

Evaluación del desempeño en métodos clásicos y datos funcionales para la estimación de datos faltantes con mediciones de Ozono en la ciudad de Santiago de Cali

Lina Marcela Henao Peña
Sergio Andrés Silva Guamanga

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

Evaluación del desempeño en métodos clásicos y datos funcionales para la estimación de datos faltantes con mediciones de Ozono en la ciudad de Santiago de Cali

Lina Marcela Henao Peña
Sergio Andrés Silva Guamanga

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

Director(a):
Johann A. Ospina, MSc

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

Dedicatoria

Dedicado a mi madre, mujer valiente y de fuerza inquebrantable, gracias por llenar mi mundo de sueños y metas, por convertirme en tu prioridad y nunca dejarme caer en los momentos difíciles, eres quien llena mi vida de colores, mi mayor inspiración y ejemplo. Para mi padre, mi guerrero, protector y confidente, gracias por recordarme día a día lo orgulloso que estabas de mi, gracias por enseñarme lo fuerte y capaz que puedo llegar a ser. Fuiste, eres y serás el amor de mi vida, mi mejor amigo, mi gran bendición. Finalmente a mi hermosa familia, quienes vierten en mi corazón alegría, amor y apoyo incondicional.

Lina Marcela Henao Peña

A Dios por brindarme salud, vida y la facultad para culminar este proyecto. A mi madre Trinidad y mi madrina Sonia, por confiar en mí y brindarme todo su apoyo incondicional en todos estos años de carrera; sin ellas no podría haber llegado hasta aquí. Por ultimo quiero dar las gracias a todas aquellas personas que durante estos años me han hecho crecer tanto en lo académico como en lo personal.

Sergio Andrés Silva Guamanga

Agradecimientos

Agradecemos a la Universidad del Valle por convertirse en nuestra alma máter, por abrirnos las puertas al conocimiento y al ser.

A la escuela de estadística y sus docentes, quienes guiaron nuestro aprendizaje y formación como profesionales y personas. A Víctor Manuel González Rojas, Rafael Antonio Klinger Angarita y Rodrigo Rada Barona por los consejos y risas que nos regalaron en el día a día, gracias por toda la dedicación y la paciencia a lo largo de estos años.

A nuestro profesor y director de trabajo de grado Johann Alexis Ospina, quien construyó junto a nosotros la presente investigación, con el mayor esmero e interés, siendo nuestro principal apoyo y guía en este proceso. Por último, agradecemos a todos nuestros compañeros porque son ustedes quienes enriquecieron nuestro aprendizaje y nos brindaron momentos inolvidables a lo largo de toda la carrera.

Resumen

En este estudio, se introducen varios métodos de imputación que han sido diseñados con base a las técnicas de análisis de datos funcionales (ADF) y se investiga la capacidad de los métodos para estimar valores faltantes óptimos, es decir, que cuenten con el mínimo error posible. Los métodos de imputación única y los métodos de imputación iterativa se realizan mediante la estimación de curvas a través de métodos de regresión y de suavizado. Por otra parte, se realizan estimaciones por medio de métodos clásicos comúnmente utilizados en investigaciones científicas por su fácil aplicabilidad y entendimiento, como lo son los métodos de medias y de regresión. El rendimiento de los métodos basados en datos funcionales y clásico, se comparan mediante un conjunto de datos reales de Ozono troposférico (datos de referencia), obtenidos de la estación Compartir, localizada en la parte oriental de la ciudad de Santiago de Cali. Mil conjuntos de datos faltantes se generan a partir de los datos de referencia con seis patrones diferentes de valores faltantes, los cuales se desarrollan con el objetivo de investigar el rendimiento de los métodos considerados. Los patrones se simulan de acuerdo con tres porcentajes (5, 10 y 15) de valores faltantes con respecto a dos tamaños diferentes (3 y 7) de longitudes de espacio máximos (puntos perdidos consecutivos). La decisión conjunta de los indicadores evaluados muestran que el método de imputación iterativa es más flexible y preciso a los otros métodos evaluados para altos porcentajes de valores aleatorios faltantes y longitudes máximas, logrando optimizar el proceso de imputación.

Palabras claves: Contaminantes criterio, Ozono troposférico, Datos faltantes, Datos funcionales (ADF), Imputación de datos faltantes.

Abstract

The use of a database with missing functions reduces the robustness in the terms of efficiency of the results and increases the possibility of poor specification in estimated models, consequences that lead to inaccurate and imprecise decisions. In this study, several imputation methods are presented that have been designed based on the functional data analysis techniques and the ability of the methods to estimate optimal missing values is studied, that is, they have the minimum possible error. The single imputation methods and the iterative imputation methods can be performed by estimating curves through regression and smoothing methods. On the other hand, the media was selected and used in scientific research for its easy application and understanding, such as media and regression methods. The performance of the functional and classical field methods is compared through a set of real tropospheric Ozone data (reference data), obtained from the station Share, located in the eastern part of the city of Santiago de Cali. The missing data sets have been generated from the reference data with six different missing value patterns, which have been developed with the aim of investigating the performance of the considered methods. The patterns are simulated according to three percentages (5, 10 and 15) of missing values with respect to two different sizes (3 and 7) of maximum space lengths (consecutive lost points). The joint decision of the evaluated indicators showed that the iterative imputation method is more flexible and accurate than other methods for high percentages of missing random values and maximum lengths, thus optimizing the imputation process.

Keywords: Criteria pollutants, Tropospheric ozone, Missing data, Functional data (ADF), Imputation of missing data.

Contenido

Agradecimientos	7
Resumen	9
Abstract	9
1. Introducción	2
2. Planteamiento del problema	4
3. Justificación	6
4. Antecedentes	7
5. Objetivos	10
5.1. Objetivo General	10
5.2. Objetivos Específicos	10
6. Marco Teórico	11
6.1. Contaminantes Criterio	11
6.2. Contaminantes criterio medidos en Cali	11
6.2.1. Ozono Troposférico	11
6.3. Datos faltantes	13
6.3.1. Patrones generadores de datos faltantes	13
6.3.2. Mecanismos generadores de datos faltantes	14
6.4. Imputación de datos faltantes	14
6.5. Eliminación de casos - No imputación	15
6.6. Imputación simple	16
6.6.1. Imputación por medias	16
6.6.2. Imputación por regresión	17
6.7. Construcción de datos funcionales	18
6.7.1. Base de funciones	18
6.7.2. Construcción de datos funcionales	19
6.7.3. Número óptimo de funciones de la base (K)	21
6.7.4. Estadística descriptiva funcional	22
6.8. Métodos de Imputación basados en técnicas de conversión de datos en la ADF	23
6.8.1. Método de datos de medias funcionales (FMeanDM) y Método de datos medianos funcionales (FMedDM)	23

6.8.2.	Método iterativo basado en la media funcional (FMeanBIM) y método iterativo basado en la mediana funcional (FMedBIM)	24
6.8.3.	Indicadores de desempeño	24
7.	Metodología	25
7.1.	Datos	25
7.2.	Verificación y elección de la base de datos	26
7.3.	Construcción de escenarios	27
7.4.	Métodos de imputación	27
7.5.	Comparación de métodos de imputación	30
8.	Resultados	31
8.1.	Análisis exploratorio de datos	31
8.1.1.	Estimación de dato faltante por regresión lineal	34
8.2.	Construcción de escenarios	37
8.3.	Construcción de datos funcionales	39
8.4.	Análisis descriptivo de datos funcionales	42
8.5.	Imputación de datos faltantes	43
8.5.1.	Métodos clásicos	43
8.5.2.	Métodos basados en datos funcionales	46
8.5.3.	Imputación media funcional y mediana funcional	46
8.5.4.	Método iterativo basado en la media funcional (FMeanBIM) y método iterativo basado en la mediana funcional (FMedBIM)	47
8.5.5.	Comparación de métodos de imputación	48
9.	Conclusiones y Recomendaciones	51
9.0.1.	Conclusiones	51
9.0.2.	Recomendaciones	52
	Anexos	53
	Bibliografía	60

Lista de Tablas

6.1. Estaciones de Monitoreo en Cali	11
7.1. Estructura datos	26
7.2. Patrones diferentes de datos generados (Datos Faltantes)	27
8.1. Distribución de datos faltantes por hora de acuerdo a los meses para el año 2015	34
8.2. Distribución de datos faltantes por hora de acuerdo a los meses para el año 2016	34
8.3. Análisis de varianza para la hora 11 para el mes de Enero 2016	35
8.4. Estimaciones de datos faltantes para la hora 11:00 del mes de Enero en el año 2016	35
8.5. Datos contaminados P15-G7 en el mes de Enero	44
8.6. Número de iteraciones que aseguran la convergencia del método	48
8.7. Rendimiento de los siete métodos de imputación	49
8.8. Estimaciones puntuales obtenidas por cada método	50
.1. Distribución de datos faltantes por hora de acuerdo a los meses para el año 2015	53
.2. Distribución de datos faltantes por hora de acuerdo a las horas para el año 2015	53
.3. Distribución de datos faltantes por hora de acuerdo a los meses para el año 2016	53
.4. Distribución de datos faltantes por hora de acuerdo a las horas para el año 2015	54
.5. Matriz $\phi(t)$	55
.6. Matriz de coeficientes ($\mathbf{C}$)	55
.7. Curva estimada $X(t)$	56
.8. Estimaciones obtenidas por los métodos evaluados para el primero de Enero del 2016	57
.9. Estimaciones obtenidas por los métodos evaluados para el siete de Enero del 2016	57
.10. Estimaciones obtenidas por los métodos evaluados para el veintitrés de Enero del 2016	57
.11. Estimaciones obtenidas por los métodos evaluados para el treinta y uno de Enero del 2016	58

Lista de Figuras

6.1. Ciclo simplificado del Ozono y procesos relacionados.	12
6.2. Patrones de datos ausentes fuente propia	13
6.3.	23
7.1. Ubicación espacial estaciones de Monitoreo	25
8.1. Mediciones de O_3 por hora	31
8.2. Boxplots de concentraciones horarias de O_3 para el año 2015 y 2016 en presencia de datos faltantes	32
8.3. Serie para el Máximo Promedio Diario para el año 2015	33
8.4. Serie para el Máximo Promedio Diario para el año 2016	33
8.5. Comportamiento promedio horario de O_3 de acuerdo a los días de la semana - Enero 2016 . .	36
8.6. Comportamiento promedio diario de O_3 para el mes de Enero	36
8.7. Contaminación del 5 % y 5 datos faltantes consecutivos	37
8.8. Contaminación del 5 % y 7 datos faltantes consecutivos	37
8.9. Contaminación del 5 % y 5 datos faltantes consecutivos	37
8.10. Contaminación del 5 % y 7 datos faltantes consecutivos	37
8.11. Contaminación del 5 % y 5 datos faltantes consecutivos	38
8.12. Contaminación del 5 % y 7 datos faltantes consecutivos	38
8.13. Validación cruzada generalizada para parámetros K y λ	39
8.14. Base B-spline de orden cuatro con sus respectivos nodos	40
8.15. Construcción de datos funcionales para mediciones de O_3	40
8.16. Primera y segunda derivada de curvas funcionales de O_3	41
8.17. Datos funcionales de O_3 con curva media y curva de desviación estándar	42
8.18. Mapa de calor para la correlación horaria	43
8.19. Mapa de calor de la matriz de correlación	45
8.20. Ejemplo intuitivo técnica de imputación media funcional y mediana funcional	46
8.21. Ejemplo intuitivo técnica de imputación media funcional y mediana funcional iterativa . . .	47
8.22. Estimaciones puntuales obtenidas por cada método para el día 24.	50
.1. Estructura de los datos para construcción del modelo	54
.2. Estimaciones puntuales obtenidas para el primer día de Enero	58
.3. Estimaciones puntuales obtenidas para el primer día de Enero	58
.4. Estimaciones puntuales obtenidas para el primer día de Enero	59
.5. Estimaciones puntuales obtenidas para el primer día de Enero	59

Capítulo 1

Introducción

De acuerdo con los análisis realizados por el Instituto de Hidrología, Meteorología y Estudios Ambientales IDEAM, la contaminación atmosférica en Colombia es uno de los problemas ambientales de mayor preocupación para las entidades ambientales, el gobierno y la población en general ya que, es el tercer factor con mayor generación de costos sociales después de la contaminación del agua y los desastres naturales. De acuerdo con lo establecido en el Título 5 del Decreto 1076 de 2015 y ajustando la definición, la contaminación atmosférica es el fenómeno de acumulación o de concentración de contaminantes, entendidos estos como fenómenos físicos, sustancias y/o elementos en estado sólido, líquido o gaseoso, causantes de efectos adversos en el medio ambiente, los recursos naturales renovables y la salud humana que solos, en combinación o como productos de reacción, se emiten al aire como resultado de actividades humanas (MinAmbiental, 2016a).

Entre los principales contaminantes se encuentra el ozono troposférico (O_3) y en particular, el ozono superficial, el cual es el principal contaminante fotoquímico y se origina principalmente en las áreas urbanas por varias fuentes de emisión, como los automóviles y la industria. La contaminación por ozono, es un problema diurno durante los días soleados, a causa de que la luz solar desempeña un papel primordial en su formación.

En altas concentraciones, el ozono superficial es causante de diversos problemas, ya que al ser un gas oxidante, reacciona fuertemente para destruir o alterar otras moléculas, actuando como un contaminante tóxico para la salud humana, produciendo daños respiratorios y pulmonares, dolores de pecho, irritación de la garganta y ojos, empeoramiento de enfermedades preexistentes del corazón, ataques de tos, entre otros síntomas. Por otra parte, este contaminante tiene efectos nocivos sobre la vegetación, ya que deteriora las hojas de los árboles y plantas, así como reduce el rendimiento de los cultivos y el crecimiento de los bosques según lo establecido por Ministerio del Medio Ambiente (MinAmbiental, 2016a).

Es por esto que surge la necesidad de controlar y monitorear los niveles de este contaminante. En la actualidad, las mediciones realizadas a variables de calidad del aire como lo es el O_3 , se realizan de forma permanente en el tiempo para ejecutar controles y seguimientos a las características que alteran el equilibrio ambiental. Sin embargo, existen interrupciones que impiden contar con un conjunto completo de datos, lo que conlleva al deterioro de cualquier tipo de análisis, resultado y decisión que se pretenda tomar. Es por esta razón que, el estudio y desarrollo de técnicas o metodologías que permitan estimar registros faltantes se han visto fortalecidos, estructurando el concepto de preparación de datos.

Es común en el campo investigativo trabajar con metodologías que permiten de forma fácil y efectiva encontrar un valor que reemplace la ausencia del registro. Dos enfoques generales para resolver este problema son la eliminación de casos y la imputación por métodos clásicos, donde la primera es la más común pero cuenta con la desventaja de perder información debido a la reducción de datos. No obstante, la continua

evolución y desenvolvimiento de la estadística en la actualidad, ha incorporado técnicas recientes para imputar valores faltantes. Éste es el caso de las técnicas de imputación basadas en el Análisis de Datos Funcionales (ADF), siendo ésta una rama de la estadística que considera un conjunto de mediciones a lo largo de una variable continua medida de forma discreta. En el ADF, los datos corresponden a un conjunto de curvas, donde la unidad elemental de información es la función completa según Ramsay and Silverman (2005).

Para la presente investigación, se desarrolla el análisis para un set de datos de O_3 , especificando de esta forma la variable de estudio, la cual es medida de forma discreta en un dominio continuo definido como el tiempo. Se pretende evaluar el problema de datos faltantes a partir de la metodología ADF junto a metodologías clásicas, que permitan generar comparaciones de interés y vislumbrar el potencial real de aplicar métodos basados en ADF.

Capítulo 2

Planteamiento del problema

La disponibilidad de un conjunto completo de datos es esencial para el desarrollo de un buen análisis estadístico que permita obtener resultados confiables en función de la calidad de los datos utilizados. Los datos faltantes son comunes cuando la información es generada de manera constante en el tiempo y/o cuando se presentan fallas en los instrumentos de medición con lo cual, se reporta la pérdida del dato y de forma general se dispone de menos observaciones a las esperadas (Yang Lenyap et al., 2015). Estas características son usuales en mediciones de calidad del aire, siendo esta un indicador de cuanto aire esté exento de polución atmosférica y apto para ser respirado, determinado por la presencia en la atmósfera de compuestos químicos. Según él Etze and French (2000) “El desarrollo de las actividades humanas producen emisiones de gases y partículas que han provocado y siguen provocando alteraciones en el complejo equilibrio que mantiene las propiedades y funciones de la atmósfera”.

Para regular la calidad del aire y evitar catástrofes ambientales existe una normatividad específica, la cual establece los límites de concentraciones permisibles de cada contaminante criterio a lo largo de cierto periodo de tiempo. El monitoreo de estos contaminantes tiene como principal objetivo evaluar la calidad del aire y verificar que los niveles especificados en la norma se cumplen, entendiendo por contaminantes criterios aquellos compuestos químicos del aire que tienen mayor impacto en la salud humana, donde puntualmente se han identificado 5 contaminantes que afectan a la salud inmediatamente desde su inhalación, a saber, Dióxido de Azufre (SO_2), Dióxido de Nitrógeno (NO_2), Partículas suspendidas (PM_{10} y $PM_{2.5}$), Monóxido de Carbono (CO) y Ozono Troposférico (O_3) (Ataz and de Mera Morales, 2004).

Es por esto que, el Sistema de Vigilancia de la Calidad del Aire de Santiago de Cali (SVCASC), el cual opera bajo la coordinación y administración del Grupo de Calidad del Aire del Departamento Administrativo de Gestión del Medio Ambiente (DAGMA), mide continuamente los contaminantes criterio en diferentes puntos de la ciudad de Santiago de Cali, a través de concentraciones medidas en promedios por hora, esperando así disponer de 24 valores diarios. Sin embargo, por diversas razones como cortes de fluido eléctrico, fallas en los equipos de medición, errores de comunicación, malas condiciones alrededor de la estación y dificultades con la fiabilidad de las redes de monitoreo, entre otras, es común encontrar días en los cuales se dispone de observaciones faltantes, incluyendo días enteros sin observaciones (Jaramillo et al., 2004).

Los datos obtenidos de los contaminantes, dadas las estaciones de observación requieren una análisis exploratorio y descriptivo antes de su uso en estudios de investigación, contemplando una metodología para evitar en mayor medida resultados propios de modelos y análisis estadísticos deficientes por la utilización de registros incompletos. Esta preocupación se ha presentado en muchos campos de estudios (Baraldi and Enders, 2010), incluyendo la investigación medioambiental (Junninen et al., 2004) y es por esto que la estimación de datos faltantes se convierte en un tema de suma importancia dentro del campo estadístico.

Varios métodos de imputación se han utilizado para abordar el problema y están profundamente discutidos en la literatura, dentro de los cuales se encuentran: métodos de eliminación del caso con valores perdidos, sustitución/imputación de éstos por valores estimados como la media de la variable, métodos basados en la predicción obtenida mediante regresión sobre las demás variables, entre otros. Este tipo de métodos clásicos suelen ser prácticos de implementar y adecuados en situaciones concretas, sin embargo presentan problemas inferenciales asociados a la ineficiencia en la estimación, aparición de sesgos, distorsión de la estructura de covarianzas; además de no incorporar la incertidumbre asociada a los datos faltantes (Medina Rivera, 2008).

En lugar de considerar los datos medidos de dicho contaminante como valores discretos, es decir, registros obtenidos en una unidad de tiempo puntual, los datos pueden ser tratados como una curva de longitud finita en un período de tiempo, dentro de este contexto aparece el Análisis de Datos Funcionales (ADF) (Ramsay and Silverman, 2005). La aplicación del ADF para el tratamiento de valores faltantes es una metodología en desarrollo (Yang Lenyap et al., 2015). Por lo tanto, es de interés evaluar este enfoque comparándolo con los métodos clásicos en diferentes escenarios: porcentajes de datos faltantes aleatorios (5 %, 10 % y 15 %) y espacios consecutivos faltantes (longitud de espacio faltante de 3 y 7 unidades horarias).

Este proyecto busca soluciones a este problema por medio de la aplicación del ADF para el tratamiento de datos faltantes del contaminante criterio O_3 , basado en una metodología estadística reciente, que permita contribuir a la solución del problema de datos faltantes en este tipo de información (Yang Lenyap et al., 2015). A partir de esto se plantean inicialmente la siguiente pregunta de investigación:

¿El análisis de Datos Funcionales presenta mejores estimaciones para los valores faltantes en comparación con los métodos clásicos, bajo diferentes escenarios de porcentajes de datos faltantes y espacios consecutivos faltantes?

Capítulo 3

Justificación

Actualmente, el O_3 está considerado como uno de los principales contaminantes atmosféricos (contaminante secundario), debido a su amplia distribución geográfica y especialmente a los efectos que provoca sobre la salud humana, los materiales, la vegetación, la biodiversidad y el funcionamiento de los ecosistemas. Además, es un gas de efecto invernadero que contribuye al calentamiento global y está directamente implicado en el cambio climático (Caballero et al., 2007).

Dicho contaminante no ejerce la función protectora que cumple en la estratosfera sino que por el contrario, dada su alta reactividad y su fuerte capacidad oxidante, cuando sus niveles se elevan puede provocar efectos adversos en los materiales, esto derivado de sus efectos corrosivos. Por otra parte, la salud humana puede verse deteriorada por los efectos que conlleva la inhalación del contaminante O_3 en las vías respiratorias, sin olvidar el deterioro constante que afecta la vegetación y los ecosistemas (Bermejo et al., 2010). Entidades y grupos de investigación como la Unidad Regional de Salud, el Ministerio de Salud y la Organización Panamericana de la Salud reconocen a este contaminante como un problema a tratar y ejercer control sobre sus efectos, problema que no solo compete a los grandes generadores de gases contaminantes (Fabricantes de automóviles, Industrias, etc) sino a toda la población en general.

Los datos de emisiones de O_3 , se usan para determinar el cumplimiento de los reglamentos de contaminación del aire, la eficacia del control, la eficiencia de producción, apoyar la investigación científica, diagnosticar las condiciones de un área antes de construir una nueva fuente de contaminación y para desarrollar modelos de dispersión de contaminantes, entre otros. La institución responsable del monitoreo de la calidad del aire SVCASC, designa los métodos de referencia para el muestreo, análisis, calibración de instrumentos y cálculo de las emisiones (Conesa Fernández et al., 1995). Sin embargo, las estaciones que aportan el insumo para el análisis correspondiente en ocasiones presentan daños, irregularidades, fallas o dejan de funcionar, situaciones que disminuyen la calidad de la información perdiendo validez, omitiendo información relevante para la investigación, alterando los resultados y conclusiones de los estudios. (Dumedah and Coulibaly, 2011).

De esta forma, es pertinente utilizar una metodología adecuada para dar solución a dicha problemática en las mediciones de O_3 diarias. En la presente investigación, se utiliza el Análisis de Datos Funcionales (ADF) para realizar la estimación de información faltante considerada en seis escenarios distintos y se genera una comparación con técnicas clásicas mediante indicadores de rendimiento, facilitando la generación de un panorama que optimice la utilización de métodos de imputación, dado el caso de porcentaje y longitudes para registros faltantes presentes en la base datos.

Por lo anterior, se pretende brindar un insumo metodológico óptimo para el análisis y la exploración de los niveles de O_3 , que cuenten con propiedades de mayor precisión y permitan la toma de decisiones correctivas. De hecho, se espera que este proyecto aporte a la gestión y monitoreo ambiental, partiendo de una visión más clara sobre el tratamiento de la información faltante.

Capítulo 4

Antecedentes

A continuación se presenta una revisión preliminar de antecedentes con respecto a la estimación de datos faltantes en mediciones de variables climáticas o investigaciones afines.

Yang Lenyap et al. (2015) en su investigación consideran datos reales del contaminante criterio PM_{10} extraído de una estación de monitoreo para la medición calidad del aire llamada Petaling Jave, localizada en la parte occidental de la península de Malasia. El conjunto de datos se encuentra conformado por mediciones del material particulado en 153 días por 24 horas registrados durante la temporada de monzones del suroeste en el año 2010. El objetivo de los autores era investigar la capacidad de los métodos de imputación única y métodos de imputación iterativa para estimar el valor perdido. Varios de los métodos de imputación se diseñaron basados en las técnicas de Análisis de Datos Funcionales (ADF), con la finalidad de comparar los resultados. Los datos faltantes se generan a partir del conjunto de datos de referencia (contaminante criterio PM_{10}) utilizando seis patrones de información perdida simulada aleatoriamente en diferentes condiciones de complejidad, con porcentajes totales del 5 %, 10 % y 15 % , de acuerdo con el número máximo de valores perdidos consecutivos por día, es decir, con 3 y 7 datos faltantes. Por último, se evalúa el desempeño de los métodos de imputación a través de tres indicadores, a saber, el coeficiente de determinación, índice de concordancia y el error absoluto medio. Los resultados mostraron que los tres mejores métodos son del enfoque con imputación iterativa. El método iterativo con el enfoque de la penalización de rugosidad presentó mayor rendimiento; mientras que la imputación única se caracterizó por presentar el menor. Todos los métodos son independientes del porcentaje faltante pero dependen del tamaño de la longitud del espacio faltante consecutivo. Por otro lado, se encontró que la imputación utilizando el método iterativo produjo una mejora en el rendimiento (precisión de la técnica para estimar el valor real) en comparación con el método de imputación no iterativa.

He et al. (2011) trabajaron con mediciones longitudinales de ingreso familiar en diferentes períodos de tiempo junto al estado de salud de los niños al interior de cada núcleo familiar como ejemplo práctico. Los autores proponen un enfoque funcional de imputación múltiple modelando los perfiles de respuesta longitudinal como curvas suaves de tiempo bajo un modelo de efectos mixtos funcionales desarrollando así, un algoritmo de muestreo de Gibbs para estimar los parámetros del modelo y las imputaciones para valores faltantes, basados en una técnica de bloqueo para una mayor eficiencia computacional. El análisis de imputación múltiple se limitó a una muestra de 2172 familias incluidas en la cohorte PSID (The Panel Study of Income Dynamics) de 1979 a 1997, y se aumentó con los datos de sus hijos recogidos en el CDS (The Child Development Supplement). Los resultados obtenidos por los autores son similares en todos los métodos. Las estimaciones de la diferencia de ingresos sugieren que los niños con mejor estado de salud son más propensos a estar en la familia de mayores ingresos. Además, los resultados del análisis de regresión logística implican que el estado de ingreso familiar en diferentes períodos de tiempo tiene efectos significativos y bastante similares

en el desarrollo de la salud de los niños. Los resultados del análisis de modelos mixtos cuantifican de forma adicional la diferencia en el crecimiento del ingreso para los dos grupos durante el período de estudio.

Junninen et al. (2004) consideraron los métodos de imputación de datos aplicados al conjunto de concentraciones de NOx , NO_2 , O_3 , PM_{10} , SO_2 y CO , clasificados como contaminantes criterio y medidos en una escala de tiempo de una hora (promedio por hora), junto con cuatro parámetros meteorológicos: velocidad del viento, dirección del viento, Temperatura y humedad relativa. La información fue obtenida de la base de datos Epidemias de la contaminación atmosférica: herramientas de modelado para mejorar la gestión de la contaminación atmosférica. Se utilizaron patrones de datos faltantes simulados aleatoriamente para evaluar los métodos en diferentes condiciones, con porcentajes de datos faltantes de 10 y 25. El principal objetivo del trabajo era comparar una serie de métodos disponibles para la imputación de los datos de calidad del aire: interpolaciones lineales, spline y vecindad más cercana, imputación basada en regresión, vecino más cercano multivariado, mapas autoorganizables y redes de retroproyección multicapa. Adicional a esto, se buscaba probar un esquema de imputación múltiple en el que se utilizaron métodos multivariados para estimar los datos faltantes. Los resultados mostraron que los métodos univariados dependen de la longitud de la brecha en el tiempo, al igual que el desempeño depende de la variable en estudio. Asimismo, los métodos multivariados tienen la ventaja sobre otros métodos ya que son menos dependientes respecto a la localización real de los valores perdidos. En general, se concluyó que es posible lograr un ligero aumento en los rendimientos mediante la combinación de los métodos multivariados; y que la forma en que debe realizarse esta combinación depende de la variable a inspeccionar. Por otra parte, se detectó que los métodos de imputación única subestiman la varianza de error de los datos faltantes y la precisión se puede mejorar sustancialmente mediante el uso de las imputaciones múltiples.

Norazian et al. (2008) trabajaron con datos de contaminación atmosférica obtenidos mediante máquinas automatizadas los cuales, con gran frecuencia contienen datos faltantes que pueden causar sesgo debido a diferencias sistemáticas entre los datos observados y los no observados. Trabajaron con datos de concentraciones de partículas PM_{10} en promedios horarios durante un año en las zonas de Seberang, Perai, Penang y Malasia. Del conjunto de concentraciones del contaminante criterio PM_{10} , se recrean al azar cinco escenarios con diferentes proporciones de valores perdidos, a saber, 5 %, 10 %, 15 %, 25 % y 40 %. En este análisis, se aplican seis técnicas de imputación única para estimar los valores faltantes. Cuatro de ellas fueron técnicas de interpolación (interpolación lineal, cuadrática, cubica y vecino más cercano) y las dos restantes fueron las técnicas de imputación media (la media antes y después y la media antes). Dado su objetivo de evaluar los métodos de imputación, se utilizaron cuatro indicadores de desempeño como lo son la precisión de predicción, el coeficiente de determinación, el error absoluto promedio y el error cuadrático medio. Los datos teóricos y observados se compararon para seleccionar el mejor método para estimar los valores faltantes. Dados los diferentes escenarios simulados dentro del margen del análisis, se producen resultados similares para todos los porcentajes de datos faltantes con los diferentes métodos o técnicas de estimación. Sin embargo, sobresale por presentar el menor error absoluto medio junto al menor error cuadrático medio, con valores altos en la precisión de la predicción y el coeficiente de determinación para cada escenario el método de media antes y después; seguido de la técnica de interpolación lineal, la cual presenta las mejores estimaciones para los valores faltantes para porcentajes inferiores al 15 % con base a la información de los indicadores de rendimiento. Ahora bien, entre las técnicas menos destacadas en este estudio, se acentúan las estimaciones arrojadas por el método de interpolación cuadrático y a su vez, por la técnica de media antes.

Caicedo and Jiménez (2016) realizaron un estudio que contempla la presencia de datos faltantes en un conjunto de concentraciones de $PM_{2.5}$, medidas de forma horaria y obtenidas a través de la estación de monitoreo Univalle ubicada en la zona sur de la ciudad de Santiago de Cali, al interior de la Universidad del Valle en el barrio Meléndez. Los datos están comprendidos entre el primer día del mes de Junio del 2013 al día treinta y uno de Mayo del año 2014. El principal objetivo de este estudio fue tratar la información

faltante mediante una propuesta de imputación. Los autores utilizaron métodos de datos funcionales (ADF) para modelar las concentraciones de material particulado aerodinámico inferior o igual a 2.5 micrómetros nominales, identificando parámetros de gran interés que aportan información para el desarrollo del método de imputación propuesto. Basados en el comportamiento de los datos observados (simetría en la forma de la distribución para las concentraciones por día) asumen un proceso gaussiano como criterio para la reconstrucción de las curvas de $PM_{2.5}$ con datos faltantes y evalúan posteriormente la calidad que poseen dichas curvas. Su principal hallazgo establece que el proceso de imputación utilizado produce los mejores resultados cuando la pérdida de información no es muy relevante, ya que logra reproducir el comportamiento bajo el cual se enmarcan las concentraciones del contaminante en estudio. Asimismo, encuentran que el método se degenera cuando se trata con días en los cuales se tiene una alta proporción de datos faltantes. Finalmente, la evaluación de la calidad de las imputaciones utilizando el método ADF conserva tanto la media como la desviación estándar funcional en el tiempo.

- Las principales diferencias entre el trabajo realizado por Caicedo and Jiménez (2016) y la presente investigación radican en:
 - No se reconstruye la curva funcional con información faltante partiendo de un proceso gaussiano como es el caso de Caicedo and Jiménez (2016). Por el contrario, se estima la información faltante partiendo del comportamiento de las curvas con información completa.
 - Se realizan estimaciones de información faltante mediante tres metodologías clásicas y cuatro funcionales, permitiendo generar comparación en el rendimiento de cada una.
 - Se consideran y estudian diversos escenarios de información faltante, donde cada uno fue simulado 1000 veces para asegurar la consistencia en los resultados, permitiendo diagnosticar la precisión de las estimaciones ante diferentes longitudes y porcentajes aleatorios de información faltante.
 - Se utilizan indicadores que evalúan el rendimiento de cada una de las metodologías propuesta en cada escenario contemplado, generando un marco de referencia para obtener mejores conclusiones.

Capítulo 5

Objetivos

5.1. Objetivo General

Implementar el Análisis de Datos Funcionales (ADF) y metodologías clásicas para estimar datos faltantes en mediciones de O_3 diarias en la estación Compartir ubicada en la ciudad Santiago de Cali, Valle del Cauca, cuando se presentan cambios en el porcentaje de datos faltantes y espacios consecutivos faltantes.

5.2. Objetivos Específicos

- Estimar datos faltantes utilizando métodos de imputación basados en el Análisis de Datos Funcionales (ADF). A saber, método de datos de media funcional (FMeanDM), método de datos de mediana funcional (FMedDM), método iterativo basado en la media funcional (FMeanBIM) y método iterativo basado en la mediana funcional (FMedBIM).
- Estimar datos faltantes utilizando métodos de imputación clásicos. A saber, método de imputación por medias, método de imputación por media antes y después y por ultimo, método de imputación por regresión.
- Comparar el desempeño de las técnicas clásicas de estimación de datos faltantes contra técnicas basadas en el Análisis de Datos Funcionales (ADF).

Capítulo 6

Marco Teórico

A continuación se presentan las definiciones teóricas de algunos de los conceptos y metodologías, que se utilizarán para el desarrollo del problema planteado.

6.1. Contaminantes Criterio

Los contaminantes criterio son aquellos agentes del aire que tienen el mayor impacto en la salud humana. Para regularlos existe una normatividad específica, que establece los límites de concentraciones permisibles de cada uno, a lo largo de cierto periodo de tiempo. El monitoreo de estos contaminantes sirve para evaluar la calidad del aire, identificando 5 contaminantes que afectan a la salud inmediatamente desde su inhalación: Dióxido de azufre (SO_2), Dióxido de Nitrógeno (NO_2), Partículas suspendidas (PM_{10} y $PM_{2.5}$), Monóxido de Carbono (CO) y Ozono troposférico (O_3).

6.2. Contaminantes criterio medidos en Cali

Actualmente en la ciudad de Cali, el grupo Calidad del Aire del DAGMA (MinAmbiental, 2016b), revisa permanentemente el estado de los equipos analizadores los contaminantes criterio funcionando con nueve (9) estaciones automáticas las cuales se listan a continuación:

Tabla 6.1: Estaciones de Monitoreo en Cali

ID	Estación	Ubicación	Variables Medidas
1	La Flora	La Flora - Zona Norte	PM_{10}
2	Obrero	Obrero - Zona Centro	PM_{10}
3	Transitoria EDB- Navarro	Poblado - Zona Oriente	$PM_{2.5}$, PM_{10}
4	Base Aérea	La Base - Zona Nororiente	$PM_{2.5}$, O_3 , SO_2
5	Pance	Pance - Zona Rural	PM_{10} , O_3
6	Univalle	Meléndez - Zona Sur	$PM_{2.5}$, O_3 , SO_2
7	Compartir	Compartir - Zona Oriente	$PM_{2.5}$, PM_{10} , O_3
8	La Ermita	La Ermita - Zona Centro	PM_{10} , SO_2
9	Cañaveralejo	Cañaveralejo - Zona Suroccidente	PM_{10} , SO_2

6.2.1. Ozono Troposférico

Según Bermejo et al. (2010) la contaminación atmosférica es uno de los principales problemas ambientales que encienden las alarmas a nivel mundial y manifestados a diversas escalas, desde eventos de categoría

local a fenómenos transcontinentales y globales que pueden llegar a cambiar las condiciones de vida del ser humano y de los ecosistemas en todo el planeta.

El Ozono como componente natural de la troposfera, se encuentra generalmente en concentraciones bajas. La generación de Ozono en la atmósfera se produce mediante reacciones químicas entre compuestos orgánicos volátiles (COV), Monóxido de Carbono (CO) y óxidos de Nitrógeno (NO_x) en presencia de la radiación solar. Estos compuestos se conocen como los precursores de la formación de Ozono. Lo descrito por Delgado Saborit (2005), en una atmósfera no alterada por la actividad humana, estas reacciones forman parte del ciclo del Carbono al transformar los COV que se emiten de forma natural (por la vegetación, la actividad biológica de las zonas húmedas, etc.) en Dióxido de Carbono y vapor de agua, como se muestra en la Figura 6.1. Pero en las regiones directamente influenciadas por las emisiones antropogénicas, los altos niveles emitidos de NO_x actúan como catalizadores para la formación de Ozono a partir de los COV, cuyos niveles a su vez también se elevan como consecuencia de la actividad urbana e industrial. De esta manera, las concentraciones de Ozono en la troposfera pueden alcanzar valores elevados, por encima del fondo natural, constituyendo un problema de contaminación atmosférica y/o ambiental.

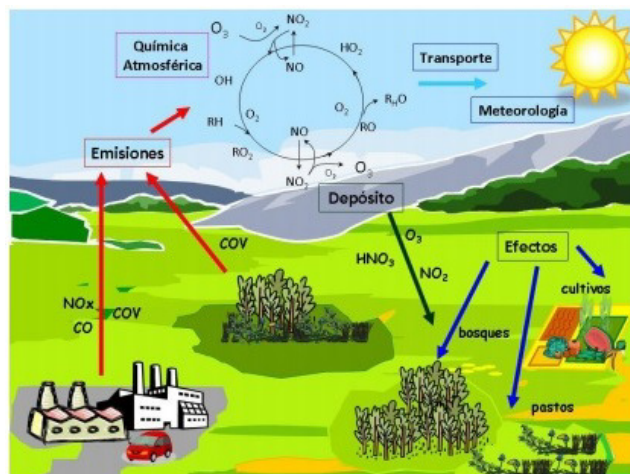


Figura 6.1: Ciclo simplificado del Ozono y procesos relacionados.

El Ozono troposférico no ejerce la función protectora que cumple en la estratosfera sino que por el contrario, dada su alta reactividad y su fuerte capacidad oxidante, cuando sus niveles se elevan por encima del fondo natural, puede provocar efectos adversos en los materiales (derivados de sus efectos corrosivos), en la salud humana (relacionados con problemas en las vías respiratorias), en la vegetación y los ecosistemas. La producción de Ozono troposférico es máxima cuando coinciden concentraciones elevadas de precursores con unas condiciones meteorológicas que favorezcan las reacciones fotoquímicas entre ellos, como son la temperatura elevada, una alta radiación solar y la ausencia de lluvias y vientos fuertes (Delgado Saborit, 2005).

El Ozono generado en la troposfera llega a alcanzar una vida media de semanas, pudiendo transportarse a grandes distancias dependiendo de las condiciones meteorológicas. De esta manera, las concentraciones de este contaminante pueden ser el resultado de la mezcla del Ozono generado a partir de emisiones locales, formado a partir de precursores emitidos en zonas lejanas, incluso procedentes de otro continente, Ozono transportado desde zonas alejadas y también de las intrusiones procedentes de la estratosfera.

El Ozono a nivel del suelo, que no debe confundirse con la capa de Ozono en la atmósfera superior es uno de los principales componentes de la niebla tóxica. Éste se forma por la reacción con la luz solar (fotoquímica) de

contaminantes como los óxidos de Nitrógeno NO_x procedentes de las emisiones de vehículos o la industria y los compuestos orgánicos volátiles COV emitidos por los vehículos, los disolventes y la industria. Los niveles de Ozono más elevados se registran durante los períodos de tiempo soleado (Shumake-Guillemot et al., 2014).

6.3. Datos faltantes

El término de datos faltantes hace referencia a la falta o ausencia de algún tipo de información sobre los fenómenos en los que se está interesado. En general, los datos faltantes introducen una dificultad para explicar y comprender dichos fenómenos sobre los cuales se desea concluir ya que, los resultados de una investigación se basan principalmente en el análisis de observaciones y con la ausencia de datos dada cualquier circunstancia, afectará la capacidad de inferir o deducir la naturaleza de los fenómenos (McKnight et al., 2007).

La presencia de datos faltantes es algo muy común. Disponer de un archivo de datos completos es ideal, sin embargo, la realidad de cualquier tipo de captura de información precede la ausencia de datos y aplicar métodos de imputación inapropiados para lograr una base de datos completa puede generar más problemas de los que resuelve (Galván, 2007). Por lo tanto, encontrar una estimación del valor ausente se ha convertido en línea de estudio y requiere minimamente el entendimiento de patrones, mecanismos y metodologías para engranar la maquinaria global que opera la imputación de valores perdidos.

6.3.1. Patrones generadores de datos faltantes

Si la base de datos se interpreta como una matriz, en donde los renglones son las unidades de observación y las columnas representan a las variables de interés, la elección del método de imputación debiera tener en cuenta el comportamiento de los datos omitidos, ya que el análisis visual permite identificar patrones como se evidencia en la Figura 6.2.

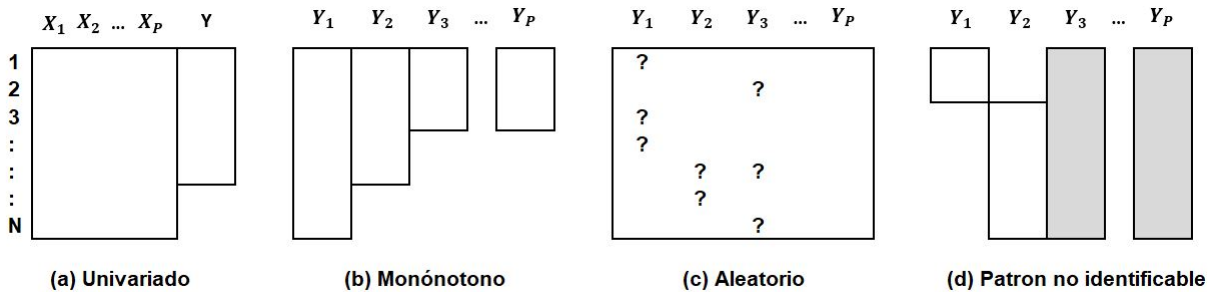


Figura 6.2: Patrones de datos ausentes fuente propia

En la Figura 6.2a, la ausencia de respuesta se concentra en la variable Y , lo cual se conoce como “patrón univariado”, mientras que en la Figura 6.2b la omisión sigue un “patrón escalonado o monótono”, que es característico de estudios longitudinales, en donde Y_p representa el valor de la variable en la p -ésima ronda, por otro lado la Figura 6.2c da cuenta de un “patrón aleatorio”, y en cualquier celda pueden existir datos faltantes; es decir, las omisiones no están dispuestas en una forma predeterminada. Finalmente, es posible que la falta de respuesta ocurra en dos o más variables en distintos períodos. Considere $Y_1, Y_2, \dots, Y_p$ y el patrón que se presenta en la Figura 6.2d, en donde Y_1 y Y_2 tienen datos para distintos registros. Esta situación corresponde a un “patrón con parámetros no identificados”.

6.3.2. Mecanismos generadores de datos faltantes

Cuando se están explorando los datos faltantes, es importante llegar a una conclusión sobre el mecanismo de la ausencia de estos datos, esto es, saber la razón por la cual los datos están perdidos. Hay varias clasificaciones de datos faltantes según el mecanismo de su ausencia. Hay datos faltantes completamente aleatorios (Missing Completely at Random, MCAR), datos faltantes aleatorios (Missing at Random, MAR) y datos faltantes no aleatorios (Missing Not a Random, MNAR) (Osborne, 2012).

La determinación del mecanismo de ausencia es importante. Si se puede inferir que los datos son faltantes al azar (MCAR ó MAR), dicha ausencia causada por fenómenos aleatorios puede ser un problema desde una perspectiva (en que casi siempre reduce el tamaño de la muestra o los grados de libertad para el análisis), pero esto no sesgaría potencialmente los resultados. Sin embargo, los datos faltantes no aleatorios (MNAR) podrían ser una influencia de sesgo muy fuerte (Rubin, 1976).

- **Missing Completely at Random (MCAR):** Los datos perdidos completamente al azar se presentan cuando la probabilidad de que una respuesta a una variable sea dato faltante es independiente tanto del valor de esta variable como del valor de otras variables del conjunto de datos, esto es que la ausencia de la información no esta originada por ninguna variable presente en la matriz de datos (Otero, 2011), es decir, si la probabilidad de ser un dato faltante es el mismo para todos los casos, entonces se dice que los datos son perdidos completamente al azar (MCAR) y esto no implica que las causas de los datos faltantes estén relacionadas con los datos (Van Buuren, 2012).
- **Missing at Random (MAR):** Los datos perdidos al azar se presentan cuando la probabilidad de que una respuesta sea dato faltante es independiente de los valores de la misma variable pero es dependiente de los valores de otras variables del conjunto de datos, es decir, la ausencia de datos está asociada a variables presentes en la matriz de datos. La probabilidad de que se produzca la ausencia de una observación depende de otras variables pero no de los valores de la variable con el valor ausente. De otra forma, sea Y_{obs} que denota los componentes observados de Y y Y_{mis} los componentes faltantes. Si la falta del dato depende solamente de los componentes Y_{obs} de Y que son observados, y no de los componentes que son faltantes. Los dos mecanismos de datos faltantes mencionados se denominan también ignorables, por cuanto producen efectos que se pueden ignorar si se controla adecuadamente por las variables que determinan la no respuesta (Otero, 2011).
- **Missing Not at Random (NMAR):** Los datos perdidos no debidos al azar se presentan cuando la probabilidad de que una respuesta a una variable sea dato faltante es dependiente de los valores de la variable, es decir, cuando la probabilidad de los datos perdidos sobre una variable Y depende de los valores de dicha variable una vez que se han controlado el resto de las variables. Por ejemplo, en el caso de tener un estudio de variables peso y edad, si los sujetos con mayores valores de peso tienen un porcentaje de datos faltantes mas elevados en esta variable para aquellos con la misma edad, entonces en este caso los datos son NMAR.

6.4. Imputación de datos faltantes

El significado propio de imputar es sustituir observaciones, ya sea por falta de información, errores en los datos u observaciones que no corresponden a la realidad (Schafer, 1999). La imputación de datos ha sido un procedimiento que ha tenido ciertos cambios y de alguna manera ha evolucionado a través del pasar de los años. Dichos cambios han desencadenado una gran cantidad de procedimientos que de alguna manera tratan de mejorar y solucionar eficientemente los posibles inconvenientes que conlleva el hecho de imputar. Las primeras técnicas de imputación tuvieron su desarrollo y aplicación a mediados de los años setenta en los cuales la imputación de datos significaba identificar y sustituir los registros sin información (Galván, 2007).

Los procedimientos más utilizados consistían en la metodología hot-deck, cold-deck y los métodos de ajuste por promedios.

Más tarde Rubin (1976), propuso un análisis de datos faltantes basado en métodos de inferencia estadística. Posteriormente, la aparición del algoritmo Expectation Maximization (EM) permitió generar estimadores robustos a partir de la aplicación del método de máxima verosimilitud (MV) (Dempster et al., 1977). Por último Rubin (2004) introdujo el concepto de imputación múltiple (IM). Después de esto se han realizado innumerables aportes e investigaciones importantes en el desarrollo de las metodologías. Los distintos métodos de imputación se pueden clasificar según dos criterios. Por un lado, pueden ser simples o múltiples. Por otro lado, pueden ser determinísticos o aleatorios (Otero, 2011):

- **Imputación simple:** Consiste en asignar un valor por cada dato faltante basándose en el valor de la propia variable o de otras variables, y con ello consiguiendo una base de datos completa.
- **Imputación múltiple:** Consiste en asignar a cada valor faltante varios valores, generando un conjunto de datos completos. En cada conjunto de datos completo se estiman los parámetros de interés y posteriormente se combinan los resultados obtenidos.
- **Métodos de imputación determinísticos:** Son aquellos que producen las mismas respuestas cuando se repite la imputación en varias unidades bajo mismas condiciones.
- **Métodos de imputación aleatorios:** Son aquellos que producen resultados diferentes cuando se repite el método de imputación bajo las mismas condiciones para la unidad.

Actualmente, la facilidad y desconocimiento de cómo tratar con información faltante conlleva a la aplicación de la metodología listwise o pairwise deletion, es decir, hacer uso de toda la información disponible, eliminando las observaciones que no tienen datos. De igual forma, es común encontrar estimaciones de valores ausentes por técnicas que sean prácticas y de fácil entendimiento como lo es la imputación simple.

6.5. Eliminación de casos - No imputación

Este procedimiento consiste en incluir en el análisis los casos que presentan observaciones completas en todas las variables. Este método solo debe utilizarse cuando el proceso de recogida de datos es aleatorio, porque en otro caso introduce sesgo. Otro inconveniente es que el tamaño muestral puede llegar a sufrir una gran reducción y afectar a la representatividad de la muestra.

- **Eliminación por Listwise (análisis de caso completo):** Elimina todos los datos de un caso que tiene uno o más valores perdidos. Esta técnica se usa comúnmente si el investigador está llevando a cabo un estudio de tratamiento y desea comparar un análisis de finalizador (eliminación en lista) versus un análisis por intención de tratar (incluye casos con datos faltantes imputados o tomados en cuenta mediante un método algorítmico) en un diseño de tratamiento. En la mayoría de los demás casos, a menudo es desventajoso utilizar la eliminación de listwise. Esto se debe a que las suposiciones de MCAR suelen ser poco frecuentes. Debido a esto, los métodos de eliminación de lista producen parámetros de sesgo y las estimaciones.
- **Eliminación por Pairwise deletion (análisis de casos disponibles):** intenta minimizar la pérdida que ocurre en la eliminación de lista. Una manera fácil de pensar cómo funciona la eliminación por parejas es pensar en una matriz de correlación. Una correlación mide la fuerza de la relación entre dos variables. Para cada par de variables para las cuales hay datos disponibles, el coeficiente de correlación tomará esos datos en cuenta. Por lo tanto, la eliminación por parejas maximiza todos los datos disponibles por análisis y análisis. Una fortaleza de esta técnica es que aumenta la potencia

en sus análisis. Aunque esta técnica suele preferirse a la eliminación de lista, también supone que los datos que faltan son MCAR. También hay desventajas. Una desventaja del uso de la eliminación por pares es que el estándar de errores calculado por la mayoría de los paquetes de software usa el tamaño de muestra promedio en todos los análisis. Esto tiende a producir un estándar de errores que se subestiman o sobreestiman.

6.6. Imputación simple

6.6.1. Imputación por medias

Este método, propuesto por primera vez por Wilks (1932), es posiblemente uno de los procedimientos de imputación más antiguo y más sencillo (Dempster et al., 1977). Los valores faltantes de una variable se sustituyen mediante la media de las unidades observadas en esa variable. Este método tiene una versión determinística y una versión aleatoria, en la que se incluye un residuo aleatorio. La imputación por media tiene dos variantes:

- **Imputación por media no condicional (MeanClas):** Consiste en estimar la media de los valores observados; es decir, si y_{ij} es el valor de la variable Y_j para la unidad i , el método de imputación por medias incondicional trata de estimar los valores faltantes y_{ij} por $\bar{y}_j^{(j)}$, la media de los valores observados de Y_j . En su aplicación se asume que los datos faltantes siguen un patrón MCAR. Este procedimiento preserva el valor medio de la variable pero los estadísticos que definen la forma de la distribución (varianza, percentiles, sesgo, etc.) pueden verse afectados, de la misma forma que también se distorsionan las relaciones entre las variables. Ha sido ampliamente documentado que su aplicación afecta la distribución de probabilidad de la variable imputada, atenúa la correlación con el resto de las variables y subestima la varianza, entre otras cosas. A pesar de los inconvenientes señalados, es común el uso de esta metodología ya que existe la falsa creencia de que en una distribución de probabilidad normal el promedio de los datos es un buen estimador de las observaciones omitidas. Asimismo, los problemas se multiplican en la medida de que la falta de respuesta se incrementa (Galván, 2007).
- **Imputación por media condicional:** Imputa medias condicionadas a valores observados. Un método común consiste en agrupar los valores observados y no observados en clases e imputar los valores faltantes por la media de los valores observados en la misma clase.
- **Media antes y después (MeanBAClas):** Sea $y_1, y_2, \dots, y_n$ una serie de tiempo con n observaciones, de los cuales k valores denotados por $y_1^*, y_2^*, \dots, y_k^*$ faltan. Por lo tanto, los datos observados con valores perdidos son:

$$y_1, y_2, \dots, y_{n_1}, y_1^*, y_{n_1+1}, y_{n_1+2}, \dots, y_{n_2}, y_2^*, y_{n_2+1}, y_{n_2+2}, \dots, y_k^*, y_n \quad (6.1)$$

Por lo tanto, el primer valor faltante ocurre después de las n_1 observaciones, el segundo valor faltante ocurre después de las n_2 observaciones, y así sucesivamente. Hay que tener en cuenta que puede haber más de una observación faltante consecutiva. El método de la media antes-después reemplaza todos los valores perdidos con la media de un dato antes del valor faltante y un dato después del valor faltante. Así, para los datos en 6.1, y_1^* será reemplazado por

$$\bar{y}_1 = \frac{y_{n_1} + y_{n_1+1}}{2} \quad (6.2)$$

y y_2^* será reemplazado por

$$\bar{y}_2 = \frac{y_{n_2} + y_{n_2+1}}{2} \quad (6.3)$$

6.6.2. Imputación por regresión

Es un método propuesto por primera vez por Buck (1960). Se emplean modelos de regresión para imputar información en la variable Y , a partir de covariables $(X_1, \dots, X_K)$ correlacionadas con Y . Este procedimiento consiste en eliminar las observaciones con datos incompletos y ajustar la ecuación de la regresión para predecir los valores faltantes. Sea n el tamaño muestral y consideremos la variable Y que presenta los primeros r valores observados y $n-r$ valores faltantes. Supongamos que las K variables, $X = (X_1, \dots, X_K)$, no presentan valores perdidos (Dempster et al., 1977). Si para el caso i tenemos que el valor Y_i no se observa, este valor faltante es imputado mediante un modelo de regresión, cuya forma general es:

$$g\{E(Y)\} = X\beta \quad Y \sim F \quad (6.4)$$

donde g se denomina función link, y F es la función de distribución. A continuación se describen los modelos posibles para el caso a trabajar, donde Y es una variable continua. Se considera que g es la identidad e Y sigue una distribución Normal, entonces se tiene un modelo de regresión lineal:

$$E(Y) = X\beta \quad Y \sim Normal \quad (6.5)$$

En este caso se encuentran dos variantes:

- **Imputación mediante regresión determinística (RegreClas):** En este caso estamos ante un modelo de regresión lineal. Es un modelo determinístico, el valor faltante es imputado usando la siguiente ecuación de regresión:

$$\hat{y}_i = \hat{\beta}_{0.1.2\dots k} + \sum_{j=1}^k \hat{\beta}_{j.1.2\dots k} x_{ij} \quad (6.6)$$

donde $\hat{\beta}_{0.1.2\dots k}$ y $\hat{\beta}_{j.1.2\dots k}$ representan los coeficientes de la regresión de Y sobre $X = (X_1, \dots, X_k)$ basada en las r observaciones completas.

- **Imputación mediante regresión estocástica:** En este caso tenemos un modelo de regresión como el definido anteriormente pero incorporando un residuo aleatorio a la predicción. Es decir, imputaremos el valor faltante mediante:

$$\hat{y}_i = \hat{\beta}_{0.1.2\dots k} + \sum_{j=1}^k \hat{\beta}_{j.1.2\dots k} x_{ij} + z_i \quad (6.7)$$

donde $z_i \sim N(0, \hat{\sigma}_{1.2\dots k})$, siendo $\hat{\sigma}_{1.2\dots k}$ la varianza residual de la regresión de Y sobre X basada en las observaciones completas.

Aunque existen otras técnicas de imputación destacadas y de mayor robustez, las desglosadas anteriormente son las más utilizadas en el campo científico junto a las listwise o pairwise deletion. La gran ventaja de la imputación simple es que se trabaja con bases de datos completos, pero este método trata los valores imputados como si fueran verdaderos y, por tanto, sobreestima la precisión ya que no tiene en cuenta la variabilidad de las componentes entre las distintas imputaciones realizadas (Galván, 2007).

6.7. Construcción de datos funcionales

El análisis de datos funcionales (ADF) es una rama de la estadística, que estudia y analiza la información contenida en curvas, superficies, o cualquier elemento que varía sobre un continuo, generalmente el tiempo (Ramsay and Silverman, 2005). Las particularidades de estos objetos, las funciones, hacen que surja todo un aparato de definiciones, teoremas y herramientas para poder estudiarlas correctamente (Manteiga and Vieu, 2007). Por ello, el primer paso será definir y comentar los conceptos básicos, las características y peculiaridades de este tipo de datos. La explosión tecnológica ha hecho del ADF una disciplina emergente dentro de la estadística con multitud de publicaciones y estudios. Prueba de este interés es la creciente presencia en revistas especializadas y la creación de un grupo de trabajo específico dentro del ERCIM (European Research Consortium for Informatics and Mathematics) (Carbonell and Stephanidis, 2012).

6.7.1. Base de funciones

En la práctica, los datos funcionales son obtenidos a partir de mediciones discretas, siendo necesario convertirlos a funciones sobre las cuales se pueda trabajar, para esto pueden emplearse diferentes tipos de ajustes. Teniendo en cuenta lo anterior, la representación de estos datos como datos funcionales a partir de una base ortonormal brinda una ventaja tanto teórica como práctica, esta base se utiliza para transformar datos discretos en datos funcionales.

- **Bases de Fourier:** la extensión de Fourier de $\chi(t)$ es de la forma:

$$\hat{\chi}(t) = c_0\phi_0 + \sum_r c_{2r-1}\phi_{2r-1}(t) + c_{2r}\phi_{2r}(t) \quad (6.8)$$

donde $\phi_0(t) = \frac{1}{\sqrt{T}}$, $\phi_{2r-1}(t) = \frac{\sin(r\omega t)}{\sqrt{T/2}}$, $\phi_{2r}(t) = \frac{\cos(r\omega t)}{\sqrt{T/2}}$ forman una base periódica de período $T = \frac{2\pi}{\omega}$.

Algunas características importantes de esta base son la fácil diferenciabilidad y la periodicidad. Dado que esta formada por senos y cosenos, siendo factible conseguir la expresión derivada en cualquier orden, además debido a sus características periódicas, se ajusta mejor a los datos que presenten esta particularidad (Ramsay and Silverman, 2005).

La expansión de Fourier es útil para funciones estables, sin grandes variaciones y con una curvatura mas o menos constante. Igualmente es conveniente que los datos muestren una periodicidad. Otra característica de este tipo de bases es que genera expansiones suaves en el sentido de diferenciabilidad, por lo que no será apropiada usarla cuando las funciones presenten discontinuidades (Ramsay and Silverman, 2005).

Bases B-Splines: estas bases combinan la eficiencia computacional de los polinomios con una mayor flexibilidad, que permite elegir un número menor de funciones (K) con respecto a las bases anteriormente mencionadas, para obtener buenos resultado (Ramsay and Silverman, 2005). En primer lugar para definir un Spline es dividir el intervalo $T = [a, b]$ en L subintervalos separados por los puntos $a = t_0, t_1, \dots, t_L = b$, en cada intervalo, el Spline es un polinomio de grado m que trata de ajustar la curva. El orden del polinomio m hace referencia al número de coeficientes que hacen falta para definirlo, es decir, estos polinomios deben coincidir en los nodos de la partición hasta la derivada de orden $m-2$, por lo que a excepción de los polinomios de orden 2, las uniones serán suaves en el sentido de diferenciabilidad. Con estas restricciones el número de grados de libertad del Spline será $m + L - 1$.

Por tanto, un elemento importante será definir el orden de los polinomios, el número de particiones y el punto en que se fijen los nodos * (Eubank, 1999).

Las bases de $B - splines$ permiten aproximar todas estas funciones facilitando así su manejo. Sus características se basan en las siguientes propiedades:

- Cada elemento de la base $\phi_k(t)$ será una función Spline de orden m y partición τ .
- Cualquier combinación lineal de funciones Spline es una función Spline.
- Cualquier función Spline de orden m sobre la partición τ se puede expresar como combinación lineal de las funciones de la base.

Sea $B_k(t, \tau)$ la representación del valor del $k - \text{ésimo}$ elemento de la base sobre la partición τ en el instante t , la función Spline $S(t)$ se representa como:

$$S(t) = \sum_{k=1}^{m+L-1} c_k B_k(t, \tau) \quad (6.9)$$

Bases con restricción: estas bases tienen en cuenta restricciones sobre la suavización de las curvas (Ramsay and Silverman, 2005). Una función suavizada positiva x se puede definir a partir de la función exponencial:

$$x(t) = e^{W(t)} \quad (6.10)$$

donde $W(t)$ corresponde a la estimación de las curvas funcionales, que dependiendo de las características de los datos observados se pueden usar base Fourier o B-spline (Ramsay and Silverman, 2005).

En ocasiones se quiere suavizar funciones estrictamente creciente por ejemplo la precipitación acumulada. En este caso Ramsay and Silverman (2005) expresa la función suave a partir la velocidad Dx como el exponencial de una función sin restricciones W :

$$Dx(t) = e^{W(t)} \quad (6.11)$$

Integrando ambos lados de esta ecuación, se obtiene:

$$x(t) = C + \int_{t_0}^t \exp[W(u)] du, \quad (6.12)$$

donde C es una constante estimada a partir de los datos.

6.7.2. Construcción de datos funcionales

En Análisis de Datos Funcionales (ADF) los datos en lugar de ser un conjunto de vectores, como en análisis multivariado clásico, son un conjunto de curvas. En la mayoría de las aplicaciones, las curvas observadas proceden de la observación de un proceso estocástico en tiempo continuo (Aguilera, 2009).

*Son puntos de control de tal manera que la función y sus dos primeras derivadas son continuas en cada punto y por lo tanto en todo su intervalo (Olaya, 2012).

El dato funcional se define a partir de una base de funciones $\{\phi_k\}_{k \in \mathbb{N}}$, tal que cualquier función puede representarse mediante una combinación lineal de ella, es decir, $x(t) = \sum_{k=1}^K c_k \phi_k(t)$, donde c_k son los coeficientes estimados. La expresión de forma matricial queda expresada como:

$$\mathbf{X}(t) = \mathbf{c}^T \Phi = \Phi^T \mathbf{c} \quad (6.13)$$

Φ es una matriz que contiene los valores de $\phi_k(t)$ y K será el número de funciones de la base. En este contexto es importante optimizar la selección del valor de K , dado que la expresión de la función seguirá siendo infinita o demasiado grande ya que este se comporta como otro parámetro del modelo midiendo el grado de suavización de la función (Ramsay and Silverman, 2005).

De la expresión 6.13 se denota que el problema consiste en estimar el vector de coeficientes $\mathbf{c}$, este vector pondera cada función de la base sobre la función de interés $x(t)$, donde Ramsay and Silverman (2005) estimaron el vector $\mathbf{c}$ a partir de mínimos cuadrados ordinarios. A continuación se presenta el proceso de estimación propuesto:

Sea y_j el valor observado de la variable Y en el tiempo $t_j \forall j = 1, \dots, n$, tal que puede ser expresados como:

$$y_j = x(t_j) + \epsilon_j = \phi^T(t_j) \mathbf{c} + \epsilon_j, \quad (6.14)$$

donde $x(t_j)$ es el valor de la función ajustada con (6.13) en el tiempo t_j y ϵ_j es el error de estimación en el momento t_j . Suponiendo que los errores se distribuyen normal, tienen media cero, varianza constante e independientes. La expresión por mínimos cuadrados ordinarios a minimizar es (Ramsay and Silverman, 2005):

$$SSE(\mathbf{c}) = \sum_j^n \left[y_j - \sum_k^K c_k \phi_k(t_j) \right]^2 = \sum_j^n [y_j - \phi^T(t_j) \mathbf{c}]^2, \quad (6.15)$$

Sea $\mathbf{Y}$ un vector conformado por todos los valores ajustados y_j , Φ^T una matriz de dimensión $n \times K$ compuesta por los valores de las K funciones base $\phi_k(t_j)$ y ϵ el vector de residuales de tamaño $n \times 1$, por lo tanto la expresión (6.14) es expresada de forma matricial como (Ramsay and Silverman, 2005):

$$\mathbf{Y} = \Phi^T \mathbf{c} + \epsilon \quad (6.16)$$

de esta manera, la expresión (6.15) se puede expresar como:

$$SSE(\mathbf{c}) = (\mathbf{Y} - \Phi^T \mathbf{c})^T (\mathbf{Y} - \Phi^T \mathbf{c}), \quad (6.17)$$

resolviendo el sistema para estimar $\mathbf{c}$, se obtiene el estimador $\hat{\mathbf{c}}$ que minimiza la suma de cuadrados de (6.17), y esta dado por:

$$\hat{\mathbf{c}} = (\Phi \Phi^T)^{-1} \Phi \mathbf{Y} \quad (6.18)$$

Finalmente los valores ajustados son expresados como:

$$\hat{\mathbf{Y}} = \Phi^T (\Phi \Phi^T)^{-1} \Phi \mathbf{Y} \quad (6.19)$$

6.7.3. Número óptimo de funciones de la base (K)

Ramsay and Silverman (2005) plantean que para seleccionar el número K se debe tener en cuenta tanto el sesgo,

$$\text{Sesgo}[\hat{x}(t)] = x(t) - E[\hat{x}(t)] \quad (6.20)$$

y la varianza de la estimación,

$$\text{Var}[\hat{x}(t)] = E\left[\{\hat{x}(t) - E[\hat{x}(t)]\}^2\right] \quad (6.21)$$

Por lo tanto, un indicador que permite controlar tanto el sesgo como la varianza es el Error Cuadrático Medio (ECM),

$$ECM[\hat{x}(t)] = \text{Sesgo}^2[\hat{x}(t)] + \text{Var}[\hat{x}(t)] \quad (6.22)$$

Esta relación indica que vale la pena tolerar un poco de sesgo, siempre y cuando haya un gran reducción de la varianza (Ramsay and Silverman, 2005). Adicional a este proceso, Ramsay and Silverman (2005) proponen la utilización de la validación cruzada generalizada (VCG).

- **Validación cruzada (VC):** es una metodología usada para las elecciones de parámetros que controlen el equilibrio entre la bondad del ajuste y la capacidad predictiva (Delicado, 2008). El método consiste en extraer de la muestra secuencialmente cada una de las observaciones x_i , estimar el modelo con los datos restantes y predecir el dato ausente. Y finalmente, comparar la predicción con el dato real. Si se realiza esto para cada valor de K , entonces se pueden construir la función:

$$ECM_{VC}(K) = \frac{1}{n} \sum_{i=1}^n [y_i - \hat{m}_{(i)}(x_i)]^2 \quad (6.23)$$

que mide el error de la predicción del estimador fuera de la muestra para cada valor de K . El valor que minimice esa función, K_{VC} , será el valor del parámetro elegido. En forma matricial se tiene que:

$$\hat{Y} = SY \quad (6.24)$$

donde los vectores columnas Y e $\hat{Y}$ tienen elementos y_i e $\hat{y}_i$, respectivamente y la matriz S va a ser una matriz de estimación. Esta matriz es análoga a la matriz HAT $H = X(X^T X)^{-1}X^T$ en regresión lineal múltiple:

$$\hat{Y}_L = X(X^T X)^{-1}X^T Y = HY \quad (6.25)$$

y se puede demostrar que:

$$ECM_{VC}(K) = \frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - \hat{y}_i}{1 - s_{ii}} \right)^2, \quad (6.26)$$

al igual que en regresión múltiple, no es necesario ajustar n regresiones no paramétricas, solo hay que ajustar la que hace intervenir todos los datos y anotar la matriz S (Delicado, 2008).

Validación cruzada generalizada (VCG): Estos modelos puede hacerse una modificación a la validación cruzada. A esta metodología se le conoce como Validación Cruzada Generalizada (VCG). Esta modificación se basa en reemplazar los valores s_{ii} de la diagonal principal de la matriz S por su valor promedio (Delicado, 2008).

$$ECM_{VCG}(K) = \frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - \hat{y}_i}{1 - \frac{\nu}{n}} \right)^2, \quad \text{donde } \nu = \text{Traza}(S) \quad (6.27)$$

Reexpresando la ecuación (6.27) se tiene que:

$$ECM_{VCG}(K) = \frac{n\hat{\sigma}_\varepsilon^2}{n - \nu} \quad (6.28)$$

donde

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n - \nu} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (6.29)$$

es una estimación de la varianza de los residuales del modelo. Y el valor de K_{VCG} que minimiza esa función, es el parámetro elegido por el criterio de validación cruzada generalizada (Delicado, 2008).

6.7.4. Estadística descriptiva funcional

Suponiendo que un dato funcional para i réplicas, llega como un conjunto de valores discretos $x_{i1}, \dots, x_{in}$, la primera tarea es convertir estos valores, utilizando métodos no paramétricos, a una función $X(t)$ calculable para cualquier valor de t (Ramsay and Silverman, 2005). Teniendo en cuenta que un primer análisis estadístico de un conjunto de datos debe empezar por un análisis exploratorio, se presentan a continuación algunas características descriptivas funcionales, propuestas en Ramsay and Silverman (2005):

Sea $x_i(t)$ $i = 1, 2, \dots, N$ una muestra de curvas o datos funcionales, la media, varianza, covarianza y correlación funcional están dadas por:

- Promedio funcional:

$$\bar{x}(t) = N^{-1} \sum_{i=1}^N x_i(t) \quad (6.30)$$

- Mediana funcional:

$$\overbrace{\begin{pmatrix} \phi_1(t_1) & \phi_2(t_1) & \cdots & \phi_k(t_1) \\ \phi_1(t_2) & \phi_2(t_2) & \cdots & \phi_k(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(t_j) & \phi_2(t_j) & \cdots & \phi_k(t_j) \end{pmatrix}}^{\Phi} \begin{matrix} \text{Mediana 1} \\ \text{Mediana 2} \\ \vdots \\ \text{Mediana } k \end{matrix} \overbrace{\begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1N} \\ c_{21} & c_{22} & \cdots & c_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ c_{k1} & c_{k2} & \cdots & c_{kN} \end{pmatrix}}^{\hat{c}} = \begin{pmatrix} x_1(t)_{Me} \\ x_2(t)_{Me} \\ \vdots \\ x_j(t)_{Me} \end{pmatrix} \quad (6.31)$$

- Varianza funcional:

$$Var_x(x(t)) = (N - 1)^{-1} \sum_{i=1}^N (x_i(t) - \bar{x}(t))^2 \quad (6.32)$$

- Covarianza funcional:

$$Cov_x(t_1, t_2) = (N - 1)^{-1} \sum_{i=1}^N (x_i(t_1) - \bar{x}(t_1))(x_i(t_2) - \bar{x}(t_2)) \quad (6.33)$$

- Correlación funcional:

$$Corr_x(t_1, t_2) = \frac{Cov_x(t_1, t_2)}{\sqrt{Var_x(t_1) Var_x(t_2)}} \quad (6.34)$$

- Covarianza cruzada funcional:

$$Cov_{x,y}(t_1, t_2) = (N - 1)^{-1} \sum_{i=1}^n (x_i(t_1) - \bar{x}(t_1)) (y_i(t_2) - \bar{y}(t_2)) \quad (6.35)$$

- Correlación cruzada funcional:

$$Corr_{x,y}(t_1, t_2) = \frac{Cov_{x,y}(t_1, t_2)}{\sqrt{Var_x(t_1) Var_y(t_2)}} \quad (6.36)$$

6.8. Métodos de Imputación basados en técnicas de conversión de datos en la ADF

6.8.1. Método de datos de medias funcionales (FMeanDM) y Método de datos medianos funcionales (FMedDM)

Estos métodos de imputación permiten reemplazar un dato faltante en el momento t_j , donde este hace referencia al tiempo en el cual fueron tomadas las observaciones y lo reemplaza por el valor correspondiente que se encuentra en la curva media funcional o mediana funcional obtenidas a partir de los datos disponibles y/o completos. La curva media funcional es la media del nivel de concentración mientras que la curva mediana funcional es el valor medio, en la columna de tiempo t_j , $j = 1, \dots, 24$ a través de la replicación de curvas diarias completas. (Pighin and Ieronutti, 2008).

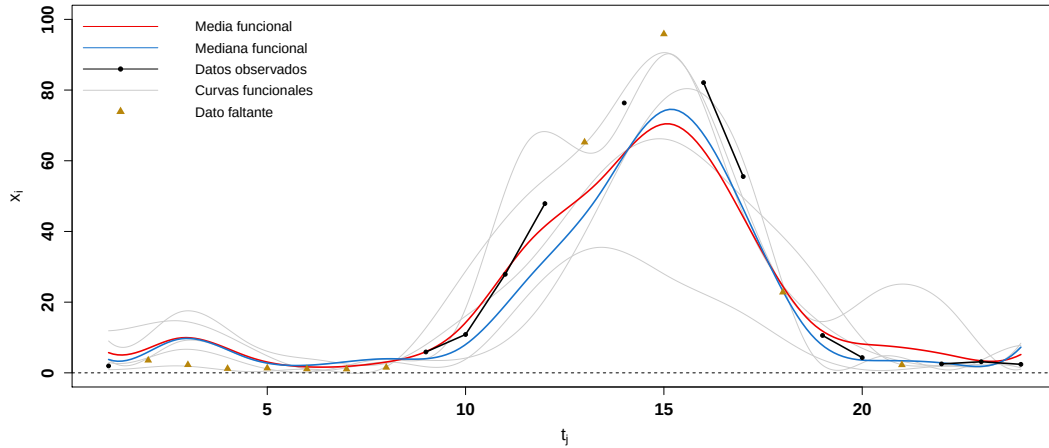


Figura 6.3:

En la Figura 6.3 se ilustra el proceso anterior descrito, donde se tiene los datos observados (Puntos negros) en presencia de datos faltantes en el momento t_j (Triángulos dorados). Por otra parte, la curva roja es la media funcional calculada mediante la expresión 6.30 y la curva azul representa la mediana funcional obtenida mediante 6.31. Según la literatura (Ramsay et al., 2011), el ADF se trabaja únicamente con información completa, es decir, esta no acepta datos faltantes (Curvas grises). Finalmente una vez estimadas la curva media funcional y mediana funcional, se reemplaza en los valores faltantes de los datos observados por el valor asociado estimado en estas curvas.

6.8.2. Método iterativo basado en la media funcional (FMeanBIM) y método iterativo basado en la mediana funcional (FMedBIM)

El enfoque iterativo para la imputación de datos faltantes busca mejorar la estrategia de imputación. Así, el proceso de iteración hará pleno uso de toda la información útil, incluyendo las instancias con los valores faltantes y mejorando el desempeño de imputación (Pighin and Ieronutti, 2008). Inicialmente, antes de comenzar la imputación, los valores faltantes deben de ser reemplazados por los valores puntuales de la media o mediana en la primera iteración y luego los valores faltantes deben de ser imputados iterativamente hasta que el algoritmo converja. La convergencia se satisface siempre que la diferencia en el error (MPAD) entre la imputación en la corriente y la iteración anterior sea como máximo 0.005. MPAD se define como la proporción media de la diferencia absoluta entre el punto del valor observado y_j y el valor ajustado $\hat{y}_j$. Con referencia a (Conte et al., 1986), MPAD también se conoce como la medida media para el error relativo de magnitud (MRE). El cálculo de MPAD para una curva en particular se lleva a cabo en la ubicación de w puntos no faltantes como se muestra:

$$MPAD = \frac{1}{w} \sum_{j=1}^w MRE = \frac{1}{w} \sum_{j=1}^w \left| \frac{y_j - \hat{y}_j}{y_j} \right| \quad (6.37)$$

6.8.3. Indicadores de desempeño

A partir de estos indicadores se optimiza (minimiza) el error de estimación del dato faltante con el objetivo de cuantificar la bondad de ajuste de la imputación, donde P es el número de imputaciones, y_p y $\hat{y}_p$ son los puntos de datos observados e imputados, respectivamente, $\bar{y}$ y $\hat{\bar{y}}$ son sus promedios, y σ_y y $\sigma_{\hat{y}}$ sus desviaciones estándar.

- **AI:** El índice de concordancia mide el acuerdo en términos de similitud entre el valor observado y el valor predicho, donde valores cercanos a uno indican un mejor ajuste y viene dado por:

$$AI = 1 - \left[\frac{\sum_{p=1}^P (y_p - \hat{y}_p)^2}{\sum_{p=1}^P (|\hat{y}_p - \bar{y}| + |y_p - \bar{y}|)^2} \right] \quad (6.38)$$

- **MAE:** El error absoluto medio es la diferencia promedio entre los valores de datos predichos y reales, varía de cero a infinito y se obtiene un ajuste perfecto cuando es igual a cero y viene dado por:

$$MAE = \frac{1}{P} \sum_{p=1}^P |(y_p - \hat{y}_p)| \quad (6.39)$$

- **R^2 :** El coeficiente de determinación (R^2) explica qué parte de la variabilidad en los datos imputados se puede explicar por el hecho de que están relacionados con los valores observados o cuán cerca están los puntos de la línea, este toma valores entre cero y uno, valores más cercanos a uno implican un mejor ajuste y viene dado por:

$$R^2 = \frac{1}{P^2} \left[\frac{\sum_{p=1}^P (\hat{y}_p - \hat{\bar{y}})(y_p - \bar{y})}{\sigma_y \sigma_{\hat{y}}} \right]^2 \quad (6.40)$$

- **SCE:** La suma de cuadrados del error (SCE) resume la diferencia entre las datos observados e imputados, cuanto menor es el valor de SCE mejor es el ajuste del modelo y viene dado por:

$$SCE = \frac{1}{P} \sum_{p=1}^P (\hat{y}_p - y_p)^2 \quad (6.41)$$

Capítulo 7

Metodología

7.1. Datos

Los datos del contaminante criterio O_3 son medidos por el Sistema de Vigilancia de la Calidad del Aire de Santiago de Cali (SVCASC), el cual tiene como principal objetivo analizar las concentraciones obtenidas continuamente de los contaminantes criterio y a su vez, la meteorología en diferentes puntos de la ciudad de Santiago de Cali para poder ejercer un control y llevar registro de los niveles de contaminación atmosférica (MinAmbiental, 2016b). Las concentraciones de O_3 son obtenidas de la estación Compartir, ubicada en la zona oriental de la ciudad de Cali, en el barrio Compartir como se muestra en la Figura 7.1, donde se mide material particulado con un diámetro aerodinámico menor o igual a 2,5 micrómetros nominales; junto al contaminante de interés.

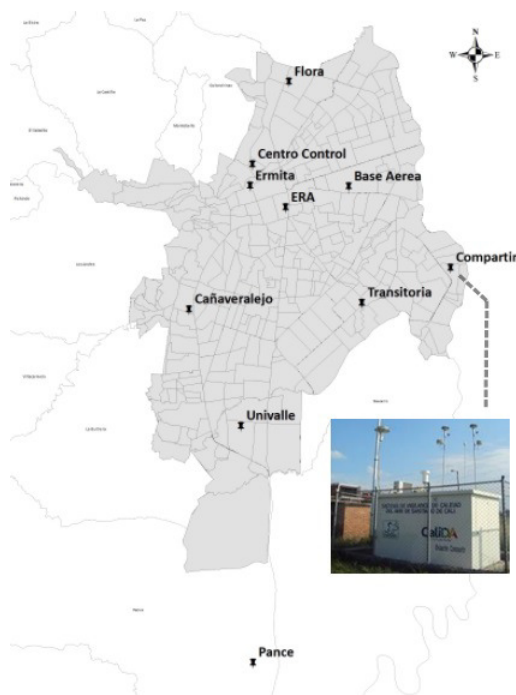


Figura 7.1: Ubicación espacial estaciones de Monitoreo

La información de las concentraciones de O_3 utilizada en la presente investigación se encuentra en forma de promedios horarios, medida en microgramos por metro cúbico ($\mu g/m^3$), correspondiente a los años 2015 y 2016. Para cada año de estudio, se cuenta con mediciones mensuales desagregada en días, es decir, los datos están disponibles para cada hora del día (24 horas), donde el tiempo esta acotado entre el 01 de Enero hasta el 31 de Diciembre, a saber, se cuenta con 365 y 366 días de observaciones. Cabe resaltar que el mes de febrero del año 2016 tiene 29 días en lugar de 28, hecho conocido como año bisiesto con periodo de oscilación de cuatro años salvo que sea año secular. De esta forma, se debería disponer con 17544 observaciones totales en ausencia de valores faltantes. Por lo anterior, queda definida la estructura de datos a evaluar como se ejemplifica de forma general en la Tabla 7.1, donde se definen las filas como los días que componen la base de datos y las columnas serán las horas del día. Así, se debería contar con una matriz de dimensiones 17544x24.

Día	Hora.1	Hora.2	Hora.3	Hora.4	...	Hora.21	Hora.22	Hora.23	Hora.24
1	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	...	$x_{1,21}$	$x_{1,22}$	$x_{1,23}$	$x_{1,24}$
1	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	...	$x_{2,21}$	$x_{2,22}$	$x_{2,23}$	$x_{2,24}$
...	...	...	...	...	...	...	...	...	...
30	$x_{30,1}$	$x_{30,2}$	$x_{30,3}$	$x_{30,4}$	...	$x_{30,21}$	$x_{30,22}$	$x_{30,23}$	$x_{30,24}$
31	$x_{31,1}$	$x_{31,2}$	$x_{31,3}$	$x_{31,4}$	...	$x_{31,21}$	$x_{31,22}$	$x_{31,23}$	$x_{31,24}$

Tabla 7.1: Estructura datos

Los datos están disponibles gracias a la gestión de la Escuela de Estadística de la Universidad del Valle, particularmente fueron tramitados por el docente Javier Olaya Ochoa. Los análisis estadísticos junto al desarrollo de la metodología propuesta se ejecutaron mediante el software estadístico R versión 3.2.5 (Gentleman and Ihaka, 2017). El análisis de datos funcionales se implementó a través de los paquetes fda (Ramsay and Silverman, 2005) y fda.usc (Bande et al., 2016).

7.2. Verificación y elección de la base de datos

El primer paso de este proyecto, consistió en realizar un análisis exploratorio de la base de datos con el principal objetivo de resumir y visualizar la información de O_3 , facilitando la identificación de tendencias o patrones que son relevantes para responder la pregunta de interés, resaltando las principales características que proveen información valiosa para comprender el fenómeno de estudio. Se cuenta con 7730 registros faltantes en la base de datos de un total 17544, esto equivale al 44.1 % de información faltante a nivel general. Por lo anterior, se procedió a identificar y seleccionar el mes que presentará la menor cantidad de valores perdidos, con la finalidad de definir la base de datos de referencia.

Es importante justificar que la escogencia de un solo mes como información de referencia se debe a que los dos años evaluados cuentan con diversidad meteorológica y climatológica, esto a que son años que dan lugar a fenómenos naturales por acoplamiento entre la atmósfera y la temperatura del Pacífico tropical, conocido como el fenómeno de el NIÑO. Se debe recordar que para el año 2016 este acontecimiento impacto fuertemente a la comunidad científica, debido a que no se estimaba que tuviera una duración tan prolongada y que su condición fuera de índole moderada a fuerte. Por lo tanto, se pueden considerar estos años como anómalos respecto a variables ambientales, climatológicas y meteorológicas (MinAmbiental, 2016b).

El mes seleccionado contiene el 0.1344 % de valores perdidos, equivalente a la ausencia de un solo promedio horario en todo el mes. Con la fuerte idea de contar con una base de referencia con información completa, se utiliza el método de imputación basado en regresión para generar la estimación del único valor faltante. Según Junninen et al. (2004), la técnica de regresión logra mejores resultados para imputar pequeñas cantidades de datos ausentes en registros de calidad del aire. Se considera el uso de datos reales y completos como principal insumo para la base de referencia con el objetivo de definir la exactitud de los métodos de imputación evaluados a través de los indicadores de desempeño.

7.3. Construcción de escenarios

Los datos faltantes se generaron a partir del conjunto de datos de referencia completo, utilizando seis escenarios de datos faltantes como se muestra en la Tabla 7.2 (contaminación de los datos de referencia). Para cada conjunto de datos generados, los patrones son diferentes en complejidad y son extraídos con porcentajes aleatorios de datos faltantes totales de 5 %, 10 % y 15 %, de acuerdo con el número máximo de valores perdidos consecutivos por filas (longitud de espacio) de 3 y 7. Tanto los porcentajes aleatorios como los espacios y/o longitudes máximas, son seleccionados en base al trabajo realizado por Yang Lenyap et al. (2015). Por lo tanto, dentro del conjunto de datos generado, las curvas pueden tener diferentes números de valores faltantes aleatorios y diferentes longitudes de espacios faltantes. Por lo tanto, se define la primera parte de la codificación del patrón como los valores que serán extraídos de forma aleatoria y la segunda parte corresponde al espacio máximo consecutivo aleatorio. Por ejemplo, el conjunto de datos con el patrón **P5-E3** consta de un 5 % de información faltante extraída de forma aleatoria y la longitud máxima de espacio por día alcanza hasta tres valores consecutivos.

Tabla 7.2: Patrones diferentes de datos generados (Datos Faltantes)

Patrones	P5-E3	P5-E7	P10-E3	P10-E7	P15-E3	P15-E7
Porcentaje Faltante	5	5	10	10	15	15
Longitud Máxima del espacio	3	7	3	7	3	7

Es relevante especificar que los escenarios serán considerados en cinco días del mes escogido, es decir, se llevará a cabo la investigación con el 84 % de los días completos, dejando de esta forma, el 16 % de los días con información faltante bajo la forma estructural especificada anteriormente. La selección de los días que serán contaminados, se realizará de forma aleatoria para proporcionar a los métodos evaluados escenarios comparables.

7.4. Métodos de imputación

Para dar cumplimiento a los objetivos planteados, después de recrear los patrones o escenarios contemplados para los valores ausentes se dispone tanto los métodos clásicos como los funcionales para estimar los valores perdidos. En esta oportunidad, serán expuestas metodologías utilizadas con gran frecuencia por científicos e investigadores en el campo clásico por su practicidad según Junninen et al. (2004).

La imputación de valores faltantes se realiza de la siguiente manera:

- **Métodos Clásicos:** A a luz de los algoritmos y definiciones planteados en el marco teórico, se desarrolla computacionalmente las siguientes formulaciones:
 - **MeanClas:** En 7.1 se presenta la estructura general de la base de datos para la aplicación del método de la media clásica, en ella se identifican las observaciones con información reportada del promedio horario de O_3 . Lo anterior, equivale al principal insumo para aplicación de la formula del promedio clásico. Para la estimación de los valores faltantes, se hace uso de la siguiente expresión 7.2, donde $d = 1, \dots, 31$ representa d-esimo día y $k = 1, \dots, 24$ hace referencia a la k-esima hora donde se presenta el dato faltante. Por otra parte, $N = 31$ es el numero total de días de la base de referencia y n es el numero de observaciones completas disponibles para realizar la imputación

en la k -ésima hora.

$$\begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ \dots & \dots & x_{dk}^* & \dots & \dots \\ x_{d1} & x_{d2} & x_{d3} & \dots & x_{dn} \end{bmatrix} \quad (7.1)$$

$$x_{dk}^* = \sum_{d=1}^N \frac{x_{dk}}{n} \quad (7.2)$$

- **MeanBAClas:** Para la imputación por este método, en primera instancia se deben identificar el día junto a la hora y/o horas en las cuales se presenta el dato faltante, una vez hecho esto, se realiza un promedio entre el dato inmediatamente anterior y posterior a esté. En 7.3 se ilustra la estructura general de la base de datos para la aplicación del método, en el cual mediante la formula 7.4 se calculan la estimación de los valores faltantes correspondientes, donde $d = 1, \dots, 31$ y $k = 1, \dots, 24$ hacen referencia al d -ésimo día y la k -ésima hora respectivamente faltante en la serie. Finalmente, n_i cuenta el número de observaciones hasta que aparece el primer dato faltante.

$$\begin{bmatrix} x_{11} & \dots & x_{1n_1} & x_{11}^* & x_{1n_1+1} & x_{1n_1+2} & \dots & x_{1n_2} & x_{12}^* & x_{1n_2+1} & \dots & x_{1k}^* & x_{1n} \\ x_{21} & \dots & x_{2n_1} & x_{21}^* & x_{2n_1+1} & x_{2n_1+2} & \dots & x_{2n_2} & x_{22}^* & x_{2n_2+1} & \dots & x_{2k}^* & x_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ x_{dn_1} & \dots & x_{dn_1} & x_{d1}^* & x_{dn_1+1} & x_{dn_1+2} & \dots & x_{dn_2} & x_{d2}^* & x_{dn_2+2} & \dots & x_{dk}^* & x_{dn} \end{bmatrix} \quad (7.3)$$

$$x_{dk}^* = \frac{x_{dn_i} + x_{dn_i+1}}{2} \quad (7.4)$$

- **RegreClas:** El primer paso para realizar la imputación por este método, es identificar el día que presente mayor correlación con el día en donde se presenta el dato faltante. Hecho esto, se procede a realizar una regresión lineal simple mediante la formula 7.5, donde $\hat{\beta}_0$ y $\hat{\beta}_1$ representan los coeficientes de la regresión de Y sobre X basada en las observaciones completas, e i indica el día que presenta la mayor correlación.

$$\hat{Y}_d = \hat{\beta}_0 + \hat{\beta}_1 X_i \quad (7.5)$$

Finalmente la estimación del valor faltante viene dada por 7.6, donde $d = 1, \dots, 31$ indica el día y $k = 1, \dots, 24$ donde se encuentra el valor ausente a imputar.

$$x_{dk}^* = \hat{Y}_{dk} \quad (7.6)$$

- **Métodos funcionales:** Para la imputación por medio de los métodos funcionales se inicia con la construcción de las curvas como se esquematiza a continuación:
- **Construcción de los datos funcionales:** Con la base de referencia completa, lo cual es el principal requisito para realizar la construcción de las curvas funcionales según Ramsay et al. (2011), se procede a definir la base que será utilizada para la estimación de las curvas de O_3 .

Al estudiar la estructura de los datos, es posible descartar que el contaminante criterio evaluado presenta un comportamiento periódico en el dominio continuo definido como las 24 horas del día, por lo cual se descarta la utilización de una base Fourier. Entre tanto, existe una base que cuenta con propiedades de fácil ajuste por integrar segmentos de polinomio unidos de extremo a extremo y de forma adicional, cuenta con mayor eficiencia computacional denominada base Bspline. La elección de esta técnica también se apoyó en estudios realizados anteriormente, en los cuales implementaban esta técnica para variables ambientales y climáticas (Yang Lenyap et al., 2015).

Para construir las curvas funcionales en torno a una base Bspline, se requiere:

- El número óptimo de funciones K que conforman la base junto al parámetro de suavización λ , haciendo uso de la técnica de validación cruzada generalizada. Lo anterior, se desarrolla con el objetivo de encontrar un equilibrio entre el sesgo y varianza. Puntualmente, si se utilizan demasiadas funciones al interior de la base, esto se traduce en un sesgo pequeño pero un aumento en la varianza muestral. Por el contrario, utilizar muy pocas funciones en la base, incrementa el sesgo pero reduce la varianza muestral, como se enmarca en la **Sección 6.6.4.**
- El orden del polinomio, definido por la ecuación 7.7. De esta forma, se requiere la definición del grado a utilizar, que según Ramsay and Silverman (2005) es común utilizar un polinomio cúbico o de grado tres, esto derivado de la continuidad de las dos primeras derivadas. Un Spline cúbico significa que la curva es visualmente suave. Por lo tanto, se define la base funcional Bspline de orden 4.

$$\text{Orden} = \text{GradoPolinomio} + 1 \quad (7.7)$$

- Por ultimo, las curvas funcionales se construyen en base a los nodos equidistantes, definidos por la ecuación 7.8, donde a este punto es posible reemplazar en la ecuación los valores y determinar el número de nodos requeridos para llevar a cabo la suavización.

$$\text{NumeroNodos} = K - \text{Orden} \quad (7.8)$$

Con todo lo anterior, se procede a estimar las curvas de O_3 mediante la expresión 7.9. A nivel computacional, lo descrito anteriormente se desarrolla a través de los paquetes *fda* y *fda.usc*.

$$x(t) = \sum_{k=1}^K c_k \phi_k(t) \quad (7.9)$$

- **FMeanDM y FMedDM:** Con las curvas funcionales de O_3 esquematizadas, es posible generar la estimación tanto de la media funcional 6.30 como de la mediana funcional 6.31, estas curvas proporcionan los valores estimados para los valores faltantes. Una vez se cuenta con este par de curvas, el valor faltante viene dado 7.10, donde $d = 1, \dots, 31$ indica el día y $k = 1, \dots, 24$ el tiempo t_k en el cual se encuentra el valor a imputar.

$$x_{dk}^* = x(t_k)_{\text{Media}} \quad o \quad x_{dk}^* = x(t_k)_{\text{Mediana}} \quad (7.10)$$

- **FMeanBIM y FMedBIM:** El algoritmo para el desarrollo de la imputación por estos métodos se describe así:
 1. El primer paso, es llevar a cabo el método de imputación anterior, es decir, reemplazar los valores faltantes en el día d y la hora k por el valor correspondiente de la media (**FMeanDM**) o mediana (**FMedDM**), fijando dichas estimaciones como la semilla del proceso.

2. El segundo paso, se lleva a cabo teniendo en cuenta que los días con información faltante, se encuentran imputados y por lo tanto completos. Por lo anterior, se construye las curva funcionales para estos días utilizando el modelo proporcionado por 7.9.
3. Se calcula el error de la suavización, renombrado como $MPAD_1$ definido por la formula 6.37.
4. A continuación, se reemplaza el valor en la ubicación ausente en el momento d del punto k con el valor obtenido en la curva ajustada.
5. Se vuelve ajustar una curva para los puntos de datos del día, utilizando el modelo proporcionado por 7.9.
6. Se calcula el error de la segunda suavización realizada, renombrado como $MPAD_2$.
7. Se precisa especificar el criterio para detener el algoritmo. Puntualmente, si se cumple 7.11, detenga la imputación; de lo contrario, repita los pasos anteriores del 4 al 7 hasta asegurar la convergencia del método a un error mínimo admisible del 0,005.

$$|MPAD_2 - MPAD_1| \leq 0.005 \quad (7.11)$$

Por lo tanto, al finalizar el algoritmo la imputación viene dada por 7.12, donde $d = 1, \dots, 31$ indica el día y $k = 1, \dots, 24$ el tiempo en el cual se encuentra el valor a imputar, x_i es la curva final obtenida mediante el proceso de iteración, ya sea partiendo de la media o mediana respectivamente.

$$x_{dk}^* = x_i(t_k) \quad (7.12)$$

7.5. Comparación de métodos de imputación

Finalmente, para seleccionar el mejor método para imputar datos faltantes, se utilizaron cuatro indicadores de rendimiento los cuales han sido especificados en la **Sección 6.8.3**, a saber, el coeficiente de determinación (R^2) especificado en 6.40, el Error Absoluto Medio (MAE) definido por 6.39, el índice de concordancia (AI) formulado por 6.38 y la suma de cuadrados del error (SCE) definido en 6.41. Según Yang Lenyap et al. (2015), se debe contar con varios indicadores de desempeño para determinar consistencia en la decisión y comparabilidad. Sin embargo, se le brinda mayor peso al indicador SCE ya que indica que tan cercano o lejano está el valor imputado del real y se le da menor relevancia al R^2 por verse influenciado por la cantidad de variables o factores que participan en la suavización según Norazian et al. (2008).

Capítulo 8

Resultados

8.1. Análisis exploratorio de datos

La finalidad del Análisis Exploratorio de Datos (AED), es examinar los datos previamente a la aplicación de cualquier técnica estadística, siendo una fase importante en toda clase de investigación, puesto que permite conocer y familiarizarse con los datos a utilizar y obtener información preliminar para encontrar la metodología estadística adecuada a aplicar dadas las manifestaciones de los datos. Adicional, el AED permite organizar y preparar los datos, detectando fallas en el diseño y recogida de la información; destacando la posibilidad de dar tratamiento y evaluación de datos ausentes e identificación de casos atípicos (Salvador and Millán, 2003).

Por exigencia de los datos funcionales, es necesario realizar una primera aproximación a los datos originales en busca de la menor cantidad de valores faltantes dado un determinado periodo de tiempo, donde se fijará la información completa de referencia para los niveles de O_3 . Por lo tanto, en este apartado se evidencia el análisis de datos faltantes realizado a la base de datos inicial de la estación Compartir para mediciones de O_3 comprendidas para los años 2015 y 2016 junto a descriptivas de la base de datos final con la cual se ejecutará el proyecto de investigación.

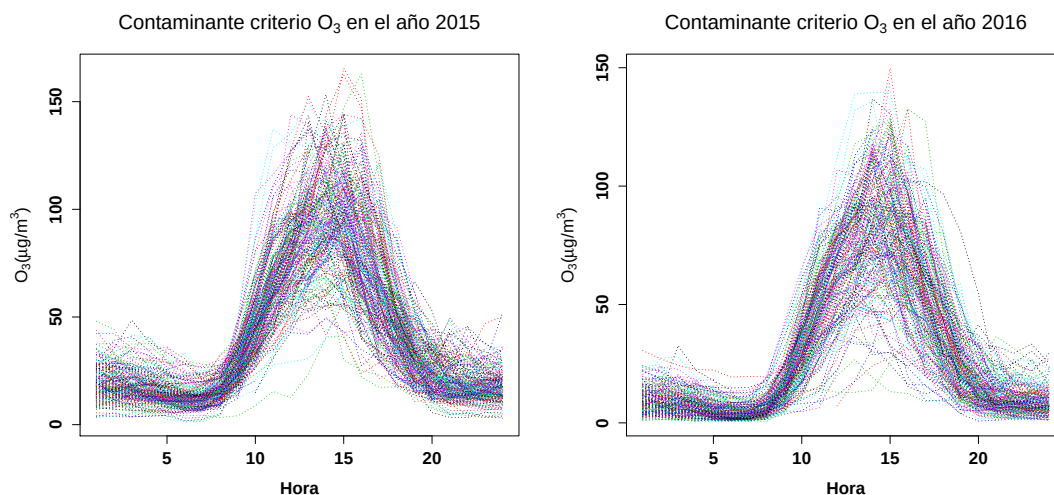


Figura 8.1: Mediciones de O_3 por hora

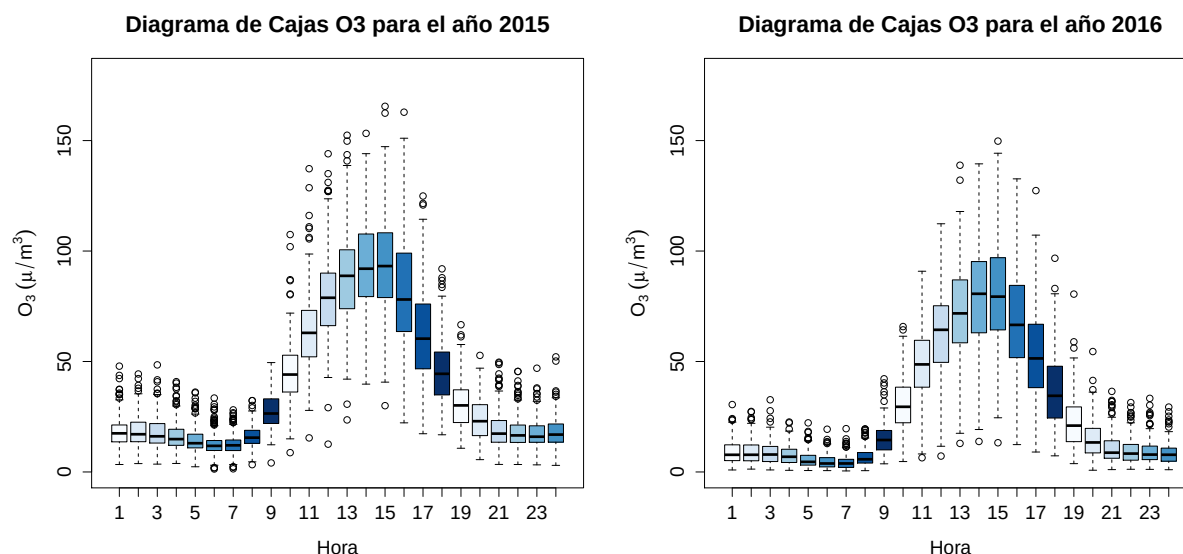


Figura 8.2: Boxplots de concentraciones horarias de O_3 para el año 2015 y 2016 en presencia de datos faltantes

A través de la Figura 8.2 es posible analizar la variabilidad de los datos medidos, donde se establecen similitudes estructurales para los años de estudio. Las concentraciones de O_3 contemplan la forma de la distribución normal asociada a una estructura gaussiana o de campana, donde el posicionamiento de los quantiles esbozan simetría respecto a la mediana de los datos (centro de la distribución) junto a las colas de la misma. Sin embargo, por la longitud de las cajas es posible afirmar que a tempranas horas de la mañana y de la noche la presencia del contaminante en el aire es menos variable en comparación con el nivel de O_3 acotada en el intervalo de tiempo comprendido desde las 10:00 hasta las 18:00, donde los valores extremos son mayores dada la magnitud de los bigotes junto a incrementos notables entre el primer, segundo y tercer quantil.

El comportamiento del contaminante O_3 para ambos años de estudio inicia con cantidades mínimas que van aumentando en el transcurso de la mañana, alcanzando valores máximos entre las doce y una de la tarde. Estos comportamientos se pueden atribuir a que el Ozono se forma en la atmósfera mediante reacciones fotoquímicas en presencia de luz solar y contaminantes precursores, siendo en estos horarios donde la radiación solar tiende a ser mayor y el flujo vehicular se encuentra en su hora pico, es decir, contribuye de manera significativa con elementos formadores del contaminante. En horas donde la temperatura y radiación solar disminuyen, los niveles de O_3 presente en el ambiente del sector de Compartir disminuyen paulatinamente, hasta alcanzar nuevamente valores mínimos en las horas de noche. De manera puntual, las cantidad máxima del contaminante registradas en la estación de monitoreo Compartir son $165.51 \mu g/m^3$ y $149.77 \mu g/m^3$ para el año 2015 y 2016 respectivamente, y valores mínimos de $1.35 \mu g/m^3$ y $0.47 \mu g/m^3$ de forma análoga.

En la Figura 8.3 y 8.4 se evidencia el promedio máximo diario de los niveles de O_3 durante todo el año 2015 y 2016 respectivamente, el cual permite monitorear el panorama real de contaminación en la ciudad de Cali sector Compartir frente a la normatividad ambiental que especifica un nivel máximo de $120 \mu g/m^3$ por hora. Es claro que en todos los meses del año 2015, la cantidad de O_3 presente en el aire supera la norma de forma significativa. Por otra parte, es posible evidenciar la ausencia de mediciones para meses como Febrero y partes de Enero, Marzo, Junio, Agosto, Octubre y Diciembre. Ahora bien, para el año 2016 se puede observar

que la cantidad de O_3 presente en los meses donde se obtuvieron registros, cumplen en mayor proporción con la norma establecida, es decir, tan solo para algunos días de los meses de Enero, Febrero, Julio y Agosto este contaminante fue superior a los $120 \mu g/m^3$ por hora. Adicional a esto, los meses de Septiembre, Octubre, Noviembre y Diciembre del año 2016 presentan valores ausentes para todos los días.

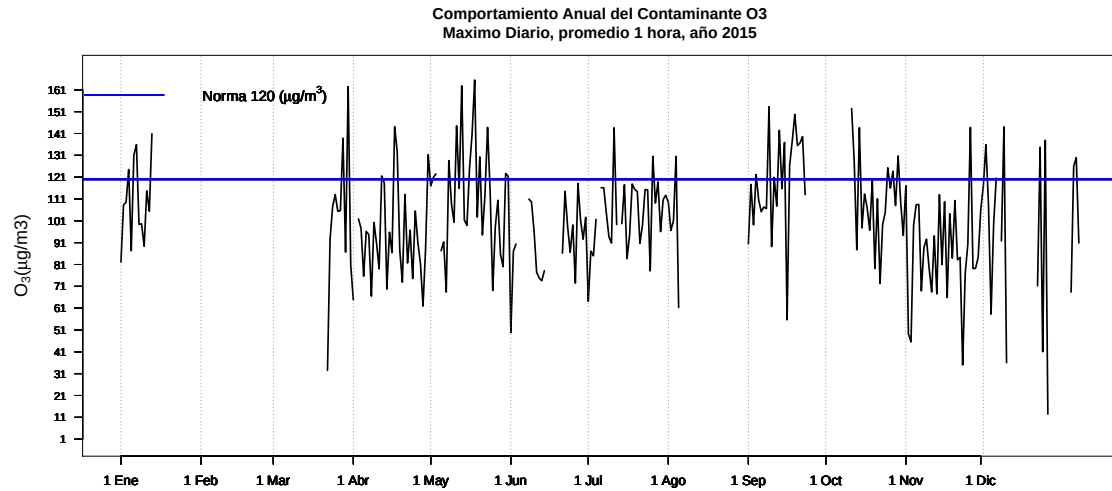


Figura 8.3: Serie para el Máximo Promedio Diario para el año 2015

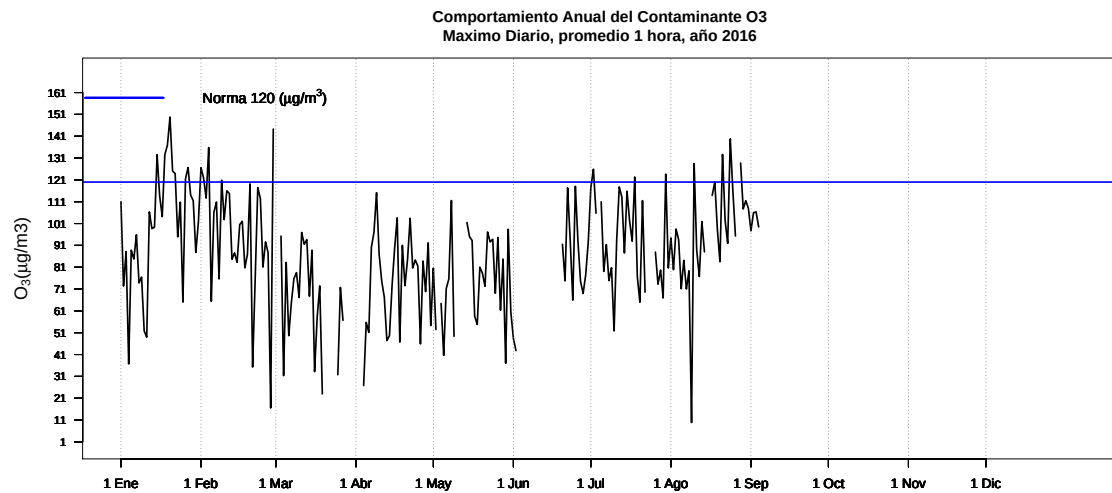


Figura 8.4: Serie para el Máximo Promedio Diario para el año 2016

Lo anterior da paso para ejecutar el análisis de valores ausentes para los años de estudio, donde a partir de la Tabla 8.1 se observan los doce meses del año (filas) y las veinticuatro horas del día (columnas) donde se especifica la frecuencia absoluta de promedios horarios faltantes del contaminante. Este análisis pretende determinar el intervalo de tiempo con menor cantidad de datos faltantes que permitan desarrollar la metodología planteada de datos funcionales. En consecuencia, los meses con mayor cantidad de valores perdidos del contaminante criterio O_3 son Febrero, Marzo, Agosto y Diciembre con porcentajes de pérdida

del 18.93 %, 17.89 %, 14.57 % y 13.96 % respectivamente para el año 2015. De igual manera, los meses con menores valores faltantes respecto a la medición del contaminante son Abril, Mayo, Julio y Noviembre con porcentajes de perdida del 0.06 %, 1.04 %, 1.55 % y 0.99 % respectivamente. En general, la base de datos para el año 2015 cuenta con un porcentaje de información faltante del 40.5 %, correspondiente a un total de 3549 registros ausentes. Las Tablas que evidencian el comportamiento porcentual de los datos perdidos por mes y hora, se encuentra en el Anexo A.

Tabla 8.1: Distribución de datos faltantes por hora de acuerdo a los meses para el año 2015

Mes	Cantidad de datos faltantes																								Total
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	
Enero	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	19	19	19	19	19	437
Febrero	28	28	28	28	28	28	28	28	28	28	28	28	28	28	28	28	28	28	28	28	28	28	28	28	672
Marzo	21	21	21	21	21	21	21	21	21	21	22	21	20	21	21	20	20	20	20	20	20	20	20	20	495
Abril	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	2
Mayo	1	1	1	1	1	1	1	1	1	1	4	1	1	1	2	2	2	2	2	2	2	2	2	2	37
Junio	11	11	11	11	11	11	11	11	11	11	10	10	9	10	10	10	10	10	10	10	10	10	10	10	249
Julio	2	2	2	2	2	2	2	2	2	2	5	3	3	2	2	2	2	2	2	2	2	2	3	3	55
Agosto	27	27	27	27	27	27	27	27	27	27	26	26	26	26	26	26	26	26	26	26	26	26	26	26	635
Septiembre	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	13	12	12	12	12	12	12	12	13	290
Octubre	6	6	6	6	6	6	6	6	6	5	5	5	5	5	6	5	5	5	4	4	4	4	4	5	125
Noviembre	1	1	1	2	1	1	1	1	1	1	2	1	2	2	2	3	2	2	2	1	1	1	1	1	35
Diciembre	21	21	22	22	22	22	22	22	22	21	22	22	22	22	22	22	21	21	21	21	21	21	21	21	517
Total	148	148	149	150	149	149	149	149	149	147	155	147	146	147	149	151	146	146	145	146	145	145	146	148	3549

Realizando el mismo análisis para el año 2016, la Tabla 8.2 evidencia los meses con mayor cantidad de valores perdidos del contaminante criterio O_3 , donde Octubre y Noviembre no presentan registro alguno equivalente al 17.79 % de la información perdida del año, seguido por los meses de Septiembre y Noviembre con el 17.22 % de datos faltantes. De igual manera, los meses con menores datos perdidos respecto a la medición del contaminante son Enero, Febrero, Abril y Julio con porcentajes de información faltante del 0.02 %, 0.65 %, 0.98 % y 2.2 % respectivamente. De forma global, el porcentaje de información perdida para el año 2016 es del 47.6 % que representa 4181 mediciones ausentes.

Tabla 8.2: Distribución de datos faltantes por hora de acuerdo a los meses para el año 2016

Mes	Cantidad de datos faltantes																								Total
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	
Enero	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Febrero	1	1	1	1	1	1	1	1	1	1	2	1	2	1	1	2	1	1	1	1	1	1	1	1	27
Marzo	13	13	13	13	14	14	14	14	14	14	15	13	14	14	14	16	14	13	13	13	13	14	14	14	329
Abril	2	2	2	2	2	2	2	2	2	4	2	2	1	1	1	3	2	1	1	1	1	1	1	1	41
Mayo	8	9	10	10	10	10	10	10	10	9	7	8	8	7	7	8	8	8	8	8	8	9	9	9	208
Junio	17	17	17	19	19	19	19	19	18	18	18	18	17	17	17	16	16	16	17	17	17	17	17	16	418
Julio	4	4	4	4	4	4	4	4	4	4	3	5	3	3	3	3	4	4	4	4	4	4	4	4	92
Agosto	6	6	6	7	8	8	8	8	7	5	4	4	4	4	4	4	4	4	5	5	5	6	7	8	137
Septiembre	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	720
Octubre	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	744
Noviembre	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	720
Diciembre	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	744
Totales	173	174	175	178	179	180	180	180	178	177	174	173	171	169	169	174	171	169	171	171	171	174	175	175	4181

De esta forma, el mes que presenta menor cantidad de valores perdidos es Enero del año 2016. Por lo tanto, la información medida en dicho mes será utilizada como base de datos de referencia después de estimar el único valor ausente. Es importante resaltar que los datos del mes es cuestión son suficientes para llevar a cabo un ADF.

8.1.1. Estimación de dato faltante por regresión lineal

Según Urrutia et al. (2010), el uso de los modelos de regresión lineal es adecuado y recomendado para un conjunto de datos con proporción de información faltantes menor o igual al 20 % que a su vez, tengan una relación lineal. Adicional, el autor recomienda el uso de modelos de regresión para variables ambientales, lo cual es preciso para el caso de estudio.

Por lo tanto, a través de un modelo de regresión lineal se estimo la hora faltante dado el día que presenta la ausencia, a saber, hora 11:00 del día 19 de Enero del 2016. La selección de la variable independiente se realizó identificando las horas que presentaran datos completos y mayor correlación con la hora donde se encontró el datos ausente, obteniendo como resultado la hora 12:00 con una correlación de 0.891 con lo cual, se ajustó el modelo de predicción y posteriormente se evaluó la significancia de los parámetros. Para el modelo obtenido, no se encuentra evidencia suficiente para rechazar la hipótesis nula de $H_0 : \beta_0 = 0$ con un valor-p del 0.171, es decir, el modelo de regresión que permite la estimación del valor ausente no cuenta con parámetro de intercepto, careciendo de sentido al análisis ya que ante la ausencia de O_3 en la hora 12:00 representa la ausencia directa de O_3 en la hora 11:00, como se muestra en la Figura .1 localizada en el Anexo A. Por lo tanto el modelo final es $Y_{11,19} = 0.7306653X_{12,19}$ con un R^2 ajustado del 0.9762, donde el coeficiente $\hat{\beta}_1$ indica que dada la cantidad de Ozono presente en la hora 12:00 del día 19 se puede esperar que el contaminante en la hora anterior sea mayor en una medida de $0.7306653 \mu g/m^3$.

Tabla 8.3: Análisis de varianza para la hora 11 para el mes de Enero 2016

	Gl	Suma cuadrados	Cuadrados medios	F	Pr(>F)
Hora 11 O_3	1	69073.65	69073.65	1229.10	0.0000
Residuales	29	1629.75	56.20		

En la Tabla 8.3 se presenta el análisis de varianza para el modelo obtenido, donde se cuenta con una desviación estándar del 7.4966 y un R^2 ajustado igual a 0.9769. El anterior coeficiente de determinación indica que aproximadamente el 98 % de la variabilidad presente en la cantidad de O_3 en la hora 11 : 00 queda explicada por el modelo de regresión obtenido. Por último, se realizó una demostración de la precisión en las estimación obtenidas por el modelo, así:

Tabla 8.4: Estimaciones de datos faltantes para la hora 11:00 del mes de Enero en el año 2016

Día	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Valor real	55.63	22.70	41.25	9.69	48.53	6.44	27.89	34.28	41.07	23.53	25.33	41.78	47.88	54.91	61.81	68.64
Valor imputado	45.27	26.28	40.82	22.05	51.33	15.82	34.99	50.24	43.96	26.97	27.44	41.48	38.89	59.92	60.79	51.43
Día	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	
Valor real	56.85	59.85	NA	58.01	51.00	59.17	42.64	48.78	49.59	55.91	62.46	58.97	51.26	48.47	63.72	
Valor imputado	53.44	54.86	52.21	58.40	58.26	54.00	53.02	51.63	43.79	48.22	62.31	51.66	58.43	52.19	54.54	

En la Tabla 8.4 se observa los días del mes de Enero del 2016, los *Valores reales* obtenidos en las mediciones realizadas (sin modificación alguna) y los *Valores estimados* producto de la demostración realizada, teniendo como resultado un modelo por día. Los distintos valores de la ultima fila fueron estimados de la siguiente manera, se realizó la estimación al día 19 hora 11:00 reconocida como un dato ausente, en la cual se presento una medición de O_3 de 52,21. Posteriormente se procedió a retirar la observación de la fila *Valor real* atribuida al día 1 y se estimo su valor el cual fue 45.27, una vez hallado éste, se retiró la observación 2 y se estimó su valor, y así sucesivamente para cada día correspondiente. Esto con el objetivo de verificar qué tan buena era la estimación realizada mediante la ecuación de regresión, es decir, verificar si se alejaba mucho o poco de su valor real. Un buen ejemplo de esto, es lo encontrado el día 12 y 20 de Enero, donde el valor estimado es relativamente próximo al valor real y por el contrario, el día 4 del mismo mes refleja un valor estimado lejano del valor real.

Descriptivas para datos completos

En la figura 8.5 se representa el promedio de los registros horarios de O_3 de acuerdo a los días de la semana, en la cual se analiza el comportamiento individual de cada día. Toda la composición de la semana inicia con bajos niveles de O_3 que alcanzan un punto mínimo a las 7:00 de la mañana, excluyendo el día lunes que presenta un incremento significativo en las primeras horas del día, lo que deja entre ver que dicho promedio

se encuentra influenciado por un valor extremo. Posteriormente, al analizar la hora 15:00 donde se presentan los máximos para todos los días de la semana es posible determinar que el día lunes cuenta con los menores niveles de Ozono mientras que los miércoles, jueves y viernes experimentan las mayores concentraciones del contaminante. En tal punto, se produce un cambio decreciente en las cantidades de O_3 para todos los días de la semana.

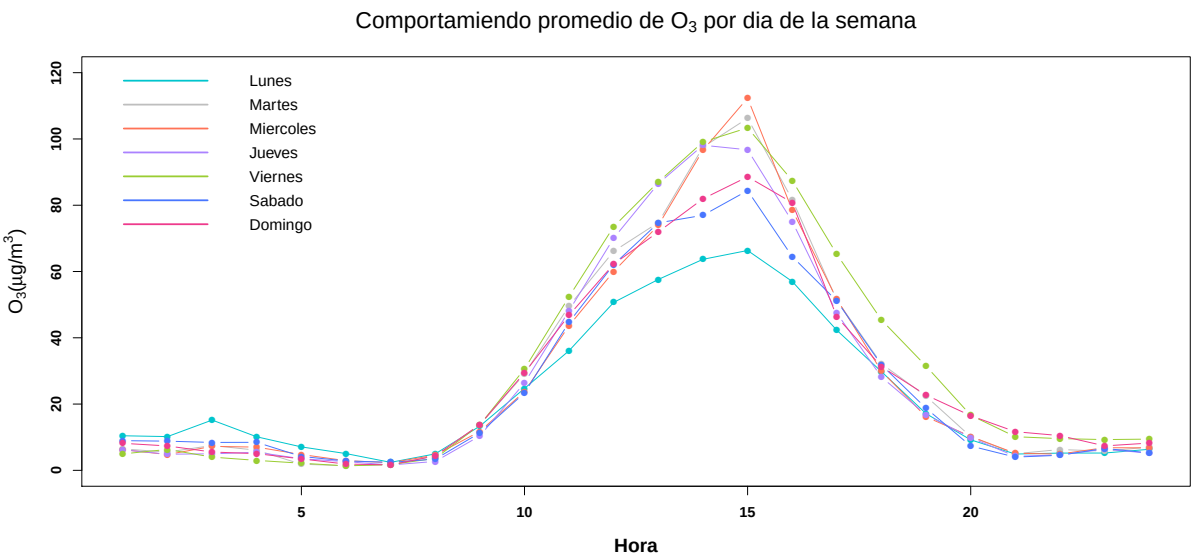


Figura 8.5: Comportamiento promedio horario de O_3 de acuerdo a los días de la semana - Enero 2016

En la gráfica 8.6 se evidencia el calendario del mes a trabajar con los promedios diarios del contaminante O_3 . La intensidad del color rojo aumenta conforme incrementa la cantidad de Ozono en el aire, por lo cual el día viernes 15 de Enero reporta altos niveles del gas medido en la estación Compartir, mientras que por otro lado el día lunes 4 de Enero se reporta la menor cantidad de O_3 .

Además, a partir del día 15 se da un incremento marcado y significativo durante todo el mes. Esto es atribuido a la culminación del periodo de vacaciones e inicio del periodo laboral, dejando como consecuencia incremento en la movilidad de la ciudad y elevando los niveles de O_3 en el ambiente.

De esta forma, se cuenta con mayor conocimiento de la base de datos a trabajar y da paso a la ejecución de la metodología planteada “Datos funcionales”.

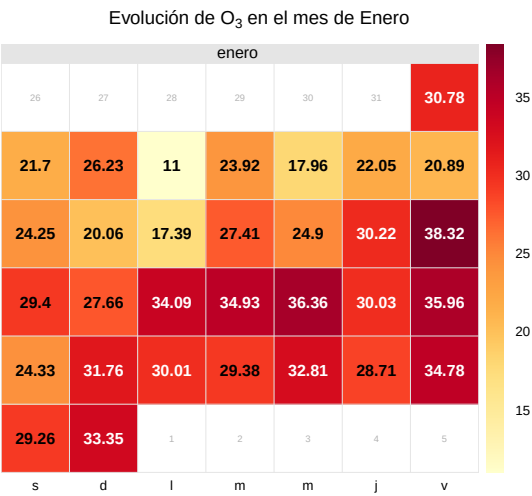


Figura 8.6: Comportamiento promedio diario de O_3 para el mes de Enero

8.2. Construcción de escenarios

El proceso de imputación provee una base de datos con 31 días, cada uno con sus 24 registros. Es así como se obtiene la base de referencia completa con un total de 744 mediciones, permitiendo la creación o construcción de los diferentes escenarios de valores ausentes. Lo anterior se ejecuta a la luz de lo descrito en la Sección 7.3, proponiendo de esta forma seis diferentes combinaciones de valores aleatorios perdidos junto a espacios consecutivos faltantes con diversas magnitudes.

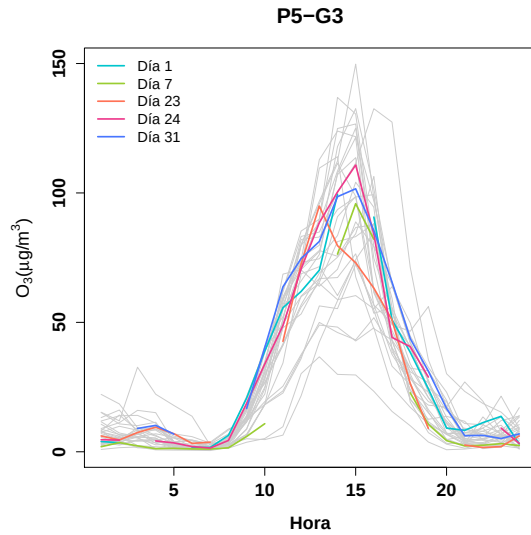


Figura 8.7: Contaminación del 5% y 5 datos faltantes consecutivos

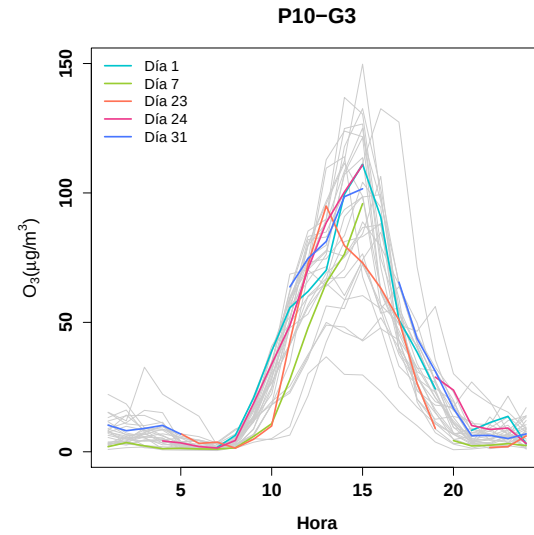


Figura 8.9: Contaminación del 5% y 5 datos faltantes consecutivos

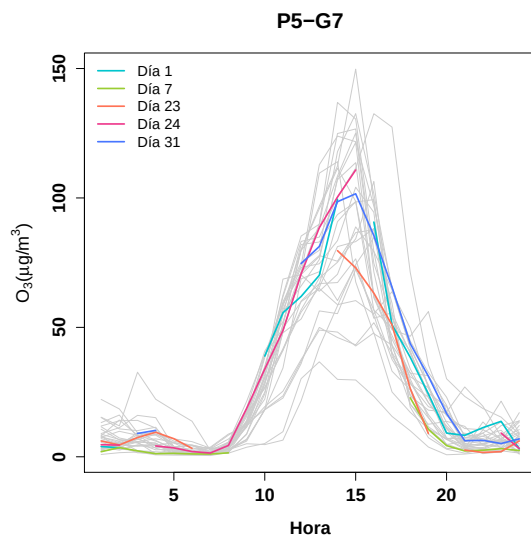


Figura 8.8: Contaminación del 5% y 7 datos faltantes consecutivos

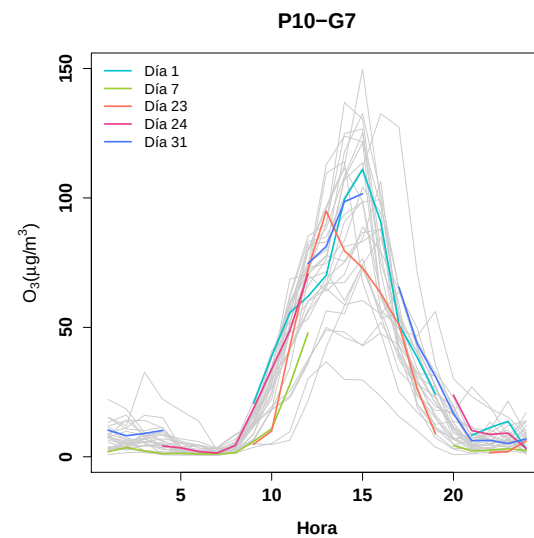


Figura 8.10: Contaminación del 5% y 7 datos faltantes consecutivos

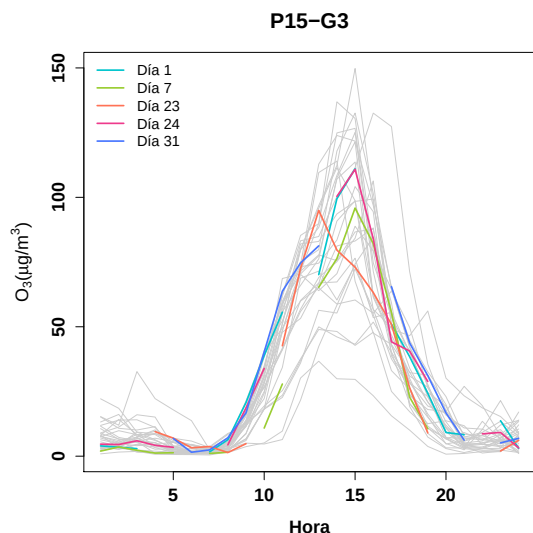


Figura 8.11: Contaminación del 5 % y 5 datos faltantes consecutivos

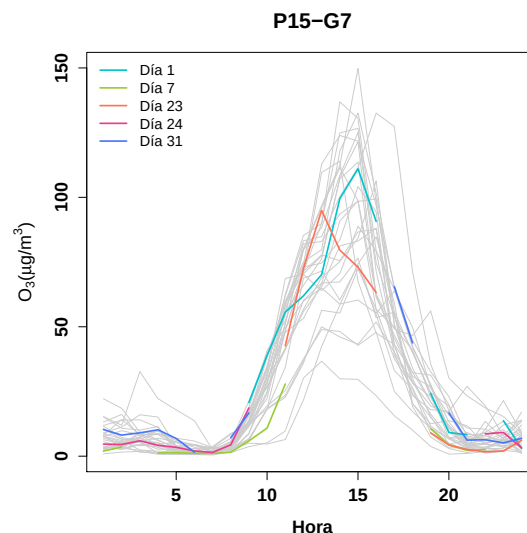


Figura 8.12: Contaminación del 5 % y 7 datos faltantes consecutivos

Es así como se fija la estructura de información perdida que será imputada por los diversos métodos a evaluar, generando cinco días aleatorios, a saber, día uno, siete, veintitrés, veinticuatro y treinta y uno existentes en el mes de Enero del 2016 en los cuales, se genera la contaminación de valores perdidos con los diferentes patrones.

Se hace hincapié en la forma como los datos son contaminados por cada uno de los escenarios propuestos. De forma específica, se toma cada uno de los días y por aleatoriedad se extrae la cantidad de datos sin tener en cuenta ningún tipo de secuencia, permitiendo que el resultado no sea previsible más que en razón de la intervención del azar. Por ende, cada día contaminado cuenta con diversos patrones con igual porcentaje de pérdida. Posteriormente, se procede a finalizar la contaminación del set de datos extrayendo de forma aleatoria y secuencial una cantidad n de mediciones de O_3 por hora definida en secciones anteriores. De esta forma, se obtiene como resultado los set de datos contaminados bajo seis escenarios. Al pasar de un escenario a otro, se espera que la cantidad de datos perdidos incremente conforme la aleatoriedad lo permita, obteniendo como máximo un 46 % aproximado de datos perdidos por día y un mínimo del 4.16 % de ausencia de mediciones diarias en todos los escenarios.

De forma general, los patrones pueden proporcionar diversos escenarios no contemplados. Por ejemplo, la contaminación que consta del 15 % de datos perdidos de forma aleatoria junto a una longitud consecutiva de siete datos (Figura 8.12), donde se definen once horas como cantidad máxima posible de valores ausentes (15 % de 24 equivale aproximadamente a 4 registros más siete datos consecutivos) y un mínimo de siete horas para el caso particular en que los datos aleatorios coincidan con los valores del espacio faltante consecutivo. Sin embargo, por aleatoriedad se puede observar como el día veinticuatro cuenta con un espacio consecutivo creado por los cuatro registros aleatorio y el espacio de siete datos horarios, obteniendo así un espacio consecutivo final de once registros ausentes. Por lo tanto, la estructura de valores ausentes en cada día es diferente y enmarca diversos retos para las técnicas de imputación.

8.3. Construcción de datos funcionales

El proceso de construcción de escenarios provee una base de datos con 26 días completos, cada uno con sus 24 registros. Es así como se obtiene la base de datos completa con un total de 624 mediciones, permitiendo la creación o construcción de los datos funcionales por día a través del método de suavización con rugosidad penalizada, este enfoque permite un control más fino sobre la cantidad de suavizaciones teniendo en cuenta la indexación de la variable observada, en este caso se tienen promedio horarios del contaminante indexados de $t = 1, 2, \dots, 24$, es decir, las observaciones discretas en las 24 horas del día se convierten en una función que permite su evaluación en cualquier momento en el espacio considerado.

La estimación de las curvas diarias se realiza por medio de un sistema de expansión de la función base, siendo esta una combinación lineal de K funciones que conforman la base. Las funciones de la base deben tener características estructurales que permitan un buen ajuste, esto hace posible lograr una aproximación satisfactoria usando un número optimo y pequeño de funciones básicas. Por lo tanto para el caso de investigación, se utiliza una combinación lineal de K funciones de base Bspline para representar las curvas diarias de O_3 , esto dadas las propiedades de fácil ajuste (polinomios por pieza) y la alta utilidad en datos no periódicos según Ramsay and Silverman (2005). Sin embargo, para estimar las curvas funcionales se debe fijar el número optimo de funciones junto al parámetro de penalización λ , entendiéndose por optimo a la minimización de la distancia entre el dato estimado y el observado. Esto es posible a través del criterio de validación cruzada generalizada, la cual calcula la suma de cuadrados del error para diferentes valores de ambos parámetros dentro del espacio considerado.

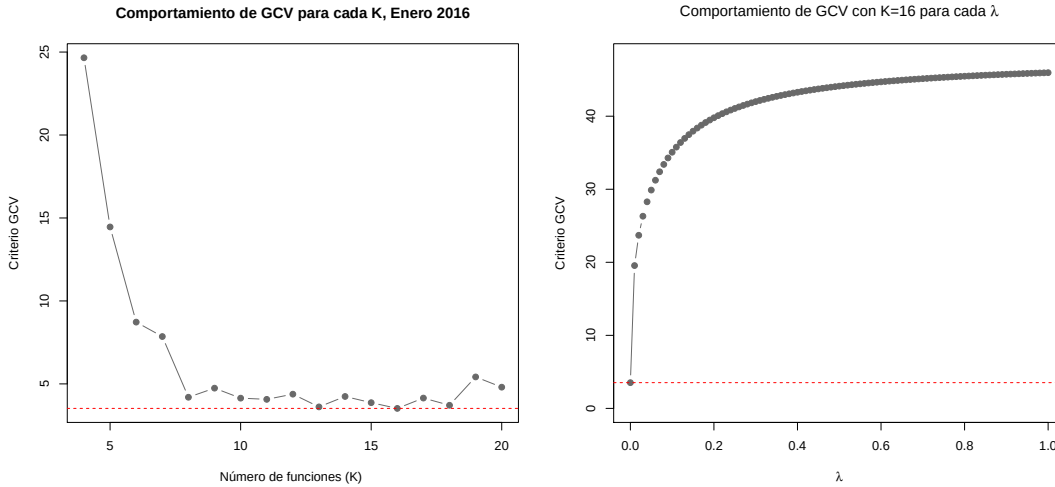


Figura 8.13: Validación cruzada generalizada para parámetros K y λ

En el lado izquierdo de la figura 8.13 se evidencia la estructura decadente del GCV, el cual a medida que aumentan el número de funciones alcanza el valor mínimo en $K = 16$ con una valor de 3.523551, definiendo de esta forma el número óptimo de funciones de la base. Por otra parte, el panel ubicado en el lado derecho visualiza el parámetro de penalización que minimiza el criterio GCV, a saber, $\lambda = 0$, siendo congruente con la baja rugosidad de las mediciones de O_3 diarias. En consecuencia, se cuenta con los parámetros requeridos para suavizar las curvas del contaminante usando el método de mínimos cuadrados conocido como suavizado de regresión.

Las funciones Bsplines se construyen dividiendo el intervalo de observación $[1, 24]$ en subintervalos, con límites en puntos llamados nodos. En cualquier subintervalo, la función Bspline es un polinomio de grado u orden fijo,

pero la naturaleza del polinomio cambia a medida que uno pasa al siguiente subintervalo. Dicho lo anterior, se define la base B-spline de orden 4, la cual cuenta con una unidad superior al grado ya que involucra el término constante, es decir, el grado más alto presente en el polinomio es 3. Adicional, una base se define en términos de un conjunto de nodos, estos están relacionados con los puntos de ruptura donde los polinomios vecinos están obligados a tener un cierto número de derivadas coincidentes asegurando de esta forma que la unión entre ellos sea de forma suave. Para el caso de interés, se definen los nodos equidistantes de esquema simple por punto de ruptura, el cual hace coincidir los valores de la función y las dos primeras derivadas, sin incluir los nodos extremos. Según Ramsay and Silverman (2005), el número de funciones de la base es equivale al orden junto con el número de nodos, de ahí se definen 12 puntos de llegada y partida de cada polinomio como se evidencia a continuación en la Figura 8.14:

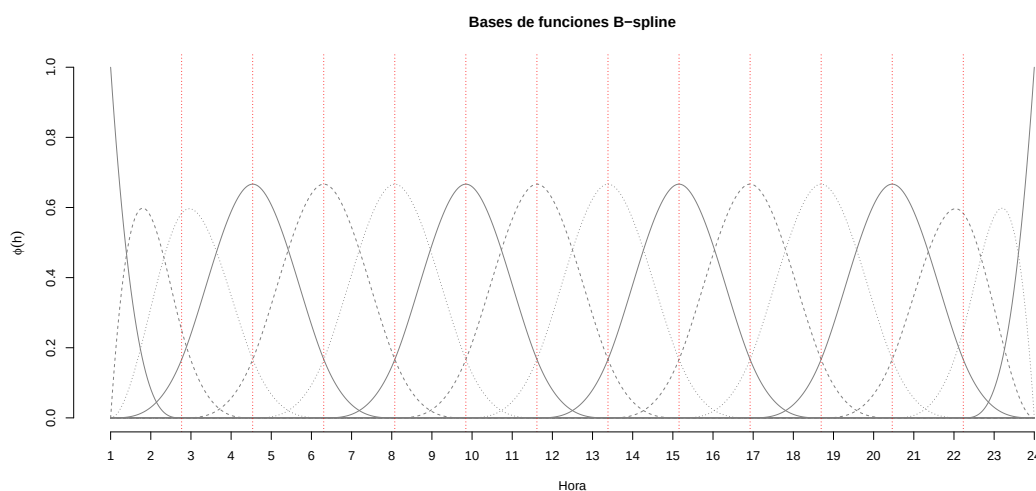


Figura 8.14: Base B-spline de orden cuatro con sus respectivos nodos

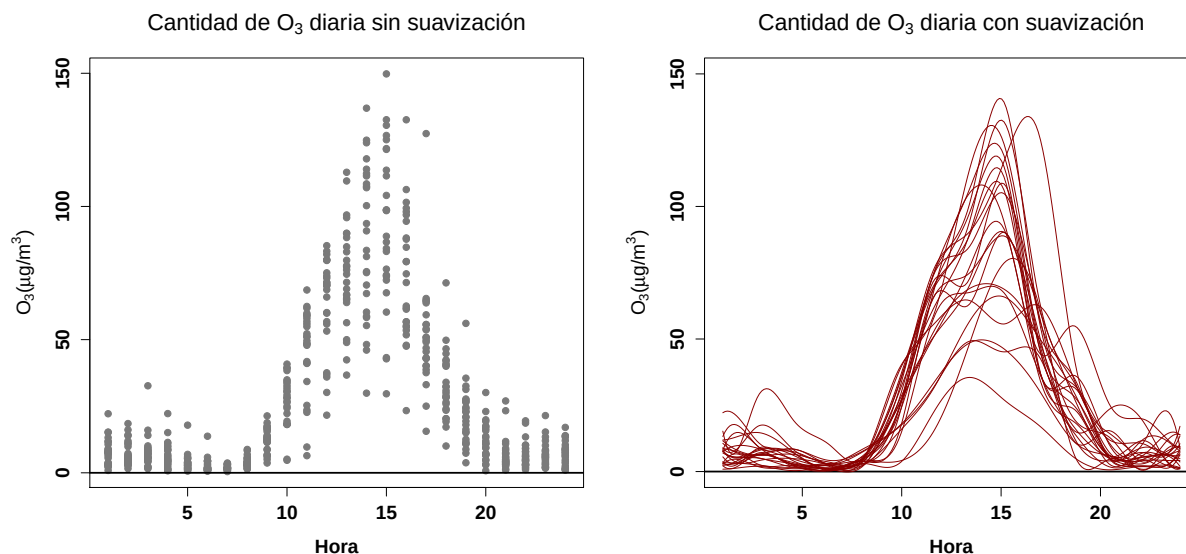


Figura 8.15: Construcción de datos funcionales para mediciones de O_3

De esta forma, se cuenta con los elementos necesarios para ejecutar la construcción de las curvas funcionales, convirtiendo datos discretos unificados por día evidenciados en el panel izquierdo de la Figura 8.15, es decir, se observan los registros en bruto del contaminante trabajado. Estos presentan picos donde las derivadas son inexistentes razón por la cual, se utiliza la suavización Bspline que convierte el dato discreto en continuo con propiedades esenciales como la existencia de derivadas, posibilidad de modelamiento, entre otras. Es así como en el panel derecho se visualizan las curvas funcionales de O_3 medidas en la estación Compartir en Enero del año 2016. La expresión matemática se esclarece en la sección 6.5.3, sin embargo, las matrices de coeficientes y funciones se anexan al final del documento. Es importante aclarar que este proceso se ejecuta con los días que cuentan con la mediciones diarias completas, es decir, se extraen los cinco días que fueron contaminados.

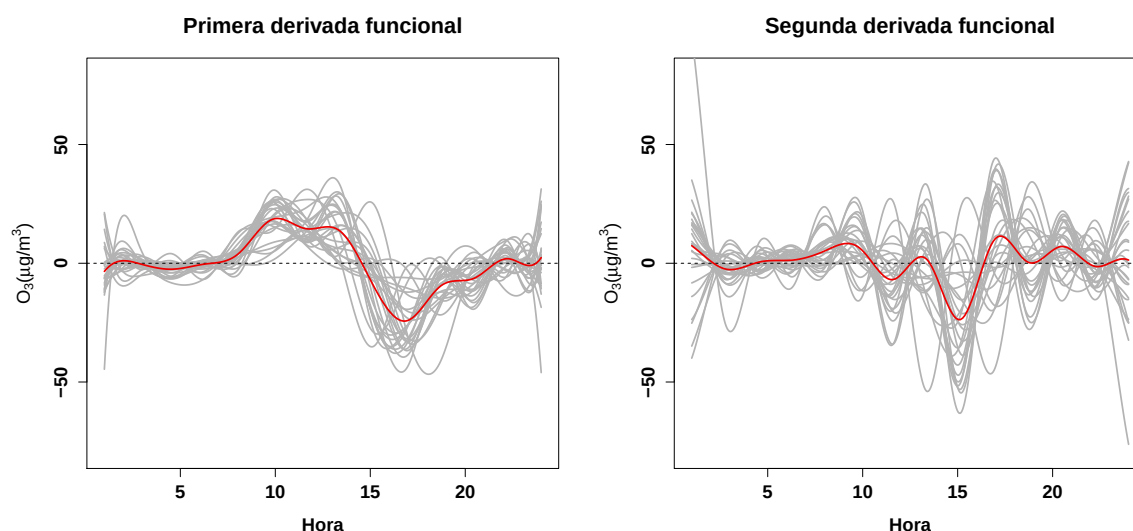


Figura 8.16: Primera y segunda derivada de curvas funcionales de O_3

Por último, la figura 8.16 visualiza la estructura funcional cuadrado integrable, permitiendo obtener la primera y segunda derivada de cada curva funcional, donde la función en tonalidad roja demarca la curva promedio de las primeras y segundas derivadas. La información extraída de la primera derivada representa la pendiente de la recta tangente de la curva funcional en cada punto del espacio, exaltando los valores extremos. De manera puntual, cada intercepción que se presente entre la curva roja con el eje de la abscisa corresponde a valores máximos o mínimos relativos donde la recta tangente a estos equivale a cero, así, se cuenta con un máximo relativo en la hora 15 junto a varios mínimos relativos en horas tempranas de la mañana al igual que en altas horas de la noche. Por otro lado, la segunda derivada representa la aceleración de la curva, de tal forma que en las horas intermedias del día (espacio de tiempo comprendido entre las 10 y 16 horas), la cantidad de O_3 se presenta en mayor medida, lo que aumenta la aceleración en la que se manifiesta el contaminante en el aire.

8.4. Análisis descriptivo de datos funcionales

Al cambiar del espacio discreto al espacio de las funciones continuas se requiere analizar y explorar el comportamiento medio de las funciones junto a la variabilidad a lo largo del día. Por lo tanto, posterior a la construcción de las curvas de O_3 diarias se extiende el análisis descriptivo clásico al campo funcional, de esta manera se construirán la curva de la media, mediana, desviación estándar y correlación por día.

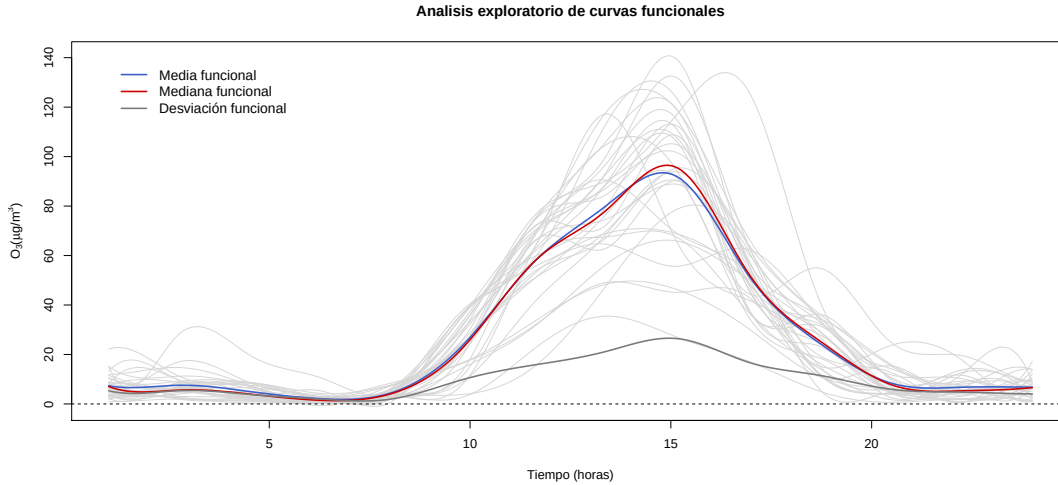


Figura 8.17: Datos funcionales de O_3 con curva media y curva de desviación estándar

En la figura 8.17 se construyen las funciones que describen el comportamiento global de las curvas de O_3 en la estación de Compartir. Como se observa, la media funcional inicia con un leve aumento del contaminante que no sobrepasa los $10 \mu g/m^3$, presentando un decaimiento a las cinco de la mañana y alcanzando su valor mínimo al rededor de las ocho. Posterior, incrementa significativamente hasta alcanzar su nivel máximo de contaminación media ($90 \mu g/m^3$) alrededor de las quince horas del día, para decaer sustancialmente hasta el mínimo local en las horas de la noche. Como es el caso en el espacio discreto, la curva media se deja influenciar por curvas extremas o atípicas (en esta ocasión no se denotaran las curvas como atípicas ya que, se requiere profundizar en conceptos y pruebas que no son de interés para la presente investigación), por lo cual se construye la curva mediana que se caracteriza por no dejarse influenciar por alguna medida diaria extrema. Por lo tanto, el comportamiento de la curva mediana al iniciar el día se encuentra por debajo de la curva media, dejando entre ver que existen días en los cuales el contaminante se encontraba en cantidades superiores o extremas a tempranas horas del día, ocasionando un jalonamiento de la curva media, caso similar ocurre al finalizar el día. Por otra parte, en las horas intermedias del día se entrecruzan las curvas mediana y media, reflejando curvas extremas tanto por encima como por debajo, es decir, existen días donde los niveles de ozono fueron extremos o mínimos.

Por otra parte, la curva que representa la desviación estándar indica poca variabilidad en horas de la mañana y en la noche. Sin embargo, entre las diez y las veinte horas de día, se observa un incremento en la variabilidad de las curvas diarias, es decir, durante este lapso de tiempo las funciones diarias tienden a presentar diversas cantidades de O_3 en el ambiente, lo cual se ve reflejado en la desviación estándar, que tiene como objetivo principal medir las discrepancias entre los comportamientos diarios del contaminante trabajado.

Por lo tanto, el mes de Enero del año 2016 se caracterizó por la entrada en la fase de madurez del Fenómeno del Niño, que originó que las temperaturas en el Valle del Cauca aumentarán hasta cuatro grados centígrados en el primer bimestre del 2016 según el MinAmbiental (2016a). Por lo anterior, con aumentos de temperatura

vienen aumentos de radiación solar, sumado la culminación del periodo vacacional en Santiago de Cali, conforman el ambiente propicio para la generación del contaminante O_3 . Este panorama, se refleja claramente en las curvas funcionales trabajadas en horas donde es mayor la radiación solar junto a la congestión vehicular en la ciudad, tiene como consecuencia principal el aumento significativo del O_3 . Esto es captado por la estación de monitoreo de Compartir, que solo dislumbra una muestra de la calidad del aire existente en toda la ciudad.

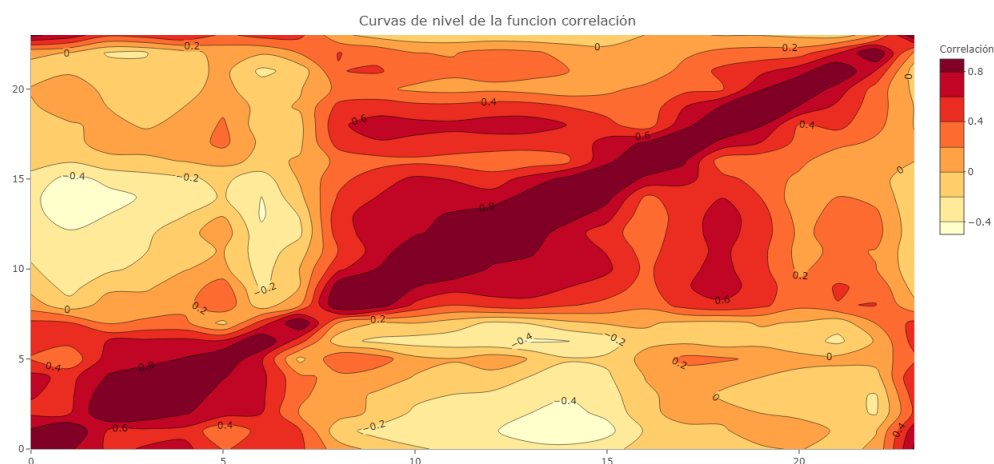


Figura 8.18: Mapa de calor para la correlación horaria

De forma intuitiva se puede observar en la Figura 8.18 la correlación horaria de las curvas funcionales. Se sabe que la más alta correlación se presenta en la diagonal, es decir, la relación existente entre cualquier hora del día con ella misma. Sin embargo, la periferia de la diagonal es donde se observan relaciones de interés. Se puede afirmar que las curvas se encuentran correlacionadas a través de la cercanía horaria. Un ejemplo claro de esto es la hora cinco del día, la cual cuenta con una mayor relación con horas que se aproximen a ella, tales como las cuatro o las seis del día. Otro ejemplo de esto es la hora diez y ocho del día, donde horas tales como la una o las cinco de la mañana no se encuentran relacionadas con el comportamiento del O_3 para la hora en cuestión, por otro lado, las horas comprendidas entre las diez y las veinte del día, cuentan con una alta correlación con esta.

8.5. Imputación de datos faltantes

Con el proceso de construcción de escenarios faltantes y estimación de las curvas funcionales culminado, el primer paso en el análisis de imputación fue generar seis patrones con 1000 conjuntos de datos incompletos para cada patrón, a partir del conjunto de referencia para garantizar la consistencia en los resultados, lo cual da paso a la imputación o estimación de valores perdidos utilizando métodos clásicos y ADF como sigue a continuación.

8.5.1. Métodos clásicos

Los métodos clásicos son contextualizados en la presente investigación como técnicas de fácil aplicabilidad y utilizadas con mayor frecuencia por investigadores en diversos campos, con el único objetivo de obtener una base de datos completa que permita la toma de decisiones confiables y certeras. Es así como de un gran universo de posibles técnicas, la imputación por media no condicional, por media antes y después y por regresión determinística son escogidas para representar la familia de metodologías clásicas.

Las ventajas de realizar el ejercicio de estimación de información faltante es obtener un conjunto de datos completos, reduciendo el sesgo debido a la no respuesta y la imputación opera sobre los datos, de forma que los resultados obtenidos por los diferentes análisis son consistentes. Sin embargo, las principales desventajas se centran en que el futuro análisis no distingue entre las imputaciones y los datos reales. Además los valores imputados pueden ser buenas estimaciones pero no son datos reales y no podemos asegurar una mejora en el sesgo respecto del sistema de datos incompletos. Es importante destacar, que lo anterior es el día a día en las investigaciones realizadas pero en esta oportunidad, conocemos los datos reales y es posible determinar técnicas que bajo ciertas condiciones pueden ser confiables y poderosas en su rol de estimar información. Con esto en mente, se llevará a cabo la explicación de las minucias de cada técnica vista desde un solo escenario propuesto, sin embargo, el mismo procedimiento es aplicado para los diversos escenarios propuestos para la base de datos.

Día	Hora	o3.1	o3.2	o3.3	o3.4	o3.5	o3.6	o3.7	o3.8	o3.9	o3.10	o3.11	o3.12	o3.13	o3.14	o3.15	o3.16	o3.17	o3.18	o3.19	o3.20	o3.21	o3.22	o3.23	o3.24
d1	4.00	3.50	2.90	NA	2.70	NA	1.70	6.30	20.70	NA	NA	NA	NA	NA	NA	NA	NA	51.00	38.50	24.30	9.10	NA	11.30	13.70	3.00
d2	9.00	11.20	16.00	15.10	3.80	4.60	2.90	2.80	6.80	18.20	22.70	36.00	56.40	55.30	72.40	56.70	49.80	41.00	22.90	9.50	3.80	1.20	1.90	8.40	
d3	0.90	1.60	1.80	0.70	0.60	0.60	0.50	1.60	12.10	30.40	41.20	55.90	65.50	81.20	83.10	88.20	43.00	23.50	18.00	17.30	27.00	19.50	7.90	1.00	
d4	11.90	13.90	14.10	11.10	3.90	1.80	1.50	1.40	3.70	5.10	9.70	30.20	36.70	29.90	29.70	23.40	15.60	10.10	3.80	0.80	1.10	1.80	1.70	7.30	
d5	3.00	4.10	7.30	3.30	1.60	1.10	1.20	2.80	6.10	18.30	48.50	70.30	63.90	70.50	88.80	79.30	42.90	24.00	15.00	6.70	2.90	2.30	2.30	1.40	
d6	3.80	4.20	9.30	6.40	2.80	2.30	3.40	2.70	5.00	4.70	6.40	21.60	42.40	58.30	74.20	84.70	59.30	19.20	7.40	2.90	1.40	2.90	3.00	7.80	
d7	2.00	3.60	2.30	1.20	1.30	1.10	1.10	1.50	5.90	10.90	27.90	47.90	65.20	76.40	95.90	NA	NA	NA	NA	NA	NA	NA	NA	2.40	
d8	3.10	6.00	2.40	1.90	1.60	1.20	1.70	2.40	6.60	18.20	34.30	68.80	71.50	60.30	73.80	54.70	33.10	25.80	8.80	3.20	3.20	3.00	6.80	1.90	
d9	8.80	6.70	5.30	2.70	2.90	1.10	2.10	4.00	16.20	26.90	41.10	60.20	66.90	67.30	76.50	53.50	55.70	20.30	11.00	4.90	2.50	10.30	21.50	9.00	
d10	22.20	18.50	9.30	8.70	5.30	4.50	3.00	6.80	12.60	18.30	23.50	36.90	50.00	46.10	43.30	51.80	40.10	20.30	12.60	13.70	11.80	4.40	8.30	14.00	
d11	7.50	9.90	8.30	6.00	3.50	2.90	3.70	8.80	11.80	19.40	25.30	37.60	48.90	48.20	42.80	47.90	25.00	17.20	10.80	6.90	3.40	3.60	5.80	9.50	
d12	15.20	10.80	10.10	11.40	1.90	1.40	0.90	5.30	16.90	26.40	41.80	56.80	69.00	75.20	84.20	106.30	50.90	30.00	15.40	4.70	6.00	9.10	3.20	12.10	
d13	8.10	3.10	4.60	4.60	5.40	2.00	1.50	2.80	11.60	25.00	47.90	53.20	66.70	90.80	98.70	56.70	37.60	26.50	11.60	6.70	9.10	6.30	8.60	4.90	
d14	7.60	6.10	7.40	10.20	3.40	1.70	1.80	3.30	13.00	28.20	54.90	82.00	83.00	93.60	98.40	99.40	50.80	28.40	19.40	16.60	4.70	2.90	5.20	8.70	
d15	3.30	3.10	2.30	1.30	1.30	0.90	1.80	3.00	13.70	34.30	61.80	83.20	84.90	100.30	111.50	132.50	127.40	71.30	35.50	17.60	8.50	7.90	5.90	3.60	
d16	7.70	5.40	5.80	5.30	1.20	0.90	1.30	4.90	17.30	32.60	68.60	70.40	77.40	107.00	113.60	61.40	45.50	30.90	26.70	10.90	4.20	3.20	2.10	6.60	
d17	2.80	4.10	1.60	1.30	1.00	0.80	0.80	1.70	8.90	24.30	56.80	73.10	74.60	83.50	104.10	94.40	38.70	28.30	23.00	10.90	3.20	13.50	6.40	1.20	
d18	7.00	5.00	5.60	1.10	3.00	1.70	2.00	4.00	16.60	34.90	59.90	75.10	79.80	117.90	132.60	101.60	64.60	42.10	27.40	13.30	4.30	5.00	6.20	5.90	
d19	5.50	3.80	5.70	5.20	2.10	1.30	1.40	5.60	18.80	40.80	52.20	71.50	90.00	136.90	130.50	79.30	64.60	42.40	27.70	9.30	7.30	9.20	13.50	7.60	
d20	10.70	6.50	6.60	8.70	6.90	5.80	3.30	8.40	12.90	31.70	58.00	79.90	96.80	112.30	149.80	98.00	55.70	40.00	24.60	19.30	5.10	5.90	12.10	13.80	
d21	11.60	7.30	5.90	6.90	5.60	3.50	1.70	2.00	8.60	28.10	51.00	79.70	88.30	108.50	125.10	71.40	40.20	28.90	16.30	7.30	4.20	5.20	6.90	13.40	
d22	6.40	4.80	4.10	4.70	2.50	1.30	1.70	3.20	15.50	39.30	59.20	73.90	112.80	123.90	121.70	96.80	65.40	44.70	32.60	23.00	7.50	7.00	4.80	6.50	
d23	6.10	4.60	7.40	9.50	7.00	3.20	3.70	1.40	4.80	10.00	42.60	NA	NA	79.60	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	6.10
d24	4.70	4.50	5.90	4.20	3.40	2.00	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	84.00	44.20	40.60	28.90	23.80	10.10	8.60	NA	3.20
d25	15.30	12.00	32.70	22.20	17.90	13.70	2.50	5.70	21.40	39.50	49.60	59.90	65.00	58.80	60.30	55.20	63.80	49.80	26.50	15.90	10.90	10.40	7.10	10.50	
d26	2.20	3.40	6.90	4.30	1.90	1.80	3.30	3.90	13.40	30.70	55.90	66.00	76.20	107.60	121.40	61.90	46.80	32.30	31.30	19.90	3.80	4.60	4.10	4.20	
d27	2.40	5.30	8.40	8.40	4.10	1.60	1.10	4.40	17.60	33.70	62.50	85.30	89.90	124.90	126.70	74.70	54.00	34.00	21.00	12.00	4.90	4.60	3.60	1.60	
d28	3.70	2.60	4.40	3.50	3.70	4.20	1.80	3.90	14.00	38.40	59.00	70.70	109.60	114.00	67.60	47.50	43.20	32.40	20.90	11.30	5.80	8.00	10.30	2.60	
d29	8.10	14.20	8.30	4.10	2.70	1.70	1.90	3.20	12.20	22.30	51.30	80.00	95.80	111.50	98.70	62.60	49.10	46.50	56.10	30.10	23.30	18.70	15.30	8.70	
d30	13.20	16.10	7.60	9.70	6.10	4.40	2.10	3.80	11.40	29.20	48.50	71.40	77.80	75.50	86.60	87.70	53.70	40.40	24.10	7.10	6.90	6.40	5.60	17.10	
d31	10.30	8.10	9.00	10.20	6.80	1.50	2.40	NA	NA	NA	NA	NA	NA	NA	NA	85.40	NA	NA	31.10	16.70	6.20	6.30	5.10	6.90	

Tabla 8.5: Datos contaminados P15-G7 en el mes de Enero

Imputación por media no condicional

Considerando el mes de Enero bajo la contaminación aplicada de valores perdidos con un 15% de datos aleatorios extraídos junto a una longitud consecutiva de 7, visualizado en la Tabla 8.5, la estimación de los valores perdidos (NA) mediante la técnica de media no condicional en una hora específica, se lleva a cabo mediante promedios de información disponible en los 31 días que conforman el mes de Enero, es decir, se tiene como principal insumo la cantidad de O_3 registrada en la misma hora para los diferentes días del mes.

De forma puntual, en la Tabla 8.5 se puede observar como el primero de enero en la hora 12 se presenta un dato ausente y la estimación de dicho valor se da mediante el promedio de la cantidad de O_3 en la misma hora (franja roja) para los 31 días del mes. Adicional, es posible detectar que días como el 23, 24 y 31 de Enero presentan registros faltantes que no permiten conocer la cantidad de contaminante que se presentó al medio día, por lo tanto, el promedio obtenido inicialmente para el primero de Enero será el mismo valor de imputación para estos días.

Y es así como se desarrolla la técnica de imputación por media no condicional para todos los escenarios contemplados, siendo una metodología de fácil alcance, sencilla y útil ante el fenómeno de registros sin información.

Imputación por media antes y después

Contemplando de igual forma el escenario explicado anteriormente, la estimación de la información perdida a través de la media antes y después se desarrolla fijando el día que presenta la ausencia en el registro de cantidad de O_3 en una hora específica y realizando un promedio entre el registro disponible en la hora inmediatamente anterior al valor perdido junto al dato disponible inmediatamente después del mismo. Ejemplificando lo anterior, la Tabla 8.5 evidencia en el día 24 un espacio consecutivo de 9 registros perdidos (franja gris), donde se realiza la estimación de dichos valores con los registros disponibles antes y después, a saber, cantidad de O_3 registrada en la hora 6 y 16 del día. Por lo tanto, se puede afirmar que el valor estimado para estas 9 horas sin información será el mismo.

Imputación por regresión determinística

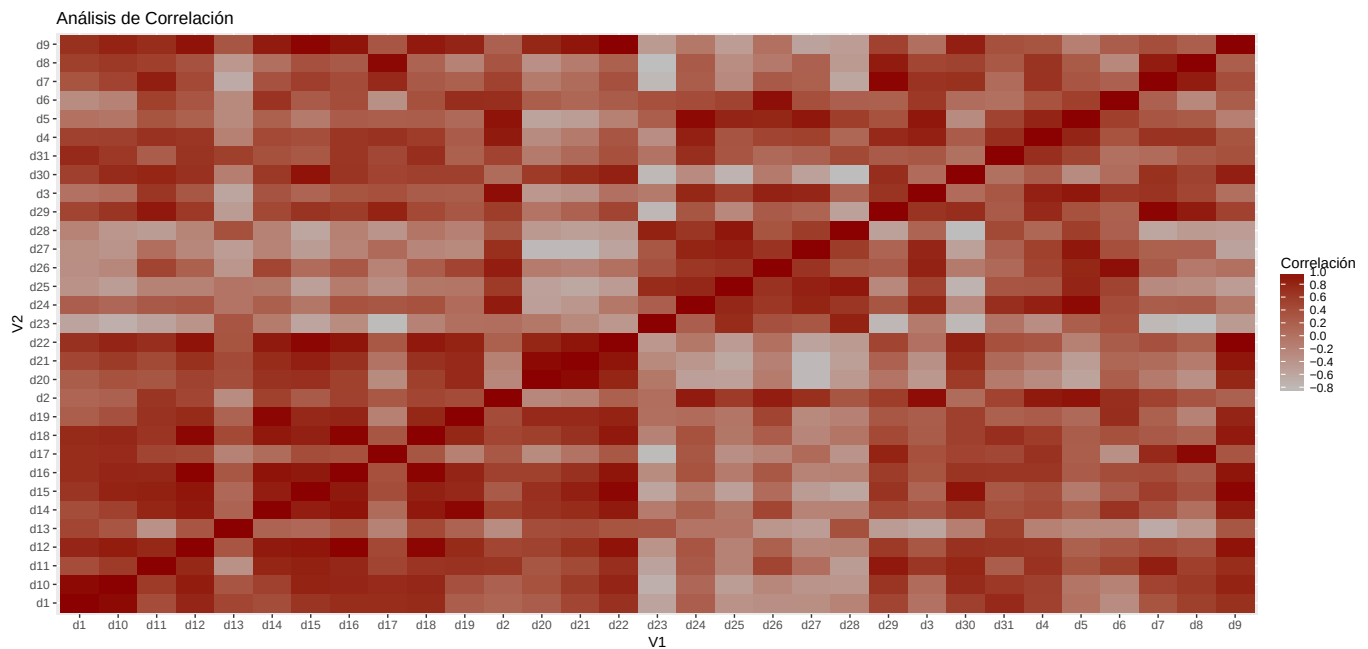


Figura 8.19: Mapa de calor de la matriz de correlación

Ejecutando la imputación por regresión determinística para modelar la variable de respuesta correspondiente al día donde se presenta la hora con ausencia de registro respecto a la cantidad promedio de O_3 , se relaciona la variable de interés con un grupo de covariables correlacionadas evaluadas en la Figura 8.19, que tienen la capacidad de estimar la cantidad del contaminante en la hora perdida. Con base al método utilizado, al igual que la técnica de imputación de medias, esta técnica sesga las medidas de asociación y variabilidad, aunque produce estimaciones insesgadas de la media cuando los mecanismos generadores del dato ausente son MCAR o MAR, esto según Junninen et al. (2004).

Por lo anterior, se requiere explorar los días que cuentan con información disponible en la hora que reporta el registro perdido con la finalidad de conocer la correlación con el día que se desea completar y diseñar de esta forma un modelo de predicción. En esta oportunidad, se tienen tantos modelos de predicción como días en presencia de datos faltantes, es decir, existe un único modelo lineal simple para cada día. Esto es generalizado para los diferentes escenarios contemplados.

A modo de ilustración, se desarrolla un modelo de predicción para el día 24 de Enero el cual, fue contaminado con 15 % de datos aleatorios extraídos junto una longitud consecutiva de 7. La variable de respuesta será el día 24 que cuenta con 10 registros perdidos y la variable independiente será el día con mayor correlación con variable dependiente, a saber, según la Tabla 8.5 la covariable para realizar la estimación será el día 5 de Enero. De esta forma, el dato faltante será estimado con el modelo obtenido. Es importante resaltar que, al ser un modelo determinístico no tiene en cuenta las propiedades que recaen en el error del modelo.

8.5.2. Métodos basados en datos funcionales

En la práctica, los datos funcionales se observan y registran discretamente, para posteriormente realizar una transformación que contempla el espacio y propiedades de las funciones. Por lo tanto, basándose en las técnicas de conversión de datos ADF, se consideran cuatro métodos de imputación. Estos métodos se clasifican en dos categorías, la imputación única y la imputación iterativa sin penalización por rugosidad.

8.5.3. Imputación media funcional y mediana funcional

Considerando la técnica de imputación única para la estimación de información perdida a través de la media y la mediana funcional, las cuales son obtenidas a partir de las curvas de los datos disponibles. Puntualmente, para este caso en particular se contó con 26 curvas con registros de cantidad de O_3 horario disponible para la obtención de este par de curvas. A la luz de la sección 6, se ejecuta el cálculo del promedio y mediana de todas las curvas con información completa con el objetivo de reemplazar el dato perdido por la cantidad de O_3 que se da en las curvas de tendencia central.

De manera puntal en la figura 8.20 se observa el día 24 junto con las medidas de tendencia central y los datos faltantes asociados al mismo exaltados mediante puntos, obtenidos mediante la contaminación aleatoria. Como se mencionó anteriormente para la estimación de los valores perdidos en el día 24, se procederá entonces a reemplazar el valor ausente por el valor ligado a la misma hora en las curvas de media o mediana funcionales.

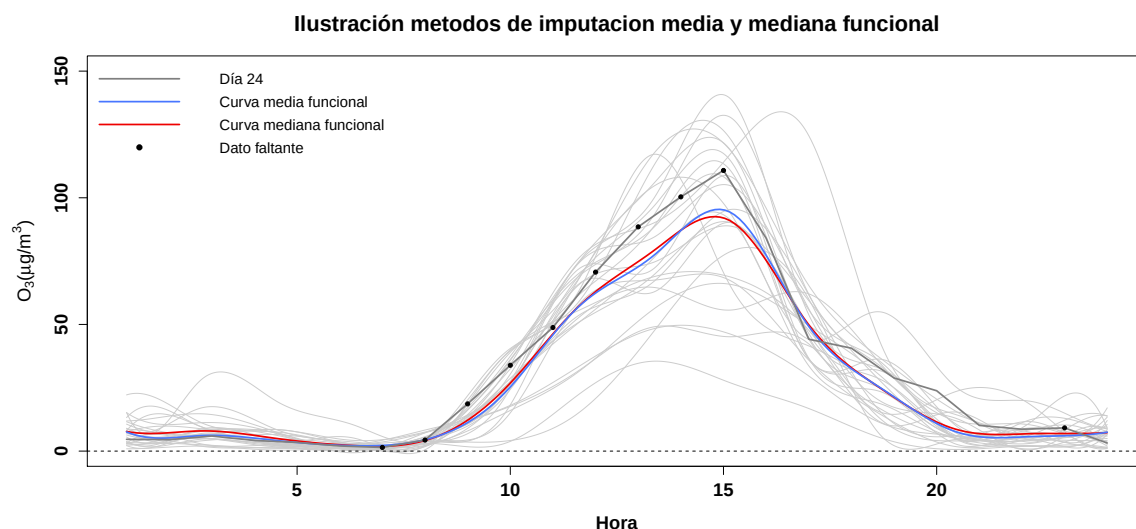


Figura 8.20: Ejemplo intuitivo técnica de imputación media funcional y mediana funcional

8.5.4. Método iterativo basado en la media funcional (FMeanBIM) y método iterativo basado en la mediana funcional (FMedBIM)

Este par de métodos desarrollan un enfoque iterativo para la imputación de datos faltantes con el objetivo de mejorar las estrategias de imputación única. El proceso inicia imputando los valores perdidos en el día con los valores asociados a la media o mediana del conjunto de datos completo disponible para la primera iteración y luego los valores perdidos se imputaran iterativamente hasta que el algoritmo converja. La convergencia se cumple siempre que la diferencia en el error (MPAD) entre la imputación en la iteración actual y la anterior sea como máximo 0,005. El MPAD se define como la proporción media de la diferencia absoluta entre el punto del valor observado y el valor ajustado.

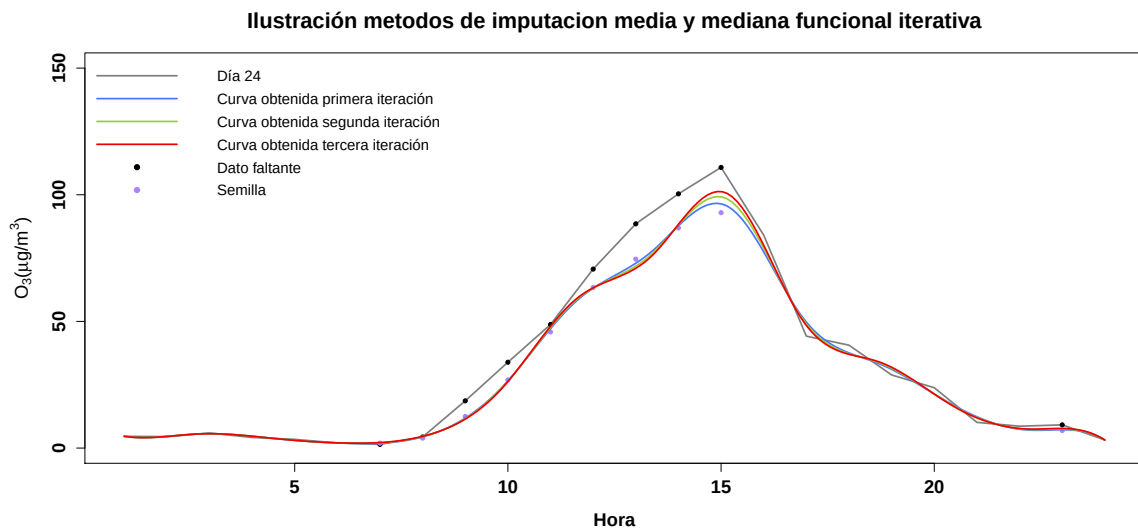


Figura 8.21: Ejemplo intuitivo técnica de imputación media funcional y mediana funcional iterativa

En la figura 8.21 se representa gráficamente la implementación de este de este tipo de métodos, tomando como punto de partida para la estimación de los registros faltantes una semilla, la cual va ser la media o mediana de la hora a tratar. Posteriormente ya habiendo estimado los registros faltantes, se ajusta una primera curva funcional sobre estos puntos (curva azul) y se calcula el indicador $MPAD_1$ para esta. Luego los valores de dicha curva se convierten en la nueva estimación del dato perdido, ajustando nuevamente sobre estos puntos una nueva curva funcional (curva verde) y calculando nuevamente el indicador $MPAD_2$ para esta nueva curva. Ahora, una vez en este punto se calcula si la diferencia absoluta entre estos dos MPADs calculados es menor a 0.005, si esto es así el método se detiene y los valores obtenidos de esta ultima curva serán la estimación para los registros faltantes, de lo contrario se continua el procedimiento. Para este caso en particular, la diferencia resultó siendo mayor, con lo cual se necesitó de una iteración mas para el método. Se procede entonces de igual manera, los valores de esta última curva funcional serán los nuevos valores para los registros faltantes y sobre estos nuevos puntos se ajusta nuevamente una curva funcional (curva roja) y se calcula su $MPAD_3$. Se evalúa de nuevo su diferencia absoluta entre estos dos últimos MPADs, obteniendo como resultado que el procedimiento se detenga y que se tome los valores de esta última curva funcional como la estimación de los valores faltantes para el día. Este tipo de métodos tiene un convergencia rápida, en ocasiones basta con dos iteraciones para que el método converga, sin embargo en algunas ocasiones se requieres de muchas mas, como se evidencia en la Tabla 8.6.

		P5G3	P5G7	P10G3	P10G7	P15G3	P15G7
Día 1	Media	29	11	41	9	19	6
	Mediana	2	2	2	2	2	2
Día 7	Media	94	28	65	20	44	23
	Mediana	6	7	7	7	10	11
Día 23	Media	31	11	32	13	25	15
	Mediana	3	5	5	5	7	10
Día 24	Media	19	5	14	6	16	5
	Mediana	4	2	4	2	4	2
Día 31	Media	25	15	26	14	25	7
	Mediana	2	2	4	2	4	3
Promedio	Media	40	14	36	12	26	11
	Mediana	3	4	4	4	5	6

Tabla 8.6: Número de iteraciones que aseguran la convergencia del método

En la Tabla 8.6 se evidencia como el método funcional iterativo por la media converge mas lento al error mínimo posible respecto a los valores reales de la curva, requiriendo de esta forma un número mayor iteraciones para asegurar la convergencia. Lo anterior, se encuentra justificado por la influencia que tiene los valores extremos sobre la medida de tendencia central media. Por otro lado, la técnica funcional iterativa por la mediana no permite dicha influencia, razón por la cual, no requiere de un gran número de iteraciones para asegurar la convergencia del método al valor que proporcione el mínimo error.

8.5.5. Comparación de métodos de imputación

Con la visión de cumplir con los objetivos planteados y darle respuesta a la pregunta de investigación, se desarrolla esta sección que permite visualizar los resultados obtenidos por los métodos clásicos y ADF, permitiendo generar comparaciones a través de indicadores de desempeño. Resumiendo lo ejecutado hasta el momento, se contaminó un set de datos completos con la finalidad de imputar o estimar los valores perdidos de cantidad promedio de O_3 que es medido en el espacio aéreo de la estación Compartir. Posteriormente, se realizan las estimaciones de tales registros mediante métodos clásicos (utilizados comúnmente) y por metodología ADF. En este punto, se cuenta con el valor imputado por las diferentes técnicas junto al dato real, lo que permite conocer cuál es la proximidad de la estimación con la realidad, evidenciando el poder de los distintos métodos de imputación en diferentes escenarios.

Patrón	Indicador	Metodos de imputación						
		No iterativos Funcionales		Iterativos Funcionales		Clasicos		
		FMeanDM	FMedDM	FMeanBIM	FMedBIM	MeanClas	MeanBAClas	RegreClas
P05-G3	AI	0,835	0,836	0,824	0,851	0,834	0,693	0,853
	SCE	70,820	71,219	77,236	52,499	71,919	211,626	62,402
	RCuadrado	0,489	0,488	0,462	0,472	0,488	0,394	0,475
	MAE	6,526	6,519	6,123	5,438	6,621	10,033	5,302
P05-G7	AI	0,938	0,937	0,927	0,939	0,938	0,489	0,935
	SCE	82,116	82,595	78,874	75,556	83,536	1078,843	105,306
	RCuadrado	0,697	0,698	0,671	0,679	0,697	0,262	0,659
	MAE	7,170	7,196	6,919	6,727	7,290	26,731	6,676
P10-G3	AI	0,904	0,902	0,900	0,918	0,904	0,819	0,917
	SCE	68,607	68,848	71,263	47,254	69,764	194,285	61,580
	RCuadrado	0,568	0,570	0,550	0,562	0,569	0,474	0,560
	MAE	6,404	6,378	5,675	5,107	6,500	9,323	5,256
P10-G7	AI	0,956	0,954	0,952	0,959	0,955	0,565	0,951
	SCE	79,201	79,593	73,154	70,164	80,644	1014,111	104,457
	RCuadrado	0,727	0,727	0,712	0,718	0,727	0,296	0,690
	MAE	7,008	7,021	6,566	6,413	7,127	25,228	6,627
P15-G3	AI	0,957	0,957	0,958	0,967	0,957	0,901	0,965
	SCE	68,018	68,002	63,254	45,493	69,212	192,885	62,880
	RCuadrado	0,679	0,679	0,667	0,676	0,679	0,582	0,668
	MAE	6,365	6,318	5,405	4,943	6,463	9,003	5,361
P15-G7	AI	0,967	0,966	0,969	0,971	0,967	0,640	0,963
	SCE	81,082	81,605	74,447	71,487	81,077	941,586	98,890
	RCuadrado	0,765	0,763	0,759	0,761	0,766	0,351	0,730
	MAE	7,154	7,084	6,510	6,319	7,214	23,423	6,725

Tabla 8.7: Rendimiento de los siete métodos de imputación

En Tabla 8.7, se visualiza los diferentes escenarios evaluados (filas) donde se construyeron los indicadores de rendimiento que permiten comparar los resultados obtenidos por los siete métodos de imputación. Es importante destacar que, estos resultados son basados en lo especificado en la Sección 6, donde se aclara la regla de decisión de cada indicador, es posible afirmar que:

- Bajo la contaminación aplicada de valores perdidos con un 5 % de datos aleatorios extraídos junto a una longitud consecutiva de 3, los indicadores consideran diferentes decisión frente al método que mejor rendimiento presenta, esto a la luz de mayor precisión o cercanía con el valor real. Sin embargo, es posible determinar indicadores de mayor interés como la suma del cuadrado de error. Por lo tanto, en este escenario tanto la regresión clásica como la iteración funcional por mediana arrojan los mejores resultados, como se muestra de forma sobresaliente (color rojo) en Tabla la 8.7.
- Cambiando al segundo escenario contemplado, el cual esta constituido con un 5 % de datos aleatorios extraídos junto una longitud consecutiva de 7, los indicadores descartan las estimaciones obtenidas por los métodos de media no condicional junto al promedio con valores antes y después, sobresaliendo de esta forma la imputación obtenida por los diferentes métodos considerados de transformación ADF y por regresión. Sin embargo, es posible afirmar que bajo porcentajes aleatorios faltantes pequeños junto a longitudes menores, los indicadores demarcan la gran utilidad de los métodos propuestos y descartando directamente los técnicas clásicas utilizadas con mayor frecuencia por los investigadores, basados en promedios.
- Al considerar el escenario intermedio, estructurado con un espacio consecutivo de longitud de 3 y 7, conservando fijo el porcentaje aleatorio de valores perdidos equivalente al 10 %, el mejor rendimiento respecto a los métodos evaluados recae sobre los diferentes variaciones de la técnica ADF. Así, la técnica iterativa utilizada con la transformación ADF refleja mayor proximidad con el valor real, con lo cual los diferentes indicadores la declaran como metodología de mejor rendimiento para la imputación.
- Por ultimo, el escenario más extremo que consta de un espacio consecutivo de longitud 3 y 7, conservando fijo el porcentaje aleatorio de valores perdidos equivalente al 15 %, se observa en los

resultados obtenidos en la Tabla 8.5, como a medida que se incrementa la cantidad de valores y longitud de espacios consecutivos, que el método que presenta mejor rendimiento pertenece al desarrollado por métodos funcionales iterativos.

A continuación se presenta un ejemplo de una posible estimación para los registros de O_3 faltantes en el día 24 del mes de Enero, bajo el escenario que se ha venido planteando de 5% de datos perdidos con una longitud de espacio de 7. Así, se obtiene el siguiente resultado:

Tabla 8.8: Estimaciones puntuales obtenidas por cada método

	Día 24									
Hora perdida	7	8	9	10	11	12	13	14	15	23
Valor real	1,45	4,38	18,64	33,89	48,78	70,67	88,52	100,36	110,80	9,16
FMeanDM	1,96	4,33	11,80	26,34	46,52	62,87	73,42	86,75	93,89	6,51
FMedDM	2,63	5,00	11,20	25,56	47,99	65,14	74,21	87,81	95,83	6,07
FMeanBIM	2,27	4,68	11,29	26,15	48,38	63,30	70,34	88,52	102,80	7,95
FMedBIM	2,08	4,49	11,28	27,68	52,93	68,93	73,62	87,85	99,58	6,62
MeanClas	1,96	3,94	12,48	26,88	45,83	63,44	74,60	86,90	92,92	6,93
MeanBAClas	43,03	43,03	43,03	43,03	43,03	43,03	43,03	43,03	43,03	5,92
RegreClas	3,15	4,37	14,56	34,21	50,61	62,78	94,92	104,06	102,26	5,70

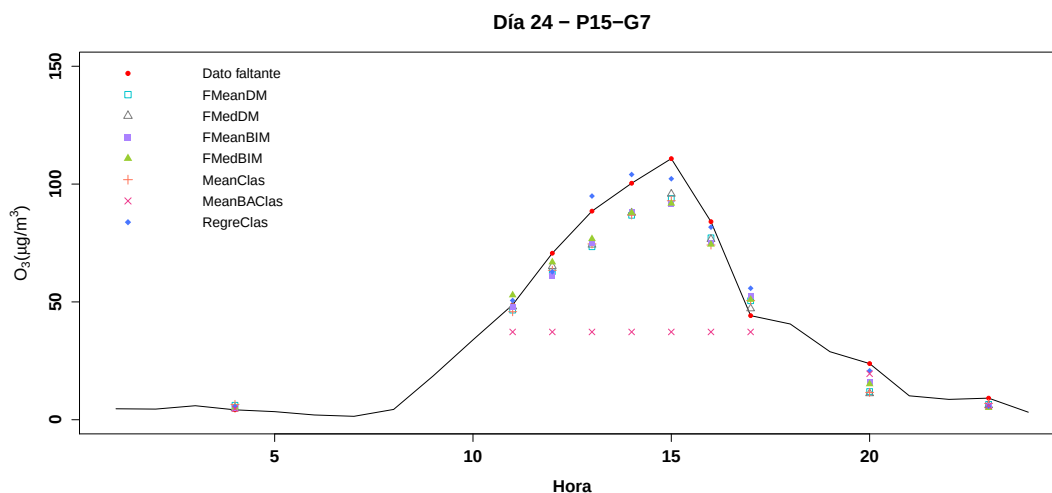


Figura 8.22: Estimaciones puntuales obtenidas por cada método para el día 24.

En Tabla 8.8 se observan los valores reales junto con los valores estimados por cada método de imputación en las horas faltantes. Esto permite contemplar la proximidad o no del valor imputado al valor real. Por otro lado en la Figura 8.22 se evidencia con mayor claridad lo anterior mencionado, donde se destaca que el método de la media antes y después, posee una estimación que se encuentra muy sesgada o muy lejana en cuanto a los valores reales. Por otra parte el método de imputación funcional iterativo FMedBIM y el método de regresión clásico RegreClas, están presentando para este escenario generado en particular, las mejores estimaciones en cuanto al valor real. Para los días 1, 7, 23, 31 se presenta un comportamiento similar en cuanto a las imputaciones generadas por los métodos, la cuales se pueden encontrar en la Sección de Anexos.

Capítulo 9

Conclusiones y Recomendaciones

9.0.1. Conclusiones

En el presente trabajo se introdujeron siete métodos de imputación para la estimación de los datos faltantes del contaminante criterio O_3 de la estación de monitoreo de calidad del aire Compartir ubicada en la ciudad Santiago de Cali, esto se realizó para los días del mes de Enero del 2016. Los métodos de imputación se clasifican en tres categorías: no iterativos funcionales, iterativos funcionales y métodos clásicos.

Los resultados obtenidos por los diferentes métodos de imputación mostraron que las técnicas de imputación por medio de metodología ADF obtienen mejores desempeños y resultados en comparación a los métodos clásicos como se evidencia en la Tabla 8.7 donde se logra contrastar los dos grupos de metodologías (clásica vs funcional). Por otro lado, se identifica que los métodos iterativos de imputación ADF al tener como punto de partida en el proceso la minimización del error entre la estimación y el valor real de la cantidad de O_3 presente en una hora determinada, tomando como función inicial el promedio o la mediana del contaminante medido diario, optimiza el rendimiento para la estimación de observaciones ausentes respecto a métodos no iterativos.

Por otra parte, a la luz de la función semilla utilizada (Media y mediana diaria) en el método iterativo se evidencian propiedades conocidas del campo clásico en el mundo de las funciones, a saber, se obtienen mejores estimaciones para datos ausentes partiendo de la mediana, la cual no se deja influenciar o afectar por curvas extremos de O_3 que se puedan presentar a lo largo del día. Se observó de forma adicional que, los métodos de imputación ADF no iterativos media (FMeanDM) y Mediana(FMedDM) presentan resultados similares en los distintos indicadores de desempeño que se evaluaron.

Del mismo modo, en los métodos de imputación clásicos se destaca el método de regresión (RegreClas), el cual presentó el mejor desempeño basado en los indicadores ya que involucra la correlación existente diaria del contaminante evaluado a lo largo del mes de Enero, seguido del método de la media (MeanClas) y por último el método de la media antes-después (MeanBAClas). Este último método en general, es el que presentó los peores resultados a la hora de estimar los valores faltantes debido a la poca eficiencia y flexibilidad presentada en las estimaciones obtenidas de los espacios consecutivos, donde generaba un único valor de O_3 , ocasionando una dependencia respecto a la posición antes y después del dato ausente para diferentes horas contiguas del día.

No obstante, es posible afirmar que a medida que aumenta la cantidad de valores faltantes al interior de la base de datos utilizada los métodos funcionales presentan un mejor desempeño y rendimiento, mientras que para bajos porcentajes de información ausente, es posible completar la base de datos con estimaciones obtenidas mediante el método de regresión o ADF iterativo. Finalmente, se encontró que todos los métodos

de imputación son independientes del porcentaje de datos faltantes pero dependen del tamaño de la longitud de espacio consecutiva que estos presenten. En general los resultados mostraron que el método de imputación iterativa FMedBIM cuenta con mayor flexible y genera estimaciones más reales frente a los otros métodos evaluados.

9.0.2. Recomendaciones

Para estudios posteriores se recomienda estudiar hasta qué punto los métodos de imputación basados en metodologías FDA logran una buena estimación del valor perdido. Puntualmente, encontrar la cantidad o porcentajes de valores ausentes que obligan al método a salirse de control, es decir, encontrar los umbrales de información faltante donde el método estima los valores de forma extremadamente sesgada, concluyendo con estimaciones que no se ajustan ni reflejan la realidad. De forma adicional, se recomienda hacer uso de diversas bases de funciones para la estimación de los valores de las horas del contaminante criterio O_3 . Por último se recomienda hacer uso de inferencias funcionales que permitan establecer si las diferencias encontradas en las funciones media, mediana y desviación estándar son significativas.

Anexos A. Análisis exploratorio de datos

A continuación, se presenta la discriminación porcentual para la distribución de valores perdidos por mes y hora trabajados para los años 2015 y 2016.

Tabla .1: Distribución de datos faltantes por hora de acuerdo a los meses para el año 2015

[illegible]

Tabla .2: Distribución de datos faltantes por hora de acuerdo a las horas para el año 2015

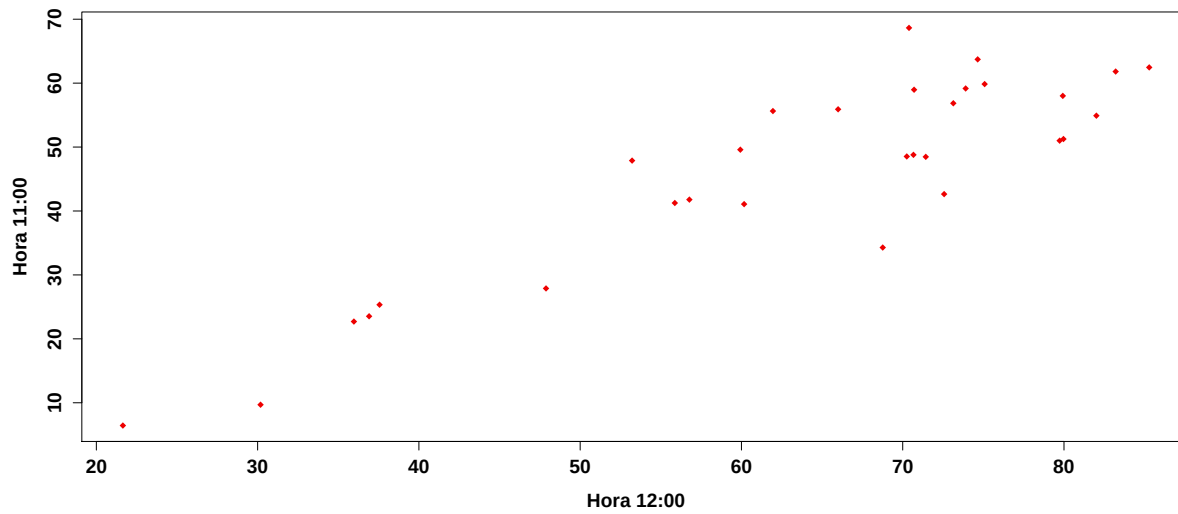
Mes	Cantidad de datos faltantes (columna)																								
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	
Enero	12.16	12.16	12.08	12.00	12.08	12.08	12.08	12.08	12.24	11.61	12.24	12.33	12.24	12.08	11.92	12.33	12.33	12.41	13.01	13.10	13.10	13.10	13.01	12.84	12.84
Febrero	18.92	18.92	18.79	18.67	18.79	18.79	18.79	18.79	18.79	18.06	19.05	19.18	19.05	18.79	18.54	19.18	19.18	19.31	19.18	19.31	19.31	19.31	19.18	18.92	18.92
Marzo	14.19	14.19	14.09	14.00	14.09	14.09	14.09	14.09	14.09	14.29	14.19	14.29	13.10	14.29	14.09	13.25	13.70	13.70	13.70	13.70	13.70	13.70	13.70	13.70	13.70
Abril	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.32	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Mayo	0.68	0.68	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.68	2.58	0.68	0.68	0.68	0.68	1.34	1.32	1.37	1.37	1.38	1.37	1.38	1.38	1.37	1.35
Junio	7.43	7.43	7.38	7.33	7.38	7.38	7.38	7.38	7.38	7.48	6.45	6.80	6.16	6.80	6.71	6.62	6.85	6.85	6.90	6.85	6.90	6.90	6.85	6.76	6.76
Julio	1.35	1.35	1.34	1.33	1.34	1.34	1.34	1.34	1.34	1.36	3.23	2.04	2.05	1.36	1.34	1.32	1.37	1.38	1.37	1.38	1.38	1.38	2.05	2.05	2.05
Agosto	18.24	18.24	18.12	18.00	18.12	18.12	18.12	18.12	18.12	18.37	17.62	17.69	17.81	17.69	17.45	17.22	17.81	17.81	17.93	17.81	17.93	17.93	17.81	17.57	17.57
Septiembre	8.11	8.11	8.05	8.00	8.05	8.05	8.05	8.05	8.05	8.16	7.74	8.16	8.22	8.16	8.05	8.61	8.22	8.22	8.28	8.22	8.28	8.28	8.22	8.78	8.78
Octubre	4.05	4.05	4.03	4.00	4.03	4.03	4.03	4.03	4.03	3.40	3.23	3.40	3.42	3.40	4.03	3.31	3.42	3.42	2.76	2.74	2.76	2.76	2.74	3.38	3.38
Noviembre	0.68	0.68	0.67	1.33	0.67	0.67	0.67	0.67	0.67	0.68	1.29	0.68	1.37	1.36	1.34	1.99	1.37	1.37	1.38	1.37	0.69	0.69	0.68	0.68	0.68
Diciembre	14.19	14.19	14.77	14.67	14.77	14.77	14.77	14.77	14.77	14.29	14.19	14.97	14.97	14.97	14.77	14.57	14.38	14.38	14.48	14.38	14.48	14.48	14.38	14.19	14.19

Tabla .3: Distribución de datos faltantes por hora de acuerdo a los meses para el año 2016

[illegible]

Tabla .4: Distribución de datos faltantes por hora de acuerdo a las horas para el año 2015

Mes	Cantidad de datos faltantes(columna)																							
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Enero	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,57	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Febrero	0,58	0,57	0,57	0,56	0,56	0,56	0,56	0,56	0,56	0,56	1,15	0,58	1,17	0,59	0,59	1,15	0,58	0,59	0,58	0,58	0,58	0,57	0,57	0,57
Marzo	7,51	7,47	7,43	7,30	7,26	7,78	7,78	7,78	7,87	7,91	8,62	7,51	8,19	8,28	8,28	9,20	8,19	7,69	7,60	7,60	7,60	8,05	8,00	8,00
Abril	1,16	1,15	1,14	1,12	1,12	1,11	1,11	1,11	1,12	2,26	1,15	1,16	0,58	0,59	0,59	1,72	1,17	0,59	0,58	0,58	0,58	0,57	0,57	0,57
Mayo	4,62	5,17	5,71	5,62	5,59	5,56	5,56	5,56	5,62	5,08	4,02	4,62	4,68	4,14	4,14	4,60	4,68	4,73	4,68	4,68	4,68	5,17	5,14	5,14
Junio	9,83	9,77	9,71	10,67	10,61	10,56	10,56	10,56	10,11	10,17	10,34	10,40	9,94	10,06	10,06	9,20	9,36	9,47	9,94	9,94	9,94	9,77	9,71	9,14
Julio	2,31	2,30	2,29	2,25	2,23	2,22	2,22	2,22	2,25	2,26	1,72	2,89	1,75	1,78	1,78	1,72	2,34	2,37	2,34	2,34	2,34	2,30	2,29	2,29
Agosto	3,47	3,45	3,43	3,93	4,47	4,44	4,44	4,44	3,93	2,82	2,30	2,31	2,34	2,37	2,37	2,30	2,34	2,37	2,92	2,92	2,92	3,45	4,00	4,57
Septiembre	17,34	17,24	17,14	16,85	16,76	16,67	16,67	16,67	16,85	16,95	17,24	17,34	17,54	17,75	17,75	17,24	17,54	17,75	17,54	17,54	17,54	17,24	17,14	17,14
Octubre	17,92	17,82	17,71	17,42	17,32	17,22	17,22	17,22	17,42	17,51	17,82	17,92	18,13	18,34	18,34	17,82	18,13	18,34	18,13	18,13	18,13	17,82	17,71	17,71
Noviembre	17,34	17,24	17,14	16,85	16,76	16,67	16,67	16,67	16,85	16,95	17,24	17,34	17,54	17,75	17,75	17,24	17,54	17,75	17,54	17,54	17,54	17,24	17,14	17,14
Diciembre	17,92	17,82	17,71	17,42	17,32	17,22	17,22	17,22	17,42	17,51	17,82	17,92	18,13	18,34	18,34	17,82	18,13	18,34	18,13	18,13	18,13	17,82	17,71	17,71

Diagrama de dispersión de las concentraciones de O_3 **Figura .1:** Estructura de los datos para construcción del modelo

Anexos B. Construcción de curvas funcionales: Coeficiente y funciones

En este capítulo se encuentran los resultados obtenidos al construir las curvas funcionales. La apreciación matemática del modelo utilizado es dispendiosa, sin embargo, en las tablas se evidencia los diferentes componente de cada curva funcional.

Dim1	Dim2	Dim3	Dim4	Dim5	Dim6	Dim7	Dim8	Dim9	Dim10	Dim11	Dim12	Dim13	Dim14	Dim15	Dim16
1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.08	0.57	0.31	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.16	0.60	0.24	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.01	0.35	0.59	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.07	0.61	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.27	0.64	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.04	0.54	0.41	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.19	0.66	0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.02	0.47	0.49	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.13	0.66	0.21	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.01	0.38	0.57	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.62	0.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.29	0.62	0.08	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.57	0.38	0.01	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.21	0.66	0.13	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.49	0.47	0.02	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.15	0.66	0.19	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.41	0.54	0.04	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.64	0.27	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.32	0.61	0.07	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.59	0.35	0.01	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.24	0.60	0.16	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.31	0.57	0.08	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Tabla .5: Matriz $\phi(t)$

	Y ₁	Y ₂	Y ₃	Y ₄	Y ₅	Y ₆	Y ₇	Y ₈	Y ₉	Y ₁₀	Y ₁₁	Y ₁₂	Y ₁₃	Y ₁₄	Y ₁₅	Y ₁₆	Y ₁₇	Y ₁₈	Y ₁₉	Y ₂₀	Y ₂₁	Y ₂₂	Y ₂₃	Y ₂₄	Y ₂₅	Y ₂₆
Dim.1	9.02	0.90	11.90	2.97	3.79	3.14	8.74	22.18	7.52	15.19	8.14	7.61	3.32	7.75	2.79	7.03	5.55	10.66	11.60	6.38	15.24	2.16	2.38	3.68	8.06	13.22
Dim.2	2.69	0.89	12.17	1.36	-1.64	10.27	7.96	25.38	10.93	8.34	1.30	3.46	3.32	2.78	6.85	4.30	-0.28	6.29	7.79	3.40	-7.36	-1.71	2.24	-0.99	19.81	23.76
Dim.3	26.80	3.26	18.00	10.28	15.23	-0.86	4.53	5.35	9.00	13.33	4.89	9.32	2.76	9.46	-0.45	6.99	10.54	5.24	4.89	6.68	47.95	13.16	11.33	8.89	6.80	2.65
Dim.4	4.75	-0.94	7.08	1.33	1.85	4.32	3.29	9.89	4.76	8.20	5.32	9.12	0.38	1.36	2.62	-0.25	0.52	10.92	7.93	1.33	13.77	-1.65	6.58	-1.26	3.34	12.13
Dim.5	4.39	2.61	-2.27	-1.20	1.81	-2.46	-0.16	0.09	0.45	-5.10	3.11	-3.48	2.11	1.42	-1.11	4.13	4.37	1.84	2.32	4.94	13.86	5.68	-0.69	11.39	0.25	-0.05
Dim.6	-0.55	-4.04	5.98	6.63	5.06	7.70	3.71	7.44	9.14	8.33	-3.26	6.74	-0.53	-0.85	3.43	-1.69	-5.78	7.20	-0.60	-7.34	-6.63	-2.77	1.54	-12.98	3.65	3.81
Dim.7	15.16	27.02	-3.23	-2.44	1.89	-1.74	22.26	15.66	14.75	18.31	25.53	11.07	23.60	31.87	6.23	30.54	47.51	16.45	14.97	41.02	44.11	28.61	28.12	47.86	13.27	14.78
Dim.8	28.60	53.52	24.36	88.84	10.69	71.44	55.86	29.86	33.54	61.61	51.50	90.56	87.47	77.50	94.26	76.13	46.34	87.16	78.01	60.09	54.53	65.17	79.70	45.09	70.80	78.20
Dim.9	59.46	64.65	40.76	43.44	46.92	64.23	71.52	56.27	52.13	57.40	73.58	72.83	84.56	79.49	48.89	79.69	122.06	83.40	89.34	128.99	71.46	81.29	102.63	152.29	112.55	67.07
Dim.10	71.52	106.89	25.63	114.98	87.70	76.66	71.57	39.65	48.31	112.01	114.00	125.60	115.49	136.22	144.01	166.44	144.79	180.53	146.63	131.22	47.11	144.07	145.98	48.96	107.64	101.62
Dim.11	49.62	49.57	18.72	38.21	74.47	33.27	54.87	55.93	32.08	66.73	18.91	53.73	156.18	11.76	25.99	45.53	39.24	29.99	13.34	59.19	74.09	8.23	27.53	44.43	21.79	59.51
Dim.12	32.37	3.39	3.31	14.41	-12.24	13.83	4.16	-3.47	6.79	5.25	23.60	11.22	20.33	46.34	31.67	38.96	45.06	39.93	33.74	36.51	31.40	52.24	36.06	25.60	73.71	26.44
Dim.13	-1.77	28.79	-0.67	3.52	11.27	-2.22	3.88	26.05	7.37	4.83	-1.58	19.09	16.08	-7.23	-4.35	-0.94	-8.73	5.38	-7.79	14.01	6.29	0.43	-2.40	2.96	12.98	-0.62
Dim.14	4.51	22.93	3.29	1.03	-4.86	5.58	4.45	-8.24	-0.73	12.20	13.83	-10.33	1.44	11.28	20.84	7.36	19.11	2.12	13.64	1.08	14.80	5.45	9.77	8.33	25.47	12.15
Dim.15	-0.67	-0.88	0.18	3.25	7.16	8.07	34.06	15.90	9.18	-2.52	5.55	13.57	8.88	-3.69	0.74	5.96	11.47	17.78	2.31	6.92	2.32	3.87	0.57	12.90	10.23	0.36
Dim.16	8.40	1.04	7.26	1.43	7.77	1.97	8.94	13.94	9.48	12.08	4.96	8.68	3.61	6.60	1.13	5.93	7.66	13.81	13.44	6.42	10.50	4.23	1.59	2.59	8.76	17.07

Tabla .6: Matriz de coeficientes (C)

Nota: La tabla .6 presenta valores del orden $1.000000e+00$ o $-3.284820e-17$, por ende los coeficientes en la tabla son aproximaciones a valores decimales.

d2	d3	d4	d5	d6	d8	d9	d10	d11	d12	d13	d14	d15	d16	d17	d18	d19	d20	d21	d22	d25	d26	d27	d28	d29	d30
9.02	0.90	11.90	2.97	3.79	3.14	8.74	22.18	7.52	15.19	8.14	7.61	3.32	7.75	2.79	7.03	5.55	10.66	11.60	6.38	15.24	2.16	2.38	3.68	8.06	13.22
10.84	1.58	13.82	4.29	4.20	6.02	6.81	18.37	9.86	10.47	3.11	5.81	3.05	5.24	4.10	5.23	3.61	6.46	7.20	4.61	12.48	3.28	5.23	2.48	14.27	15.92
17.55	1.87	14.42	6.67	9.25	2.21	4.79	9.73	8.30	11.28	4.40	8.31	2.28	6.42	1.49	4.82	6.36	6.77	6.10	4.86	30.67	7.17	8.70	4.84	8.11	8.39
12.40	0.74	10.40	4.31	6.49	2.18	3.56	7.87	6.04	9.24	5.02	8.44	1.33	4.20	1.37	2.55	4.22	8.40	6.56	3.41	25.55	3.93	7.80	2.99	4.49	8.22
6.10	0.48	4.80	1.13	2.75	1.80	2.26	6.42	3.67	4.26	4.56	5.06	1.10	1.92	1.22	1.64	2.42	7.60	5.89	2.82	16.04	1.71	4.54	3.46	2.58	7.54
4.04	1.04	1.01	0.22	2.14	0.30	1.13	3.39	2.42	-0.28	3.10	0.84	1.40	1.20	0.31	2.42	2.40	4.76	3.54	2.82	11.94	2.95	1.47	5.73	1.40	3.56
2.49	0.00	1.45	2.09	3.14	1.95	1.78	3.62	4.31	1.12	0.81	1.32	1.18	0.80	0.96	1.85	0.51	4.52	1.46	0.14	5.78	2.18	0.79	1.31	1.89	2.13
2.68	1.75	3.08	3.83	3.98	4.40	5.69	7.25	8.31	7.24	2.14	5.44	3.49	4.35	2.98	4.11	3.91	7.53	2.22	2.04	4.65	3.41	5.00	0.51	4.41	4.68
7.97	12.74	1.73	3.97	3.57	4.37	14.01	11.88	12.32	14.26	12.33	10.67	13.49	17.16	6.88	16.14	21.89	13.55	8.99	18.30	20.19	14.45	16.45	18.81	9.92	10.90
16.04	28.74	3.84	18.21	4.18	15.10	27.09	17.65	18.05	26.30	27.43	27.50	34.19	37.47	24.68	36.19	40.50	30.38	26.47	38.97	39.91	32.44	35.77	39.56	24.35	26.94
24.72	43.56	14.50	51.48	8.98	42.85	43.45	25.53	27.09	44.58	42.27	58.94	62.45	59.70	58.07	58.42	49.92	59.55	54.03	55.56	50.93	51.54	60.61	50.69	50.39	53.06
36.67	54.76	26.98	68.25	20.77	63.49	57.81	36.50	37.53	57.00	56.02	79.06	81.56	74.54	73.98	73.69	68.85	80.56	76.42	78.94	58.66	67.12	82.43	76.82	78.53	69.91
51.28	64.69	34.66	62.42	39.45	67.23	66.84	47.11	46.30	62.93	70.24	82.15	87.78	83.35	69.75	85.49	101.50	92.15	90.46	108.77	64.49	81.48	99.22	112.35	99.72	73.01
62.55	80.08	34.10	72.69	60.94	69.07	70.70	48.73	49.68	78.41	87.54	93.57	96.95	100.48	86.98	112.26	126.62	120.11	110.05	126.16	61.44	103.89	117.52	107.31	108.12	80.66
66.16	90.59	27.99	89.96	77.30	68.50	69.44	45.27	47.06	94.59	93.29	105.21	114.04	108.31	108.72	132.56	126.53	140.67	117.48	121.59	55.73	113.42	121.69	70.45	97.80	88.90
60.32	77.32	22.37	75.72	78.76	55.02	62.57	46.84	40.09	87.69	67.13	88.81	131.96	75.31	84.78	105.78	93.30	105.59	81.17	95.91	59.95	77.68	87.82	48.87	67.18	79.83
49.55	49.19	16.81	44.90	59.98	35.92	47.71	42.31	29.66	61.69	33.67	56.17	124.52	36.46	44.28	61.92	55.73	53.83	36.64	65.40	62.07	36.38	46.42	41.53	44.14	59.39
38.53	24.26	9.69	24.73	25.10	21.80	25.56	22.36	17.57	31.44	21.64	30.04	76.67	31.08	29.12	41.42	41.65	35.97	24.96	45.88	48.07	33.21	32.22	32.68	50.53	39.70
24.88	14.50	3.70	13.73	2.17	11.38	8.85	9.96	9.32	10.92	16.45	17.29	31.95	28.80	21.54	28.93	30.18	29.77	20.75	32.62	28.71	34.28	25.00	21.33	52.63	22.33
9.81	20.27	0.93	6.96	2.79	3.58	4.16	14.32	6.71	5.64	7.63	14.68	16.88	11.33	9.04	12.61	10.62	16.37	7.10	20.52	15.15	17.49	10.90	10.74	33.42	9.14
2.34	25.11	0.94	3.26	4.30	1.47	4.31	12.37	4.53	7.37	5.26	8.35	11.16	2.25	6.49	4.24	4.14	6.27	2.09	10.71	10.64	5.11	4.02	6.17	20.72	5.36
2.17	20.41	1.83	2.00	0.97	4.12	9.18	3.93	2.84	8.02	8.79	0.65	6.17	4.40	11.51	5.15	11.21	5.49	6.66	5.15	10.72	4.01	5.35	7.80	20.00	7.16
1.67	7.64	1.71	2.41	3.56	6.48	21.80	8.47	6.04	3.52	7.89	5.84	6.33	1.75	6.92	6.19	12.95	12.17	6.48	5.26	7.03	4.29	3.45	10.32	14.98	5.40
8.40	1.04	7.26	1.43	7.77	1.97	8.94	13.94	9.48	12.08	4.96	8.68	3.61	6.60	1.13	5.93	7.66	13.81	13.44	6.42	10.50	4.23	1.59	2.59	8.76	17.07

Tabla .7: Curva estimada $X(t)$

Anexos C. Comparación de valores estimandos contra valores reales

En esta sección, se puede observar los valores puntuales estimados a través de los métodos trabajados a lo largo de la presente investigación para los todos los días contaminados.

Tabla .8: Estimaciones obtenidas por los métodos evaluados para el primero de Enero del 2016

	Día 1									
Hora perdida	4	6	10	11	12	13	14	15	16	21
Valor real	2,94	1,69	38,89	55,63	61,95	70,14	99,47	111,03	90,68	8,28
FMeanDM	6,13	2,40	26,34	46,52	62,87	73,42	86,75	93,89	77,19	7,37
FMedDM	4,78	2,39	25,56	47,99	65,14	74,21	87,81	95,83	76,59	6,16
FMeanBIM	4,32	0,10	32,03	46,54	61,06	75,05	88,07	91,56	74,72	5,01
FMedBIM	3,50	0,82	36,48	52,10	65,50	77,11	88,43	91,33	74,43	4,70
MeanClas	6,50	2,65	26,88	45,83	63,44	74,60	86,90	92,92	74,15	6,80
MeanBAClas	2,81	2,18	35,86	35,86	35,86	35,86	35,86	35,86	35,86	10,20
RegreClas	7,98	8,62	39,55	58,20	68,83	104,04	108,04	66,00	47,85	10,00

Tabla .9: Estimaciones obtenidas por los métodos evaluados para el siete de Enero del 2016

	Día 7							
Hora perdida	16	17	18	19	20	21	22	23
Valor real	82,11	55,54	22,80	10,58	4,32	2,26	2,53	3,14
FMeanDM	77,19	50,50	33,04	21,45	11,96	7,37	6,87	6,51
FMedDM	76,59	47,18	30,73	20,74	11,15	6,16	5,81	6,07
FMeanBIM	77,21	49,66	32,64	21,48	11,51	6,74	6,85	6,92
FMedBIM	76,23	47,91	31,32	20,68	10,54	5,29	5,25	6,12
MeanClas	74,15	50,64	32,71	21,16	11,60	6,80	6,80	6,93
MeanBAClas	49,14	49,14	49,14	49,14	49,14	49,14	49,14	49,14
RegreClas	66,23	35,91	24,68	13,63	9,88	-0,31	0,24	4,66

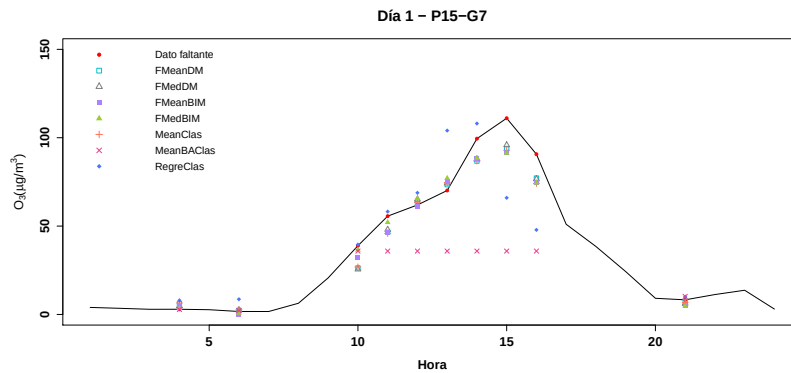
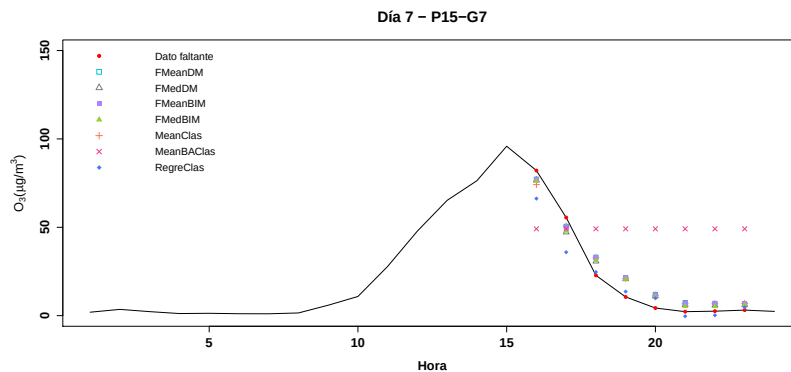
Tabla .10: Estimaciones obtenidas por los métodos evaluados para el veintitrés de Enero del 2016

	Día 23										
Hora perdida	12	13	15	16	17	18	19	20	21	22	23
Valor real	72,57	94,94	73,00	63,23	50,92	26,41	9,01	4,36	2,59	1,67	2,00
FMeanDM	62,87	73,42	93,89	77,19	50,50	33,04	21,45	11,96	7,37	6,87	6,51
FMedDM	65,14	74,21	95,83	76,59	47,18	30,73	20,74	11,15	6,16	5,81	6,07
FMeanBIM	64,89	70,89	89,88	75,45	50,13	32,92	21,13	11,51	6,92	6,73	6,95
FMedBIM	64,79	70,55	87,88	73,85	49,20	31,81	19,89	10,59	5,67	4,99	6,19
MeanClas	63,44	74,60	92,92	74,15	50,64	32,71	21,16	11,60	6,80	6,80	6,93
MeanBAClas	61,14	61,14	42,88	42,88	42,88	42,88	42,88	42,88	42,88	42,88	42,88
RegreClas	61,97	75,02	77,41	47,68	36,54	34,43	42,31	20,92	15,29	11,46	8,71

Tabla .11: Estimaciones obtenidas por los métodos evaluados para el treinta y uno de Enero del 2016

	Día 31									
Hora perdida	8	9	10	11	12	13	14	15	17	18
Valor real	7,12	16,72	40,29	63,72	74,65	81,26	98,55	101,61	65,52	43,74
FMeanDM	4,33	11,80	26,34	46,52	62,87	73,42	86,75	93,89	50,50	33,04
FMedDM	5,00	11,20	25,56	47,99	65,14	74,21	87,81	95,83	47,18	30,73
FMeanBIM	6,49	12,16	25,32	48,02	63,78	70,66	87,88	102,50	54,75	41,03
FMedBIM	5,49	11,73	27,14	52,91	69,25	73,38	87,60	100,35	51,48	37,68
MeanClas	3,94	12,48	26,88	45,83	63,44	74,60	86,90	92,92	50,64	32,71
MeanBAClas	43,94	43,94	43,94	43,94	43,94	43,94	43,94	43,94	58,27	58,27
RegreClas	9,70	21,53	36,19	70,53	72,19	78,87	107,15	113,42	48,44	34,54

De forma adicional, se muestra gráficamente el comportamiento de las estimaciones obtenidas por los difernetes métodos y es posible generar la comprobación contra los valores reales de O_3 por día.

**Figura .2:** Estimaciones puntuales obtenidas para el primer día de Enero**Figura .3:** Estimaciones puntuales obtenidas para el primer día de Enero

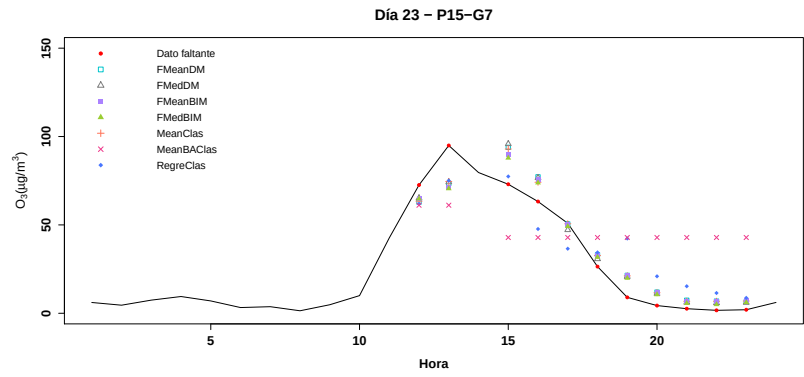


Figura .4: Estimaciones puntuales obtenidas para el primer día de Enero

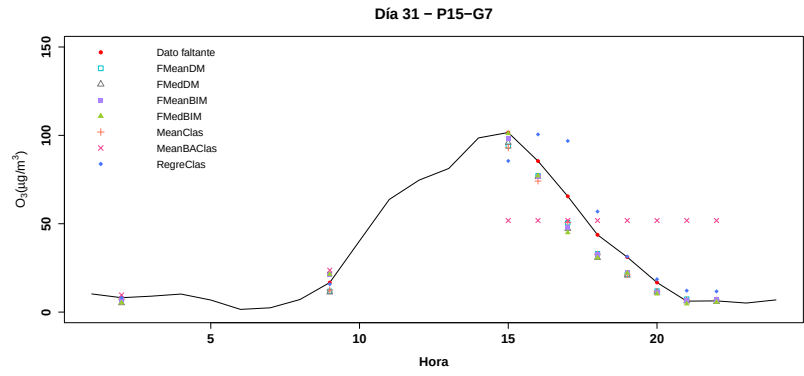


Figura .5: Estimaciones puntuales obtenidas para el primer día de Enero

Bibliografía

- Aguilera, M. d. C. (2009). Estimación penalizada de datos funcionales. Tesis de Maestría - Universidad de Granada.
- Ataz, E. M. and de Mera Morales, Y. D. (2004). *Contaminación atmosférica*, volume 45. Universidad de Castilla La Mancha.
- Bande, M., Oviedo de la Fuente, M., Galeano, P., Nieto, A., and Garcia, E. (2016). Functional data analysis and utilities for statistical computing. *R: A Language and Environment for Statistical Computing*.
- Baraldi, A. N. and Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology*, 48(1):5–37.
- Bermejo, V., Alonso, R., Elvira, S., Rábago, I., and García, M. (2010). El ozono troposférico y sus efectos en la vegetación. *Ministerio de Medio Ambiente Y Medio Rural Y Marino. Madrid*, 74p.
- Caballero, M., Lozano, S., and Ortega, B. (2007). Efecto invernadero, calentamiento global y cambio climático: una perspectiva desde las ciencias de la tierra. *Revista digital universitaria*, 8(10):1–12.
- Caicedo, F. and Jiménez, C. (2016). *Imputación basada en análisis de datos funcionales de observaciones faltantes de contaminación atmosférica por partículas finas suspendidas en el aire*. Pregrado, Universidad del Valle.
- Carbonell, N. and Stephanidis, C. (2012). European research consortium for informatics and mathematics workshop adjunct proceedings. pages 210–p.
- Conesa Fernández, V., Ros Garro, V., Conesa Ripoll, V., and Conesa Ripoll, L. A. (1995). *Guía metodológica para la evaluación del impacto ambiental*. Mundi-prensa.
- Conte, S. D., Dunsmore, H. E., and Shen, V. Y. (1986). *Software engineering metrics and models*. Benjamin-Cummings Publishing Co., Inc.
- Delgado Saborit, J. M. (2005). Validación e implementación de técnicas de captación pasiva para el estudio de los niveles y efectos de ozono troposférico y dióxido de nitrógeno en un área costera mediterránea.
- Delicado, P. (2008). Curso de modelos no paramétricos. *Departamento de Estadística Investigación Operativa, Universidad Politécnica de Cataluña*. http://www-eio.upc.es/~delicado/docencia/Apuntes_Models_No_Parametrics.pdf.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.
- Dumedah, G. and Coulibaly, P. (2011). Evaluation of statistical methods for infilling missing values in high-resolution soil moisture data. *Journal of Hydrology*, 400(1):95–102.

- Etze, R. A. and French, J. G. (2000). Contaminación del aire. *Impacto de los desastres en la salud public salud publica*, page 337.
- Eubank, R. L. (1999). *Nonparametric regression and spline smoothing*. CRC press. 360p, New York, USA.
- Galván, M. (2007). *Imputación de datos: teoría y práctica*, volume 54. United Nations Publications.
- Gentleman, R. and Ihaka, R. (2017). R: A language and environment for statistical computing.
- He, Y., Yucel, R., and Raghunathan, T. E. (2011). A functional multiple imputation approach to incomplete longitudinal data. *Statistics in medicine*, 30(10):1137–1156.
- Jaramillo, M., Núñez, M., Ocampo, W., Pérez, D., and Portilla, G. (2004). Inventario de emisiones de contaminantes atmosféricos convencionales en la zona de cali-yumbo. *Revista Facultad de Ingeniería Universidad de Antioquia*, 31:38–48.
- Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., and Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. *Atmospheric Environment*, 38(18):2895–2907.
- Manteiga, W. G. and Vieu, P. (2007). Statistics for functional data. *Computational Statistics & Data Analysis*, 51(10):4788–4792.
- McKnight, P. E., McKnight, K. M., Sidani, S., and Figueredo, A. J. (2007). *Missing data: A gentle introduction*. Guilford Press.
- Medina Rivera, R. (2008). Estimación estadística de valores faltantes en series históricas de lluvia. Master’s thesis, Universidad Tecnológica de Pereira, Facultad de Ingeniería Industrial.
- MinAmbiental (2016a). Boletín climatológico del mes de Enero. Technical report, Ideam, Santiago de Cali.
- MinAmbiental (2016b). Boletín Mensual de Calidad del Aire de Santiago de Cali. Technical report, DAGMA, Santiago de Cali.
- Norazian, M. N., Shukri, Y. A., Azam, R. N., Mohd Mustafa Al Bakri, A., et al. (2008). Estimation of missing values in air pollution data using single imputation techniques.
- Olaya, J. (2012). *Métodos de regresión no paramétrica*. Editorial: Universidad del Valle. 109p, Cali, Colombia.
- Osborne, J. W. (2012). *Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data*. Sage.
- Otero, G. (2011). Imputación de datos faltantes en un sistema de información sobre conductas de riesgo. *Técnicas estadísticas*. Universidad de la Curuña. España.
- Pighin, M. and Ieronutti, L. (2008). A methodology supporting the design and evaluating the final quality of data warehouses. *International Journal of Data Warehousing and Mining (IJDWM)*, 4(3):15–34.
- Ramsay, J. and Silverman, B. (2005). *Functional data analysis*. Springer. 428p, New York, USA.
- Ramsay, J., Wickham, H., Graves, S., and Hooker, G. (2011). Fda: functional data analysis. *R package version*, 2(6).
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.
- Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons.
- Salvador, R. and Millán, M. (2003). Análisis histórico de las brisas en castellón. *Tethys*, 2:37–51.

- Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical methods in medical research*, 8(1):3–15.
- Shumake-Guillemot, J., Jalkanen, L., and Adhair-Rohani, H. (2014). Calidad del aire y salud humana, prioridad para una acción conjunta. *Boletín de la OMM*.
- Urrutia, J. A., Palomino, R., and Salazar, H. D. (2010). Metodología para la imputación de datos faltantes en meteorología. *Scientia et technica*, 3(46):44–49.
- Van Buuren, S. (2012). *Flexible imputation of missing data*. CRC press.
- Yang Lenyap, K. U., Shaadan, N., Deni, S. M., and Jemain, A. A. (2015). Application of functional data analysis for the treatment of missing air quality data. *Sains Malaysiana*, 44(10):1531–1540.