

Análisis estadístico de datos funcionales y su aplicación en el estudio del PM10 en Bogotá, Colombia

Sebastián Calcetero Maria Elsa Correal

Departamento de Ingeniería Industrial
Universidad de los Andes

Diciembre, 2017

Outline

- 1 Introducción
- 2 Marco conceptual de datos funcionales
 - Curvas aleatorias
 - Suavizamiento
 - Componentes principales funcionales
 - Series de Tiempo Funcionales
 - Regresión Funcional
- 3 Análisis estadístico del PM10 como un dato funcional
 - El PM10
 - Análisis descriptivo funcional
 - Componentes principales funcionales
 - Análisis de serie de tiempo funcional
 - Regresión funcional con errores FARMA
- 4 Conclusiones

Introducción

Contexto

- El análisis de datos funcionales ha crecido en los últimos años dado su potencial de aplicación en diversas áreas, en especial en aquellas en que se requiere de modelos parsimoniosos debido a la gran cantidad de información disponible.
- Un dato funcional se puede entender como la **observación de una función $X(s)$, $s \in S$** , donde S corresponde a un intervalo continuo como el tiempo o el espacio.

Introducción

Contexto

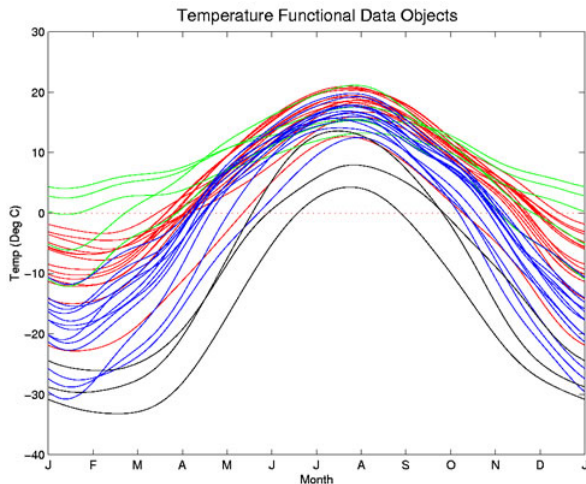


Figura: Ejemplo de datos funcionales

Introducción

Objetivos

- Se han realizado diversos estudios para el óptimo tratamiento de los datos funcionales. De forma general, las monografías que más impacto han tenido en la comunidad académica sobre el tema de datos funcionales son [2], [5], [1],[3] y [4].
- El **objetivo** de este proyecto es por un lado mostrar un marco conceptual en el que se evidencien las ideas de mayor influencia que han llevado al desarrollo de los datos funcionales en los últimos años, y por otro ilustrar estas pueden llevar a conclusiones de interés en el contexto del PM10.

Outline

- 1 Introducción
- 2 Marco conceptual de datos funcionales
 - Curvas aleatorias
 - Suavizamiento
 - Componentes principales funcionales
 - Series de Tiempo Funcionales
 - Regresión Funcional
- 3 Análisis estadístico del PM10 como un dato funcional
 - El PM10
 - Análisis descriptivo funcional
 - Componentes principales funcionales
 - Análisis de serie de tiempo funcional
 - Regresión funcional con errores FARMA
- 4 Conclusiones

Curvas aleatorias

Definición

- Un dato funcional se puede entender como la realización de un **proceso estocástico** en tiempo continuo cuyas trayectorias pertenecen al espacio $L_2(S)$. Una base de datos funcionales no es más que diferentes realizaciones de un proceso estocástico dado.
- Con base a esto, se pueden definir los análogos a la media, la varianza y la correlación para variables aleatorias funcionales. Su interpretación puntual es la misma que se da con datos escalares.

- Función de Media

$$\mu(s) = E(X(s)) = \int X^{(\omega)}(s) d\mathcal{P}(\omega)$$

.

- Función de Covarianza

$$c(s, r) = E(X(s)X(r)) - E(X(s))E(X(r))$$

- Función de Correlación

$$\rho(s, r) = \frac{c(s, r)}{\sqrt{c(s, s)c(r, r)}}$$

Curvas aleatorias

Base funcional

Un concepto básico, pero fundamental del espacio L_2 es el de base:

Base Funcional

Una sucesión de funciones $\{\phi_j(s)\}_{j=1}^{\infty}$ es una *base* de L_2 si toda función $X(s) \in L_2$ puede representarse de forma única como:

$$X(s) = \sum_{j=1}^{\infty} a_j \phi_j(s)$$

para algunos escalares $\{a_j\}_{j=1}^{\infty}$

Esta representación para las funciones es lo que determina todo el paradigma de tratamiento para los datos funcionales en este documento. En particular, **llevar un problema de datos funcionales, a uno de estadística multivariada.**

Outline

- 1 Introducción
- 2 Marco conceptual de datos funcionales
 - Curvas aleatorias
 - **Suavizamiento**
 - Componentes principales funcionales
 - Series de Tiempo Funcionales
 - Regresión Funcional
- 3 Análisis estadístico del PM10 como un dato funcional
 - El PM10
 - Análisis descriptivo funcional
 - Componentes principales funcionales
 - Análisis de serie de tiempo funcional
 - Regresión funcional con errores FARMA
- 4 Conclusiones

Suavizamiento

Motivación

- En la práctica no se observa como tal la forma funcional del dato $X(s)$, sino su valor para ciertos puntos esparcidos en el dominio S .
- Se observan los valores asociados a $X(s_1), \dots, X(s_m)$ para algunos $s_1, \dots, s_m \in S$.
- Por lo tanto, la primera etapa en un análisis de datos funcionales consiste en recuperar la función $X(s)$ a partir de dichas observaciones puntuales, proceso que se conoce como **suavizamiento** de la función.

Suavizamiento

Idea

- El enfoque consiste en utilizar la estructura de espacio vectorial de $L_2(S)$ para recrear la función con **bases funcionales**.
- En ese orden de ideas, se plantea un modelo de la forma:

$$X(s_l) = \sum_{j=1}^K a_j \phi_j(s_l) + \varepsilon_l = \phi^T(s) \mathbf{a} + \varepsilon_l, \forall l \in \{1, \dots, m\}$$

La base es truncada a un número finito de términos K , y el modelo corresponde a una **regresión lineal**.

- Las bases más populares para realizar el suavizamiento son las bases de *Fourier* y *B-splines*

Suavizamiento

Idea

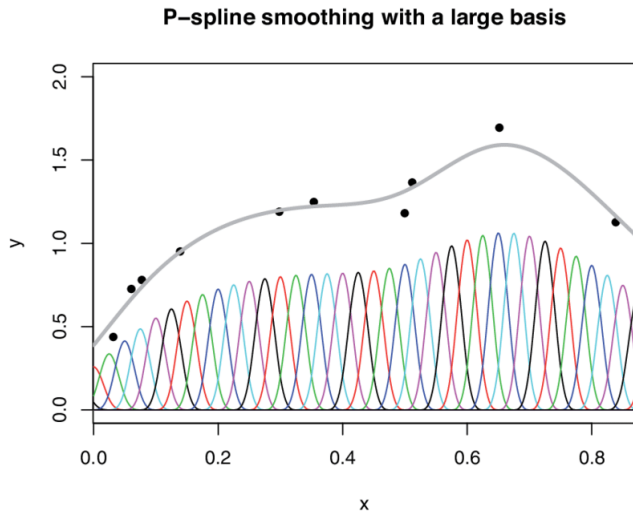


Figura: Ejemplo de suavizamiento

Outline

- 1 Introducción
- 2 Marco conceptual de datos funcionales
 - Curvas aleatorias
 - Suavizamiento
 - Componentes principales funcionales
 - Series de Tiempo Funcionales
 - Regresión Funcional
- 3 Análisis estadístico del PM10 como un dato funcional
 - El PM10
 - Análisis descriptivo funcional
 - Componentes principales funcionales
 - Análisis de serie de tiempo funcional
 - Regresión funcional con errores FARMA
- 4 Conclusiones

Componentes principales funcionales

Motivación

- Al realizar el suavizamiento se utiliza una base arbitraria y poco informativa. Una pregunta natural se asocia a encontrar una base “óptima”, que permita la interpretación de los términos de la base de forma eficiente.
- El objetivo es hallar $\xi_j(s)$ y α_{ij} tal que se tenga la siguiente descomposición de forma óptima:

$$X_i(s) = \mu(s) + \sum_{j=1}^P \alpha_{ij} \xi_j(s)$$

Es decir, un comportamiento “común” $\mu(s)$, que es afectado por distintas fuentes de variación $\xi_j(s)$ según un peso α_{ij} .

Componentes principales funcionales

Estimación

- Para un P dado, este problema es equivalente a encontrar coeficientes $\{a_j\}$ y funciones $\{\xi_j\}$ que minimicen la distancia:

$$\frac{1}{n} \sum_{i=1}^n \|X_i(s) - \sum_{j=1}^P \alpha_{ij} \xi_j(s)\|^2$$

- Esta situación es equivalente al problema de *componentes principales* del contexto multivariado. Más aún, de forma análoga, se tiene que la solución se resume en hallar la descomposición espectral del operador de covarianza:

$$\int c(s, r) \xi_j(r) dr = \lambda_j \xi_j(s)$$

Componentes principales funcionales

Estimación

- Esto se puede hacer utilizando nuevamente el concepto de *base funcional*. Considerando la descomposición $\xi_j(s) = \phi^T(s)\mathbf{b}_j$, para unos coeficientes desconocidos \mathbf{b}_j , y después de cierta álgebra, el problema se resume al siguiente:

$$\left(\Phi^{1/2}\mathbf{A}\Phi^{1/2}\right)\mathbf{u}_j = \lambda_j\mathbf{u}_j$$

$$\mathbf{u}_j^T\mathbf{u}_j = 1$$

$$\mathbf{u}_j^T\mathbf{u}_k = 0, j \neq k$$

donde $\mathbf{A} = \frac{1}{n} \sum_{i=1}^n \mathbf{a}_i \mathbf{a}_i^T$, $\Phi = \int \phi(r)\phi^T(r)dr$ y $\mathbf{u}_j = \Phi^{1/2}\mathbf{b}_j$.

- Note que la situación se reduce a encontrar las componentes principales de un **problema multivariado clásico**.

Outline

- 1 Introducción
- 2 Marco conceptual de datos funcionales
 - Curvas aleatorias
 - Suavizamiento
 - Componentes principales funcionales
 - **Series de Tiempo Funcionales**
 - Regresión Funcional
- 3 Análisis estadístico del PM10 como un dato funcional
 - El PM10
 - Análisis descriptivo funcional
 - Componentes principales funcionales
 - Análisis de serie de tiempo funcional
 - Regresión funcional con errores FARMA
- 4 Conclusiones

Series de Tiempo Funcionales

Contextualización

- En la literatura se puede observar que la teoría de series de tiempo funcionales aún se encuentra en desarrollo. Se han desarrollado avances, pero aún es un tema abierto.
- De forma breve, se han estudiado modelos *autoregresivos* análogos a los modelos $ARMA(p, q)$ para datos escalares. Un ejemplo es el ARH(1):

$$Y_t(s) = \int \psi(s, r) Y_{t-1}(r) dr + \varepsilon_t(s)$$

- A nivel práctico, se han desarrollado *estrategias* para realizar pronósticos como la denominada metodología de *Hyndman-Ullah* [4].

Series de Tiempo Funcionales

Método de Hyndman-Ullah

- La idea consiste en realizar la descomposición de la serie de tiempo funcional en la *base* de sus componentes principales, para luego pronosticar las series de tiempo de los *puntajes*.
- Como ejemplo, considere figurativamente el proceso $FARMA(P,Q)$:

$$Y_t(s) = \sum_{i=1}^P \left(\int \psi_i(s, r) Y_{t-i}(r) dr \right) + \varepsilon_t(s) + \sum_{i=1}^Q \left(\int \theta_i(s, r) \varepsilon_{t-i}(r) dr \right)$$

Para estimarlo, represente los parámetros funcionales en términos de la **bases dada por los componentes principales**:

$$\psi_i(s, r) = \xi_Y^T(s) \Psi_i \xi_Y(r), \theta_i(s, r) = \xi_Y^T(s) \Theta_i \xi_Y(r) \text{ y} \\ \varepsilon_t(s) = \xi_Y^T(s) \mathbf{e}_t .$$

Series de Tiempo Funcionales

Estimación modelo FARMA

- Reemplazando en el modelo $FARMA(P,Q)$ se tiene que:

$$\xi_Y^T(s)\alpha_t^Y = \xi_Y^T(s) \left[\sum_{i=1}^P \Psi_i \left(\int \xi_Y(r) \xi_Y^T(r) dr \right) \alpha_{t-i}^Y + \mathbf{e}_t + \sum_{i=1}^Q \Theta_i \left(\int \xi_Y(r) \xi_Y^T(r) dr \right) \mathbf{e}_{t-i} \right]$$

$$\alpha_t^Y = \sum_{i=1}^P \Psi_i \alpha_{t-i}^Y + \mathbf{e}_t + \sum_{i=1}^Q \Theta_i \mathbf{e}_{t-i}$$

- La expresión anterior corresponde a un modelo $VARMA(P,Q)$, donde la serie multivariada corresponde a los puntajes α_t^Y .
- Por tanto, estrategias de identificación para modelos series de tiempo multivariadas aplican para determinar implícitamente un modelo adecuado para la serie de tiempo funcional.

Outline

- 1 Introducción
- 2 Marco conceptual de datos funcionales
 - Curvas aleatorias
 - Suavizamiento
 - Componentes principales funcionales
 - Series de Tiempo Funcionales
 - **Regresión Funcional**
- 3 Análisis estadístico del PM10 como un dato funcional
 - El PM10
 - Análisis descriptivo funcional
 - Componentes principales funcionales
 - Análisis de serie de tiempo funcional
 - Regresión funcional con errores FARMA
- 4 Conclusiones

Regresión Funcional

Motivación

- El modelo de regresión funcional de interés para este proyecto se expresa de la forma:

$$Y(s) = \beta_0(s) + \int \beta_1(s, r)X(r)dr + \varepsilon(s)$$

donde $\varepsilon(s)$ representan errores aleatorios funcionales iid de media 0.

- El intercepto $\beta_0(s)$ es una función que hace alusión a la media del proceso Y condicionada a un efecto nulo de $X(s)$.
- El efecto de cada uno de los valores puntuales $X(r)$ en el valor $Y(s)$ se da a través de un ponderador bivariado $\beta_1(s, r)$, que luego es integrado en un único efecto agregado.

Regresión Funcional

Estimación

- Representando los componentes funcionales en términos de la bases dada por los componentes principales de X y Y (i.e $\beta_0(s) = \xi_Y^T(s)\mathbf{b}_0$, $\beta_1(s, r) = \xi_Y^T(s)\mathbf{B}_1\xi_X(r)$, $\varepsilon(s) = \xi_Y^T(s)\mathbf{e}$), y reemplazando en el modelo de regresión funcional, se tiene que:

$$\xi_Y^T(s)\alpha^Y = \xi_Y^T(s)\mathbf{b}_0 + \int \xi_Y^T(s)\mathbf{B}_1\xi_X(r)\xi_X^T(r)\alpha^X dr + \xi_Y^T(s)\mathbf{e}$$

$$\xi_Y^T(s)\alpha^Y = \xi_Y^T(s) \left(\mathbf{b}_0 + \mathbf{B}_1 \left(\int \xi_X(r)\xi_X^T(r)dr \right) \alpha^X + \mathbf{e} \right)$$

$$\alpha^Y = \mathbf{b}_0 + \mathbf{B}_1\alpha^X + \mathbf{e}$$

Regresión Funcional

Inferencia

- En ese orden de ideas, se tiene una regresión multivariada que se da directamente sobre los *puntajes* de las componentes α^Y vs α^X , y que se puede estimar utilizando mínimos cuadrados ordinarios.
- La *inferencia estadística* para la regresión funcional se basa en la inferencia estadística del modelo de regresión multivariado equivalente, en donde el proceso de inferencia ya es rutinario.
- Note que la situación se reduce, nuevamente, a un **problema multivariado clásico**.

Outline

- 1 Introducción
- 2 Marco conceptual de datos funcionales
 - Curvas aleatorias
 - Suavizamiento
 - Componentes principales funcionales
 - Series de Tiempo Funcionales
 - Regresión Funcional
- 3 Análisis estadístico del PM10 como un dato funcional**
 - El PM10**
 - Análisis descriptivo funcional
 - Componentes principales funcionales
 - Análisis de serie de tiempo funcional
 - Regresión funcional con errores FARMA
- 4 Conclusiones

El PM10

Contexto

- El PM10 es el *material particulado* que se encuentra suspendido en la atmósfera con un diámetro aerodinámico igual o inferior a los 10 micrómetros. El ser humano puede inhalar fácilmente las partículas y así generar problemas respiratorios.
- En Bogotá, la Secretaría Distrital de Ambiente regula la calidad del aire por medio de la *Red de Monitoreo de Calidad de Aire (RMCA)* conformada por 12 estaciones repartidas a lo largo de la ciudad.

El PM10

Objetivos

- El propósito de esta sección es ilustrar la aplicación de las técnicas de datos funcionales ya descritas para el análisis del PM10 en Bogotá.
- Se tiene información horaria para los años 2015 y 2016 del PM10. Después de un preprocesamiento de los datos, se trabajará con la estación de *Guaymaral*.
- El PM10 existe para cada instante del tiempo, luego es un proceso funcional cuyo dominio es el tiempo. **En este estudio, una curva mostrará el nivel de PM10 durante el día.** Es decir, el conjunto de datos funcionales estará dado por $\{PM10_t(s), s \in S = [0, 24), t \in T = \{1, 2, \dots, 730\}\}$.

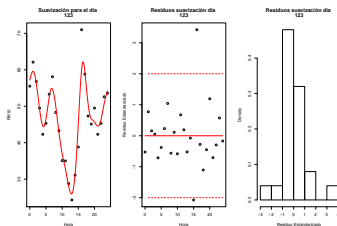
Outline

- 1 Introducción
- 2 Marco conceptual de datos funcionales
 - Curvas aleatorias
 - Suavizamiento
 - Componentes principales funcionales
 - Series de Tiempo Funcionales
 - Regresión Funcional
- 3 **Análisis estadístico del PM10 como un dato funcional**
 - El PM10
 - **Análisis descriptivo funcional**
 - Componentes principales funcionales
 - Análisis de serie de tiempo funcional
 - Regresión funcional con errores FARMA
- 4 Conclusiones

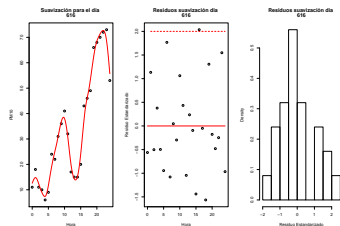
Análisis descriptivo funcional

Suavizamiento

- El suavizamiento se hace de forma *simultanea* para todas las curvas utilizando la misma base, así como el mismo parámetro de flexibilidad.
- Los resultados sugieren que la metodología de suavizamiento más apropiada corresponde a *B-Splines regularizados*.



(a) 3 de Mayo de 2015



(b) 7 de Septiembre de 2016

Figura: Ejemplos del suavizamiento del PM10

Análisis descriptivo funcional

El comportamiento medio del PM10 se puede entender según la intensidad de tráfico en la ciudad. En horas pico se tienen mayores valores de PM10, mientras que en horas valle se tienen bajos.

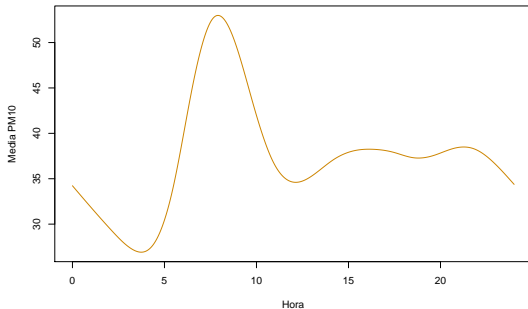
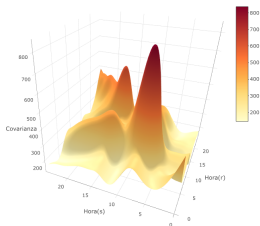


Figura: Función de media del PM10

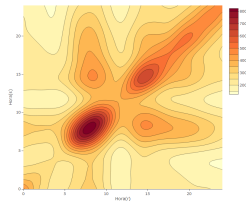
Análisis descriptivo funcional

Superficie de covarianza

La varianza del PM10 no es constante a lo largo del día, y su intensidad también es proporcional al comportamiento del tráfico. Existe una covarianza considerable al rededor de las 8:00 y las 15:00. Esto sugiere que altos niveles de PM10 en la mañana suelen estar acompañados a altos niveles de PM10 en la tarde.



(a) Gráfico de superficie



(b) Gráfico de contorno

Figura: Función de covarianza del PM10

Análisis descriptivo funcional

Estacionalidad Semanal

El PM10 tiene un comportamiento similar entre semana, mientras que para el fin de semana se presentan niveles más bajos, en especial el domingo cuando hay poco tráfico.

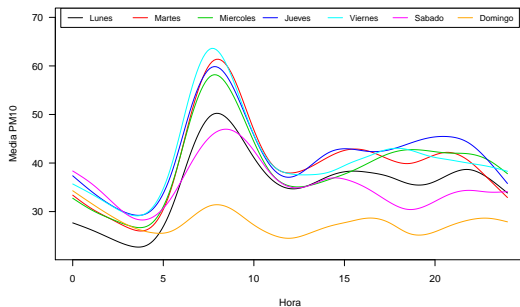


Figura: Función de media del PM10 por día

Outline

- 1 Introducción
- 2 Marco conceptual de datos funcionales
 - Curvas aleatorias
 - Suavizamiento
 - Componentes principales funcionales
 - Series de Tiempo Funcionales
 - Regresión Funcional
- 3** **Análisis estadístico del PM10 como un dato funcional**
 - El PM10
 - Análisis descriptivo funcional
 - Componentes principales funcionales**
 - Análisis de serie de tiempo funcional
 - Regresión funcional con errores FARMA
- 4 Conclusiones

Componentes principales funcionales

Selección número de componentes

Para realizar la selección del número de componentes a utilizar, se tiene en cuenta tres criterios de importancia: Variación explicada, reconstrucción de los datos originales y su interpretación.

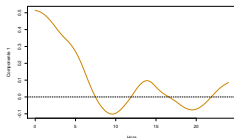
Tabla: Varianza explicada por los componentes principales funcionales rotados del PM10

Componente	Varianza	Prop. Var.	Prop. Acumulada
1	1356,453	0,115	0,115
2	2828,270	0,239	0,354
3	2045,592	0,173	0,527
4	2726,274	0,231	0,757
5	1751,873	0,148	0,905

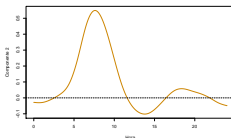
Componentes principales funcionales

Interpretación

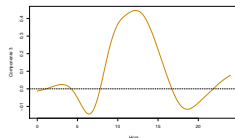
Cada componente explica la variabilidad en una franjas horaria.



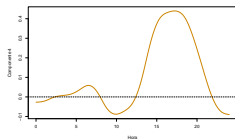
(a) C1: Madrugada



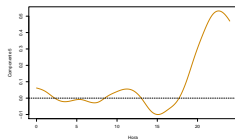
(b) C2: Mañana



(c) C3: Mediodía



(d) C4: Tarde



(e) C5: Noche

Figura: Componentes principales funcionales rotados del PM10

Outline

- 1 Introducción
- 2 Marco conceptual de datos funcionales
 - Curvas aleatorias
 - Suavizamiento
 - Componentes principales funcionales
 - Series de Tiempo Funcionales
 - Regresión Funcional
- 3** **Análisis estadístico del PM10 como un dato funcional**
 - El PM10
 - Análisis descriptivo funcional
 - Componentes principales funcionales
 - Análisis de serie de tiempo funcional**
 - Regresión funcional con errores FARMA
- 4 Conclusiones

Análisis de serie de tiempo funcional

Correlograma de los puntajes

Existe alta dependencia del PM10 con sus valores en días pasados, así como un claro patrón estacional semanal del PM10.

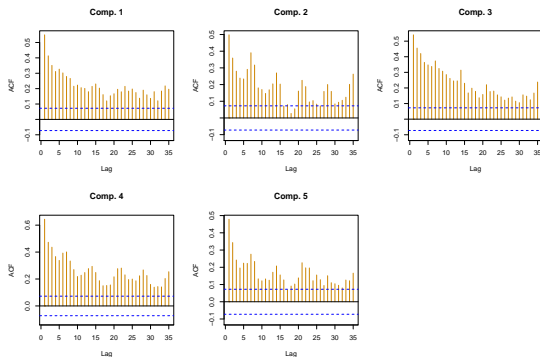


Figura: Correlogramas de los puntajes de los componentes principales del PM10

Análisis de serie de tiempo funcional

Estimación modelo FARMA

La metodología de *Box-Jenkins* sugiere la necesidad de tomar la primera diferencia estacional $\nabla_7 \alpha_t$, y utilizar un modelo *VARMA* $(1, 0) \times (0, 1)_7$ para dicha diferencia.

$$\nabla_7 \alpha_t = \begin{bmatrix} 0,4 & -0,0 & 0,1 & 0,0 & 0,4 \\ 0,1 & 0,4 & 0,0 & 0,1 & 0,3 \\ 0,1 & 0,0 & 0,5 & 0,2 & 0,1 \\ 0,0 & 0,0 & 0,1 & 0,6 & 0,2 \\ 0,0 & 0,1 & 0,1 & 0,2 & 0,4 \end{bmatrix} \nabla_7 \alpha_{t-1} + \mathbf{e}_t + \begin{bmatrix} -0,9 & -0,0 & -0,0 & -0,0 & -0,0 \\ -0,0 & -0,9 & -0,0 & -0,0 & 0,1 \\ -0,0 & 0,0 & -0,9 & -0,0 & 0,0 \\ -0,1 & 0,1 & -0,0 & -0,9 & 0,0 \\ -0,1 & 0,1 & 0,0 & -0,0 & -0,9 \end{bmatrix} \mathbf{e}_{t-7}$$

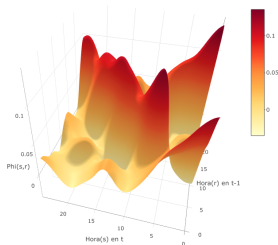
Por tanto el modelo FARMA resultante toma al forma:

$$PM10_t(s) = PM10_{t-7}(s) + \int \psi(s, r) \nabla_7 PM10_{t-1}(r) dr + \varepsilon_t(s) + \int \theta(s, r) \varepsilon_{t-7}(r) dr$$

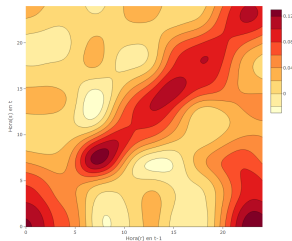
Análisis de serie de tiempo funcional

Superficie $\psi(s, r)$ modelo FARMA

$\psi(s, r)$ toma valores significativamente diferentes de 0 en la diagonal, y en la zona asociada a la noche del día $t - 1$ y la madrugada del día t . Esto es claro ya que el nivel de PM10 en la madrugada dependerá de su valor en las horas de la noche anterior.



(a) Gráfico de superficie



(b) Gráfico de contorno

Figura: Superficie $\psi(s, r)$ estimada

Análisis de serie de tiempo funcional

Pronósticos

El pronóstico es similar a la media por días, haciendo notar que efectivamente el modelo incluye la estacionalidad semanal en la serie. Aún así, las curvas difieren por los valores recientes.

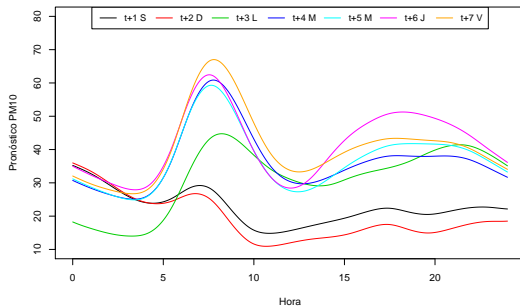


Figura: Pronóstico del PM10 para la siguiente semana

Outline

- 1 Introducción
- 2 Marco conceptual de datos funcionales
 - Curvas aleatorias
 - Suavizamiento
 - Componentes principales funcionales
 - Series de Tiempo Funcionales
 - Regresión Funcional
- 3 Análisis estadístico del PM10 como un dato funcional**
 - El PM10
 - Análisis descriptivo funcional
 - Componentes principales funcionales
 - Análisis de serie de tiempo funcional
 - Regresión funcional con errores FARMA**
- 4 Conclusiones

Regresión funcional con errores FARMA

Especificación

- Se pretende entender el papel de otras variables climáticas en el comportamiento del PM10. En este caso, se estudian **la Temperatura y la Humedad**.
- En principio se propone estimar el modelo:

$$\widetilde{PM}_{10}^t(s) = \int \beta_1(s, r) \widetilde{TEMP}_{t-1}(r) dr + \int \beta_2(s, r) \widetilde{HUM}_{t-1}(r) dr + \varepsilon_t(s)$$

donde $\tilde{X}(s) = X(s) - \mu_X(s)$

- Observe que se utilizan las variables explicativas rezagadas en 1 periodo. Esto es para mantener la coherencia en la forma estructural del modelo.

Regresión funcional con errores FARMA

Algunas consideraciones

- El PM10 tiene una estacionalidad semanal, por tanto se procede a utilizar un modelo que la incluya. Para esto, el modelo de regresión multivariada respectivo se estima con errores VARMA.
- La metodología de *Box-Jenkins* sugirió utilizar un modelo de regresión con errores VARMA $(1, 0, 0) \times (0, 1, 1)_7$. Se asume que la estructura de las matrices de coeficientes del VARMA para los residuales es *diagonal* dada la no convergencia del modelo.
- La prueba de significancia sugiere que las variables temperatura y humedad son estadísticamente relevantes para explicar el comportamiento del PM10.

Regresión funcional con errores FARMA

Estimación

El modelo de regresión multivariado estimado es:

$$\alpha_t^{PM10} = \begin{bmatrix} -0,4 & -0,5 & 1,1 & -0,2 & 0,2 \\ 3,7 & 1,3 & 1,5 & 2,3 & -2,8 \\ 0,8 & 0,1 & 0,8 & 1,9 & 0,2 \\ -2,2 & 0,5 & 0,1 & -1,9 & 1,1 \\ 1,3 & -0,4 & -1,3 & 1,6 & 0,7 \end{bmatrix} \alpha_{t-1}^{TEMP} + \begin{bmatrix} 0,0 & -0,3 & 0,2 & 0,3 & 0,1 \\ 0,5 & -0,1 & 0,5 & -0,1 & 0,9 \\ 0,4 & -0,5 & 0,2 & 0,4 & 0,3 \\ -0,5 & -0,1 & -0,0 & 0,3 & -0,2 \\ 0,6 & -0,2 & 0,2 & -0,2 & 0,0 \end{bmatrix} \alpha_{t-1}^{HUM} + \mathbf{e}_t$$

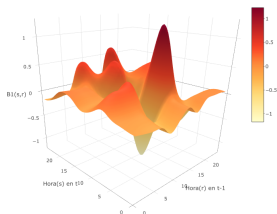
donde $\nabla_7 \mathbf{e}_t = \Phi \nabla_7 \mathbf{e}_{t-1} + \mathbf{v}_t + \Theta \mathbf{v}_{t-7}$, siendo \mathbf{v} un ruido blanco multivariado. Finalmente, el respectivo modelo de regresión funcional toma la forma:

$$\begin{cases} \widetilde{PM}_{10}^t(s) = \int \beta_1(s, r) \widetilde{TEMP}_{t-1}(r) dr + \int \beta_2(s, r) \widetilde{HUM}_{t-1}(r) dr + \varepsilon_t(s) \\ \varepsilon_t(s) = \varepsilon_{t-7}(s) + \int \phi(s, r) \nabla_7 \varepsilon_{t-1}(r) dr + \nu_t(s) + \int \theta(s, r) \nu_{t-7}(r) dr \end{cases}$$

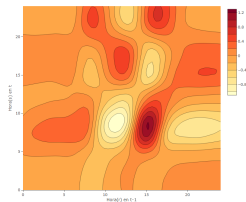
Regresión funcional con errores FARMA

Interpretación $\beta_1(s, r)$

Gran parte de la superficie es casi 0, y los valores no nulos se dan en las 8:00 ~ 9:00 y 15:00 ~ 16:00. El efecto de la temperatura del día anterior realiza una ponderación entre una hora en la mañana y una en la tarde. Los valores de la superficie tienden a estar concentrados hacia magnitud positivas, luego incrementos de la temperatura aumentan el valor del PM10.



(a) Gráfico de superficie



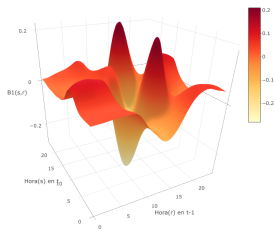
(b) Gráfico de contorno

Figura: Superficie $\beta_1(s, r)$, variable Temperatura

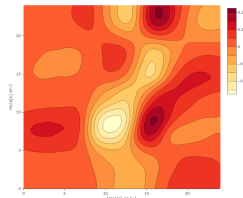
Regresión funcional con errores FARMA

Interpretación $\beta_2(s, r)$

El efecto de la humedad sobre el PM10 es similar al dado por la temperatura. En la base inferior de la superficie, valores de humedad por encima del promedio reducen el PM10 en la mañana, mientras que para las demás horas no sucede. Luego, el efecto reductor de la humedad sobre el PM10 se da exclusivamente en la mañana.



(a) Gráfico de superficie



(b) Gráfico de contorno

Figura: Superficie $\beta_2(s, r)$, variable Humedad

Regresión funcional con errores FARMA

Calidad del ajuste

El modelo parece estar bien especificado, y cumple los supuestos de regresión. Si bien la temperatura y la humedad están asociadas con el PM10, no son suficientes para explicar el total de su comportamiento.

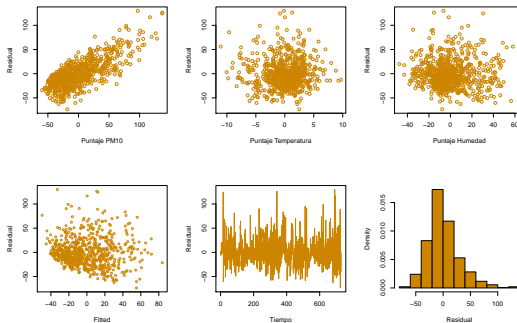


Figura: Gráficos de residuales del modelo de regresión multivariado





Conclusiones

- La metodología de datos funcionales permite desarrollar análisis a profundidad, permitiendo evidenciar relaciones casi imposibles de detectar con el uso de metodologías “clásicas”.
- Los modelos basados en datos funcionales no asumen grandes supuestos y el costo computacional adicional es casi imperceptible.
- La aplicación de datos funcionales para el PM10 permite explicar su comportamiento durante todo el día, incluyendo su codependencia entre horas. Es posible identificar los patrones estacionales del PM10, así como su importancia para ser incluidos en otros estudios.
- Los componentes principales funcionales definen *índices* horarios para el PM10, de forma óptima. Los puntajes de los componentes miden como difiere la contaminación al comportamiento usual, y pueden utilizarse en la definición de políticas públicas para la contaminación del aire.

Conclusiones

- Usando datos funcionales, es posible lograr pronósticos del PM10 más ambiciosos que a los que se llegarían con metodologías tradicionales. Los modelos funcionales permiten incorporar la correlación dentro y entre curvas sin dificultad, al tiempo que se mantiene un modelo parsimonioso.
- Los modelos de regresión funcional permiten evidenciar la influencia de las variables explicativas sobre el PM10 para cada una de las horas del día en todas las combinaciones posibles, hecho que permite obtener conclusiones más refinadas.
- Trabajos futuros deben evaluar como los resultados encontrados para el PM10 pueden aplicarse de forma efectiva en la realidad. Del mismo modo, es necesario considerar otros enfoques que pueden ayudar en el análisis explicativo del PM10.

Referencias

-  Ferraty F. y Vieu P. *Nonparametric Functional Data Analysis*. Springer Series in Statistics, 2006.
-  Ramsay J. y Silverman B. W. *Functional Data Analysis*. Springer, 2005.
-  Horváth L. y Kokoszka P. *Introduction to Functional Data Analysis*. Springer-Verlag New York, 2012.
-  Kokoszka P. y Reimherr M. *Inference for Functional Data with Applications*. Chapman Hall/CRC Texts in Statistical Science, 2017.
-  Hsing T. y Eubank R. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley Series in Probability y Statistics, 2015.