UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Facultad de Estadística
Trabajo de Grado
Agosto 2018

Combinación de Análisis de Datos Funcionales Dinámicos y métodos de Análisis de Conglomerados jerárquicos para evaluar la probabilidad de transición por medio de cadenas de Márkov

Combination of Dynamic Functional Data Analysis and Hierarchical Conglomerate Analysis methods to evaluate probability of transition through Márkov chains

Daniel Ricardo Torralba Barreto.[a]
danieltorralba@usantotomas.edu.co

Wilmer Pineda Rios.[b]
wilmerpineda@usantotomas.edu.co

Resumen

El presente documento proporciona una estrategia metodológica estadística para el agrupamiento de series dependientes, usando métodos de Análisis de componentes funcionales dinámicos - ACPFD, que permita conocer la probabilidad de transición de una curva media de un grupo a otra curva media, por medio del uso de métodos de clasificación jerárquicos para el agrupamiento de curvas dependientes. Se describen los pasos para su implementación en el ambiente de R y se ilustra el uso por medio de un ejemplo de datos.

***Palabras clave*:** análisis multivariado, Análisis de componentes funcionales,cadenas de markow, clasidicación jerárquica, datos funcionales..

Abstract

This document shows an statistical methodology stretegy to grouping dependent series, using Dinamic Functional Components Analysis methods, that allow to know the transition probability of an average curve from one group to another average curve, making use of the hierarchical analysis for the grouping of dependent curves. The steps for its implementation are described in the R environment and the use is illustrated by means of a data example.

***Keywords*:** Multivariate analysis, Analysis of functional components, Markow chains, Hierarchical classification, functional data.

1. Introducción

El estudio a profundidad del uso eficiente de recursos y la estimación de determinantes que permitan predecir la ocurrencia de un evento dadas unas condiciones contingentes no observables, representan uno de los principales retos del trabajo empírico en diversas disciplinas. Esto se debe a que los determinantes de un estado contingente se encuentran más allá del modelamiento de pronóstico, dado que es necesario reconocer las características de los fenómenos observados que permitan identificar los diversos ciclos o patrones emergentes, así como su probabilidad de cambio de un estado a otro.

Desde esta perspectiva, se hace necesario reconocer que los cambios tecnológicos observados en las últimas décadas han desencadenado un sin número de cambios en la manera en que se interrelacionan las

[a] Estudiante de Estadística, U. Santo Tomás, sede Bogotá
[b] Profesor Asociado. Faculta de Estadística U. Santo Tomás, sede Bogotá

sociedades contemporáneas, acompañadas de tecnologías cada vez más eficientes y de mayor capacidad. Aunque estas nuevas tecnologías son la principal herramienta para facilitar las dinámicas de interacción, también se establecen como herramientas para el monitoreo y medición de características observables útiles para la toma de decisiones.

Esta evolución tecnológica ha permitido que el análisis estadístico cambie en algunos de los paradigmas clásico, donde se cuenta con conjuntos de datos muy grandes observados en un intervalo de tiempo continuo, como por ejemplo en el campo de la medicina, donde un cardiograma obtiene cerca de 10.000 observaciones en tan solo un minuto, permitiendo dibujar una curva para observar el comportamiento cardiovascular.

En este sentido, se han planteado diversos métodos estadísticos que permiten develar patrones de series de tiempo. Sin embargo, la literatura muestra que uno de los métodos más utilizados recientemente es el Análisis de Datos Funcionales -ADF-, destacando i) el uso de una función generada por los datos en vez un conjunto ordenado de datos, ii) representar los datos de manera que ayuden a un mayor análisis, iii) mostrar los datos para resaltar varias características iv) estudiar fuentes importantes de patrones y variaciones entre los datos, v) para comparar dos o más conjuntos de datos con respecto a ciertos tipos de variaciones.

Existen diversas aplicaciones en los campos de agronomía, como por ejemplo el estudio de Hamdan et al. (2013), en el que observan las actividades de recursos hídricos afectadas por los elementos del cambio climático. En el campo de economía, Borrajo (2017) observó por medio del análisis de datos funcionales patrones de consumo de tarjetas de crédito.

Por otro lado, se debe indicar que el método de ADF no permite observar la dependencia dinámica de toda la serie de tiempo, generando reducciones de dimensiones inadecuadas en un ajuste de series de tiempo. No obstante, Hörmann et al. (2015) realizan una propuesta denominada Análisis de Componentes Principales Funcional Dinámico -ACPFD-, basada en la teoría de Brillinger de componentes principales dinámicos funcionales, que se fundamenta en un enfoque de dominio de frecuencia, permitiendo observar la estructura de dependencia en series de tiempo.

A menudo se utilizan como un paso preliminar técnicas de reducción de datos dependiendo de su naturaleza, por ejemplo, para datos de series de tiempo, es usado comúnmente el método ADF, con el fin de reducir el problema de los datos con dimensional infinita a uno finito. Sin embargo, la reducción de datos por medio de este método es ineficiente dado que, como se mencionó anteriormente, este no permite observar la dependencia dinámica de la serie de tiempo.

Al realizar la revisión bibliográfica, se observan estudios de series de tiempo basadas en el Análisis de Componentes Principales Funcional Dinamico - ACPFD - para la identificación la reducción de datos y para el análisis de agrupamientos se basan en el Análisis de Datos Funcional -ADF-.

Por lo tanto, una adecuada reducción de datos para generar agrupamientos de series de tiempo debe realizarse por medio de la implementación de ACPFD, generando grupos consistentes, homogéneos e independientes, para posteriormente implementar métodos de agrupamientos como agrupamientos jerárquico o k-medias.

En este sentido, el presente documento busca propone una estrategia para la combinación del método ACPFD, el análisis de conglomerados jerárquica e implementación de cadenas de Márkov para evaluar la probabilidad de transición de una curva media pueda cambiar de agrupamiento.

Este documento se encuentra divido en 3 secciones; la primera sección es la anterior introducción, la sección dos se resume la estrategia propuesta, y finalmente en la sección cuatro se realiza una implementación del método.

2. 2. Estrategia combinada.

En esta sección se presentan los métodos que sustentan la estrategia propuesta y como estos es articulan, tal como se presentan en la figura 1.

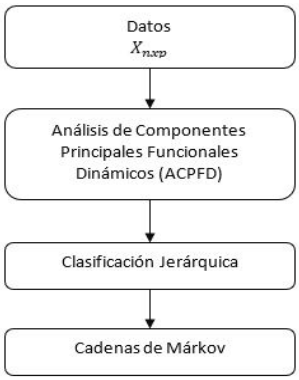


Figura 1: Diagrama de flujo para implementación de estrategia de combinación del método ACPFD, el análisis de conglomerados jerárquica e implementación de cadenas de Márkov

2.1. Datos funcionales

En la actualidad existe fenómenos que son observados en un intervalo de tiempo continuo, como por ejemplo en el campo de la medicina, donde un cardiograma obtiene cerca de 10.000 observaciones en tan solo un minuto, permitiendo configurar una curva para evaluar el comportamiento cardiovascular.

Este tipo de datos son comunes en las diversas áreas del conocimiento, como en finanzas donde los datos de cotización de una acción en la bolsa de valores forman una curva del comportamiento de la acción, o en el área de las telecomunicaciones, donde se observan datos continuos del comportamiento de red que configuran una curva de consumo de red, entre otras.

Frente a este nuevo paradigma, surge como alternativa estadística el análisis de datos funcionales, el cual permite abordar los mismos problemas de la estadística clásica desde otra perspectiva, tales como; i) la exploración y descripción de datos para destacar o develar características relevantes, ii) explicar y modelar la relación entre una variable dependiente y una independiente, iii) métodos de clasificación supervisada o no supervisada de un conjunto de datos, así como iv) contraste de hipótesis y predicción.

Como lo define Febrero (2008), tomando como referente a Ferraty & Vieu (2006), una variable aleatoria X se dice que es una variable funcional si toma valores en un espacio funcional ϵ (espacio normado o seminormado completo). Además, un conjunto de datos funcionales $x_1, \ldots, x_n$, es la observación de n variables funcionales $x_1, \ldots, x_n$, idénticamente distribuidas.

Febrero (2008), indica que los datos funcionales son usados comúnmente en el espacio $L^2[S]$, lo cual indica que las funciones de cuadrado integrable en el intervalo $S = [a, b] \subset \mathbb{R}$.

Del mismo modo Febrero (2008) destaca que al trabajar con datos funcionales, se debe encontrar una representación adecuada para los datos, dado que si se observar los datos de manera directa en una gráfica de dos ejes, este no permite observar patrones o características relevantes, por lo cual, se hace necesario la limpieza de las curvas por medio de un cambio de escala.

Este cambio de escala se realiza por medio del uso de bases ortonormales, por su eficiencia en espacios $L^2[S]$, permitiendo así la representación de cada dato funcional en la base usando aquellas coordenadas que son más significativas. Debido a la alta dimensionalidad de los datos funcionales, se elige un número K para representar los datos en un subespacio.

El parámetro K, se configura como un parámetro de suavizamiento de los datos funcionales, aunque no existe una regla universal para la selección adecuada del paramento K, sin embargo, el uso de un k pequeño garantiza un modelo manejable, pero con perdida de información relevante y el uso de un k grande garantiza una muy buena representación de los datos, pero con dimensiones muy grandes (Febrero 2008).

Del mismo modo no existe una regla general para la selección de la base, sin embargo como lo sugiere Febrero (2008), para datos de tipo periódico se suele implementar la base Fourier y para los datos no periódicos se implementa la base B-SPLINE .

Finalmente existe una técnica basada en la expansión de karhunen loéve, denominada Componentes Principales Funcionales (FPCA), la cual es una extensión del método de análisis de componentes principales multivariante, las cuales haciendo uso del operador de momentos de segundo orden muestral permite calcular las autofunciones y autovalores para generar una base ortonormal adaptada para los datos.

2.2. Producto escalar

El análisis descriptivo funcional, como lo describe Ramsay & Silverman (2007), puede provenir de un vector o una función que se puede describir como un producto escalar.

Pérez (2004), define como producto escalar euclidiano de elementos x e y del espacio vectorial E, denotado como $< x, y >$, el cual se puede interpretar como medida de asociación entre dos elementos del espacio vectorial observado.

Definición 1:

$$<\cdot,\cdot>:\mathrm{E}x\mathrm{E}\rightarrow \Re(\mathrm{x,y}) \rightarrow<\mathrm{x,y}> \tag{1}$$

Propiedades

1. Simetría: $< x, y >=< y, x >$ para todo x e y.

2. Positividad: $<x,x>\geq 0 para todo x, con$ <x,y>= 0 $\Leftrightarrow$x= 0.

3. Bilinealidad: Para todo a,b,$\in \Re$ <ax+by,z>= a <x,y>+b<y,z>

El producto escalar euclidiano general se puede expresar como, $< x, y >_w$, la cual dependerá del elemento del espacio vectorial observado, es decir, será una aplicación sobre el espacio que satisfacen las tres propiedades. (Pérez 2004).

En el contexto de análisis de datos funcionales Pérez (2004), define W como una función de pesos definida positiva y los elementos x e y son funciones de $\Re$.

Definición 2:

Sea el espacio de funciones de cuadrado integrable.

$$\mathrm{L}^2 = \{\,\mathrm{f}\colon \Re \to \Re : \int_{\Re} \mathrm{f}^2(t)\mathrm{y}(t)dt < \infty \tag{2}$$

Por lo tanto, el producto escalar euclidiano en L^2 se puede definir como:$< x, y >= xy = (t)y(t)dt$.

Es decir, que el producto escalar se puede observar como una relación entre las funciones $x(t)$ e $y(t)$.

Propiedades

1. Simetría: $<x,y>=x´y = \int \mathrm{x}(t)\mathrm{y}(t)dt = \int \mathrm{y}(t)\mathrm{x}(t)dt =< y, x >$

2. Positividad: $<x,x>= x´x = \int \mathrm{x}(t)\mathrm{x}(t)dt \geq 0$

3. Bilinealidad: Para todo $a, b \in \Re (ax + ay)'z = \int \left[a \cdot x(t) + b \cdot y(t)\right] z(t)dt = a < x, z > +b < y, z >$.

EL producto escalar general busca dar mayor peso a unos intervalos de (t) y quitar relevancia a otros, por medio de la función w(t).

2.3. Estadisticos descriptivos univariados

Como en la mayoria de los métodos estadísticos, una de las principales tareas es lograr resumir los datos en algunas medidas, permitiendo describir su comportamiento, el análisis de datos funcionales tambien permite realizar este tipo de deducciones, por medio del anális de la media y varianza de una función $x(t)$.

Las expresiones presentadas a continuación son extraidas de Ríos et al. (2017) , basados en Ramsay & Silverman (2005).

Definición 3:

Media de una función $x(t)$ o tambien llamado valor medio de la función x(t) como:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} f_i(t) \tag{3}$$

La media de x(t)(x) representa la tendencia central y el nivel medio de todos los valores de $x(t)$.

Definición 4:

La varianza de x(t), se denota como

$$S^2(x(t)) = \frac{1}{n-1}\sum_{j=1}^{n}(x_j(t) - \overline{x(t)})^2 \tag{4}$$

Como lo indica Pérez (2004), representa la variación media al cuadrado de todos los valores de la función respecto a su valor medio, es decir que una función que presente una alta variabilidad mostrara una área grande entre la función y la función de valor medio.

Definición 5:

La función de desviación estandar muestral de $x(t)$ se denota como:

$$Stdev_{x(t)}(t) = \sqrt{Var_{x(t)})} \tag{5}$$

Definición 6:

Si $x_1(t), x_2(t), ..., x_N(t)$ es una muestra de funciones de una función aleatoria $x(t)$, definida en $[0, T]$, entonces podremos definir la función de covarianza muestral de $x(t)$ entre t_1 y t_2 como:

$$Cov_{x(t)}(t_1, t_2) = \frac{1}{n-1}\sum_{j=1}^{N}(x_j(t_1) - \overline{x}(t_1))'(x_i(t_2) - \overline{x}(t_2)) \tag{6}$$

Definición 7:

Si $x_1(t), x_2(t), ..., x_N(t)$ es una muestra de funciones de una función aleatoria $x(t)$, definida en $[0, T]$, entonces podremos definir la función de correlación muestral de $x(t)$ entre t_1 y t_2 como:

$$\text{Corr}_{x(t)}(t_1, t_2) = \frac{Cov_{x(t)}(t_1,t_2)}{\sqrt{Var_{x(t)}(t_1 \cdot Var_{x(t)(t_2)})}} \tag{7}$$

2.4. Análisis de Componentes Principales Funcionales Dinámicos (ACPFD)

Los datos funcionales usualmente generan curvas aleatorias principalmente uniformes, donde se asume que $U \sim [0, 1]$, donde cada observación es una curva, es decir por cada observación se cuenta con una función, lo cual implica que se observan datos con dimensión intrínsecamente infinita, lo cual es el principal objetivo del Análisis Funcional de componentes principales (ACPF).

El ACPF, es una analogía del análisis de componentes principales (ACP) multivariados, lo cual implica que dependan de una descomposición propia de la función de covarianza Ramsay & Silverman (2007).

Hörmann et al. (2015), señala que el método de ACPF está diseñado principalmente para observaciones independientes, es decir+ que este método operada de manera estática, omitiendo la información potencial y valiosa presentada en los datos pasados de la serie funcional, lo cual puede llevar a conclusiones parciales o complemente erróneas.

Por lo tanto, para contemplar la dependencia de la serie funcional surge como respuesta el Análisis de Componentes Principales Funcionales Dinámicos-ACPFD-, con el objetivo de convertir una serie temporal funcional en una serie temporal vectorial, donde los componentes generados están incorrelacionados y presenten la mayor parte de la dinámica y varianza de proceso original.

Hörmann et al. (2015), presenta las siguientes definiciones que soportan el ACPFD.

Definición 8:

Asumiendo que es posible seleccionar algunas funciones$(\Phi_{m\ell} : \ell \in \mathbb{Z})$, tal que, $\Phi^*_m(\theta) := \varphi_m(\theta)$

$$\Phi^*_m(\theta) := \sum_{\ell \in \mathbb{Z}} \Phi_{m\ell} e^{\iota \ell \theta} = \varphi_m(\theta) \tag{8}$$

Donde $\Phi^*_m(\theta)$ es la función de valores propios y $\Phi_{m\ell}$ son los coeficientes de la expansión de Fourier, of $varphi_m(\theta)$ como una función de θ. (Para una explicación más detallada de la expansión de Fourier, ver Hörmann et al. (2015), página 9.)

Definición 9:

Asumiendo que $X_t : t \in \mathbb{Z}$, es un proceso estacionario con media cero y con valores en L_H^2,la m-ésima puntuación del componente principal funcional dinámico (ACPFD) de x_t es,

$$Y_{mt} := \sum_{\ell \in \mathbb{Z}} < X_{t-\ell, \Phi_{m\ell}} > t \in \mathbb{Z}, m \geq 1. \tag{9}$$

Donde $\Phi_m := (\Phi_{m\ell} : \ell \in \mathbb{Z})$

Del mismo modo (Hörmann et al. 2015) destaca cinco propiedades relevantes:

Propiedades fundamentales

1. Las funciones propias $\varphi_m(\theta)$ son Hermitianas, y por lo tanto Y_t es real.
2. Si $C_h = 0$ para $h \neq 0$, las puntuaciones ACPFD coinciden con las puntuaciones de FPC estáticas.
3. La serie que define Y_{mt} es media-cuadrada convergente.
4. Para $m \neq m'$, las puntuaciones ACPFD Y_{mt} y $Y_{m's}$ no están correlacionadas para todos los s, t.

   Dado que los componentes generados por el ACPFD están incorrelacionados, es posible usarlos para análisis posteriores en métodos estadísticos clásicos que requieran esta condición, adicionalmente de manera análoga al método de ACPF, las curvas se pueden reconstruir haciendo uso de la expansión de Karhunen-Loèv.
5. La variación a largo plazo de la n-ésima secuencia de puntuación ACPFD es:
   $\lim_{x\to\infty} \frac{1}{n} Var(Y_{m1} + ... + Y_{mn}) = 2\Pi\lambda_m(0)$

2.5. Clasificación Jerárquica

El método descrito de ACPFD, puede ser aprovechado para la segmentación de curvas funcionales, por medio de la incorrelación de los vectores generados, es posible el uso del método multivariante de clasificación jerárquica, garantizando las propiedades necesarias para su implementación.

La combinación de los dos métodos se da de la siguiente manera:

1. Obtención de los componentes principales dinámicos por medio de la implementación del método de ACPFD.
2. Se efectúa una clasificación ascendente jerárquica donde los elementos terminales del árbol son las n clases de la partición de las curvas funcionales iniciales. El árbol correspondiente se construye según el método de clasificación deseado, los más comunes son:
3. A partir de la clasificación de cada curva y dada su naturaleza dependiente se reconstruye las curvas funcionales, con el objetivo de conocer la agrupación de cada curva en un grupo.

El resultado de esta combinación permite conocer la estructura o ciclo de curva funcional a lo largo de la serie funcional, permitiendo identificar agrupamientos deseables en un fenómeno.

Agregación	Formula
Salto mínimo o más cercano	$\delta_{\min}(A,B) = \min\{d(\mathbf{x}_i,\mathbf{x}_j)\|\mathbf{x}_i \in A, \mathbf{x}_j \in B\}$
Salto máximo	$\delta_{\max}(A,B) = \max\{d(\mathbf{x}_i,\mathbf{x}_j)\|\mathbf{x}_i \in A, \mathbf{x}_j \in B\}$
Salto promedio	$\delta_{\text{prom}}(A,B) = \frac{1}{\|A\| \times \|B\|} \sum_{\substack{\mathbf{x}_i \in A \\ \mathbf{x}_j \in B}} d(\mathbf{x}_i,\mathbf{x}_j)$
Ward	$\delta_{\text{ward}}(A,B) = \frac{\|A\|\|B\|}{\|A\|+\|B\|}\|\|\mathbf{g}_A - \mathbf{g}_B\|\|^2$

Tabla 1: Métodos de clasificación jerárquica.

2.6. Cadenas de Márkov.

Finalmente, haciendo uso de la agrupación de las curvas funcionales generadas por la clasificación jerárquica, se construye la matriz de transición, la cual formalmente se construye por medio de la probabilidad de que el sistema esté en un estado particular en un periodo de observación, depende solamente de su estado en el periodo de observación inmediatamente anterior.

Probabilidad de transición, lo define Kemeny & Snell (1983) como la probabilidad de ir del estado i al estado j en n unidades de tiempo. Se denota como t_{ij}, además, esta probabilidad no cambia con el tiempo.

Propiedades:

- Al ser una probabilidad se tiene que
$$0 \leq t_{ij} \leq 1$$
$$1 \leq i, j \leq n$$
- Si el sistema está en el estado j en cierto periodo de observación, entonces debe estar en alguno de los n estados
$$t_{1j} + t_{2j} + \cdots + t_{nj} = 1$$

Estas probabilidades se deben incorporar en una matriz cuadrada de transición $T = [t_{ij}]$, cuyas entradas son no negativas y según la propiedad anterior las filas suman 1.

Vector de estado del proceso de Markov

Sea

$$x^k = \begin{bmatrix} p_1^{(k)} \\ p_2^{(k)} \\ \vdots \\ p_n^{(k)} \end{bmatrix}$$

donde

- $k \geq 0$ y es definido como el periodo de observación.
- $p_j^{(k)}$ es la probabilidad de que el sistema se encuentre el estado j en el periodo k. Cuando $k = 0$ el vector x^0 es definido como el vector de estado incial.

Proceso de Markov Regular

Una matriz de transición T de un proceso de Markov es regular si todas las entradas de alguna potencia de T son positivas. Un proceso de Markov es regular si su matriz de transición es regular (Kemeny & Snell 1983).

Teorema 1

Si T es la matriz de transición de un proceso de Markov regular, entonces:

(a) A medida que $n \to \infty$, T^n tiende a una matriz

$$A = \begin{bmatrix} u_1 & u_1 & \cdots & u_1 \\ u_2 & u_2 & \cdots & u_2 \\ \vdots & \vdots & & \vdots \\ u_n & u_n & \cdots & u_n \end{bmatrix}$$

donde todas las columnas de la matriz son idénticas.

(b) Toda columna de A es un vector de probabilidad tal que todos sus componentes con positivos. Es decir, $u_i > 0$ $1 \leq i \leq n$ y $u_1 + u_2 + \cdots + u_n = 1$.

Teorema 2

Si T es una matriz de transición regular y A y u son como en el *Teorema 1*, entonces: (a) Para cualquier vector de probabilidad x, $T^n x \to u$ conforme $n \to \infty$, de modo que u es un vector de estado estacionario. (Kemeny & Snell 1983). (b) El vector de estado estacionario u es el único vector de probabilidad que satisface la ecuación matricial $Tu = u$.

Procedimientos para calcular el vector de estado estacionario

1. El primer procedimiento para calcular el vector de estado estacionario u de una matriz de transición regular T es el siguiente. Paso 1. Calcular las potencias $T^n x$, donde x es cualquier vector de probabilidad. Paso 2. u es el límite de las potencias $T^n x$.
2. El segundo procedimiento para calcular el vector de estado estacionario u de una matriz de transición regular T es el siguiente. Paso 1. Resolver el sistema homogéneo:

$$(I_n - T)u = 0$$

Paso 2. De la infinidad de soluciones obtenidas en el paso 1, determinamos una única solución u, al exigir que sus componentes satisfagan la ecuación $u_1 + u_2 + \cdots + u_n = 1$.

3. Aplicación del método

3.1. Análisis de precipitaciones del Rio Suta.

Objetivo: Conocer la probabilidad de cambio la clasificación según la intensidad de la precipitación del rio Suta en la estación hidrológica del municipio de Tausa respecto a la estructura de los datos, usando como marco común los cinco (5) estados contingentes esperados propuesto por Viñas Rubio & López (2015).

Datos: Los datos usados para este análisis son tomados de la página de datos abiertos del gobierno nacional. La fuente responsable de los datos es la Corporación Autónoma Regional de Cundinamarca - CAR.

Clase
Debiles
Moderadas
Fuertes
Muy fuertes
Torrenciales

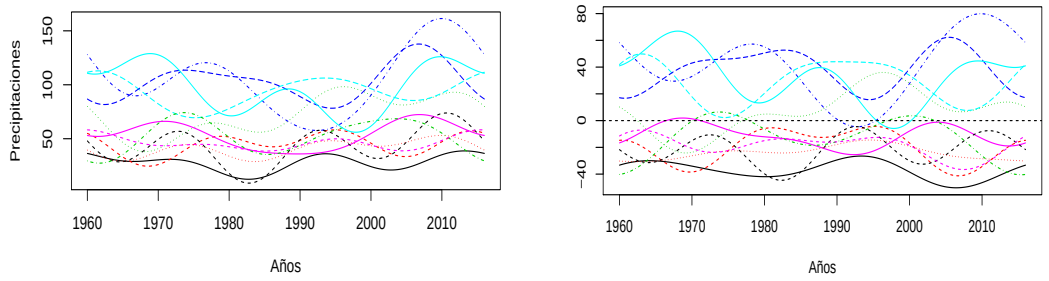
Tabla 2: Clasificación de la precipitación según la intensidad.

Los datos muestran el volumen de precipitaciones mensuales del municipio de Tausa desde el año 1960 a 2016, obtenidos en la red de estaciones hidrológicas de la CAR.

Procedimiento de análisis:

El primer paso consiste en la construcción de la serie de datos a datos funcionales, transformando los datos por medio de la base B-SPLINE, como se presenta en la figura 2. Donde se observa la construcción de 12 curvas funcionales en un intervalo de tiempo de 1960 a 2016.

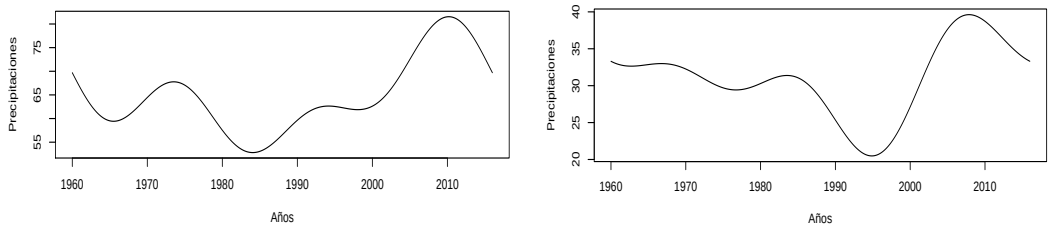
Se destaca que han existido precipitaciones del rio altas o torrenciales en la ventana de observación principalmente en los meses de octubre y diciembre, además de precipitaciones bajas o débiles en los meses de enero y julio.



(a) Datos Suavizados Funcionales Dinamico (b) Datos Centrados Dinamicos

Figura 2: Datos funcionales con base B-SPLINE vs Datos funcionales centrados con base B-SPLINE

Estadísticos descriptivos



(a) Función de curva media (b) Función de curva de desviación

Figura 3: Estadísticos funcionales descriptivos

Al observar la figura 3 de estádisticos descriptivos, se evidencia que los años con menor precipitación fueron en 1983, 1987, 2001 y 2015, por otro lado los años con mayor precipitación fueron en los años 1979,2005 y 2011. Además los años con mayor volatilidad en las precipitaciones observadas fueron 1963,1965,1993 y 2011.

Al observar la figura 4 se evidencia que la función de correlación muestral presenta altas correlaciones con los tres años anteriores o siguientes, además se evidencio que los años 1960 hasta 1963 presentan una correlación superior a 0.98 con los años 2012 a 2015.

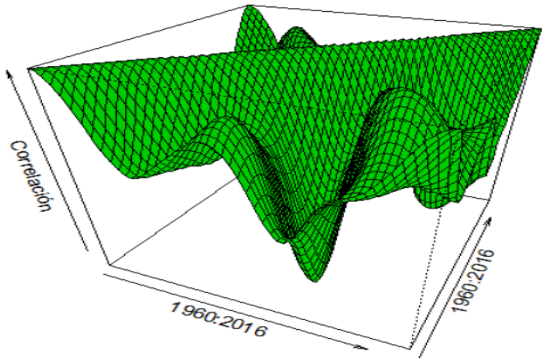


Figura 4: Función de correlación muestral

Análisis de Componente Principales Dinámicos

Al generar el análisis de componente principales dinámicos, se presentan los dos primeros componentes dinámicos funcionales, dado que acumulan el 95.7 % de la varianza acumulada, las cuales se presentan en la figura 5. Se destaca que el primer componente acumula el 74.3 % de la varianza de las curvas funcionales, la cual explica principalmente precipitaciones elevadas y constantes hasta el año 1980, luego una caída precipitaciones cercanas al año 1995, cuando las precipitaciones continúan precipitaciones elevadas.

El segundo componente acumula 21.4 % de la varianza restante, explica las precipitaciones bajas con un pico entre 1980 y el 2000.

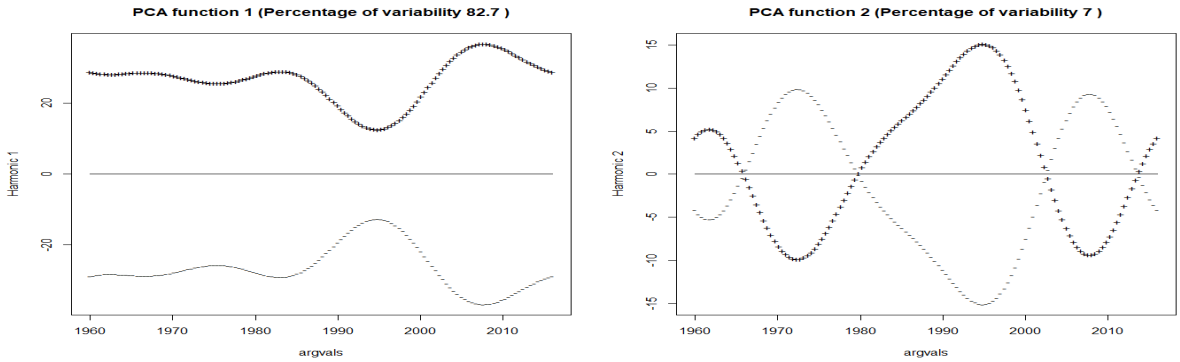
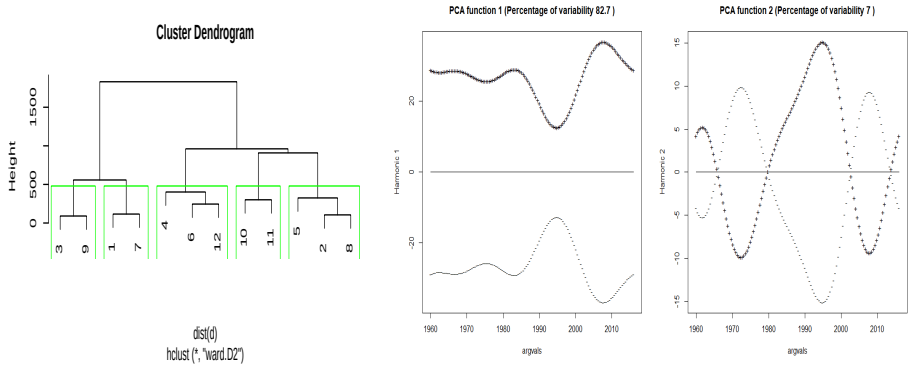


Figura 5: Componentes principales dinámicos

Con el objetivo de agrupar las curvas funcionales, se realizó el método de clasificación haciendo uso de los dos primeros componentes, permitiendo conocer los meses en los cuales se presentan el mismo comportamiento. En la figura 6 se presenta el dendrograma y las curvas funcionales de cada mes y su posible grupo.

Se evidencia que los meses de enero y julio presentan el mismo comportamiento de precipitaciones bajas, como octubre y diciembre presenta precipitaciones altas.



(a) Dendograma (b) Curvas segmentadas

Figura 6: Dendograma de las curvas funcionales vs agrupaciones de curvas funcionales

Construcción de matriz de transición:

Una vez obtenida la agrupación jerárquica adecuada, se observa el grupo de clasificación del mes observado, como se evidencia en la tabla 3.

Enero	Febrero	Marzo	Abril	Mayo	Junio
1	2	3	4	2	4
Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre
1	2	3	5	5	4

Tabla 3: Clasificación de los meses observados.

Posteriormente, haciendo uso de la clasificación y de las propiedades de dependencia entre los meses observados, se procede a la construcción de la matriz de transición por medio de las frecuencias de ocurrencia de los estados contingentes generados, como se presentan en la figura 6.

Se observa en la figura 7, que la curva funcional de enero se encuentra asignada en la clase 1, es decir de precipitaciones débiles, del mismo modo se encuentran la curva de julio. La mayoría de las curvas presentan cambios en su estado respecto al siguiente mes con excepción de octubre.

Las demás curvas presentan para el siguiente mes un cambio en su estado contingente, por ejemplo, de marzo a abril pasa de estar en un estado de precipitaciones fuertes a precipitaciones muy fuertes.

Se debe destacar que no se evidencio cambios del estado de precipitaciones débiles a precipitaciones fuertes ni torrenciales, lo cual puede corresponder la temporalidad de la información.

Finalmente, al observar la matriz de transición presentan en la figura 8, se evidencia que la probabilidad de estar en el estado de precipitaciones débiles a muy fuertes es del 33.33 %, además al largo plazo la probabilidad de estar en el estado de precipitaciones débiles es de 13.04 %

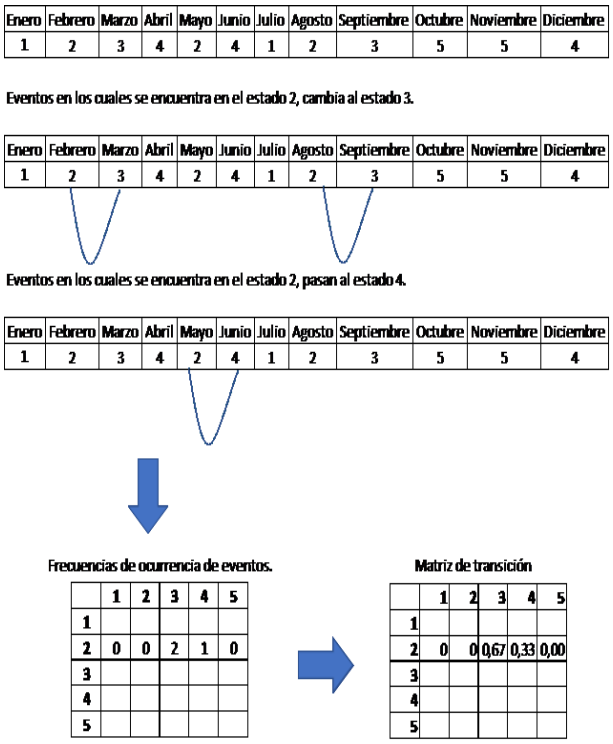


Figura 7: Construcción de la matriz de transición.

Matriz de transición

	1	2	3	4	5
1	0	1	0	0	0
2	0	0	0,67	0,33	0
3	0	0	0	0,5	0,5
4	0,5	0,5	0	0	0
5	0	0	0	0,5	0,5

Estado estacionario

1	2	3	4	5
0,13	0,26	0,17	0,26	0,17

Figura 8: Matriz de transición y estados estacionarios.

Referencias

Borrajo, L. L. (2017), *Aplicación de análisis de datos funcionales en medios de pago*, Universidad de Santiago de Compostela.

Febrero, M. (2008), 'A present overview on functional data analysis.', *Boletín de Estadística e Investigación Operativa. BEIO* **24**(1), 6–12.

Ferraty, F. & Vieu, P. (2006), *Nonparametric functional data analysis: theory and practice*, Springer Science & Business Media.

Hamdan, M. F., Suhaila, J. & Jemain, A. A. (2013), 'Functional data analysis technique on daily rainfall data: A case study at north region of peninsular malaysia', *Matematika* **29**, 233–240.

Hörmann, S., Kidziński, Ł. & Hallin, M. (2015), 'Dynamic functional principal components', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **77**(2), 319–348.

Kemeny, J. G. & Snell, J. L. (1983), *Finite Markov chains: with a new appendix"Generalization of a fundamental matrix*

, Springer.

Pérez, V. N. (2004), Análisis de datos funcionales: implementación y aplicaciones, PhD thesis, Universitat Politècnica de Catalunya. Facultat de Matemàtiques i Estadística. Departament d'Estadística i Investigació Operativa, 2004 (Llicenciatura de Ciències i Tècniques Estadístiques).

Ramsay, J. O. & Silverman, B. W. (2005), *Functional data analysis*, Springer.

Ramsay, J. O. & Silverman, B. W. (2007), *Applied functional data analysis: methods and case studies*, Springer.

Ríos, W. P., Ramírez, A. C. & Escobar, O. G. (2017), 'Análisis de datos funcionales aplicado en electroencefalogramas: agrupamiento por k-medias funcional', *Comunicaciones en Estadística* **10**(1), 129–144.

Viñas Rubio, J. M. & López, M. (2015), 'Nuevo manual de uso de términos meteorológicos de aemet'.