = \dfrac{R}{\sigma}$; en este caso se tiene que $d_2 = 1.128$.

De acuerdo con lo anterior se tiene que los limites de control para medidas individuales

están dados por

$$LCI = \bar{X} - 3\frac{\bar{R}}{1.128}, \quad LC = \bar{X}, \quad LCS = \bar{X} + 3\frac{\bar{R}}{1.128}. \tag{1.4.4}$$

De manera gráfica esto se puede observar en la figura 1.4.1

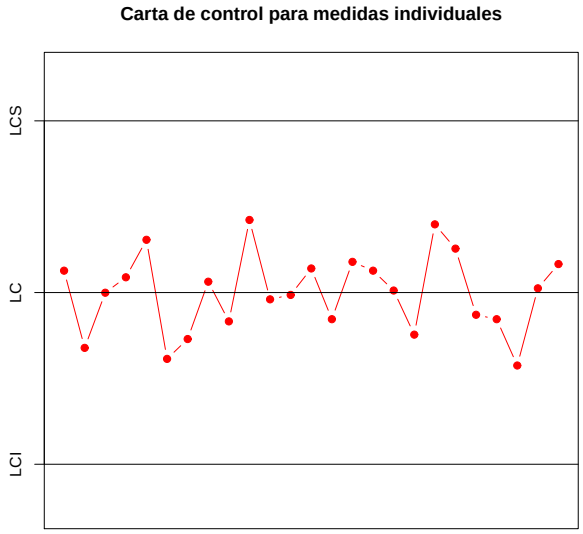


Figura 1.4.1: Carta de control para medidas individuales

Cuando se trabaja con este tipo de carta de control una manera de monitorear la variabilidad del proceso es usar una carta de rangos móviles de orden dos: Esto como complemento a la carta de control para medidas individuales, cuyos límites de control están dados por

$$LCI = 0, \quad LC = \bar{R}, \quad LCS = \bar{R} + 3\sigma_R. \tag{1.4.5}$$

Aquí $\sigma_R = d_3\sigma$ es una constante que depende del tamaño del subgrupo y corresponde a la desviación estándar del rango relativo, $q = \dfrac{R}{\sigma}$, en este caso $d_3 = 0.853$. Esta constante d_3 viene tabulada en la mayoría de los libros de control estadístico de procesos.

De acuerdo a lo anterior se tiene que

$$LCI = 0, \quad LC = \bar{R}, \quad LCS = \bar{R} + 3\frac{d_3}{d_2}\sigma. \tag{1.4.6}$$

Esto se puede apreciar en la figura 1.4.2

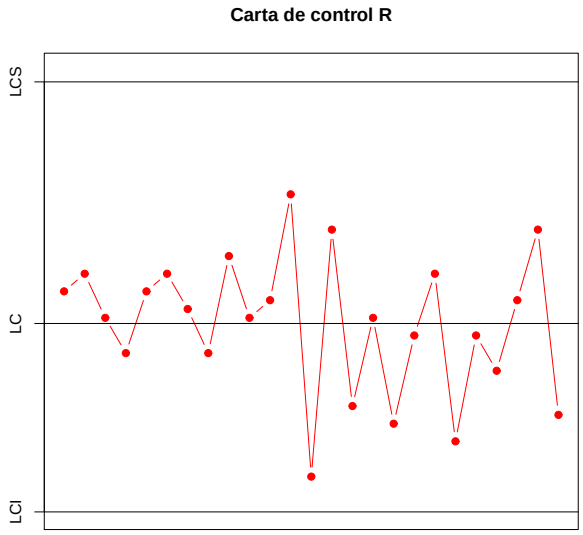


Figura 1.4.2: Carta de control para rangos móviles

En recientes estudios se ha demostrado que las cartas de control para medidas individuales son bastante robustas tanto para detectar cambios en la media como en la dispersión del proceso.

Cabe mencionar que las cartas de control para medidas individuales son una gran alternativa para el monitoreo de procesos lentos, sin embargo, el desvío de la distribución de las observaciones de la normal puede afectar los criterios de interpretación de la carta.

1.5. Longitud promedio de Corrida (ARL)

Una manera de comparar la eficiencia de las cartas de control es a partir de la longitud promedio de corrida (ARL por sus siglas en inglés).

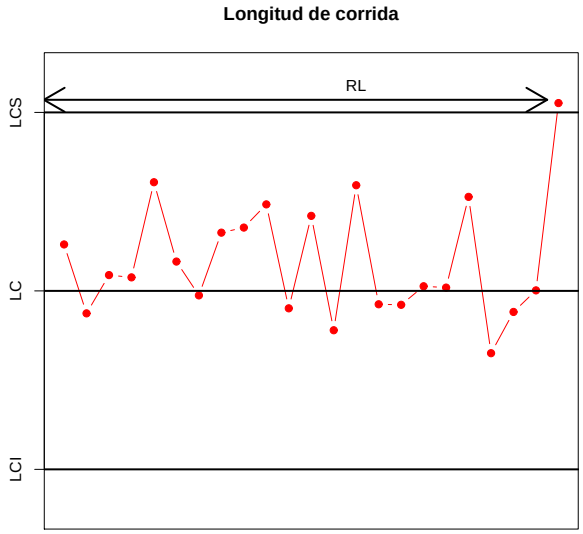


Figura 1.5.1: Longitud de corrida

El ARL se define como el número promedio de puntos antes de que la carta de control dé una señal de fuera de control, sin que haya ocurrido algún cambio en el proceso. Es decir, esta señal de fuera de control se debe sólo al azar del proceso.

Si definimos RL como el número de puntos antes de obtener una señal de fuera de control, ver figura 1.5.1, entonces RL tendrá distribución geométrica de parámetro p, es decir

$$P(RL = x) = (1-p)^{x-1}p, \quad x = 1, 2, ... \tag{1.5.1}$$

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

$$p = P(X > LCS \ \ ó \ \ X < LCI). \tag{1.5.2}$$

De acuerdo con esto se sigue que $E(RL) = \frac{1}{p}$ y $var(RL) = \frac{(1-p)}{p^2}$, de donde se obtiene que bajo condiciones de control estadístico, se tiene

$$ARL = \frac{1}{p} \tag{1.5.3}$$

Por ejemplo bajo distribución normal y usando un esquema de cartas tipo Shewhart se tiene que $p = 0.0027$ y $ARL = 370.4$, es decir cada 370 puntos la carta de control dará una señal de fuera de control a pesar de que no hayan ocurrido cambios en el proceso. De esta manera mediante el ARL es posible la comparación de cartas, siendo la mejor aquella que envíe menos señales de fuera de control falsas, es decir, aquella con el ARL mayor.

Capítulo 2

Análisis de datos funcionales

2.1. Introducción

Durante los últimos años los avances tecnológicos se han visto en aumento, provocando un gran impacto en diversas áreas de investigación. Éste tipo de impacto ha mejorado los instrumentos de medición, haciéndolos más rápidos y precisos.

En el área de estadística estos cambios no han sido menores. Actualmente se ha empezado a trabajar con grandes bases de datos que corresponden a observaciones de variables aleatorias tomadas sobre intervalos de tiempo, donde el resultado de dicha medición es una curva que representa a la muestra concreta que ha sido evaluada. Este tipo de datos son llamados datos funcionales.

Ante estos nuevos retos surge como respuesta la estadística de datos funcionales, la cual define a un dato como una función en un intervalo de tiempo. En el presente capítulo se dará una breve introducción acerca de la estadística de datos funcionales.

2.2. Datos funcionales

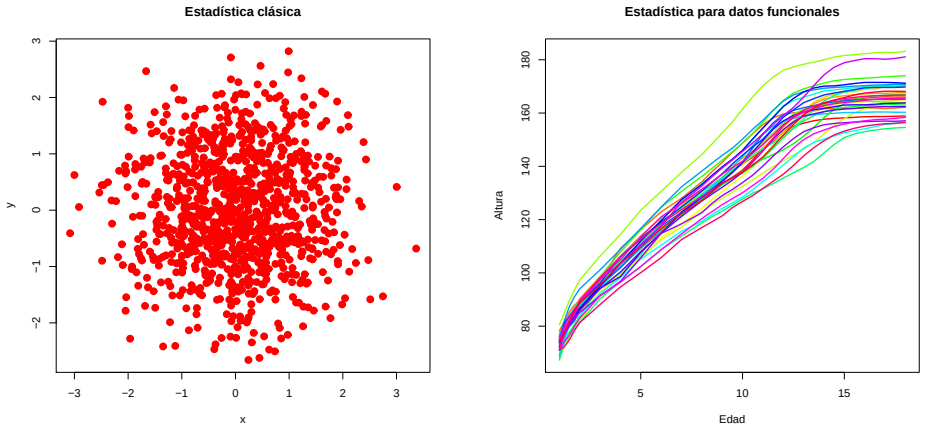
El análisis de datos funcionales es una metodología relativamente reciente impulsada principalmente por los trabajos de Ramsay y Silverman (1997) [12].

En esencia los problemas a los que se enfrenta la estadística de datos funcionales son

los mismos a los que se enfrenta la estadística clásica, cuyos objetivos se pueden listar como:

- Lograr una representación que capture las características del conjunto de datos.
- Estudiar fuentes importantes de patrones y variación entre los datos.
- Explicar la variabilidad de una respuesta mediante el uso de variables independientes.
- Contraste, validación y predicción.
- Métodos de clasificación de un conjunto de datos respecto a alguna característica.

Como se mencionó anteriormente, ahora tenemos el caso en el que el conjunto de observaciones para el análisis es un conjunto de funciones y no de datos como usualmente se hace, en la gráfica 2.2.1 se puede observar las diferencias entre ambos enfoques.



a) b)

Figura 2.2.1: Tipos de análisis: a) Estadística clásica, b) Estadística para datos funcionales

Siguiendo la definición de Ferraty y Vieu (2006) [7] se tiene que la definición de dato funcional esta dada por

Definición 2.2.1 *Una variable aleatoria X se dice que es una variable funcional si toma valores en un espacio funcional ξ (Espacio normado o semi-normado completo).*

De acuerdo con esto decimos que, *un conjunto de datos funcionales* $\{X_1, ..., X_n\}$ *es la observación de n variables funcionales* $X_1, ..., X_n$ *idénticamente distribuidas.*

Esto puede ser aplicado a muchos tipos de espacios. En particular, $\mathbb{R}^p$ con las métricas usuales será un espacio funcional y por lo tanto puede deducirse que toda técnica que se desarrolle para datos funcionales puede ser aplicada con ciertas garantías en el caso multivariado. El reverso generalmente no es cierto.

A manera de ejemplo considere los datos provenientes del libro de Ramsay y Silverman (2005)[12], los cuales muestran las curvas de crecimiento de 10 niñas con mediciones en 31 puntos entre 1 y 18 años de edad, dichas mediciones no son igualmente espaciadas. Este estudio corresponde a un estudio de crecimiento realizado en Berkeley, en la gráfica 2.2.2 se observan los datos correspondientes a este estudio.

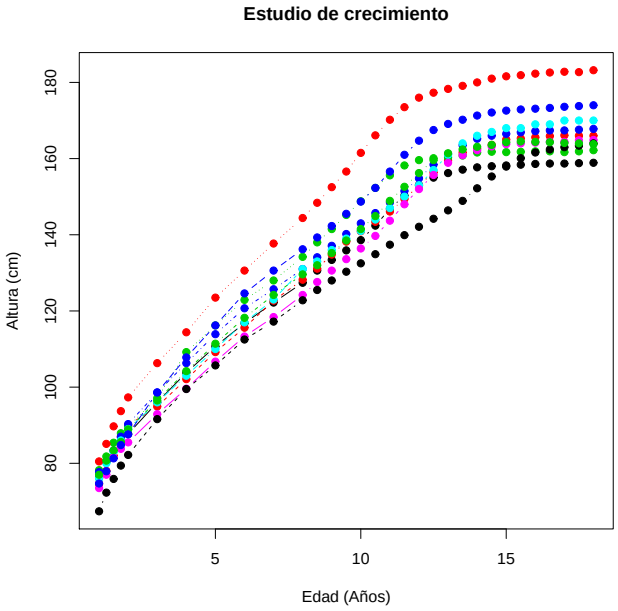


Figura 2.2.2: Estudio de crecimiento

Como puede observarse del ejemplo acerca del estudio de crecimiento, los datos llegan de manera discreta, por lo cual la primera tarea en el análisis de datos funcionales es convertir estos datos a funciones $X(t)$ para algún argumento deseado t. Si se asume que

las observaciónes discretas no contienen errores, entonces el proceso de conversión es una simple interpolación, mientras que si se cuenta con algún error observacional que debe ser removido, este procesos involucra un suavizamiento.

2.3. Bases

Como se ha visto, los datos funcionales corresponden a observaciones a través del tiempo de una variable aleatoria. Estos valores recolectados corresponden a discretizaciones de curvas a las cuales llamamos datos funcionales.

De acuerdo con lo anterior, la representación de un dato funcional mediante una base ortonormal proporcionará una ventaja tanto teórica como practica, sirviendo como puente entre la discretización del dato funcional y su verdadera forma funcional.

Definición 2.3.1 (Bases) *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 quiera, mediante una combinación lineal de* K *de ellas con* K *suficientemente grande. De esta forma, la observación funcional puede aproximarse como* $X(t) \approx \sum_{k=1}^{K} c_k\phi_k(t)$.

De manera matricial esto puede escribirse como:

$$X(t) = \mathbf{c}^T\mathbf{\Phi} \tag{2.3.1}$$

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

La elección de la base más adecuada y de K es crucial, sin embargo, no existe una regla que permita esta elección de manera óptima y universal. La elección de la base dependerá de la naturaleza del problema, por ejemplo si se esta trabajando con datos periódicos, es mejor el uso de una base de Fourier, mientras que si los datos son no periódicos puede usarse una base de Splines o bien una base Wavelet.

2.3.1. Splines

En el presente trabajo se usará una base de Splines, por lo que se dará una breve introducción acerca de ellos

Definición 2.3.2 (Splines) *La función* $\phi : [a,b] \longrightarrow \mathbb{R}$ *es un spline de grado* p *con nodos en* $t_1, ..., t_k$ *si se verifica lo siguiente:*

1. $a < t_1 < \cdots < t_k < b$ *(denotemos* $t_0 = a, t_{k+1} = b$*)*
2. *En cada intervalo* $[t_j, t_{j+1}]$, $j = 0, ..., k$, ϕ *es un polinomio de grado* p *o inferior.*
3. *La función* ϕ *tiene* $(p-1)$ *derivadas en* $[a,b]$*, (es decir, los polinomios que definen la función* ϕ *en los intervalos* $[t_{j-1}, t_j]$ *y* $[t_j, t_{j+1}]$ *enlazan bien en* t_j*)*

Una vez definida la función de splines, definamos una base de splines, dentro de las cuales resaltan las B-splines las cuales cumplen las siguientes propiedades

- Cada función base $\phi_k(t)$ es una función spline de orden p y con τ nodos.
- Dado que un múltiplo de splines es spline y la suma de splines es también spline, cualquier combinación lineal de $\phi_k(t)$ será también un spline.
- Cualquier spline de orden p y con τ nodos puede ser expresado como una combinación lineal de funciones base $\phi_k(t)$.

Este tipo de bases desarrolladas por de Boor (2001) son las más populares y estan disponibles en una gran cantidad de Software, incluyendo R. En la gráfica 2.3.1 presentamos trece funciones B-spline para un spline de orden tres definidas por nueve nodos equiespaciados.

Un caso especial resulta cuando se hace uso de splines suaviados, los cuales agregan un termino que penaliza la falta de suavidad por parte de la curva.

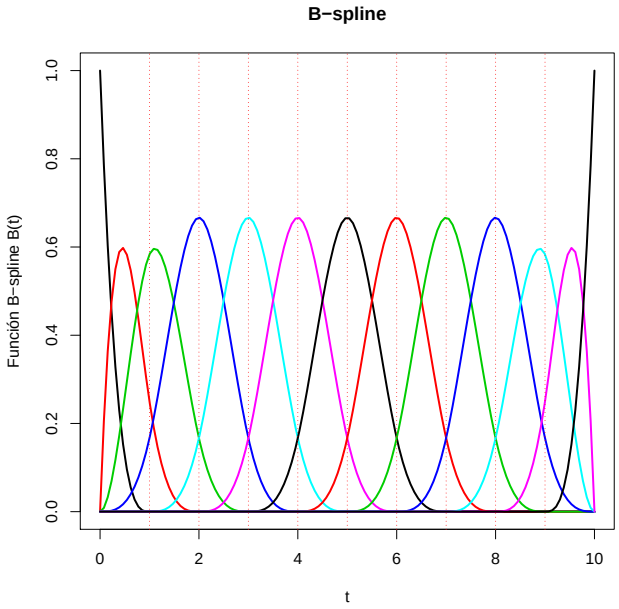


Figura 2.3.1: B-spline

2.3.2. Splines suavizados

La metodología de spline suavizado ajusta la curva $x(t)$ proveniente de las observaciones $y_j = x(t_j) + \epsilon_j$ tomando en cuenta los posibles conflictos que pueden surgir en la estimación. Por un lado se desea asegurar un buen ajuste de la curva a los datos. Por otro lado no se desea que el ajuste sea tan bueno si este produce una curva excesivamente rugosa o muy variable localmente.

El cuadrado de la segunda derivada, $[D^2x(t)]^2$, al tiempo t es llamado la curvatura de la función al tiempo t. Entonces una medida de la rugosidad de una función es la integral del cuadrado de la segunda derivada,

$$PEN_2(x) = \int [D^2x(s)]^2 ds. \tag{2.3.2}$$

En funciones altamente variables puede esperarse que los valores de $PEN_2(x)$ sean altos, ya que su segunda derivada será grande al menos sobre un rango de interés.

Dado que en varias aplicaciones de datos funcionales, las derivadas son de gran interés, la expresión (2.3.2) puede no ser adecuada, puesto que sólo controla la curvatura de la función original. Por tanto, si queremos estudiar la derivada de orden m, se debe penalizar las derivadas de orden $m+2$ para controlar la curvatura de derivadas de alto orden (Ramsay y Silverman, 2005[12]).

De acuerdo con esto, mediante la generalización (2.3.2) que permite una derivada, $D^m x$, de orden arbitrario se obtiene la siguiente penalización

$$PEN_m(x) = \int [D^m x(s)]^2 ds. \tag{2.3.3}$$

Por ejemplo para estimar la aceleración es mejor usar.

$$PEN_4(x) = \int [D^4 x(s)]^2 ds, \tag{2.3.4}$$

puesto que con esto controlamos la curvatura de $D^2 x$.

2.3.3. Estimación de parámetros

Recuerde que los datos con los que se trabaja en datos funcionales provienen de discretizaciones de funciones. Entonces el objetivo es ajustar las observaciones discretas y_j, $j = 1, 2, ..., n$ usando el modelo $y_j = x(t_j) + \epsilon_j$, donde $x(t)$ es aproximado mediante una base como se mencionó en la sección 2.3.1.

Un simple suavizamiento se obtiene determinando los coeficientes c_k por mínimos cuadrados ordinarios

$$SMSSE(\mathbf{y} \mid \mathbf{c}) = \sum_{j=1}^{n} [y_j - \sum_{k=1}^{K} c_k \phi_k(t_j)]^2. \tag{2.3.5}$$

De manera matricial se tiene

$$SMSSE(\mathbf{y} \mid \mathbf{c}) = (\mathbf{y} - \mathbf{\Phi c})^T (\mathbf{y} - \mathbf{\Phi c}), \tag{2.3.6}$$

entonces derivando (2.3.6) con respecto a $\mathbf{c}$ e igualando a cero se obtiene la siguiente ecuación

$$2\mathbf{\Phi}\mathbf{\Phi}^T\mathbf{c} - 2\mathbf{\Phi}^T\mathbf{y} = \mathbf{0}, \tag{2.3.7}$$

resolviendo para $\mathbf{c}$ se obtiene un estimador de $\hat{\mathbf{c}}$ que minimiza la suma de cuadrados dada en (2.3.6), y esta dado por

$$\hat{\mathbf{c}} = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{y}. \tag{2.3.8}$$

De acuerdo con esto el vector de valores ajustados $\hat{\mathbf{y}}$ es

$$\hat{\mathbf{y}} = \mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{y} \tag{2.3.9}$$

El uso de mínimos cuadrados ordinarios es apropiado en situaciones donde se asuma que los residuales ϵ_j son independientes e idénticamente distribuidos con media cero y varianza constante σ^2. En el caso que exista una estructura de correlación entre los errores, es necesaria la extensión del criterio de mínimos cuadrados ordinarios al caso de mínimos cuadrados ponderados. Mediante esta extensión se pretende incorporar la estructura de correlación en la estimación de los parámetros. En este caso se desea minimizar

$$SMSSE(\mathbf{y} \mid \mathbf{c}) = (\mathbf{y} - \mathbf{\Phi}\mathbf{c})^T\mathbf{W}(\mathbf{y} - \mathbf{\Phi}\mathbf{c}), \tag{2.3.10}$$

donde $\mathbf{W}$ es una matriz simétrica y positiva definida que permite pesos desiguales para los residuales.

Si la matriz de varianzas-covarianzas de los residuales ϵ_j, $\mathbf{\Sigma}_e$, es conocida, entonces

$$\mathbf{W} = \Sigma_e^{-1}. \tag{2.3.11}$$

Cuando la estimación de $\mathbf{\Sigma}_e$ no es posible, la covarianza entre los errores se asume como cero y $\mathbf{W}$ será una matriz diagonal, cuyos elementos estarán dados por el recíproco de la varianza del error asociado a y_j.

Minimizando (2.3.10) se obtiene que el estimador de mínimos cuadrados ponderados esta dado por

$$\hat{\mathbf{c}} = (\mathbf{\Phi}^T\mathbf{W}\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{W}\mathbf{y}. \tag{2.3.12}$$

2.3.4. Estimación de parámetros penalizados

Con el fin de permitir que la penalización de la rugosidad, $PEN_m(x)$, juegue un rol en la determinación de $x(s)$, se define una expresión que regule el ajuste de los datos contra el suavizamiento de la curva; ésta es conocida como mininos cuadrados penalizados,

$$PENSEE_\lambda(\mathbf{y} \mid \mathbf{c}) = (\mathbf{y} - \mathbf{\Phi c})^T\mathbf{W}(\mathbf{y} - \mathbf{\Phi c}) + \lambda \times PEN_m(x). \tag{2.3.13}$$

El parámetro de suavizamiento λ mide la importancia relativa entre el ajuste de los datos, medido por la suma de cuadrados de residuales, y la variabilidad de la función $x(t)$, medido por $PEN_m(x)$.

A medida que λ va creciendo, las funciones que no son lineales se vuelven más influyentes sobre la penalización por rugosidad, a través de $PEN_m(x)$, y consecuentemente $PENSEE_\lambda$ hace más énfasis en el suavizamiento de $x(t)$ y menos en el ajuste de los datos.

Por otra parte para valores muy pequeños de λ la curva tiende a ser más variable ya que hay menos penalización en la rugosidad, y cuando $\lambda \to 0$ la curva $x(t)$ es aproximada mediante la interpolación de los datos que satisface la ecuación $x(t_j) = y_j$ para toda j. Aún en este caso límite, la curva ajustada es la mejor curva suave y dos veces derivable que interpola a los datos, (Ramsay y Silverman, 2005[12]).

Podemos escribir a $PEN_m(x)$ de manera matricial, como sigue

$$\begin{aligned} PEN_m(x) &= \int [D^m x(s)]^2 ds, \\ &= \int [D^m \mathbf{c}^T \mathbf{\Phi}(s)]^2 ds, \\ &= \int \mathbf{c}^T D^m \mathbf{\Phi}(s) D^m \mathbf{\Phi}^T \mathbf{c} ds, \\ &= \mathbf{c}^T [\int D^m \mathbf{\Phi}(s) D^m \mathbf{\Phi}^T ds] \mathbf{c}, \\ &= \mathbf{c}^T \mathbf{R} \mathbf{c}, \end{aligned} \tag{2.3.14}$$

donde

$$\mathbf{R} = \int D^m \mathbf{\Phi}(s) D^m \mathbf{\Phi}^T ds. \tag{2.3.15}$$

Finalmente agregando la suma de cuadrados $SSE(\mathbf{y} \mid \mathbf{c})$ y PEN_m multiplicado por λ se obtiene

$$PENSEE_\lambda(\mathbf{y} \mid \mathbf{c}) = (\mathbf{y} - \mathbf{\Phi c})^T \mathbf{W} (\mathbf{y} - \mathbf{\Phi c}) + \lambda \mathbf{c}^T \mathbf{R} \mathbf{c}. \tag{2.3.16}$$

Derivando con respecto al vector de parámetros $\mathbf{c}$ e igualando a cero, se tiene

$$-2\mathbf{\Phi}^T \mathbf{W} \mathbf{y} + \mathbf{\Phi}^T \mathbf{W} \mathbf{\Phi c} + \lambda \mathbf{R c} = \mathbf{0}, \tag{2.3.17}$$

de donde se obtiene que un estimador de $\mathbf{c}$ esta dado por

$$\hat{\mathbf{c}} = (\mathbf{\Phi}^T \mathbf{W} \mathbf{\Phi} + \lambda \mathbf{R})^{-1} \mathbf{\Phi}^T \mathbf{W} \mathbf{y}. \tag{2.3.18}$$

De acuerdo con esto se tiene que los valores ajustados $\hat{y}$ son

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{\Phi} (\mathbf{\Phi}^T \mathbf{W} \mathbf{\Phi} + \lambda \mathbf{R})^{-1} \mathbf{\Phi}^T \mathbf{W} \mathbf{y}, \\ &= \mathbf{S}_{\lambda\phi} \mathbf{y}. \end{aligned} \tag{2.3.19}$$

donde $\mathbf{S}_{\lambda\phi}$ es la matriz " hat", simétrica pero a diferencia del caso de regresión estándar

$\mathbf{S}_{\lambda\phi}$ no es idempotente, es decir

$$\mathbf{S}_{\lambda\phi}\mathbf{S}_{\lambda\phi} \neq \mathbf{S}_{\lambda\phi}.$$

Se debe tener precaución al elegir el parámetro de suavizamiento, λ. Esta elección debe ser de manera que se obtenga un nivel aceptable de rugosidad en las curvas estimadas y no perder fuentes de variación importantes.

Existen varios procedimientos para la elección del parámetro λ. Estos métodos son descritos en el libro de Ramsay y Silverman (2007)[12]. Sin embargo, el procedimiento de suavizamiento esta determinado por los intereses del estudio que se esté realizando.

Continuando con el ejemplo del estudio de crecimiento en Berkeley en la figura 2.3.2 mostramos el ajuste de un spline de orden 6 con nodos en los puntos de medición de la altura. Dado que es de interés observar la aceleración, D^2x debemos penalizar la estimación por D^4x y para ello es necesario que el orden del spline sea 6.

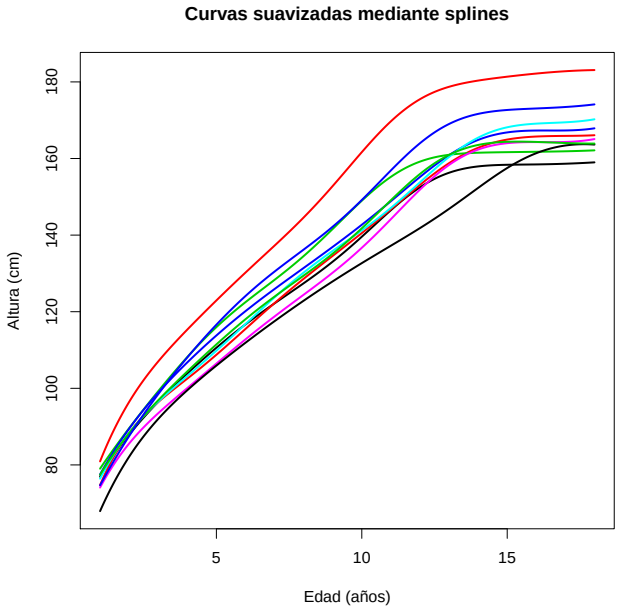


Figura 2.3.2: Ajuste de base de splines

2.4. Exploración de la variabilidad en datos funcionales

Al igual que en el caso de la estadística clásica, la variabilidad de un conjunto de datos funcionales resulta muy interesante. Para ello son necesarias técnicas que nos permitan determinar y observar las fuentes de variación. Entre las principales técnicas en el caso de datos funcionales se tiene

- Estadísticas descriptivas funcionales.
- Análisis de componentes principales funcionales.
- Análisis de derivadas.

2.4.1. Estadísticas descriptivas funcionales

Como en cualquier otro análisis de datos las estadísticas descriptivas son una gran herramienta para la descripción y exploración de datos. En el caso de datos funcionales estas estadísticas son definidas como funciones, donde la función media está dada por:

$$\bar{x}(t) = \frac{1}{N}\sum_{i=1}^{N} x_i(t). \tag{2.4.1}$$

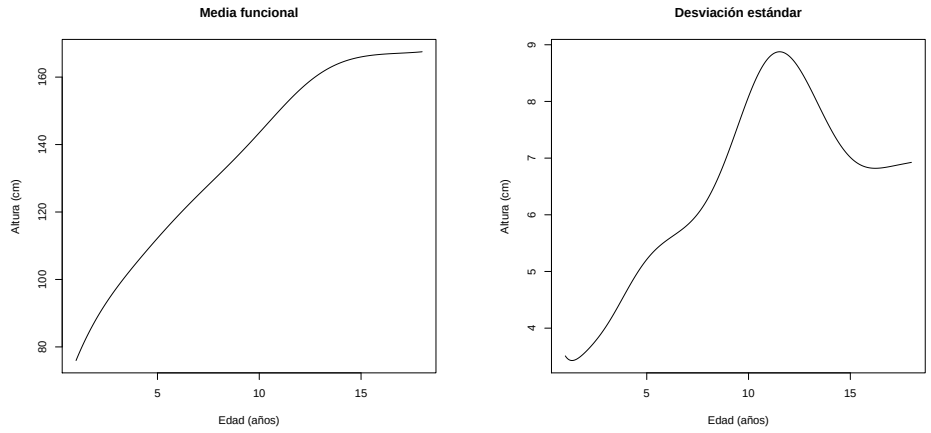
Esto es, el promedio de las funciones punto a punto, sobre todas las replicas del proceso. De igual manera definimos la función varianza

$$var[x(t)] = \frac{1}{N-1}\sum_{i=1}^{N}[x_i(t) - \bar{x}(t)]^2, \tag{2.4.2}$$

donde la desviación estándar se define como la raíz cuadrada de la función varianza.

En la figura 3.4.2 se muestran la media y la varianza para los datos del estudio de crecimiento de Berkeley

Definamos ahora la función de covarianza funcional, la cual mide el grado de dependencia a través de los diferentes argumentos y es calculada para toda t_1 y t_2 mediante la expresión



a) b)

Figura 2.4.1: Estadísticas descriptivas: a) Media funcional, b) Desviación estándar funcional

$$Cov\left[x\left(t_1,t_2\right)\right]=\frac{1}{N-1}\sum_{i=1}^{N}\left[x_i\left(t_1\right)-\bar{x}\left(t_1\right)\right]\left[x_i\left(t_2\right)-\bar{x}\left(t_2\right)\right], \tag{2.4.3}$$

de donde se obtiene que la función correlación esta dada por

$$Corr\left[x\left(t_1,t_2\right)\right]=\frac{Cov\left[x\left(t_1,t_2\right)\right]}{\sqrt{var\left[x\left(t_1\right)\right]var\left[x\left(t_2\right)\right]}}. \tag{2.4.4}$$

2.4.2. Análisis de derivadas

Otra herramienta de gran ayuda en el análisis de datos funcionales es el uso de las derivadas, las cuales son ampliamente usadas tanto en los métodos exploratorios como en el desarrollo detallado de la metodología de datos funcionales. Esta metodología puede ser usada para la detección de componentes de variación.

El uso de esta metodología juega también un rol muy importante en la regulación o suavizamiento de una curva ya que con ella se puede observar realmente que tan suave es una curva.

Además mediante el análisis de derivadas pueden observarse características no observables directamente en los datos originales. Por ejemplo considere el estudio de crecimiento en Berkeley, al observar la segunda derivada, figura 2.4.2, podemos observar que el crec-

imiento se acelera durante la infancia y durante la pubertad, cosa que no se puede observar a partir de los datos, ver figura 2.3.2.

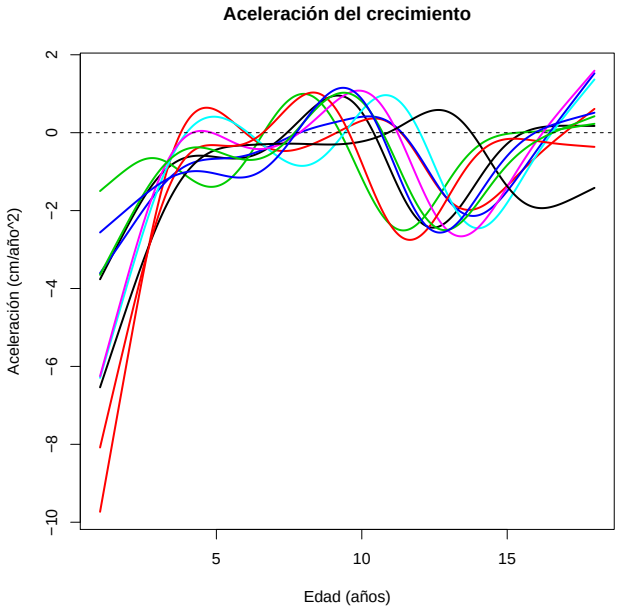


Figura 2.4.2: Análisis de derivadas

El análisis diferencial principal es un método usado para identificar importantes componentes de variabilidad. El desarrollo de esta técnica puede verse en Ramsay y Silverman (2005) [12].

2.5. Bootstrap suavizado

El bootstrap es una técnica para la creación de muestras artificiales, las cuales pueden ser usadas para la estimación de densidades.

Considere $X_1, ..., X_n$ una muestra de datos *iid* tal que $X_i \sim f(x)$ donde frecuentemente $f(x)$ es desconocida por lo cual hay que estimarla mediante algún método de estimación de densidad.

Mediante el boostrap se pretende generar valores provenientes de $f(x)$. Existen dos metodologías las cuales se pueden distinguir de la siguiente manera:

- **El bootstrap clásico:** Usualmente funciona de la siguiente manera: Sea $X_1, ..., X_n$ una muestra *iid* con densidad $f(x)$. $X_1^*, ..., X_n^*$ es una muestra boostrap de tamaño n, la cual es tomada con remplazo de la muestra original

- **El boostrap suavizado:** Consiste en obtener muestras de tamaño n de la función de densidad estimada $\hat{f}(x)$. Entonces $X_1^*, ..., X_n^*$ *iid* según $X^* \sim \hat{f}(x)$ es una muestra boostrap.

En este caso usaremos el estimador Kernel para densidades el cual se define como:

Definición 2.5.1 *Dado un kernel K y un número positivo h, llamado ancho de banda, el estimador de densidades kernel es definido como:*

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big) \tag{2.5.1}$$

tal que $K(x) \geq 0$ y $\int K(x)dx = 1, \int xK(x)dx = 0$ y $\int x^2K(x)dx > 0$.

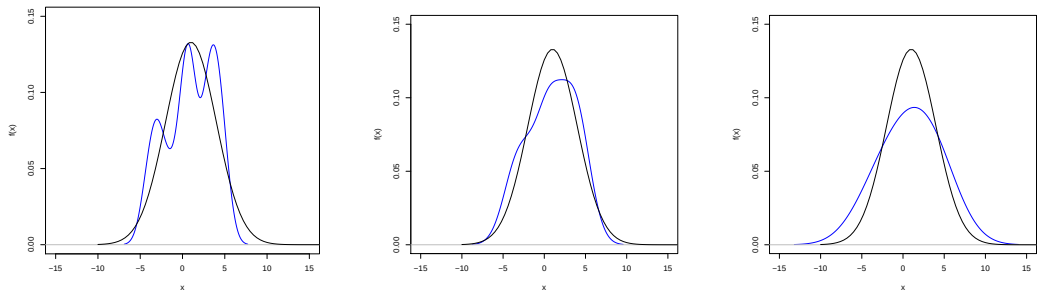
El valor del ancho de banda, h, debe ser elegido cuidadosamente ya que valores muy grandes de h producen una función de densidad estimada demasiado suave, mientras que valores muy pequeños de h producen una función de densidad estimada demasiado rugosa. La elección óptima de h debe realizarse de modo que este valor minimice el error cuadrado medio integrado que se define como:

$$ECMI(\hat{f}) = \int_{\mathbb{R}} E(\hat{f}(x) - f(x))^2. \tag{2.5.2}$$

En la figura 2.5.1 se puede observar la manera en la cual influye el ancho de banda , h, en la estimación de densidades vía kernel.

La extensión al caso multivariado de este estimador kernel esta dado por

$$\hat{f}(\mathbf{x}) = \frac{1}{n|\mathbf{H}|} \sum_{i=1}^{n} K_d\Big(\mathbf{H^{-1}}(\mathbf{x} - \mathbf{x}_i)\Big) \tag{2.5.3}$$



a) b) c)

Figura 2.5.1: Influencia del parámetro h: a) h=0.9, b) h=1.519 (óptimo), c) h=3

donde K_d es una función de densidad simétrica y de dimensión d. En este caso la forma en que se introduce el grado de suavizamiento es mediante lo que se conoce como matriz ventana $\mathbf{H}$, donde $\mathbf{H}$ es una matriz $d \times d$ no singular.

2.5.1. Elección de la matriz ventana

Los elementos de la matriz $\mathbf{H}$ en la práctica se estiman de la siguiente manera:

- $\mathbf{H} = h\mathbf{C}^{1/2}$, donde $\mathbf{C}$ es la matriz de covarianza de los datos. Esta elección considera las correlaciones entre los elementos del vector $\mathbf{x}$.

- En el caso en que $Cov(\mathbf{x}) = 0$ y $Var(x_i) = Var(x_j)$ esto se reduce a $\mathbf{H} = h\mathbf{I_d}$.

- En el caso en que $Cov(\mathbf{x}) = 0$ y $Var(x_i) \neq Var(x_j)$ esto se reduce a $\mathbf{H} = diag(h_1, ..., h_d)$.

En todos estos casos h es un parámetro de suavizamiento, el cual puede ser elegido por algún criterio análogo a los vistos en dimensión uno.

Bowman y Foster (1993)[2] proponen un h el cual minimiza el ECMI y es extendible a cualquier dimensión

$$h = \left[\frac{4}{(d+2)n}\right]^{1/(d+4)}. \tag{2.5.4}$$

En este caso utilizaremos un kernel gaussiano dado por

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\{\frac{-x^2}{2}\}, \quad x > 0. \tag{2.5.5}$$

2.6. Conjuntos de confianza

En el caso de datos funcionales no hablamos de intervalos de confianza en el sentido estricto, en este caso nos referimos a conjuntos de confianza, esto se debe a las distintas formas que pueden tomar los datos funcionales, ya que estos no aparecen en una forma ordenada y generalmente se entrelazan entre si.

Definimos un conjunto de confianza para el funcional $X(t)$ con nivel α, como el conjunto de curvas c las cuales tienen la misma distribución que $X(t)$ tal que

$$CS(x) = \{c : d(x, c) < D_\alpha\} \tag{2.6.1}$$

donde d es una distancia funcional y D_α es tal que $P(CS(x)) = \alpha$. Esto es, dada una muestra original $X_1(t), X_2(t), ..., X_n(t)$ se extraen muestras bootstrap denotadas por $X_1^*(t), X_2^*(t), ..., X_n^*(t)$. Entonces un conjunto del $\alpha\,\%$ de confianza basado en el estadístico $T(X_1, X_2, ...X_n)$ es definido calculando el valor de D_α tal que el $\alpha\,\%$ de replicas $T(X_1^*, X_2^*, ...X_n^*)$ se localicen dentro de una distancia menor que D_α de su promedio, es decir D_α es el radio del conjunto de confianza bootstrap centrado en la media de T. En este caso para medir la distancia D_α se usará una métrica L^2 la cual para una curva $x(t)$ está dada por

$$||x(t)||_2 = \Big(\int_{t_{\text{mín}}}^{t_{\text{máx}}} x(t)^2 dt \Big)^{1/2} \tag{2.6.2}$$

Capítulo 3

Cartas de control para datos funcionales

3.1. Introducción

En el presente trabajo se realizó un análisis de datos de motores empleando un enfoque de datos funcionales mediante una propuesta de cartas de control para datos funcionales. Con este enfoque se pretende captar mucha más información que con técnicas tradicionales. Los datos fueron tomados del trabajo de Amiri et al. (2009)[1].

3.2. Descripción del problema

Una de las características de calidad más importantes en un motor de automóvil es la relación entre el torque (N$\times$ m) producido por el motor y la velocidad del motor en revoluciones por minuto (RPM).

Como se mencionó anteriormente los datos fueron tomados del trabajo de Amiri et al. (2009)[1] y corresponden a mediciones del torque del motor de un automóvil (TU3 de autos Peugeot). Los motores fueron puestos en marcha a diferentes RPM (variable explicativa) y los valores del torque (variable de respuesta) fueron registrados. El la figura 3.2.1 se presentan los datos correspondientes a 26 motores evaluados a distintos niveles de RPM.

De acuerdo con lo anterior si el proceso de manufactura está bajo control, los perfiles que describen la relación entre RPM y torque deben ser muy similares, mientras que si algún motor presenta anomalías el perfil correspondiente a este motor debe diferir de los

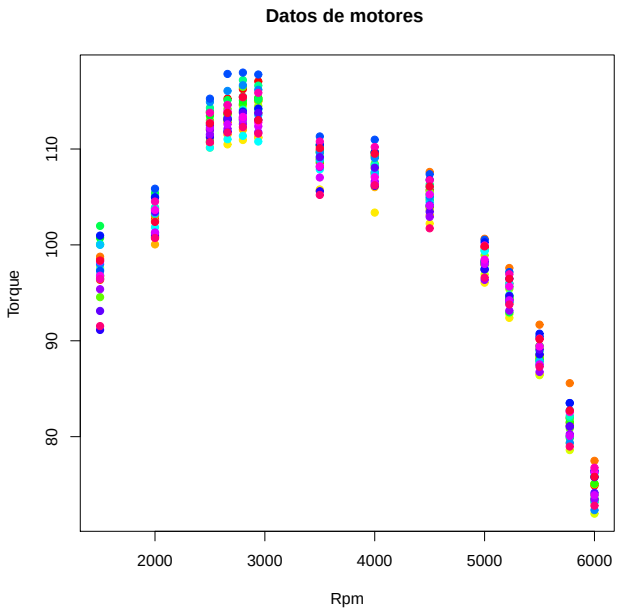


Figura 3.2.1: Datos de motores

demás. La idea de este análisis es la implemetación de una técnica para el monitoreo de la calidad del motor.

3.3. Propuestas realizadas

En el trabajo de Amiri et al. (2009)[1] se propone usar una técnica multivariada de control de calidad. Para ello comienzan por ajustar un modelo paramétrico de segundo orden a los datos, ver figura 3.3.1.

La idea propuesta en el trabajo de Amiri et al. (2009) es ajustar un modelo de efectos mixtos de segundo orden de la forma

$$\mathbf{y}_j = \mathbf{X}\beta + \mathbf{Z}\mathbf{b}_j + \epsilon_j, \tag{3.3.1}$$

donde $\mathbf{y}_j$ es el vector respuesta del $j-simo$ perfil, $\mathbf{X}$ es la matriz de observaciones, β es el vector de parámetros fijos, $\mathbf{Z}$ matriz asociada con los efectos aleatorios, $\mathbf{b}_j$ vector de efectos aleatorios. Se asume que los efectos aleatorios son normalmente distribuidos, $\mathbf{b}_j \sim N_p(\mathbf{0}, \mathbf{D})$ donde $\mathbf{0}$ es un vector de ceros y $\mathbf{D}$ es una matriz diagonal de varianzas-

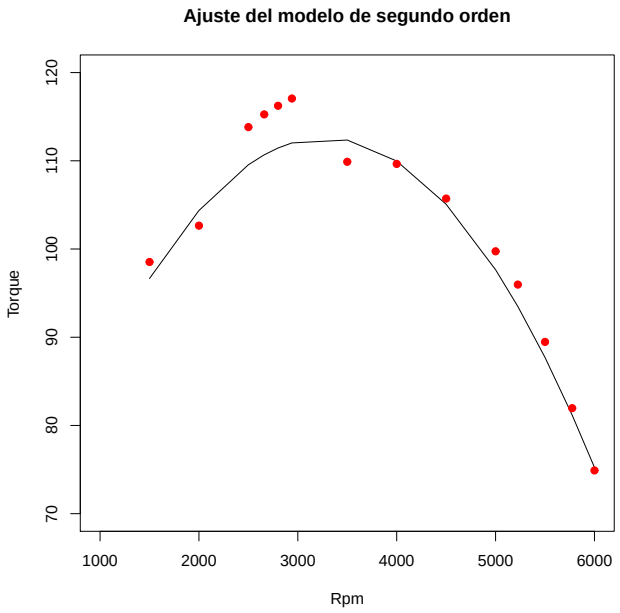


Figura 3.3.1: Ajuste del modelo de segundo orden

covarianzas positiva definida. También se asume que los errores siguen una distribución normal, esto es $\epsilon_j \sim N_p(\mathbf{0}, \mathbf{\Sigma})$ donde $\mathbf{\Sigma}$ es una matriz positiva definida a la cual se le permite una estructura de correlación $AR(1)$ y además se asume $Cov(\epsilon_j, \mathbf{b}_j) = 0$.

A partir de lo anterior fue construida una carta de control multivariada, T^2 para los parámetros de efectos aleatorios estimados, mediante la cual se monitoreará el proceso. Si algún motor presenta algún defecto los parámetros de éste se verán afectados por lo que mostrara una señal de fuera de control.

Mediante esta técnica se observó que los motores están bajo control, por lo cual a través estos datos fue realizada la **Fase I** para proseguir a monitorear el proceso en la **Fase II**. El objetivo de este trabajo es la propuesta de un método alternativo a este procedimiento, para mayor información ver el trabajo de Amiri et al. (2009)[1].

3.4. Cartas de control para datos funcionales

El objetivo principal de este trabajo es la construcción de una carta de control así como la evaluación de su desempeño. Esto se hará mediante la utilización de un enfoque de datos funcionales.

Como puede observarse, los datos de motores están dados en forma discreta por lo cual para iniciar el análisis usando un enfoque de datos funcionales, hay que buscar una representación de los datos que muestre su verdadera forma funcional. En este caso el uso de una base B-spline, como las definidas en la sección 2.3.1, permite una mejor representación de los datos que el ajuste propuesto por Amiri et al.(2009)[1].

Mediante el paquete *fda* de *R* fue definida una base B-spline que aproxima la forma funcional de los datos. Esta base es definida por splines cúbicos cuyos nodos fueron definidos de tal forma que se agruparan más donde exista mayor variabilidad por parte de la curva, los nodos fueron definidos en $2000, 2500, 2660, 2940, 3500, 4000, 4500, 5775$. En la figura 3.4.1 se muestra el ajuste del B-spline de splines cúbicos, donde las lineas punteadas muestran los puntos donde fueron colocados los nodos.

Note que en este caso no se tiene gran interés en el análisis de derivadas de los datos por lo cual se ha tomado el parámetro de suavizamiento, λ como cero. Esto reduce la estimación de los coeficientes, $\hat{\mathbf{c}}$, definidos en (2.3.18) al caso de estimación por mínimos cuadrados dada en (2.3.12). En el cuadro 3.4.1 son presentados los valores de $\hat{\mathbf{c}}$.

Al igual que el caso de la estadística clásica, las estadísticas descriptivas son de gran ayuda para un primer análisis de los datos. En la figura 3.4.2 se muestran la media y desviación estándar para los datos correspondientes a los motores. En esta figura puede observarse que la variabilidad del proceso es mayor al inicio y a la mitad del estudio.

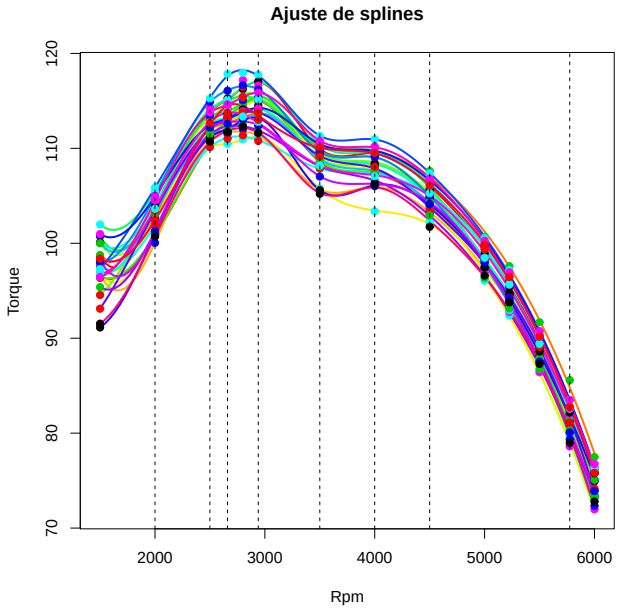
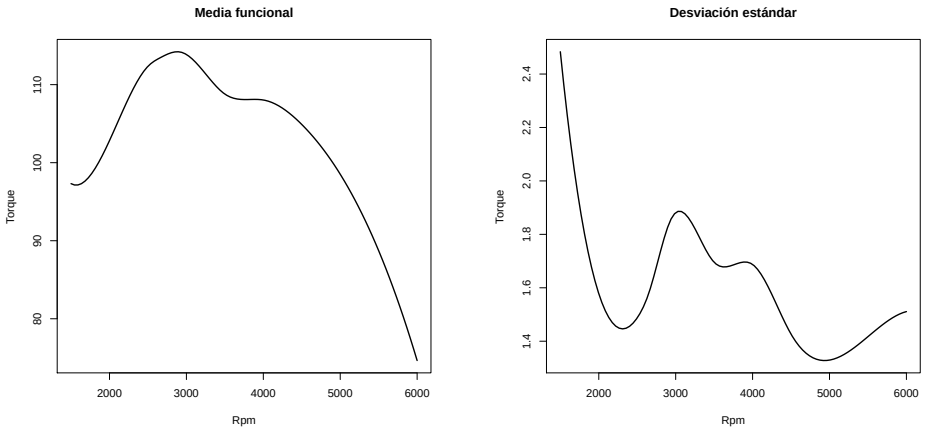


Figura 3.4.1: Ajuste de B-spline

	Nodos									
No. De Motor	**1500**	**2000**	**2500**	**2660**	**2940**	**3500**	**4000**	**4500**	**5775**	**6000**
329	98.530	92.947	112.718	115.232	119.394	107.043	110.847	103.334	92.795	74.731
449	96.350	94.738	106.524	113.811	116.852	106.364	108.269	101.886	90.966	74.896
529	98.770	94.838	112.008	112.584	115.034	108.848	110.386	105.425	94.568	77.867
642	96.700	89.354	111.194	111.756	112.047	103.451	107.085	101.408	90.538	74.890
724	96.750	91.799	110.852	110.270	112.613	104.587	103.312	102.692	87.579	73.126
803	97.610	94.186	111.523	113.093	117.471	106.687	108.961	101.196	88.860	71.780
930	100.061	93.704	114.412	112.549	116.096	106.959	109.113	103.950	91.440	75.044
1148	94.549	98.203	109.504	115.409	116.928	107.486	109.554	102.803	90.648	73.905
1171	96.480	94.861	106.660	113.953	117.011	106.499	108.409	102.015	91.083	74.986
1516	100.730	94.227	113.983	113.632	117.650	107.364	110.392	101.965	91.051	73.102
1791	101.980	98.662	112.313	115.471	115.784	108.726	110.911	103.362	88.520	73.557
2600	96.829	95.608	113.279	115.251	119.298	107.874	110.637	104.421	91.227	75.688
3100	100.070	96.527	111.826	113.425	115.493	107.432	109.310	102.548	91.120	75.772
3720	97.440	95.552	108.647	111.510	110.880	107.037	107.771	104.319	92.629	76.076
4025	100.000	95.325	111.873	111.653	115.776	107.046	108.484	102.295	90.330	73.083
4068	97.980	97.908	113.195	116.497	117.148	107.333	110.419	101.792	90.315	72.184
4926	97.290	101.871	110.777	118.630	118.024	109.232	112.131	104.852	93.355	76.401
5155	100.970	98.447	112.206	113.054	115.752	108.924	110.576	104.736	93.217	76.427
6143	91.130	93.653	110.651	112.072	115.301	102.367	107.378	102.887	90.036	75.495
6844	93.110	98.739	110.291	113.543	114.645	107.759	108.922	100.488	92.731	73.911
7811	95.380	92.861	110.649	112.589	111.992	107.279	107.289	100.600	89.425	73.483
8007	98.280	90.752	112.172	112.813	113.116	105.458	106.742	104.001	89.192	73.725
8623	96.790	97.184	111.271	114.233	112.343	107.219	107.529	104.419	91.052	76.307
9388	96.450	97.471	113.365	114.563	117.584	108.846	111.080	104.919	92.364	76.478
9404	91.530	94.192	109.047	112.235	112.379	102.668	107.497	99.095	91.314	72.253
10430	98.369	96.029	108.504	115.519	112.587	109.295	110.240	104.082	93.072	75.712

Cuadro 3.4.1: Coeficientes B-spline estimados



a) b)

Figura 3.4.2: Estadísticas descriptivas: a) Media funcional, b) Desviación estándar funcional

Una vez encontrada la mejor representación de los datos a través del ajuste de la B-spline pasemos a construir los límites de control de la carta para lo cual haremos uso de la metodología bootstrap.

Para la obtención de muestras bootstrap funcionales, comencemos por la obtención de muestras discretizadas para cada perfil. Note que el uso de técnicas usuales de bootstrap generará muestras sobre un rango muy limitado de valores. Ante estas limitaciones el uso de bootstrap suavizado mejora la obtención de dichas muestras.

En la generación de muestras bootstrap suavizadas fue utilizado un kernel multivariado, como el definido en 2.5.3. El Kernel usado fue un kernel gaussiano, definido 2.5.5, de dimensión $d = 14$.

En este caso la elección de la matriz ventana, $\mathbf{H}$, será de la forma $h\mathbf{\Sigma}_x^{1/2}$ donde $\mathbf{\Sigma}_x$ es la matriz de varianzas-covarianzas del vector $x(t_1), ..., x(t_n)$ y h es el parámetro de suavizamiento.

De acuerdo con Cuevas et al. (2006)[3] la generación de estas muestras es de la siguiente

manera:

- Generar B muestras bootstrap denotadas por $x_i^b(t_i)$ $i = 1, 2, ..., n$ y $b = 1, 2, ..., B$
- Obtener muestras suavizadas

$$y_i^b(t_i) = x_i^b(t_i) + z_i^b(t_i), \tag{3.4.1}$$

donde $z_i^b(t_i)$ es tal que $z_i^b(t_1), ..., z_i^b(t_n)$ tiene distribución normal con vector de medias $\mathbf{0}$ y matriz de varianzas- covarianzas $h\mathbf{\Sigma}_x$.

Dada la expresión 2.5.4 propuesta por Bowman y Foster (1993)[2] se tiene que el h óptimo está dado por

$$h = 0.7725785. \tag{3.4.2}$$

En la figura 3.4.3 son presentadas 1000 simulaciones realizadas mediante el bootstrap suavizado a las que posteriormente fue ajustada una B-spline definida por splines cúbicos.

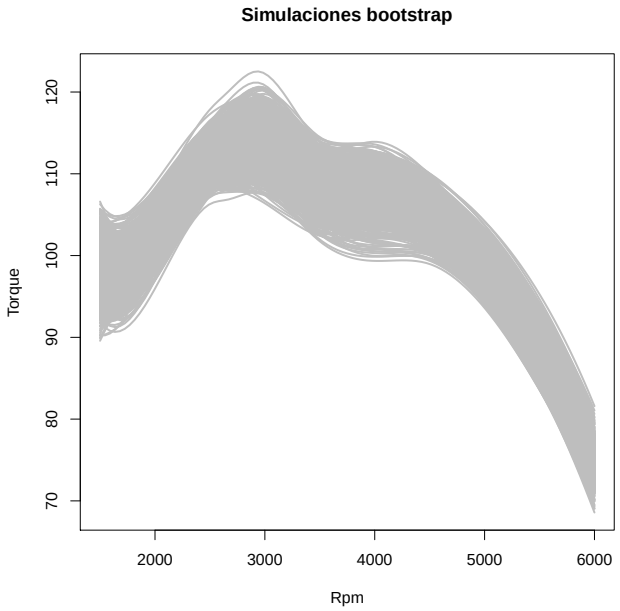


Figura 3.4.3: Simulaciones bootstrap

Una vez obtenidas las simulaciones bootstrap es necesario fijar el valor D_α tal que

$P(CS(x)) = \alpha$ y así construir un conjunto de confianza como se menciona en la sección 2.6.

Considerando 10000 muestras bootstrap y tomando un nivel $\alpha = 0.95$ se obtuvo un conjunto del 95 % de confianza, esto es, el valor $D_{.95}$ es fijado de tal forma que el 95 % de las muestras bootstrap se encuentren a una distancia $D_{.95}$ de la media funcional de los datos de motores, es decir, este será un conjunto de confianza centrado en la media funcional. En la figura 3.4.4 se puede observar un conjunto del 95 % de confianza para los datos de motores.

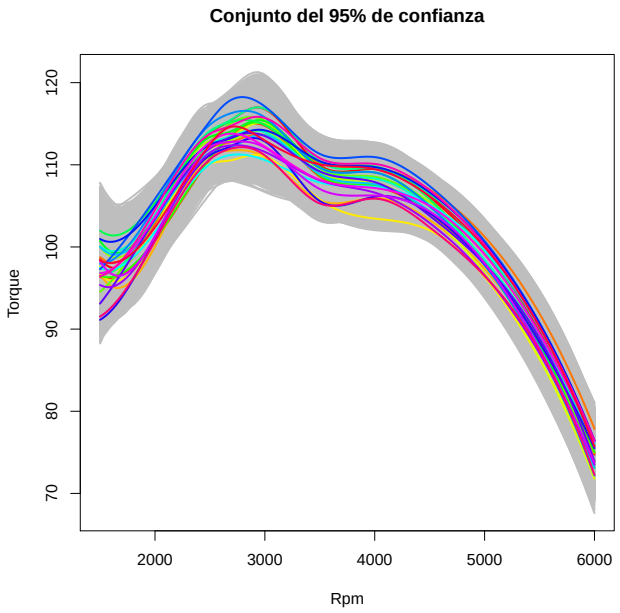


Figura 3.4.4: Conjunto de confianza

Note que todas las curvas correspondientes a los datos de motores, están dentro del conjunto de confianza, lo cual indica que el proceso está bajo control. Dado que el proceso se encuentra bajo control y no presenta mucha variabilidad, este conjunto de confianza puede ser usado para el monitoreo del proceso en la **Fase II**.

Una vez construido el conjunto de confianza es necesario evaluar su desempeño para evitar que envíe señales falsas de fuera de control. Como se mencionó en la sección 1.5,

el ARL es una muy buena forma de evaluar el desempeño de la carta de control. En la figura 3.4.5 es presentada la idea de longitud de corrida para datos funcionales.

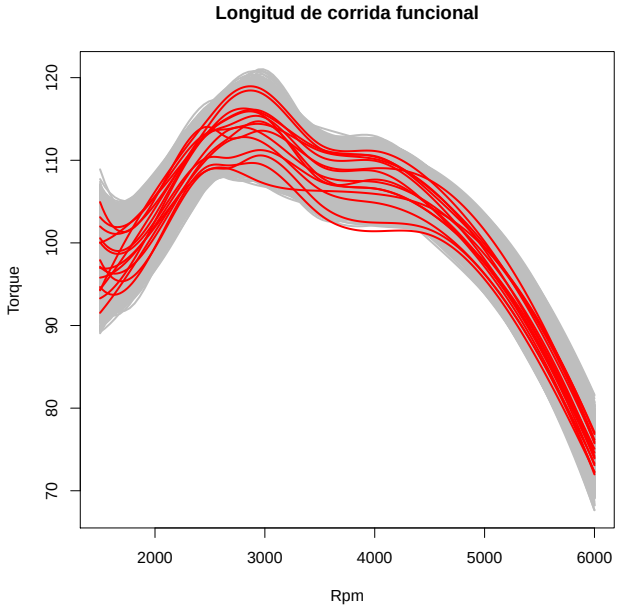


Figura 3.4.5: Longitud de corrida

Para el cálculo del ARL fueron simuladas curvas mediante el bootstrap suavizado, estas curvas fueron simuladas hasta hasta obtener una señal de fuera de control (RL). Después de repetir este proceso 10000 es posible obtener la distribución empírica de RL, que como se definió en la sección 1.5, tiene distribución geométrica con parámetro $p = 1/ARL$. De acuerdo con lo anterior se tiene que la longitud promedio de corrida (ARL) esta dada por

$$ARL = 29.5607, \tag{3.4.3}$$

de donde se sigue que la probabilidad de que la carta de control envíe una señal de fuera de control es

$$p = 0.0338287. \tag{3.4.4}$$

Es decir, se esperaría que en promedio cada 29 curvas el proceso mostrará una señal de fuera de control sin que realmente esta exista.

Nivel de confianza	ARL	Probabilidad de fuera de control
0.9	16.7211	0.05980468
0.95	29.5607	0.0338287
0.99	103.6727	0.00964574
0.9975	141.9561	0.007044431

Cuadro 3.4.2: Relación Confianza-ARL

En la figura 3.4.6 se puede observar que efectivamente la distribución del la longitud de corrida (RL) es geométrica de parámetro $p = 1/ARL$.

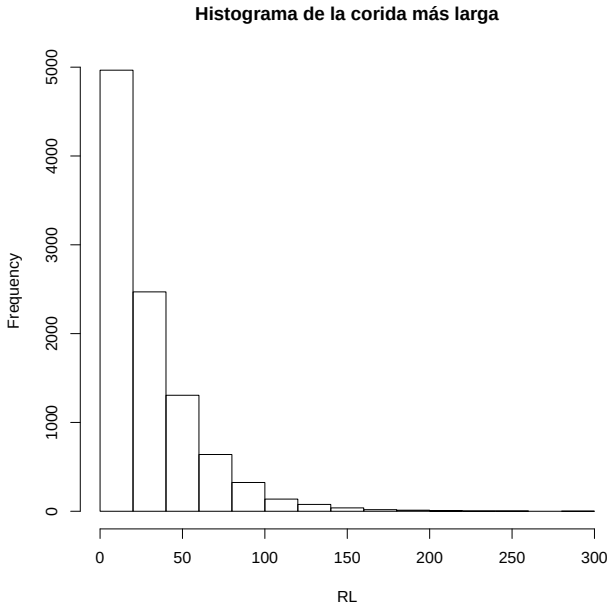


Figura 3.4.6: Distribución de la corrida mas larga

Recuerde que la amplitud del conjunto de confianza quedará determinada por el porcentaje de variabilidad que se desee cubrir y el ARL deseado. En otras palabras, si se desea cubrir un alto porcentaje de variabilidad, el conjunto de confianza será muy amplio, por tanto el ARL será grande con lo que no habrá muchas señales en falso; en cambio si el porcentaje es pequeño, en conjunto de confianza será muy estrechos y el ARL pequeño causando muchas señales falsas. En el cuadro 3.4.2 se muestra la relación entre el nivel de confianza del conjunto y el ARL obtenido.

Otra manera de obtener estas cartas de control es construir los límites de control usando la idea propuesta por Shewhart (1924) y la extensión del teorema de Chebishev 1.2.1. Es decir, la amplitud del intervalo de control estará dada en términos de la desviación estándar. Por ejemplo, considere la idea de las cartas tipo Shewhart, esto es

$$LCI = \mu_w - 3\sigma_w, \quad LC = \mu_w, \quad LCS = \mu_w + 3\sigma_w, \tag{3.4.5}$$

donde μ_w y σ_w correspondes a la media y desviación estándar para datos funcionales, los cuales fueron definidos en la sección 2.4. En la figura 3.4.7 se propone la extensión de las cartas tipo Shewhart al caso de datos funcionales.

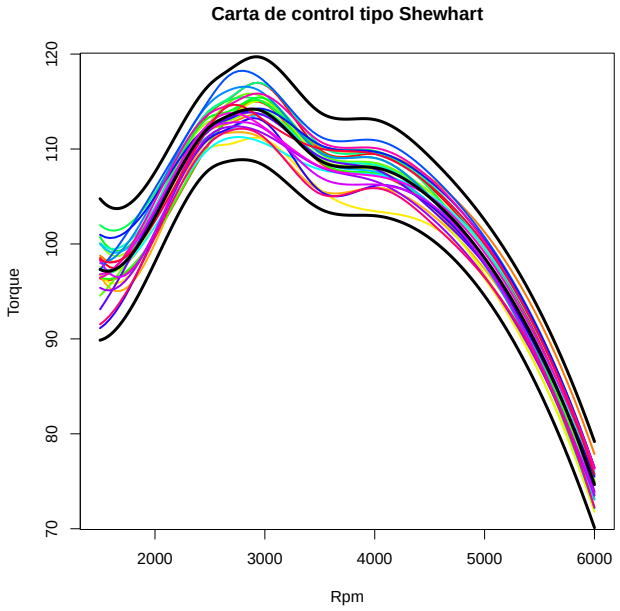


Figura 3.4.7: Carta tipo Shewhart funcional

En el cuadro 3.4.3 se muestra la relación entre el ARL de la carta, y la amplitud del intervalo en términos de la desviación estándar. Los valores de este cuadro fueron calculados simulando trayectorias de motores mediante el bootstrap suavizado, hasta obtener una señal de fuera de control (RL), es decir, que estas excedan en algún punto los limites de control. Finalmente repitiendo este proceso 10000 veces, es posible obtener la distribución

Amplitud del conjunto de confianza	**ARL**	**Probabilidad de fuera de control**
σ	1.039	0.96246391
$\mathbf{2}\sigma$	1.9068	0.52443885
$\mathbf{3}\sigma$	7.7515	0.12900729
$\mathbf{3.5}\sigma$	21.0268	0.04755835
$\mathbf{4}\sigma$	70.3506	0.01421452
$\mathbf{4.5}\sigma$	293.0233	0.0034127
$\mathbf{5}\sigma$	1549.756	0.000645263

Cuadro 3.4.3: Amplitud del conjunto de confianza

empírica de RL, lo cual permite obtener la longitud promedio de corrida (ARL).

Note que el ARL se incrementa conforme se aumenta la amplitud del intervalo y con ello la probabilidad de obtener una señal de fuera de control falsa disminuye. Sin embargo, un ARL muy grande involucra un conjunto de confianza demasiado amplio lo cual reducirá la precisión de la estimación

Capítulo 4

Conclusiones

Como fue mencionado a lo largo de este trabajo, el monitoreo de un proceso de producción permite detectar cuando un proceso se ha salido de control, esto es, identificar en que punto del proceso las condiciones con las que se trabaja han sufrido cambios provocando anomalías en la producción. Una vez detectadas la señales de fuera de control es posibles corregirlas para mantener el proceso bajo control.

Las herramientas tradicionales para el monitoreo de procesos han sido las cartas de control univariadas y multivariadas. Estas cartas son de gran utilidad cuando la característica de calidad se expresa por medio de una o más variables. Sin embargo, en muchas ocasiones la calidad de un producto o proceso está dada por la relación funcional entre una variable respuesta y una o más covariables, por tanto es necesaria la implementación de una herramienta que permita el monitoreo del proceso bajo este esquema.

La estadística de datos funcionales permite trabajar con funciones en lugar de simples datos, univariados o multivariados, este tipo de análisis ofrece una gran variedad de herramientas para el análisis de grandes conjuntos de datos y por tanto se ha convertido en una valiosa herramienta de análisis.

En este trabajo se propuso una carta de control, para monitorear una característica de calidad funcional de motores de un automóvil (TU3 de autos Peugeot). En estos, la calidad se expresa por la relación entre el torque (N$\times$ m) y la velocidad (Rpm) a la que se gira.

Las funciones que aquí se presentan fueron analizadas previamente por Amiri et al. (2009)[1], a través de una carta de control multivariada para los parámetros de un modelo de efectos mixtos, ajustado a las trayectorias de los motores.

Aquí fueron modeladas las trayectorias de los motores, mediante la definición de una base B-spline que aproxima la forma funcional de los datos, a las cuales posteriormente se les construyó una cartas de control mediante la metodología bootstrap.

Entre las herramientas usadas en la simulación de las trayectorias de motores, el bootstrap suavizado resulta una herramienta poderosa en la creación de muestras artificiales, ya que amplia el rango de posibles valores sobre los cuales realizar el re-muestreo.

Las cartas de control para datos funcionales propuestas en este trabajo, permiten el monitoreo de un proceso cuando la característica de calidad esté en función de una variable explicativa. A diferencia de la propuesta realizada por Amiri et al. (2009)[1], este tipo de cartas de control permite observar todos los trayectorias a lo largo del periodo de observación, y de esta manera identificar los periodos donde el proceso presenta una mayor variabilidad.

Otro aspecto relevante de las cartas de control para datos funcionales es el fácil cálculo del ARL y con ello evaluar el desempeño de la carta. En el problema de datos de motores el ARL obtenido resulta bastante bueno, ya que se esperan pocas señales de fuera de control falsas y una mejor detección de verdaderas señales de fuera de control. Usando un esquema de las cartas tipo Shewhart y de acuerdo al cuadro 3.4.3 la amplitud del intervalo será elegida dependiendo del ARL que se desee obtener.

En resumen, la estadística de datos funcionales ofrece una muy atractiva alternativa para el análisis de grandes conjuntos de datos. Las cartas de control para datos funcionales propuestas ofrece una muy buena alternativa para el monitoreo de procesos, permitiendo la incorporación de variables explicativas en el monitoreo de alguna característica de calidad.

Bibliografía

[1] Amiri, A., Jensen, W. A. and Kazemzadeh, R. B., *A case study on monitoring polinomial profiles in the automotive industry*, Qual. Reliab. Engng. Int. 26, 509-520,2009.

[2] Bowman A.W., Foster P.J., *Adaptive smoothing and density-based tests of multivariate normality*, Journal of the American Statistical Association, Vol. 88, No. 442,529-537, 1993.

[3] Cuevas A., Febrero-Bande, M., Fraiman R., *On the use of bootstrap for estimating functions with functional data*, Computational Statistics and Data Analysis 51: 1063-1074, 2006.

[4] de Boor, C., *A practical guide to splines*, Revised Edition, Springer, New York, 2001.

[5] Febrero-Bande, M., Galeano, P., and Gonzalez-Manteiga, W., *Outlier detection in functional data by depth measures with application to identify abnormal NOx levels*, Environmetrics 19, 4, 331-345, 2008.

[6] Febrero-Bande, M., *A present overview on functional data analysis*, Boletín De La Sociedad De Estadística E Investigación Operativa Vol. 24, Pp. 7-14, 2008.

[7] Ferraty F. y Vieu Ph., *Nonparametric functional data analysis*, Springer, New York, 2006.

[8] Fraiman R., Muniz G., *Trimmed means for functional data*, Test 10: 419-440, 2001.

[9] Gutiérrez, P. H. y De la Vara, S. R., *Control estadístico de calidad y seis sigma*, Mcgraw-Hill Interamericana, México, 2004.

[10] Montgomery D., *Control estadístico de la calidad*, Limusa-Wiley, México, 2004, 3a ed.

[11] Ramsay J.O, Hooker G. and Graves S., *Functional data analysis with R and Matlab*, Springer, New York, 2009.

[12] Ramsay J.O and Silverman B.W., *Functional data analysis*, Springer, New York, 2005, 2a ed.

[13] Wasserman, L., *All of statistics: A concise course in statistical inference*, Springer-Verlag. New York, NY, 2004