ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

ANÁLISIS DE DATOS FUNCIONALES APLICADO A LA DISTRIBUCIÓN DE LA POBLACIÓN ECUATORIANA

TESIS PREVIA A LA OBTENCIÓN DEL GRADO DE MAGISTER EN ESTADÍSTICA APLICADA

IVÁN CRISTIAN NAULA REINA
inaula@hotmail.com

Director: Dr. LUIS ALCIDES HORNA HUARACA
luis.horna@enp.edu.ec

Quito, Mayo 2016

DECLARACIÓN

Yo, Iván Cristian Naula Reina, declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentada para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

La Escuela Politécnica Nacional puede hacer uso de los derechos correspondientes a este trabajo, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.

Iván Cristian Naula Reina

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por Iván Cristian Naula Reina, bajo mi supervisión.

Dr. Luis Horna Huaraca

DIRECTOR

AGRADECIMIENTOS

A mis profesores, y en especial al Dr. Marco Calahorrano y a mi director de tesis Dr. Luis Horna por su aporte con las revisiones y sugerencias para alcanzar el objetivo planteado.

DEDICATORIA

A mi familia

Iván Cristian Naula Reina

ÍNDICE DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

RESUMEN

En el presente documento se muestra la aplicación del Análisis Funcional a la Estadística habitual y Multivariante mediante el Análisis de Datos Funcionales (ADF) en particular, el Análisis de Componentes Principales Funcionales (ACPF), con el cual se describe el comportamiento de la población ecuatoriana por estructura de edad representada en dos componentes principales. Los datos utilizados para este estudio son las proyecciones de población realizada por el INEC a partir del Censo de Población y Vivienda 2010; esta proyección está desagregada por edad y sexo para las 24 provincias del Ecuador.

Palabras clave: Análisis de Datos Funcionales, Componentes Principales Funcionales.

ABSTRACT

This document presents the application of Functional Analysis to the usual and Multivariate Statistics by means of Funtional Data Analysis (FDA) in particular, the Functional Principal Componentes Analysis (FPCA), which describes the behavior from 2010 to 2020 of the Ecuadorian population by age represented in two main components. The data used for this study are the population projections made by INEC from Population and Housing 2010 Census; This projection is dissaggregated by age and sex for the 24 provinces of Ecuador.

Keywords: Functional Data Analysis, Functional Principal Components.

1 ASPECTOS GENERALES

1.1 PRESENTACIÓN DEL PROBLEMA

El análisis de datos funcionales es una rama de la estadística en la que los elementos de la muestra son funciones aleatorias. Este trabajo trata sobre el estudio de la distribución de la población en el Ecuador con respecto a su estructura demográfica por provincia, sexo y edad, haciendo énfasis en el análisis descriptivo y de componentes principales desde la perspectiva del análisis funcional.

1.2 JUSTIFICACIÓN

Puesto que el análisis de datos funcionales es una herramienta estadística relativamente nueva y poca aplicada en el Ecuador, nace la necesidad de comprender la fundamentación teórica y aplicar esta herramienta, específicamente mediante el uso del análisis de componentes principales y el análisis descriptivo desde la perspectiva del análisis funcional, para realizar el estudio del comportamiento de la población del Ecuador por provincia, género y edad; y así contribuir al fortalecimiento del estudio del análisis multivalente en el país. Los resultados obtenidos podrán ser de interés para el INEC y en general para el gobierno nacional, además de utilidad para la comunidad académica ecuatoriana.

1.3 OBJETIVOS

1.3.1 OBJETIVOS GENERALES

Analizar el comportamiento de la distribución de la población ecuatoriana con respecto a su estructura demográfica por provincia, sexo, edad y su evolución, desde la perspectiva del análisis de datos funcional, haciendo énfasis en el análisis descriptivo y de componentes principales funcionales.

1.3.2 OBJETIVOS ESPECÍFICOS

- Estudiar y aplicar la teoría del análisis de datos funcionales y en particular, el análisis descriptivo y el análisis de componentes principales.
- Estudiar y aplicar la teoría de interpoladores que permita transformar los datos discretos en funciones.
- Implementar algoritmos en el lenguaje de programación R, que permita realizar la transformación de los datos, el análisis descriptivo y el análisis de componentes principales.
- Analizar los resultados.

1.4 METODOLOGÍA PROPUESTA

Los datos provienen del Censo de Población y Vivienda del 2010 y de reportes estadísticos de las proyecciones de la población ecuatoriana por edades simples y sexo de cada provincia desde el 2010 hasta el 2020.

Con estos datos se construirán las pirámides de población que luego serán transformadas en funciones con segunda derivada continua mediante funciones interpoladoras, logrando así tener funciones que pertenecen a L^2.

La ventaja del método de interpolación mediante splines es el uso de polinomios de grado bajo para producir globalmente interpolantes suaves. Dada una función $f \in C^2([a,b])$, para construir la función spline de interpolación S de grado 3, aplicamos las condiciones de las funciones splines al polinomio cúbico.

Se consideran 24 provincias y la proyección de 10 años para cada una. Se tiene entonces 264 funciones, con las cuales se procederá a realizar un suavizamiento con el objetivo de minimizar errores en los cálculos de cada una de estas funciones. Para la suavización de los datos se utilizaran los B-Splines.

A continuación se realizará un análisis descriptivo sobre la población mediante los cálculos de la media y varianza funcional, esto con el fin de observar la tendencia central y la dispersión de los datos.

Luego se aplicará un análisis de componentes principales funcional cuyo objetivo es de reducir la dimensión del espacio puesto que las variables aleatorias funcionales que se tratarán se encuentran en espacios de dimensión infinita.

Finalmente, por medio del análisis de componentes principales funcionales se interpretaran las pautas de comportamiento de las frecuencias por edades de las pirámides de población, para esto, encontraremos funciones (componentes principales funcionales) las cuales resumen el comportamiento de la variabilidad de las funciones.

Se desarrollará algoritmos en R, que permita la creación de las pirámides de población y su transformación en funciones, considerando las edades de los hombres como positivas y las edades de las mujeres como negativas.

En el capítulo 1 se presentan el problema, la justificación, los objetivos y la metodología utilizada en el desarrollo de la tesis. En el capítulo 2 se expone una introducción a la teoría del análisis de datos funcionales. En el capítulo 3 se exhibe la teoría de interpolación, en particular la interpolación con Splines cúbicos. Ésta será útil para la transformación de los datos discretos en datos funcionales. Dado que el análisis de datos funcionales es una herramienta novedosa de análisis, en el capítulo 4 se elabora un conjunto de algoritmos con los que podemos realizar los cálculos. Estos algoritmos serán implementados en el lenguaje de programación R. En el capítulo 5, se muestra la aplicación del análisis de datos funcionales, empezando con un análisis descriptivo funcional para finalizar con el cálculo de componentes principales funcionales y su respectivo análisis y, finalmente, en el capítulo 6, se presentan las conclusiones y recomendaciones de este trabajo.

2 INTRODUCCIÓN

2.1 ANÁLISIS DESCRIPTIVO FUNCIONAL

2.1.1 PRODUCTO ESCALAR Y DEFINICIONES

Las medidas del análisis descriptivo pueden ser expresadas con notación algebraica. Ramsay y Silverman las definen así para que luego sean fácilmente adaptables a la naturaleza de nuestros datos: variables, vectores o funciones. La clave de todo esto es cómo definimos para cada elemento (variable, vector o función) el producto escalar.

Empezaremos con la definición de producto escalar y con sus propiedades. Tengamos en cuenta que ahora los elementos pueden ser de cualquier tipo: vectores o funciones.

Este capítulo se fundamenta principalmente en [1].

DEFINICIÓN 2.1:

El *producto escalar euclidiano* de elementos x e y del espacio vectorial E, denotado como $\langle x, y \rangle$, es una aplicación

$$\langle \cdot, \cdot \rangle : E \times E \to \mathrm{R}$$

$$(x, y) \to \langle x, y \rangle$$

que satisface las siguientes propiedades

1) Simetría: $\langle x, y \rangle = \langle y, x \rangle$ para todo x e y en E

2) Positividad: $\langle x, x \rangle \geq 0$ para todo $x \in E$, con $\langle x, x \rangle = 0 \Leftrightarrow x = 0$

3) Bilinealidad: Para todo $a, b \in \mathrm{R}$, $\langle ax + by, z \rangle = a\langle x, y \rangle + b\langle y, z \rangle$, para todo $x, y, z \in E$.

El producto escalar se puede interpretar como medida de cantidad de asociación entre dos elementos del espacio vectorial en el que trabajamos. También si lo aplicamos a un mismo elemento es una medida de magnitud del mismo elemento.

Podemos ampliar el producto escalar euclidiano si introducimos un operador W. En este caso denotamos el producto escalar general como $\langle x, y \rangle_w$ y su definición dependerá del elemento del espacio vectorial con el que trabajamos. El producto escalar general será pues la aplicación sobre el espacio E que satisface las propiedades de simetría, positividad y bilinealidad al incorporar W.

DEFINICIÓN 2.2:

La *norma de un elemento* x del espacio, se define como:

$$\|x\| = \sqrt{\langle x, x \rangle}$$

PROPIEDADES:

1) $\|x\| \geq 0$ y $\|x\| = 0 \Leftrightarrow x = 0$

2) $\|ax\| = |a|\|x\| \quad \forall a \in \mathrm{R}$

3) $\|x + y\| \leq \|x\| + \|y\|$

4) $|\langle x, y \rangle| \leq \|x\|\|y\| = \sqrt{\langle x, x \rangle \langle y, y \rangle}$ (*desigualdad de Cauchy-Schwartz*).

COROLARIO:

De la *desigualdad de Cauchy-Schwartz* podemos deducir la *desigualdad del coseno*, que es la siguiente:

$$-1 \leq \frac{\langle x, y \rangle}{\|x\|\|y\|} \leq 1$$

DEFINICIÓN 2.3:

Diremos que dos elementos x e y del espacio son *ortogonales* si

$$\langle x, y \rangle = 0$$

DEFINICIÓN 2.4:

El *ángulo* θ *entre los elementos* x e y los definimos como:

$$\theta = \arccos \frac{\langle x, y \rangle}{\|x\| \cdot \|y\|}$$

Una interpretación de la norma de un elemento x puede ser la de la magnitud del elemento x dentro del espacio vectorial. La desigualdad de Cauchy-Schwartz nos está indicando que el valor absoluto del producto escalar de dos elementos está acotado por el producto de las normas de los elementos. Si el valor absoluto del producto escalar de estos dos elementos se aproxima a la cota entonces los elementos están definidos en direcciones del espacio vectorial semejantes, por lo que podemos decir que el grado de asociación entre ellos es grande. Sin embargo si el producto escalar es próximo a cero el grado de asociación es pequeño y las direcciones que definen los elementos son casi ortogonales.

DEFINICIÓN 2.5:

La *distancia entre* x *e* y se define como:

$$d_{x,y} = \|x - y\| = \sqrt{\langle x - y, x - y \rangle}$$

Si las direcciones y los sentidos que definen x e y son muy parecidos y además la norma de los elementos es parecida entonces el tamaño de $x - y$ será pequeño y por tanto la distancia entre x e y será corta. Tanto si las direcciones son perpendiculares como si las direcciones son iguales pero de sentido contrario, las distancias se harán grandes.

2.1.2 ESTADÍSTICOS DESCRIPTIVOS EN NOTACIÓN DE PRODUCTO ESCALAR

Es posible generalizar definiciones para los estadísticos básicos de tendencia central y de dispersión si los definimos a través del producto escalar, sin especificar el espacio vectorial en el que estamos trabajando. Veremos las definiciones más importantes:

DEFINICIÓN 2.6:

Definimos la *media* como la cantidad obtenida al realizar el siguiente producto escalar:

$$\bar{x} = \frac{1}{N}\langle x, \mathbf{1} \rangle,$$

donde $\mathbf{1}$ es elemento unidad y $N = \langle \mathbf{1}, \mathbf{1} \rangle = \|1\|^2$.

DEFINICIÓN 2.7:

Definimos la *varianza* como la cantidad obtenida al realizar la siguiente operación:

$$S_x^2 = \frac{1}{N} \left\langle x - \bar{x}\mathbf{1}, x - \bar{x}\mathbf{1} \right\rangle = \frac{1}{N} \|x - x\mathbf{1}\|^2$$

donde $\bar{x}\mathbf{1} = (\bar{x}, \bar{x}, .., \bar{x})$

Se puede interpretar esta medida como el módulo de la diferencia entre el elemento x respecto al elemento media.

DEFINICIÓN 2.8:

Definimos la *covarianza entre* x *e* y como la cantidad obtenida al realizar el siguiente producto escalar:

$$S_{x,y} = \frac{1}{N} \left\langle x - \bar{x}\mathbf{1}, y - \bar{y}\mathbf{1} \right\rangle$$

DEFINICIÓN 2.9:

Definimos *la correlación entre* x *e* y como la cantidad obtenida al realizar el siguiente producto escalar:

$$\begin{aligned} r_{x,y} &= \frac{S_{x,y}}{S_x S_y} \\ &= \frac{\left\langle x - \bar{x}\mathbf{1}, y - \bar{y}\mathbf{1} \right\rangle}{\left\| x - \bar{x}\mathbf{1} \right\| \left\| y - \bar{y}\mathbf{1} \right\|} \\ &= \cos\left(x - \bar{x}\mathbf{1}, y - \bar{y}\mathbf{1} \right) \end{aligned}$$

La *correlación* x *e* y es la proporción de la magnitud explicada por la relación de direcciones entre x e y respecto al total de magnitud de estos dos elementos $\left(S_x, S_y\right)$.

2.1.3 PRODUCTO ESCALAR. DEFINICIONES PARA DATOS FUNCIONALES

Si nos referimos al análisis funcional de datos nuestros elementos x e y son funciones de R en R y W es una función de peso definida positiva. En adelante todas las integrales serán definidas sobre R, salvo que explícitamente se indique lo contrario. En este caso:

DEFINICIÓN 2.10:

Sea el espacio de funciones de cuadrado integrable,

$$L^2 = \left\{ f \ : \mathrm{R} \to \mathrm{R} : \int_{\mathrm{R}} f^2(t)dt < \infty \right\}$$

Definimos el *producto escalar euclidiano* en L^2 como:

$$\langle x, y \rangle = \int_{\mathrm{R}} x(t) y(t) dt$$

OBSERVACIÓN: El espacio de funciones de cuadrado integrable L^2 es un espacio de Hilbert.

Cuando nos referimos al espacio L^2 podemos imaginar que estamos trabajando con un vector de infinitas componentes, una para cada valor $t \in \mathrm{R}.$ El producto escalar entonces será la suma del producto de las infinitas componentes de x e y. Teniendo en cuenta que la distancia entre componente y componente es infinitesimal, el producto escalar se expresa a través de integrales. Observemos que si $x(t)$ crece a la vez que $y(t)$ en los mismos intervalos de t entonces el producto escalar se hará grande; mientras que si cuando $x(t)$ es grande e $y(t)$ es pequeño, el producto no se hace grande. Deducimos del punto anterior que el concepto de producto escalar es el de una magnitud de la relación entre las funciones $x(t)$ e $y(t)$. Veremos que $\langle x, y \rangle$ es, en efecto, un producto escalar:

1) Simetría: $\langle x, y \rangle = \int_{\mathrm{R}} x(t)y(t)dt = \int_{\mathrm{R}} y(t)x(t)dt = \langle y, x \rangle$
2) Positividad: $\langle x, x \rangle = \int_{\mathrm{R}} x(t)x(t)dt \geq 0$ y $\langle x, x \rangle = \int_{\mathrm{R}} x(t)x(t)dt = 0 \Leftrightarrow x = 0$
3) Bilinealidad: Para todo

$$(ax + by)z = \int_{\mathrm{R}} \left[a.x(t) + b.y(t) \right] z(t)dt = \int_{\mathrm{R}} \left[a.x(t).z(t) + b.y(t) \right] dt =$$

$$\int_{\mathrm{R}} a.x(t).z(t)dt + \int_{\mathrm{R}} b.y(t).z(t)dt = a\langle x, z \rangle + b\langle y, z \rangle$$

DEFINICIÓN 2.11:

Definimos *el producto escalar general en el espacio* L^2 como:

$\langle x, y \rangle_w = \int_R w(t)x(t)y(t)dt$ donde $w(t) \in L^2$

El producto escalar general lo que hace es ponderar por una función $w(t)$ para dar más peso a unos intervalos de t y restárselo a otros.

PROPIEDADES:

1) Simetría:

$$\langle x, y \rangle_w = \int_R w(t)x(t)y(t)dt = \int_R w(t)y(t)x(t)dt = \langle y, x \rangle_w$$

2) Positividad:

$$\langle x, x \rangle_w = \int_R w(t)x(t)x(t)dt \geq 0 \quad \text{y} \quad \langle x, x \rangle_w = \int_R w(t)x(t)x(t)dt = 0 \Leftrightarrow x = 0$$

3) Bilinealidad : Para todo $a, b \in \mathrm{R}$

$$\langle ax + by, z \rangle_w = \int_R w(t)\left[a.x(t) + b.y(t)\right]z(t)dt =$$

$$= \int_R \left[a.w(t).x(t).z(t) + b.w(t).y(t).z(t)\right]dt =$$

$$= a\int_R w(t).x(t).z(t)dt + b\int_R w(t).y(t).z(t)dt = a\langle x, z \rangle_w + b\langle y, z \rangle_w$$

DEFINICIÓN 2.12:

La *L^2 – norma de* $x(t)$ se define como:

$$\|x\| = \sqrt{\langle x, x \rangle} = \sqrt{\|x\|^2} = \sqrt{\int_R x^2(t)dt}$$

La norma nos da la magnitud del elemento $x(t)$ dentro del espacio. Las propiedades de la norma se cumplen de la misma manera en este espacio. Existe también en el caso de las funciones el concepto de ángulo derivada de la desigualdad de Cauchy-Schwarts:

DEFINICIÓN 2.13:

El *ángulo* θ *entre las funciones* $x(t)$ *e* $y(t)$ los definimos como:

$$\theta = \arccos\frac{\langle x, y \rangle}{\|x\|\|y\|}$$

$$= \frac{\int_R x(t)y(t)dt}{\sqrt{\int_R x(t)^2 dt \int_R y(t)^2 dt}}$$

Explica la relación de dependencia entre $x(t)$ e $y(t)$.

DEFINICIÓN 2.14:

Diremos que una función $x(t)$ es *ortogonal* a $y(t)$ si:

$$\langle x, y \rangle = \int_{\mathrm{R}} x(t)y(t)dt = 0$$

2.1.4 ESTADÍSTICOS DESCRIPTIVOS EN ANÁLISIS FUNCIONAL DE DATOS

Estadísticos sobre una función

En el análisis funcional de datos hay que hacer distinciones entre los diferentes estadísticos, en función de si se definen para estudiar la tendencia central y la dispersión dentro de un solo elemento (estadísticos sobre una función), o bien si se define para estudiar una muestra de funciones (estadísticos de una muestra de una función aleatoria) o bien si se define para estudiar muestras de dos o más funciones aleatorias. La primera clase de estadísticos pretende resumir la información de toda una función con pocas medidas. La segunda clase de estadísticos intentan caracterizar la función aleatoria mientras que la tercera clase intenta captar la relación entre dos o más funciones aleatorias. En este punto exponemos los primeros.

Hallar la media, la varianza y la covarianza de una función $x(t)$ es simplemente un ejercicio de deducción si tenemos en cuenta las definiciones de estas cantidades en notación algebraica y la definición del producto escalar (2.10). Para simplificar los conceptos supondremos que el intervalo en el que mueve t será $[0,T]$. Por lo tanto:

DEFINICIÓN 2.14:

Definimos *la media de la función* $x(t)$ *o también valor medio de la función* $x(t)$ como:

$$\bar{x} = \frac{1}{T}\langle x, \mathbf{1}\rangle = \frac{1}{\int_0^T \mathbf{1}(t)dt}\int_0^T x(t)\mathbf{1}(t)dt$$

donde la función $\mathbf{1}(t) = 1 \ \forall t \in \mathrm{R}.$

La media de $x(t)$ ($\bar{x}$) va a representar la tendencia central y el nivel medio de todos los valores de $x(t).$ No hay que confundir este concepto $(\bar{x} \in \mathrm{R})$ con los conceptos de función media y función valor medio, que veremos a continuación.

DEFINICIÓN 2.15:

La *función valor medio de* $x(t),$ que denotaremos por $\bar{x}\mathbf{1}(t)$, es aquella cuyo valor $\forall t \in [0,T]$ es el de la media de

Esta definición va a tener sentido a la hora de expresar la varianza y la covarianza de una función.

PROPIEDAD:

Sea $x(t)$ una función continua definida en el intervalo $[a,b]$ y sea $\bar{x}$ su valor medio. Definimos ahora la función valor medio $f_{\bar{x}}(t) = \bar{x} \quad \forall t \in [a,b]$. Entonces se cumple que:

$$\int_a^b x(t)dt = \int_a^b f_{\bar{x}}(t)dt$$

Es decir, el área que encierra la función $x(t)$ es la misma que el área que encierra la función de valor constante igual al valor medio de la función en el intervalo $[a,b]$.

PRUEBA:

$\int_a^b x(t)dt - \int_a^b f\bar{x}(t)dt = \int_a^b x(t)dt - f_{\bar{x}}(t)\int_a^b dt =$

$= \int_a^b x(t)dt - \left(\frac{1}{\int_a^b dt}\int_a^b x(t)dt\right)\int_a^b dt =$

$= \int_a^b x(t)dt - \frac{\int_a^b dt}{\int_a^b dt}\int_a^b x(t)dt = 0,$

como queríamos demostrar.

DEFINICIÓN 2.16:

La *varianza de* $x(t)$, que denotamos con $S^2_{x(t)}$ se define como:

$$S^2_{x(t)} = \frac{1}{T}\left\langle x-\bar{x}\mathbf{1}, x-\bar{x}\mathbf{1}\right\rangle$$
$$= \frac{1}{\int_0^T dt}\int_0^T \left(x(t)-\bar{x}\mathbf{1}(t)\right)^2 dt$$

Esta cantidad representa la variación media (al cuadrado) de todos los valores de la función respecto a su valor medio. Gráficamente una función con mucha variabilidad será aquella que englobe un área grande entre dicha función y la función valor medio.

DEFINICIÓN 2.17:

La *covarianza entre dos funciones* $x(t)$ *e* $y(t)$, que denotamos con $S_{x(t),y(t)}$ se define como:

$$S_{x(t),y(t)} = \frac{1}{T}\left\langle x-\bar{x}\mathbf{1}, y-\bar{y}\mathbf{1}\right\rangle$$
$$= \frac{1}{\int_0^T dt}\int_0^T \left(x(t)-\bar{x}\mathbf{1}(t))(y(t)-\bar{y}\mathbf{1}(t)\right)dt$$

DEFINICIÓN 2.18:

La *correlación entre dos funciones* $x(t)$ *e* $y(t)$, que denotamos como $r_{x(t),y(t)}$ se define como:

$$r_{x(t),y(t)} = \frac{S_{x(t),y(t)}}{S_{x(t)}S_{y(t)}} =$$
$$= \frac{\frac{1}{T} < x-\bar{x}\mathbf{1}, y-\bar{y}\mathbf{1} >}{\frac{1}{T} < x-\bar{x}\mathbf{1}, x-\bar{x}\mathbf{1} > \frac{1}{T} < y-\bar{y}1, y-\bar{y}\mathbf{1} >} =$$
$$= \frac{\frac{1}{\int_0^T dt}\int_0^T \left(\left(x(t)-\bar{x}\mathbf{1}(t)\right)\left(y(t)-\bar{y}\mathbf{1}(t)\right)\right)dt}{\frac{1}{\int_0^T \mathbf{1}(t)dt}\int_0^T \left(x(t)-\bar{x}\mathbf{1}(t)\right)^2 dt \frac{1}{\int_0^T \mathbf{1}(t)dt}\int_0^T \left(y(t)-\bar{y}\mathbf{1}(t)\right)^2 dt}$$

En este caso la función covarianza mide la relación entre dos funciones $x(t)$ e $y(t)$ y la correlación se sitúa entre -1 y $+1$. Digamos que la covarianza es el producto de dos áreas: la comprendida entre la función $x(t)$ y su valor medio (Ax), y la comprendida entre la función $y(t)$ y su valor medio (Ay). Se pueden dar tres

situaciones $r \approx +1;\ r \approx -1$ y $r \approx 0$. Cuando $r \approx +1$ tendremos que en el mismo rango de $t \in \mathrm{R}$, Ax tiene el mismo signo que Ay. Cuando $r \approx -1$, Ax y Ay tienen signo distinto. Por último cuando $r \approx 0$ tenemos en unos intervalos de $t \in \mathrm{R}$ $Ax \approx 0$ y por tanto $Ax.Ay \approx 0$, y en otros intervalos $Ay \approx 0$ y por tanto $Ax.Ay \approx 0$. También entre funciones periódicas si una tiene un periodo múltiplo de dos del otro también su correlación será cero.

El valor medio, la varianza y la correlación nos van a servir para identificar bien el comportamiento particular de cada una de nuestras funciones y sus relaciones dos a dos.

Estadísticos de muestras de una función aleatoria

Se llama función aleatoria a aquella cuyo valor, para cada valor del argumento (o de los argumentos) es una variable aleatoria. La función aleatoria se puede considerar como el conjunto de magnitudes aleatorias $X_t = X(t)$, $\alpha < t < \beta$, que representan los valores de la misma para diferentes valores de t. Esto quiere decir que la función aleatoria es equivalente a un conjunto infinito de variables aleatorias. Sabemos que, para muestras de valores en R la media, la varianza y la covarianza son valores también en R. Cuando trabajamos con muestras cuyos elementos son los vectores en R^n, la naturaleza de los estadísticos varía; la media también es un vector pero luego tenemos expresada de forma compacta tanto la varianza como la covarianza, mediante lo que denominamos matriz de varianzas covarianzas.

En el análisis funcional sucede algo parecido a lo que sucede con el caso de los vectores. Podemos pensar que en análisis funcional trabajamos con los vectores con infinitas componentes, una para cada $t \in \mathrm{R}$. Tendremos así una expresión de la función media y varianza como funciones a lo largo de t (vector con infinitas componentes) y las funciones covarianza y correlación entre diferentes funciones serán funciones que van de R^2 en R.

DEFINICIÓN 2.19:

Sea $x_1(t), x_2(t), ..., x_n(t)$ una muestra de funciones de una función aleatoria $x(t)$, definida en $[0,T]$. Definimos la *función media muestral de* $x(t)$ como:

$$\bar{x}(t) = \frac{1}{n}\sum_{i=1}^{n} x_i(t)$$

Si fijamos el valor de t en un punto concreto y evaluamos ahí todas las funciones entonces obtenemos una muestra de tamaño n de la cual podemos extraer su media. La *función media muestral de* $x(t)$ nos va a dar la media de esos valores en ese valor de t de forma explícita en una sola función.

DEFINICIÓN 2.20:

Sea $x_1(t), x_2(t), ..., x_n(t)$ una muestra de funciones de una función aleatoria $x(t)$, definida en $[0,T]$. Definimos la *función varianza muestral de* $x(t)$ como:

$$Var_{x(t)}(t) = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i(t) - \bar{x}(t)\right)^2$$

Y definimos la *función desviación estándar muestral de* $x(t)$ como:

$$Stdev_{x(t)}(t) = \sqrt{Var_{x(t)}(t)}$$

De la misma manera que la función media muestral de $x(t)$ nos indica la tendencia central de las funciones en un t dado, las funciones varianza muestral y desviación estándar muestral de $x(t)$ nos cuantifican el valor medio al cuadrado y el valor medio respectivamente de las desviaciones respecto la media en t.

DEFINICIÓN 2.21:

Sea $x_1(t), x_2(t), ..., x_n(t)$ una muestra de funciones de una función aleatoria $x(t)$, definida en $[0,T]$. La *función covarianza muestral de* $x(t)$ *entre* t_1 *y* t_2 será:

$$Cov_{x(t)}(t_1, t_2) = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i(t_1) - \bar{x}(t_1)\right)\left(x_i(t_2) - \bar{x}(t_2)\right)$$

DEFINICIÓN 2.22:

Sea $x_1(t), x_2(t), ..., x_n(t)$ una muestra de funciones de una función aleatoria $x(t)$, definida en $[0,T]$. La *función correlación muestral de* $x(t)$ *entre* t_1 *y* t_2 será:

$$Corr_{x(t)}(t_1,t_2)=\frac{Cov_{x(t)}(t_1,t_2)}{\sqrt{Var_{x(t)}(t_1)Var_{x(t)}(t_2)}}$$

Estas dos medidas nos van a indicar la magnitud de la relación entre el comportamiento de la función $x(t)$ en el valor t_1 y el comportamiento de la función $x(t)$ en el valor t_2. En realidad si fijamos t_1 y t_2 podríamos pensar que tenemos dos muestras del mismo tamaño: $x_1(t_1),x_2(t_1),...,x_n(t_1)$, y $x_1(t_2),x_2(t_2),...,x_n(t_2)$ y podríamos calcular la covarianza o el coeficiente de correlación entre las dos muestras. Con esto resumimos la dependencia de los registros a través de los distintos t.

Estadísticos de muestras de dos o más funciones aleatorias

Supongamos que ahora tenemos muestras de tamaño n de dos funciones aleatorias $x(t)$ e $y(t)$ y queremos saber qué relación tiene una con otra. Queremos saber la magnitud de la dependencia entre una y otra función para t_1 y t_2 fijados.

DEFINICIÓN 2.23:

Sean dos funciones aleatorias $x(t)$ e $y(t)$ y sean:

$x_1(t),x_2(t),...,x_n(t)$ muestra de $x(t)$

$y_1(t),y_2(t),...,y_n(t)$ muestra de $y(t)$

entonces la *función de covarianza* cruzada de $x(t)$ e $y(t)$ en t_1, t_2 se define como:

$$Cov_{x(t),y(t)}(t_1,t_2)=\frac{1}{n-1}\sum_{i=1}^{n}\left[x_i(t_1)-\bar{x}(t_1)\right]\left[y_i(t_2)-\bar{y}(t_2)\right]$$

DEFINICIÓN 2.24:

Sea la situación de la definición 2.23. Entonces la *función de correlación* cruzada de $x(t)$ e $y(t)$ en t_1, t_2 se define como:

$$Corr_{x(t),y(t)}(t_1,t_2)=\frac{Cov_{x(t),y(t)}(t_1,t_2)}{\sqrt{Var_{x(t)}(t_1)Var_{y(t)}(t_2)}}$$

Observemos que si fijamos t_1 y t_2 lo único que estamos haciendo es calcular el coeficiente de correlación entra la muestra $x_1(t), x_2(t), ..., x_n(t)$ y la muestra $y_1(t), y_2(t), ..., y_n(t)$.

2.2 ANÁLISIS DE COMPONENTES PRINCIPALES EN ESPACIO CON PRODUCTO ESCALAR

El análisis de componentes principales en espacios con producto escalar (ACPPE) definido, considera que todos los elementos del espacio en el que trabajamos (sea R^n o sea L^2) se pueden expresar como combinación lineal de unos pocos elementos. Es decir, $\widehat{x_i} = \sum_{k=1}^{K} f_{i,k}\, \xi_k$ donde K es menor que la dimensión del espacio.

El objetivo del ACPPE es el de condensar la máxima información dentro de la combinación lineal con el menor número de elementos posible. En el ACPPE, por tanto, buscamos en primer lugar una nueva base de elementos $\xi = (\xi_1, \xi_2, ..., \xi_n)$ que contenga la máxima variabilidad posible en su primera componente. De la variabilidad restante, es decir, del espacio ortogonal al generado por el elemento ξ_1, buscamos que la segunda componente contenga la máxima posible, y así sucesivamente hasta formar una base con los elementos ortogonales generadores de todo el espacio. La transformación con la que conseguimos una base con las propiedades anteriormente descritas en la resultante de obtener los elementos propios. El concepto de elemento propio no es más que la generalización del concepto de vector propio pero definido para cualquier espacio a través de productos escalares.

Resuelto su cálculo a este nivel general posteriormente se trataría simplemente de adaptar la notación a la naturaleza del espacio con el que trabajaremos. En primer lugar introduciremos unas definiciones, luego veremos cómo se resuelve el problema de elementos propios y en el punto siguiente nos plantearemos el problema de buscar funciones propias como caso particular del problema que planteamos ahora.

2.2.1 PROYECCIONES EXPRESADAS COMO PRODUCTOS ESCALARES

Sean $v_1, v_{2,}..., v_n$ elementos del espacio E. Entonces podemos considerar el conjunto de elementos posibles de E que se pueden obtener mediante combinaciones lineales de estos elementos generadores, que llamaremos V .

$x \in E$ si $x = v'\alpha$ donde $v = (v_1, v_{2,}..., v_n)'$, $\alpha = (\alpha_1, \alpha_2, ..., \alpha_n)'$ $(\forall i : \alpha_i \in \mathrm{R})$

Este conjunto de elementos V, es un subespacio vectorial, pues cumple las siguientes propiedades:

1) El subespacio contiene el elemento nulo 0.
2) Sean $x, y \in V$ entonces $x + y \in V$
3) Sean $\lambda \in \mathrm{R}$ y $x \in V$, entonces $\lambda x \in V.$

DEFINICIÓN 2.25:

Sea P un operador y $v_1, v_{2,}..., v_n$ elementos del espacio vectorial E que define un subespacio vectorial V. Entonces P es un operador de proyección ortogonal hacia V si cumple que:

1) $\forall z \in E$, $Pz \in V$. Es decir, para todo elemento z del espacio E, su proyección Pz es uno de los elementos del subespacio V:

$$Pz = v'c, \text{ para algún } c \in R^n$$

2) El proyector P aplicado a un elemento del subespacio vectorial da como resultado el mismo elemento y:

$$\forall y \in V, \ Py = y$$

3) $\forall z$, el residuo $z - Pz$ es ortogonal a todo elemento $v \in V$:

$$< v'\alpha, z - Pz >= 0, \forall \alpha \in R^n$$

En resumen, con Pz obtenemos el elemento de V que está más cerca de z.

PROPOSICIÓN: La demostración se la omite y se puede encontrar en [1].

Sea:

z Elemento del espacio vectorial E

$v = (v_1, v_{2,}, ..., v_n)$ vector que contiene los elementos generadores del subespacio vectorial de V. Entonces la proyección ortogonal de z sobre el subespacio V se halla de la siguiente manera:

$$Pz = v'K^{+} < v, z >,$$

donde:

$K = vv' = \left(k_{ij}\right)_{ij} = \left(< v_i, v_j >\right)_{ij}$ matriz nxn de productos escalares de elementos que generan V,

$< v, z >$ es el producto escalar entre elementos del espacio .

$K^{+} =$ Inversa generalizada de K $(KK^{+}K = K)$

2.2.2 ELEMENTOS PROPIOS

DEFINICIÓN 2.26:

Sea Π un operador lineal definido en un espacio E con producto escalar (un endomorfismo),

$\Pi : E \to E$

$e \to g = \Pi e$,

que cumple:

Π *es un operador definido no negativo* $< e, \Pi e > \geq 0$ y,

Π *es un operador simétrico* $< e, \Pi f > = < \Pi e, f > = < f, \Pi e >$.

Diremos que u es un *elemento propio de valor propio* λ respecto de Π si se cumple que:

$$\Pi u = \lambda u.$$

El conjunto de todos los valores propios respecto a Π forman una base que posee las propiedades que nos interesan para el análisis de componentes principales.

PROPOSICIÓN: La demostración se la omite y se puede encontrar en [1].

Sea Π un operador lineal endomórfico y simétrico y sea x un elemento del espacio E. Sea el problema de maximización con restricción siguiente:

$$Max < x, \Pi x > \text{ sujeto a } \| x \| = 1,$$

Entonces, la solución del problema es el elemento propio u de Π con mayor valor propio λ.

2.2.3 EL PROBLEMA DEL ANÁLISIS DE COMPONENTES PRINCIPALES MEDIANTE EL PRODUCTO ESCALAR

Sea la representación

$$\hat{x_i} = \sum_{k=1}^{K} f_{i,k}\,\xi_k$$

Queremos aproximar el valor del elemento x_i a través de la combinación lineal anterior. Es decir, que la diferencia entre el elemento y su aproximación sea mínima. Supongamos que $k=1$. Veremos entonces que se trata de resolver un problema de maximización de la covarianza sujeto a unas restricciones:

$$\min_{\xi\in E}\left\{\min_{f_E\in \mathrm{R}} \sum_{i=1}^{n} < (x_i - f_i\xi),(x_i - f_i\xi) > \right\} =$$

$$= \min_{\xi\in E}\left\{\min_{f_E\in \mathrm{R}} \sum_{i=1}^{n} < x_i, x_i > + f_i^2 < \xi,\xi > -2f_i < x_i,\xi > \right\} =$$

(derivando e igualando a cero: $< \xi,\xi >= 1$ y $2f_i = 2 < x_i,\xi > \Rightarrow f_i =< x_i,\xi >$)

$$= \min_{\xi\in E}\left\{\sum_{i=1}^{n} < x_i, x_i > + < x_i,\xi >^2 -2 < x_i,\xi >^2 \right\}$$

$$= \min_{\xi\in E}\left\{\sum_{i=1}^{n} < x_i, x_i > - < x_i,\xi >^2 \right\} \Leftrightarrow$$

$$\Leftrightarrow \max_{\xi\in E} \sum_{i=1}^{n} < x_i,\xi >^2$$

Así pues, el problema del ACPPE se reduce a un problema de maximización de la función de covarianza, sujeto a restricciones de normal igual a uno y ortogonalidad cuya solución es el cálculo de elementos propios.

PROPOSICIÓN: La demostración se la omite y se puede encontrar en [1].

Sea Π un operador lineal endómorfico y simétrico y sea x un elemento del espacio E. Sean los problemas sucesivos siguientes:

1) Sea U_1 el espacio generado por el elemento u_1 solución del problema:

   $Max < x, \Pi x >$ sujeto a $\| x \| = 1$

2) Sea U_2 el espacio generado por el elemento u_2 solución del problema:

   $Max < x, \Pi x >$ sujeto a $\| x \| = 1$ y $< x, u_1 >= 0$.

3) Sea U_i el espacio generado por el elemento u_i solución del problema:

$Max\langle x, \Pi i \rangle$ sujeto a $\| x \| = 1$ y $< x, u_j >= 0$ para $j < i.$

Entonces la solución $u_1, .. u_i, .. u_n$ de los problemas sucesivos anteriores corresponde a los elementos propios del operador lineal.

COROLARIO:

La solución de los problemas sucesivos anteriores utilizando como operador lineal el operador covarianza corresponde a los elementos propios de dicho operador, y por tanto solucionan el problema ACPPE.

2.3 ANÁLISIS DE COMPONENTES PRINCIPALES FUNCIONALES

Una vez resuelto el problema del ACPPE, hemos de adaptar la notación al espacio vectorial L^2. Para ello definiremos en primer lugar algunos conceptos y posteriormente resolveremos el problema del análisis de componentes principales funcionales (ACPF).

El ACPF pretende explicar el conjunto de funciones de la muestra a partir de unas pocas funciones:

$$\widehat{x_i}(t) = \sum_{k=1}^{K} f_{i,k}\, \xi_k(t)$$

Pero, ¿cuáles son estas funciones?. Tal como se deduce del apartado anterior se puede caracterizar cualquier función como combinación lineal de las funciones propias. El problema de calcular valores y vectores propios, o bien valores y funciones propias, no es más que la aplicación del problema general de buscar elementos propios en un espacio donde se trabaja con vectores en R^n o con funciones L^2 respectivamente.

2.3.1 INTRODUCCIÓN Y DEFINICIONES EN EL ESPACIO FUNCIONAL L^2

Recordemos el planteamiento en el análisis de componentes principales. Sea X_{nxk} una matriz que representa una muestra o una población de tamaño n representada en vectores de dimensión k. Queremos realizar un cambio de base que nos permita explicar en pocas componentes la máxima información (variabilidad) existente.

$$A : \mathrm{R}^n \to \mathrm{R}^n$$
$$x \to y = Ax$$

Lo que busca el ACP clásico es buscar un cambio de base. Ese cambio de base pretende que la proyección de los puntos en ese nuevo eje contenga la máxima varianza posible. Posteriormente vuelve a buscar lo mismo para el espacio ortogonal al primer eje, y así sucesivamente.

El análisis de componentes principales funcional (ACPF) pretende lo mismo: buscar un cambio de base de manera que en la primera componente se explique la máxima variabilidad, en la segunda la máxima variabilidad del espacio ortogonal a la primera, etc. La única novedad es que ahora nuestra base actual contiene elementos que son funciones, en vez de vectores.

DEFINICIÓN 2.27:

Sea $\Pi : \mathrm{R}^2 \to \mathrm{R}$. Dada $f \in L^2$, la función Π permite definir otra función de L^2, a la que llamaremos $\Pi(f)$:

$$\Pi(f) : \mathrm{R}^2 \to \mathrm{R}$$
$$s :\to \Pi(f)(s) = \int_{\mathrm{R}} \Pi(s,t) f(t) dt$$

Así, Π define un endomorfismo en L^2, al que también llamaremos Π:

$$\Pi : L^2 \to L^2$$
$$f \to \Pi(f)$$

PROPOSICIÓN:

El anterior endomorfismo es lineal.

DEMOSTRACIÓN:

Sean $\alpha, \beta \in \mathrm{R}$, f y $g \in L^2$. Sea $s \in \mathrm{R}$. Entonces,

$$\Pi(\alpha f + \beta g)(s) = \int_{\mathrm{R}} \Pi(s,t)(\alpha f(t) + \beta g(t))dt =$$

$$= \alpha \int_{\mathrm{R}} \Pi(s,t) f(t)dt + \beta \int_{\mathrm{R}} \Pi(s,t) g(t)dt =$$

$$= \alpha \Pi(f)(s) + \beta \Pi(g)(s)$$

como queríamos demostrar.

OBSERVACIÓN:

- Podríamos considerar que Π es una matriz de infinitas filas e infinitas columnas (una para cada real).
- La función covarianza muestral permite definir un operador lineal endomórfico en L^2.

DEFINICIÓN 2.28:

Sea Π un operador lineal definido en el espacio L^2 (un endomorfismo),

$$\Pi : L^2 \to L^2$$

$$f \to g = \Pi f$$

Entonces diremos que Π *es un operador definido positivo* si

$$< f, \Pi f >\geq 0 \quad \forall f \text{ y } < f, \Pi f >= 0 \Leftrightarrow f = 0.$$

DEFINICIÓN 2.29:

Sea Π un operador lineal definido positivo en el espacio L^2 (un endomorfismo), entonces Π es *simétrico* si:

$$\langle f, \Pi g \rangle = \langle \Pi f, g \rangle = \langle g, \Pi f \rangle$$

DEFINICIÓN 2.30:

Sea M el operador lineal endomórfico:

$$M : L^2 \to L^2$$

$$f \to g = Mf$$

Entonces diremos que f *es una función propia sobre* M si $\exists \lambda \in \mathrm{R}$ tal que:

$$Mf = \lambda f$$

2.3.2 EL PROBLEMA DEL ANÁLISIS DE COMPONENTES FUNCIONALES

La mejor manera para interpretar qué es lo que hace ACPF es la de explicarlo como una expansión del análisis de componentes principales (ACP) con datos multivariantes.

- En el caso del ACP tenemos que: $X = (x_1, x_{2,\ldots\ldots,} x_p)$ es una muestra de vectores en R^p centrados. Es decir, que el vector media de toda la matriz X resulta ser el vector nulo. Con el ACP buscamos expresar las funciones como combinación lineal de unas variables llamadas *componentes principales*. Es por eso que queremos buscar los coeficientes de cada x_p

$$f_{i,1} = \sum_{j=1}^{p} \xi_{j,1} x_{i,j} =< \xi_{1,} x_i > i = 1,...n$$

  tal que $\frac{1}{n}\sum_{i=1}^{n} f_{i,j}^2 = \frac{1}{n}\sum_{i=1}^{n} \left\langle \xi_1, x_i \right\rangle^2$ sea máximo sujeto a la restricción de que la norma del vector ξ_1 sea igual a uno. Es decir que buscamos una dirección ξ_1 de la nube de puntos tal que la varianza de los puntos proyectados sobre ésta sea la máxima entre todas las direcciones posibles. Del espacio ortogonal restante podemos volver a realizar la misma operación y así sucesivamente hasta obtener la base $\xi_{1,}\xi_{2,}\ldots\ldots,\xi_p$.

- Sean $x_1(t),\ x_2(t),\ldots,x_n(t)$ funciones centradas (su integral en R es cero) y definidas en L^2. Estas funciones definen un espacio que tiene dimensión n. Con el ACPF buscamos para cada una de las funciones un valor resultado de aplicar el producto escalar general al espacio L^2: hemos de adaptar a la naturaleza del espacio la definición de producto escalar. Es por eso que si antes x_i e ξ_j eran vectores y el producto escalar se expresaba como un sumatorio, ahora x_i e ξ_j serán funciones y el producto escalar se expresará como una integral:

$$f_{i,1} = \left\langle \xi_1, x_i \right\rangle = \int_{-\infty}^{+\infty} \xi_1(s) x_i(s) ds$$

  Y buscaremos la ξ_i, tal que $\frac{1}{n}\sum_{i=1}^{n} f_{i,1}^2 = \frac{1}{n}\sum_{i=1}^{n} \left\langle \xi_1, x_i \right\rangle^2$ sea máximo sujeto a que $\left\|\xi_1\right\| = \int_{-\infty}^{+\infty} \xi_1(s) ds = 1$.

Del espacio ortogonal restante podemos volver a realizar la misma operación y así sucesivamente hasta obtener la base $\xi_{1,}\xi_2,....,\xi_p$. Formalizando, lo que queremos resolver es esta sucesión de problemas de maximización continua con restricciones (problema ACPF):

- *problema* $(1)\left\{\begin{matrix} Max_{\xi_1\in L^2}\ \sum_{i=1}^{n}\left\langle \xi_1, x_i\right\rangle^2 \\ \text{sujeto a } \left\|\xi_1\right\| = 1 \end{matrix}\right\}$
- *problema* $(2)\left\{\begin{matrix} Max_{\xi_2\in L^2}\ \sum_{i=1}^{n}\left\langle \xi_2, x_i\right\rangle^2 \\ \text{sujeto a } \left\|\xi_2\right\| = 1 \text{ y } \left\langle \xi_{1,}\xi_2\right\rangle = 0 \end{matrix}\right\}$
- *problema* $(j)\left\{\begin{matrix} Max_{\xi_j\in L^2}\ \sum_{i=1}^{n}\left\langle \xi_j, x_i\right\rangle^2 \\ sujetoa\left\|\xi_j\right\| = 1 \text{ y } \left\langle \xi_{k,}\xi_j\right\rangle = 0 \text{ para } k<j \end{matrix}\right\}$

PROPOSICIÓN:

Sea el problema ACPF y sean $x_1(t), x_2(t),..., x_n(t)$ funciones definidas en R. Sea además $v(s,t) = \frac{1}{n}\sum_{i=1}^{n} x_i(s)x_i(t),$ donde

$V \ : \ L^2(\mathrm{R},\mathrm{R}) \to \mathrm{R}$

$(s,t) \to v(s,t)$

y donde,

$v(s,t) : \mathrm{R} \to \mathrm{R}$

$s \to \int_{\mathrm{R}} v(s,t)y(t)dt,$ que es función de s .

Entonces la solución del problema ACPF es la solución de la ecuación propia siguiente:

$$\int_{\mathrm{R}} v(s,t)\xi(t)dt = \left\langle v(s,\cdot),\xi\right\rangle = \lambda\xi(s)$$

que corresponde al vector de las funciones propias de $x_1(t),\ x_2(t),....,x_n(t)$ sobre el operador covarianza $v(s,t).$

2.3.3 INTERPRETACIÓN

Como resultado del problema de maximización anteriormente planteado obtenemos la ecuación:

$$\widehat{x_i}(t) = \sum_{k=1}^{K} f_{i,k}\, \xi_k(t),$$

donde los coeficientes son $f_{i,k}$ y las funciones componentes principales son $\xi_i(t)$. Con estas últimas podremos explicar resumidamente cada x_i.

Como hemos dicho antes queremos representar x_i a partir de las mínimas funciones posibles. Para saber cuál es el porcentaje de representación con un número de componentes en concreto utilizamos los valores propios.

- En ACP: La suma de los valores propios corresponden a la traza de la matriz de varianzas-covarianzas, que es la varianza total.

  $traza(V) = \sum_{j=1}^{p} V(x_j) = \sum_{j=1}^{p} \sigma_j^2 =$

  $= \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} (x_{i,j} - \bar{x}_j)^2 =$

  $= \frac{1}{n} \sum_i \sum_j (x_{i,j} - \bar{x}_j) = \sum_j^p \lambda_j$

- En ACPF: De la misma manera que en ACP, la suma de los valores propios corresponde a la varianza total.

  $\int_{\mathrm{R}} v(t,t)dt = \frac{1}{n} \sum_{i=1}^{n} \left(\int_{\mathrm{R}} x_i(t)^2 dt \right) = \sum_{j=1}^{n} \lambda_j$

DEFINICIÓN 2.31:

Sea l el número de primeras componentes con las que queremos representar x_i, entonces el *porcentaje de representación* será el siguiente:

$$\%representacion = 100 \times \frac{\sum_{i=1}^{l} \lambda_i}{\sum_{i=1}^{n} \lambda_i}$$

Uno de los objetivos del ACPF es el de poder representar como puntos en el espacio de dos o tres dimensiones la totalidad de la muestra. Las coordenadas de los puntos serían los coeficientes $\left(f_{i,1}, f_{i,2}, f_{i,3}\right)$ y los ejes representan las direcciones de crecimiento de las funciones propias $\xi_1(t)$, $\xi_2(t)$ y $\xi_3(t)$. Para hacer una interpretación completa que sitúe en el espacio cada una de las x_i's

hemos de interpretar a qué pauta de comportamiento corresponde cada una de las componentes principales $\xi_i(t)$. En este caso no podemos utilizar el mismo método que en ACP ya que no tenemos representación de las componentes en función de las p variables, que en este caso sería $p=\infty$. Lo que hacemos en este caso para interpretar las componentes principales funcionales es representar la función media juntamente con la función media $\pm C\xi(t)$, donde C es una constante que facilita la interpretación del gráfico.

2.3.4 ACPF PARA FUNCIONES REPRESENTADAS A TRAVÉS DE FUNCIONES BASE CONOCIDAS

En muchas ocasiones nos vamos a encontrar con que nuestras funciones están expresadas como una combinación lineal de una base de funciones. Para este caso en particular tenemos una adaptación del ACPF.

Sea $\phi_1(t), \phi_2(t),...,\phi_k(t)$ una base de funciones, entonces tendremos representadas nuestras funciones de la siguiente manera:

$$x_i(t) = \sum_{i=1}^{k} c_{i,k}\phi_k(t)$$

En forma matricial estaríamos diciendo lo siguiente: $x = C\phi$ donde x y ϕ son vectores de funciones y C es una matriz $N \times K$ que contiene los coeficientes de la combinación lineal.

La función de covarianza teniendo en cuenta esta representación sería:

$$v(s,t) = \frac{1}{n}\phi(s)C'C\phi(t)$$

La ecuación propia a resolver es:

$$\int_{\mathrm{R}} v(s,t)\xi(t)dt = \langle v(s,\cdot),\xi\rangle = \lambda\xi(s)$$

Supondremos también que las funciones propias pueden ser expresadas en la base anterior. Es decir, $\xi_i(s) = \sum_{k=1}^{K} b_k\phi_k(s)$

Considerando la expresión de la varianza actual:

$\int_{\mathrm{R}} v(s,t)\xi(t)dt = \int_{\mathrm{R}} \frac{1}{n}\phi'(s)C'C\phi(t)\phi'(t)bdt = \phi'(s)N^{-1}C'CWb$

Con lo que la ecuación propia que queda por resolver es:

$$\phi'(s)N^{-1}C'CWb = \lambda b$$

Donde $W = (w_{k_1,k_2})_{k_1,k_2} = (\int_{\mathrm{R}} \phi_{k_1}(t)\phi_{k_2}(t)dt)_{k_1,k_2}$

Como esto es para todo $s \in \mathrm{R}$ entonces tendremos que todo se reduce a calcular los vectores propios de una matriz A:

$$N^{-1}C'CWb = \lambda b \Leftrightarrow Ab = \lambda b$$

3 ESTIMACIÓN DE FUNCIONES POR SPLINES

3.1 REPRESENTACIÓN DE DATOS E INTERPOLACIÓN

Para el análisis de datos funcionales teórico es preciso conocer la forma explícita de cada una de las funciones. Sin embargo en la mayoría de las situaciones reales no conoceremos esta forma. Es más, como resultado de nuestro experimento o de la observación vamos a obtener simplemente los datos observados de la función a lo largo de su eje de ordenadas. Podemos decir que en general esta será nuestra situación.

SITUACIÓN:

- Sea $f_1(x), f_2(x),..., f_N(x)$ una muestra de funciones aleatorias.
- Para cada $f_i(x)$ habremos evaluado la función obteniendo $y_{i1}, y_{i2},..., y_{in}$ como imagen de la misma evaluada en la serie de valores $x_{i1}, x_{i2,}....., x_{in}$.

Ante esta situación nos tenemos que plantear cómo transformar estos datos en una buena estimación de la función observada. Desgraciadamente a la hora de tomar estos valores también podemos cometer errores: error observacional. Por lo tanto habremos de considerar para la estimación de la función si el error observacional puede ser o no obviado.

Existen varios métodos de conseguir funciones a partir de esta situación: el polinomio inteporlador de Lagrange, splines interpoladores (Bonet) o regresión no paramétrica (Lange K.), entre otros. En todos ellos el planteamiento de la situación será el siguiente:

PROBLEMA:

Convertir los valores $y_{i1,} y_{i2},....., y_{in}$ de alguna manera eficiente en una función $f_i(x)$ tal que $y = f_i(x_{i,j}) + \varepsilon_{i,j}$, donde $\varepsilon_{i,j}$ es un error observacional aleatorio con $E\left[\varepsilon_{i,j}\right] = 0$

Para la situación en que sabemos que el error observacional es despreciable usaremos el spline interpolador, por las buenas propiedades que tiene. Mientras que en el caso en que tengamos que tener en cuenta el error observacional o de medida se puede usar la regresión no paramétrica por splines. Este capítulo se fundamenta principalmente en [12].

Los splines se pueden definir de forma genérica y luego en particular se puede definir una clase de splines. La definición genérica es la siguiente:

DEFINICIÓN 3.1:

Sea:

- $S(x)$ una función definida en $[a,b]$,
- $a = x_0 < x_1 < .... < x_n = b$ conjunto de puntos ordenados.

Entonces se dice que $S(x)$ es un *spline de grado p y nodos* $x_0 < x_1 < \ldots < x_n$ si se verifica que:

a) $S(x)$ es un polinomio de grado menor o igual a p en cada intervalo $[x_i, x_{i+1}]$,

b) La función $S(x)$ tiene derivadas hasta de orden $(p-1)$ continuas en $[a,b]$.

Cuando tenemos un gran número de evaluaciones de la función, herramientas como el polinomio interpolador suelen dar problemas, ya que acaban dando como solución polinomios de muy alto grado, que suelen ser muy poco "suaves". Entre todas las posibilidades (Polinomios de Lagrange, Interpolación de Hermite, etc...) nos dedicaremos al uso de splines, ya que son la función interpoladora que minimiza la integral de la segunda derivada al cuadrado entre todas las funciones de L^2, tal como describen P.J Green y B.W.Silverman $[2]$

Es por la propiedad anterior por la que los splines interpoladores dan como resultado estimaciones de la función más "suave", y no precisa de polinomios de alto grado. En particular son los splines de tercer grado los que consideraremos por sus buenas propiedades.

3.2 SPLINES CÚBICOS

Una spline es una función definida a trozos sobre intervalos de $\mathbb{R}$ que se unen entre si obedeciendo a ciertas condiciones de regularidad. La terminología fue introducida por I. J. Schoenberg (1946).

El nombre de spline proviene del nombre del instrumento mecánico del mismo nombre que consiste en un alambre flexible que puede ser utilizado para dibujar curvas suaves a través de puntos asignados. Esta clase de instrumentos fueron utilizados para dibujo técnico en las industrias aeronáuticas, automotriz, naval, etc.

DEFINICIÓN 3.2:

Sea

- $x_0 < x_1 < ...x_n$ conjunto de puntos ordenados (llamados también knots).
- $f_0, f_1, \ldots, f_n$ valores de la función evaluada en los puntos anteriores:

  $f_i = f(x_i)$, $i = 0,..., n.$

Entonces el *spline cúbico interpolador* $S(x)$ es una función definida en el intervalo $[x_{0,} x_n]$ con las siguientes propiedades:

- $S(x)$ es un polinomio cúbico en cada intervalo $[x_{j,} x_{j+1}]$.
- $S(x_i) = f_i$ en cada nodo x_i.
- La segunda derivada $S''(x)$ existe y es continua a lo largo del intervalo $[x_{0,} x_n].$
- En los nodos extremos $S''(x_0) = S''(x_n) = 0.$

PROPOSICIÓN:

Existe exactamente un único spline en $[x_{0,} x_n]$ que satisfaga las propiedades anteriores.

DEMOSTRACIÓN (intuitiva):

De la definición anterior podemos deducir que un spline cúbico es la solución de un sistema de ecuaciones. Si el sistema de ecuaciones es compatible determinado entonces habremos demostrado que es único.

Tenemos que $S(x)$ es:

$S_0(x) = S_{0,1} + S_{0,2}x + S_{0,3}x^2 + S_{0,4}x^3$ para el intervalo $[x_0, x_1]$

$S_1(x) = S_{1,1} + S_{1,2}x + S_{1,3}x^2 + S_{1,4}x^3$ para el intervalo $[x_1, x_2]$

...

$S_{n-1}(x) = S_{n-1,1} + S_{n-1,2}x + S_{n-1,3}x^2 + S_{n-1,4}x^3$ para el intervalo $[x_{n-1}, x_n]$

Por lo que tenemos que calcular $4n$ coeficientes.

Sobre la base de la definición encontramos que tiene que haber una serie de restricciones.

- Restricciones de interpolación: El spline ha de pasar por los puntos (x_i, f_i)

  $S_0(x_0) = f_0, \quad S_0(x_1) = f_1 \quad ... \quad S_{n-1}(x_n) = f_n$ ($n+1$ restricciones)

- Restricciones de continuidad:

  $S_0(x_1) = S_1(x_1),\ S_1(x_2) = S_2(x_2) \quad ... \quad S_{n-2}(x_{n-1}) = S_{n-1}(x_{n-1})$ ($n-1$ restricciones)

- Restricciones de derivada continua:

  $S_0'(x_1) = S_1'(x_1),\ S_1'(x_2) = S_2'(x_2) \quad ... \quad S_{n-2}'(x_{n-1}) = S_{n-1}'(x_{n-1})$ ($n-1$ restricciones)

- Restricciones de segunda derivada continua:

  $S_0''(x_1) = S_1''(x_1) \quad S_1''(x_2) = S_2''(x_2) \quad ... \quad S_{n-2}''(x_{n-1}) = S_{n-1}''(x_{n-1})$ ($n-1$ restricciones)

- Restricciones de punto extremo (el spline se convierte en rectas fuera de $[x_0, x_n]$):

  $S_0''(x_0) = 0$ y $S_{n-1}''(x_n) = 0$ (2 restricciones)

Por lo tanto tenemos un sistema de ecuaciones con $4n$ incógnitas y $4n$ ecuaciones, lo que lo convierte en un sistema compatible determinado. Por tanto tiene una única solución, como queríamos demostrar.

Otra de las ventajas de los splines es la de que pertenece al espacio L^2 . Además entre todas las funciones de ese espacio el spline es el que minimiza la siguiente cantidad: $\int_R (f''(x))^2\, dx$, que es una medida global de la curvatura de la función. Las funciones que oscilan mucho entre los puntos a interpolar tienen pendientes muy cambiantes y segundas derivadas muy altas. Esto se traduce en lo que gráficamente llamamos ser una función "poco suave" (con altibajos bruscos).

PROPOSICIÓN: (P.J.Green & B.W.Silverman)

Sea $S(x)$ el spline interpolador de una función $f(x)$ en los nodos $x_0 < x_1 < ... < x_n$. Si $g(x)$ es cualquier otra función dos veces diferenciable continua e interpoladora de $f(x)$ en ese nodo, entonces:

$$\int_{x_0}^{x_n} g''(x)^2\,dx \geq \int_{x_0}^{x_n} s''(x)^2\,dx$$

con igualdad si y sólo si $g(x) = f(x)$

Podemos decir que el spline interpolador es la función ideal para utilizar tanto por pertenecer al espacio L^2 como por poseer de entre todas las funciones que pertenecen a L^2, la propiedad de ser la más "suave".

ESPACIO DE FUNCIONES SPLINES

Sean $n \in \mathbb{N}.$ Se llama conjunto de nodos un conjunto de puntos $\tau(n) = \{x_j\}_{j=0,\ldots,n}$, donde $a = x_0 < x_1 < \cdots < x_n = b.$ Estos nodos forman una partición del intervalo $[a,b] \subset \mathbb{R}$ en subintervalos $[x_{j-1}, x_j]$, $j = 1,\ldots,n.$ Los puntos $x_1,\ldots,x_{n-1}$ se llaman nodos interiores, y los puntos $x_0 = a$ y $x_n = b$ se llaman nodos frontera.

Un conjunto de puntos $S_n = \{(x_i, y_i) \mid x_i \in \tau(n),\ y_i \in \mathbb{R},\ i = 0,1,\ldots,n\},$ se llama conjunto de puntos de base.

Notamos con P_m el espacio de polinomios de grado $\leq m.$ Se designa con $C^{-1}([a,b])$ el espacio de funciones continuas a trozos en $[a,b]$. Se denota con $C^{k-1}([a,b])$, $a \leq k \leq m,$ el espacio de funciones que poseen derivadas continuas hasta el orden $k-1$ en $[a,b]$ ($C^0([a,b]) = C([a,b])$ es el espacio de las funciones$[a,b]$).

DEFINICIÓN 3.3:

Sea $m \in \mathbb{N}$. Una función $S : [a,b] \to \mathbb{R}$ se llama función spline polinomial de grado m si ella posee las propiedades siguientes:

- $S \in C^{m-1}([a,b]);$
- $S \in P_m$ para $x \in [x_{j-1}, x_j]$, $j = 1,...,n$

Denotamos con $S_m(\tau(n))$ el conjunto de todas las funciones splines polinomiales de grado m asociadas a la subdivisión $\tau(n)$ de $[a,b]$.

En lo sucesivo nos limitaremos a los splines polinomiales y nos referiremos a ellas simplemente como splines.

BASE DE $S_m(\tau(n))$

El conjunto $S_m(\tau(n))$ provisto de las operaciones habituales entre funciones (adición y producto de un número real por una función) es un espacio vectorial real de dimensión $m+n$ y una base de dicho espacio es el conjunto de funciones

$$\{p_0, p_1, \ldots, p_m, q_{m,1}, \ldots, q_{m,n-1}\},$$

donde

$$p_i(x) = x^i, \quad i = 0,1,\ldots,m,$$

y la familia de funciones $\{q_{m,j} \mid j = 0,\ldots,n-1\}$ definidas como sigue:

$$q_{m,j}(x) = \begin{cases} (x - x_j)^m, & \text{si } x \in [x_j, b], \\ 0, & \text{si } x \in [a, x_j[, \end{cases}$$

son splines de grado m asociadas a la subdivisión $\tau(n)$ del intervalo $[a,b]$ y $m \in \mathbb{N}$.

Se puede probar que toda función $S \in S_m(\tau(n))$ se escribe de manera única en la forma

$$S(x) = a_0 + \sum_{i=1}^{m} a_i x^j + \sum_{j=1}^{n-1} b_j q_{m,j}(x) \quad x \in [a,b],$$

donde $a_0, a_i, b_j \in \mathbb{R}, \quad i = 1,\ldots,m; \quad j = 2,\ldots,m.$

3.3 CÁLCULO DE SPLINES CÚBICOS

Consideremos $f \in C^2([a,b]), \quad \tau(n)$ una subdivisión de $[a,b]$ y

$$S_n = \{(x_i, f(x_i)) \mid x_i \in \tau(n), \quad i = 0,\ldots,n\},$$

un conjunto de puntos de base.

Puesto que $\dim S_3(\tau(n)) = n+3,$ si se requiere interpolar en cada uno de los $n+1$ nodos $x_0,\ldots,x_n,$ entonces quedan 2 parámetros libres que pueden ser utilizados en los tipos de splines siguientes:

a) Interpolación con condiciones de frontera de Hermite.

Hallar $S \in S_3(\tau(n))$ tal que

i. $S(x_j) = f(x_j), \quad j = 0,1,\ldots,n.$

ii. $S'(a) = f'(a).$

iii. $S'(b) = f'(b).$

b) Interpolación con condiciones de frontera naturales.

Suponemos que $n \geq 2.$

Hallar $S \in S_3(\tau(n))$ tal que

i. $S(x_j) = f(x_j), \quad j = 0,1,\ldots,n.$

ii. $S''(a) = S''(b) = 0.$

c) Interpolación con condiciones de frontera periódicas

$(f(a) = f(b)$ y $f'(a) = f'(b)).$

Hallar $S \in S_3(\tau(n))$ tal que

i. $S(x_j) = f(x_j), \quad j = 0,1,\ldots,n.$

ii. $S'(a) = S'(b).$

iii. $S''(a) = S''(b).$

Con el propósito de mostrar que los problemas a), b) y c) tienen solución única, enunciamos la propiedad siguiente de las splines cúbicas, conocida como relación integral.

RELACIÓN INTEGRAL

Sea $f \in C^2([a,b])$ y $S \in S_3(\tau(n))$ una función spline de interpolación de f tal que la diferencia $E(x) = f(x) - S(x)$ $x \in [a,b]$, satisface la condición de frontera

$$S''(a)E'(a) = S''(b)E'(b).$$

Entonces

$$\int_a^b [f''(x)]^2 dx = \int_a^b [f''(x) - S''(x)]^2 dx + \int_a^b [S''(x)]^2 dx.$$

De esta relación, vemos que

i. Si $E'(a) = E'(b) = 0,$ entonces se tiene las splines del tipo a).

ii. Si $S''(a) = S''(b) = 0,$ entonces se tiene las splines del tipo b).

iii. Si $S''(a)=S''(b)$ y $E'(a)=E'(b)$, corresponden entonces a las splines del tipo c).

Usando la relación integral se prueba que los problemas de interpolación a), b) y c) tienen siempre una única solución $S \in S_3(\tau(n))$.

CONSTRUCCIÓN

Dada una función $f \in C^2([a,b])$, para construir la función spline de interpolación S, aplicamos las condiciones de las funciones splines y de las splines cúbicas al polinomio cúbico siguiente:

$$S_j(x)=a_j+b_j(x-x_j)+c_j(x-x_j)^2+d_j(x-x_j)^3, \quad x\in[x_j,x_{j+1}], \quad j=0,1,\ldots,n-1,$$
$$S(x)=S_j(x), \quad x\in[x_j,x_{j+1}], \quad j=0,1,\ldots,n-1.$$

Por i) de los tipos de splines a), b) y c), se tiene

$$S_j(x_j)=a_j=f(x_j), \quad j=0,1,\ldots,n-1,$$

y ponemos $a_n=f(x_n)$.

De la definición de función spline (continuidad en cada nodo), se obtiene

$$\begin{aligned} a_{j+1} &= S_{j+1}(x_{j+1})=S_j(x_{j+1}) \\ &= a_j+b_j(x_{j+1}-x_j)+c_j(x_{j+1}-x_j)^2+d_j(x_{j+1}-x_j)^3, \end{aligned}$$

para $j=0,1,\ldots,n-1$.

Notamos con $h_j=x_{j+1}-x_j$, $j=0,1,\ldots,n-1$. Entonces la relación precedente se escribe

$$a_{j+1}=a_j+b_jh_j+c_jh_j^2+d_jh_j^3.$$

La derivada de $S_j(x)$ es la función

$$S_j'(x)=b_j+2c_j(x-x_j)+3d_j(x-x_j)^2, \quad x\in[x_j,x_{j+1}], \quad j=0,1,\ldots,n-1,$$

de donde

$$S_j'(x_j)=b_j, \quad j=0,1,\ldots,n-1.$$

Definimos $b_n=S'(x_n)$.

Por la continuidad de S_j' en cada nodo x_j, tenemos

$$b_{j+1}=S_{j+1}'(x_{j+1})=S_j'(x_{j+1})=b_j+2c_jh_j+3d_jh_j^2, \quad j=0,1,\ldots,n-1.$$

La derivada segunda de $S_j(x)$ está dada por

$$S_j''(x) = 2c_j + 6d_j(x - x_j), \quad x \in [x_j, x_{j+1}], \quad j = 0,1,\ldots,n-1,$$

de donde

$$S_j''(x_j) = 2c_j, \quad j = 0,1,\ldots,n-1,$$

y definimos $c_n = \frac{1}{2}S''(x_n)$.

Nuevamente, utilizando la continuidad de $S_j''(x)$ en cada nodo x_j, tenemos:

$$c_{j+1} = \frac{1}{2}S_{j+1}''(x_{j+1}) = \frac{1}{2}S_j''(x_{j+1}) = \frac{1}{2}(2c_j + 6d_j h_j) = c_j + 3d_j h_j, \quad j = 0,1,\ldots,n.$$

Obtengamos relaciones que liguen los coeficientes b_j, c_j, d_j en términos de los datos $a_j = f(x_j), \quad j = 0,1,\ldots,n.$

Resulta

$$d_j = \frac{c_{j+1} - c_j}{3h_j},$$

con lo cual

$$a_{j+1} = a_j + b_j h_j + c_j h_j^2 + \frac{c_{j+1} - c_j}{3h_j}h_j^3 = a_j + b_j h_j + \frac{1}{3}(2c_j + c_{j+1})h_j^2,$$

$$b_{j+1} = b_j + 2c_j h_j + 3\frac{c_{j+1} - c_j}{3h_j}h_j^2 = b_j + (c_j + c_{j+1})h_j, \quad j = 0,1,\ldots,n-1.$$

Para obtener la relación final entre los coeficientes, de la igualdad

$$a_{j+1} = a_j + b_j h_j + \frac{1}{3}(2c_j + c_{j+1})h_j^2,$$

obtenemos

$$b_j = \frac{1}{h_j}(a_{j+1} - a_j) - \frac{h_j}{3}(2c_j + c_{j+1}), \quad j = 0,1,\ldots,n-1,$$

y para $j = 0,1,\ldots,n$:

$$b_{j-1} = \frac{1}{h_{j-1}}(a_j - a_{j-1}) - \frac{h_{j-1}}{3}(2c_{j-1} + c_j),$$

con lo cual la relación

$$b_{j+1} = b_j + (c_j + c_{j+1})h_j,$$

se expresa en la forma

$$b_j = b_{j-1} + (c_{j-1} + c_j)h_{j-1}, \quad j = 1,\ldots,n,$$

y

$$\frac{1}{h_j}(a_{j+1}-a_j)-\frac{h_j}{3}(2c_j+c_{j+1})=\frac{1}{h_{j-1}}(a_j-a_{j-1})-\frac{h_{j-1}}{3}(2c_{j-1}+c_j)+(c_{j-1}+c_j)h_{j-1},$$

de donde

$$\frac{1}{h_j}(a_{j+1}-a_j)-\frac{1}{h_{j-1}}(a_j-a_{j-1})=\frac{h_j}{3}(2c_j+c_{j+1})-\frac{h_{j-1}}{3}(2c_{j-1}+c_j)+(c_{j-1}+c_j)h_{j-1}$$

$$=\frac{1}{3}h_{j-1}c_{j-1}+\frac{2}{3}(h_{j-1}+h_j)c_j+\frac{1}{3}h_jc_{j+1},$$

o bien

$$(h_{j-1},2(h_{j-1}+h_j),h_j)\begin{pmatrix}c_{j-1}\\c_j\\c_{j+1}\end{pmatrix}=\frac{3}{h_j}(a_{j+1}-a_j)-\frac{3}{h_{j-1}}(a_j-a_{j-1}),\qquad j=1,\ldots,n-1.$$

Ponemos $\vec{c}^{\,t}=(c_0,c_1,\ldots,c_n).$ El sistema de ecuaciones precedente involucra únicamente el vector $\vec{c},$ las longitudes de los subintervalos $[x_{j-1},x_j],\quad j=1,\ldots,n$ y los valores de f en los puntos $\tau(n)=\{x_j\}_{j=1,\ldots,n}$ de la subdivisión de $[a,b].$

3.3.1 INTERPOLACIÓN CON CONDICIONES DE FRONTERA DE HERMITE

Sea $f\in C^{(4)}([a,b]),$ f tiene una única spline cúbica de interpolación $S\in S_3(\tau(n))$ que satisface las condiciones de frontera de Hermite $S'(a)=f(a)$ y $S'(b)=f'(b).$ En efecto,

$$S'(a)=S'(x_0)=b_0=f'(a),$$

y para $j=0,\quad b_0$ está dado por

$$b_0=\frac{1}{h_0}(a_1-a_0)-\frac{h_0}{3}(2c_0+c_1),$$

resulta que

$$2h_0c_0+h_0c_1=-3f'(a)+\frac{3}{h_0}(a_1-a_0).$$

De manera similar, tenemos

$$S'(b)=S'(x_n)=b_n=f'(b).$$

Como

$$b_n = b_{n-1} + h_{n-1}(c_{n-1} + c_n),$$

y

$$b_{n-1} = \frac{1}{h_{n-1}}(a_n - a_{n-1}) - \frac{h_{n-1}}{3}(2c_{n-1} + c_n),$$

se tiene entonces que

$$\begin{aligned} f'(b) &= \frac{1}{h_{n-1}}(a_n - a_{n-1}) - \frac{h_{n-1}}{3}(2c_{n-1} + c_n) + h_{n-1}(c_{n-1} + c_n) \\ &= \frac{1}{h_{n-1}}(a_n - a_{n-1}) + \frac{h_{n-1}}{3}(c_{n-1} + 2c_n), \end{aligned}$$

con lo cual

$$h_{n-1}c_{n-1} + 2h_{n-1}c_n = 3f'(b) - \frac{3}{h_{n-1}}(a_n - a_{n-1}).$$

En resumen, tenemos que

$$2h_0c_0 + h_0c_1 = -3f'(a) + \frac{3}{h_0}(a_1 - a_0),$$

$$(h_{j-1}, 2(h_{j-1} + h_j), h_j)\begin{pmatrix} c_{j-1} \\ c_j \\ c_{j+1} \end{pmatrix} = \frac{3}{h_j}(a_{j+1} - a_j) - \frac{3}{h_{j+1}}(a_j - a_{j-1}), \quad j = 1, \ldots, n-1,$$

$$h_{n-1}c_{n-1} + 2h_{n-1}c_n = 3f'(b) - \frac{3}{h_{n-1}}(a_n - a_{n-1}),$$

que puede expresarse en forma compacta como un sistema de ecuaciones

$$A\vec{C} = \vec{b},$$

donde

$$A = \begin{pmatrix} 2h_0 & h_0 & 0 & 0 & \cdots & 0 \\ h_0 & 2(h_0 + h_1) & h_1 & 0 & \cdots & 0 \\ 0 & h_1 & 2(h_1 + h_2) & h_2 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & & \\ \vdots & & \ddots & \ddots & \ddots & h_{n-1} \\ 0 & \cdots & \cdots & \cdots & h_{n-1} & 2h_{n-1} \end{pmatrix},$$

$$\vec{b} = \begin{pmatrix} -3f'(a) + \dfrac{3}{h_0}(a_1 - a_0) \\ \dfrac{3}{h_1}(a_2 - a_1) - \dfrac{3}{h_0}(a_1 - a_0) \\ \vdots \\ \dfrac{3}{h_{n-1}}(a_n - a_{n-1}) - \dfrac{3}{h_{n-2}}(a_{n-1} - a_{n-2}) \\ 3f'(b) - \dfrac{3}{h_{n-1}}(a_n - a_{n-1}) \end{pmatrix}.$$

La matriz A es simétrica, estrictamente diagonalmente dominante, por lo tanto el sistema de ecuaciones precedente tiene solución única.

El método de resolución numérica que puede utilizarse es el de factorización LU de Crout o de Doolitle.

Una vez calculados los coeficientes $c_0, c_1, \ldots, c_n,$ los coeficientes b_j se calculan usando la relación

$$b_j = \frac{1}{h_j}(a_{j+1} - a_j) - \frac{h_j}{3}(2c_j + c_{j+1}), \qquad j = 1, \ldots, n-1,$$

y los coeficientes d_j por

$$d_j = \frac{c_{j+1} - c_j}{3h_j}, \qquad j = 0, 1, \ldots, n-1.$$

Finalmente, se define $S(x) = S_j(x), \quad x \in [x_{j-1}, x_j], \quad j = 1, \ldots, n.$

El error de interpolación en la norma $L^\infty(a,b)$ satisface la desigualdad siguiente;

$$\|f - S\|_{L^\infty(a,b)} \leq \frac{5}{384} M h^4,$$

donde $M = \left\|f^{(4)}\right\|_{L^\infty(a,b)}, \quad h = \max_{j=0,1,\ldots,n} h_j.$

Es claro que $\|f - S\|_{L^\infty(a,b)} \underset{h \to 0}{\to} 0;$ es decir que S converge uniformemente a f cuando $h \to 0.$

3.3.2 INTERPOLACIÓN CON CONDICIONES DE FRONTERA NATURALES

Sea $f \in C^{(4)}([a,b]),$ tiene una única spline cúbica de interpolación $S \in S_3(\tau(n))$ que satisface las condiciones de frontera naturales $S''(a) = S''(b) = 0.$ Efectivamente,

$$0 = S''(a) = S''(x_0) = 2c_0 + 6d_0(x_0 - x_0),$$

de donde $c_0 = 0,$

$$c_n = \frac{S''(x_n)}{2} = \frac{S''(b)}{2} = 0.$$

Así, $c_0 = 0,\ c_n = 0.$ Para $j = 1,\ldots,n-1,$ tenemos

$$(h_{j-1}, 2(h_{j-1} + h_j), h_j)\begin{pmatrix} c_{j-1} \\ c_j \\ c_{j+1} \end{pmatrix} = \frac{3}{h_j}(a_{j+1} - a_j) - \frac{3}{h_{j-1}}(a_j - a_{j-1}),$$

que podemos escribir como un sistema de ecuaciones lineales

$$A\vec{C} = \vec{b},$$

donde

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & \ldots & 0 \\ h_0 & 2(h_0 + h_1) & h_1 & 0 & \ldots & 0 \\ 0 & h_1 & 2(h_1 + h_2) & h_2 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & & h_{n-2} & 2(h_{n-2} + h_{n-1}) & h_{n-1} \\ 0 & 0 & \ldots & \ldots & 0 & 1 \end{pmatrix}$$

$$\vec{b} = \begin{pmatrix} 0 \\ \frac{3}{h_1}(a_2 - a_1) - \frac{3}{h_0}(a_1 - a_0) \\ \vdots \\ \frac{3}{h_{n-1}}(a_n - a_{n-1}) - \frac{3}{h_{n-2}}(a_{n-1} - a_{n-2}) \\ 0 \end{pmatrix}.$$

La matriz A es estrictamente diagonalmente dominante. Esto implica que el sistema de ecuaciones precedente tiene solución única. El método numérico de resolución de tal sistema es el de factorización de Crout o de Doolitle.

Sea $\vec{C}^t = (c_0, c_1, \ldots, c_n)$ la solución del sistema de ecuaciones $A\vec{C} = \vec{b}$. Los coeficientes b_j y d_j se calculan usando las fórmulas siguientes:

$$b_j = \frac{a_{j+1} - a_j}{h_j} - \frac{h_j}{3}(c_{j+1} + 2c_j) \quad j = 0,1,\ldots,n-1,$$

$$d_j = \frac{c_{j+1} - c_j}{3h_j} \quad j = 0,1,\ldots,n-1.$$

Note que $b_n = S'(b)$ y $d_n = 0.$

Definimos $S(x) = S_j(x) \quad x \in [x_{j-1}, x_j], \quad j = 1,\ldots,n.$

Se tiene entonces la siguiente estimación de error

$$\|f - S\|_{L^\infty(a,b)} \leq C\|f^{(4)}\|_{L^\infty(a,b)} h^4,$$

donde $C > 0$ es una constante independiente de n y $h = \max_{j=1,\ldots,n} h_j.$

3.3.3 INTERPOLACIÓN CON CONDICIONES DE FRONTERA PERIÓDICAS

Sea $f \in C^4([a,b]),$ f tiene una única spline cúbica de interpolación $S \in S_3(\tau(n))$ que satisface las condiciones de frontera $S'(a) = S'(b), \quad S''(a) = S''(b).$

Mediante un razonamiento similar a los dos casos a) y b), se obtiene el sistema de ecuaciones lineales siguiente:

$$A\vec{C} = \vec{b},$$

donde

$$A = \begin{pmatrix} 2(h_0 - h_{n-1}) & h_0 & 0 & \ldots & -h_{n-1} & 0 \\ h_0 & 2(h_0 + h_1) & h_1 & \ldots & 0 & 0 \\ 0 & h_1 & 2(h_1 + h_2) & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & h_{n-2} & 2(h_{n-2} + h_{n-1}) & h_{n-1} \\ 0 & h_0 & \ldots & 0 & -h_{n-1} & 2(h_0 - h_{n-1}) \end{pmatrix}$$

$$\vec{b} = \begin{pmatrix} \frac{3}{h_{n-1}}(a_n - a_{n-1}) - \frac{3}{h_0}(a_1 - a_0) \\ \frac{3}{h_1}(a_2 - a_1) - \frac{3}{h_0}(a_1 - a_0) \\ \vdots \\ \frac{3}{h_{n-1}}(a_n - a_{n-1}) - \frac{3}{h_{n-2}}(a_{n-1} - a_{n-2}) \\ \frac{3}{h_{n-1}}(a_n - a_{n-1}) - \frac{3}{h_0}(a_1 - a_0) \end{pmatrix}.$$

El valor de $S(x)$ para $x \in [a,b]$ se obtiene de manera análoga a los casos a) y b). El error de interpolación es idéntico al caso b).

3.4 B-SPLINES

En las secciones precedentes construimos los espacios de splines $S_m(\tau(n))$ para una subdivisión dada $\tau(n) = \{x_j\}_{j=0,\ldots,n}$. Estos espacios tienen dimensión $m+n$ y una base de $S_m(\tau(n))$ en la familia de funciones

$$\{p_0, p_1, \ldots, p_m, q_{m,1}, \ldots, q_{m,n}\}.$$

En esta sección discutiremos bases alternativas para espacios de splines mejor adaptadas a los aspectos numéricos. Estas funciones fueron introducidas por Schoenberg y las denominó Curvas básicas de Splines que en la actualidad se conocen simplemente como B-Splines.

Notamos con $\tau_\infty = \{x_j\}_{j\in\mathbb{Z}}$ una subdivisión de $\mathbb{R}$ tal que $x_j \underset{j\to-\infty}{\to} -\infty$, $x_j \underset{j\to+\infty}{\to} +\infty$, y $x_j < x_{j+1}$, $\forall j \in \mathbb{Z}$.

DEFINICIÓN 3.4:

Sea τ_∞ una subdivisión de $\mathbb{R}$. Se nota con $B_{m,j}$ la función de $\mathbb{R}$ en $\mathbb{R}$ tal que:

i. $B_{m,j}(x) = 0$ si $x \in \mathbb{R} - [x_j, x_{j+m+1}]$, $j \in \mathbb{Z}$;

ii. $B_{m,j} \in P_m$ sobre cada subintervalo $[x_i, x_{j+1}]$, $i = j,\ldots, j+m+1$;

iii. $\int_{-\infty}^{+\infty} B_{m,j}(x)dx = \int_{x_j}^{x_{j+m+1}} B_{m,j}(x)dx = 1$

Las funciones $B_{m,j}$ se llaman B - Splines. La condición iii) se conoce con el nombre de condición de normalización. Se puede probar que existe una única función $B_{m,j}$ que verifica i), ii) y iii).

Las funciones $\{B_{m,j} \mid j \in \mathbb{Z}\}$ forman una base del espacio de splines $S_m(\tau_\infty)$.

3.4.1 INTERPOLACIONES MEDIANTE B-SPLINES CÚBICAS

Sea $f : [a,b] \to \mathbb{R}$ una función definida en $[a,b]$. Buscamos una función $S \in S_3(\tau(n))$ tal que

$$S(x_j) = f(x_j), \quad j = 0,1,\ldots,n,$$

donde $\tau(n)$ es una subdivisión en puntos igualmente espaciados.

Para lograrlo, necesitamos los valores de las B-splines $B_{3,j}, \quad j = -3,\ldots,n-1$ en los nodos $x_0 = a, x_1, \ldots, x_n = b,$ así como los valores de las derivadas $B'_{3,j}$ o $B''_{3,j}$ en $x_0 = a$ para $j = -3,-2,-1$ y en $x_n = b$ para $j = n-3, n-2, n-1.$

En la tabla siguiente se ilustran estos valores:

	x_j	x_{j+1}	x_{j+2}	x_{j+3}	x_{j+4}
$B_{3,j}(x)$	0	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$	0
$B'_{3,j}(x)$	0	$\frac{1}{2h}$	0	$-\frac{1}{2h}$	0
$B''_{3,j}(x)$	0	$\frac{1}{h^2}$	$-\frac{2}{h^2}$	$\frac{1}{h^2}$	0

.

Sea $S \in S_3(\tau(n))$. Supongamos que

$$S(x) = \sum_{j=-3}^{n-1} \alpha_j B_{3,j}(x), \quad x \in [a,b].$$

Los problemas a), b) y c) discutidos en la sección de interpolación mediante splines cúbicas se escriben como sigue:

$$\sum_{j=-3}^{n-1} \alpha_j B_{3,j}(x_k) = f(x_k), \quad k = 0,1,\ldots,n.$$

Condiciones de frontera:

a) $\sum_{j=k-3}^{k-1} \alpha_j B'_{3,j}(x_k) = f'(x_k), \quad k = 0, n;$

b) $\sum_{j=-3}^{-1} \alpha_j B''_{3,j}(x_k) = 0, \quad k = 0, n;$

c) $\sum_{j=-3}^{-1} \alpha_j B'_{3,j}(a) = \sum_{j=n-3}^{n-1} \alpha_j B'_{3,j}(b);$

$\sum_{j=-3}^{-1} \alpha_j B''_{3,j}(a) = \sum_{j=n-3}^{n-1} \alpha_j B''_{3,j}(b).$

El sistema de ecuaciones resultante

$$A\vec{C} = \vec{b},$$

para los tres problemas, tiene las formas siguientes:

1. **Condiciones de frontera de Hermite**

$$A = \frac{1}{6}\begin{pmatrix} -\frac{3}{h} & 0 & \frac{3}{h} & 0 & \cdots & 0 \\ 1 & 4 & 1 & 0 & \cdots & 0 \\ 0 & 1 & 4 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \cdots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & & & 1 & 4 & 1 \\ 0 & \cdots & \cdots & -\frac{3}{h} & 0 & \frac{3}{h} \end{pmatrix},$$

$\vec{b}^{\,t} = (f'(a), f(x_0), \ldots, f(x_n), f'(b)).$

2. **Splines naturales**

$$A=\frac{1}{6}\begin{pmatrix} \frac{6}{h^2} & -\frac{12}{h^2} & \frac{6}{h^2} & 0 & \cdots & 0 \\ 1 & 4 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \cdots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & & & 1 & 4 & 1 \\ 0 & \cdots & \cdots & \frac{6}{h^2} & -\frac{12}{h^2} & \frac{6}{h^2} \end{pmatrix},$$

$\vec{b}^{\,t}=(0,f(x_0),\ldots,f(x_n),0).$

3. **Splines periódicas**

$$A=\frac{1}{6}\begin{pmatrix} -\frac{3}{h} & 0 & \frac{3}{h} & 0 & \cdots & 0 & \frac{3}{h} & 0 & -\frac{3}{h} \\ \frac{6}{h^2} & -\frac{12}{h^2} & \frac{6}{h^2} & 0 & \cdots & 0 & -\frac{6}{h^2} & \frac{12}{h^2} & -\frac{6}{h^2} \\ 1 & 4 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 1 & 4 & 1 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & & \ddots & \ddots & \ddots & & & & \vdots \\ \vdots & & & \ddots & \ddots & \ddots & & & \vdots \\ \vdots & & & & \ddots & \ddots & \ddots & & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & 1 & 4 & 1 & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 1 & 4 & 1 \end{pmatrix},$$

$\vec{b}^{\,t}=(0,0,f(x_0),\ldots,f(x_{n-1}),f(a)).$

4 ASPECTOS COMPUTACIONALES

Para poder llevar a la práctica los métodos de análisis de datos funcionales es fundamental el uso de un software estadístico potente y flexible. Se utiliza el software R por dos razones fundamentales, la primera es que es el paquete más usado por los investigadores en ADF y la segunda que es de uso libre y permite el desarrollo de código propio para aquellos métodos no implementados directamente en sus librerías.

Algunas librerías se usarán en la aplicación, como por ejemplo:

- Libería fda: implementa las técnicas de análisis de datos funcionales del libro de Ramsay y Silverman (2007).
- Librería fda.usc: integra y complementa la librería fda con medidas de profundidad, detección de outliers funcionales, modelos de regresión funcional y métodos de clasificación de un conjunto de datos funcionales.
- Librería fds: conjuntos de datos funcionales.

Las funciones de R que se aplicarán pertenecen en su mayoría a la librería fda que sirve como referencia a investigadores de todo el mundo (Ramsay et al., 2009). Se hará uso también de la librería fda.usc para el análisis exploratorio de las curvas muestrales (Febrero-Bande y Oviedo de la Fuente, 2012).

4.1 LECTURA DE DATOS

Las siguientes líneas de código, cargan los datos de frecuencias de la población ecuatoriana desde un archivo plano.

```
azuayFrec<- read.delim("C:/CodigosACPF/DatosProvin/azuayFrec.txt", dec=",")
```

De manera similar procedemos con la carga de las frecuencias de las demás provincias.

En el archivo plano se encuentra las frecuencias de los hombre y mujeres ordenados en forma ascendente de acuerdo a la edad, desde el 2010 hasta el 2020 (Hombre10 Mujer10 Hombre11 Mujer11 ...).

4.2 PREPARACIÓN DE DATOS

Se implementan algunas funciones que nos permiten preparar los datos para su posterior procesamiento.
Las siguientes funciones son complementarias para invertir el orden de un vector y concatenar dos vectores respectivamente.

```
intercambia<-function(y){
k=length(y)
vector<-rep(0,k)
for (j in k : 1){
vector[j]=y[k+1-j]
}
end
return(vector)

}

concatenar<-function(Prov){
dimen=dim(Prov)
Provin<-matrix(0,nrow=dimen[1]*2,ncol=dimen[2]/2)
k=0
for(i in 1:(dimen[2]/2))
{
Provin[,i]<-c(Prov[,2*i],Prov[,2*i-1])
}
end
return(Provin)

}
```

El siguiente código invierte el orden de las frecuencias para las mujeres para cada año y provincia, como ejemplo se hace para Azuay (de mayor edad a menor).

```
tam<-dim(azuayFrec)
```

```
for(i in 1:round(tam[2]/2)){
 y<-azuayFrec[,2*i]; azuayFrec[,2*i]<-intercambia(y)

}
```

Luego concatenamos el vector de frecuencias de las mujeres y hombres para cada año, como ejemplo se hace para azuay (primero las mujeres).

```
Prov<-azuayFrec; azuay<-concatenar(Prov)
```

4.3 PIRÁMIDE DE POBLACIÓN

Se grafica la pirámide de población, se toma como ejemplo los datos de la provincia de Pichincha del año 2015.

```
pichinchaFrec1<-read.delim("C:/CodigosACPF/DatosProvin/pichinchaFrec.txt", dec=",")
mujeres<-pichinchaFrec1$Mujer15
hombres<-pichinchaFrec1$Hombre15
pob<-c(mujeres,hombres)
plot.new()
amplitud<-1
escalax<-0.002
edadmax<-99
region<-"Pichincha"
max1<-max(c(hombres,mujeres))
n<-length(hombres)
min.x<--(max1%/%escalax+1)*escalax
max.x<-(max1%/%escalax+1)*escalax
plot(0,0,type="n",xaxt='n',yaxt='n',ylim=c(0,edadmax),
xlim=c(min.x,max.x),xlab="",ylab="")
ejex1<-seq(0,max1,by=escalax)
ejex2<--ejex1[order(-ejex1)]
ejex<-c(ejex2,ejex1)
axis(1,at=ejex,labels=as.character(abs(ejex)),
cex.axis=0.8,las=1)
```

```
ejey<-c(seq(0,edadmax,by=amplitud))
axis(2,at=ejey,labels=as.character(ejey),cex.axis=0.3,las=1)
for(i in 1:n){
  x1<-0
  x2<-hombres[i]; x3<--mujeres[i]
  y1<-(i-1)*amplitud
  y2<-y1+amplitud
  rect(x1,y1,x2,y2,col='BurlyWood')
  rect(x1,y1,x3,y2,col='antiqueWhite')
}
x.l1<--max1/16-1.5*escalax
x.l2<-max1/16+escalax
title(main=paste("Pirámide de Población",sep="\n"),ylab="Edad",xlab="Frecuencia")
legend(x.l1,edadmax+5,"mujeres",bty="n",xjust=0.91)
legend(x.l2,edadmax+5,"hombres",bty="n")
```

Graficamos la pirámide de población abatida, se toma como ejemplo los datos de la provincia de Pichincha del año 2015.

```
pichinchaFrec1<-read.delim("C:/CodigosACPF/DatosProvin/pichinchaFrec.txt", dec=",")
mujeres<-(pichinchaFrec1$Mujer15)
hombres<-pichinchaFrec1$Hombre15
pob<-c(mujeres,hombres)
plot.new()
escalax<-1
edadmax<-0.002
region<-"Azuay"
max1<-max(c(hombres,mujeres))
n<-length(hombres)
plot(0,0,type="n",xaxt='n',yaxt='n',ylim=c(0,max1),
xlim=c(-99,99),xlab="",ylab="")
ejex1<-seq(0,n,by=escalax)
ejex2<--ejex1[order(-ejex1)]
ejex<-c(ejex2,ejex1)
axis(1,at=ejex,labels=as.character(abs(ejex)),cex.axis=0.8,las=1)
ejey<-c(seq(0,max1,by=0.002))
```

```
axis(2,at=ejey,labels=as.character(ejey),cex.axis=0.5,las=1)
for(i in 1:n){
  x1<-0
  x2<-hombres[i]; x3<-mujeres[i]
  rect(i-1,x1,i,x2,col='BurlyWood')
  rect(1-i,x1,-i,x3,col='antiqueWhite')
}
x.l1<--max1-50
x.l2<-max1+50
title(main=paste("Pirámide Abatida",sep="\n"),ylab="frecuencia",xlab="edad")
legend(x.l1,max1,"mujeres",bty="n",xjust=0.91)
legend(x.l2,max1,"hombres",bty="n")
```

4.4 REPRESENTACIÓN DE DATOS FUNCIONALES

Cargamos los datos de todas las provincias en una matriz de 200x264 y lo transformamos, donde cada columna representa un dato funcional.

```
DatosEcu<-matrix(c(azuay,bolivar,caniar,carchi,chimborazo,
 cotopaxi,elOro,esmeraldas,galapagos,guayas,imbabura,
 loja,losRios,manabi,moSantiago,napo,orellana,pastaza,pichincha,
 staElena,stoDomingo,sucumbios,tungurahua,zamChinchipe),ncol=264,nrow=200)
par(mfrow=c(1,1))
t<-c(seq(from=-99,to=0,by=1),seq(from=0,to=99,by=1))
x<-fdata(t(DatosEcu),t)
```

Se realiza el suavizamiento de los datos con B-Splines y su representación en bases de funciones, utilizando el criterio de validación cruzada (GCV).

```
result.np1 <- min.basis(x,type.CV = GCV.S, type.basis="bspline",verbose=TRUE)
```

4.5 ANÁLISIS DESCRIPTIVO

Se calcula la media y la varianza funcional de los datos originales.

```
par(mfrow=c(1,1))
plot(func.mean(x), main = "Media Funcional ",xlab="Edad",ylab="Frecuencia",lwd=2)
legend("topleft", c("mean"),lwd=3,bty="n",text.col=1,col=1)
par(mfrow=c(1,1))
plot(func.var(x), main = "Varianza Funcional ",xlab="Edad",ylab="Frecuencia",lwd=1)
legend("topleft", c("varianza"),lwd=1,bty="n",text.col=1,col=1)
```

4.6 ANÁLISIS DE COMPONENTES PRINCIPALES FUNCIONALES

Primero se realiza un suavizamiento de los datos y se grafican las dos primeras componentes principales.

```
par(mfrow=c(1,1))
t1=seq(from=-99,to=100,by=1)
bspl <- create.bspline.basis(c(-99,100),nbasis=31,norder=4)
harmaccelLfd <- vec2Lfd(c(0,0.005), c(-100,100))
harmfdPar <- fdPar(bspl, harmaccelLfd, lambda=1e5)
daytempfd <- smooth.basis(t1, DatosEcu, bspl)$fd
numComp=2
daytemppcaobj <- pca.fd(daytempfd, nharm=numComp, harmfdPar)
op <- par(mfrow=c(1,1))
plot.pca.fd(daytemppcaobj,ylab="Función ")
```

Graficamos las dos funciones propias asociadas a las componentes principales.

```
plot(daytemppcaobj$harmonics,main="Componentes Principales Funcionales",lty=1)
legend("topleft",c("CP1","CP2"), title="Componente",text.col=c(1,2),col=c(1,2))
```

Obtenemos algunos resultados como: funciones propias, valores propios, la varianza explicada por cada función propia, las puntuaciones en las componentes principales, y la media funcional.

```
eigenFunctions<-data.frame(daytemppcaobj$harmonics$coef)
completeEigenValues<-data.frame(daytemppcaobj$values)
VarianceExplained<-data.frame(daytemppcaobj$varprop)
scoresPrincipalComponent<-data.frame(daytemppcaobj$scores)
functionalMean<-data.frame(daytemppcaobj$meanfd$coefs)
```

Finalmente guardamos los resultados anteriores en un archivo de Excel.

```
library(XLConnect)
wb=loadWorkbook("Resultados.xls",create=TRUE)
createSheet(wb,c("eigenFunctions","completeEigenValues","VarianceExplained",
            "scoresPrincipalComponent","functionalMean"))
saveWorkbook(wb)
createName(wb,name=c("eigenFunctions","completeEigenValues","VarianceExplained",
             "scoresPrincipalComponent","functionalMean"),
formula=c("eigenFunctions!$B$1","completeEigenValues!$B$1","VarianceExplained!$B$1
", "scoresPrincipalComponent!$B$1","functionalMean!$B$1"))
writeNamedRegion(wb,eigenFunctions,name="eigenFunctions")
writeNamedRegion(wb,completeEigenValues,name="completeEigenValues")
writeNamedRegion(wb,VarianceExplained,name="VarianceExplained")
writeNamedRegion(wb,scoresPrincipalComponent,name="scoresPrincipalComponent")
writeNamedRegion(wb,functionalMean,name="functionalMean")
saveWorkbook(wb)
```

5 APLICACIÓN DEL ACPF

En demografía, la pirámide de población es una herramienta que permite al investigador saber cómo es la distribución de la edad de la población así como algunas de sus características demográficas sobre la estructura de la población.

DEFINICIÓN: *Una pirámide de población es un gráfico compuesto por dos histogramas, cuyos datos son: la edad (eje y) y el sexo (eje x), donde se coloca a la derecha a los hombres y a la izquierda las mujeres.*

La ventaja de la pirámide de población es que se puede observar de manera simultánea la distribución de la población por edad y sexo. Para construir las pirámides se utilizó la frecuencia de la edad para evitar trabajar con cantidades muy grandes.
Con ayuda del ACPF se interpretará el comportamiento de la población mediante componentes principales funcionales, para esto es necesario un tratamiento previo a los datos.

5.1 REPRESENTACIÓN DE LOS DATOS

Los datos corresponden a la proyección de la población $2010-2020$ realizado por el INEC. Estos datos, están desagregados por edades simple, sexo y por provincia para cada año, es decir, se tienen 264 pirámides de población. Cada pirámide está conformada por la frecuencia de las edades de 0 a 99 años tanto para hombres como mujeres.
Ahora, para poder trabajar con estos datos es necesario realizar una preparación de los mismos, que será de gran utilidad para la conversión de los datos en su forma funcional.

5.1.1 TRANSFORMACIÓN DE PIRÁMIDES EN FUNCIONES

Para graficar las pirámides de población se calcularon las frecuencias para cada edad, tanto para hombres y mujeres, representados a la derecha e izquierda respectivamente.

Para utilizar el ACPF se transformará las pirámides en funciones mediante el abatimiento como se muestra en la siguiente figura.

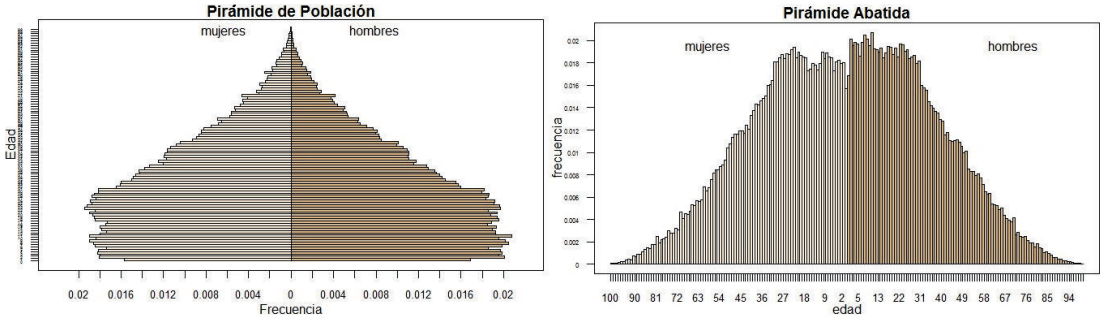

Figura 1 – Pirámide de población

En la siguiente figura se presenta el conjunto de datos con los cuales se va a trabajar.

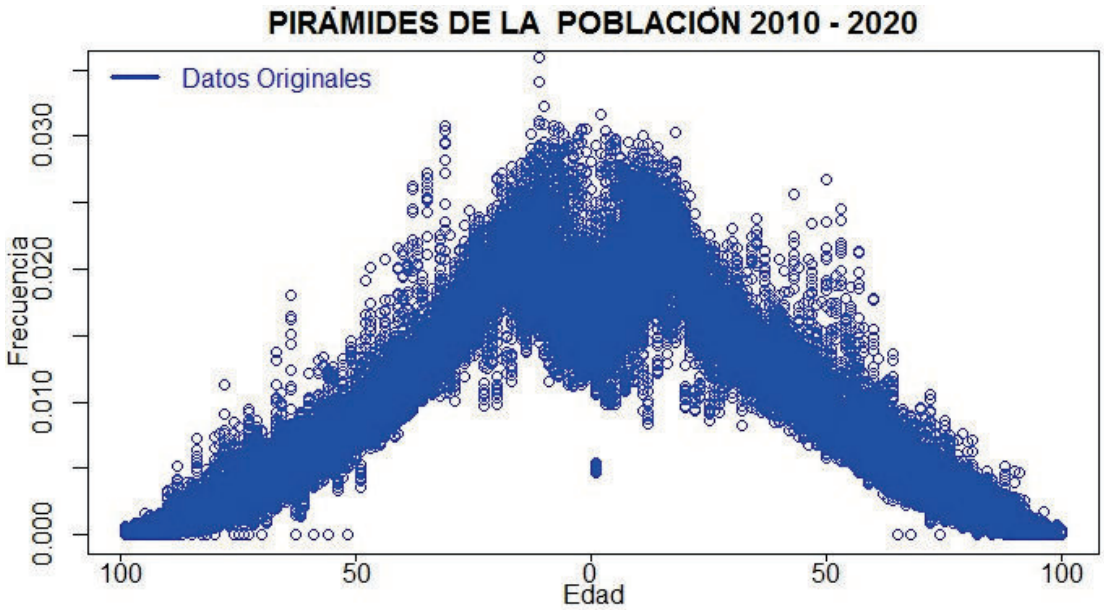

Figura 2 – Pirámides de población abatidas

Luego procedemos a su representación funcional, la cual se realiza con ayuda del paquete estadístico R.

5.1.2 REPRESENTACIÓN FUNCIONAL

Recordemos que X es la variable funcional que toma valores en un espacio vectorial normado (o semi-normado) V. Si V es un espacio de Hilbert, entonces, el dato funcional se lo puede representar aproximadamente mediante una base [8]. Los datos funcionales se encuentran discretizados en un conjunto de puntos $\{t_i\}_{i=1,\dots k}$ (no necesariamente equidistantes); por lo tanto se tiene k evaluaciones para cada una de las n observaciones, para cada columna de la matriz de 200 filas y 264 columnas para este caso, donde cada columna representa la pirámide de población de una provincia en un año determinado.

En la figura siguiente se presentan las 264 pirámides abatidas en su forma funcional, donde la edad con signo es la variable aleatoria continua que tiene definida una función distinta para cada provincia. Se han resaltado las funciones correspondientes a las provincias de Azuay (color rojo), Galápagos y Pichincha correspondientes al año 2014.

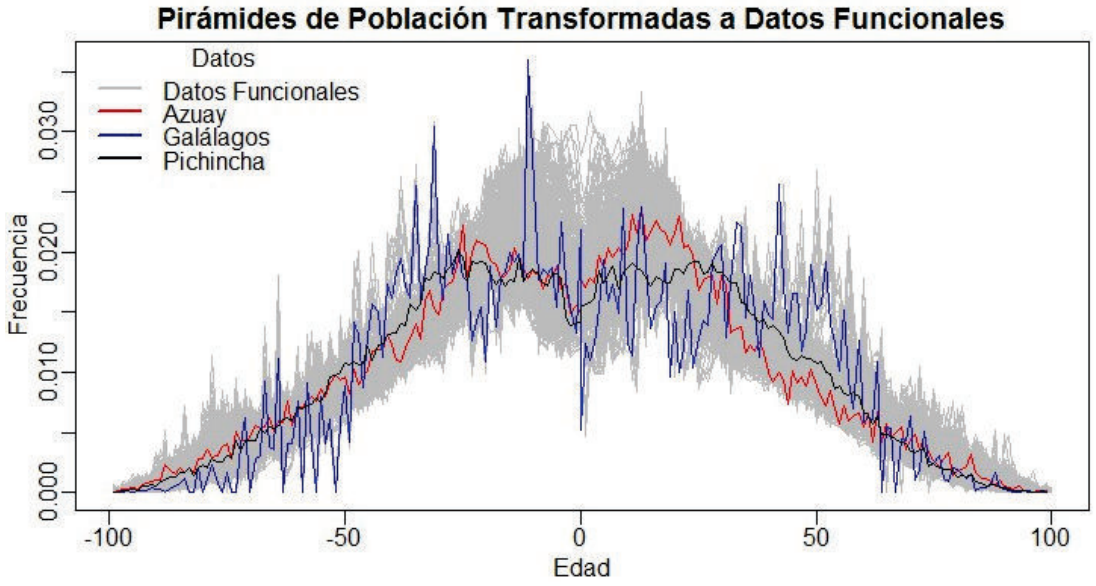

Figura 3 – Pirámides de población abatidas transformadas a datos funcionales

5.1.3 SUAVIZADO DE LOS DATOS FUNCIONALES

Supongamos que nuestro dato funcional $Y(t)$ es observado mediante el modelo $Y(t_j) = X(t_j) + \varepsilon(t_j)$, donde, $Y(t_j)$ es el dato observado y $\varepsilon(t)$ el error producido por la matriz de covarianzas que es independiente de $X(t)$; para obtener el signo de $X(t)$ es necesario realizar un suavizamiento lineal:

$$\hat{X} = \sum_{i=1}^{n} s_{ij} y_i$$

donde s_{ij} es el peso que el punto t_j da al punto t_i y y_i es el valor observado de la variable Y en el punto t_i (respectivamente, $x_i = X(t_i)$); para esto se implementarán dos procedimientos:

- Representación en una base (penalizada) de L^2 y,
- Suavizado del núcleo o kernel smoothing.

Para el suavizado de los datos se utilizarán la B-Splines para aprovechar sus buenas propiedades; se presentan dos metodologías de suavizamiento con el fin de mostrar las diferencias que existen entre ellas.

Representación en bases

Una curva puede ser representada por bases cuando se asume que los datos pertenecen al espacio L^2 . Recordemos que, una base es un conjunto de funciones conocidas $\{\phi_k\}_{k\in\mathrm{N}}$ tal que cualquier función puede ser aproximada por una combinación lineal de n funciones con n lo suficientemente grande.

Para ello se trunca en un cierto n de modo que el error sea despreciable, es decir,

$$X(t) = \sum_{k\in\mathrm{N}} c_k \phi_k \; ? \sum_{k=1}^{n} c_k \phi_k(t) = \mathbf{c}^{\mathrm{T}} \mathbf{\Phi},$$

donde c_i son los coeficientes de la nueva base. Usando esta representación, la matriz de proyección (o suavizamiento) está dada por: $S = \Phi(\Phi^T W \Phi)^{-1} \Phi^T W$ con $traza(S) = k$ grados de libertad [9].

Para la elección del número de elementos de la base se utilizó el criterio de validación cruzada.

Criterios de validación

La dimensión infinita (o la gran dimensión) de un dato funcional es difícil manipular, por lo tanto, estos datos se proyectan sobre un espacio de dimensión finita es cual es más tratable.

La elección del número de parámetros de la base y la base más apropiada para los datos observados es importante. La decisión sobre qué base se debe más al estudio que se realiza se basa en la característica de los datos; por ejemplo, si los datos son periódicos se utiliza una base de "Fourier" y "B-Splines" para datos no recurrentes como en este caso.

Entre los diferentes criterios de selección del parámetro $\upsilon = (k, \lambda)$, tenemos: el "criterio de validación cruzada (C.V)" y el "criterio de validación cruzada generalizada (GCV)".

En el software R, el paquete **fda.usc** incorpora la función **min.basis()**, la cual representa un dato funcional en bases. La elección de un número adecuado de bases $\upsilon_1 = k$ y el parámetro de penalización $\upsilon_2 = \lambda$ están incluidos en este proceso. Los dos criterios están definidos como sigue:

Validación cruzada:

$$CV(v) = \frac{1}{k}\sum_{i=1}^{k}\frac{\left(y_i - \hat{r}_{-i}^{v}(x_i)\right)^2}{1 - S_{ii}}w(x_i),$$

donde $\hat{r}_{-i}^{v}(x_i)$ es el estimador (o la predicción) sobre el punto t_i el cual no utiliza la observación i-ésima, el par (x_i, y_i); S_{ii} es el elemento i-ésimo de la matriz de suavizamiento S (con $\upsilon = traza(S)$) y $\omega(x_i)$ es el peso del dato x en el punto t_i. Este criterio está implementado en la función **CV.S().**

Validación cruzada generalizada:

$$GCV(v) = \frac{1}{k}\sum_{i=1}^{k}\left(y_i - \hat{r}_{-i}^{v}(x_i)\right)^2 w(x_i)\,\Xi\,(v),$$

donde Ξ denota el tipo de función penalizadora.

El criterio de validación cruzada generalizada está implementado en la función **GCV.S()** con los siguientes tipos de funciones penalizadoras Ξ [9] .

- Validación cruzada generalizada (GCV): $\Xi(\upsilon) = (1 - tr(S)k^{-1})^{-2}$.
- Criterio de información Akaike (AIC): $\Xi(\upsilon) = \exp(2tr(S)k^{-1})$.
- Predicción finita del error (FPE): $\Xi(\upsilon) = \frac{1+tr(S)k^{-1}}{1-tr(S)k^{-1}}$.
- Modelo selector de Shibata (Shibata): $\Xi(\upsilon) = 1 + 2tr(S)k^{-1}$.

En las siguientes figuras se muestran las imágenes mediante los diferentes criterios de validación cruzada generalizada (GCV).

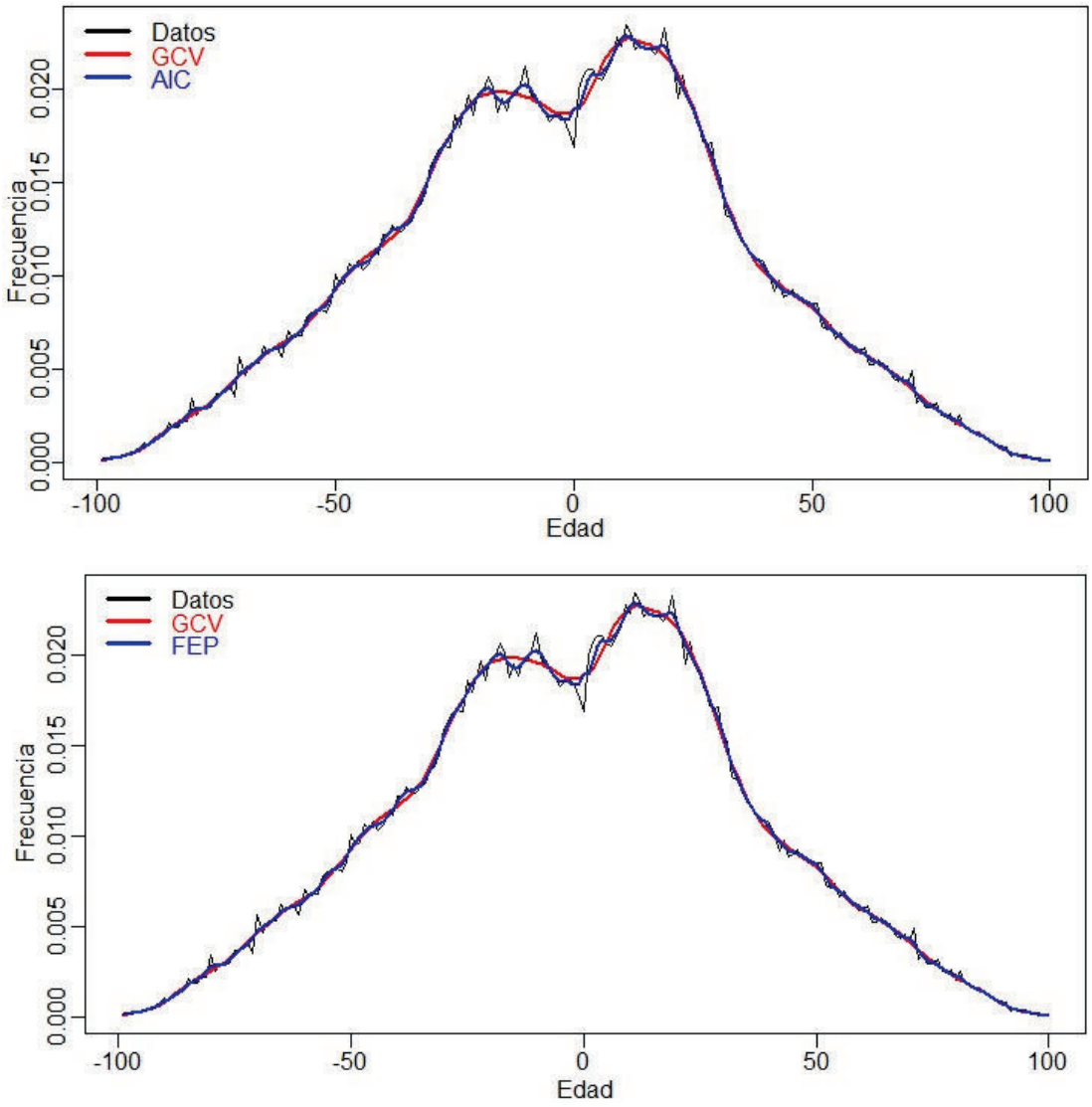

Figura 4 – GCV y AIC, GCV y FEP

En la figura se puede observar las diferencias que existe entre el criterio de validación cruzada generalizada (GCV) con el criterio de información de Akaike (AIC) y el criterio de predicción finita del error (FPE); los tres criterios no presentan mayor diferencia en la suavidad de los datos. La diferencia entre el GCV, AIC y FEP es que el primero presenta una oscilación menor con respecto a los otros criterios de validación, además, nótese que AIC y el FPE tienen un gran parecido como se muestra en la figura anterior.

En la figura siguiente puede verse que no existe mayor diferencia el GCV y el modelo selector Shibata además, la oscilación es similar en ambos casos.

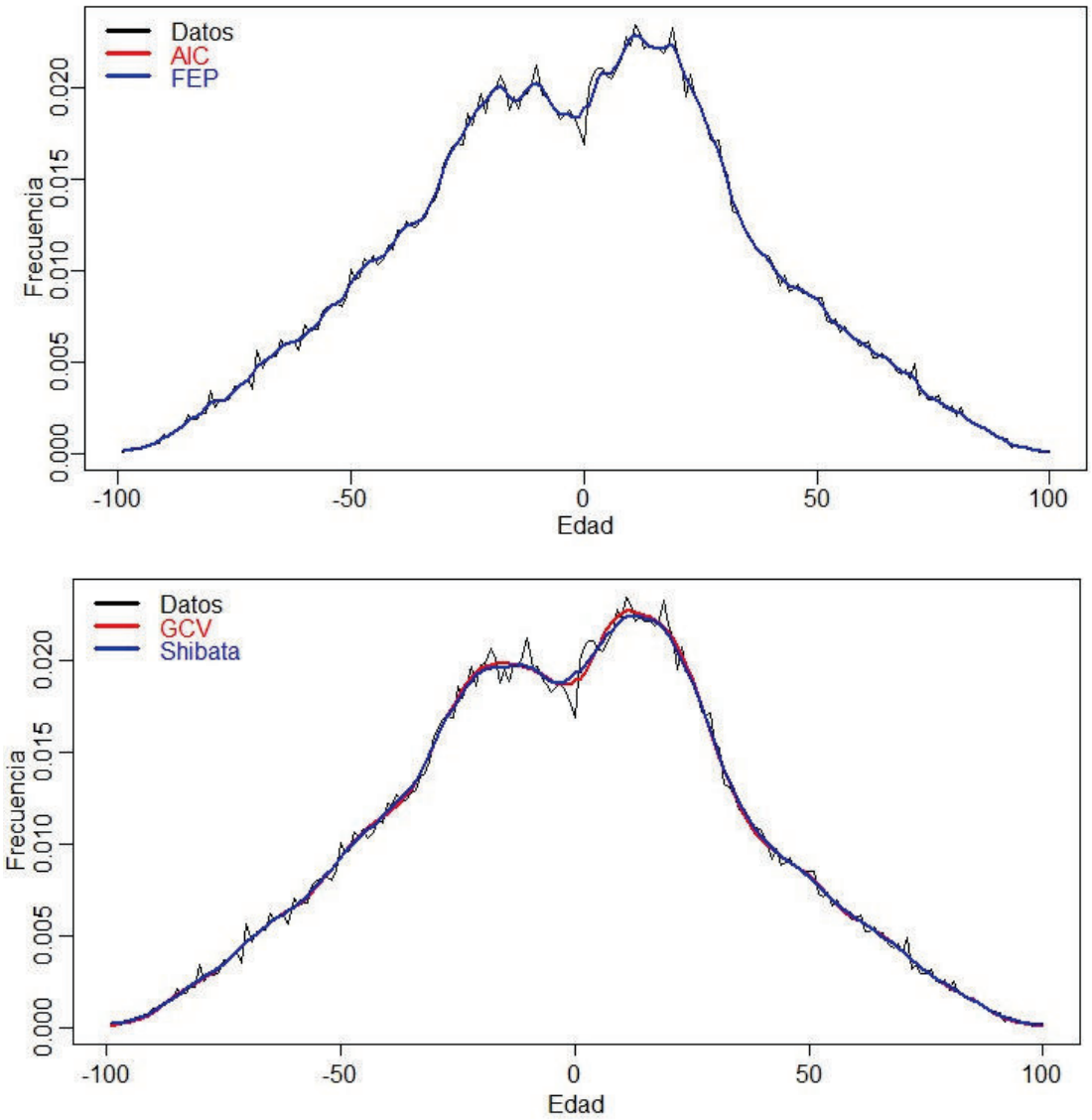

Figura 5 – GCV y Shibata, AIC y FEP

Ahora utilizando los criterios anteriores se obtuvo que el número óptimo de elementos de la base es 31 (B-Splines) por el criterio de GCV y por el criterio de CV se obtuvo que el número óptimo es 29 (B-Splines).

En la figura siguiente se presentan los distintos tipos de suavizamiento mediante el criterio de GCV.

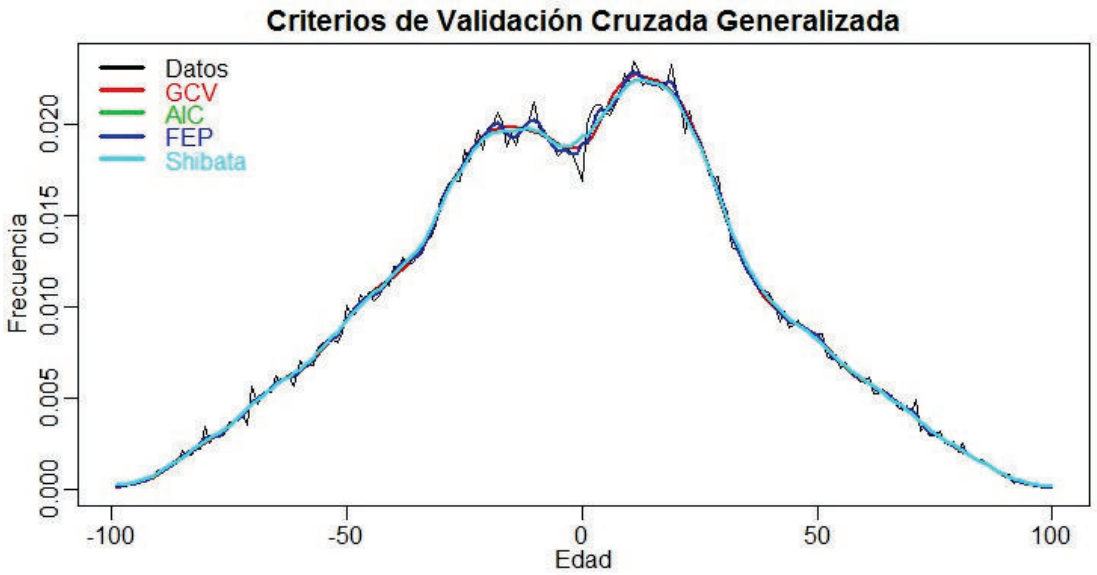

Figura 6 – Datos funcionales suavizados

Luego de haber seleccionado el número de elementos de la base, se procede a la suavización de todos los datos como se muestra en la figura, en la cual se han resaltado las provincias del Azuay, Galápagos y Pichincha correspondientes al año 2014.

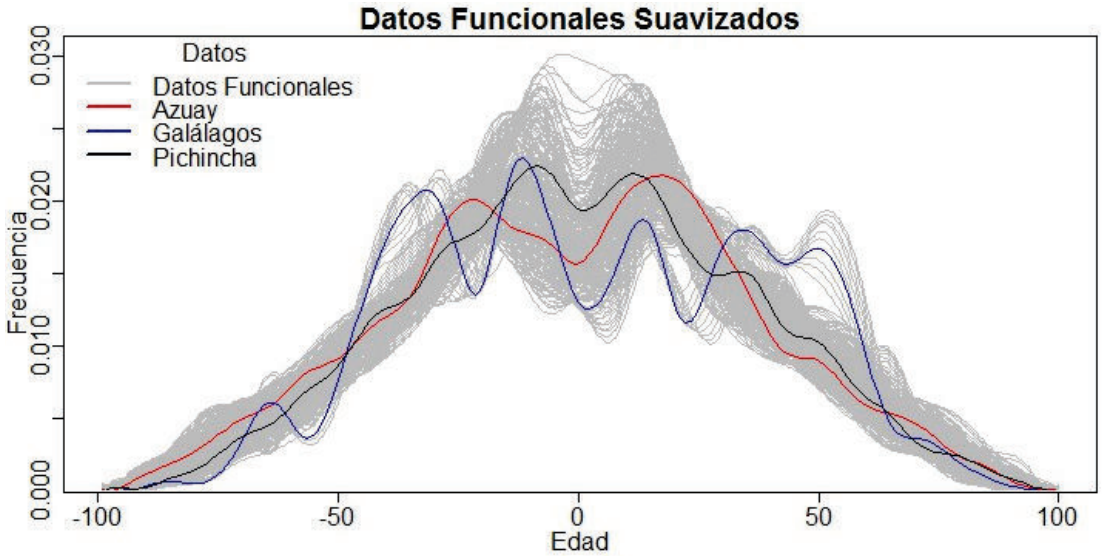

Figura 7 – Conjunto de datos funcionales suavizados

5.2 RESULTADOS

Para el análisis descriptivo (exploratorio) de los datos funcionales no se trabaja con la representación en base de funciones, sino solo es su representación funcional, ya que los resultados estarían condicionados por el error de representación.

5.2.1 ANÁLISIS DESCRIPTIVO

A continuación, se realiza un análisis descriptivo de la población mediante el cálculo de la media y la varianza funcional, esto con el fin de observar la tendencia central y la dispersión de los datos.

Media Funcional

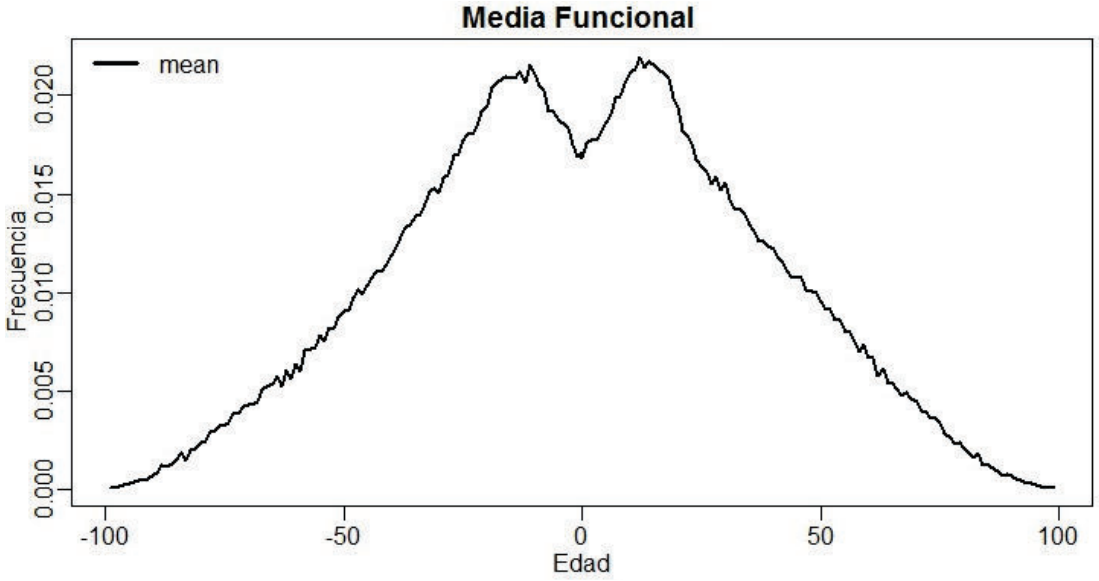

Figura 8 – Media funcional

Igual que en todas las provincias del Ecuador, la media presenta una deformación en la base de la pirámide, lo que significa que las tasas de fecundidad se están reduciendo. En $[10]$ puede verse los resultados de la proyección $2010-2020$ mostrada en grupos etáreos (quinquenales), en la que se observa claramente que la población entre 0 y 5 años cada vez es menor disminuyendo su tasa de crecimiento de 1.39 en el 2010 a 1.20 en el 2020.

Varianza Funcional

En la figura siguiente puede verse que la varianza es mayor entre los 0 y 5 años y entre 40 y 80 años tanto como hombres y mujeres; esta diferencia puede ser debida a la mala declaración de la edad, en particular de Galápagos y las provincias del Oriente que presentan la mayor cantidad de variación con respecto a las demás provincias.

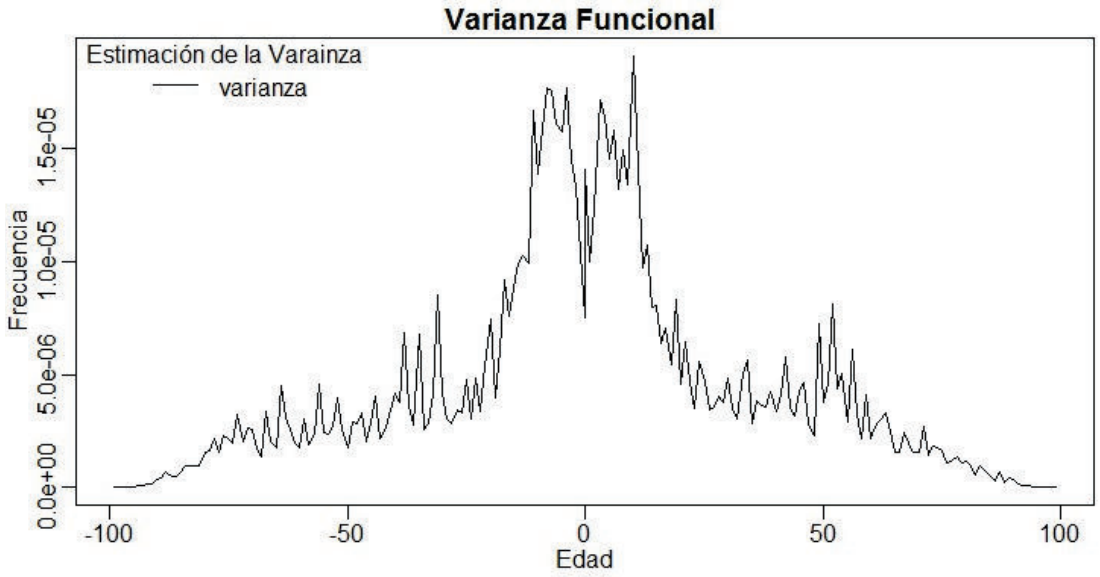

Figura 9 – Varianza funcional

5.2.2 COMPONENTES PRINCIPALES FUNCIONALES

En esta sección se muestran los resultados del análisis en componentes principales funcionales, para lo cual, se hace un suavizamiento previo; este suavizamiento fue hecho mediante una base B-Splines.

El objetivo del análisis en componentes principales funcionales es la de reducir la dimensión del espacio, puesto que las variables aleatorias funcionales aquí tratadas se encuentran en espacios de dimensión infinita.

Valores propios

Mediante el criterio de validación cruzada se tiene que el número óptimo de elementos de la base (B-Splines) es 31, por lo tanto la dimensión del espacio se redujo al número de elementos de la base.

Ahora, para el cálculo de los valores propios y funciones propias se usó la función **pca.fd** con la cual se obtuvieron los siguientes resultados:

Tabla 1- Valores Propios para cada B-Spline

B-Spl	Val.Pro
bsp4.1	2,06e-4
bsp4.2	2,20e-5
bsp4.3	5,47e-6
bsp4.4	1,34e-6

Puesto que la dimensión del espacio es 31 se obtuvieron igual número de valores propios, de los cuales sólo presentamos los cuatro primeros.

Variabilidad explicada

Por medio del ACPF se interpretará las pautas de comportamiento de las frecuencias por edades de la pirámides de población, para esto, encontramos funciones (componentes principales funcionales) las cuales resumen el comportamiento de la variabilidad de las funciones; a partir de eso se calculó los coeficientes que indican como aumentó o disminuyó la población con respecto a la media poblacional.

Para observar el porcentaje de la variabilidad explicada, se presenta los primeros 4 valores propios en la siguiente tabla:

Tabla 2- Varianza explicada por cada función propia

Valor Propio	Varianza Explicada
λ_1	$87,18\%$
λ_2	$9,33\%$
λ_3	$2,32\%$
λ_4	$0,57\%$

Los dos primeros valores propios explican más del 96% de la variabilidad, por lo que se considera dos componentes principales (funcionales) que se encuentran graficadas a continuación:

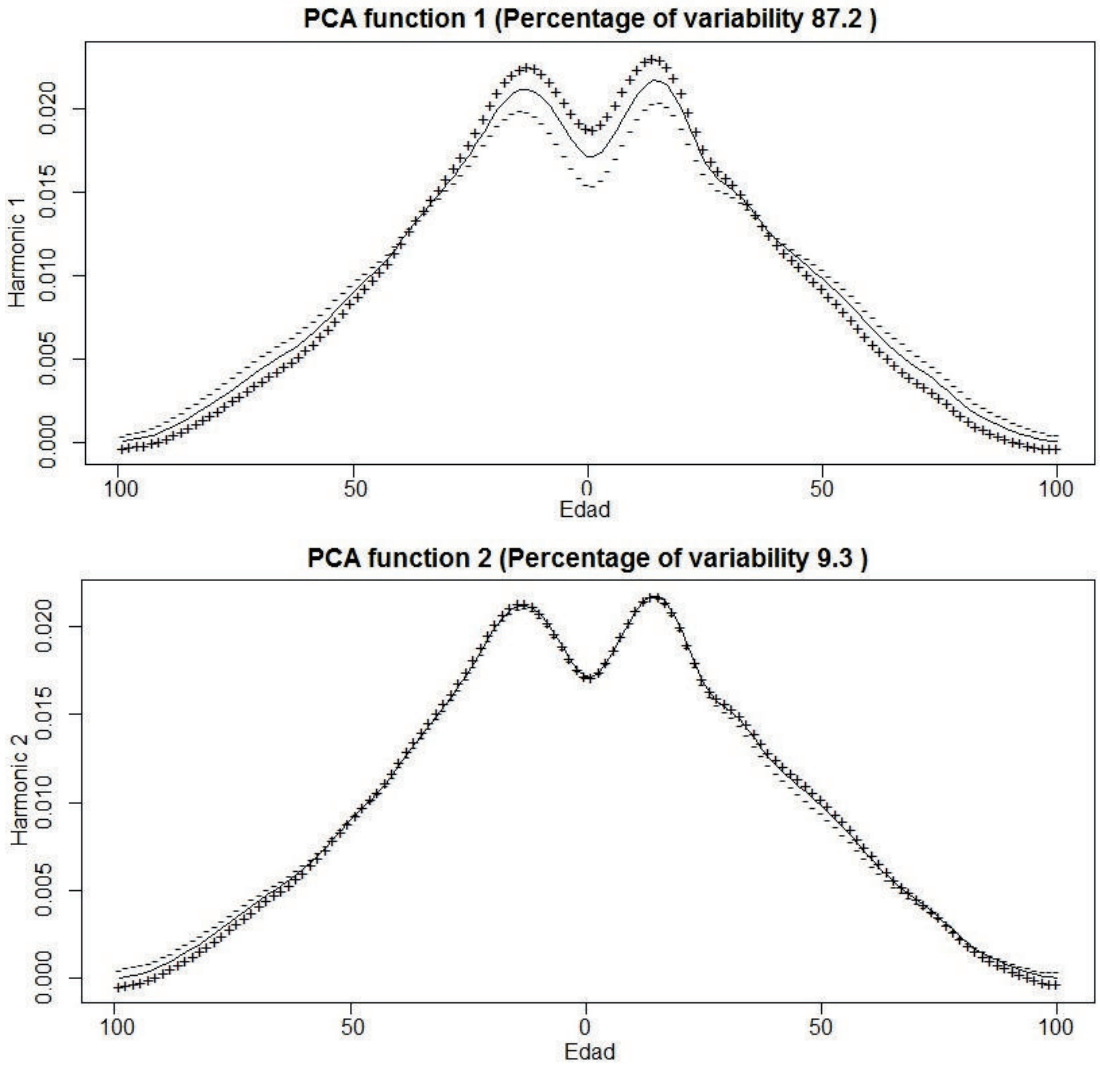

Figura 10 – Componentes principales funcionales

Estas componentes han sido sometidas a un suavizamiento donde los signos " + " y " - " ayudarán a la interpretación de las componentes principales.

Funciones propias

En la siguiente tabla se muestran los coeficientes de las funciones propias para las 2 componentes principales funcionales obtenidas.

Tabla 3- Coeficientes de las funciones propias

B-Spl	FPC1	FPC2	B-Spl	FPC1	FPC2	B-Spl	FPC1	FPC2
bsp4.1	-2,57e-2	-1,04e-1	bsp4.13	2,06e-2	3,86e-2	bsp4.25	-2,50e-2	8,77e-2
bsp4.2	-2,82e-2	-1,03e-1	bsp4.14	4,74e-2	3,84e-2	bsp4,26	-4,15e-2	8.67e-2
bsp4.3	-3,31e-2	-1,01e-1	bsp4.15	7,41e-2	3,42e-2	bsp4,27	-5,22e-2	7,43e-2
bsp4.4	-4,01e-2	-9,81e-2	bsp4.16	9,72e-2	2,74e-2	bsp4.28	-5,74e-2	5,34e-2
bsp4.5	-4,63e-2	-9,28e-2	bsp4.17	1,13e-1	1,78e-2	bsp4.29	-5,80e-2	2,85e-2
bsp4.6	-5,08e-2	-8,34e-2	bsp4.18	1,19e-1	7,47e-3	bsp4.30	-5,51e-2	3,64e-3
bsp4.7	-5,30e-2	-6,88e-2	bsp4.19	1,14e-1	1,02e-3	bsp4.31	-4,97e-2	-1,91e-2
bsp4.8	-5,19e-2	-4,96e-2	bsp4.20	9,93e-2	2,49e-3			
bsp4.9	-4,70e-2	-2,72e--2	bsp4.21	7,68e-2	1,37e-2			
bsp4.10	-3,76e-2	-3,84e-3	bsp4.22	5,01e-2	3,36e-2			
bsp4.11	-2,31e-2	1,71e-2	bsp4.23	2,24e-2	5,72e-2			
bsp4.12	-3,49e-3	3,19e-2	bsp4.24	-3,30e-3	7,70e-2			

Los gráficos de las funciones principales o funciones propias asociadas a las componentes se presentan a continuación:

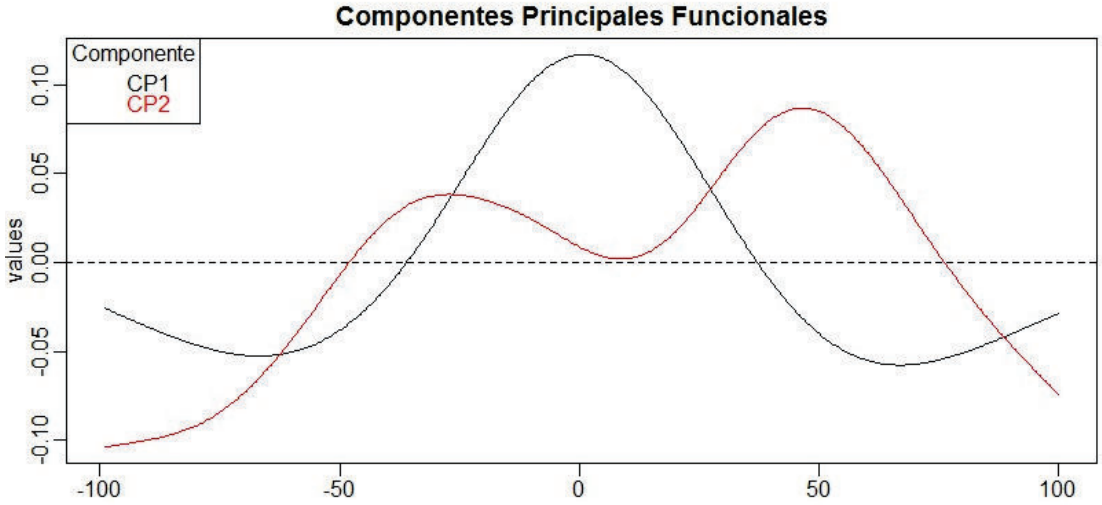

Figura 11 – Funciones propias

En la siguiente sección se muestran los coeficientes para cada provincia con el fin de entender el comportamiento de la población del Ecuador.

Interpretación de las componentes

En la Tabla 5 se muestran los coeficientes de las dos componentes principales en la que se puede observar que existen valores positivos y negativos los cuales darán el comportamiento de la forma de la pirámide de población para cada provincia en un año específico. Ahora, en la Tabla 5.4 se muestra la codificación de cada provincia con el objetivo de simplificar la escritura, es decir, si nos referimos a C.16 significa que se está haciendo referencia a la provincia de Carchi en el año 2016.

Tabla 4- Códigos por provincia

Código	Provincia	Código	Provincia	Código	Provincia
A	Azuay	W	Galápagos	Q	Orellana
B	Bolívar	G	Guayas	S	Pastaza
U	Cañar	I	Imbabura	P	Pichincha
C	Carchi	L	Loja	Y	Santa Elena
H	Chimborazo	R	Los Ríos	J	Santo Domingo
X	Cotopaxi	M	Manabí	K	Sucumbíos
O	El Oro	V	Morona Santiago	T	Tungurahua
E	Esmeraldas	N	Napo	Z	Zamora Chinchipe

Ahora se presentan los resultados y la interpretación de cada una de las provincias para ciertos años de la proyección.

Tabla 5- Matriz de scores de las funciones principales

Prov.	CPF1	CPF2	Prov.	CPF1	CPF2	Prov.	CPF1	CPF2	Prov.	CPF1	CPF2
A.10	4,82e-3	-4,37e-3	A.14	-2,98e-3	-3,25e-3	A.18	-1,14e-2	-2,76e-3	A.20	-1,56e-2	-2,58e-3
B.10	4,52e-3	-7,45e-3	B.14	-5,43e-3	-7,52e-3	B.18	-1,59e-2	-8,25e-3	B.20	-2,12e-2	-8,87e-3
U.10	8,43e-3	-7,49e-3	U.14	-2,63e-3	-6,73e-3	U.18	-1,01e-2	-6,50e-3	U.20	-1,50e-2	-6,34e-3
C.10	-1,35e-3	-2,96e-3	C.14	-1,36e-2	-1,58e-3	C.18	-2,56e-2	-3,60e-3	C.20	-3,15e-2	-4,76e-3
H.10	3,54e-3	-6,25e-3	H.14	-3,94e-3	-5,20e-3	H.18	-1,21e-2	-4,86e-3	H.20	-1,62e-2	-4,83e-3
X.10	9,48e-3	-4,27e-3	X.14	7,69e-4	-3,70e-3	X.18	-8,31e-3	-3,99e-3	X.20	-1,63e-2	-4,52e-3
O.10	1,57e-3	1,94e-3	O.14	-7,77e-3	3,77e-3	O.18	-1,72e-2	2,90e-3	O.20	-2,18e-2	2,25e-3
E.10	2,02e-2	1,41e-4	E.14	1,01e-2	2,06e-3	E.18	-6,16e-4	2,43e-3	E.20	-6,00e-3	2,57e-3
W.10	-3,14e-3	9,27e-3	W.14	-1,47e-2	1,46e-2	W.18	-2,65e-2	1,35e-2	W.20	-3,18e-2	1,31e-2
G.10	1,33e-3	2,39e-3	G.14	-6,51e-3	5,12e-3	G.18	-1,48e-2	4,93e-3	G.20	-1,89e-2	4,61e-3
I.10	4,32e-3	-3,89e-3	I.14	-2,13e-3	-4,13e-3	I.18	-9,41e-3	-4,34e-3	I.20	-1,32e-2	-4,56e-3
L.10	2,58e-3	-5,97e-3	L.14	-5,66e-3	-6,16e-3	L.18	-1,49e2	-6,06e-3	L.20	-1,96e-2	-6,22e-3
R.10	9,39e-3	1,44e-3	R.14	6,47e-4	3,33e-3	R.18	-8,75e-3	2,94e-3	R.20	-1,35e-2	2,50e-3
M.10	6,17e-3	-1,08e-4	M.14	-3,01e-3	3,39e-3	M.18	-1,33e-2	2,88e-3	M.20	-1,84e-2	2,50e-3
V.10	3,65e-2	-2,50e-3	V.14	2,74e-2	-1,75e-3	V.18	1,75e-2	-1,14e-3	V.20	1,24e-2	-8,95e-4
N.10	2,94e-2	9,03e-4	N.14	2,26e-2	4,63e-4	N.18	1,51e-2	5,00e-4	N.20	1,13e-2	3,80e-4
O.10	3,19e-2	4,44e-3	O.14	2,56e-2	2,84e-3	O.18	1,91e-2	2,67e-3	O.20	1,57e-2	2,55e-3
S.10	2,71e-2	9,18e-5	S.14	1,69e-2	2,85e-4	S.18	6,66e-3	3,23e-4	S.20	1,53e-3	1,22e-4
P.10	-1,55e-3	1,95e-3	P.14	-1,12e-2	3,87e-3	P.18	-2,08e-2	4,02e-3	P.20	-2,55e-2	3,82e-3
Y.10	1,27e-2	1,14e-3	Y.14	5,21e-3	1,96e-3	Y.18	-2,43e-3	1,32e-3	Y.20	-6,25e-3	7,30e-4
J.10	1,46e-2	2,14e-3	J.14	5,98e-3	3,43e-3	J.18	-3,11e-3	3,53e-3	J.20	-7,67e-3	3,42e-3
K.10	2,39e-2	5,14e-3	K.14	1,60e-2	3,44e-3	K.18	7,90e-3	2,77e-3	K.20	3,85e-3	2,39e-3
T.10	-3,73e-3	-2,38e-3	T.14	-1,19e-2	-1,89e-3	T.18	-2,02e-2	-2,36e-3	T.20	-2,43e-2	-2,84e-3
Z.10	2,58e-2	-7,36e-4	Z.14	1,57e-2	-1,05e-3	Z.18	5,22e-3	-1,29e-3	Z.20	2,78e-6	-1,58e-3

La interpretación de los resultados que se presentan a continuación se la hace con respecto a la media funcional.

- Para el caso que se tenga un coeficiente positivo en la primera *CPF* (Componente Principal Funcional) tendrá frecuencias superiores hasta los 33 años, entre 34 y 40 años será igual y de 41 en adelante será menor en los hombres. En las mujeres, mayor hasta los 32 años, igual de 33 a 40 y menor de 41 años en adelante.
- Para el caso que se tenga un coeficiente negativo en la primera *CPF* tendrá frecuencias inferiores hasta los 33 años, entre 34 y 40 años será igual y de 41 en adelante será mayor en los hombres. En las mujeres, menor hasta los 32 años, igual de 33 a 40 y mayor de 41 años en adelante.

- Una provincia con coeficiente positivo en la segunda *CPF*, no presentará diferencias en las edades hasta los 27 años y de 66 a 88 años, de 28 a 65 años será menor, y de 89 años en adelante será mayor para los hombres. Para las mujeres será igual hasta los 59 años, y menor de 60 años en adelante.
- Una provincia con coeficiente negativo en la segunda *CPF*, no presentará diferencias en las edades de 0 a 27 años y de 66 a 88 años, de 28 a 65 será menor, y de 89 años en adelante será mayor para los hombres. Para las mujeres será igual hasta los 59 años, y mayor de 60 años en adelante.

En la siguiente tabla se resume lo antes descrito.

Tabla 6- Interpretación de los signos en las componentes principales

	CPF1				**CPF2**			
	+		**-**		**+**		**-**	
	H	***M***	***H***	***M***	***H***	***M***	***H***	***M***
>	0-33	0-32	41-99	41-99	89-99		89-99	60-99
<	41-99	41-99	0-33	0-32	28-65	60-99	28-65	
=	34-40	33-40	34-40	33-40	0-27, 66-88	0-59	0-27, 66-88	0-59

En algunos casos, como se muestra en la tabla va a existir aparentemente ciertos conflictos con las interpretaciones dadas en los ítems anteriores, por ejemplo, si tenemos que la primera componente tiene coeficiente positivo y la segunda componente principal tienen coeficiente negativo, la interpretación dependerá de que tan grande es el coeficiente de cada componente principal, es decir, si la CPF1>CPF2, predomina la primera componente; la forma de la pirámide dependerá de que tan grande es el coeficiente de las componentes principales funcionales.

En las tablas anteriores se han omitido los resultados de los años 2011-2013, 2015-2017 y 2019, esto se lo hizo con el objeto de mejorar la visualización de los resultados, además, los resultados omitidos tienen el mismo comportamiento a los mostrados, por lo que sí es posible dar un análisis completo sobre la dinámica de la población con los resultados presentados.

Presentamos los resultados de la varianza, para cada provincia para ciertos años, omitiendo los años antes mencionados, esto con el fin de observar cuales son las provincias que tienen un comportamiento similar.

Tabla 7- Varianzas Funcionales

Prov.	var	Prov.	var	Prov.	var	Prov.	var
A.10	5,46e-5	A.14	5,00e-5	A.18	4,50e-5	A.20	4,27e-5
B.10	5,36e-5	B.14	4,56e-5	B.18	3,75e-5	B.20	3,35e-5
U.10	5,80e-5	U.14	5,07e-5	U.18	4,54e-5	U.20	4,20e-5
C.10	4,87e-5	C.14	4,18e-5	C.18	3,42e-5	C.20	3,04e-5
H.10	5,24e-5	H.14	4,71e-5	H.18	4,22e-5	H.20	3,98e-5
X.10	5,84e-5	X.14	5,08e-5	X.18	4,35e-5	X.20	3,97e-5
O.10	5,33e-5	O.14	5,00e-5	O.18	4,32e-5	O.20	4,01e-5
E.10	7,20e-5	E.14	6,36e-5	E.18	5,50e-5	E.20	5,11e-5
W.10	5,91e-5	W.14	6,35e-5	W.18	5,77e-5	W.20	5,57e-5
G.10	5,36e-5	G.14	4,90e-5	G.18	4,39e-5	G.20	4,15e-5
I.10	5,34e-5	I.14	4,82e-5	I.18	4,27e-5	I.20	3,99e-5
L.10	5,10e-5	L.14	4,56e-5	L.18	3,98e-5	L.20	3,70e-5
R.10	6,01e-5	R.14	5,50e-5	R.18	4,79e-5	R.20	4,45e-5
M.10	5,65e-4	M.14	5,43e-5	M.18	4,74e-5	M.20	4,41e-5
V.10	9,39e-5	V.14	8,18e-5	V.18	7,09e-5	V.20	6,56e-5
N.10	8,37e-5	N.14	7,54e-5	N.18	6,71e-5	N.20	6,29e-5
O.10	8,80e-5	O.14	8,08e-5	O.18	7,43e-5	O.20	7,11e-5
S.10	8,00e-5	S.14	6,91e-5	S.18	5,90e-5	S.20	5,42e-5
P.10	5,20e-5	P.14	4,68e-5	P.18	4,21e-5	P.20	3,99e-5
Y.10	6,33e-5	Y.14	5,82e-5	Y.18	5,13e-5	Y.20	4,80e-5
J.10	6,59e-5	J.14	5,99e-5	J.18	5,26e-5	J.20	4,91e-5
K.10	7,79e-5	K.14	7,01e-5	K.18	6,27e-5	K.20	5,90e-5
T.10	4,69e-5	T.14	4,16e-5	T.18	3,63e-5	T.20	3,37e-5
Z.10	7,90e-5	Z.14	7,00e-5	Z.18	6,00e-5	Z.20	5,48e-5

Nótese que las provincias de la región amazónica son las que mayor varianza presentan, en particular las provincias de Morona Santiago, Orellana y Napo, lo que significa una mayor dispersión de los datos; esta dispersión es probablemente a la mala declaración de la edad y la dificultad de recolección en esta región.

Finalmente se presenta la estimación de los coeficientes de regresión $\beta(t)$ para un modelo lineal funcional con respuesta escalar $[11]$, usando las puntuaciones de

las dos primeras componentes principales funcionales. Las líneas punteadas representan el intervalo de confianza del 95% para los valores de

La regresión de las puntuaciones de los componentes principales tiene el siguiente modelo:

$$y_i = \beta_0 + \sum c_{ij}\beta_j + \varepsilon_i$$

donde $c_{ij} = \int_R \xi_j(t)(x_i(t) - \bar{x}(t))dt$. Se tiene entonces que

$$y_i = \beta_0 + \int_R \sum \beta_j \xi_j(t)(x_i(t) - \bar{x}(t))dt + \varepsilon_i$$

Esto nos da

$$\beta(t) = \sum \beta_j \xi_j(t)$$

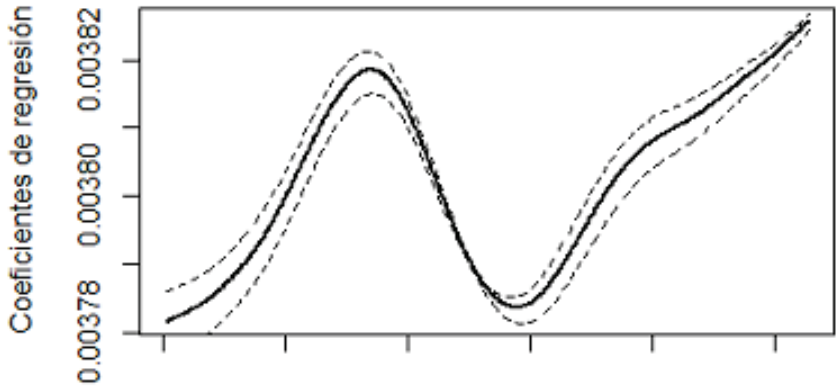

Figura 12 – Estimación de β(t) e intervalo de confianza del 95%

6 CONCLUSIONES Y RECOMENDACIONES

1. El análisis de datos funcional es una rama de la estadística que va en crecimiento tanto en las aplicaciones como en la teoría. Puede ser aplicable a todo objeto de estudio que pueda ser representado como funciones.
2. La etapa de registro es un punto débil ya que si ésta se lleva a cabo de manera incorrecta, los estadísticos descriptivos pueden llegar a variar mucho.
3. Hemos construido funciones en base a las pirámides de población para cada una de las provincias mediante técnicas de interpolación.
4. Se realiza el suavizamiento de los datos con B-Splines y su representación en bases de funciones, utilizando el criterio de validación cruzada generalizada.
5. Se aplicó un análisis de componentes principales funcionales (ACPF) a los datos de la proyección de la población ecuatoriana, con lo cual se logró realizar un análisis con respecto a su estructura demográfica para cada provincia.
6. Con el ACPF conseguimos reducir la dimensionalidad del problema original, ya que con las dos primeras componentes principales obtenidas somos capaces de explicar el 96.5% de la variabilidad. Se logró resumir los datos de una matriz de 264x200 a una de 2x200. Además, se interpreta los resultados obtenidos al aplicar el ACPF.
7. En la función media puede verse que, existe una deformación en la base de la pirámide ya que el grupo comprendido entre los 0 a 5 años lo que significa que la tasa de fecundidad irá disminuyendo con el paso de los años.
8. Se recomienda seguir profundizando con este enfoque alternativo de análisis de datos como por ejemplo: intervalos de confianza, pruebas de hipótesis, ACP para datos mixtos y métodos no paramétricos. Además, se sugiere tomar las medidas necesarias con el objetivo de tener buenos datos para su análisis.

REFERENCIAS

[1] Ramsay, J.0. & Silverman, B.W. (1997). Functional Data Analysis. Springer. New York.

[2] Green, P.J & Silverman, B.W. (1994). Nonparametric Regression and Generalized Linear Models, Chapman and Hall, London.

[3] Lange, K.(1998). Numerical Analysis for Statisticians. Springer. New York.

[4] Bonet, C.; Jorha, A.; Martinez-Seara, M.T.; Masdemont, J.; Ollé M.; Susin A.; Valencia M. (1994) Calcul numeric. Edicions UPc. Barcelona.

[5] R Development Core Team (2004). R:A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3, URL

[6] Kneip, A. & Utikal K.J. (June 2001) Inference for Density Families Using Functional Principal Component Analysis. Journal of the American Statistical Association, Vol. 96, N°.454.

[7] Locantore, N. Marron, IS. Simpson, D.G. Tripoli,N. Zhang, J.T. Cohen, K.L. (1999). Robust principal component análisis for functional data. Test-Sociedad de Estadística e Investigación Operativa. Vol 8, N°. 1, pp 1-73.

[8] Febrero-Blande, M. y Oviedo, M., Statistical Computing in Functional Data Analysis, The R Package fda.usc, Journal of Statistical Software, (2012), pp. 3-10.

[9] Oviedo, M., Utilities for Statistical Computing in Functional Data Analysis, The R Package fda.usc, Universidad de Santiago de Compostela, (2011), pp. 17-22.

[10] Antamba, L. y Medina, P., Movilidad endógena y variaciones demográficas: Una aplicación para Ecuador, Revista Analítika, INEC, Ecuador - Quito, (2011). Vol. 7, ISSN 1390 - 6208, pp. 51-71.

[11] Ramsay, J.0. & Graves, S. (2009). Functional Data Analysis with R and MATLAB. Springer. New York.

[12] Benalcázar, H., Análisis Numérico, Preprinter, Quito, 2008