



ESTIMACIÓN FUNCIONAL ROBUSTA DE LA MEDIA DE LA HUMEDAD RELATIVA

Programa Académico de Estadística

**Daniel Arturo Sánchez Ortíz
Michael Steven Rodriguez Caceres**

Universidad del Valle
Facultad de Ingeniería, Escuela de Estadística
Santiago de Cali, Colombia
2018

**ESTIMACIÓN FUNCIONAL ROBUSTA DE LA
MEDIA DE LA HUMEDAD RELATIVA**
Programa Académico de Estadística

Daniel Arturo Sánchez Ortíz
Michael Steven Rodriguez Caceres

Trabajo de grado presentado como requisito parcial para optar al título de:
Estadístico

Director:
Javier Olaya Ochoa

Universidad del Valle
Facultad de Ingeniería, Escuela de Estadística
Santiago de Cali, Colombia
2018

Dedicatoria

A Dios por ser mi compañía, darme la fuerza y sabiduría cada día para alcanzar mis metas, en especial este logro que hasta hoy es el más importante. A mis padres por siempre apoyarme y por el gran esfuerzo que hicieron para que yo fuera la persona que soy hoy en día; en verdad agradezco que hayan puesto toda su confianza en mí. Finalmente le doy las gracias a todas esas personas que no menciono pero que llegaron y me aportaron tantas cosas en este largo camino, en especial a mi amiga Evelyn Bedoya D., por extenderme la mano cada vez que lo necesite, brindarme su hermosa amistad y confianza y mi amigo Emerson Caravali, por su disposición y ayuda en todo momento.

Michael S. Rodriguez .

A Dios por regalarme salud, vida y entendimiento durante todo este trayecto. A mi madre y padre por su infinito amor y apoyo desinteresado, por ser el motor que me motiva cada día. A mis hermanas por estar siempre presentes en los buenos y malos momentos. A mi gran familia porque sé que se enorgullecen de este gran logro como si fuera propio, además de estar pendientes apoyándome en todo este proceso. Finalmente a todos los amigos que me brindaron su alegría para hacer este camino mucho más ameno.

Daniel A. Sanchez

Agradecimientos

A nuestro profesor y director de trabajo de grado Javier Olaya Ochoa, que confió en nosotros y nos dio la oportunidad de realizar el presente trabajo, además, nos acompañó con su conocimiento y asesoría en todas las etapas del mismo. Y al Departamento Administrativo de Gestión del Medio Ambiente (DAGMA), por brindarnos y permitirnos utilizar sus datos como principal insumo de nuestra investigación.

Agradecemos a la Universidad del Valle y a la escuela de estadística, por brindarnos las herramientas necesarias y el conocimiento, que nos formaron no solo como profesionales estadísticos, si no como personas.

Resumen

Las observaciones atípicas pueden generar estimaciones erróneas o que no concuerdan con la realidad, por lo que es de gran importancia la implementación de técnicas que identifiquen y traten de una forma adecuada dichos valores. En el siguiente trabajo se logra establecer una metodología para reducir el impacto generado por las observaciones atípicas de la Humedad Relativa en el promedio del año 2016 en la ciudad de Santiago de Cali, Colombia. Para esto se aprovechan las ventajas generadas al utilizar el análisis de datos funcionales, de donde se desprenden conceptos como el de profundidad funcional, utilizado para identificar los días del año que presentaron un patrón diferente al convencional y posteriormente mediante teoría de datos escalares buscar una estimación de la media que no se vea afectada por dichos valores. Se estableció que realizando un recorte del 16 % de las curvas menos profundas se logra una disminución en la variabilidad y una estimación robusta de la media funcional.

Palabras claves: Humedad relativa, estimación robusta, datos funcionales, profundidad funcional, media recortada

Abstract

Atypical observations can generate erroneous estimates or that do not agree with reality, so it is of great importance to implement techniques that identify and treat these values in an appropriate way. In the following work, it is possible to establish a technique to reduce the impact generated by the atypical observations of Relative Humidity in the 2016 average in the city of Santiago de Cali, Colombia. For this, the advantages generated by using the functional data analysis are exploited, from which concepts such as functional depth are derived, used to identify the days of the year that presented a different pattern to the conventional one and later, using scalar data theory, to search for an estimate of the average that is not affected by these values. It was established that by making a cut of 16 % of the less depth curves, a decrease in variability and a robust estimate of the functional mean were achieved.

Keywords: relative humidity, robust estimate, functional data, depth functional, trimmed mean

Contenido

Resumen	vii
1. Introducción	1
1.1. Planteamiento del problema	1
1.2. Justificación	2
1.3. Objetivos	3
1.3.1. General	3
1.3.2. Específicos	3
1.4. Antecedentes	3
2. Marco teórico y definiciones	6
2.1. Marco teórico conceptual	6
2.1.1. Zona de estudio	6
2.1.2. Clima	6
2.1.3. Humedad relativa	7
2.1.4. Temperatura	7
2.2. Marco teórico estadístico	8
2.2.1. Análisis de Datos Funcionales (ADF)	8
2.2.2. Construcción de Datos Funcionales	8
2.2.3. Estimación robusta	11
2.2.4. Datos atípicos	12
2.2.5. Profundidad funcional	13
2.2.6. Box Plot funcional	14
2.2.7. Valores Atípicos Funcionales	15
3. Metodología	16
3.1. Validación e inspección de los datos	16
3.2. Construcción de las curvas diarias de la humedad relativa	17
3.3. Detección de las curvas diarias de humedad relativa atípicas	17
3.4. Estimación robusta funcional	18
4. Resultados	19
4.1. Depuración de los datos y análisis exploratorio	19
4.2. Construcción de datos funcionales	27

4.3. Valores Atípicos Funcionales	31
4.4. Media funcional robusta	36
5. Conclusiones y recomendaciones	39
5.1. Conclusiones	39
5.2. Recomendaciones	41
A. Anexo: BoxPlot Funcional del porcentaje de humedad relativa año 2016	42
B. Anexo: BoxPlot Funcional del porcentaje de humedad relativa para cada uno de los meses del año 2016	43
Bibliografía	46

Lista de Figuras

4-1. Comportamiento del Porcentaje de Humedad Relativa y su Promedio (Linea Negra) año 2016	22
4-2. Variación de la Humedad relativa por Hora	22
4-3. Diagramas de caja de la Humedad Relativa por mes	23
4-4. Comportamiento del porcentaje de Humedad Relativa y el promedio para cada mes del año 2016	26
4-5. Diagrama de puntos de Humedad Relativa vs Temperatura	26
4-6. Curvas diarias de porcentaje de humedad relativa para cada mes del año 2016	30
4-7. Curvas diarias y curva media funcional (En rojo) y curva de desviación estándar (En naranja) y curvas atípicas (En azul) de porcentaje de humedad relativa para cada mes del año 2016	33
4-8. Curvas diarias de porcentaje de Humedad Realtiva del año 2016	34
4-9. Curvas diarias, curva media funcional (rojo),curvas atípicas (azul), curvas de desviación estándar (naranja) del porcentaje de humedad relativa para el año 2016	35
4-10.Estimación de la curva media y desviación estándar de porcentaje de Humedad Relativa (En rojo) y las curvas medias y desviación estándar robustas funcionales.	36
4-11.Desviación estándar a diferentes niveles de recorte	37
4-12.Curva media funcional (Roja), Curva media funcional robusta (Negra) del porcentaje de humedad relativa y su intervalo de confianza	38
A-1. Box-Plot funcional del porcentaje de humedad relativa año 2016	42
B-1. Box-Plots funcionales por meses del año 2016	45

Lista de Tablas

4-1. Balance general del porcentaje de datos recolectados en la estación Compartir durante el 2016.	20
4-2. Días validos por mes año 2016	20
4-3. Resumen descriptivo Humedad relativa por hora	21
4-4. Resumen descriptivo Humedad relativa por mes	23
4-5. Comparación Fourier con B-Spline	27

1. Introducción

1.1. Planteamiento del problema

Mundialmente distintas organizaciones están encargadas de monitorear y registrar a diario datos referentes a variables del clima y la calidad del aire, información utilizada para sintetizar aspectos de fenómenos que resultan importantes para uno o más propósitos analíticos y de monitoreo en el tiempo. En este sentido, bajo la coordinación y administración del Grupo de Calidad del Aire del Departamento Administrativo de Gestión del Medio Ambiente (DAGMA) opera el Sistema de Vigilancia de la Calidad del Aire (SVCASC) encargado de realizar estas mediciones en diferentes puntos de la ciudad Santiago de Cali-Colombia.

No obstante, entre las observaciones registradas se pueden presentar algunas que no siguen el patrón de las demás y no son errores de medición, por ejemplo, en ocasiones se registran días en los que la temperatura obedece a fenómenos naturales y desciende o asciende abruptamente, generando inconvenientes en las predicciones o cualquier tipo de análisis estadístico que se realice, ya que estas observaciones, que podrían considerarse atípicas, frecuentemente traen consecuencias sobre las estimaciones (principalmente de la media), puesto que pueden “*halar*” el estimador hacia ellas.

Para este estudio se utilizará una base de datos proporcionada por el SVCASC la cual tiene registrados grandes cantidades de datos, ya que sus mediciones se hacen cada hora durante todos los días del año. De esta forma, y sabiendo que estas variables climáticas arrojan resultados continuamente (cada instante de tiempo) se pueden transformar en datos funcionales usando técnicas de suavizado (“El análisis de datos funcionales es aquella parte de la estadística que trabaja con muestras de funciones aleatorias” (Ramsay y Silverman, 2005)), aprovechando las ventajas que este tipo de datos proporciona.

Así que, el objetivo principal de este proyecto será proponer una estimación usando métodos robustos de la media funcional de la Humedad relativa registrada en un determinado periodo y sitio de la ciudad, permitiendo tomar buenas decisiones basadas en los resultados y/o conclusiones obtenidos, ya que los métodos robustos se han caracterizado por mejorar las estimaciones teniendo en cuenta la presencia de datos atípicos.

1.2. Justificación

Es de gran importancia realizar distintos análisis que permitan conocer u obtener un indicador que represente de manera precisa el comportamiento de los datos estudiados. En el caso del estudio de variables climáticas se debe tener un mayor cuidado ya que al presentar un reporte que no esté acorde con la realidad, se puede incidir en una falsa alarma de algún fenómeno o peor aún en la no alerta temprana del mismo.

En Santiago de Cali-Colombia, el Sistema de Vigilancia de la Calidad del Aire (SVCASC) toma registros con el fin de conocer el estado de la calidad del aire y el comportamiento de la meteorología en las zonas de influencia de las estaciones de monitoreo. En la actualidad cuentan con nueve estaciones automáticas, que están ubicadas alrededor de la ciudad las cuales toman registro a cada minuto del día, de distintas variables de la calidad del aire y climatológicas.

Esta información la emplea el DAGMA en sus distintas actividades como entidad encargada del control ambiental en el municipio, ya sea ejecutando las políticas y funciones precisas, sin embargo esta información no es solo materia de estudio del DAGMA, sino que también es de gran utilidad en distintos sectores como: económico, energético, industrial, entre otros, en donde el entendimiento del comportamiento de las distintas variables climatológicas es de vital importancia para el desarrollo y desempeño de sus actividades. De aquí, la importancia de considerar las observaciones no convencionales a la hora de realizar el análisis, obtener resultados confiables y no cometer errores a la hora de tomar cualquier decisión.

Los patrones históricos del comportamiento del clima han presentado cambios, el principal indicador de esto es el incremento del calentamiento global que al parecer cada vez es más alto (Berruezo y Jiménez, 2017), situación donde se refleja que el comportamiento del clima en la ciudad pueda presentar distintas anomalías en un día. Con esta información se realizan diferentes estudios y análisis, en donde la validez de estos depende en gran medida de la calidad de los datos y la manera en que se procesó la información. La validez de estos estudios podría estar en juego si se pasa por alto o no se considera el problema de los datos atípicos (Outliers), ya que se omitiría información importante y los resultados y conclusiones de la investigación no serían confiables.

Por lo anterior, es pertinente evaluar una estimación robusta de la media funcional para la solución al problema del registro de posibles observaciones atípicas en variables climatológicas, para esto se debe tener una medida con la que se pueden identificar y seleccionar las observaciones atípicas funcionales para así tratarlas de diferente manera a las convencionales.

1.3. Objetivos

1.3.1. General

Estimar la media funcional de la humedad relativa promedio horaria utilizando un método robusto, con los datos suministrados por el Sistema de Vigilancia de la Calidad del Aire de Santiago de Cali (SVCASC), para el año 2016 en la estación compartir.

1.3.2. Específicos

- Construir Datos funcionales (curvas) diarias de la humedad relativa promedio horaria de la estación de monitoreo Compartir de la ciudad Santiago de Cali.
- Establecer mediante técnicas de detección de Outliers cuales de estos datos funcionales diarios son observaciones atípicas.
- Estimar la media de la humedad relativa promedio horaria con un método robusto de análisis de datos funcionales.
- Comparar la estimación robusta con la estimación clásica de la media funcional de la humedad relativa promedio horaria, para la estación de monitoreo Compartir ubicada en la Ciudad de Santiago de Cali.

1.4. Antecedentes

El empleo de métodos robustos para datos funcionales ha sido tratado por distintos autores, los cuales combinaron estas técnicas para diferentes aplicaciones. A continuación se citan algunos trabajos describiendo brevemente sus metodologías. Se presenta además investigaciones realizadas en torno a las variables climatológicas.

Hyndman y Ullah (2007) proponen un nuevo método para pronosticar la mortalidad por edades y las tasas de fecundidad observadas con el tiempo. Este enfoque permite funciones suaves de edad, es robusto para los años periféricos debido a guerras y epidemias, y proporciona un marco de modelado que se adapta fácilmente para permitir restricciones y otra información. Las ideas de análisis de datos funcionales, suavizado no paramétrico y estadísticas robustas se combinan para formar una metodología que es ampliamente aplicable a cualquier serie de datos funcionales observados discretamente y posiblemente con error. La metodología se aplica a los datos de mortalidad franceses y a los datos de fertilidad australianos, y los pronósticos obtenidos son superiores a los del método Lee-Carter (LC) comúnmente utilizado en la predicción de la mortalidad y la fertilidad.

Crambes *et al.* (2008) se centran en las propiedades asintóticas de una estimación no paramétrica condicional de una variable valorada real con una covariable funcional. Presentan resultados relacionados con la convergencia en probabilidad, normalidad asintótica. Utilizando estimación robusta la cual proporciona un enfoque alternativo a los métodos clásicos que no se afecta indebidamente por la presencia de valores atípicos.

Torres *et al.* (2011) utilizaron los datos funcionales para construir una técnica para detectar valores atípicos, comparándolos con los resultados obtenidos por métodos clásicos. Los autores realizan la investigación haciendo uso de los datos de emisión de gases en una zona urbana de la ciudad de Oviedo, encontrando que su propuesta identificó la existencia de valores extremos en algunos casos donde las técnicas estándar no lo hacían.

Sawant *et al.* (2012) proponen un análisis de componentes principales funcionales robusto para encontrar las combinaciones lineales de las variables originales que contienen la mayor parte de información, incluso si existen valores atípicos funcionales, aplicando este estudio en un extenso trabajo de simulación y dos conjuntos de datos de la quimiometría y el medio ambiente.

Olaya (2012) analiza el comportamiento del Ozono y como se ve afectado por variables meteorologicas en la ciudad de Cali. En este caso la propuesta para modelar este contaminante va orientada a la utilización de un modelo aditivo generalizado (GAM) semiparametrico, que usa funciones suavizadas por medio de splines de las variables regresoras. Se obtiene que todos los coeficientes de regresión estimados son altamente significativos, aunque una validación de supuestos indica que el modelo no parece ser el más adecuado, se deja con el fin de ilustrar el beneficio que trae la utilización de modelos GAM.

Boente y Vahnovan (2016) adaptan Modelos lineales parciales para tratar con covariables funcionales capturando tanto las ventajas de un modelo semi-lineal como las de modelado no paramétrico para datos funcionales. Para resolver el problema de las observaciones atípicas cuando las covariables del componente no paramétrico son funcionales, se introducen estimaciones robustas para el parámetro de regresión y el operador de regresión. Se estudian los resultados de consistencia de los estimadores robustos y la distribución asintótica del estimador de parámetros de regresión. Los experimentos numéricos muestran que los estimadores resultantes tienen buenas propiedades de robustez. Los beneficios de considerar estimadores robustos también se ilustran en un conjunto de datos reales donde el ajuste robusto revela la presencia de valores atípicos influyentes.

2. Marco teórico y definiciones

2.1. Marco teórico conceptual

2.1.1. Zona de estudio

Santiago de Cali, capital del Valle del Cauca, está situada en la región sur del valle entre la cordillera occidental y la cordillera central de los andes. La posición geográfica de la ciudad: Latitud, al norte Cerro de Tatama 5° 00' 30", al Sur La Balsa 3° 05' 35". Longitud, al Este Páramo de Barragán 75° 41' 32", al oeste Bocas del Naya 77° 00' 33". Altitud: Farallones de Cali 4.080 m.s.n.m y una altitud media de 1018 m.s.n.m en el área metropolitana. Cali limita al norte con los municipios de La Cumbre y Yumbo, al oriente con los municipios de Palmira, Candelaria y Puerto Tejada, al sur con el municipio de Jamundí y al occidente con los municipios de Buenaventura y Dagua (Alcaldía, 2004).

Santiago de Cali cuenta con un clima de sabana tropical, en promedio la precipitación anual va desde los 900 mm en las zonas más secas hasta los 1.800 mm en las zonas más lluviosas, con 1.000 mm promedio sobre la mayor parte del área Metropolitana de Cali. La temperatura media es de 23.1 °C (73.6 °F) con un mínimo promedio de 15 °C (66 °F) y un máximo promedio de 32 °C (86 °F), con un máximo absoluto de 36 °C y mínimo absoluto de 13 °C. Las estaciones secas van de diciembre a febrero y de julio a agosto y la estación de lluvias de marzo a mayo y de septiembre a noviembre.

2.1.2. Clima

La interacción entre distintos factores atmosféricos, biofísicos y geográficos dan como resultado el clima en las diferentes partes del mundo, estos agentes como la temperatura, la presión atmosférica, el viento, la humedad y la lluvia pueden cambiar en el tiempo y el espacio debido a que son fenómenos naturales. A partir de estas combinaciones se determina el clima de un lugar, por ejemplo, el clima de las montañas por lo general es frío, en cambio a menor altitud, como es el caso de las costas, el clima es cálido y con temporadas de sequías (Martínez *et al.*, 2004).

Ademas, “el clima se suele definir en sentido restringido como el estado promedio del tiempo”,

es decir, el clima una descripción estadística del tiempo atmosférico en términos de los valores medios y de la variabilidad de los factores correspondientes durante períodos que pueden abarcar desde meses hasta miles o millones de años (IPCC, 2013). El período de promedio habitual es de 30 años, según la definición de la Organización Meteorológica Mundial.

2.1.3. Humedad relativa

La humedad relativa es la cantidad de humedad en el aire, comparado con la que el aire puede mantener a esa temperatura. Se define como la razón entre la presión parcial del vapor de agua y la presión de vapor saturado a una temperatura dada (Nave, 2005). Generalmente se expresa como porcentaje.

$$\text{Humedad Relativa} = \frac{\text{Densidad de vapor de } H_2O \text{ actual}}{\text{Densidad de vapor de } H_2O \text{ saturado}} \times 100 \quad (2-1)$$

En consecuencia, cuando la humedad relativa es cercana al 100 % el aire retiene casi todo el vapor de agua que puede.

Los humanos son sensibles a la humedad. Por lo general, una humedad relativa del 40-50 % es óptima tanto para la salud como para la comodidad. La humedad alta, particularmente en un día caluroso, reduce la evaporación de la humedad de la piel, que es uno de los mecanismos vitales del cuerpo para regular la temperatura corporal. Por otra parte, la humedad muy baja reseca la piel y las membranas mucosas (Douglas, 2006).

2.1.4. Temperatura

Tambutti y Muñoz (2002) lo definen como la magnitud física que mide cuán caliente o frío está un objeto. Generalmente es medido mediante un termómetro. Una temperatura elevada es asociada con el concepto de calor mientras que una temperatura baja, es asociada con frío.

En términos técnicos, la temperatura es una magnitud escalar relacionada con la energía interna de un sistema termodinámico, definida por el principio cero de la termodinámica. Más específicamente, está relacionada directamente con la parte de la energía interna conocida como energía cinética, que es la energía asociada a los movimientos de las partículas del sistema, sea en un sentido traslacional, rotacional, o en forma de vibraciones. A medida que sea mayor la energía cinética de un sistema, se observa que este se encuentra más caliente; es decir, que su temperatura es mayor (Yunus, 2006).

2.2. Marco teórico estadístico

2.2.1. Análisis de Datos Funcionales (ADF)

El Análisis de Datos Funcionales es una alternativa novedosa de gran utilidad que está siendo implementada por investigadores en diversos campos del conocimiento, principalmente gracias a los aportes de Ramsay y Silverman (2005). Esta consiste en trabajar con muestras de funciones y no con datos escalares. Dichas funciones son cuadrado integrables pertenecientes al espacio L^2 las cuales tienen la gran ventaja de representar fielmente el comportamiento de una variable observada en función del tiempo, espacio u otra variable continua sin necesidad de verificar hipótesis bastante restrictivas, como cuando se emplean otras técnicas.

Cuando se realizan análisis en variables continuas como la Humedad relativa, la cual en un día cualquiera presenta 24 mediciones (una por hora), por los métodos tradicionales se podría presentar un resumen estadístico que contenga las medidas de tendencia central y de dispersión, sin embargo este resumen presenta limitaciones a la hora de observar la evolución por hora de esta variable a lo largo del día. Ya que entre mediciones se presentan infinitos registros que por restricciones del número de mediciones que se puedan tomar no se tienen en cuenta. Es por esto que representar la humedad relativa diaria mediante una curva que contemple el comportamiento continuo de dicha variable resulta una muy buena alternativa para evitar las limitaciones mencionadas. El interés se centrará entonces en estimar dicha curva a la cual se llamará dato funcional.

2.2.2. Construcción de Datos Funcionales

Para explicar el proceso de construcción de los datos funcionales es necesario definir algunos conceptos como el de variable aleatoria funcional, la cual es definida por Ferraty y Vieu (2006) de la siguiente manera: “Una variable aleatoria X se llama variable funcional (*f.v.*) si toma valores en un espacio infinito (o espacio funcional). Una observación x de X se denomina dato funcional”. Teniendo claro esto, se busca reconstruir la observación funcional utilizando las observaciones discretas y reconociendo su posible error de medición.

Un dato funcional se puede representar de la siguiente forma: sea $x = x(t); t \in \tau$ un dato funcional y sean $y_1, y_2, \dots, y_n$ las n observaciones de la función x para los valores $t = 1, t = 2, \dots, t = n$ donde t corresponde usualmente a tiempo. Bajo el supuesto de que x es una función suave, es posible utilizar la amplia teoría desarrollada para la regresión en el contexto funcional, al representar a $x(t)$ como una combinación lineal de funciones de una base, como

se muestra a continuación:

$$x(t) = \sum_{k=1}^K c_k \phi_k(t) \quad (2-2)$$

Donde $\phi_k = (\phi_1(t), \dots, \phi_k(t))$ son K funciones conocidas de una determinada base de funciones y los c_k son los coeficientes desconocidos en la combinación lineal los cuales se deben determinar por el criterio de mínimos cuadrados, representada en forma matricial se tiene que:

$$SMSSE(y|c) = (y - \Phi c)'(y - \Phi c). \quad (2-3)$$

Derivando la ecuación 2-3 respecto a c se tiene la ecuación:

$$2\Phi\Phi' - 2\Phi'y = 0. \quad (2-4)$$

Se obtienen entonces los coeficientes estimados $\hat{c}$ que minimizan la suma de cuadrados del error, los cuales se representan así:

$$\hat{c} = (\Phi'\Phi)^{-1}\Phi'y. \quad (2-5)$$

Los valores ajustados en los nodos de observación son:

$$\hat{y} = \Phi\hat{c} = \Phi(\Phi'\Phi)^{-1}\Phi'y \quad (2-6)$$

Obteniendo la curva ajustada:

$$\hat{y}(t) = \hat{c}'\phi(t) \quad (2-7)$$

Sin embargo, para obtener buenos resultados al crear las curvas ajustadas es necesario hacerse preguntas como, ¿Qué familia de funciones base utilizar? y ¿Cual es el número (K) óptimo de funciones?. Preguntas que se discutirán a continuación.

Selección de la familia de función base

Un sistema de bases de funciones es un conjunto conocido de funciones matemáticamente independientes que tienen la propiedad de poder representar arbitrariamente bien cualquier función, mediante una suma ponderada o combinación lineal de un número K suficientemente

grande de funciones de dicha base (Ramsay y Silverman, 2005). Para elegir la base de función se debe tener en cuenta las características de los datos y buscar la similitud la base de función. Por ejemplo, la base B-spline se utilizan principalmente para información funcional sin ninguna variación fuertemente cíclica, mientras que la base de Fourier tiende a ser utilizada para describir datos periódicos. Si esta similitud existe, el número de funciones que definirán la base se reduce. Por lo tanto, se debe buscar una base de función capaz de mantener el número K de funciones tan menor a n como sea posible, sin perder de vista los detalles de ajuste.

Elección del número de funciones base

El grado de suavización de los datos y_j en oposición a la interpolación está determinado por el número K de funciones básicas. Se obtiene una representación exacta o interpolación cuando $K = n$, y un exceso de suavización cuando se tiene un número pequeño de bases de función. Sin embargo en ambos extremos se induciría en errores ya que en el primer caso se incluiría la variabilidad y en el segundo se ignoraría ciertas características o detalles que pueden brindar gran información en el fenómeno estudiado mediante la reconstrucción de los datos funcionales. En consecuencia, no se define un número fijo de parámetros K , sino un parámetro que se elige de acuerdo con las características de los datos (Ramsay y Silverman, 2005).

Por lo tanto, para elegir la dimensión de la expansión K se debe encontrar un equilibrio de los extremos mencionados. Para ello Ramsay y Silverman (2005) mencionan al Error Cuadrático Medio (ECM), debido a que este se puede expresar como la suma de un componente que cuantifica la varianza y otro que cuantifica el sesgo de la estimación de $x(t)$ representadas en la siguiente expresión:

$$ECM[\hat{x}(t)] = Sesgo^2[\hat{x}(t)] + Var[\hat{x}(t)] \quad (2-8)$$

Del mismo modo como se indicó anteriormente pero en otros términos, la elección de un tamaño grande de la base de funciones producirá un sesgo cercano a cero, e igual a cero cuando $K = n$, y para conservar el equilibrio de la igualdad (2-8), la varianza se incrementará en gran medida. Lo que se busca es una disminución significativa de la varianza muestral así esto conlleve a un incremento en el sesgo, siempre y cuando este incremento sea aceptable. Una herramienta muy importante para encontrar dicho equilibrio entre otros parámetros es la Validación Cruzada Generalizada (GCV), descrita a continuación.

Validación Cruzada Generalizada (GCV)

Para explicar la Validación Cruzada Generalizada, desarrollada por Craven y Wahba (1979), es importante describir la Validación Cruzada (CV), esta consiste en dejar por fuera una observación en la estimación de x y calcular su error de estimación, haciendo esto para todos

los datos, uno a la vez, y al promedio de los cuadrados de todos estos errores se le conoce como validación cruzada. Se busca entonces el parámetro de suavizado λ_i que minimiza la expresión:

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_{(i)})^2 \quad (2-9)$$

Donde el sub-índice (i) indica que en la estimación del dato funcional x se omitió la observación i . Para encontrar entonces un λ óptimo, se debe calcular el criterio 2-9 para una secuencia de λ 's y elegir aquel que arroje su valor mínimo. Debido a la carga computacional y a que en la practica no solo se utiliza un dato funcional sino muestras de ellos se emplea una simplificación del anterior método, la cual propone hacer:

$$CV2(\lambda) = \frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - \hat{y}_{(i)}}{1 - S_{\lambda ii}} \right)^2 \quad (2-10)$$

Siendo $S_{\lambda ii}$ la diagonal de la matriz de suavizado $S\lambda = \Phi(\Phi'W\Phi + \lambda R)^{-1}\Phi'W$, donde $R = \int D^2\phi(s)D^2\phi'(s)ds$ y contiene la integral de la segunda derivada de las funciones bases que conforman las curvas. Este cambio es equivalente a calcular los residuales en cada ajuste con cada dato omitido. Por lo tanto, se evitará suavizar n veces un dato funcional y se libera gran parte del esfuerzo de procesamiento.

A esta variación y reemplazando en la ecuación 2-10 el termino $1 - S_{\lambda ii}$ por su promedio $n^{-1}traza(I - S_\lambda)$ y escribiendo el numerador de la misma como la suma de cuadrados del error (SSE), se obtiene la expresión que representa la Validación Cruzada Generalizada (GCV):

$$GCV(\lambda) = \frac{n^{-1}SSE}{[n^{-1}traza(I - S_\lambda)]^2} \quad (2-11)$$

Se puede concluir que la GCV es más efectiva que la CV, ya que presenta menor tendencia al sobre suavizado además de corregir los dos problemas mencionados.

2.2.3. Estimación robusta

La robustez de un método de estimación se refiere a su condición para obtener estimaciones insensibles ante posibles violaciones de alguno de los supuestos fijados al especificar un modelo, en particular, el relativo a la distribución admitida para la perturbación aleatoria (Zamar, 1994). En sentido general Kendall y Buckland (1960) brindan una definición similar en el diccionario de términos estadísticos “Si las inferencias se ven poco afectadas por desviarse de esas suposiciones, por ejemplo, si la significancia de las pruebas varían poco

cuando la población se aleja bastante de la normalidad, se dice que las pruebas de las inferencias son robustas. En un sentido bastante más general, un procedimiento estadístico se describe como robusto si no es muy sensible a la desviación de las suposiciones sobre de lo que depende”.

Suponiendo que se quiere estimar la media muestral, que comúnmente se calcula sumando los elementos y dividiendo entre las n observaciones de la muestra, los datos reales con frecuencia contienen valores atípicos, que pueden ser errores graves u observaciones excepcionales. En esta situación, las estimaciones no son confiables para la toma de decisiones. Para remediar este problema, se han desarrollado técnicas estadísticas robustas que dan una respuesta confiable cuando los datos están contaminados. Un estimador menos conocido, especialmente en física e ingeniería, es la mediana de la muestra, en donde se ordena la muestra de menor a mayor y el estimador es la observación del medio (si el número de observaciones es par, se toma el promedio de las dos observaciones en el medio). Tanto la media muestral como la mediana muestral son estimadores de ubicación, ya que miden la posición general de los datos.

Se dice que la mediana muestral es un estimador robusto, ya que con frecuencia un valor atípico no tiene efecto en la estimación, a diferencia de la media muestral que es muy sensible a los valores atípicos. En muchos cursos de estadística clásica se prefiere la media muestral por varias razones, como su facilidad de cálculo y el hecho de que se presta mucho más a las manipulaciones matemáticas elementales. Su justificación usual es que es óptima en distribuciones normales, pero este es un razonamiento circular porque Gauss en realidad introdujo la distribución normal como el mejor marco para la media de la muestra (Rousseeuw, 1990). En el campo de las estadísticas robustas, se intenta construir técnicas que no se vean afectadas en gran medida por las desviaciones de la normalidad.

Para una muestra $x_1, x_2, \dots, x_n$ contaminada con valores atípicos, una media robusta es la media de la parte central $1 - 2\alpha$ de la distribución, por lo que las observaciones αn se recortan de cada extremo, es decir, para estimar la media se determina la fracción (%) de observaciones que se recortarán a cada extremo de la muestra x antes de calcular la media. La información perdida al no considerar estos valores atípicos, que pueden deberse a errores de digitación o circunstancias excepcionales, está contenida en la variabilidad de la muestra, por lo que este estimador lo que busca es disminuir el efecto de los valores atípicos sobre el estimador obteniendo una mejora en la variación de la muestra sin perder tanta información.

2.2.4. Datos atípicos

La definición de *dato atípico* mas citada en la literatura la presenta Grubbs (1969): “una observación periférica, o atípica, es una que parece desviarse marcadamente de los otros

miembros de la muestra en la que ocurre”. Sin embargo, esta definición esta basada en muestras y un modelo o población no está involucrado en la definición. Por lo tanto Liu (2014) amplían este concepto tomando definiciones de diferentes autores y presentan la siguiente: “los valores atípicos son puntos de datos que se desvían marcadamente de (i) la muestra en la que ocurren o (ii) el modelo supuesto, ya sea una distribución de probabilidad o un modelo estadístico como el modelo de regresión de mínimos cuadrados”.

Los valores atípicos pueden llegar a sesgar las estimaciones, provocando, por ejemplo, desplazamiento en la media hacia la izquierda o derecha. Cohen *et al.* (2013) mostraron que con un solo valor no convencional en una muestra, el coeficiente de regresión no fue estadísticamente significativo y el R^2 disminuyó desde 0.43 en los datos originales (sin valores atípicos) hasta cero en presencia del valor atípico. Tales observaciones extremas pueden estar reflejando alguna alteración en la característica medida, o bien pueden ser el resultado de un error en la medición, tabulación de los datos, acontecimientos extraordinarios, valores extremos y en el peor de los casos, causas no conocidas. Este tipo de datos debe ser tratado y analizado por separado al resto de los datos.

2.2.5. Profundidad funcional

La medida de profundidad se introdujo originalmente en el análisis multivariado para medir la centralidad de un punto con respecto a una nube de puntos. La profundidad proporciona una forma de ordenar puntos en un espacio euclidiano desde el centro hasta la periferia, de tal manera que los puntos más cercanos al centro tendrán mayor profundidad (Torres *et al.*, 2011). La noción de profundidad se ha extendido recientemente a los datos funcionales. La profundidad funcional mide la centralidad de una curva x_i dentro de un conjunto de curvas $x_1, \dots, x_n$. Las medidas de profundidad mas populares según Torres *et al.* (2011) son:

- **Fraiman-Muniz depth (FMD) (Fraiman y Muniz, 2001):** sea $F_{n,t}(x_i(t))$ la función de distribución empírica acumulada para los valores de las curvas $\{x_i(t)\}_{i=1}^n$ en un instante de tiempo $t \in [a, b]$ dada por:

$$F_{n,t}(x_i(t)) = \frac{1}{n} \sum_{k=1}^n I(x_k(t) \leq x_i(t)) \quad (2-12)$$

Donde $I(\cdot)$ es la función indicadora. La FMD para una curva x_i con respecto al conjunto $x_1, \dots, x_n$, está dada por:

$$FMD_n(x_i(t)) = \int_a^b D_n(x_i(t)) dt \quad (2-13)$$

donde $D_n(x_i(t))$ es la profundidad de punto de $x_i(t)$, $\forall t \in [a, b]$ dada por:

$$D_n(x_i(t)) = 1 - \left| \frac{1}{2} - F_{n,t}(x_i(t)) \right| \quad (2-14)$$

- **Profundidad H-modal (HMD) (Cuevas *et al.*, 2006)**: el modo funcional (que se basa en el concepto de moda) se define como la curva más densamente rodeada por las otras curvas de la muestra. HMD se expresa como:

$$MD_n(x_i, h) = \sum_{k=1}^n K\left(\frac{\|x_i - x_k\|}{h}\right) \quad (2-15)$$

Donde K es una función kernel, $\| \cdot \|$ es una norma en un espacio funcional y h es el parámetro de ancho de banda. Una de las normas más ampliamente utilizados para un espacio funcional es L^2 , dado como:

$$\|x_i(t) - x_j(t)\| = \left(\int_a^b (x_i(t) - x_j(t))^2 dt \right)^{1/2} \quad (2-16)$$

2.2.6. Box Plot funcional

La extensión de esta técnica en el contexto funcional es introducida por Sun y Genton (2011) quienes se basan en el concepto de profundidad de banda (López-Pintado y Romo, 2009) que permite ordenar una muestra de datos funcionales desde el punto central hacia afuera en ambas direcciones, deduciéndose de este orden los cuantiles funcionales y una medida de centralidad. Sea $x_1(t), x_2(t), \dots, x_n(t)$ una muestra de datos funcionales en $t \in \tau$ y τ un intervalo en $\mathfrak{R}$, la gráfica de una función x en un subespacio del plano que se define $G(x) = \{(t, x(t)) : t \in \tau\}$, la banda en $\mathfrak{R}^2$ delimitada por las curvas $x_{i_1}, x_{i_2}, \dots, x_{i_k}$ está definida como:

$$B(x_{i_1}, x_{i_2}, \dots, x_{i_k}) = \left\{ (t, y) : t \in \tau, \min_{r=1, \dots, k} x_{i_r}(t) \leq y \leq \max_{r=1, \dots, k} x_{i_r}(t) \right\} \quad (2-17)$$

Esta banda no es sino la región delimitada por el mínimo y el máximo de las k curvas que la conforman para cada $t \in \tau$. Se define además para cualquier curva x en $x_1, \dots, x_n$ y un valor de j fijo, con $2 \leq j \leq n$, la profundidad de banda (BD, band depth) como la cantidad

$$BD_n^{(j)} = \binom{n}{j}^{-1} \sum_{1 \leq i_1 < i_2 < \dots < i_j \leq n} \mathbf{I} \{G(x) \subseteq B(x_{i_1}, x_{i_2}, \dots, x_{i_j})\} \quad (2-18)$$

Esta cantidad representa la proporción de bandas $B(x_{i_1}, x_{i_2}, \dots, x_{i_j})$ determinadas por las j diferentes curvas $x_{i_1}, x_{i_2}, \dots, x_{i_j}$ que contienen completamente la gráfica de x dividida por la cantidad de bandas de tamaño j que se pueden formar con los n datos funcionales, y donde $\mathbf{I}$ es la función indicadora. Cuanto mayor sea el valor de BD mayor será la posición central de la curva, y la curva con la profundidad de banda más grande corresponde a la mediana de la muestra.

2.2.7. Valores Atípicos Funcionales

Un conjunto de muestras funcionales puede tener elementos que aunque no incorporen error en sí mismos, pueden presentar patrones diferentes de los otros. Las mediciones de profundidad descritas anteriormente, utilizadas para identificar valores atípicos en muestras funcionales, permiten establecer conjuntos de observaciones a lo largo del tiempo ajustadas a curvas, en lugar de sólo los valores medios en el intervalo de tiempo de medición.

Profundidad y valores atípicos son conceptos inversos; por lo tanto, un valor atípico para una muestra funcional tendrá una profundidad considerablemente menor. Se buscan las curvas con menor profundidad para identificar los valores atípicos funcionales.

3. Metodología

En la realización de este proyecto es necesario para cada etapa realizar una minuciosa revisión bibliográfica, donde los principales temas de búsqueda serán el análisis de datos funcionales, la estimación robusta y como integrar estas técnicas. Además, temas referentes a las variables climáticas para poder conocer y comprender el comportamiento de estas, en especial de la humedad relativa, variable medida por el Sistema de Vigilancia de la Calidad del Aire de Santiago de Cali. Con el fin de investigar el comportamiento de las curvas diarias de humedad relativa, se propone un análisis exploratorio en donde se detecten las curvas atípicas y el efecto de estas sobre la curva media de humedad relativa y para esto, como ya se mencionó anteriormente, se trabajó con los datos registrados por el sistema de vigilancia de calidad del aire de Santiago de Cali en la estación Compartir, ubicada en el oriente de la ciudad.

3.1. Validación e inspección de los datos

Específicamente antes de aplicar las técnicas estadísticas que se proponen en este estudio, se realiza un análisis exploratorio a los datos suministrados por el DAGMA que corresponden a registros promedios de las 24 horas del día (horarios) de diferentes variables de la calidad del aire y del clima. De esta manera y teniendo en cuenta el fin del estudio, conocer y resaltar el comportamiento y posibles relaciones de las diferentes variables del clima contenidas, principalmente de la humedad relativa. Lo ideal es contar con 24 observaciones diarias para los 365 días del año en cuestión, no obstante, la calidad de la base de datos en cuanto a la recolección de estos se ve afectada por las diferentes anomalías, como fallas en el sistema de energía o mantenimiento en la estación de monitoreo que perjudican los instrumentos de medición. Este tipo de sucesos ocasiona que la base tenga datos faltantes, que puede ser desde el dato de una hora en específico hasta los datos de un mes completo. De esta manera, surge la necesidad de realizar un control de calidad de la base de datos con el fin de aprovechar de la mejor manera la información disponible.

En primera instancia se inspeccionó la base y se procedió a identificar, para cada mes, los días que tengan el registro completo de la información con el fin de contar con una mejor muestra para la construcción de las curvas diarias de humedad relativa para el periodo y lugar de estudio establecido.

3.2. Construcción de las curvas diarias de la humedad relativa

Se procedió a construir las curvas o los datos funcionales diarios de la humedad relativa y de esa forma reflejar el comportamiento para cada día sin perder información. Esta técnica se puede desarrollar mediante métodos que se escogen según el comportamiento de los errores observacionales, entre los métodos se encuentran el polinomio interpolador de LaGrange, splines interpoladores o regresión no paramétrica, entre otros (Navarro Pérez, 2004), teniendo en cuenta que la humedad relativa sigue un comportamiento cíclico y según la revisión bibliográfica realizada para la reconstrucción de las curvas se propuso las bases B-spline y Fourier, ya que estas familias de funciones son usadas con mayor frecuencia para modelar el comportamiento de variables climatológicas por su tolerancia y demás propiedades.

Para la elección de la familia de funciones, el número de funciones base y parámetro de suavización óptimo, se utilizó la técnica de validación cruzada generalizada descrita en secciones anteriores. Luego de obtener las curvas diarias de humedad relativa se realizó un análisis descriptivo funcional, se calculan medidas de tendencia central y medidas de dispersión funcional y además se indaga la variabilidad de la muestra de funciones.

3.3. Detección de las curvas diarias de humedad relativa atípicas

Con el fin de cumplir el propósito de este estudio la detección de los valores atípicos se trabajó diferente a la convencional ya que para cada día se tienen curvas o funciones y no datos escalares, para esto se tuvieron en cuenta dos métodos mencionados por Billor (2016): Profundidad funcional y Box-Plot funcional. Estas herramientas nos permitieron identificar, para cada mes, las curvas diarias de humedad relativa atípicas o con comportamientos no convencionales que pueden ser de forma o valores numéricamente distantes del resto.

Entre las medidas de profundidad funcional más populares están la *Fraiman - Muniz (FMD)* y *H-modal (HMD)* que miden la centralidad de cada curva dentro de la muestra curvas y ya que profundidad y valores atípicos son conceptos inversos, nos permitieron establecer el conjunto de curvas funcionales con menor profundidad para denominarlas como curvas diarias de humedad relativa atípicas. Además, se usaron herramientas gráficas como el Box-Plot funcional, que es una extensión de la técnica convencional para estudiar la variabilidad de la muestra de funciones y evidenciar las posibles curvas diarias atípicas para cada mes. Para este estudio se buscará el método más potente para identificar las curvas atípicas de estos dos mencionados y se trabajará con este.

3.4. Estimación robusta funcional

Posteriormente se extendió el concepto de robustez a los datos funcionales, esta técnica da como resultado la estimación de la curva media de humedad relativa diaria para el periodo de estudio con la ventaja de que se mitiga el impacto ejercido por los comportamientos del conjunto de curvas identificadas como atípicas. Para esto, después de identificar las curvas atípicas se organizaron de menor a mayor utilizando la medida de profundidad funcional para determinar las curvas extremas en el conjunto de curvas atípicas de la muestra, es decir, las curvas con un valor de profundidad funcional menor son las que están más alejadas del centro. Esto con el fin de realizar una estimación de la curva media de porcentaje de humedad relativa diaria recortando un porcentaje adecuado de curvas atípicas, de tal manera que no se pierda mucha información y se logre poca variabilidad en la muestra y además, mitigar el impacto de estas curvas atípicas sobre el estimador.

Para determinar el porcentaje óptimo de recorte, se procedió a variar el mismo para identificar el cambio sobre la desviación estándar del porcentaje de humedad relativa, es decir, se observó en qué punto se estabiliza la curva de desviación estándar o el cambio de la variabilidad no es tan significativo.

4. Resultados

Este capítulo consta de cuatro secciones en las que se presentan los resultados obtenidos de manera detallada a lo largo del estudio.

La primer sección está conformada por el análisis exploratorio de los datos mediante técnicas y herramientas de estadística clásica, en esta etapa se manipula por primera vez la base de datos con el fin de identificar los patrones de comportamiento y similitud, además de conocer la calidad y cantidad de la información contenida en esta. Todo esto, teniendo en cuenta los meses y las 24 horas para cada día del periodo de estudio.

Para la segunda sección, después de determinar la familia de funciones, el número de bases y parámetro de penalización se reconstruyen las curvas diarias de humedad relativa para cada mes.

Una vez obtenidas las curvas diarias de humedad relativa, en la tercera sección se exponen los resultados del análisis a través de los métodos anteriormente mencionados, para determinar cuáles de estas curvas son atípicas, con el objetivo de indagar la procedencia y el impacto de estas sobre los estimadores, en especial el de la curva diaria media de humedad relativa. Por último, se presenta la curva diaria media de humedad relativa obtenida por un método robusto, estudiando el rendimiento de este estimador y comparándolo con la curva diaria media de humedad relativa clásica para el periodo de estudio establecido.

4.1. Depuración de los datos y análisis exploratorio

El propósito del Análisis Exploratorio de Datos es inspeccionar o estudiar los datos previamente a la aplicación de cualquier técnica estadística, de esta manera entender el comportamiento de estos y las posibles relaciones entre las variables analizadas. Este análisis permite preparar y organizar los datos, detectar fallas en el diseño y la recolección de estos, además, se puede realizar un tratamiento y evaluación de datos ausentes e identificación de casos atípicos (Figueras y Gargallo, 2003). Esta fase es de gran importancia en toda investigación, puesto que permite comprender los datos en estudio y familiarizarse con ellos, también obtener información preliminar para encontrar las metodologías estadísticas adecuadas para el estudio, además de cómo proceder teniendo en cuenta los resultados de esta fase. Con este fin, a continuación se muestra un resumen del análisis exploratorio de la Humedad relativa, variable monitoreada por el SVCASC en la estación Compartir en el año 2016.

Para este año la cantidad de datos recolectados de la Humedad relativa es 8172, es decir se tiene un 93 % de datos válidos, siendo este un buen balance del registro de datos, además se tiene en cuenta que a excepción de la variable Temperatura, las variables climatológicas que se monitorearon también presentan un buen porcentaje de datos recolectados. Al realizar el análisis del porcentaje de los datos faltantes (tabla 4-1) se evidencia la dificultad para llevar registro y el estado de las máquinas de medición de la Humedad relativa y el resto de variables climatológicas y contaminantes criterio, en el transcurso del año.

Tabla 4-1.: Balance general del porcentaje de datos recolectados en la estación Compartir durante el 2016.

	<i>Temperatura (C°)</i>	<i>Humedad (%)</i>	<i>Rad. Solar (Watt/M2)</i>	<i>Lluvia (mm)</i>
PROMEDIO ANUAL	27,96	58,89	93,79	0,23
N° DATOS VALIDOS	3461	8172	6212	7193
% DATOS VALIDOS	39,40 %	93,03 %	70,72 %	81,89 %

Teniendo en cuenta los objetivos, y ya que para realizar el análisis de datos funcionales se debe contar con información completa en este estudio se utilizaron los días del mes en los que se contaba con la totalidad de sus mediciones (24 registros). En la Tabla 4-2 se detalla cuantos días válidos hay por mes y el porcentaje de datos válidos para el periodo de estudio.

Tabla 4-2.: Días validos por mes año 2016

	Días Validos	Días Totales	% Días Validos
Enero	31	31	100 %
Febrero	27	29	93,10 %
Marzo	25	31	80,65 %
Abril	27	30	90,00 %
Mayo	27	31	87,10 %
Junio	11	30	36,67 %
Julio	31	31	100 %
Agosto	25	31	80,65 %
Septiembre	29	30	96,67 %
Octubre	30	30	100 %
Noviembre	28	30	93,33 %
Diciembre	30	31	96,77 %
TOTAL	321	365	87,95 %

Analizando todas las mediciones realizadas, el promedio y la desviación estándar de

Humedad relativa por día o las 24 horas del día, se observa que la humedad relativa empieza a decaer a partir de las ocho de la mañana (8:00am) hasta aproximadamente las dos de la tarde (14:00pm), donde se encuentra el punto mínimo de humedad relativa. Se puede afirmar que esta tendencia es producto del aumento de la temperatura en el punto de monitoreo, ya que la humedad relativa y temperatura están inversamente correlacionada, es decir a mayor temperatura es menor el porcentaje de humedad relativa en el ambiente.

Por otra parte, se puede observar que en general se presenta poca dispersión en los registros, sin embargo en la mañana entre las 10:00-11:00 y en la tarde entre las 17:00-20:00 los datos presentan una mayor variación respecto a la media. Además, a partir de las diez, en especial en la tarde, se presentan un mayor número de mediciones que posiblemente son atípicas (figura 4-2), de modo que el promedio del porcentaje de humedad relativa para estas horas se ve afectado e indicando un resultado no confiable.

Tabla 4-3.: Resumen descriptivo Humedad relativa por hora

Hora	Promedio	Desviación	CV	Mínimo	Máximo
1	70,66	6,14	9 %	59	84
2	72,19	5,82	8 %	64	87
3	71,79	4,55	6 %	62	79
4	73,52	4,53	6 %	66	82
5	73,69	5,00	7 %	64	84
6	74,58	5,53	7 %	66	86
7	76,00	6,03	8 %	65	89
8	72,24	5,12	7 %	57	81
9	64,91	6,72	10 %	55	82
10	60,18	9,53	16 %	47	79
11	51,76	9,83	19 %	34	83
12	43,23	4,80	11 %	30	56
13	42,83	8,08	19 %	30	72
14	38,03	7,37	19 %	26	54
15	39,66	7,81	20 %	30	73
16	38,69	7,13	18 %	23	51
17	41,82	9,78	23 %	32	78
18	44,58	10,31	23 %	24	79
19	51,47	11,29	22 %	39	76
20	54,93	9,98	18 %	42	79
21	59,97	8,80	15 %	48	87
22	64,81	7,18	11 %	53	83
23	63,57	5,80	9 %	50	80
24	70,00	6,98	10 %	55	84

Aquí es importante resaltar que se presentó un mínimo de humedad relativa de 23% a las cuatro de la tarde (16:00pm), esto ocurrió el 19 de febrero, de igual manera el valor máximo de humedad relativa fue de 89% a las siete y ocho de la mañana del 29 y 31 de mayo respectivamente, las cuales son posibles mediciones atípicas. Esta información se describe en la tabla 4-3 y se ilustra en las figuras 4-1 y 4-2 donde también se observa que el porcentaje diario de Humedad Relativa es volátil, es decir se presentan cambios que gráficamente parecen significativo de un día a otro.

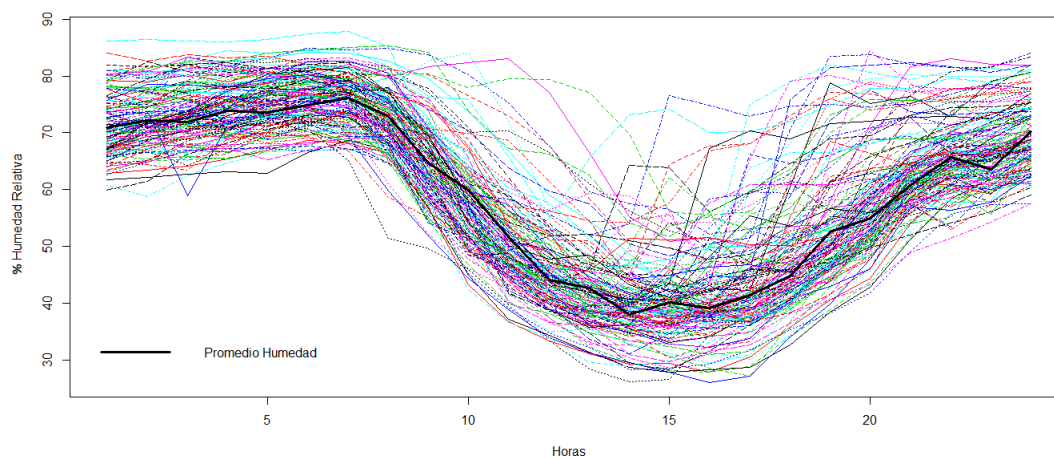


Figura 4-1.: Comportamiento del Porcentaje de Humedad Relativa y su Promedio (Linea Negra) año 2016

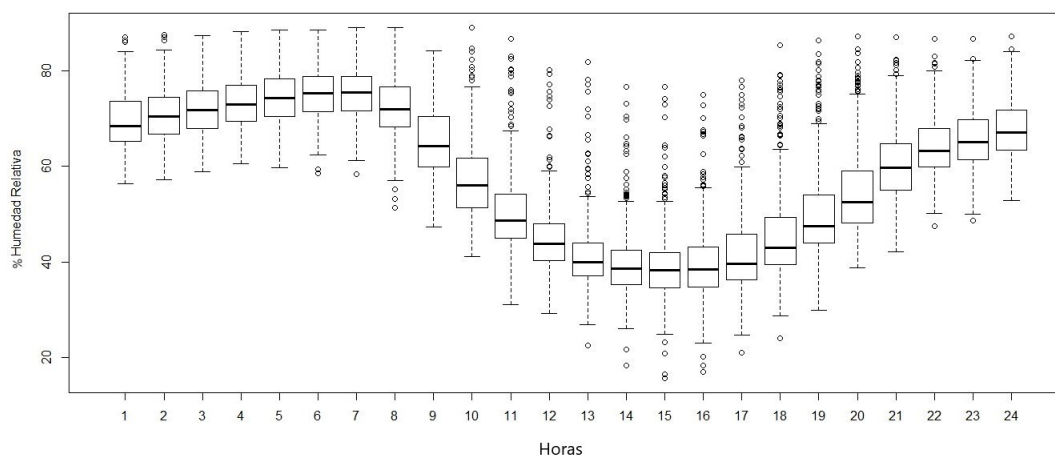


Figura 4-2.: Variación de la Humedad relativa por Hora

De igual manera, según la información descrita en la tabla 4-4 e ilustrada en la figura 4-3, se examina el comportamiento del porcentaje de humedad relativa en todo el año mes

a mes. En el diagrama de dispersión o Boxplot, para todo el conjunto de datos se observa que a excepción del mes de Mayo la variación respecto a la media de los registros es poca y en este mes se presenta el aumento o mayor porcentaje de humedad relativa.

Además se observa que a lo largo del año se presentaron cambios en el porcentaje de humedad relativa, como resultado de los picos y valles en la zona; por ejemplo, a finales de enero y durante el mes de febrero la temperatura aumentó, posteriormente hasta mediados del mes de julio la temperatura disminuye y al finalizar el año se estabiliza. Por otro lado, cabe resaltar que en el mes de junio se recolectó el menor porcentaje de datos válidos.

Tabla 4-4.: Resumen descriptivo Humedad relativa por mes

Mes	Promedio	Desviación	CV	Mínimo	Máximo
Enero	59,70	13,96	23,39 %	33	83
Febrero	53,26	15,93	29,91 %	23	78
Marzo	59,67	15,51	25,99 %	31	83
Abril	60,12	14,93	24,83 %	33	81
Mayo	63,46	16,88	26,60 %	38	89
Junio	59,00	13,16	22,31 %	39	79
Julio	55,74	13,46	24,15 %	31	80
Agosto	51,31	13,18	25,69 %	24	73
Septiembre	61,82	13,89	22,47 %	36	84
Octubre	60,35	14,32	23,73 %	35	82
Noviembre	60,12	15,97	26,56 %	26	84
Diciembre	61,53	14,37	23,35 %	37	85

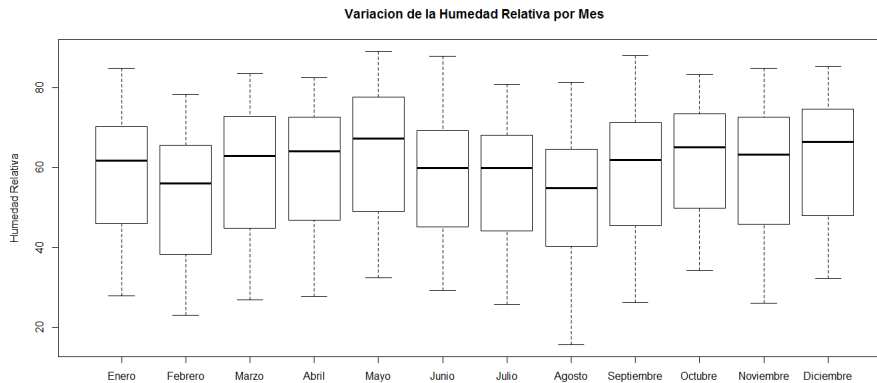
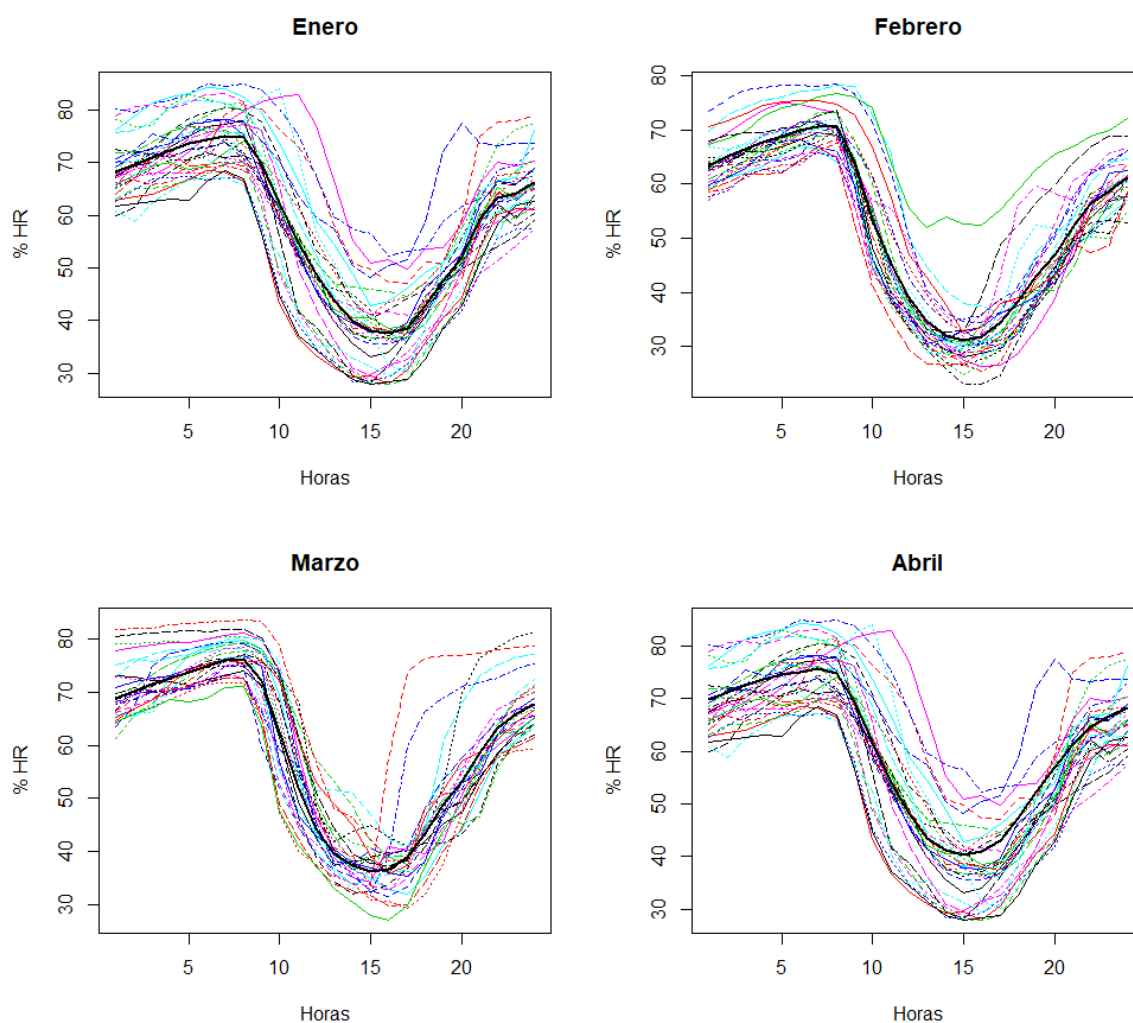
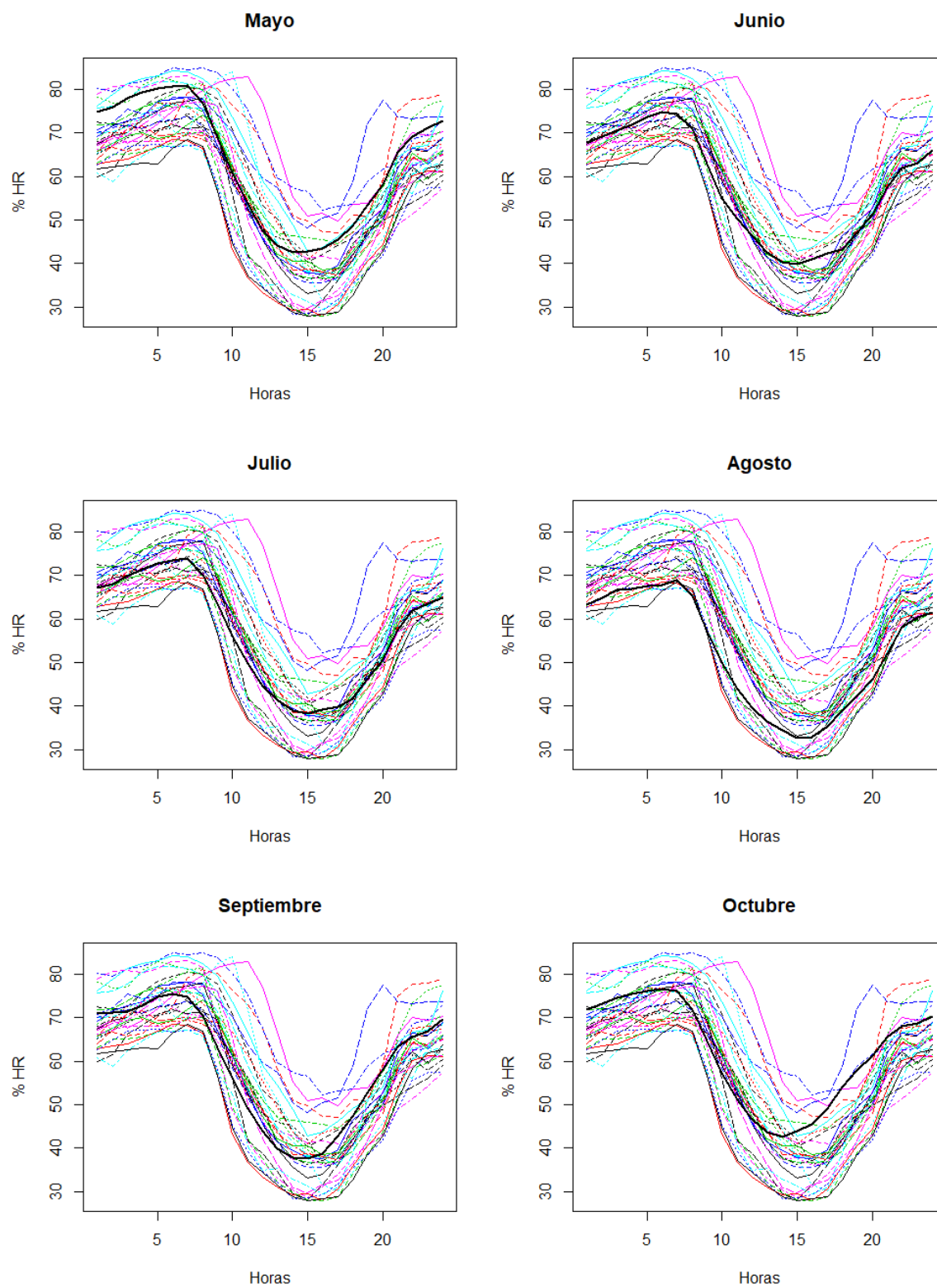


Figura 4-3.: Diagramas de caja de la Humedad Relativa por mes

las figuras 4-4 dan una visión del comportamiento del porcentaje de humedad relativa respecto a las 24 horas del día para cada mes, en donde se puede percibir en primera instancia los comportamientos no convencionales; como es el caso de enero y febrero que se observan algunos días con un comportamiento diferente entre las 11:00am y las 19:00pm. Además, se evidencia la influencia del error de medición, que es natural en el registro de información, dando sentido al análisis de datos funcionales en este tipo de problemas.





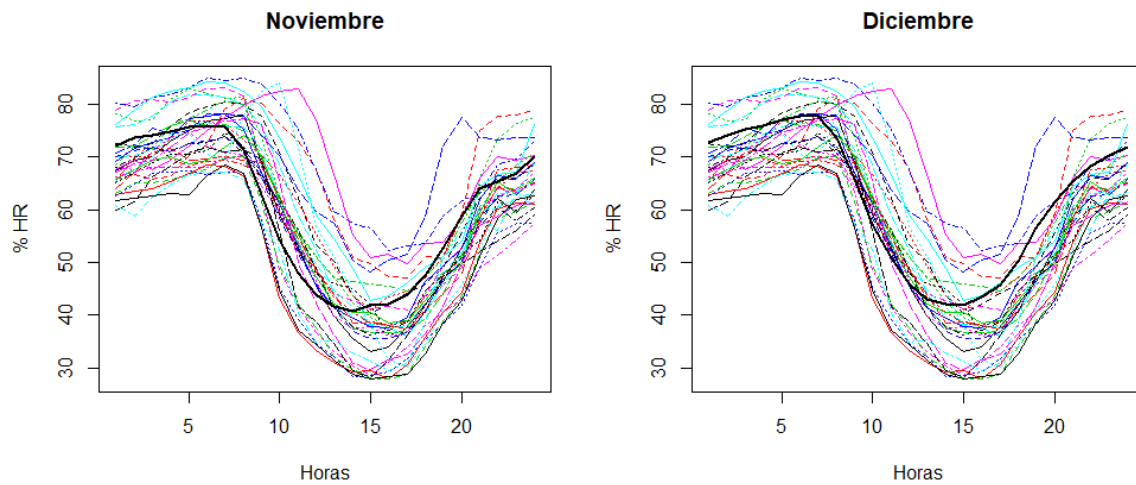


Figura 4-4.: Comportamiento del porcentaje de Humedad Relativa y el promedio para cada mes del año 2016

Como se ilustra en la figura 4-5 La Temperatura y la Humedad Relativa presentan un fuerte correlación negativa ($-0,9689$), esta correlación es evidente ya que la humedad relativa es el cociente entre el contenido de vapor de agua de la masa de aire y la temperatura y si la temperatura del aire aumenta y el contenido de humedad permanece constante, la humedad relativa disminuye.

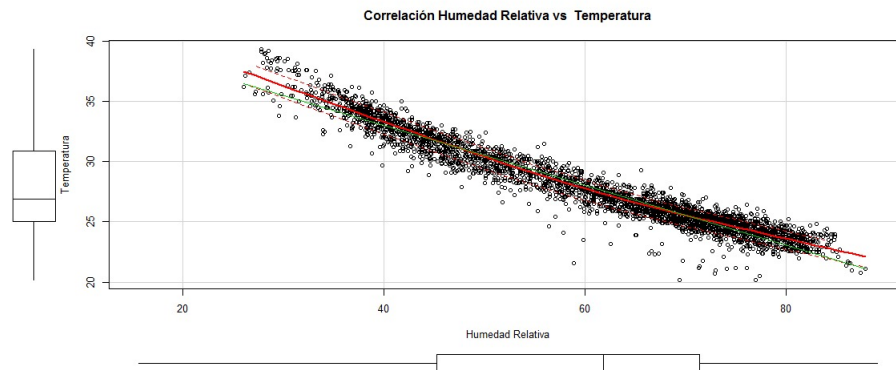


Figura 4-5.: Diagrama de puntos de Humedad Relativa vs Temperatura

4.2. Construcción de datos funcionales

Para crear los datos funcionales a partir de datos escalares, como se mencionó anteriormente, se tendrán en cuenta los días con los registros completos para cada hora (24 registros) y para esto se empleó el paquete *fda* desarrollado por Ramsay *et al.* (2017). Se buscó para las bases de Fourier y B-Spline el número óptimo de funciones base para que representen los datos y así construir los datos funcionales por día, para ello se utilizó el criterio de Validación Cruzada Generalizada, los resultados se presentan a continuación:

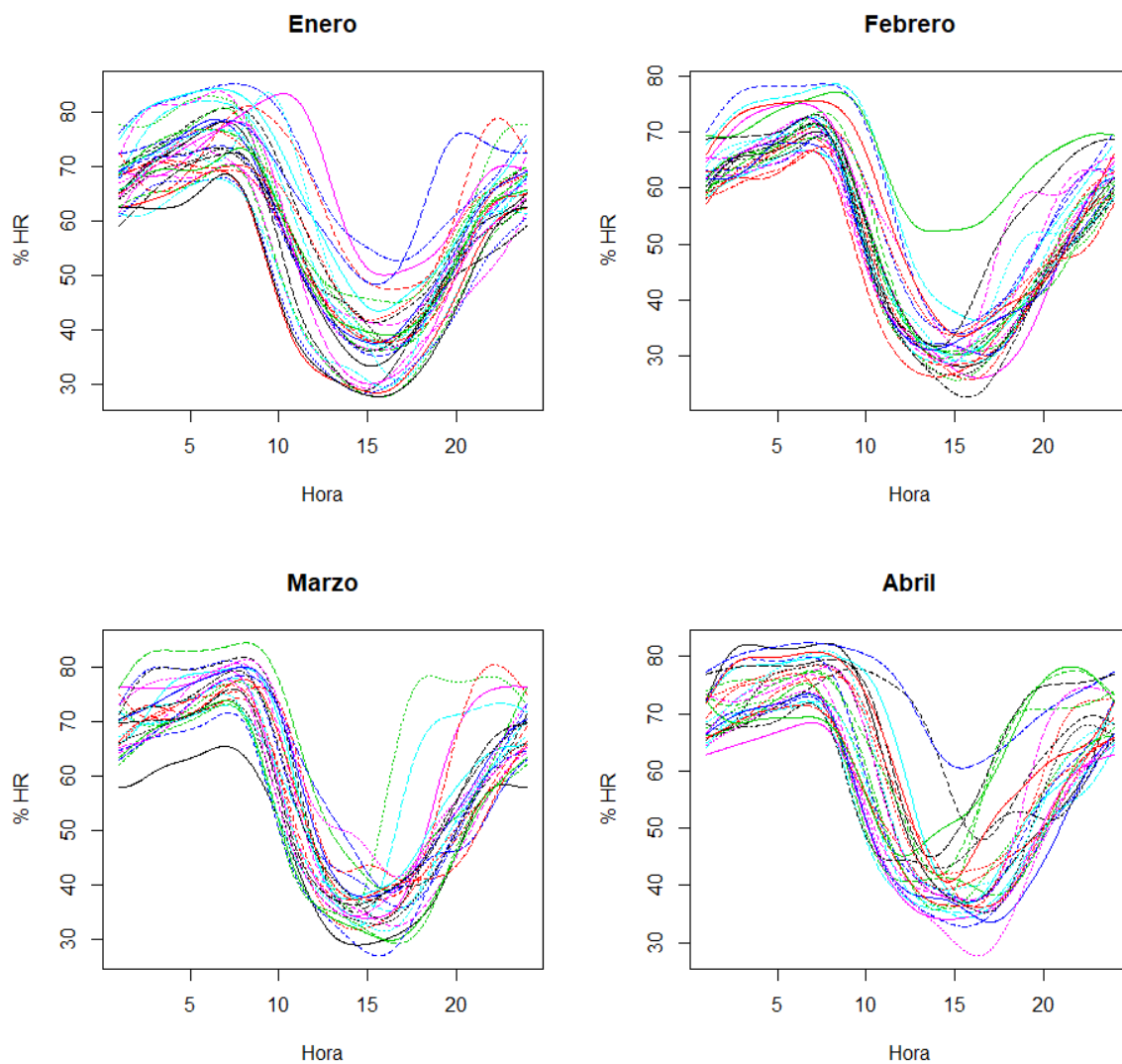
Tabla 4-5.: Comparación Fourier con B-Spline

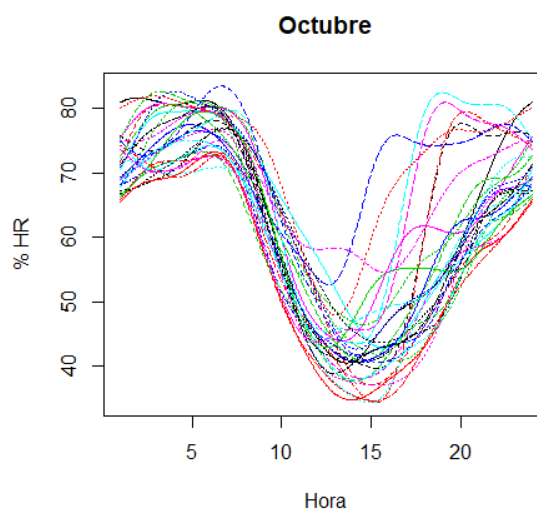
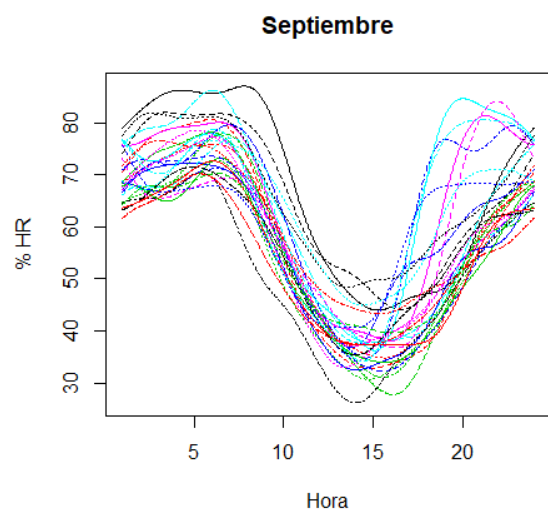
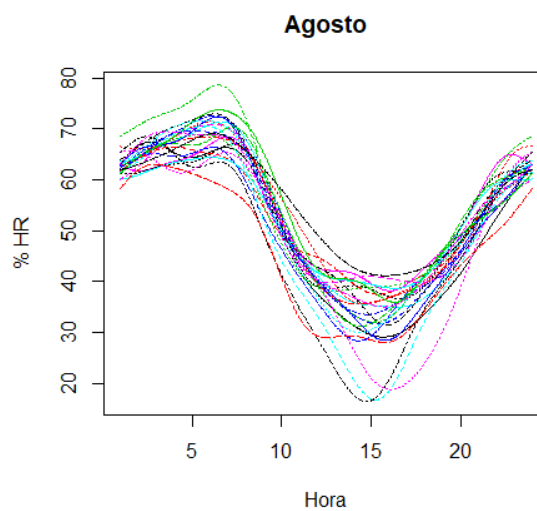
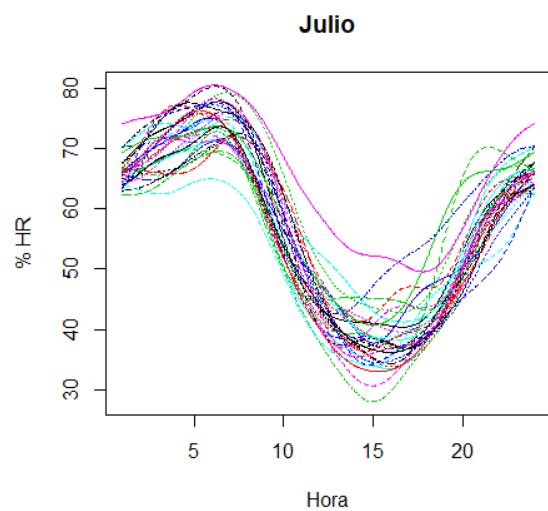
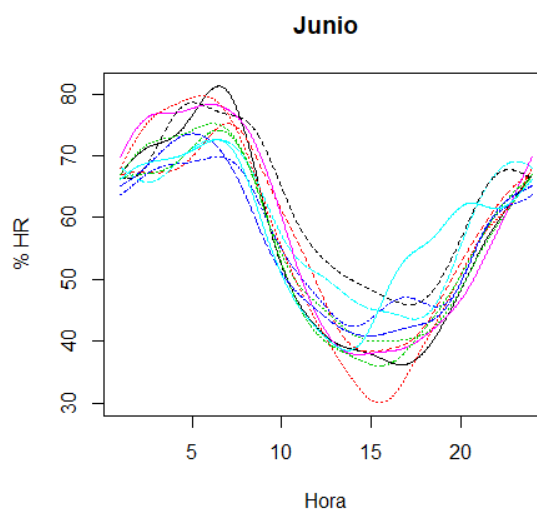
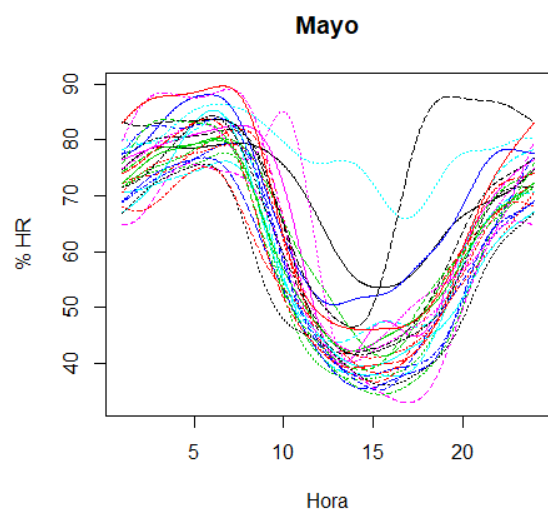
Mes	Bases de Fourier			B Spline		
	N Funciones Opt	Lambda Opt	GCV Opt	N Funciones Opt	Lambda Opt	GCV Opt
Enero	15	0,12	3,44	20	0,02	2,01
Febrero	17	0,09	3,06	23	0,01	1,66
Marzo	15	0,09	4,54	21	0,02	2,41
Abril	15	0,14	3,96	23	0,02	1,8
Mayo	15	0,1	6,11	20	0,04	5,09
Junio	15	0,24	5,73	18	0,06	3,67
Julio	17	0,11	3,82	24	0,07	3,75
Agosto	17	0,11	3,67	17	0,07	3,51
Septiembre	15	0,22	7,83	23	0,04	5,36
Octubre	17	0,15	7,08	23	0,04	5,61
Noviembre	17	0,11	7,27	23	0,02	4,87
Diciembre	17	0,1	5,2	23	0,04	4,49

A pesar de que son muy similares para este estudio se decide trabajar con las bases de Fourier ya que requiere menos funciones que con Bspline para optimizar el error cuadrático medio, además de que estas bases se ajustan mejor cuando se tienen datos periódicos o cíclicos, como lo es el caso del comportamiento del porcentaje diario de humedad relativa. Determinada la familia de funciones base con las que se van a construir los datos funcionales, en la tabla 4-5 se presenta para cada mes el número de funciones base y el parámetro de suavización λ óptimos, sin embargo, para la suavización se optó por elegir el número de funciones modal (*15 funciones*) debido a que no hay mucha diferencia con respecto al criterio, y un parámetro de penalización global (*0.151*), esto con el fin de que el análisis sea comparable mes a mes. A continuación se presentan las curvas suavizadas con bases de Fourier, en donde se puede apreciar el comportamiento diario del porcentaje de humedad relativa para cada mes en el periodo año 2016, en la estación Compartir de la ciudad Santiago de Cali.

En general se puede apreciar un comportamiento similar para cada mes, las horas del día donde se presentan picos o valores más altos del porcentaje de humedad relativa, en la mañana y en la noche, esto debido a que en estas horas la temperatura baja y la

disminución de porcentaje de humedad relativa en las horas de la tarde, aproximadamente hasta las 3:00pm que es donde la temperatura alcanza su punto más alto. Otra característica importante en las gráficas, son las curvas que tienen a simple vista una forma (comportamiento) diferente, es decir, que sobresalen a las demás por encima o por debajo; las cuales pueden estar relacionadas con algún cambio brusco en el clima de ese día, y que se pretende más adelante determinar si son o no curvas atípicas.





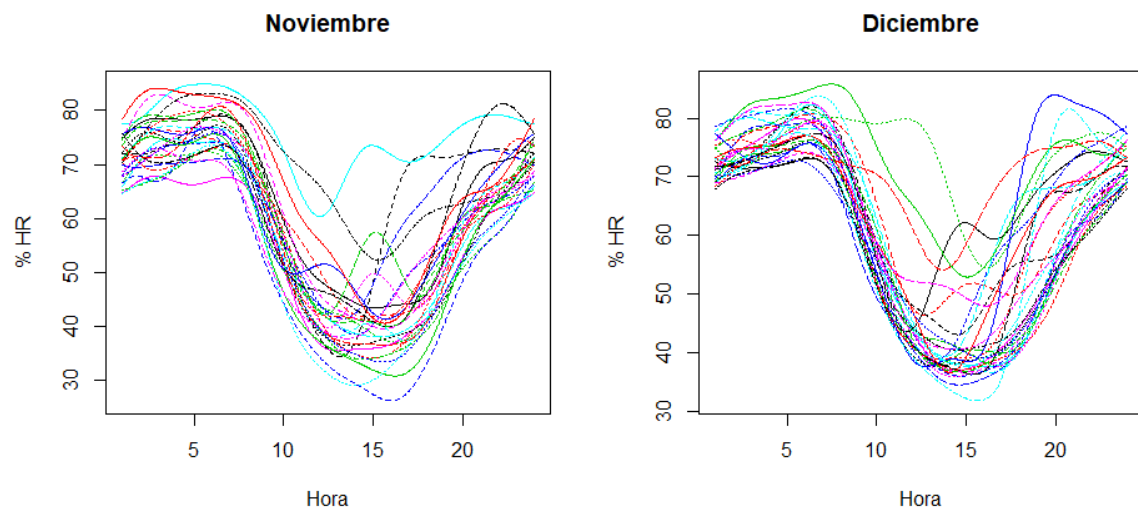
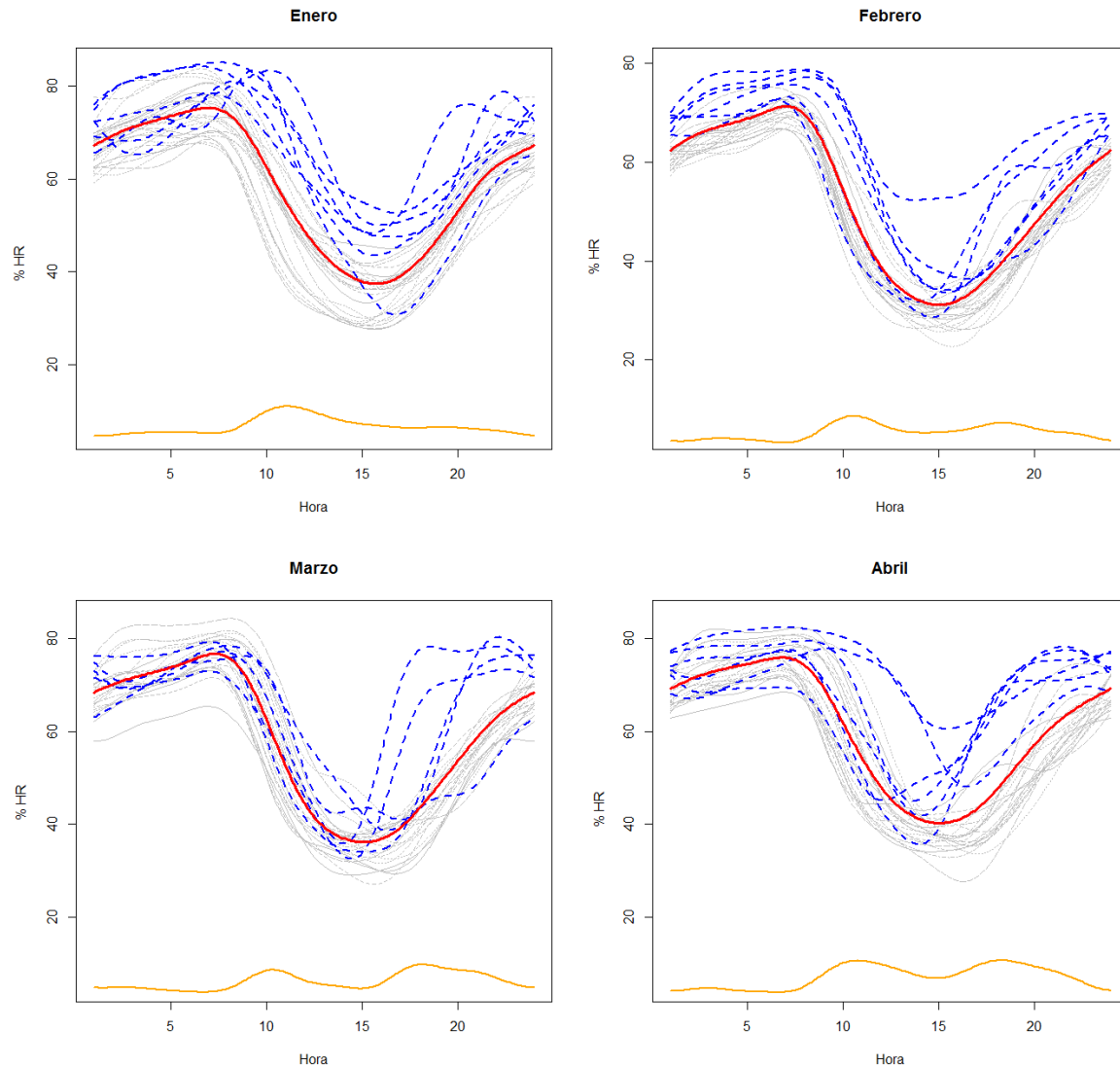
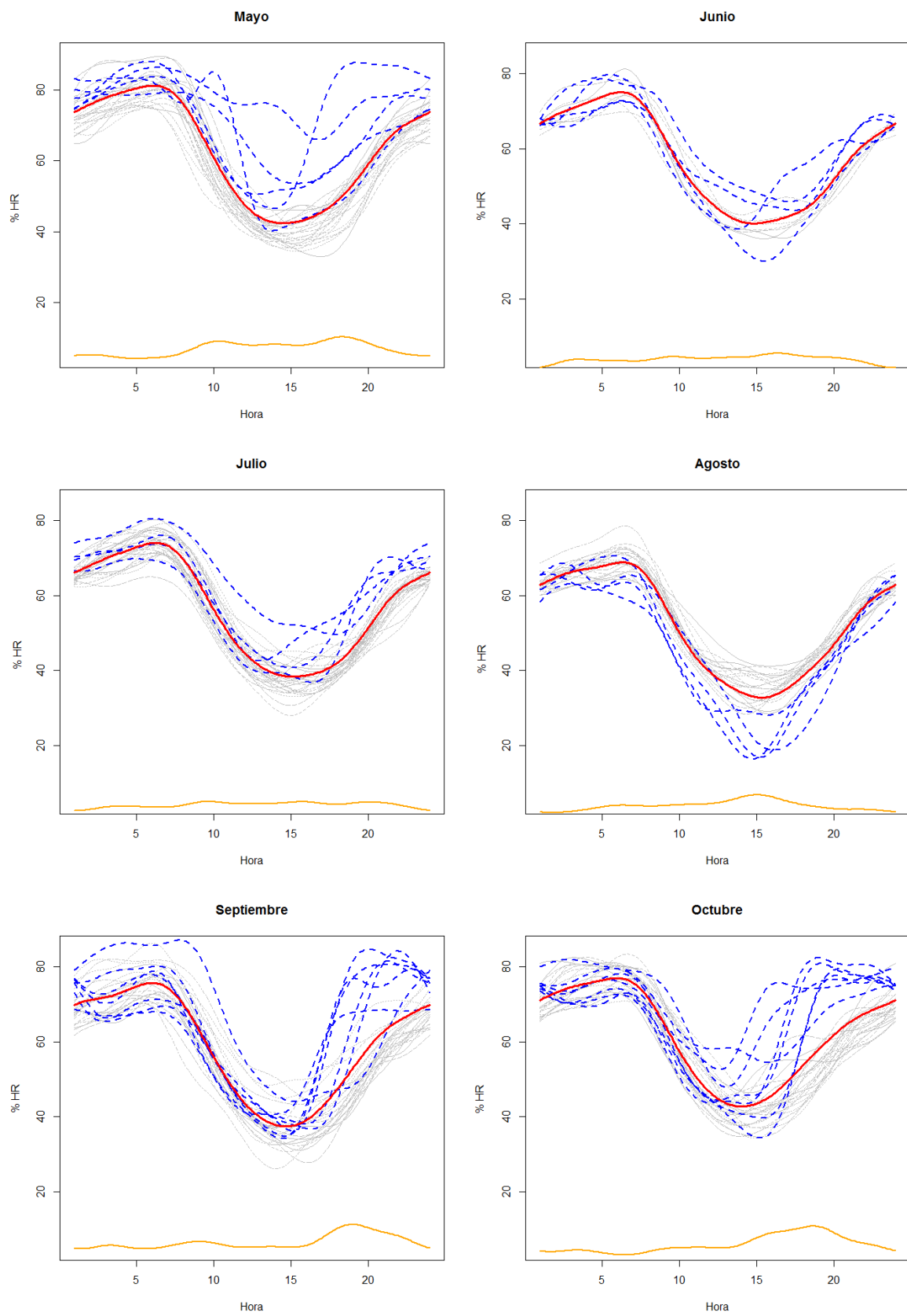


Figura 4-6.: Curvas diarias de porcentaje de humedad relativa para cada mes del año 2016

4.3. Valores Atípicos Funcionales

Para realizar la detección de elementos que presenten patrones diferentes de los otros o curvas atípicas en la muestra funcional para cada mes se utilizaron dos técnicas, los Box-plots funcionales y uno basado en la profundidad funcional, sin embargo el análisis se realizó con el método donde se obtuvieron mejores resultados, en este caso el de la profundidad funcional y el método de Box-Plots funcionales se agrega en la sección de Anexos A. Para determinar las curvas con menor profundidad funcional se utilizó la función *Outliers.depth.trim* del paquete *fda.usc* de Febrero-Bande y Oviedo de la Fuente (2012), la cual brinda el valor de profundidad H-modal para cada muestra y además el umbral para la detección de las curvas atípicas, basada en el bootstrap, como ya se mencionó anteriormente, se buscan las curvas con menor profundidad y se determinará este conjunto como el de datos funcionales atípicos.





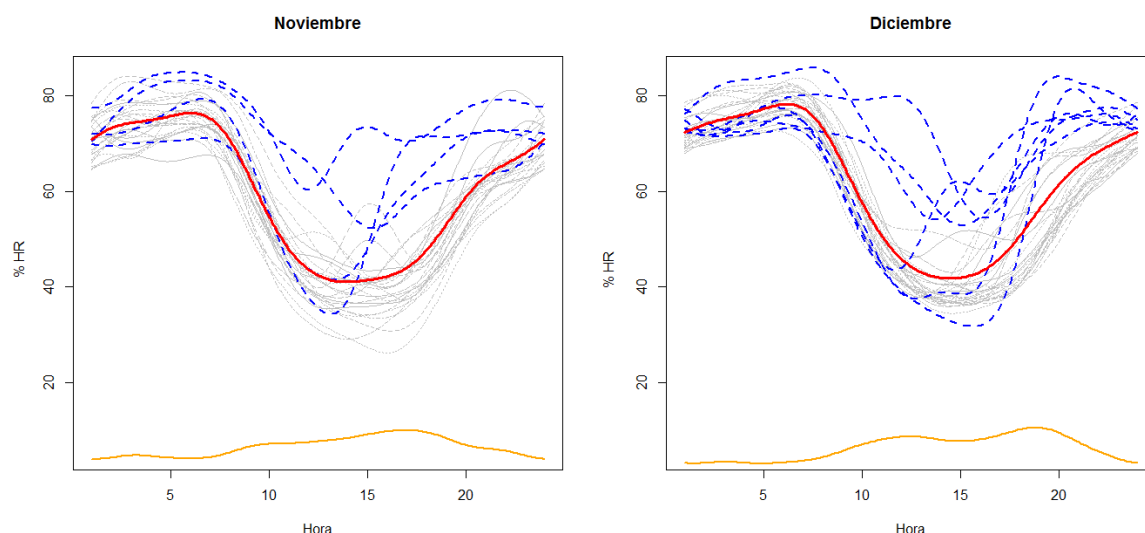


Figura 4-7.: Curvas diarias y curva media funcional (En rojo) y curva de desviación estándar (En naranja) y curvas atípicas (En azul) de porcentaje de humedad relativa para cada mes del año 2016

En las figuras que se presentan en esta sección (4.3) se observa, para cada mes, las curvas atípicas (en Azul) según la medida de profundidad HMD y la curva diaria media del porcentaje de humedad relativa (en Rojo). En general, la cantidad de curvas diarias atípicas por mes es superior a 4; pero uno de los rasgos que expone este análisis es que, a excepción del mes de Agosto, todas las curvas atípicas están por encima de la curva media lo que daría un indicio del efecto de estas curvas sobre el estimador. Respecto a la cantidad de curvas diarias atípicas por mes, el mes de Octubre es el que presenta más cantidad de curvas diarias atípicas, con 7 días y los meses Junio, Julio, Agosto y Noviembre son los que tienen la menor cantidad de días atípicos, con 4 curvas para cada uno. También se observa que en el mes de Diciembre las curvas diarias atípicas son más variables o con cambios más abruptos. Por otro lado, entre las curvas atípicas que se detectaron se ve que algunas de ellas se pueden caracterizar como valor atípico de forma y otros como extremos.

Respecto a la curva diaria media de porcentaje de humedad relativa clásica de cada mes, se puede percibir los cambios o efectos de las curvas diarias atípicas sobre esta, que como ya se mencionó anteriormente, puede estar siendo halada hacia arriba, como es el caso del mes de Diciembre. Sin embargo, este hecho va a ser materia de estudio en la siguiente sección. Por otra parte, se evidencia los momentos del día en que se presentan los picos debido al aumento y disminución de temperatura, siendo alrededor de las 6 de la mañana el porcentaje de humedad relativa más alto. Además, respecto a la función desviación estándar (En naranja) para cada mes se refleja una variabilidad distinta, en donde las curvas atípicas son el factor del aumento de variabilidad, por ejemplo, para el mes de agosto es más variable alrededor de las 3:00pm hora en que se determinaron varias curvas diarias atípicas por

su diferencia con las demás por alcanzar valores mínimos de porcentaje de humedad relativa.

Ahora, teniendo en cuenta que el objetivo de este estudio es el estimador de la media funcional robusta para el año 2016, se extiende este análisis para todo el año, en donde se cuenta con un total de 321 días con los 24 registros, uno para cada hora, completos. Se procede entonces a crear los datos funcionales por día a través del método que se utilizó anteriormente, de igual forma se utiliza la familia de funciones Fourier y la función *min.basis* de la librería *fds.usc* para obtener los valores óptimos para el numero de funciones y el parámetro de suavización λ . De esta manera, para el año completo con 15 funciones y parámetro de suavización $\lambda = 0,13$ se minimiza el criterio CGV.

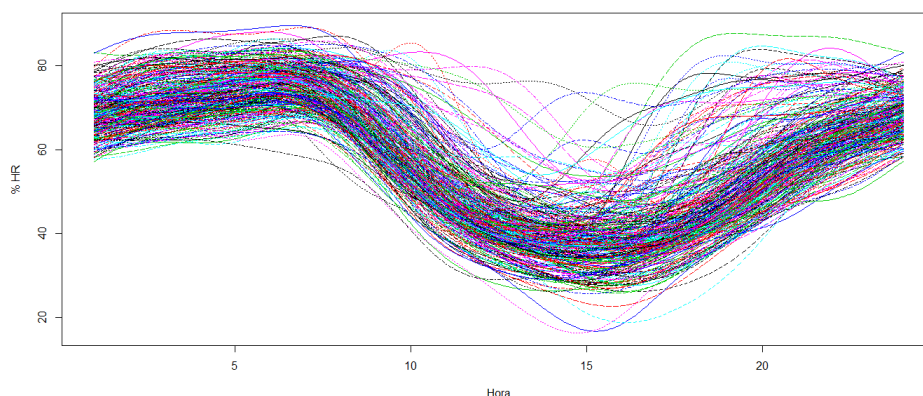


Figura 4-8.: Curvas diarias de porcentaje de Humedad Realtiva del año 2016

En la Figura 4-8 se puede apreciar el comportamiento diario de porcentaje de humedad relativa para los 321 días del año 2016, en la estación de monitoreo Compartir de la ciudad de Santiago de Cali. Como rasgo general se puede observar los momentos del día en que el porcentaje de humedad relativa aumenta hasta alcanzar su máximo (pico) y disminuye debido al aumento de la temperatura. Otra característica que se resalta en la gráfica y que es de gran interés son las curvas que sobresalen a las demás por encima y por debajo entre las 11:00am y 6:00pm aproximadamente, días que se relacionan con anomalías presentadas en el clima de la zona.

Con el fin de estudiar el efecto de estas curvas, al igual que el inicio de esta sección, se procede a determinar si son curvas diarias de humedad relativa atípicas y se analizará posteriormente su impacto sobre la curva media diaria de humedad relativa para el año 2016.

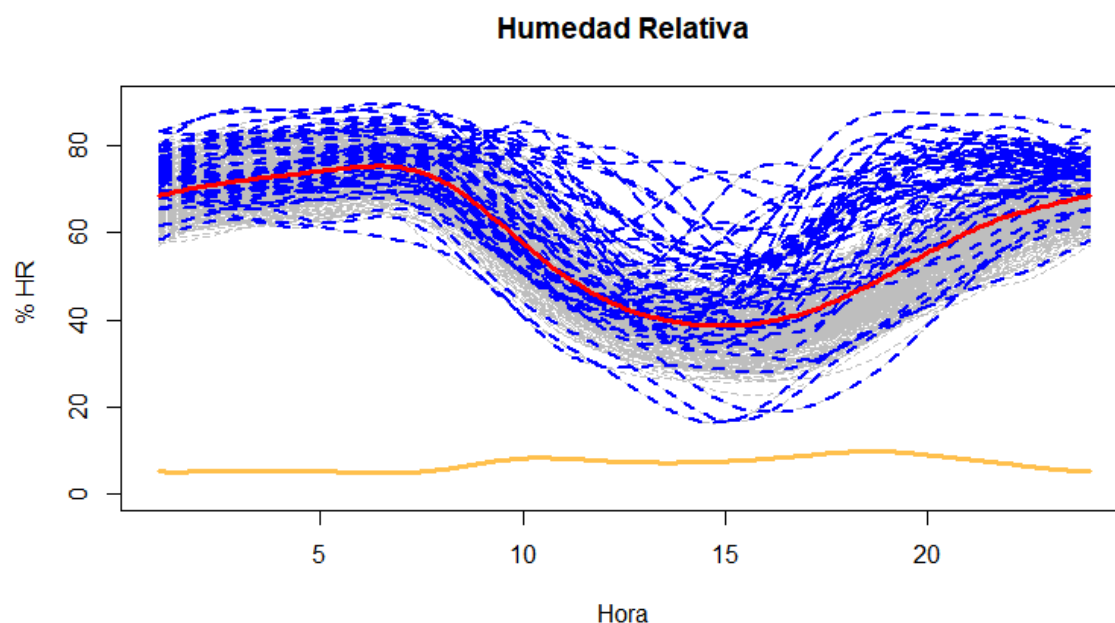


Figura 4-9.: Curvas diarias, curva media funcional (rojo), curvas atípicas (azul), curvas de desviación estándar (naranja) del porcentaje de humedad relativa para el año 2016

Para hacer el análisis de las curvas diarias de porcentaje humedad relativa atípicas en la Figura 4-9 se muestran las que se determinaron (en Azul) según el umbral para la detección generado por la función utilizada, basada en la medida de profundidad HMD, para un total de 53 días o curvas. Al igual que el análisis que se realizó mes a mes, un rasgo general que se observa es que la mayoría de las curvas atípicas están por encima de la curva media de porcentaje de humedad relativa (en Rojo), hecho que será estudiado para determinar si tiene algún efecto sobre este estimador. Por otra parte se observa la variabilidad (en negro) que hay entre las curvas, y entre las 10:00am y 7:00pm hay un cambio notable en las funciones de porcentaje de humedad relativa, se puede notar que este incremento de variabilidad es debido a las curvas diarias atípicas, rango de tiempo en donde normalmente tiende a aumentar la temperatura en la zona; este comportamiento puede ser producto de la disminución de la temperatura causado por una “lluvia inesperada”.

Además, observando la curva media de porcentaje de humedad relativa de cada mes y la anual, también se tiene la misma percepción en donde queda evidenciado los momentos del día en que el porcentaje de humedad relativa alcanza su máximo en la zona, sin embargo, la curva media de porcentaje de humedad relativa anual no tiene cambios tan marcados como los que se observan en cada mes, como se mencionó anteriormente, esto puede ser por la variabilidad de las curvas atípicas en el intervalo de tiempo “crítico” y su efecto sobre este estimador.

4.4. Media funcional robusta

Para cumplir el objetivo principal de este estudio se emplearon técnicas de estadística clásica para datos escalares como el “Trim” o “Recorte”, en donde se organiza la muestra de menor a mayor y se recorta o se quita de la muestra un porcentaje de valores por derecha e izquierda; teniendo en cuenta que estos valores se consideran como observaciones extremas, además, se usaron herramientas mencionadas anteriormente para identificar los casos atípicos y extremos como la profundidad funcional.

Se logra establecer una técnica que consiste en realizar un recorte en la muestra de curvas utilizando el concepto de profundidad funcional con el que se determinó que tan alejados están las curvas del centro, se organizaron las profundidades de las curvas de menor a mayor y recordando que a menor profundidad la curva se encuentra más alejada del centro se realiza un recorte, para esto, se establecieron unos porcentajes de recorte con el objetivo de conocer el efecto causado en las estimaciones y lograr identificar dónde se puede alcanzar un equilibrio perdiendo la menor cantidad de información y consiguiendo una disminución en la variabilidad de la muestra de curvas diarias y estimaciones más confiables y robustas, como se muestra en la figura 4-10.

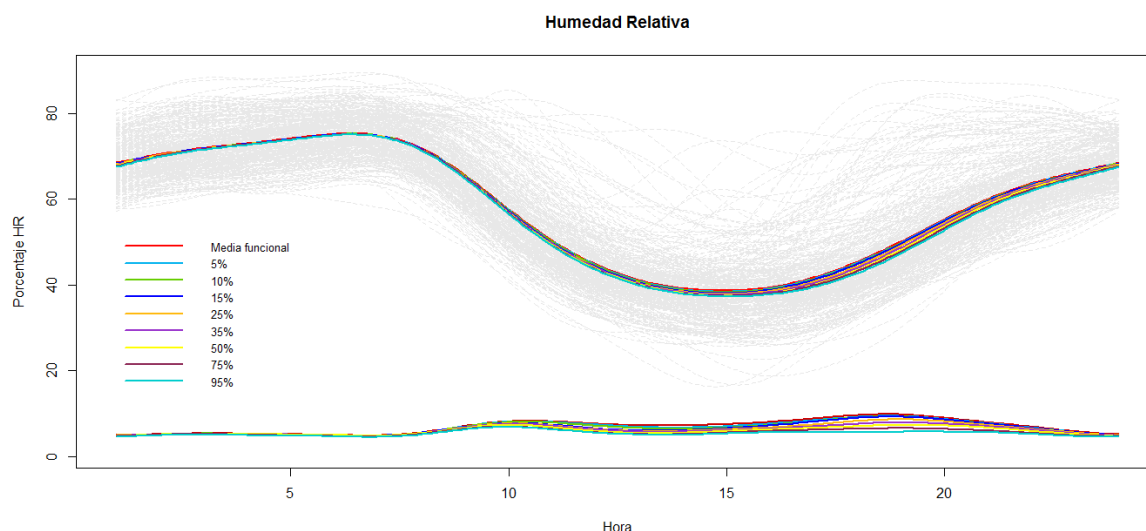


Figura 4-10.: Estimación de la curva media y desviación estándar de porcentaje de Humedad Relativa (En rojo) y las curvas medias y desviación estándar robustas funcionales.

Sin embargo a simple vista no es posible identificar en qué punto se estabiliza o que tanta variabilidad se logra disminuir al ir realizando diferentes niveles de recorte y dada la importancia de establecer un porcentaje de recorte de la muestra de curvas, se procede a variar el porcentaje de recorte y se observa el comportamiento de la variabilidad del conjunto de curvas de porcentaje de humedad relativa.

Para esto, en la figura 4-11 se pueden identificar los cambios a diferentes niveles de recorte, en la curva negra de la parte superior se encuentra la desviación estándar al recortar el 5 % de las curvas, obteniendo el mayor nivel de variabilidad. Posteriormente se realizan disminuciones de un punto porcentual como se observa en las curvas grises y de cinco puntos en las curvas negras, hasta llegar a un porcentaje de recorte del 25 %, observada en la curva de menor variabilidad.

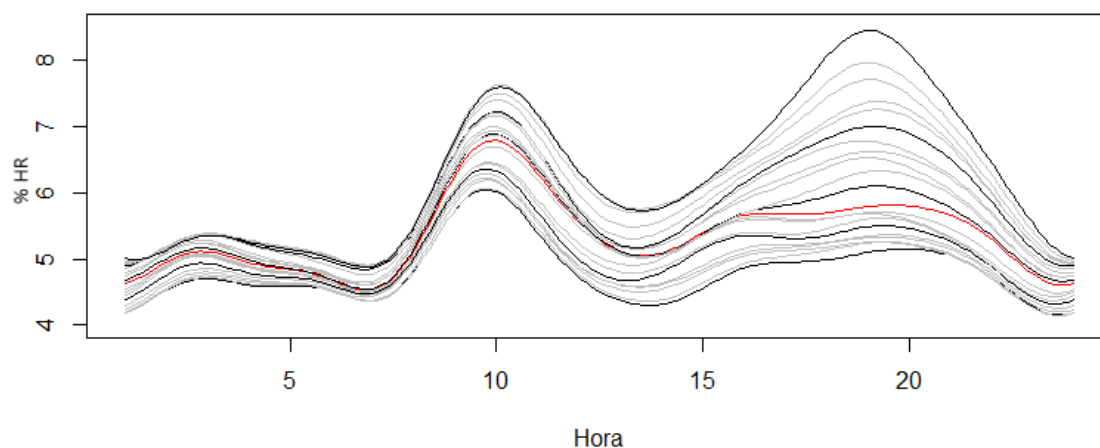


Figura 4-11.: Desviación estándar a diferentes niveles de recorte

Aunque siempre se identifican cambios al aumentar el porcentaje de recorte, se puede percibir que al realizar un recorte del 16 % (curva roja) se obtienen bajos niveles de desviación estándar, similares a los de porcentajes superiores de recorte, es decir, se tiende a estabilizar la variación en la muestra de curvas de porcentaje de humedad relativa. Por lo tanto se decide trabajar con este 16 %, tomando este porcentaje como el óptimo para estimar la media funcional robusta.

De esta manera, en la figura 4-12 se observa e identifica el ajuste que presenta la curva media al ser recortada con el 16 % y se estima su intervalo de confianza, identificando que en las horas donde se presentaba mayor variabilidad, entre 15:00 y 21:00 horas aproximadamente, esta baja notablemente en comparación con la curva media de porcentaje de humedad relativa funcional. Un aspecto importante es que como se identificó anteriormente la gran mayoría de curvas atípicas se encontraban en la parte superior o con porcentajes de Humedad relativa mayores, haciendo que la curva media fuera halada hacia arriba, esto es corregido y ajustado con la curva media funcional robusta ya que, como se observa en todas las horas, esta curva está por debajo de la curva media de porcentaje de humedad relativa funcional de esta.

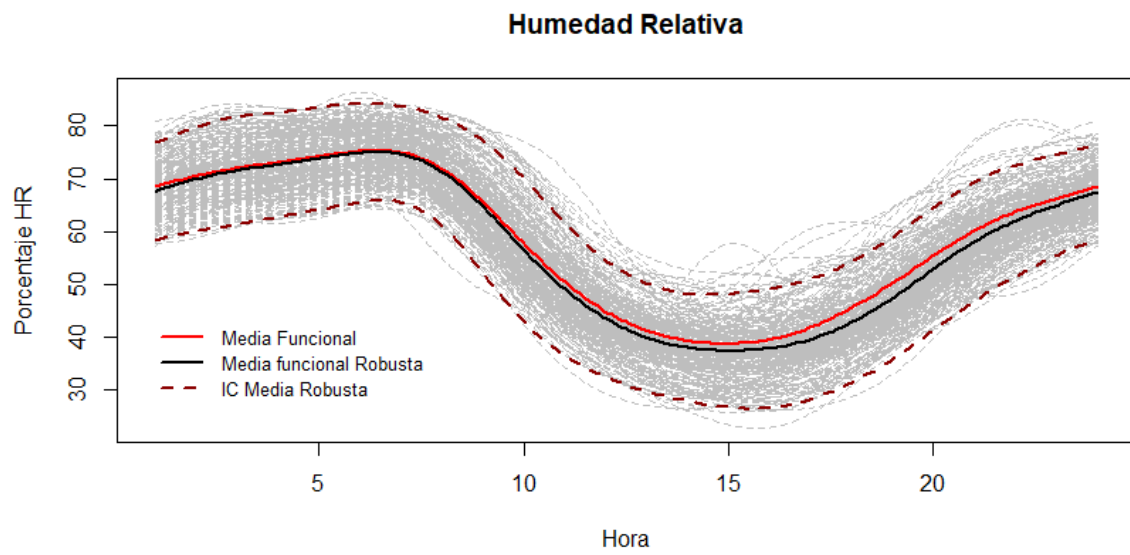


Figura 4-12.: Curva media funcional (Roja), Curva media funcional robusta (Negra) del porcentaje de humedad relativa y su intervalo de confianza

5. Conclusiones y recomendaciones

5.1. Conclusiones

En el presente trabajo de grado se buscaba un estimador de la media, en donde se mitigara el impacto de las observaciones atípicas presentes en la muestra, de esta manera caracterizar el comportamiento durante el día del porcentaje de humedad relativa, utilizando datos recolectados en la estación de monitoreo de calidad del aire ubicada en el barrio compartir, al oriente de la ciudad Santiago de Cali, para el año 2016. Para ello se utilizó la teoría de datos funcionales y varias herramientas de análisis de su conjunto, además, se extendió el concepto de robustez en datos escalares para obtener una estimación confiable de la curva media de porcentaje de humedad relativa para el periodo de estudio. Las curvas diarias de porcentaje de humedad relativa se reconstruyeron utilizando bases de Fourier de orden 15 y parámetro $\lambda = 0, 13$. De esta manera, después de realizar el análisis a través de las diferentes técnicas que se utilizaron para finalizar este estudio, a continuación se presentan las conclusiones obtenidas.

De esta manera, antes de resaltar los resultados y hallazgos de este estudio primero se dice que respecto al objetivo principal de este trabajo, con la metodología que se propuso se cumplió a totalidad. Haciendo uso de este conjunto de técnicas estadísticas para el análisis se obtuvo un estimador del porcentaje de humedad relativa utilizando datos recolectados en la estación de monitoreo de calidad del aire ubicada en el barrio compartir, al oriente de la ciudad Santiago de Cali. Para finalmente, con resultados confiables y acordes a la realidad, caracterizar el comportamiento de esta variable climatológica para el año 2016. Por otra parte, vale la pena destacar la ventaja de utilizar el análisis de datos funcionales con este tipo de datos, primero porque nos brinda mucha más información y herramientas para estudiar el comportamiento en variables meteorológicas medidas a través del tiempo y segundo, la posibilidad de cumplir la propiedad de orden para datos funcionales, recordando que con una de las técnicas usadas en este estudio en donde se le asocio una medida escalar a cada función (Medida de profundidad funcional) permitió ordenar el conjunto de curvas según que tan alejadas están del centro.

Uno de los principales rasgos encontrado en la muestra de curvas funcionales del porcentaje de humedad relativa promedio horaria, que para el periodo de estudio fue un total de 321 días, es la existencia de dos momentos del día, en la mañana y en la noche, en que el porcentaje de humedad relativa alcanza su máximo, este comportamiento responde a la relación inversa de la humedad relativa, es decir, se alcanzan estos puntos máximos debido a que esa hora la

temperatura disminuye. El otro rasgo de gran interés es la variabilidad, entre las 11:00am y 6:00pm aproximadamente, del conjunto de curvas que sobresalen por encima y por debajo debido a las anomalías que se presentan en el clima de la zona, otra característica en términos de variabilidad es que se observa que en este lapso de tiempo se presenta la mayor variación del día según la función de desviación estándar. Este comportamiento nos brindó la primer alerta de aquellas funciones que se debían estudiar para determinar el conjunto de curvas atípicas diarias del porcentaje de humedad relativa. También se analizó el comportamiento del porcentaje de humedad relativa promedio horaria de los meses del año, esto con el fin de explorar algún rasgo de relevancia pero como lo fue para el análisis del año completo, se observaron las mismas caracterizas a lo largo del día.

Para la muestra de curvas de porcentaje de humedad relativa promedio horario, se hizo el estudio de funciones atípicas por medio de los BoxPlot funcionales y medidas de profundidad funcional HMD, determinando un total de 53 curvas diarias atípicas y se observó que la mayoría de las curvas estaban por encima de la curva media de porcentaje de humedad relativa. Después de estudiar este hecho, se determinó que la curva media de porcentaje de humedad relativa estaba siendo afectada, halándola hacia arriba. A través del análisis de datos funcionales, medida de profundidad funcional HMD y el concepto de estimación robusta para datos escalares, se utilizó una técnica, producto de la extensión de los conceptos, para estimar la curva media de porcentaje de humedad relativa. Se determinó el 16 % como el porcentaje óptimo de recorte de curvas extremas, para mitigar el impacto del efecto de las curvas funcionales atípicas sobre la estimación de la curva media y una buena variabilidad para el conjunto de curvas restantes. Es decir, haciendo uso de estas técnicas en conjunto se obtuvo un estimador confiable de la curva media de porcentaje de humedad relativa promedio horario. Para el año 2016 el comportamiento de la curva media de porcentaje de humedad relativa promedio horario, deja en evidencia las dos horas del día en donde se registra el porcentaje de humedad relativa máximo, alrededor de las 7:00am y 11:00pm, además alrededor de las 3:00pm se registra el menor porcentaje de humedad relativa. A lo largo del día, este comportamiento responde a la inversa relación con la temperatura, variando con el aumento o disminución de esta.

5.2. Recomendaciones

En estudios posteriores se recomienda extender el concepto de robustez y el análisis de datos funcionales con otros estimadores como la mediana, mencionado en este documento, y la regresión, ya que en datos reales siempre hay presencia de valores atípicos. Además, estudiar el comportamiento de otras variables climatológicas que pueden brindar resultados de importancia para diferentes sectores.

Finalmente se recomienda realizar inferencias funcionales que permitan establecer si las diferencias encontradas en las estimaciones de la media funcional y la robusta son significativas, a su vez, evaluar la robustez del estimador frente a diferentes escenarios como por ejemplo la introducción (simular) de nuevas observaciones extremas que puedan generar ruido en la muestra y el estimador.

A. Anexo: BoxPlot Funcional del porcentaje de humedad relativa año 2016

Como se puede observar el BoxPlot funcional es mucho menos potente para identificar curvas atípicas que otros métodos como el de la profundidad funcional, en este caso solo 5 curvas fueron detectadas como atípicas.

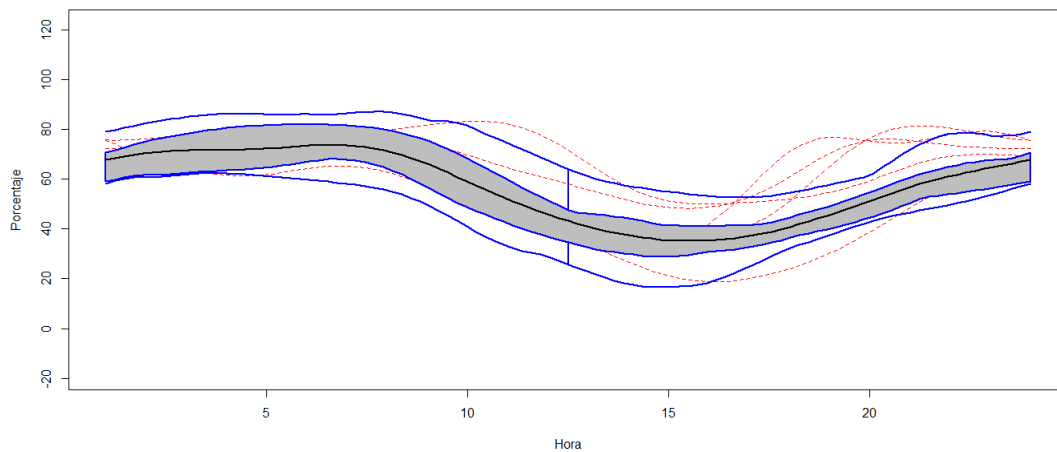
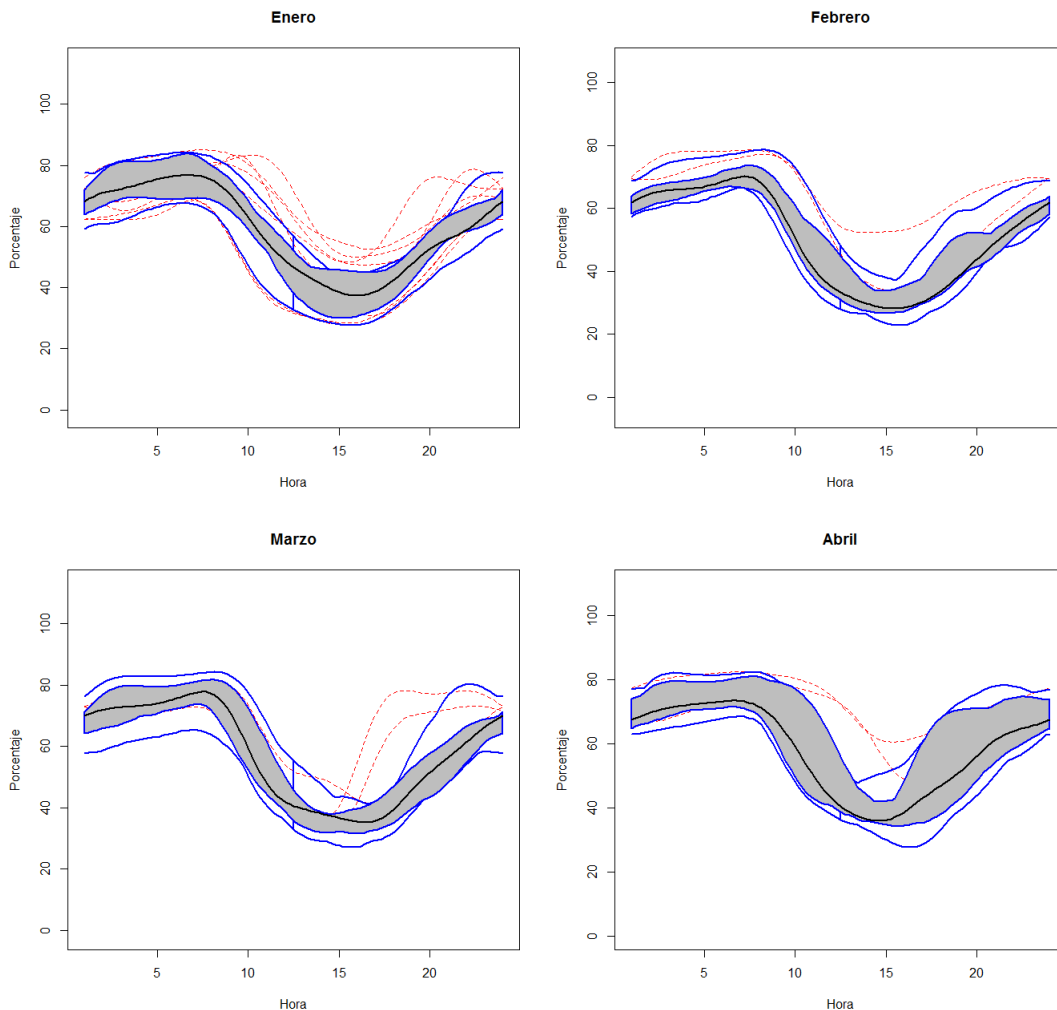


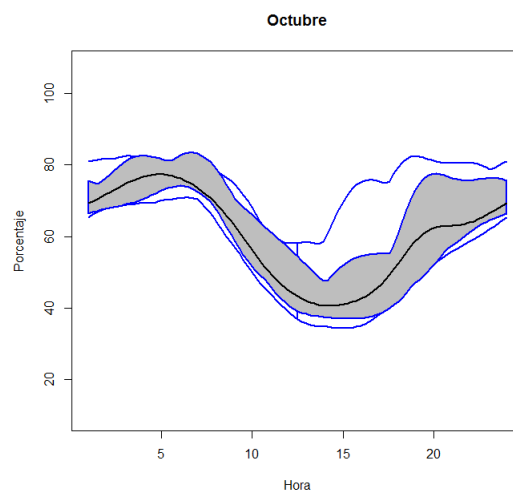
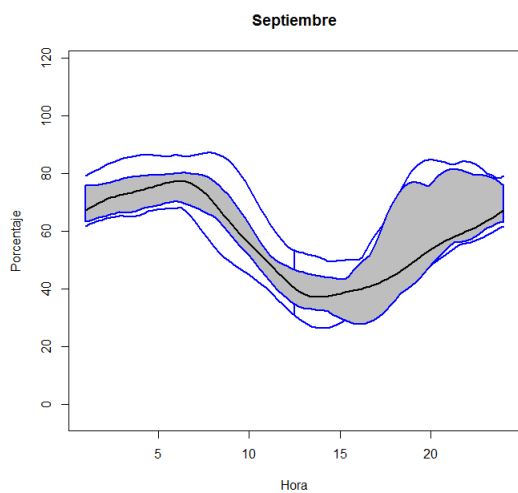
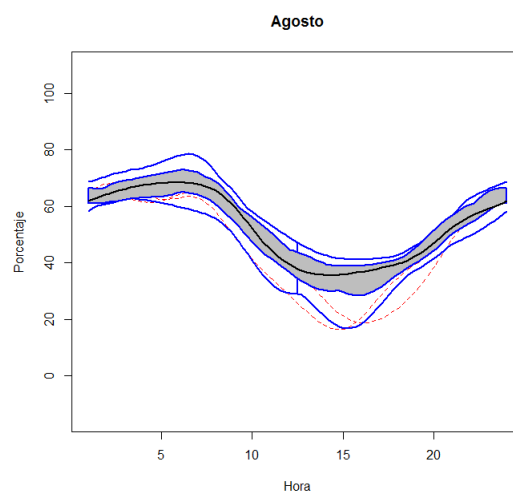
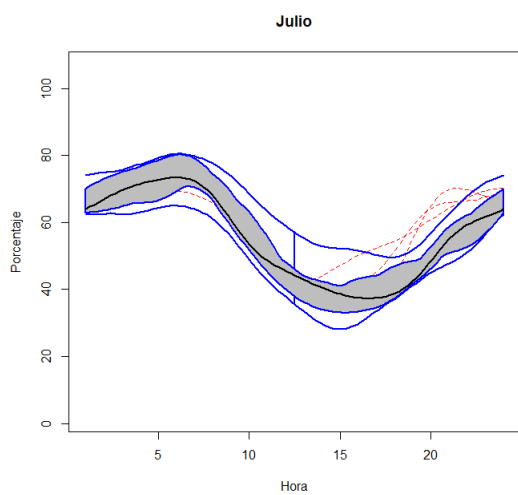
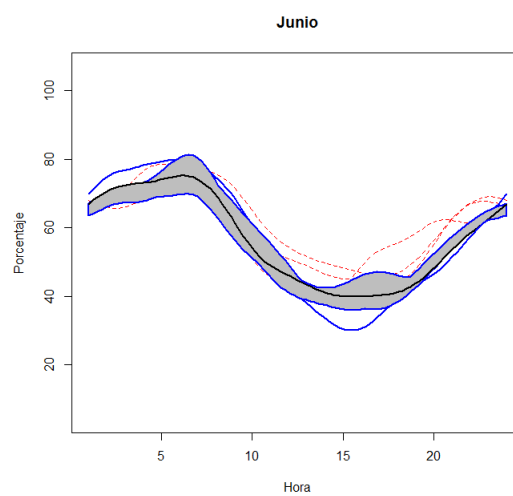
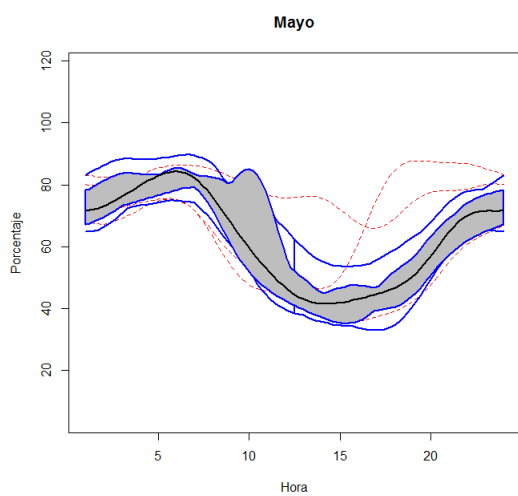
Figura A-1.: Box-Plot funcional del porcentaje de humedad relativa año 2016

En la Figura A-1 se muestra el BoxPlot para la muestra completa del año 2016 donde se puede observar que el comportamiento de la mediana funcional fluctúa alrededor del 60%, similar al comportamiento de la media funcional. Respecto a la identificación de curvas funcionales atípicas, esta técnica también está basada en la profundidad funcional (*Band Depth*) sin embargo los BoxPlot funcionales pasan por alto algunas curvas con comportamientos muy diferentes al de las demás, debido a que muchas de estas no sobrepasan en ningún momento la región comprendida por 1,5 veces de la región central. A pesar de esto, se puede observar que el comportamiento no es muy variado para la muestra de datos funcionales del año 2016 y la detección de 5 curvas o funciones atípicas, que corresponden a diferencias de porcentaje y forma de las funciones.

B. Anexo: BoxPlot Funcional del porcentaje de humedad relativa para cada uno de los meses del año 2016

Para cada uno de los meses del año 2016 el análisis del BoxPlot funcional es implícito, es decir, debido al interés de este trabajo solo se analizó la figura anual (A-1) y para cada mes del año el análisis es similar.





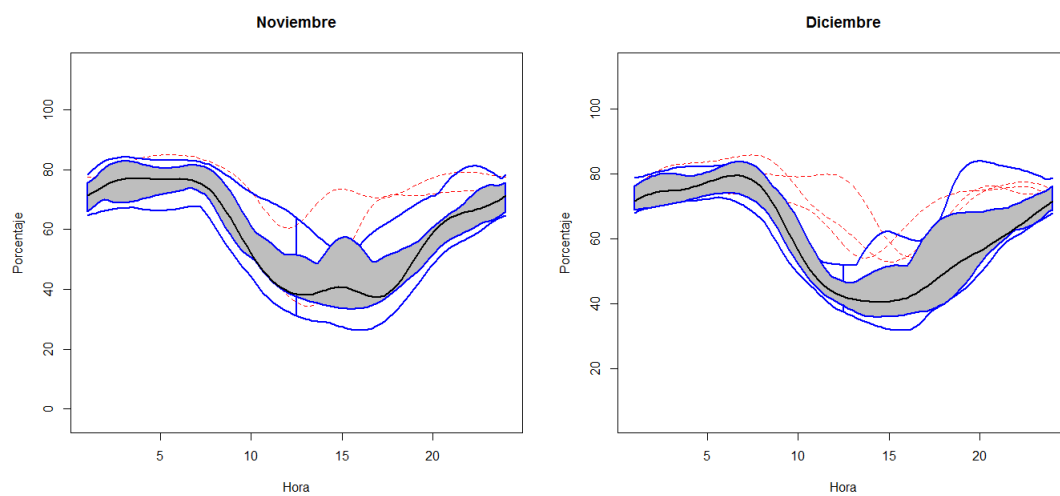


Figura B-1.: Box-Plots funcionales por meses del año 2016

Bibliografía

- Alcaldía (2004). Datos de cali y el valle del cauca. http://www.cali.gov.co/publicaciones/datos_de_cali_y_el_valle_del_cauca_pub. Accessed: 2016-11-16.
- Berruezo, J. A. y Jiménez, J. D. (2017). Situación del convenio marco de naciones unidas sobre el cambio climático. resumen de las cumbres de parís, cop21 y de marrakech, cop22. *Revista de Salud Ambiental*, 17(1):34–39.
- Billor, N. (2016). Outlier detection. En *Simposio Internacional de Estadística, Combinatorics, Mathematics and Applications*.
- Boente, G. y Vahnovan, A. (2016). Robust estimators in semi-functional partial linear regression models. *Journal of Multivariate Analysis*.
- Cohen, J., Cohen, P., West, S. G., y Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge.
- Crambes, C., Delsol, L., y Laksaci, A. (2008). Robust nonparametric estimation for functional data. *Journal of Nonparametric Statistics*, 20(7):573–598.
- Craven, P. y Wahba, G. (1979). Smoothing noisy data with splin functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31:377–403.
- Cuevas, A., Febrero, M., y Fraiman, R. (2006). On the use of the bootstrap for estimating functions with functional data. *Computational statistics & data analysis*, 51(2):1063–1074.
- Douglas, C. (2006). Física. principios con aplicaciones.
- Febrero-Bande, M. y Oviedo de la Fuente, M. (2012). Statistical computing in functional data analysis: The R package fda.usc. *Journal of Statistical Software*, 51(4):1–28.
- Ferraty, F. y Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media.
- Figueras, S. y Gargallo, P. (2003). Análisis exploratorio de datos,[en línea] 5campus. com. *Estadística*,<http://www.5campus.com/leccion/aed>.
- Fraiman, R. y Muniz, G. (2001). Trimmed means for functional data. *Test*, 10(2):419–440.

- Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, 11(1):1–21.
- Hyndman, R. J. y Ullah, M. S. (2007). Robust forecasting of mortality and fertility rates: a functional data approach. *Computational Statistics & Data Analysis*, 51(10):4942–4956.
- IPCC (2013). Cambio climático 2013. bases físicas. contribución del grupo de trabajo i al quinto informe de evaluación del grupo intergubernamental de expertos sobre el cambio climático. *Cambridge University Press*.
- Kendall, M. G. y Buckland, W. R. (1960). *A dictionary of statistical terms: prepared for the International Statistical Institute with the assistance of the UNESCO*.
- Liu, Y. (2014). *Outliers*, pp. 4542–4546. Springer Netherlands, Dordrecht.
- López-Pintado, S. y Romo, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association*, 104(486):718–734.
- Martínez, M. F., Osnaya, P., y Bremauntz, A. F. (2004). *Cambio climático: una visión desde México*. Instituto Nacional de Ecología.
- Navarro Pérez, V. (2004). Análisis de datos funcionales: implementación y aplicaciones.
- Nave, C. R. (2005). Hyperphysics. *Report, ECE dept. Georgia State University*.
- Olaya, J. (2012). Modelos aditivos generalizados gam: Aplicación al estudio de la contaminación del aire por ozono. *V Escuela de Verano*.
- Ramsay, J. O. y Silverman, B. (2005). *Functional data analysis*. Springer Series in Statistics, second edición.
- Ramsay, J. O., Wickham, H., Graves, S., y Hooker, G. (2017). *fda: Functional Data Analysis*. R package version 2.4.7.
- Rousseeuw, P. J. (1990). Robust estimation and identifying outliers.
- Sawant, P., Billor, N., y Shin, H. (2012). Functional outlier detection with robust functional principal component analysis. *Computational Statistics*, 27(1):83–102.
- Sun, Y. y Genton, M. G. (2011). Functional boxplots. *Journal of Computational and Graphical Statistics*, 20(2):316–334.
- Tambutti, R. y Muñoz, H. (2002). *Introducción a la física ya la química 1*. Editorial Limusa.
- Torres, J. M., Nieto, P. G., Alejano, L., y Reyes, A. (2011). Detection of outliers in gas emissions from urban areas using functional data analysis. *Journal of hazardous materials*, 186(1):144–149.

Yunus, A. (2006). Cengel. *MICHAEL A., Boles. Termodinámica. Sexta Edición. Editorial Mc Graw Hill.*

Zamar, R. H. (1994). Estimación robusta. *Estadística española*, 36(137):327–387.