



MODELACIÓN MEDIANTE SERIES DE TIEMPO DE LA
CONTAMINACIÓN POR $PM_{2.5}$

ANDERSON CASTAÑEDA TRUJILLO

ESCUELA DE ESTADÍSTICA
FACULTAD DE INGENIERÍA
UNIVERSIDAD DEL VALLE
SANTIAGO DE CALI

2016

**MODELACIÓN MEDIANTE SERIES DE TIEMPO DE LA
CONTAMINACIÓN POR PM_{2.5}**

ANDERSON CASTAÑEDA TRUJILLO

**Propuesta de proyecto de grado presentado
como requisito para optar al título de
ESTADÍSTICA**

DIRECTOR:

JAVIER OLAYA OCHOA

**ESCUELA DE ESTADÍSTICA
FACULTAD DE INGENIERÍA
UNIVERSIDAD DEL VALLE
SANTIAGO DE CALI**

2016

*“A mis padres por ser el motor de
mi vida y la fuerza que me impulso
a salir adelante. ”*

Anderson Castañeda Trujillo

Agradecimientos

Doy las gracias principalmente a Dios porque guiado por su amor, esperanza y sabiduría logré terminar esta etapa de mi vida. A mi Familia por apoyarme día a día, creer en mí y ayudarme a salir adelante. A cada profesor que formó parte de este camino, en especial al profesor Javier Olaya, mi Tutor de tesis, que con su infinito conocimiento no solo guió este trabajo, sino que ayudó a construir los cimientos de un buen profesional. De igual manera, una mención muy especial al profesor Wilmar Torres, que a pesar de no formar parte de la escuela de Estadística, siempre me brindó su apoyo y conocimientos de la mejor manera para sacar este proyecto adelante. Por último, mil y mil gracias a todos mis compañeros de carrera, mis amigos y las personas que con su compañía y motivación hicieron de este sueño una realidad.

Resumen

El material particulado de diámetro aerodinámico menor que 2.5 micrómetros ($PM_{2.5}$) ha sido catalogado como uno de los contaminantes de mayor impacto en la salud humana, siendo considerado un potencial generador de graves enfermedades cardiovasculares y respiratorias. En este trabajo se propone la construcción de un modelo de series de tiempo que permita realizar el pronóstico de las concentraciones de material particulado fino, generando así la posibilidad de contar con alertas tempranas que sirvan de insumo en políticas y acciones para el beneficio de la comunidad. En el periodo de estudio se identificaron un total de 27 registros diarios faltantes, generando la necesidad de aplicar métodos de imputación que permita completar la serie de tiempo, este proceso se desarrolla en dos etapas, la primera corresponde a la implementación de un modelo de Mínimos Cuadrados Ponderados tomando como referencia una variable altamente correlacionada con el $PM_{2.5}$ como es el caso del material particulado de diámetro aerodinámico menor que 10 micrómetros (PM_{10}); la segunda etapa se basa en la metodologías propuesta por Castaño, E (2007), el cual propone tratar una observación faltante como un valor Atípico Aditivo. Una vez completa la serie se implementa la metodología Box y Jenkins para la construcción del modelo final, teniendo en cuenta las adecuaciones necesarias como la inclusión de observaciones atípicas y cambios estructurales. Como resultado se obtuvo un modelo de medias móviles estacional multiplicativo de la forma $ARMA(0, 1)(0, 1)[7]$ con sus respectivas variables exógenas, el cual presento pronósticos plausibles en cuanto a nivel, siguiendo el comportamiento esperado de la serie.

Palabras Clave: *Material Particulado, Imputación de Datos, Contaminación Atmosférica, Análisis de Intervención, Observaciones atípicas, Cambio de nivel*

Tabla de Contenido

Introducción	1
1. Planteamiento del problema	3
2. Justificación	6
3. Objetivos	8
3.1. Objetivo General	8
3.2. Objetivos Específicos	8
4. Antecedentes	9
5. Marco Teórico	12
5.1. Contaminación Atmosférica	12
5.2. Material Particulado menor a 2.5 micras ($PM_{2.5}$)	12
5.3. Monitoreo de la calidad del aire en Santiago de Cali	13
5.4. Imputación de Datos	13
5.4.1. Modelos de regresión lineal	14
5.4.2. Reconstrucción de series de tiempo en presencia de datos atípicos . .	15
5.5. Análisis de intervención	18
5.5.1. Técnica del análisis de intervención	18

5.5.2.	Valores atípicos	19
5.5.3.	Atípicos aditivos (AO)	20
5.5.4.	Atípicos innovativos (IO)	21
5.5.5.	Cambio de nivel (LS)	21
5.6.	Series de Tiempo	22
5.6.1.	Ruido Blanco (RB)	23
5.6.2.	Identificación de la serie de tiempo	23
5.6.3.	Determinación de la transformación para estabilizar la varianza . . .	23
5.6.4.	Estabilización en nivel de la serie	24
5.6.5.	Identificación de la estructura ARMA	25
5.6.6.	Modelos Autorregresivos AR(p)	25
5.6.7.	Modelo de medias móviles MA(q)	26
5.6.8.	Estimación de los parámetros del modelo	27
5.6.9.	Verificación de la serie de tiempo	28
6.	Metodología	32
6.1.	Revisión bibliográfica	32
6.2.	Obtención y depuración de la base de datos	33
6.3.	Análisis exploratorio de los datos	33
6.4.	Reconstrucción de la base de datos	34
6.4.1.	Etapa 1: Regresión lineal por Mínimos Cuadrados Ponderados . . .	34
6.4.2.	Etapa 2: Estimación de datos faltantes por el método de Castaño . .	34
6.4.3.	Etapa 3: Estimación de datos faltantes por Sub-Series de tiempo . .	35
6.5.	Construcción y validación del modelo	35
6.6.	Pronósticos de la serie tiempo	36

7. Resultados	37
7.1. Análisis Exploratorio	37
7.1.1. Análisis de la serie por tramos	41
7.1.2. Transformación de la Serie del $PM_{2.5}$	43
7.2. Reconstrucción de la serie del $PM_{2.5}$	45
7.2.1. Regresión por Mínimos Cuadrados Ponderados (MCP)	46
7.2.2. Análisis de intervención como método de imputación	47
7.3. Modelación de la serie reconstruida	51
7.3.1. Varianza Constante y evaluación de existencia de observaciones atípicas	56
7.3.2. Contraste de Ljung-Box sobre las autocorrelaciones	58
7.3.3. Normalidad de los residuales del modelo	59
8. Conclusiones	62
9. Recomendaciones	64
Anexos	65
A. Relación contaminantes y variables meteorológicas	65
A.1. Rosa de viento de la estación Pance - Nivel de polución del PM_{10}	66
A.2. Correlación del $PM_{2.5}$ con las variables meteorológicas	66
A.3. Dirección del viento estación Univalle y estación de Pance	67
B. Análisis de los residuales de los tramos de la serie	68
B.1. FAC, FACP y evaluación de los supuestos del Tramo 1	68
B.2. FAC, FACP y evaluación de los supuestos del Tramo 2	69
B.3. FAC, FACP y evaluación de la Subserie	70

C. Análisis de los residuales del modelo filtrado completo	71
C.1. FAC, FACP y evaluación de los supuestos de la serie completa	71
C.2. Intervalos para el ajuste de la serie de tiempo	72

Lista de Tablas

5.1. Normatividad $PM_{2.5}$ Resolución 610 del 2010	13
5.2. Daniel Peña 2005. Comportamiento de la FAC y la FACP	25
6.1. Normatividad $PM_{2.5}$ Resolución 610 del 2010	33
7.1. Estadísticas descriptivas para el PM2.5 Horario	37
7.2. Estadísticas descriptivas para el $PM_{2.5}$ Diario	39
7.3. Estadísticas descriptivas por Tramo	42
7.4. Parametros de la Regresión por Mínimos Cuadrados Ponderados	47
7.5. Intervenciones Tramo 1	48
7.6. Observaciones Faltantes Estimadas	49
7.7. Intervenciones Tramo 2	50
7.8. Intervenciones de los datos faltantes del Tramo 2	50
7.9. Intervenciones en la serie completa del $PM_{2.5}$	52
7.10. Función de Autocorrelación Extendida	53
7.11. Criterios de selección del modelo	54
7.12. Parámetros de las intervenciones Completas $PM_{2.5}$	55
7.13. Pronósticos modelo ARMA(0,1)(0,1)[7] $PM_{2.5}$	60
A.1. Supuestos Serie Filtrada Completa	65

A.2. Correlación entre contaminante y variables meteorológicas	66
B.1. Supuestos Tramo 1	69
B.2. Supuestos Tramo 2	69
B.3. Supuestos Series Subtramo	70
B.4. Pronósticos modelo MA(1) para la imputación de 2 datos faltantes	70
C.1. Supuestos Serie Filtrada Completa	72

Lista de Figuras

7.1. Serie de tiempo horaria para el $PM_{2.5}$	38
7.2. Diagrama de cajas horarias para el $PM_{2.5}$	39
7.3. Concentraciones promedios diarios de $PM_{2.5}$	40
7.4. Concentraciones diarias de $PM_{2.5}$ por día de la semana	41
7.5. Concentraciones $PM_{2.5}$ por Tramos	42
7.6. Transformación Box-Cox para cada tramo	43
7.7. Histogramas $PM_{2.5}$ con transformaciones	44
7.8. Serie Original	45
7.9. Serie Trans Raíz Cuadrada	45
7.10. Estimaciones $PM_{2.5}$ por MCP	46
7.11. Imputación mediante MCP	46
7.12. Histogramas $PM_{2.5}$ con transformaciones	48
7.13. Tramo 2 recortado para análisis de intervención	49
7.14. Valores imputados de $PM_{2.5}$ según método empleado	51
7.15. Serie de $PM_{2.5}$ completa	52
7.16. Serie Filtrada - FAC y FACP	53
7.17. Raíces inversas del polinomio MA	55
7.18. FAC y FACP residuales del modelo $ARMA(0,1)(0,1)[7]$ mas las intervenciones	56

7.19. Supuesto varianza constante	57
7.20. Supuesto varianza constante	58
7.21. Gráfica de P-Valores de la prueba de Ljung-Box con k autocorrelaciones . .	59
7.22. Q-Q Plot y Histograma para evaluar normalidad en los residuales	60
7.23. Predicciones modelo SARIMA(0,0,1)(0,0,1)[7]	61
A.1. Rosa de viento de la estación Pance	66
A.2. Mapa ubicación de las estaciones y dirección del viento	67
B.1. Serie Filtrada, FAC y FACP Tramo 1	68
B.2. Serie Filtrada, FAC y FACP Tramo 2	69
B.3. FAC y FACP Subserie	70
C.1. FAC y FACP residuales modelo ARMA(0,1)(0,1)[7]	71
C.2. Intervalos de ajuste Serie Completa	72

Introducción

La contaminación atmosférica a nivel mundial sigue en aumento, la industrialización de las ciudades y el crecimiento del parque automotor contribuyen a que la polución en el aire empeore, superando los estándares recomendados por las entidades ambientales y afectando a más del 90 % de la población mundial según reportes de la Organización Mundial para la Salud (OMS), cobrando aproximadamente 7 millones de muertes cada año (OMS 2014). La EPA y la OMS hacen especial énfasis en controlar aquellos contaminantes que por su reducido diámetro son capaces de afectar el organismo directamente, como es el caso del Material Particulado Menor a 2.5 micras ($PM_{2.5}$), el cual por su tamaño al ser inhalado alcanza zonas periféricas de los bronquios afectando el intercambio de gases, aumentando las probabilidades de enfermedades cardiovasculares y respiratorias, así como el cáncer de pulmón y en caso de estar sometidos a altos niveles por un tiempo prolongado puede llegar a ocasionar la muerte (EPA 2005).

El principal objetivo de estudio es ajustar un modelo estadístico que permita realizar estimaciones de la contaminación por $PM_{2.5}$ en el municipio de Santiago de Cali, empleando a su vez las herramientas estadísticas complementarias para la reconstrucción de la serie de datos y las validaciones necesarias. Considerando apropiado utilizar la metodología introducida por Box y Jenkins para el modelamiento de la serie de tiempo, tomando como referencia estudios aplicados tanto en el ámbito nacional e internacional, como son los publicados por Castaño, E (2007), Kumar & Ridder (2014), García & Gonzalez (2014), Chen & Liu (1993), McAleer & Chan (2006), Cadena & García (2014), entre otros autores.

Una de las dificultades más relevantes a la hora de utilizar series de tiempo es la presencia de valores faltantes, donde las fallas en los instrumentos de medición, la ausencia de energía eléctrica o mantenimiento de los equipos son los principales motivos que generan vacíos en la continuidad de la información. La reconstrucción de la serie de tiempo se realiza en dos fases, la primera consiste en emplear modelos de regresión lineal (Robustos en presencia de heterocedasticidad) tomando como referencia una variable altamente correlacionada con el contaminante de estudio y una segunda fase basada en el procedimiento planteado por

Castaño, E (2007), el cual propone tratar las observaciones faltantes como valores atípicos aditivos.

Finalmente, con la serie reconstruida, se plantea el análisis de intervención para la identificación de observaciones atípicas y cambios estructurales, que serán excluidos de la serie de tiempo para la modelación del ruido de la serie, planteando un conjunto de modelos por medio de la metodología Box y Jenkins, evaluando por los criterios de información y los supuestos de los residuales el mejor modelo que permita las predicciones más precisas.

Capítulo 1

Planteamiento del problema

El aire limpio es un derecho del cual deben gozar los seres humanos en el día a día, sin embargo, en las últimas décadas la contaminación atmosférica ha crecido alarmantemente, al punto de ser considerada una de las mayores problemáticas en la actualidad, pasando de una tasa de 2 millones de personas muertas anualmente en el 2005 a 7 millones en el 2012 (OMS 2005). Las principales fuentes de contaminación atmosférica en el planeta son atribuidas a las grandes industrias y su falta de control ambiental, la deforestación, el tránsito pesado en las ciudades, las construcciones y demoliciones, los productos químicos entre otras.

La Agencia de Protección Ambiental de los Estados Unidos (EPA por sus siglas en inglés), se ha convertido en un referente internacional, realizando investigaciones y estableciendo normas para la conservación de un medio ambiente sano que garantice la protección de la salud en la sociedad. La EPA basándose en estudios ambientales propone el Índice de Calidad del Aire (ICA) como una medida representativa de la calidad atmosférica en las ciudades, conformado por 6 contaminantes prioritarios a ser regulados: Ozono (O_3), Material particulado de diámetro menor a diez micras (PM_{10}) y menor a 2.5 micras ($PM_{2.5}$), Dióxido de Azufre (SO_2), Dióxido de Nitrógeno (NO_2) y el Monóxido de Carbono (CO) (EPA 2009) .

Las Partículas Suspendidas Totales (PST) oscilan entre 0.0002 y 500 micrómetros (μm), siendo las menores a 10 micras y 2.5 micras las de mayor frecuencia en la medición y monitoreo por las autoridades ambientales debido a el daño que pueden ocasionar al organismo humano al estar expuesto en un tiempo considerable. Por lo tanto, el material particulado menor a dos micras ($PM_{2.5}$) es el contaminante más peligroso de los contaminante prioritarios medidos por la EPA, formado principalmente por los procesos de combustión (OMS

2005). El $PM_{2.5}$ por su diámetro aerodinámico se convierte en un contaminante peligroso para la salud humana, debido a que al ser inhalado puede alcanzar zonas periféricas de los bronquiolos y alterar el intercambio pulmonar de gases. Además la exposición crónica al $PM_{2.5}$ genera enfermedades cardiovasculares y respiratorias, como es el caso del cáncer de pulmón (OMS 2005).

En Colombia el Ministerio del Medio Ambiente y Desarrollo Sostenible ajusta la norma ambiental planteada por la EPA para el manejo de la calidad del aire a los estándares nacionales, impartiendo por medio de leyes, resoluciones, acuerdos, decretos, entre otros, la normatividad con los lineamientos necesarios para el monitoreo y control de la calidad atmosférica en el país. Estos lineamientos son dirigidos hacia todas las corporaciones y autoridades ambientales, con el propósito de reglamentar las mediciones de las emisiones de contaminantes en las ciudades. En Santiago de Cali el Departamento Administrativo de la Gestión del Medio Ambiente (DAGMA) es la máxima autoridad ambiental, encargada de monitorear, regular y ejecutar acciones que protejan los recursos naturales y el desarrollo sostenible del medio ambiente. El DAGMA mediante su grupo de Calidad del Aire realiza las mediciones de los contaminantes prioritarios considerados por la EPA y las variables meteorológicas más importantes que influyen directamente en su propagación y aumento; las mediciones se hacen a través de 9 estaciones de monitoreo de calidad del aire distribuidas estratégicamente por la ciudad, con el fin de medir todas las zonas (Norte, Sur, Oriente y Occidente) y así obtener resultados representativos del aire que respiran los ciudadanos.

La inclusión de medidores automáticos de partículas finas ($PM_{2.5}$) en la ciudad de Santiago de Cali se realizó en el año 2013, implementando esta tecnología en dos estaciones de monitoreo, la primera ubicada en la Universidad del Valle y la segunda en la Base Aérea. Estas estaciones de monitoreo atmosférico realizan mediciones de los contaminantes cada hora, datos que son enviados por medio de una conexión de banda ancha al centro de control ubicado en el grupo de calidad del aire del DAGMA; el cual procesa los datos y realiza reportes mensuales que publica en su portal virtual para que la ciudadanía esté enterada del control que se ejerce a la contaminación atmosférica. No obstante, la información que se brinda es actual y de días previos, es decir, no se realizan predicciones a cerca de un comportamiento del contaminante en el futuro.

Dada la necesidad de realizar predicciones de las concentraciones de ($PM_{2.5}$), se hace indispensable contar con un modelo estadístico que se ajuste a la estructura del contaminante, teniendo en cuenta los factores externos como datos atípicos y cambio estructurales y los problemas de estacionalidad que puedan presentarse por las variaciones del contaminante en algunos días de la semana, considerando la metodología Box y Jenkins como una alter-

nativa plausible para la construcción del modelo deseado que permita realizar pronósticos acertados, apoyándose de otras herramientas estadísticas que aporten a la reconstrucción de los datos, manteniendo la estructura temporal de estos.

Capítulo 2

Justificación

La contaminación atmosférica se ha convertido en una de las principales causas de muerte prematura en el mundo y un factor determinante en la generación de enfermedades respiratorias. Motivo por el cual, entidades a nivel mundial como la Organización Mundial para la Salud (OMS) y la Agencia de Protección Ambiental de los Estados Unidos (EPA), prendieron las alarmas por el incremento de los principales contaminantes que afectan la salud humana y deterioran la calidad del medio ambiente.

Así mismo, la Agencia Internacional para la Investigación sobre el Cáncer (AIRC) ha realizado una serie de estudios en los últimos años, para comprobar la relación que tiene la contaminación atmosférica con la presencia de cáncer de pulmón en personas expuestas prolongadamente a altos niveles de contaminación. En el 2010 resultados publicados por la IARC indicaron que 223,000 muertes por cáncer de pulmón se debieron a la contaminación atmosférica (IACR 2013). En un panorama nacional, en Colombia mueren más de seis mil personas al año por la contaminación del aire en las grandes ciudades, además de dejar pérdidas millonarias por la ausencia laboral a causa de las enfermedades respiratorias y los gastos que estos generan a las entidades de salud del país (Minambiente 2002).

Uno de los mayores generadores de contaminación ambiental en las ciudades es el uso vehículos motorizados, influyendo drásticamente el tipo y la calidad de combustible, el estado técnico mecánico del vehículo y el aumento del parque automotor. En Santiago de Cali el número de vehículos ha incrementado drásticamente en los últimos años, pasando de 285,335 vehículos en el 2004 a 449,517 vehículos en el 2011. En el año 2012 la ciudad contaba con un parque automotor de aproximadamente 597,968 vehículos, distribuido en 150,978 motos, 55,130 vehículos de transporte público, 3422 vehículos de transporte oficial y 388,438 vehículos particulares (Cali Como Vamos 2013), responsables de generar un 90 %

de las emisiones de NO_X , CO y CO_2 , un 76% de los Compuestos Orgánicos Volátiles (COVS) y un 26% de las Partículas Suspendidas Totales (PST) (Informe de Gestión DAGMA). Otras fuentes relevantes en las emisiones atmosféricas en Santiago de Cali son las fuentes fijas atribuidas al sector industrial, las cuales generan aproximadamente 11422 Ton/año de PST, 5,799 Ton/año de PM_{10} , 573 Ton/año de $PM_{2.5}$ y 690 Ton/año de SO_X (DAGMA 2013).

Es importante tener en cuenta que la exposición constante a la contaminación atmosférica, presenta un peligro para la humanidad, considerando las partículas finas como el contaminante más dañino según la Agencia de Protección Ambiental Estadounidense (EPA). Este Contaminante puede generar problemas a la salud como ataques cardíacos no mortales, latidos irregulares del corazón, agravamiento del asma y en algunos casos la muerte prematura en personas con enfermedad cardíaca o pulmonar. Motivos por los cuales se hace necesario para la ciudad implementar políticas y estrategias que permitan regular la exposición de la población a las partículas finas, una alternativa plausible es la creación de modelos estadísticos que permitan predecir los niveles de $PM_{2.5}$ en los siguientes días, mediante el monitoreo de su tendencia y la generación de acciones correctivas, cuando se presente un crecimiento por encima de la norma Colombiana.

En este proyecto se implementa la metodología desarrollado por Box y Jenkins para la modelación del contaminante ($PM_{2.5}$), tomado como referencia estudios que relacionan la contaminación atmosférica con esta metodología estadística como es el caso de Kumar & Ridder (2014), Krasnov, Katra & Friger (2015), McAleer & Chan (2006), Iglesias, Jorquera & Palma (2006), Agudelo, Teixeira & Pereira (2014), García & Gonzalez (2014), Cadena & García (2014) entre otros autores, los cuales resaltan la importancia de realizar control de los contaminantes atmosféricos por medio de modelos que permitan predecir las concentraciones futuras de estos para la toma de decisiones que disminuyan el riesgo de afecciones a la salud. De igual manera, se hace indispensable emplear herramientas estadísticas que permitan reconstruir la serie del $PM_{2.5}$ y a su vez tener en cuenta las variables exógenas implícitas en la estructura de los datos, para lo cual se emplea las metodologías propuestas por Chen & Liu (1990) y Castaño, E (2007).

Capítulo 3

Objetivos

3.1. Objetivo General

Pronosticar las concentraciones diarias de material particulado menor a 2.5 micrómetros ($PM_{2.5}$) en la ciudad de Santiago de Cali.

3.2. Objetivos Específicos

- Caracterizar el comportamiento del $PM_{2.5}$ entre días de la semana, meses y estaciones de monitoreo.
- Aplicar métodos de imputación de datos para completar la serie del contaminante $PM_{2.5}$.
- Abordar los problemas de intervención que presente la serie de tiempo.
- Proponer un modelo de series de tiempo que permita explicar el comportamiento de las concentraciones promedio diarias del contaminante $PM_{2.5}$.

Capítulo 4

Antecedentes

- ***Iglesias et al. (2006)***: Los autores plantean una metodología para realizar inferencia acerca de los coeficientes para los modelos de memoria larga (ARFIMA) cuando se tiene bases de datos con valores faltantes mediante la utilización de métodos clásicos y bayesianos. El procedimiento es aplicado a la modelación de la contaminación atmosférica, en especial a las fuentes que contribuyen la generación de material particulado (PM) debido a la peligrosidad de este contaminantes para la salud humana. Los datos son brindados por la Red de Monitoreo de Calidad del Aire de Santiago de Chile (MACAM), los contaminantes que se tiene en cuenta en el estudio son los siguientes: Partículas Finas ($PM_{2.5}$), Monóxido de Carbono, Dióxido de Azufre, Dióxido de Nitrógeno, Óxido Nítrico (NO), Hidrocarburos Totales de THC, Ozono y algunas variables meteorológicas como la precipitación y las velocidad del viento. Uno de los grandes inconvenientes en estas bases de datos son la cantidad de datos faltantes, debido a que los registros se toman en el periodo comprendido entre Enero de 1989 a Diciembre de 1995 contando solo con 531 casos diarios. La metodología aplicada se divide en las siguientes etapas: a) Se usan los residuales obtenidos de la estimación de Mínimos Cuadrados Ordinales para estimar los parámetros del modelo de error b) se aplica una metodología llamada el filtro de Kalman para estimar el modelo ARFIMA en presencia de datos faltantes c) se utiliza un modelo de regresión gaussiana con una matriz de covarianzas siguiendo las especificaciones ARFIMA d) se aplican métodos clásicos y bayesianos para la estimación de los coeficientes de regresión del modelo adoptado.
- ***Castañó, E (2007)***: En su artículo denominado Reconstrucción de datos de series de tiempo: Una aplicación a la Demanda horaria de la Electricidad presenta un procedimiento que permite estimar el efecto de las intervenciones, de las observaciones

atípicas, estimar las observaciones faltantes y simultáneamente identificar el proceso ARIMA que generó el proceso aleatorio. Como resultado se tiene una serie reconstruida con estimaciones óptimas para los datos faltantes y para los efectos de las observaciones atípicas. La metodología aplicada se divide en dos etapas: La primera consiste en la caracterización de una observación faltante como una observación atípica adictiva por medio de un modelo ARIMA de orden alto y la segunda etapa se construye el modelo usando las estimaciones preliminares y se realizan de nuevo las estimaciones de las observaciones atípicas y faltantes. La aplicación del procedimiento se realiza a una serie de datos sobre la demanda de electricidad horaria, la cual presenta los tres problemas mencionados anteriormente. La finalidad del estudio es construir un modelo para pronosticar la demanda futura de cada una de las horas del día. Como conclusiones los autores resaltan la importancia de la aplicación de esta metodología comparado con los métodos de imputación de datos comúnmente utilizados como son los realizados por promedio de observaciones vecinas o estacionalmente vecinas, debido a que estas metodologías pueden generar problemas en la identificación si existen observaciones atípicas.

- **Valdes & Torres (2009):** Plantean una serie de tiempo para modelar el precio diario de energía eléctrica en Colombia entre el 1 de Enero de 2002 y el 30 de Junio de 2008. Con el fin de tener en cuenta el comportamiento de los precios en cada una de las horas del día, los autores deciden trabajar con los precios diarios de energía eléctrica ponderados por la demanda, debido a que el comportamiento del precio varía según la hora del día. La modelación de la serie de tiempo se realizó mediante un modelo Autorregresivo Integrado de Media Móvil Estacional (SARIMA), en donde el modelo ARIMA (1,1,2)(1,1,1)[7] es el que posee un mejor ajuste y predicciones más acertadas. Para corregir el problema de múltiple estacionalidad los autores deciden implementar un modelo de series de tiempo por día, generando ventajas en las predicciones del precio diario de energía, ya que se contarían con 28 estimaciones en vez de las 4 que brinda el modelo completo. Como conclusiones se tiene que los modelos por días presentan mejores predicciones que las brindadas por el modelo completo, en especial los días correspondientes a los fines de semana (Sábado y Domingo) dando estimaciones plausibles con un intervalo de confianza más ajustados a la realidad.
- **Montoya & Orobio (2009):** Modelan el comportamiento de la concentración promedio horaria del material particulado menor a 10 micrómetros (PM_{10}) durante 24 horas teniendo en cuenta la estación que reporta los datos, el tipo de día (Ordinario, fin de semana y festivo) y el año. La modelación se llevó a cabo mediante regresión no paramétrica utilizando curvas de suavización LOESS en la ciudad de Santiago de

Cali para un periodo comprendido desde el 2003 al 2006. Los resultados obtenidos en el análisis muestran que la mayor concentración de PM_{10} se presentó en el año 2003, los días más contaminados fueron los ordinarios incluyendo el día sábado y la hora del día de mayor concentración de material particulado fue a las 10 de la mañana.

- ***García & Gonzalez (2014):*** Modelan las concentraciones máximas del Ozono (O_3) en la zona norte de Santiago de Cali, en la estación de monitoreo del Éxito de la Flora, para el periodo comprendido entre el 2010 y el 2011, haciendo uso de la metodología de series temporales desarrollada por Box y Jenkins. Debido a la presencia de valores faltantes y como condición necesaria para la utilización de los modelos ARIMA, es indispensable contar con una serie de datos completa, por tal razón los autores emplean el procedimiento propuesto por Castaño (2007), la cual permite identificar los efectos de las intervenciones de las observaciones atípicas y cambios estructurales. Los resultados obtenidos muestran que el modelo que mejor explica la variabilidad de las concentraciones máximas de ozono es el modelo de medias móviles de orden 1 (MA(1)) con sus respectivas observaciones atípicas e intervenciones, concluyendo que la concentración del máximo por día del ozono troposférico depende no solo de la información del pasado, sino de la nueva información que se genera el mismo día.
- ***Cadena & García (2014):*** Realizan la modelación del Dióxido de Azufre (SO_2) en Santiago de Cali tomando como periodo de referencia entre el 1 de Julio del 2010 hasta el 30 de Junio del 2011. Este contaminante perjudica principalmente a los grupos más susceptibles como los niños, ancianos y aquellos con enfermedades cardíacas o respiratorias. El modelo propuesto se ajusta por medio de la metodología Box y Jenkins, no obstante, por la presencia de datos faltantes es necesario realizar técnicas de imputación de datos, aplicando el procedimiento descrito por Castaño (2007) el cual permite realizar la reconstrucción de la serie y a su vez la detección de valores atípicos y cambios estructurales. Debido al comportamiento del SO_2 y la influencia que tiene la temperatura y las variaciones climatológicas en Santiago de Cali (Temporadas de sequías y de lluvias) que afectan las concentraciones del contaminante en cuanto a nivel, los autores deciden dividir la serie en tres segmentos para la imputación de valores faltantes y el análisis de intervención previo. El modelo final corresponde a un proceso de MA(2) con sus respectivas variables exógenas (Intervenciones), comportandose sus residuales como ruido blanco y obteniendo un ajuste del 89.2 % de sus valores estimados a los observados.

Capítulo 5

Marco Teórico

Este capítulo está conformado por dos partes, la primera describe los principales conceptos ambientales abordados en el estudio y una segunda parte que contiene las metodologías estadísticas empleadas para la depuración y el procesamiento de la información que lleven a la construcción de un modelo pertinente utilizando la herramienta de series de tiempo.

5.1. Contaminación Atmosférica

Se entiende como el fenómeno de acumulación o concentración de sustancias y elementos en estado sólido, líquido o gaseoso causante de problemas en el medio ambiente (Resolución 610 de 2010). Son formados por la intervención humana en las diversas actividades del diario vivir, principalmente las generadas por el sector industrial que implican proceso de combustión y la contaminación generada por el parque automotor. Los contaminantes más comunes emitidos a la atmósfera son: El Dióxido de Nitrógeno, Monóxido de Carbono, Dióxido de Azufre, Ozono, Material Particulado menor a 10 micras y a 2.5 micras (EPA), siendo el material particulado el contaminantes más dañino para la salud humana (OMS, 2005).

5.2. Material Particulado menor a 2.5 micras ($PM_{2.5}$)

También conocido como partículas finas, compuesto principalmente por sulfatos, nitratos, amoníaco, cloruro sódico, carbón, polvo de minerales y agua. El material particulado está compuesto por una mezcla de partículas líquidas y solidas suspendidas en el aire.

Son consideradas peligrosas para la salud humana debido a que al ser inhaladas pueden alcanzar zonas periféricas de los bronquiolos y alterar el intercambio pulmonar de gases. La exposición prolongada a este contaminante aumenta el riesgo de enfermedades cardiovasculares y respiratorias, así como el cáncer de pulmón y en algunos casos la muerte (OMS 2011). La norma nacional que regula los niveles de $PM_{2.5}$ es la Resolución 610 del 2010, la cual indica los niveles permisibles para cada contaminante y la periodicidad con que debe ser medida. En la siguiente tabla se enseñan los límites estipulados para el material particulado:

Contaminante	Nivel Máximo Permissible	Tiempo de Exposición
Material	25	Anual
Particulado 2.5	50	24 Horas

Tabla 5.1: Normatividad $PM_{2.5}$ Resolución 610 del 2010

5.3. Monitoreo de la calidad del aire en Santiago de Cali

El Departamento Administrativo de la Gestión del Medio Ambiente (DAGMA) es la autoridad ambiental de Santiago de Cali encargada de monitorear, controlar y proteger los recursos naturales del municipio. Desde el año 2003 el DAGMA mediante el Sistema de Vigilancia de la Calidad del Aire (SVCASC) monitorea la calidad atmosférica, contando actualmente con 8 estaciones de monitoreo de los principales contaminantes según lo determinado por el ministerio en la resolución 610 del 2010. El material particulado menor a 2.5 micras es monitoreada en dos estaciones del SVCASC ubicadas en el campus de la Universidad del Valle y en el interior de la Base Aérea, contando con datos horarios desde el año 2013, aunque debido a los mantenimientos que se le ha realizado a los equipos más las dificultades técnicas afrontadas no es posible contar con una serie de datos constantes, generando así registros faltantes.

5.4. Imputación de Datos

En los estudios sobre contaminación atmosférica es común encontrar registros faltantes, generados por problemas técnicos o logísticos en las estaciones de monitoreo, como es el caso de mantenimiento a los equipos de medición, fallas en los instrumentos, falta de energía eléctrica, entre otros. Con el fin de involucrar metodologías estadísticas en el análisis o modelamiento de contaminación atmosférica es necesario utilizar técnicas que permitan

llenar esos espacios faltantes con valores plausibles y pertinentes que sean acorde a la estructura de datos que se está manejando con el mínimo sesgo posible que no afecte la eficiencia de las inferencias del estudio (Wei 2006).

En el desarrollo de este proyecto se plantean dos métodos de imputación de valores faltantes, el primero hace alusión a un método recursivo como es el caso del modelo de regresión simple, al tener una variable altamente correlacionada con el $PM_{2.5}$ como es el caso del PM_{10} , mientras que el segundo método consiste en emplear el procedimiento planteado por Chen y liu (1993) y retomado por Castaño (2007) en su artículo “Reconstrucción de datos de series de tiempo: una aplicación a la demanda horaria de la electricidad”.

5.4.1. Modelos de regresión lineal

Los modelos de regresión lineal simple, son modelos matemáticos que representan el comportamiento de una variable Y , a través de una variable X , con el margen de error que generan las perturbaciones aleatorias ε_i , que reflejan todos los factores (Distintos de X) que influyen sobre la variable endógena, pero que ninguno de ellos es relevante individualmente. En la regresión lineal existen diferentes tipos de procedimientos para la estimación de parámetros como: Estimación por mínimos cuadrados ordinarios, estimación por máxima verosimilitud y los métodos de estimadores insesgados de varianza mínima.

Conforme a lo planteado por Gujarati (2010), el procedimientos de Mínimos Cuadrados atribuido a Carl Friedrich Gauss, consiste en minimizar la suma de los cuadrados de las distancias verticales entre los valores de los datos y la regresión estimada, es decir, minimizar la suma de los residuos al cuadrado, teniendo como residuo la diferencia entre los valores observados y los estimados por el modelo.

Mínimos Cuadrados Ponderados (MCP)

Uno de los supuestos de la mayoría de métodos de modelación lineal, es que cada dato proporciona información igualmente precisa sobre la parte determinística de la variación total del proceso. En otras palabras, la desviación estándar del término de error es constante en todos los valores del predictor o variable explicativa (Gujarati 2010). Aunque es común que esta hipótesis no siempre se cumpla, y que la variación del promedio de las observaciones sea distinta en diferentes puntos del proceso, en donde, en situaciones como esta los MCP puede maximizar la eficiencia de estimación de los parámetros, dándole

a cada observación una ponderación adecuada, la cual cuando σ_i es conocida los pesos corresponderán al inverso de la varianza, donde:

$$\omega_{i,j} = \frac{1}{\hat{\sigma}_i^2} = \left[\frac{\sum_{j=1}^{n_i} (y_{i,j} - \bar{y}_i)^2}{n_i - 1} \right]^{-1} \quad (5.1)$$

Quedando el modelo de regresión de la siguiente manera:

$$\frac{Y_i}{\sigma_i} = \hat{\beta}_1 \left(\frac{1}{\sigma_i} \right) + \hat{\beta}_2 \left(\frac{X_i}{\sigma_i} \right) + \left(\frac{\varepsilon_i}{\sigma_i} \right) \quad (5.2)$$

5.4.2. Reconstrucción de series de tiempo en presencia de datos atípicos

Se plantean una metodología alterna para el manejo de datos faltantes aplicados en las series de tiempo, la metodología es propuesta por Elkin Castaño (2007) retomando el estudio de autores como Chen y Liu (1993) la cual se divide en dos etapas:

Caracterización de una observación faltante como una observación atípica aditiva

Suponga que Z_t es una serie de tiempo que sigue un proceso ARIMA de la forma:

$$Z_t = \frac{\theta(B)}{\delta(B)\phi(B)} a_t$$

donde B es el operador de rezagos usual, $\theta(B)$ es el polinomio de medias móviles con todas sus raíces fuera del círculo unidad, $\phi(B)$ es el polinomio autorregresivo y $\delta(B)$ es el polinomio de diferencias. En primer lugar, se considera el caso donde hay solamente una observación faltante. Para esto, suponga que la serie se observó durante T períodos y que no se encuentra disponible la observación para el periodo $t = t^*$, $1 < t^* < T$. Una caracterización natural de un valor faltante es describirlo como una observación atípica aditiva (Chen & Liu 1990). Suponiendo que en el periodo $t = t^*$ ocurre una observación atípica aditiva, podemos representar la serie observada como:

$$NZ_t = \begin{cases} Z_t & \text{si } t \neq t^* \\ AF_t^{t^*} + Z_t & \text{si } t = t^* \end{cases} \quad (5.3)$$

Donde A indica la desviación desde el verdadero valor de Z_t , y la variable $F_t^{t^*}$ toma el

valor de uno cuando $t = t^*$ y de cero en otro caso. El valor ajustado para NZ_t^* según (Chen & Liu 1990) es el siguiente:

$$NZ_t^* = \frac{\sum_{j=1}^{t-1} \left[\sum_{k=j}^{n-t^*} \pi_k \pi_{k-j} \right] NZ_{t^*-j} + \sum_{j=1}^{n-t^*} \left[\sum_{k=j}^{n-t^*} \pi_k \pi_{k-j} \right] NZ_{t^*+j}}{\sum_{k=1}^{n-t^*}} \quad (5.4)$$

Donde los coeficientes π son obtenidos del polinomio autorregresivos $\pi(B) = 1 - \pi_1 B - \pi_2 B^2 - \dots = \phi(B)/\theta(B)$.

El resultado anterior sugiere que podemos estimar un valor faltante en una serie de tiempo, tratándola como si fuera una observación atípica adictiva. por tanto, si conocemos a $\theta(B)$, $\phi(B)$ y $\delta(B)$, el procedimiento para estimar el valor desconocido de Z_t^* consiste en asignar un valor atípico cualquiera a la observación faltante y estimar el modelo intervenido:

$$NZ_t = \sum_{j=1}^m AF_t^{t^*} + \frac{\theta(B)}{\delta(B)\phi(B)} a_t \quad (5.5)$$

Sin embargo, el problema en la práctica es más complicado pues, en general, no se dispone del conocimiento de $\theta(B)$, $\phi(B)$ y $\delta(B)$, y debemos tratar de identificarlos a partir de la información completa inicial (Castaño, E 2007).

El procedimiento de reconstrucción

A continuación se enumeran las etapas a seguir en la metodología propuesta por Chen & Liu (1993) y retomado por (Castaño, E 2007) para la reconstrucción de la serie de tiempo:

1. A cada una de las m observaciones faltantes asigne un valor que sea atípico. Aproxime a Z_t usando las intervenciones de un proceso puro autorregresivo de orden alto. En la elección del orden, se debe tener en cuenta la frecuencia del período de observación, y si el proceso estacional o no. Z_t puede aproximarse como un proceso autorregresivo puro:

$$Z_t = \sum_{i=1}^k v_i(B) I_t^{t_j^*} \sum_{j=1}^m AF_t^{t^*} + \frac{1}{\delta(B)\phi(B)} a_t \quad (5.6)$$

Donde $v(B)$ es la función de transferencia correspondiente a la i -ésima intervención, $i = 1, \dots, k$, $I_t^{t_i}$ es una variable indicadora la cual toma el valor de 1 si $t = t_i$ y cero en otro caso; $F_t^{(t_j^*)}$ es otra variable indicadora que toma el valor de 1 cuando el periodo es $t = t_i^*$, el periodo donde ocurre la j -ésima observación faltante.

2. Especifique un nuevo modelo en el cual se adicionan h observaciones atípicas detectadas. El nuevo modelo por estimar es:

$$Z_t = \sum_{i=1}^k v_i(B) I_t^{t_i} + \sum_{j=1}^m A F_t^{t_j^*} + \sum_{s=1}^h v_s^*(B) O_t^{(t_s)} + \frac{1}{\delta(B)\phi(B)} a_t \quad (5.7)$$

donde $v_i(B)$ es la función de transferencia para la s -ésima observación atípica detectada y $O_t^{t_s}$ es una variable indicadora que toma el valor de 1 si $t = t_s$, el periodo donde ocurre la s -ésima observación atípica.

3. Filtre a Z_t por el modelo estimado sin incluir el modelo del ruido. Esto es lo mismo que estimar el modelo restringido:

$$Z_t = \sum_{i=1}^k v_i(B) I_t^{t_i} + \sum_{j=1}^m A F_t^{t_j^*} + \sum_{s=1}^h v_s^*(B) O_t^{(t_s)} + \eta_t \quad (5.8)$$

donde el valor de los parámetros es restringido a ser igual a los estimados en la etapa anterior. Obtenga los residuales e identifique el modelo ARIMA correspondiente. Sobre esta serie de residuales pueden emplearse las técnicas de identificación de Box-Jenkins para los modelos ARIMA.

4. Suponga que el modelo identificado es de la forma:

$$\eta_t = \frac{\theta(B)}{\delta(B)\phi(B)} a_t \quad (5.9)$$

Para tener una estimación mas refinada de las observaciones faltantes a:

$$Z_t = \sum_{i=1}^k v_i(B) I_t^{t_i} + \sum_{j=1}^m A F_t^{t_j^*} + \sum_{s=1}^h v_s^*(B) O_t^{(t_s)} + \frac{\theta(B)}{\delta(B)\phi(B)} a_t \quad (5.10)$$

El valor de Z_t para el periodo $t = t_j$ menos la estimación de A_j es la estimación óptima de Z_{tj*} , $j = 1, 2, 3, \dots, m$, (es decir, $\hat{Z}_{tj*} = Z_t - \hat{A}_j$, $j = 1, 2, 3, \dots, m$).

5.5. Análisis de intervención

Las series de tiempo son a veces influenciadas por eventos externos, tales como: paros laborales, crisis económicas, políticas repentinas, inesperadas olas de calor o frío, errores no notados de digitación y registro, entre otros. Estos eventos externos son comúnmente llamados intervenciones y crean observaciones falsas que son inconsistentes con el resto de la serie. Tales observaciones son usualmente referidas como valores atípicos, cuando la medida de duración y las causas de interrupción son conocidas, sus efectos pueden ser considerados mediante el uso de un modelo de intervención (Wei 2006). Si modelamos estos efectos en la serie podemos mejorar la precisión de la estimación de los parámetros y de las predicciones (Peña 2005). Es importante llevar a cabo el análisis de intervención debido a que con frecuencia al construir modelos para series de tiempo se pueden generar uno o más residuos mayores a lo esperado, si dichos residuos pueden explicarse por la ocurrencia de intervenciones, se podría reducir su influencia en los resultados del modelo. Además existen otras situaciones en las cuales los efectos de los eventos externos impiden siquiera la identificación de un modelo para representar la serie (Guerrero 2003).

5.5.1. Técnica del análisis de intervención

Box & Tiao (1975), denominaron análisis de intervención a la inclusión en un modelo de series temporales de variables ficticias para representar sucesos que producen efectos deterministas. Las variables ficticias más utilizadas para representar sucesos cualitativos que afectan a la serie son de dos tipos: variables de impulso y variables escalón. Las variables impulso representan sucesos que ocurren únicamente en un instante, por ejemplo, un accidente, un error de medida o una huelga. Las variables escalón representan acontecimientos que comienzan en un instante desconocido y se mantienen a partir de ese instante, como una subida de precios, un cambio legal, un cambio de base en un índice etc (Peña 2005).

En particular, una serie de tiempo estacionaria W_t está constituida por dos partes, una determinística que refleja esencialmente el nivel de la serie μ_w y otra estocástica que da origen a las fluctuaciones alrededor de dicho nivel. Se pueden distinguir tres tipos de efectos:

1. Aquellos que dejan sentir como una elevación o caída momentánea del nivel y que desaparecen sin influir sobre el comportamiento posterior de la serie.
2. Los que ejercen una influencia sostenida (no momentáneamente) sobre el nivel de la serie, pero dejan intacta la estructura básica de su parte estocástica.

3. Los efectos que, independientemente de influir o no sobre la parte determinística, si alteran la estructura de la parte estocástica.

Las intervenciones que causan los efectos tipo (1) y tipo (2), impulso y escalón respectivamente, pueden estudiarse mediante el análisis de intervención. El análisis de intervención puede verse como una extensión del análisis de procesos ARIMA, ya que un modelo para una serie $T(Z_t)$ que contenga los efectos de una intervención, se expresa como:

$$T(Z_t) = \varepsilon_{I,t} + N_t \quad (5.11)$$

donde N_t se asocia con un modelo ARIMA, estacionario e invertible, que representa la parte estocástica de la serie y el cual puede expresarse como:

$$\phi(B)\nabla^d N_t = \theta_0 + \theta(B)a_t \quad (5.12)$$

Con N_t medida como desviaciones respecto a su media si $d=0$, mientras que $\varepsilon_{I,t}$ es una función que permite representar los efectos de la intervención; dicha función debe ser tal que permita apreciar los cambios ocurridos en la serie no sólo en el momento que ocurrió la intervención, sino los que se dieron posteriormente, pero son atribuibles a ella, es decir se requiere que $\varepsilon_{I,t}$ sea en realidad un modelo dinámico de la intervención (Guerrero 2003).

5.5.2. Valores atípicos

Con mucha frecuencia ocurren en las series reales hechos puntuales que desconocemos. Las observaciones afectadas por estas intervenciones pueden presentar una estructura distinta de las demás y aparecer como datos atípicos (Chen & Liu 1993). Es importante ser capaz de identificar estas situaciones desconocidas y separarlas de la dinámica habitual de la serie porque:

1. Si sus efectos son grandes, pueden sesgar la estimación de los parámetros, lo que produciría malas predicciones futuras.
2. Si el suceso ha ocurrido en la última parte de la serie y alguna observación afectada se utiliza para generar predicciones estas no serán buenas, incluso aunque los parámetros estén bien estimados.

3. Si estos sucesos atípicos pueden volver a aparecer en el futuro y los identificamos y estimamos sus efectos, podemos incorporar esta información en las predicciones y obtener intervalos de predicción más realistas.

Un valor se considera como atípico si tienen pocas posibilidades de ocurrir dada la estructura habitual de evolución de la serie. Entonces, la serie puede considerarse como la suma de dos series distintas. La primera es la generada por las circunstancias habituales, entendiendo por ello las que están presentes en el 95 % o 99 % de los datos observados. La segunda es la formada por datos que corresponden a situaciones peculiares específicas (Peña 2005).

Podemos considerar dos casos básicos de atípicos: los aditivos, que son análogos a los atípicos que ocurren en datos estáticos y los innovativos, que son específicos de situaciones dinámicas. Una serie puede sufrir también otros efectos, como cambios de nivel o tendencia (Peña 2005).

5.5.3. Atípicos aditivos (AO)

Se dirá que ha ocurrido un AO en el instante h si el valor de la serie se genera en ese instante de manera distinta al resto. En general si el dato Z_h tiene propiedades distintas del resto por un cambio en las condiciones externas o del sistema de medición, diremos que este dato es atípico. El modelo que seguirá una serie observada Z_t , afectada por un AO en $t=h$ será:

$$Z_t = \begin{cases} y_t & t \neq h \\ y_t + \omega_A & t = h \end{cases} \quad (5.13)$$

donde Y_t es la serie no contaminada por atípicos que suponemos sigue un modelo ARIMA, entonces el modelo que sigue la serie observada, Z_t , es:

$$Z_t = \omega_A I_t^{(h)} + \psi(B)a_t \quad (5.14)$$

donde $I_t^{(h)} = 0, t \neq h; I_t^{(h)} = 1, t = h$. Este modelo es idéntico al de análisis de intervención con un impulso.

Es importante tener en cuenta que los atípicos aditivos generan dos tipos de efectos. El primero son los efectos en los residuos, cuando un AO deja una huella en la serie debido a

la alteración de su valor en un punto; cuando desconozcamos su presencia y construyamos un modelo ARIMA, esta alteración puede detectarse con los residuos del modelo. Cuando hay varios AO, sus efectos se superponen y pueden modificar mucho la estructura de la serie. El segundo son los efectos en la estimación de los parámetros, cuando un AO sesga los coeficientes de autocorrelación y la estimación de los parámetros del proceso hacia cero. En general los efectos dependen mucho del tamaño muestral y para tamaños muestrales muy grandes un AO de tamaño moderado puede tener un efecto pequeño (Peña 2005).

5.5.4. Atípicos innovativos (IO)

Se dirá que ha ocurrido un IO, cuando la innovación en un punto está directamente afectada por una cantidad desconocida debida a un suceso imprevisto. Podemos decir que un IO es un incremento del error de previsión de la serie en un punto debido a un cambio en las condiciones externas (Peña 2005). El modelo para la serie que sufre un IO de magnitud ω_t en el instante h es:

$$Z_t = \psi(B)(\omega_I I_t^{(h)} + a_t) \quad (5.15)$$

Es importante tener en cuenta que los atípicos innovativos generan dos tipos de efectos. El primero, los efectos sobre la serie y el segundo los efectos en la estimación, en ambos casos dependen de la estacionariedad de la serie y de los parámetros autorregresivos y de media móvil de los modelos. Supongamos que la serie es estacionaria:

$$Z_t = y_t + \psi(B)\omega_I I_t^{(h)} \quad (5.16)$$

donde $y_t = \psi(B)a_t$ es la serie sin contaminar que sigue el modelo ARMA. Según esta ecuación, la relación entre las observaciones contaminadas, Z_t , y las originales, y_t , es:

$$Z_t = \begin{cases} y_t & t < h \\ y_t + \omega_I \psi_j & t = h + j \quad j \geq 0 \end{cases} \quad (5.17)$$

5.5.5. Cambio de nivel (LS)

Otro efecto importante que puede aparecer sobre una serie temporal es el cambio de nivel. Peña (2005) menciona, que una serie ha sufrido un cambio de nivel en el instante h si sigue

el modelo:

$$Z_t = \omega_L S_t^{(h)} + \psi(B)a_t \quad (5.18)$$

Donde $S_t^{(h)}$ es la variable escalón, definida por $S_t^{(h)} = 1$ si $t \geq h$ y cero en otro caso. Los valores de la serie observada estarán relacionados con la serie sin contaminar por el cambio de nivel mediante:

$$Z_t = \begin{cases} y_t & t < h \\ y_t + \omega_L & t \geq h \end{cases} \quad (5.19)$$

Si el proceso es estacionario, un cambio de nivel convertirá a la serie en no estacionaria, ya que la esperanza de cada observación será μ antes del cambio de nivel y $\mu + WL$ después. Si la serie es no estacionaria el cambio de nivel puede ser equivalente a un IO. Un efecto adicional que puede considerarse es el cambio transitorio (TC), definido por:

$$Z_t = \frac{\omega_{TC}}{1 - \delta B} I_t^{(h)} + \psi(B)a_t \quad (5.20)$$

Vemos que si $\delta = 1$ el modelo es del cambio de nivel, LS, ya que $\nabla^{-1} I_t^{(h)} = S_t^{(h)}$, mientras que si $\delta = 0$ tenemos un AO. En la práctica, este tipo de atípico se utiliza fijando el valor de δ , habitualmente 0.7, de manera que no pueda confundirse con los AO o LS (Peña 2005).

5.6. Series de Tiempo

La metodología a aplicar en este trabajo se basa en los modelos de series de tiempo desarrollados por Box y Jenkins (1970), la cual se desarrolla en cuatro etapas fundamentales (Peña 2005):

1. Identificación de un posible modelo dentro de la clase ARIMA, es decir, identificar los valores p, d, y q que especifiquen un modelo de series de tiempo ARIMA.
2. Estimación de los parámetros involucrados en el modelo mediante técnicas no lineales.
3. Verificación del modelo, evaluándolo mediante el ajuste que este proporcione a los valores reales y la validación de los supuestos implícitos en la series de tiempo. De

no cumplirse los supuestos se realizan las modificaciones necesarias y se repiten las etapas anteriores hasta que la verificación indique resultados plausibles.

4. Uso del modelo, en donde se tendrá en cuenta los fines para el cual se ajustó la serie de tiempo, los cuales pueden ser: Pronósticos, simulación, control o explicación de un fenómeno de estudio.

5.6.1. Ruido Blanco (RB)

Se denomina ruido blanco a un tipo especial de proceso estocástico que cumple las siguientes condiciones:

- $E[Z_t] = 0, t = 1, 2, \dots$
- $Var(Z_t) = \sigma^2, t = 1, 2, \dots$
- $Cov(Z_t, Z_{t-k}), k = \pm 1, \pm 2, \dots$

Es decir, la esperanza es siempre constante e igual a cero, la varianza constante y las variables del proceso están incorrelacionadas para todos los retardos, lo que equivale a que $a_t \sim Niid(0, \sigma^2)$, para todo t .

5.6.2. Identificación de la serie de tiempo

Consiste en determinar los órdenes de los polinomios autorregresivos y de media móvil, además del número de veces que debe ser diferenciado de ser requerido. Es decir, se identifica una serie estacionaria en función de la serie original, que tenga una representación ARMA (p, q) (Guerrero 2003).

5.6.3. Determinación de la transformación para estabilizar la varianza

Con el fin de estabilizar la variabilidad de la serie se propone aplicar una transformación que disminuya su volatilidad, dicho método sugiere elegir la potencia λ de tal manera que satisfaga la relación:

$$\sigma_t / \mu_t = \text{Constante para } t = 1, 2, \dots, N$$

En donde σ_t y μ_t representan la desviación estándar y la media de Z_t , y N hace referencia a la cantidad de observaciones con los que cuenta la serie de tiempo. Guerrero (1993) sugiere un método para estabilizar la varianza descrito a continuación:

- Divídanse las observaciones en H grupos que contengan $R = (N - n)$ observaciones contiguas cada uno, dejando por fuera un total e n observaciones ($0 \leq n < R$).
- Se calcula la media y la desviación estándar de cada grupo para posteriormente determinar su coeficiente de variación.
- Se construye una tabla con valores equidistantes de la potencia estabilizadora de la serie λ (Lambda).
- Determinar el menor coeficiente de variación asociado a la potencia Lambda.

La transformación de la serie será la siguiente:

$$T(Z_t) = \begin{cases} Z_t^\lambda & \text{si } \lambda \neq 0 \\ \log(Z_t) & \text{si } \lambda = 0 \end{cases}$$

5.6.4. Estabilización en nivel de la serie

Para estabilizar la media se aplicarán diferenciaciones regulares y estacionales de ser necesario. La decisión de aplicar estas diferencias pueden determinarse si se observa en el gráfico de la serie de tiempo un comportamiento creciente y apoyándose en la Función de Autocorrelación Muestral (FAC). Según Guerrero (2003), si la Función de Autocorrelación Muestral decae rápidamente a cero es indicativo de que la serie es estacionaria en cuanto a nivel se refiere y no es necesario aplicarle diferenciaciones. Existen otros métodos para identificar si la serie no es estacionaria en media, como es el caso de la prueba de Dickey Fuller, la cual contrasta la siguiente prueba de hipótesis:

$$\begin{aligned} H_o &: \phi(B)(1 - B)T(Z_t) = \theta(B)a_t \\ H_a &: \phi(B)(1 - \rho B)T(Z_t) = \theta(B)a_t \end{aligned}$$

Si $p = 1$ la serie tiene una raíz unitaria y por tanto no es estacionaria, lo cual indica que es necesario aplicarle al menos una diferenciación para estabilizar la serie.

5.6.5. Identificación de la estructura ARMA

La identificación de los órdenes p y q del modelo se realiza comparando las funciones estimadas de autocorrelación simple y parcial con las funciones teóricas de los procesos ARMA (Guerrero 2003). A continuación se enseña las características principales para identificar el modelo ARMA a utilizar:

Proceso	FAC	FACP
AR(p)	Convergencia a cero con comportamiento indicado por la ecuación: $\phi(B)\rho = 0$, para $k \geq p$.	Solamente las primeras p autocorrelaciones parciales son distintas de cero
MA(q)	Solo las primeras q autocorrelaciones son distintas de cero	Sucesión infinita convergente a cero
ARMA(p,q)	Comportamiento irregular de las primeras q autocorrelaciones y después convergencia a cero de acuerdo con $\phi(B)\rho_k = 0$, para $p > k$	Sucesión infinita convergente a cero

Tabla 5.2: Daniel Peña 2005. Comportamiento de la FAC y la FACP

5.6.6. Modelos Autorregresivos AR(p)

Se define un modelo como autorregresivo si la variable endógena de un período t es explicada por las observaciones de ella misma correspondientes a períodos anteriores añadiéndose como en los modelos estructurales un término de error. En el caso de procesos estacionarios con distribución normal, la teoría estadística de los procesos estocásticos dice que, bajo determinadas condiciones previas, toda Z_t puede expresarse como una combinación lineal de sus valores pasados (parte sistemática) más un término de error (innovación)(Arce & Mahía 2000).

Los modelos autorregresivos se abrevian con la palabra AR seguido del orden del modelo: AR(1), AR(2),....etc. El orden del modelo expresa el número de observaciones retrasadas de la series temporal analizada que intervienen en la ecuación.

La expresión genérica de un modelo autorregresivo de orden p AR(P) es la siguiente:

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \dots + \phi_p Z_{t-p} + a_t \quad (5.21)$$

El término de error de los modelos de este tipo se denomina generalmente ruido blanco cuando cumple las tres hipótesis básicas tradicionales mencionadas al principio del texto:

- media nula
- varianza constante
- covarianza nula entre errores correspondientes a observaciones diferentes

5.6.7. Modelo de medias móviles MA(q)

Un modelo de los denominados de medias móviles es aquel que explica el valor de una determinada variable en un período t en función de un término independiente y una sucesión de errores correspondientes a períodos precedentes, ponderados convenientemente. Estos modelos se denotan normalmente con las siglas MA, seguidos, como en el caso de los modelos autorregresivos, del orden entre paréntesis (Arce & Mahía 2000). Así, un modelo con q términos de error MA(q) respondería a la siguiente expresión:

$$Z_t = a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \dots + \theta_p a_{t-p} \quad (5.22)$$

Modelos Autorregresivos Integrados de Medias Móviles ARIMA(p,d,q)

Los modelos autorregresivos e integrados de medias móviles (ARIMA) pueden ser vistos como una generalización de los modelos ARMA. Yaglom (1985) sugirió la posibilidad de que cierto tipo de no estacionariedad mostrado por algunas series de tiempo, podía representarse mediante la simple toma sucesiva de diferencias de la serie original. En estos modelos se aplica el operador diferencia para eliminar la posible tendencia polinomial de orden d presente en la serie bajo análisis. La notación empleada para los modelos ARIMA es la siguiente:

$$\phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \dots + \phi_p Z_{t-p} + Z_t = \theta_1 a_{t-1} + \theta_2 a_{t-2} + \dots + \theta_p a_{t-p} + a_t \quad (5.23)$$

Modelo Multiplicativo Estacional

Los modelos SARIMA multiplicativos, son una extensión de la aplicabilidad de los modelos ARIMA para cubrir la necesidad del análisis de las series de tiempo estacionales.

La estacionalidad hace que la media de las observaciones no sea constante, sin embargo evoluciona en forma predecible de acuerdo a un patrón cíclico. La estacionalidad se da periódicamente dependiendo de la naturaleza de la serie, por ejemplo, una serie estacional con observaciones mensuales que presenta similitud en los comportamientos del mismo mes en años consecutivos tiene un periodo estacional de 12. La incorporación de la estacionalidad en un modelo ARIMA puede ser de manera multiplicativa o aditiva (Peña 2005). El caso más habitual es incorporarla en forma multiplicativa, con lo que resulta el modelo ARIMA estacional Multiplicativo. Estos modelos se basan en la idea de convertir las series no estacionarias en estacionarias mediante diferencias regulares y de eliminar la estacionalidad de la serie mediante una diferencia estacional con un periodo estacional S . Al ser multiplicativos se componen de dos partes: la parte regular (ARIMA convencional) y la parte estacional:

$$\phi(B)\Phi(B^E)\nabla^d\nabla_E^D(Z_t - \mu) = \theta(B)\Theta(B^E)a_t \quad (5.24)$$

El cual es denotado como $ARIMA(p, d, q)(P, D, Q)_E$, donde P es el orden del polinomio de la parte autorregresiva estacional, D es la diferencia estacional, Q es el orden del polinomio de la parte de promedio móvil estacional, y S es el periodo estacional.

5.6.8. Estimación de los parámetros del modelo

Para realizar la etapa de estimación se presupone que se ha identificado a un modelo adecuado, valorando la presencia de estacionariedad, estacionalidad y aplicando los métodos pertinentes para su corrección. Por tanto se procede a encontrar los mejores valores de los parámetros que permita al modelo representar adecuadamente la serie trabajada. Es decir, una vez conocidos los órdenes de los polinomios autorregresivos y de promedios móviles p y q , y el grado de diferenciación d , se postula el siguiente modelo:

$$\phi(B)\nabla^dT(Z_t) = \theta_o + \theta(B)a_t$$

El cual resulta apropiado para representar la serie estudiada, por tanto se requiere asignarle valores pertinentes a $\theta_1, \theta_2, \dots, \theta_N$ y $\phi_1, \phi_2, \dots, \phi_N$, mediante el método de máxima verosimilitud.

5.6.9. Verificación de la serie de tiempo

La metodología Box Jenkins se basa en la idea de que todo modelo es erróneo, debido a que los modelos son representaciones simplificadas de la realidad. Por tal razón es común plantear varios modelos y se elegirá el modelo que presente menos fallas, en donde cada modelo se debe comparar mediante los supuestos de la series de tiempo y elegir el que menos supuestos viole o el que más se adecúe a las necesidades del autor (Peña 2005). Una de las maneras más fáciles de detectar las violaciones a los supuestos es mediante el análisis de los residuos del modelo, considerando un residuo como aquella parte de las observaciones que no es explicada por el modelo, es decir, los residuales miden la discrepancia entre los valores observados y los valores estimados por el modelo. A continuación se plantean los supuestos de los procesos at los cuales pueden verificar y corregirse para tener como resultado final el modelo más pertinente para la serie de datos evaluada:

Supuesto 1: Los a_t tienen media cero.

Se calcula la media aritmética y la desviación estándar de los residuos de la siguiente manera:

$$m(\hat{a}) = \sum_{t=t'}^N \frac{\hat{a}_t}{(N - d - p)} \quad (5.25)$$

$$\sigma_a = \sqrt{\sum_{t=t'}^N \frac{(\hat{a}_t - m(\hat{a}))^2}{(N - d - p - q)}} \quad (5.26)$$

En donde $t' = d + p + 1$, para construir el cociente $\sqrt{(N - d - p)}m(\hat{a})/\sigma_a$. Si el valor absoluto de dicho cociente es menor que dos, se dirá que no hay evidencia de que la media del proceso de ruido blanco es distinta de cero, y por ende no se rechaza el supuesto, por el contrario, si $|\sqrt{(N - d - p)}m(\hat{a})/\sigma_a| \geq 2$, la media de los residuales es significativamente distinta de cero, lo cual indica que se ha violado el supuesto (Guerrero 2003).

Supuesto 2: Los a_t tiene varianza constante.

Se puede realizar mediante una verificación visual al elaborar una gráfica de los residuos versus el tiempo para observar si la varianza parece ser o no constante. No obstante, existen métodos formales para evaluar este supuesto, tal es el caso del contraste de Breusch-Pagan el cual comprueba si es posible encontrar un conjunto de variables Z que sirvan para explicar la evolución de la varianza de las perturbaciones aleatorias estimadas a partir del cuadro de los errores del modelo inicial sobre el que se pretende comprobar si existe o no

heterocedasticidad. Se calcula el estadístico de prueba bajo la hipótesis nula de homoscedasticidad, se demuestra que el producto del tamaño de la muestra (n) multiplicado por el R^2 sigue asintóticamente una distribución Chi-Cuadrado con un número de grados de libertad igual al número de rezagos de los residuos incluidos en la regresión $nR^2 = \chi_p^2$ (Wei 2006).

Supuesto 3: Los a_t siguen una distribución normal para todo t .

Se puede realizar una gráfica de residuos versus tiempo ó un histograma para visualizar la forma de la distribución. Se espera que aproximadamente un 95 % de las observaciones este dentro de un intervalo que este dos desviaciones estándar por encima y debajo de la media $(-2\widehat{\sigma}_a, 2\widehat{\sigma}_a)$.

No obstante, existen pruebas formales para probar este supuesto. Tal es el caso de la Test de Jarque Bera, la cual es considerada una prueba de normalidad para grandes muestras, es decir, un proceso estadístico utilizado para determinar si una muestra o cualquier grupos de datos se ajustan a una distribución normal. La prueba de JB toma el principio de “qué tanto se desvían los coeficientes de asimetría y curtosis de los residuos, con los coeficientes de asimetría y curtosis de las distribución normal” (Wei 2006). El estadístico de JB esta dado por $n/6(s^2 + 1/4k^2)$, donde s es atribuido al coeficiente de Asimetría y k al coeficiente de Curtosis, el estadístico se comprara con una distribución Chi-Cuadrado con 2 grados de libertad.

Supuesto 4: Las variables aleatorias a_t son mutuamente independientes.

Debido a que la independencia no implica autocorrelación, se requiere que $\rho_k(a) = 0$ para todo $k \neq 0$. Se puede verificar calculando primero la FAC muestral de los residuos $r_k(\widehat{a})$. Si $|r_k(\widehat{a})| \geq 2/\sqrt{N-d-p}$ se dirá que la autocorrelación k -ésima es significativamente distinta de cero. en donde:

$$r_k(\widehat{a}) = \frac{\sum_{t=t'}^{N-k} \widehat{a}_t \widehat{a}_{t+k}}{\sum_{t=t'}^N \widehat{a}_t^2} \quad (5.27)$$

No obstante, esta prueba no es válida para correlaciones pequeñas $k \leq 3$, por tanto es conveniente realizar una prueba conjunta de las primeras K autocorrelaciones, como es el caso del estadístico Q de Box y Pierce (1970):

$$Q = (N-d-p)(N-d-p+2) \sum_{k=1}^K \frac{r_k^2(\widehat{a})}{N-d-p-k} \quad (5.28)$$

En donde si K mayor que 20 sigue aproximadamente una distribución chi-cuadrado con $K - p - q$ grados de libertad, de aquí que el valor Q calculado debe ser comparado con los valores de la chi-cuadrado (Guerrero 2003).

La prueba de Wang, Van Gelder, J & Ma (2005) es una variación de la prueba de Ljung-Box para probar homocedasticidad condicional (Efectos ARCH), al elevar las autocorrelaciones al cuadrado y evaluar el siguiente estadístico de prueba:

$$Q = N(N + 2) \sum_{k=1}^K \frac{r_k^2(a)^2}{N - k} \quad (5.29)$$

Donde, N corresponde al tamaño de la muestra, r_k^2 son las autocorrelaciones al cuadrado de los k rezagos de los residuales al cuadrado. Se evalúa la hipótesis nula de que las k autocorrelaciones son igual a cero, comparando el estadístico de prueba con una Chi-Cuadrado con (K) grados de libertad.

Supuesto 5: No existe observaciones atípicas.

Por medio de la gráfica de residuos versus el tiempo es posible visualizar si existen observaciones atípicas, ya que se pueden considerar observaciones anómalas aquellas que su probabilidad de ocurrencia sea muy pequeña. Una regla adecuada para identificar observaciones atípicas son aquellas que los residuales sean superiores ó inferiores a $(-3\hat{\sigma}_a, 3\hat{\sigma}_a)$ según sea el caso.

Supuesto 6: El modelo considerado es parsimonioso.

Induce a optar por un modelo sencillo en vez de uno complejo, es decir, que tenga la cantidad justa de parámetros, sin necesidad de llegar a la sobreparametrización ó subparametrización. Para verificar este supuesto se construye intervalos de aproximadamente un 95 % de confianza para cada uno de los parámetros como lo muestra la ecuación 5 y se observa si el valor cero se encuentra en el intervalo.

$$\left(\hat{\theta} - 2\sqrt{\widehat{Var}(\hat{\theta})}, \hat{\theta} + 2\sqrt{\widehat{Var}(\hat{\theta})} \right) \quad (5.30)$$

Aunque puede existir casos en que el conocimiento del fenómeno indique que el parámetro debe aparecer, aun cuando una prueba de hipótesis indique lo contrario.

Supuesto 7: El modelo es admisible.

Este supuesto puede verificarse de acuerdo a las condiciones de estacionariedad y/o invertibilidad correspondientes a los polinomios de retraso que intervengan, es decir, que los parámetros se encuentren dentro de las regiones admisibles.

Supuesto 8: El modelo es estable en lo parámetros.

La principal causa de inestabilidad es la redundancia de parámetros, de tal forma que un cambio en un parámetro puede compensarse mediante un cambio en otro, sin que la suma de cuadrados se altere. Esto implica que se debe estar alerta a la presencia de alta correlación (Negativa y positiva) entre los parámetros estimados, ya que son posible causa de inestabilidad.

$$r_{(\hat{\phi}, \hat{\theta})} = \frac{\widehat{Cov}(\hat{\phi}, \hat{\theta})}{\sqrt{\widehat{Var}(\hat{\phi})\widehat{Var}(\hat{\theta})}} \quad (5.31)$$

Capítulo 6

Metodología

En este capítulo se describe la metodología empleada para llevar a cabo este proyecto, enfatizando en el paso a paso de cada procedimiento para la construcción del modelo de series de tiempo, partiendo desde la revisión de la bibliografía, el tratamiento de los datos, los métodos de imputación de valores faltantes empleados, el análisis de intervención, la metodología Box y Jenkins y los supuestos correspondientes para la correcta especificación del modelo.

6.1. Revisión bibliográfica

Durante el desarrollo del proyecto se realiza la búsqueda en medios físicos y virtuales sobre documentación que enriquezca el aprendizaje en el análisis del contaminante y de igual manera en las metodologías estadísticas a emplear. Las principales fuentes de información revisada para la comprensión del comportamiento y efectos en la salud del material particulado menor a 2.5 micras son: la Agencia Ambiental de Protección del Medio Ambiente de Estados Unidos de Norte América (EPA-USA), la Organización mundial para la salud (OMS), el Ministerio del Medio Ambiente y Desarrollo Sostenible de Colombia, La autoridad ambiental municipal (DAGMA), entre otras. Por otro lado, las fuentes de información de las metodologías estadísticas empleadas, están enfocadas a los artículos científicos y tesis de grado en las cuales se involucren la metodología Box y Jenkins para la modelación de contaminantes atmosféricos, donde se destacan autores como Cadena & García (2014), García & Gonzalez (2014), Castaño, E (2007), Chen & Liu (1993), Iglesias et al. (2006) entre otros.

6.2. Obtención y depuración de la base de datos

Los datos son suministrados por el Sistema de Vigilancia de Calidad del Aire perteneciente al Departamento Administrativo de la Gestión del Medio Ambiente (DAGMA), tomando como referencia la estación ubicada al interior del campus de la Universidad del Valle (Sur de la Ciudad). El reporte del contaminante $PM_{2.5}$ es entregado como promedio horario, el cual es la desagregación mínima reportada al público por la autoridad ambiental.

Se toma como periodo de referencia un año calendario entre el 1 de Octubre del 2013 al 30 de Septiembre del año 2014, para un total de 8760 datos horarios de los cuales 545 corresponden a valores faltantes, es decir el 6.22% de los datos registrados. La cantidad de datos faltantes contenidos en la base de datos y su distribución es el tiempo dificulta la aplicación de los métodos de imputación comúnmente usados en la reconstrucción de series de tiempo, razón por la cual se decide utilizar los promedios diarios según la Resolución 610 del 2010 expedida por el Ministerio de Ambiente y Desarrollo Sostenible, donde se mencionan las formas de representación del contaminante, el cálculo empleado y los límites permisibles como se muestra en la Tabla 6.1.

Tabla 6.1: Normatividad $PM_{2.5}$ Resolución 610 del 2010

Contaminante	Nivel Máximo Permissible	Tiempo de Exposición	Tipo de Calculo
Material	25	Anual	Promedio Aritmético
Particulado 2.5	50	24 Horas	Promedio Aritmético
Material	50	Anual	Promedio Aritmético
Particulado 10	100	24 Horas	Promedio Aritmético

Se aclara que las mediciones de los instrumentos de monitoreo arrojan los valores de $PM_{2.5}$ en microgramos por metro cúbico ($\mu g/m^3$), por lo tanto no hay que realizar proceso de conversión de unidades, como es el caso de los gases Dióxido de Azufre, Dióxido de Nitrógeno, Óxidos de Carbono y Ozono que están dados en partículas por millón (ppm) y si es necesario hacer el cambio de unidades de medida.

6.3. Análisis exploratorio de los datos

El análisis exploratorio se realiza con el fin de observar el comportamiento del contaminante $PM_{2.5}$ a través del tiempo, identificando así excedencias a la normatividad y los efectos generados en la salud. De igual manera, se caracterizó el Material Particulado menor a 2.5 micras por días de la semana, días ordinarios, días festivos y por algunas condiciones climatológicas.

Al realizar los descriptivos básicos de tendencia central en conjunto con métodos de visualización como gráficos lineales y diagramas de cajas y alambres, se logró identificar observaciones atípicas y cambios estructurales transitorios y permanentes en la serie temporal, que serán tenidos en cuenta para el proceso de modelación. Todo el análisis de la información se desarrolla en el Software Estadístico Libre R, con sus diversos paquetes que permite el uso de metodologías estadísticas modernas.

6.4. Reconstrucción de la base de datos

Con el fin de solucionar los problemas de datos faltantes, se hace uso de metodologías estadísticas de imputación de datos, en donde, en el caso de este proyecto en particular se divide en tres etapas la reconstrucción de la serie, en la primer etapa se utiliza un método recursivo de imputación por medio de la regresión lineal simple utilizando Mínimos Cuadrados Ponderados debido a los problemas de heterocedasticidad, una segunda etapa empleando el procedimiento desarrollado por Chen & Liu (1993) y retomado por Castaño, E (2007) y una etapa final utilizando una Sub-Serie de tiempo, con la cual se pronostican los valores ausentes cercanos al cambio de nivel.

6.4.1. Etapa 1: Regresión lineal por Mínimos Cuadrados Ponderados

Como primer etapa de imputación de datos, se estiman los valores de $PM_{2.5}$ por medio de la regresión lineal simple tomando como variable independiente un contaminante altamente correlacionado como es el caso del Materia Particulado menor a 10 micras. Los parámetros del modelo se plantean inicialmente por medio de los Mínimos Cuadrados Ordinarios, pero debido a los problemas de heterocedasticidad observados se decide emplear metodologías de estimación más robusta como es el caso de los Mínimos Cuadrados Ponderados. Se aclara que ambos contaminantes deben estar en la misma escala de tiempo, por este motivo se calcula el promedio diario de cada variable implícita en la regresión.

6.4.2. Etapa 2: Estimación de datos faltantes por el método de Castaño

La segunda etapa de imputación de datos se realiza por medio de la metodología propuesta por Castaño, E (2007), en donde a cada valor faltante se le asigna un valor Atípico Aditivo (AO) de $100 \mu g/m^3$. Posteriormente se realiza la identificación de las intervenciones ajustando un modelo autoregresivo de orden alto ($n \leq T^{1/3} = 7$) con el cual se hará el barrido

de cada una de las observaciones que conforman la serie de tiempo, detectando las observaciones atípicas incluidas en los registros faltantes y los demás valores que el procedimiento considere como observaciones poco probables y que son naturales del proceso.

Una vez estimado el modelo, se debe verificar que no existan nuevas observaciones atípicas y que los residuales se comporten como ruido blanco, en caso de no ser así se debe iniciar de nuevo el proceso de identificación de valores atípicos. Se estima el modelo restringido, resultante de filtrar la serie Z_t sin incluir el modelo de ruido, donde se obtienen sus residuales y se identifica el modelo ARIMA correspondiente por medio de la metodología Box-Jenkins.

6.4.3. Etapa 3: Estimación de datos faltantes por Sub-Series de tiempo

Debido a la presencia de datos faltantes cercanos al cambio de nivel, que no pudieron ser imputados en la etapa 1 por la regresión lineal, al existir valores faltantes para el PM10 en ese mismo instante de tiempo, ni por la etapa 2 al tener problemas de convergencia por el cambio estructural con el método de Castaño, E (2007), se decide realizar una sub-serie de tiempo, estimando con el tramo 2 de la serie (Completo por las etapas previas) los últimos valores faltantes.

6.5. Construcción y validación del modelo

La identificación de Valores Atípicos (AO), Valores Inovacionales (IO), Cambio de Nivel Temporal (TC) y Cambio de Nivel Permanente (LS), permitió tener una serie de datos libre de eventos exógenos que generen perturbaciones y afecten la identificación de un modelo correcto. La identificación del modelo de ruido se realiza empleando la metodología ARIMA desarrollada por Box y Jenkins, en donde por medio de la Función de Correlación (FAC), Función de Correlación Parcial (FACP) y la Función de Autocorrelación Extendida (FACE) se identifica un conjunto de modelos posibles, eligiendo el mejor modelo por medio de diferentes criterios como el Criterio de Información de Akaike (AIC), el Criterio de Información Bayesiano (BIC), el Criterio de Log-Verosimilitud y el Criterio de Varianza Mínima.

Para la validación de los supuestos del modelo se emplean las pruebas estadísticas comúnmente usadas en las series de tiempo, tal es el caso de la prueba de Jarque Bera para normalidad, la prueba de Ljung Box para el análisis de independencia entre las observaciones, la

gráfica de residuales contra el tiempo para observar si la varianza presenta algún comportamiento, además de evaluar la presencia de observaciones atípicas y la significancia de los parámetros empleados en cada modelo.

6.6. Pronósticos de la serie tiempo

Una vez identificado el modelo de la serie filtrada, se realizan los pronósticos de las concentraciones diarias de $PM_{2.5}$ mediante el paquete Forecast de R desarrollado por Hyndman (2006), teniendo en cuenta los efectos de las variables exógenas determinadas en el análisis de intervención. Vale la pena resaltar, que para realizar los pronósticos es necesario primero aplicar una transformación inversa a la serie, ya que las predicciones transformadas carecen de interpretabilidad, para lo cual es necesario considerar el factor de corrección propuesto por Guerrero (2003).

Capítulo 7

Resultados

7.1. Análisis Exploratorio

Los datos fueron suministrados por el Sistema de Vigilancia de Calidad del Aire de Santiago de Cali, tomando como referencia un año calendario entre Octubre del año 2013 a Septiembre del año 2014, contando con un total de 8760 registros de concentraciones horarias de PM2.5 en la estación de monitoreo ubicada en la Universidad del Valle.

En el periodo de estudio se encontraron un total de 545 datos faltantes equivalentes a un 6.22% de la cantidad de datos totales, estos vacíos de información son atribuibles a diversos factores como problemas en los instrumentos de medición, fallas en la continuidad de la energía eléctrica en el sector y el mantenimiento o cambio de equipos.

Tabla 7.1: Estadísticas descriptivas para el PM2.5 Horario

Mínimo	Cuartil 1	Mediana	Mean	Cuartil 2	Máximo	Desv.Estandar	NA
0	8	14	15.15	20	617	11.65	545

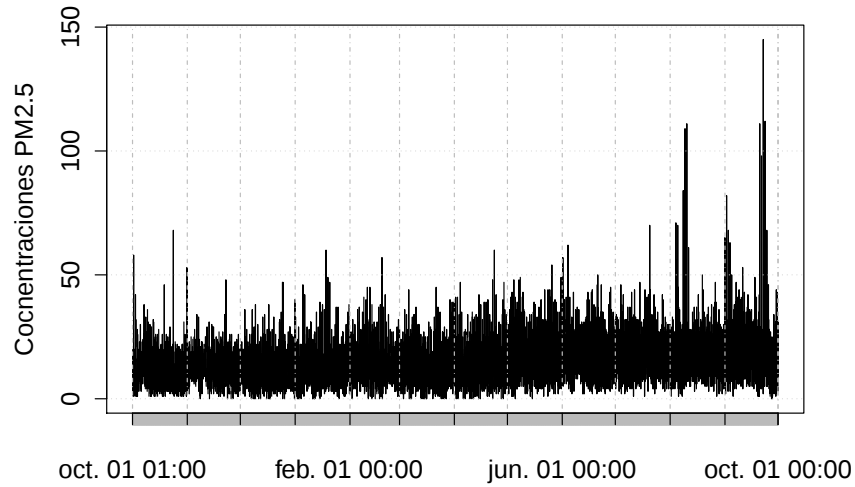


Figura 7.1: Serie de tiempo horaria para el $PM_{2.5}$

En la Figura 7.1 se observa las concentraciones horarias del $PM_{2.5}$ en el periodo de estudio, en el cual fueron excluidos dos valores atípicos momentáneamente para efectos de visualización del gráfico, correspondientes a dos quemas realizadas en la zona sur de la ciudad que ocasionaron un aumento drástico en los niveles del material particulado. Los niveles horarios del $PM_{2.5}$ se mantuvieron por debajo de $50\mu g/m^3$ el 99% de las veces, promediando concentraciones de $15\mu g/m^3$ con una desviación estándar elevada debido a la presencia de observaciones atípicas en la serie de tiempo. El valor mínimo en la serie de datos es de cero lo cual no indica ausencia del contaminante, sino mas bien concentraciones tan pequeñas de material particulado que no son detectables por el instrumento de medición.

El material particulado presenta un comportamiento homogéneo en el transcurso del día como lo muestra la Figura 7.2, incrementando brevemente su concentración después de las 7 de la mañana y disminuyendo a partir de las 5 de la tarde. Las horas de mayor dispersión vienen dadas por la jornada activa laboral y estudiantil, entre las 8 am y las 5 pm, posiblemente por el aumento del flujo vehicular por ser una zona académica y de sector empresarial industrial.

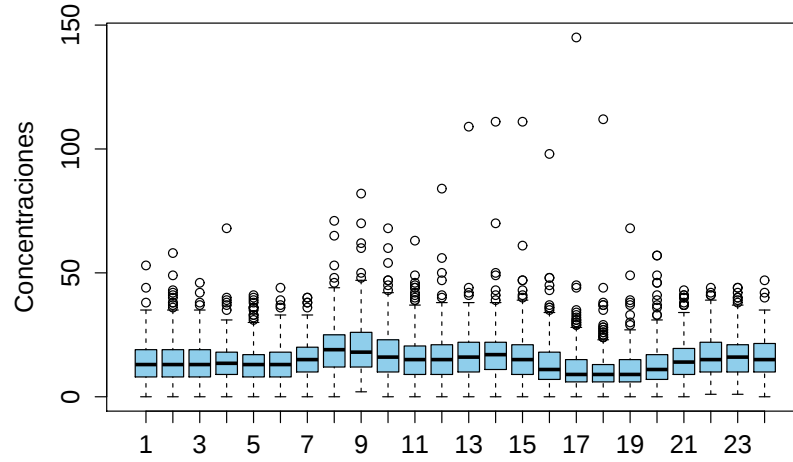


Figura 7.2: Diagrama de cajas horarias para el $PM_{2.5}$

La cantidad de datos faltantes y la distribución temporal de estos a lo largo de la serie de estudios dificulta la aplicación de métodos de imputación de datos normalmente utilizados en la reconstrucción de series de tiempo, como es el caso de la metodología empleada por Castaño, E (2007) en su artículo de reconstrucción de la serie de tiempo de la demanda eléctrica en Colombia, motivo por el cual se hace necesario agregar los datos en una escala menos detallada de información como es el caso de los promedios diarios. Se toma como referencia la Resolución 610 del 2010, la cual indica que para el cálculo de un promedio diario es necesario contar con al menos un 75 % de mediciones válidas en el día.

Tabla 7.2: Estadísticas descriptivas para el $PM_{2.5}$ Diario

Mínimo	Cuartil 1	Mediana	Mean	Cuartil 2	Máximo	Desv.Estandar	NA
3.20	11.75	14.75	15.22	18.10	71.21	5.68	27

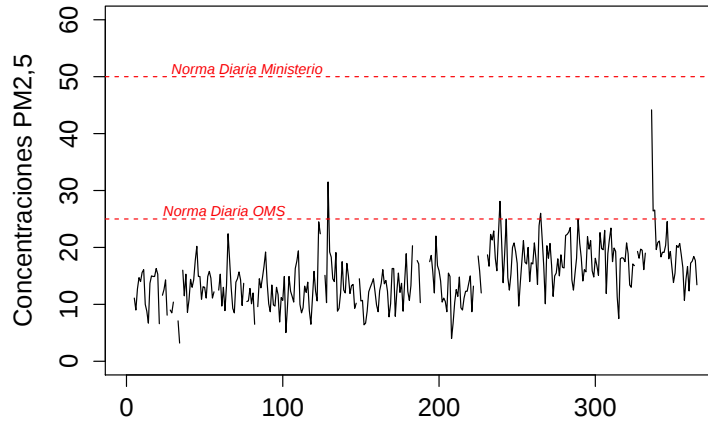


Figura 7.3: Concentraciones promedios diarios de $PM_{2.5}$

Con los datos agregados en promedios aritméticos diarios se disminuye la cantidad de valores ausentes, pasando de 545 a 27 datos faltantes, equivalente al 7.39 % de la cantidad total de datos diarios en el año de estudio (365 datos año), facilitando la reconstrucción de la serie de tiempo por medio de las metodologías a emplear ya mencionadas con anterioridad. Otro aspecto importante a resaltar al utilizar promedios diarios es la ganancia en la interpretabilidad de la información, al ser comparados con los estándares permisibles de contaminación establecidos por la Agencia de Protección Ambiental de Estados Unidos (EPA en inglés), adoptados y ajustados por el Ministerio de Medio Ambiente Vivienda y Desarrollo Territorial, para la condiciones climáticas y ambientales de Colombia en la Resolución 610 del 2010 "Por la cual se modifica la Resolución 601 del 4 de abril de 2006", en donde se estipulan los niveles máximos permisibles de cada contaminante y que en el caso del Material Particulado menor a 2.5 Micras corresponden a $50\mu g/m^3$ para un periodo de exposición diario.

Como se observa en la Figura 7.3, el $PM_{2.5}$ presenta un incremento en su valor promedio de $5.32\mu g/m^3$ a partir del 13 mayo del 2013, pasando de $12.82\mu g/m^3$ a $18.14\mu g/m^3$ y oscilando la serie alrededor de este valor por el tiempo restante de estudio, lo cual puede generar implicaciones en la salud de los habitantes, debido a que estudios demuestran que incrementos del material particulado fino aumentan el número de hospitalizaciones por problemas respiratorios en especial en la población sensible (niños y ancianos) que permanece sometida a estos niveles de contaminación por un tiempo prolongado (Romero, Román, Limón & Praena 2015). La única excedencia a la normatividad Colombiana ($50\mu g/m^3$) se

dio el día 29 de Agosto del 2014 alcanzado un valor promedio de $71.21\mu g/m^3$, atribuible a una quema de pasto realizada en la zona sur de la ciudad (ElPaís 2014).

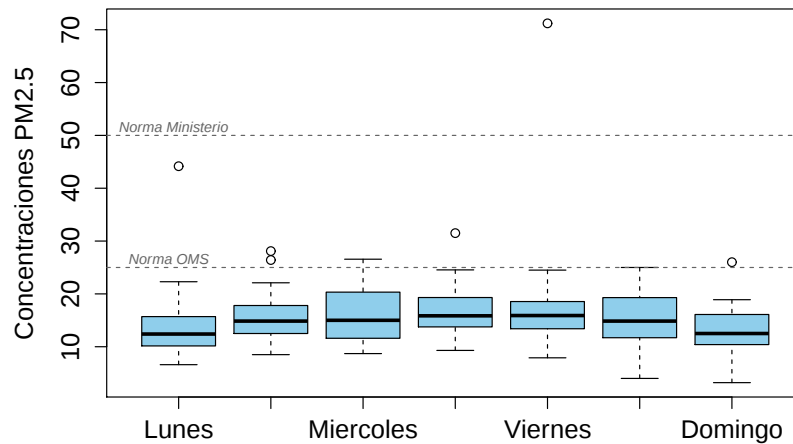


Figura 7.4: Concentraciones diarias de $PM_{2.5}$ por día de la semana

En la Figura 7.4 se puede apreciar el comportamiento de los niveles promedio de $PM_{2.5}$ en el transcurso de la semana, en donde los días domingo y lunes se presenta una leve disminución en las concentraciones del contaminante. Un aspecto importante a resaltar es que el valor promedio diario del $PM_{2.5}$ en la mayoría de registros no sobrepasa la norma nacional ni la estipulada por la OMS de ($10\mu g/m^3$), estos valores se deben interpretar, no como un umbral de material particulado por debajo del cual ningún daño a la salud se observa, sino mas bien como el valor recomendado que representa un objetivo aceptable y alcanzable para minimizar los efectos de salud en el contexto de las limitaciones locales, las capacidades y prioridades de salud pública (Murcia 2010).

7.1.1. Análisis de la serie por tramos

La serie del $PM_{2.5}$ presenta un cambio estructural a partir del mes de mayo del 2014, debido a un posible ajuste metodológico en el cálculo del contaminante, al observar que en la base de datos de las concentraciones horarias del $PM_{2.5}$ hasta el 29 de Mayo del 2013 se evidencia la presencia de concentraciones que toman el valor de cero y posterior a esta fecha dejan de aparecer dichos valores, lo que lleva a concluir que la exclusión de los ceros en el cálculo del material particulado generó un aumento del contaminante, descartando

un incremento por alguna fuente externa de contaminación como se había planteado en un inicio. Por este motivo, se plantea un análisis del contaminante antes y después del cambio de nivel, como se muestra en la Figura 7.5, logrando así observar la tendencia, variabilidad y el porcentaje de datos faltantes en ambos tramos de la serie.

Tabla 7.3: Estadísticas descriptivas por Tramo

	Mínimo	Cuartil 1	Mediana	Mean	Cuartil 2	Máximo	Desv.Estandar	NA
Tramo 1	3.20	10.30	12.60	12.83	14.90	31.50	3.72	23
Tramo 2	7.50	15.90	18.10	18.72	20.70	71.21	6.21	4

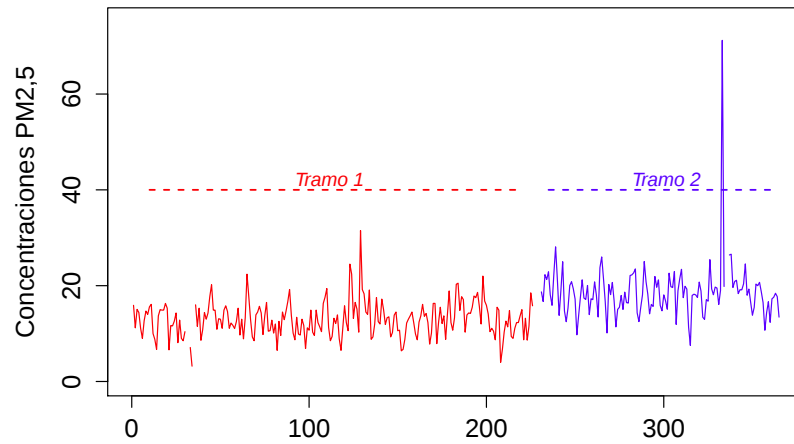


Figura 7.5: Concentraciones $PM_{2.5}$ por Tramos

En la Tabla 7.3 se observan las estadísticas descriptivas para cada uno de los tramos, resaltando que el promedio del Tramo 2 es mayor que el del tramo 1 al igual que su variabilidad. De igual manera el tramo 2 presenta una leve asimetría positiva, posiblemente por la presencia de datos atípicos. Por otro lado la cantidad de datos faltantes parecen estar concentrados en el primer tramo de la serie del $PM_{2.5}$, lo que indica que los esfuerzos de la autoridad ambiental por mejorar la continuidad de las mediciones del $PM_{2.5}$ en la ciudad está dando resultados, disminuyendo la cantidad de valores ausentes en el transcurso del tiempo.

7.1.2. Transformación de la Serie del $PM_{2.5}$

El material particulado es un contaminante de alta volatilidad en el tiempo, donde las características geográficas de Santiago de Cali y sus fuertes vientos son posiblemente la principal razón de la variabilidad de las concentraciones entre los días. Con el fin de reducir la heterocedasticidad y solucionar a su vez problemas de normalidad en la serie (Ver supuestos del modelo final) es necesario encontrar una transformación que estabilice la serie.

Una de las alternativas planteadas para solucionar este problema es el uso de la transformación paramétrica de potencia propuesta por Box & Cox (1964), basada bajo el supuesto de normalidad. Sin embargo, Castaño (2011) teniendo en cuenta los estudios desarrollados por otros autores como Carroll & Raymond (1980), Bickel & Doksum (1981), Powell (1991), Chamberlain (1994), Buchinsky (1995), entre otros, señalan que dicha transformación no es robusta cuando existen observaciones atípicas y cambios estructurales en los datos. Razón por la cual, se decide realizar la transformación Box-Cox excluyendo los puntos atípicos más representativos de la serie y dividiendo el análisis de la prueba por cada tramo de estudio, con el fin de observar si el comportamiento de la variabilidad en cada tramo arrojaba valores de λ equivalentes o por si lo contrario cada subserie de estudio presenta su propia volatilidad.

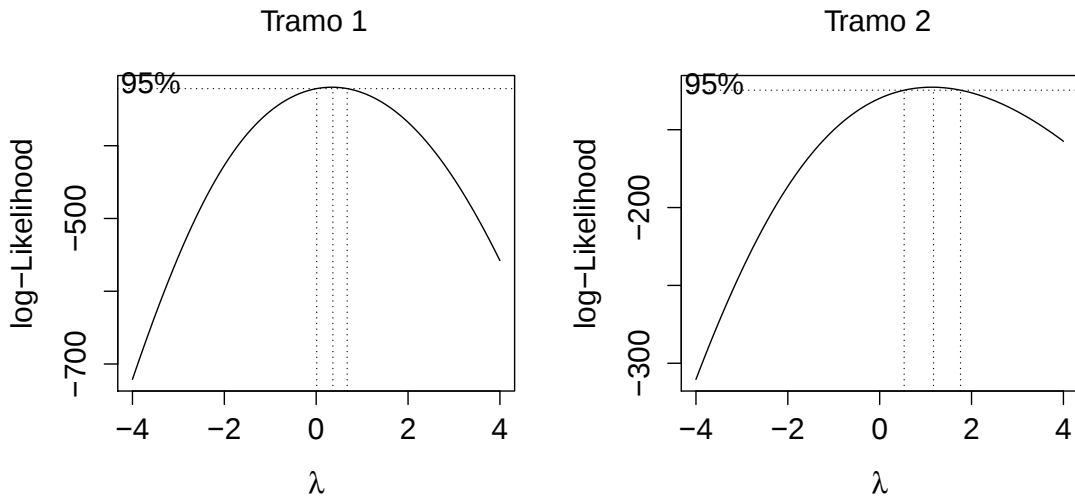


Figura 7.6: Transformación Box-Cox para cada tramo

En la Figura 7.6 se evidencia un valor de λ para cada tramo de la serie, indicando que la volatilidad del contaminante antes y después del cambio de nivel es diferente, dificultando

estabilizar la serie mediante el criterio propuesto por Box-Cox. Siendo poco conveniente emplear el criterio anteriormente descrito, pero manteniendo la necesidad de corregir la volatilidad y amortiguar los problemas de normalidad, se propone utilizar una transformación que mejore la asimetría de la serie (Levemente asimétrica negativa) aplicando las transformaciones más frecuentes como se muestra en la Figura 7.7.

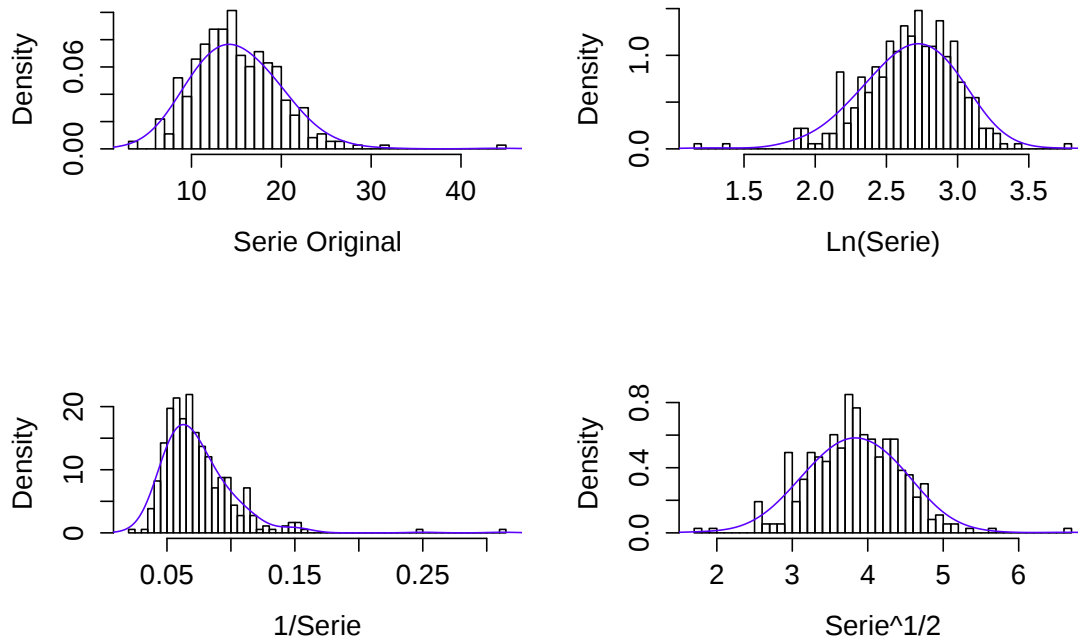


Figura 7.7: Histogramas $PM_{2.5}$ con transformaciones

Como muestra la Figura 7.7, la transformación más adecuada para corregir la leve asimetría de la serie es mediante la raíz cuadrada de sus observaciones. Al aplicar la transformación se obtiene la serie presentada en la Figura 7.9, vale la pena resaltar que tanto la imputación de los valores faltantes por el método de Castaño como la construcción del modelo final se realizará basados en la serie transformada.

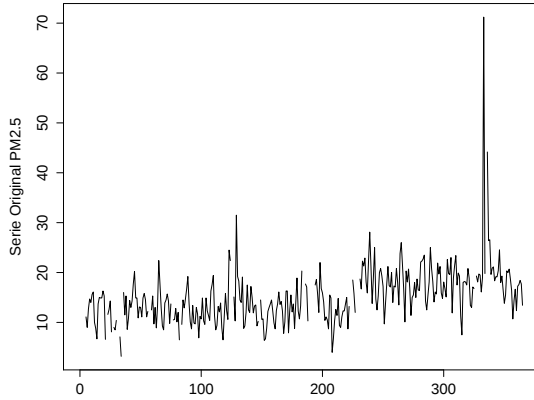


Figura 7.8: Serie Original

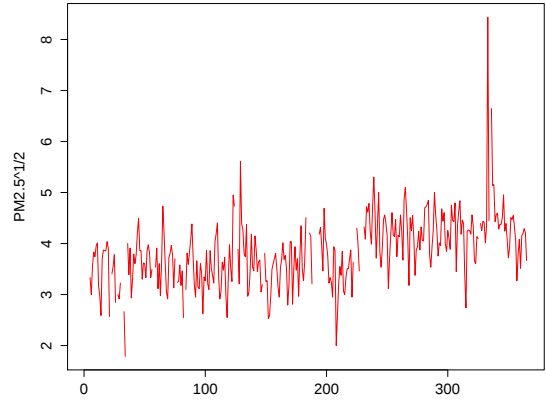


Figura 7.9: Serie Trans Raíz Cuadrada

7.2. Reconstrucción de la serie del $PM_{2.5}$

En la literatura existen variedad de métodos matemáticos y estadísticos para la imputación de valores faltantes que permita la reconstrucción de información, aunque la elección del mejor método dependerá de algunas características importantes como: el tipo de información, fines de uso, cantidad de registros ausentes, entre otros (Castaño 2011). En el caso del proyecto de estudio se cuenta con un total de 27 valores faltantes distribuidos a lo largo de la serie y en algunos casos ubicados de manera consecutiva, lo que hace necesario la utilización de métodos recursivos que permitan completar la serie del material fino.

Se plantearon tres etapas para la reconstrucción de la serie de tiempo, la primera corresponde a la implementación de un modelo de Mínimos Cuadrados Ponderados, tomando como referencia una variable altamente correlacionada con el $PM_{2.5}$ como es el caso del Material Particulado menor a 10 micras (PM_{10}), la segunda etapa consiste en emplear el método propuesto por Chen & Liu (1993) y retomado por Castaño, E (2007), en el cual se considera que una observación faltante puede ser estimada de manera óptima al asignarle un valor tal que pueda ser identificada como una observación atípica aditiva, por último se emplean subseries de tiempo mediante la metodología ARIMA en pequeños tramos de información completa para estimar los registros faltantes que no pudieron ser imputados por los dos métodos anteriormente planteados.

7.2.1. Regresión por Mínimos Cuadrados Ponderados (MCP)

La existencia de correlación entre el Material Particulado menor a 10 y 2.5 micras, ha permitido a varios autores generar estimaciones de razón de cambio para predecir un tipo de partículas suspendidas cuando hay ausencia de otra, tales son los caso de estudios desarrollados por Rojas & Galvis (2005), Echeverri & Maya (2008)), Rojano, Angulo & Restrepo (2013), entre otros. Es importante resaltar que las concentraciones del material particulado se comportan de manera diferente a partir de las condiciones meteorológicas que estén sometidas, la estructura geológica de la ciudad y las fuentes generadoras cercanas.

Con el fin de imputar los datos faltantes de $PM_{2.5}$ se utiliza la regresión por Mínimos Cuadrados Ponderados, partiendo de una variable altamente correlacionada como el PM_{10} , ver figura 7.10. Se hace claridad que las concentraciones del PM_{10} no se miden en la misma estación donde se registra las del material particulado fino, debido a que el protocolo de monitoreo empleado por el DAGMA basado en la Resolución 610 del 2010, plantea realizar mediciones de $PM_{2.5}$ solo en las estaciones que hayan evidenciado excedencias del PM_{10} , no obstante, por el diámetro del material particulado menor a 2.5 micras y su facilidad para suspenderse en el aire y ser arrastradas por las corrientes de viento se espera que una estación cercana capture la mayor cantidad de partículas que fueron registradas en la estación ubicada en UNIVALLE.

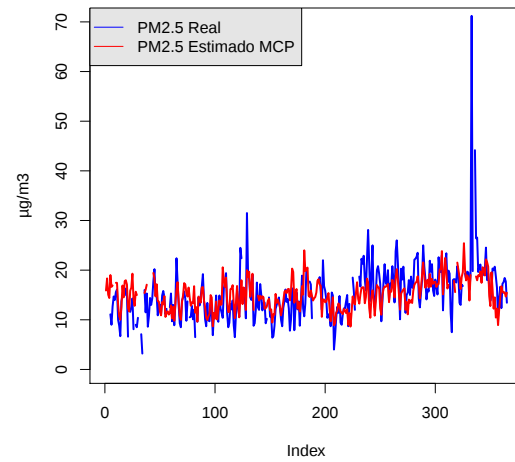
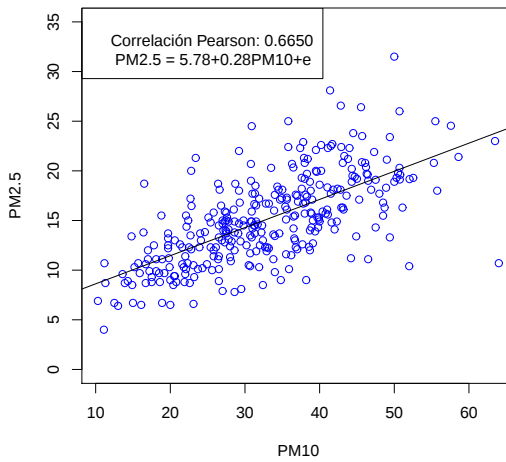


Figura 7.10: Estimaciones $PM_{2.5}$ por MCP Figura 7.11: Imputación mediante MCP

Tabla 7.4: Parametros de la Regresión por Mínimos Cuadrados Ponderados

	Estimate	Std. Error	t value	Pr(t)
(Intercept)	5.7422	0.0355	161.68	0.0000
PM10	0.2852	0.0013	213.94	0.0000

Como se puede observar en la Figura 7.11, las estimaciones de $PM_{2.5}$ por regresión de Mínimos Cuadrados Ponderados tiene un ajuste aceptable que permite realizar la imputación de los valores faltantes, no obstante, en los casos donde exista ausencia del $PM_{2.5}$ y a su vez del PM_{10} no es posible predecir estos valores, por lo que es necesario emplear otro método que permita la imputación de los 6 registros faltantes con el fin de reconstruir la serie temporal.

7.2.2. Análisis de intervención como método de imputación

El proceso de reconstrucción de la serie en esta fase se basa en el método aplicado por Elkin Castaño en su artículo "Reconstrucción de datos de series de tiempo: una aplicación a la demanda horaria de la electricidad", donde retomando la metodología de Chen y Lin en 1990 se realiza la imputación de los valores faltantes por medio de un proceso de análisis e intervención, asignándoles valores atípicos aditivos (AO por sus siglas en inglés) a los registros ausentes de la serie de tiempo. Es importante resaltar que por problemas de convergencia el análisis de esta sección solo se utiliza en casos donde existan suficientes registros a la derecha e izquierda del valor faltante.

La identificación de los valores Atípicos Aditivos se realiza por medio del software estadístico libre R, mediante el paquete **tsoutliers** desarrollado por Javier Lopez de Lacalle (2015), el cual aplica el procedimiento planteado por Chen y Liu en el año 1993. Se toma el valor de $100\mu g/m^3$ como valor atípico para el desarrollo de la metodología propuesta, luego se procede a utilizar un modelo autorregresivo de orden alto ($\sqrt[3]{n}$) ajustado a la cantidad de datos utilizados en cada tramo.

Tramo 1 de la serie del $PM_{2.5}$:

Se trabaja con un total de 224 datos en el periodo de estudio comprendido entre el 1 de Octubre del 2013 al 12 de Mayo del 2014, donde se imputaron un total de 20 valores faltantes en la primer etapa por medio de la regresión de MCP, persistiendo 3 valores faltantes a los cuales se les asigna un valor Atípico Aditivo de 10 considerada una concentración poco probable para la serie transformada del $PM_{2.5}$ (Ver Figura 7.12).

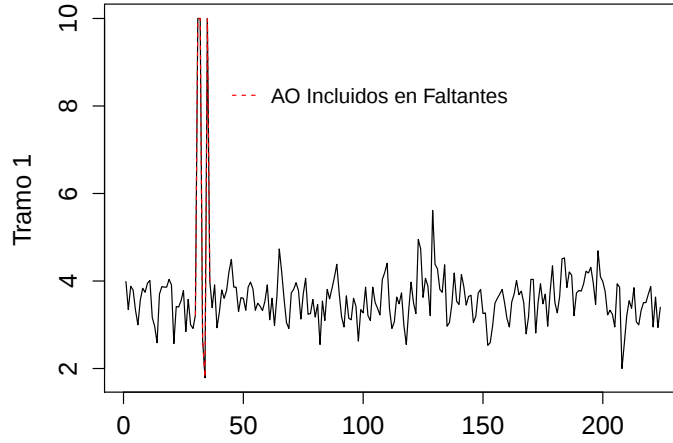


Figura 7.12: Histogramas $PM_{2.5}$ con transformaciones

Se emplea un proceso autorregresivo AR(6) como modelo inicial para la detección de las intervenciones en la serie de tiempo, en la Tabla 7.5 se presentan los valores atípicos y cambios estructurales identificados y estimados mediante el paquete estadístico tsoutliers, proceso que es alcanzado con tres iteraciones.

Tabla 7.5: Intervenciones Tramo 1

Etiqueta	Tipo de Intervención	Fecha de Intervención
31	AO	31/10/2013
32	AO	01/11/2013
34	AO	03/11/2013
35	AO	04/11/2013
123	TC	31/01/2014
129	TC	06/02/2014

Se realiza el filtrado de la serie, construyendo el modelo de ruido mediante la metodología Box y Jenkins, donde se identifica un modelo SARIMA(0,0,1)(0,0,1)[7] correctamente especificado, debido a que sus residuales tienden a comportandose como ruido blanco (Ver Anexo B.1). Las imputaciones de los datos faltantes resultan de la diferencia entre los valores originales al comienzo de la serie y las estimaciones del punto intervenido como se presenta en la Tabla 7.6.

Tabla 7.6: Observaciones Faltantes Estimadas

Etiqueta	Tipo de Intervención	Fecha de Intervención	Valor Original (Transf)	Estimación	Valor Imputado
31	AO	31/10/2013	$\sqrt[2]{100}$	6.536	3.464
32	AO	01/11/2013	$\sqrt[2]{100}$	6.752	3.248
35	AO	04/11/2013	$\sqrt[2]{100}$	6.360	3.64

Tramo 2 de la serie del $PM_{2.5}$:

El segundo tramo contiene 141 datos de los cuales 4 corresponden a valores faltantes, no obstante dos de estos cuatro valores se encuentran cercanos al cambio de nivel, generando problemas de convergencia en el método de imputación de datos empleado por Castaño. Para llevar a cabo el análisis de intervención e imputación de valores faltantes en el tramo 2 se trabaja desde las observaciones posteriores a los datos faltantes adyacentes al cambio estructural como se observa en la Figura 7.13, el cual es señalado por la línea punteada roja.

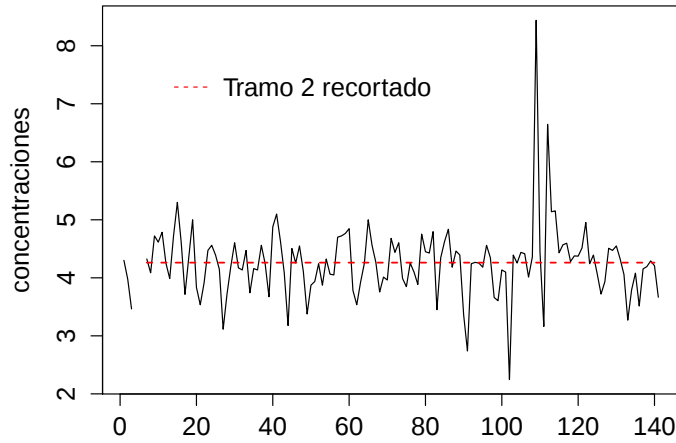


Figura 7.13: Tramo 2 recortado para análisis de intervención

El análisis de intervención de tramo recortado se realiza de manera similar al de la serie del tramo 1, incluyendo concentraciones atípicas (Asignándole valor de 10) en las observaciones faltantes y utilizando un modelo autorregresivo de orden 5 para la detección de outliers y cambios estructurales. La Tabla 7.7 muestra los valores a intervenir después de dos iteraciones.

Tabla 7.7: Intervenciones Tramo 2

Etiqueta	Tipo de Intervención	Fecha de Intervención
91	AO	11/08/2014
109	AO	29/08/2014
111	AO	31/08/2014
112	TC	01/09/2014

A priori se clasifica la concentración de $PM_{2.5}$ del día 29 de Agosto del 2014 como un valor atípico aditivo, con el fin de evitar identificaciones inadecuadas de ese punto por parte del programa estadístico *tsoutliers*. Por otro lado se observa en la Tabla 7.7 un cambio temporal, dados una la variación brusca en la serie que influyen en un instante de tiempo t y en las siguientes $t + k$ observaciones adyacentes, estabilizándose a su nivel promedio rápidamente. A continuación se presentan las estimaciones de los valores faltantes tratados como Atípicos Aditivos y la estimación del Atípico Aditivo real del 29 de Agosto:

Tabla 7.8: Intervenciones de los datos faltantes del Tramo 2

Etiqueta	Tipo de Intervención	Fecha de Intervención	Valor Original	Estimación	Valor Imputado
109	AO	29/08/2014	$\sqrt[3]{100}$	4.134	5.866
111	AO	31/08/2014	$\sqrt[3]{100}$	5.470	4.53

A partir de la serie filtrada se postula un modelo para el ruido haciendo uso de la metodología Box y Jenkins (Guerrero 2003), en la cual por medio de la función de autocorrelación simple (ACF) y la función de autocorrelación parcial (PACF) se identifica un proceso AR-MA(1,2), cumpliendo con los supuestos básicos, indicando que sus residuales se comportan como ruido blanco (Ver Anexo B.2).

Por ultimo, para la estimación de los dos datos faltantes cercanos al cambio de nivel y que no pudieron ser imputados por el método de castaño por presentar problemas de convergencia, se opta de manera recursiva, por ajustar una Sub-serie de tiempo a la parte completa del tramo 2, la cual corresponde a la línea punteada de la Figura 7.13, realizando pronósticos con este modelo para así completar el tramo alto de la serie de tiempo. El modelo ajustado, la validación de los supuestos y los pronósticos se puede consultar en el Anexo B.3.

Concluida la etapa de reconstrucción de la serie por la regresión de Mínimos Cuadrados Ponderados y el método empleado por Castaño, se procede a unir los dos tramos como se observa en la Figura 7.14, donde se concluye que los métodos de imputación tuvieron estimaciones plausibles acorde con el comportamiento esperado del Material Particulado menor a 2.5 micras.

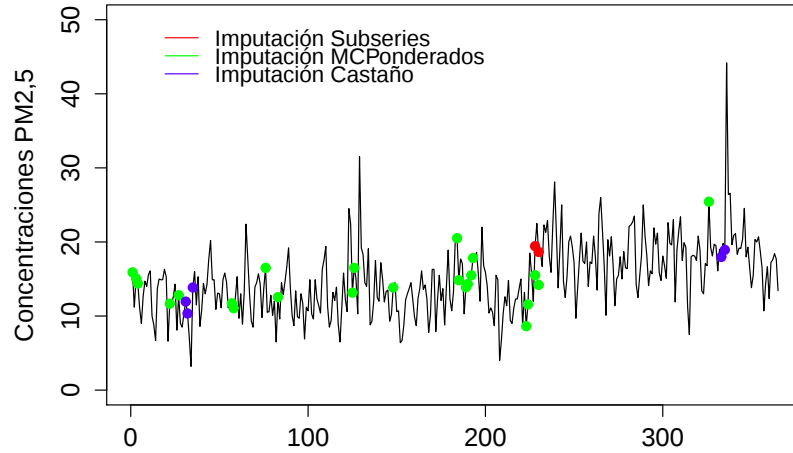


Figura 7.14: Valores imputados de $PM_{2.5}$ según método empleado

7.3. Modelación de la serie reconstruida

Para la construcción del modelo final se tienen en cuenta las intervenciones anteriormente planteadas en la etapa de reconstrucción por tramos, excluyendo los Atípicos Aditivos incluidos en las observaciones faltantes, donde en su lugar se incluirán las estimaciones de estos valores que pueden ser observadas en las Tablas 7.6 y 7.8. De igual manera el valor Atípico identificado para el 29 de Agosto del 2014 es reemplazado por el valor estimado del proceso propuesto por Castaño, E (2007), con el fin de evitar identificaciones erradas en las intervenciones cercanas a este valor.

En resultados anteriores se había evidenciado un cambio estructural en la serie el día 13 de Mayo del 2014, asociado posiblemente a ajustes en la metodología de cálculo en el promedio diario del $PM_{2.5}$ (Ausencia de ceros después de esta fecha). Por tal motivo se incluirá en el análisis de intervención de la serie completa un Cambio de Nivel Permanente (LS) a partir de esta fecha.

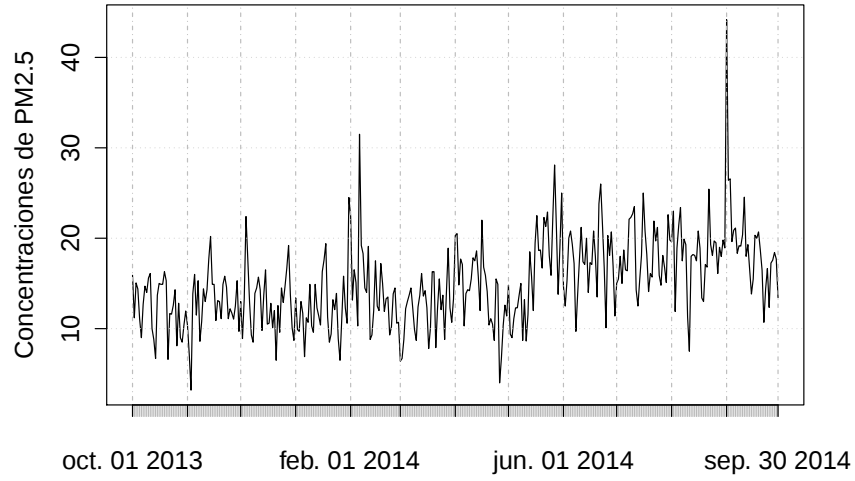


Figura 7.15: Serie de $PM_{2.5}$ completa

Al igual que la modelación por tramos, se plantea un modelo autorregresivo de orden alto, donde el orden del modelo está dado por $T^{1/3}$, con el cual se realizará la identificación de valores atípicos y cambios estructurales. Las intervenciones identificadas por el paquete estadístico tsoutliers con un total de 4 iteraciones se pueden observar en la Tabla 7.9.

Tabla 7.9: Intervenciones en la serie completa del $PM_{2.5}$

Etiqueta	Tipo de Intervención	Fecha de Intervención
34	AO	29/08/2014
123	TC	31/01/2014
129	TC	06/02/2014
208	AO	26/04/2014
209	AO	27/04/2014
228	LS	13/05/2014
315	AO	11/08/2014
336	TC	01/09/2014

Se presentaron un total de cuatro valores Atípicos Aditivos (AO), tres Cambios Temporales (TC) y un Cambio de Nivel (LS), donde al incluir estos factores en el modelo se obtiene la Serie Filtrada como se muestra en la Figura 7.16, con la cual, basados en la FAC, la FACP y la FAC Extendida se identifica el modelo de ruido.

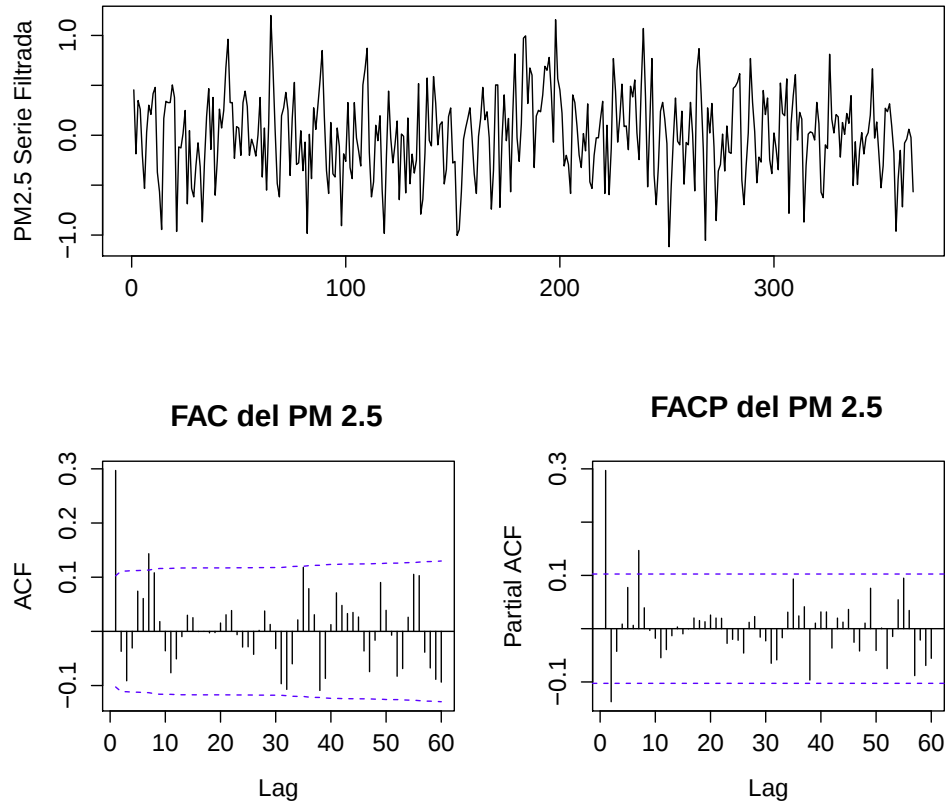


Figura 7.16: Serie Filtrada - FAC y FACP

Tabla 7.10: Función de Autocorrelación Extendida

AR/MA	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	x	o	o	o	o	o	x	x	o	o	o	o	o	o
1	x	x	o	o	o	o	o	o	o	o	o	o	o	o
2	x	o	x	o	o	o	o	o	o	o	o	o	o	o
3	x	o	x	o	o	o	o	o	o	o	o	o	o	o
4	x	x	x	o	x	o	o	o	o	o	o	o	o	o
5	o	x	o	x	x	o	o	o	o	o	o	o	o	o
6	o	x	o	x	x	x	x	o	o	o	o	o	o	o
7	x	o	o	x	o	o	o	o	o	o	o	o	o	o

Como se observa en la Figura 7.16 los efectos generados por los datos atípicos y cambios estructurales han sido corregidos al realizar el proceso de reconstrucción de la serie. Se observa de igual manera que la serie oscila alrededor de cero, ya que el efecto del intercepto fue incluido en la serie filtrada. Analizando la función de autocorrelación simple y parcial,

se postulan algunos modelos que pueden adaptarse a la estructura de la serie filtrada del contaminante $PM_{2.5}$, eligiendo como mejor modelo aquel que presente menor AIC, menor BIC y mayor Log-Verosimilitud.

Tabla 7.11: Criterios de selección del modelo

Modelo	AIC	BIC	Log-Verosimilitud
MA(0,1)	374.51	382.31	-185.26
ARMA(0,1)(0,1)[7]	371.10	382.8	-182.55
ARMA(1,2)(0,1)[7]	372.86	392.36	-181.43

Según el criterio de información de Akaike y el criterio de Log-Verosimilitud, el mejor modelo que se ajusta a la serie filtrada del $PM_{2.5}$ es el ARMA(0,1)(0,1)[7], sin embargo el criterio de información bayesiano sugiere al modelo MA(1) como el más adecuado, aunque un análisis detallado de los supuestos de cada modelo indica que el mejor modelo es atribuido al proceso ARMA estacional multiplicativo (Ver Anexo C). Además, retomando el análisis descriptivo realizado en capítulos anteriores, se observa que el contaminante $PM_{2.5}$ tiene a disminuir su concentración diaria el día domingo y lunes, marcando un leve efecto estacional.

El modelo final esta dado por:

$$\begin{aligned}
Z_t^* = & \theta_0 + AO_1 I_t^{34} + AO_2 I_t^{208} + AO_3 I_t^{209} + AO_4 I_t^{315} + \\
& \frac{1}{1 - \delta B} (TC_1 I_t^{123} + TC_2 I_t^{129} + TC_3 I_t^{336}) \\
& \frac{1}{1 - B} LS I_t^{228} + (1 - \theta_1 B) \\
& (1 - \Theta_1 B^7) a_t
\end{aligned} \tag{7.1}$$

En donde, $Z_t^* = \sqrt[2]{Z_t}$, θ_0 corresponde al intercepto, θ_1 representa el parámetro de media móvil, Θ_1 el parámetro de media móvil estacional, AO_i son las observaciones Atípicas Aditivas, TC_i son los cambios temporales del proceso, LS es el cambio de nivel permanente y a_t es la sucesión de ruido blanco. Las estimaciones de cada uno de los parámetros se muestra en la Tabla 7.12:

Tabla 7.12: Parámetros de las intervenciones Completas $PM_{2.5}$

Parámetro	Valor	Desv. Est	T-value
θ_0	3.5354	0.0401	88.1646
AO	-1.5661	0.3776	-4.1475
TC	1.4094	0.3467	4.0652
TC	1.7748	0.3464	5.1236
AO	-1.541	0.3972	-3.8797
AO	-0.7708	0.398	-1.9367
LS	0.6937	0.0646	10.7384
AO	-1.1986	0.3753	-3.1937
TC	2.045	0.3487	5.8646
θ_1	0.3323	0.0509	6.5285
Θ_1	0.1194	0.052	2.2962

Como se observa en la Figura 7.17, las raíces inversas tanto reales como imaginarias del polinomio característico son menores que 1, es decir que están dentro del círculo unitario, se puede concluir que el modelo es estable e invertible (Cavazos & Rivas 2009). Por otro lado, la serie estudiada obedece a un proceso $ARMA(0,1)(0,1)_7$, lo que indica que una observación en un tiempo t depende la de innovación actual mas la relación con la innovación de el periodo anterior, adicionando un efecto estacional cada siete días, es decir, la contaminación por $PM_{2.5}$ depende de la contaminación generada el día anterior, presentando un descenso en sus concentraciones los fines de semana.

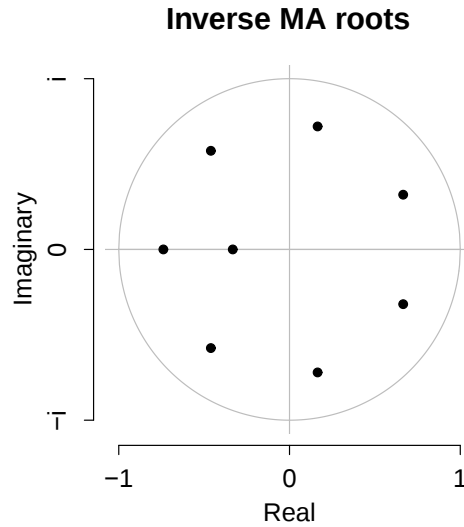


Figura 7.17: Raíces inversas del polinomio MA

La verificación del modelo final estimado por la metodología Box y Jenkins se realiza mediante el análisis de sus residuales, los cuales deben tener las siguientes características: Media marginal igual a cero, varianza marginal constante, falta de correlación para cualquier retardo y normalidad.

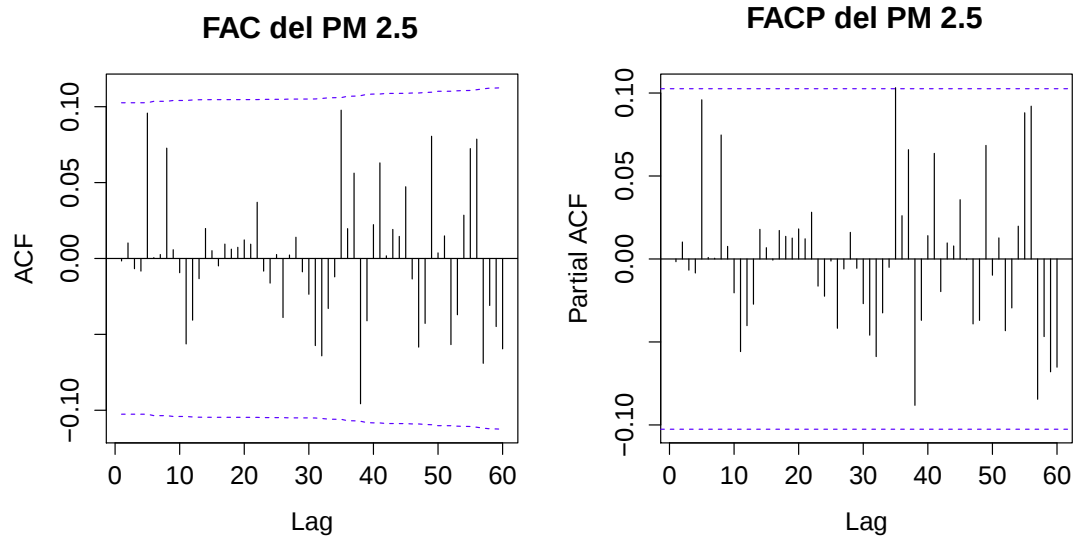


Figura 7.18: FAC y FACP residuales del modelo ARMA(0,1)(0,1)[7] mas las intervenciones

7.3.1. Varianza Constante y evaluación de existencia de observaciones atípicas

Se verifica en la Figura 7.19 que el comportamiento de la varianza es constante a lo largo del tiempo, oscilando alrededor de cero sin presentar tendencia alguna, vale la pena recordar que la serie estudiada se le aplicó la transformación $\sqrt[2]{Z_t}$ con el fin de estabilizar la varianza. Por otro lado, se evidencia la presencia de un dato atípico, el cual supera el límite establecido de $\pm 3\widehat{\sigma}_a$.

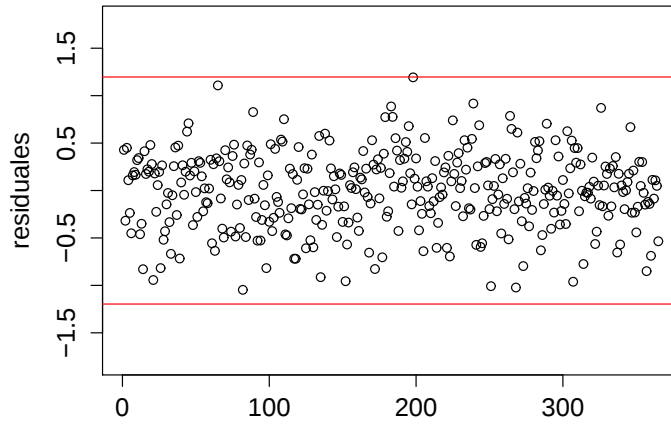


Figura 7.19: Supuesto varianza constante

Con el fin de contar con un soporte formal para este supuesto se realiza la prueba de Breusch-Pagan, partiendo de las siguiente prueba de hipótesis:

H₀: La varianza de los residuales es constante (Homocedasticidad)
H_a: La varianza de los residuales no es constante (Heterocedasticidad)

En donde se compara el estadístico de prueba $BP=0.63409$ con una Chi-Cuadrado con un grado de libertad, o lo que es equivalente su valor $P = 0.4259$ con un nivel de significancia del 5 %, concluyendo que la varianza se mantiene constante a lo largo de las observaciones.

Para observar si la serie temporal posee efectos ARCH, se realiza la prueba de McLeod-Li (1983), la cual esta basada en la prueba de Ljung-Box, tomando como entrada los residuales al cuadrado. Como se observa en la Figura 7.20, los valores P de la prueba para cualquier rezago son superiores al nivel de significancia del 5 %, concluyendo que no existen evidencias suficientes para rechazar la hipótesis nula de que las autocorrelaciones al cuadrado son iguales a cero, es decir, la serie del $PM_{2.5}$ no presenta problemas de heterocedasticidad condicional.

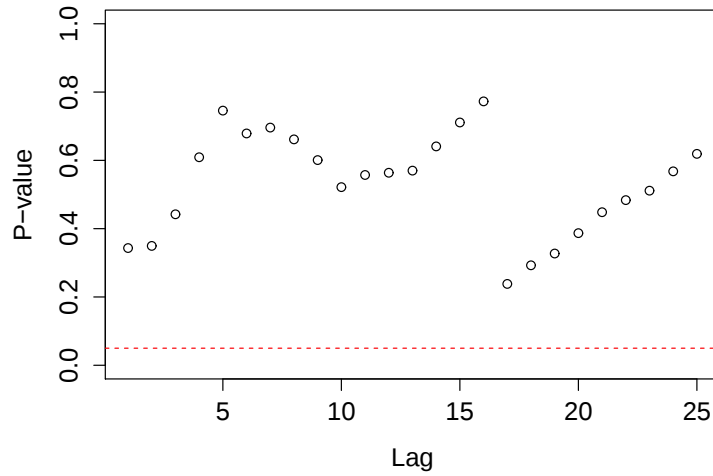


Figura 7.20: Supuesto varianza constante

7.3.2. Contraste de Ljung-Box sobre las autocorrelaciones

Se realiza la prueba de Ljung-Box a cada conjunto de rezagos, observando en la Figura 7.21 que los valores P se alejan del nivel de significancia del 5 % a medida que la cantidad de autocorrelaciones en conjunto aumentan, corroborando que los residuales del modelo se comportan como ruido blanco. Vale la pena aclarar que según Guerrero (2003) cuando K es suficientemente grande ($K > 20$), el estadístico de prueba se debe comparar con una distribución Chi-Cuadrado con $K-p-q$ grados de libertad, donde K es la cantidad de rezagos evaluados conjuntamente, p es el orden de la parte autorregresiva del modelo y q el orden del proceso de media móvil.

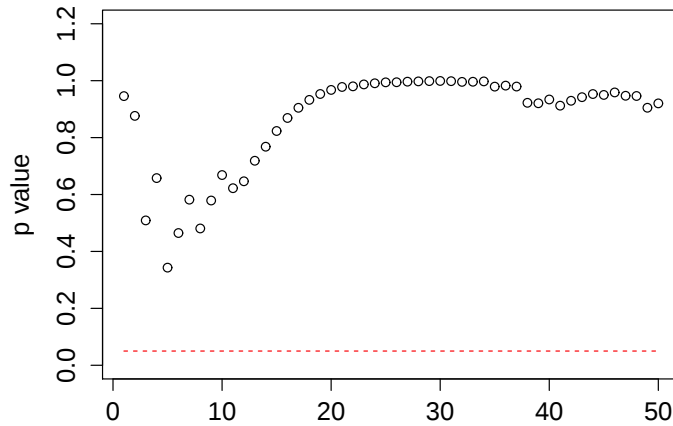


Figura 7.21: Gráfica de P-Valores de la prueba de Ljung-Box con k autocorrelaciones

7.3.3. Normalidad de los residuales del modelo

Como se demuestra en la Figura 7.22, el gráfico Q-Q plot indica que la distribución de los residuales es normal, debido a que aproximadamente el 95 % de las observaciones están contenidas por la distribución teórica de estudio. No obstante, se plantea la prueba de Jarque-Bera para darle más soporte al análisis, partiendo de las siguientes hipótesis:

Ho: Los residuales se distribuyen normal

Ha: Los residuales no se distribuyen normal

El valor P arrojado por la prueba de Jarque-Bera es de 0.4626, mayor que el nivel de significancia planteado de 5 %, concluyendo a un nivel de confianza del 95 % que los residuales siguen una distribución normal. De igual manera el p-valor para el parámetro de simetría es de 0.2259, indicando que la distribución es simétrica, corroborando gráficamente con el histograma presentado en la Figura 7.22

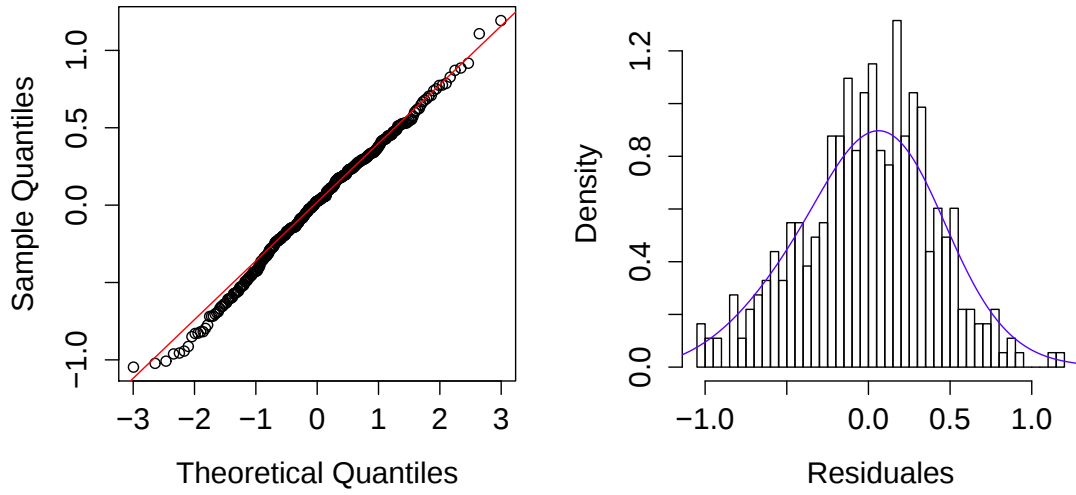


Figura 7.22: Q-Q Plot y Histograma para evaluar normalidad en los residuales

Para realizar los pronósticos del modelo es necesario primero aplicar una transformación inversa a la serie, ya que los pronósticos transformados carecen de interpretabilidad. Para calcular la transformación se emplea el factor de corrección propuesto en (Guerrero 2003), con el fin de reducir el sesgo que se genera al pasar de $T(Z_t)$ a la serie original Z_t .

$$\hat{c}_\lambda(h) = \begin{cases} 1/2 + \sqrt{1 - \frac{2(\lambda-1)}{\lambda} [\hat{T}(Z_t)(h)]^{-2} \hat{Var}[e_t(h)]/2} \\ \exp \hat{Var}[e_t(h)]/2 \end{cases} \quad (7.2)$$

Los pronósticos generados para la Serie original se presentan en la Tabla B.4

Tabla 7.13: Pronósticos modelo ARMA(0,1)(0,1)[7] $PM_{2.5}$

Dia	Fecha	Valor Real	Pronostico	$e_t(h)^2$	Intervalos	
					Lo 95	Hi 95
Jueves	01/10/2014	17.79	14.14	13.32	10.56	23.17
Viernes	02/10/2014	24.75	17.16	57.54	11.02	24.67
Sabado	03/10/2014	21.08	17.79	10.85	11.52	25.41
Domingo	04/10/2014	13.54	17.86	18.63	11.57	25.50
Lunes	05/10/2014	7.33	17.99	113.53	11.68	25.65
Martes	06/10/2014	11.23	17.99	45.71	11.68	25.66
Miércoles	07/10/2014	9.86	17.38	56.57	11.19	24.93

Los pronósticos del modelo final tienden a subestimar las concentraciones reales los tres primeros días del mes de Octubre y sobreestimar los siguientes tres días, no obstante 4 de los 7 intervalos de predicción contienen el valor real de los promedios diarios del $PM_{2.5}$ a un nivel de confianza del 95 %. El Error Cuadrático Medio (ECM) para los pronósticos es de 45.16, el cual esta distante de cero, evidenciando la baja capacidad predictiva del modelo planteado, esto debido a que al estar compuesto por un parámetro de media móvil, la capacidad del pronóstico se ve reducida al efecto de la observación del día anterior y el factor estacional que generará la disminución presentada por el $PM_{2.5}$ los fines de semana.

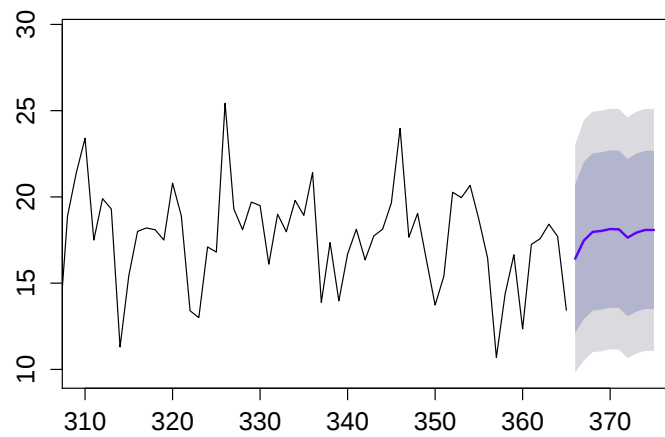


Figura 7.23: Predicciones modelo SARIMA(0,0,1)(0,0,1)[7]

De igual manera, se concluye que los pronósticos tienen valores aceptables en media, siguiendo la tendencia de la serie como se ve en la Figura 7.23, no obstante la variabilidad del contaminante entre días nos es recogida en su totalidad por el modelo propuesto, razón por la cual se recomienda la inclusión de variables o métodos que abarquen la volatilidad no modelada en este proyecto.

Capítulo 8

Conclusiones

Cuando se analizó el comportamiento de la serie de datos, se identificó un incremento en el nivel de $PM_{2.5}$ en el mes de mayo, que podría atribuirse a alguna novedad como: la presencia de una fuente externa de contaminación, el cambio de metodología de cálculo de los promedios diarios ó el mantenimiento de los instrumentos de medición. Es por esta razón, que las entidades responsables deben ser cuidadosos cuando se evidencia algún tipo de alteración en el comportamiento habitual del contaminante, con el propósito de dar claridad acerca del motivo de las variaciones y evitar conclusiones prematuras sobre un posible aumento en los niveles de contaminación en Cali.

Es necesario identificar previo a la modelación los eventos exógenos que puedan generar ruido en la identificación de la serie, como es el caso de los valores atípicos aditivos, atípicos Innovativos, los cambios estructurales o los cambios de nivel, debido a que pueden influir en la parte estocástica y determinística de la serie afectando la identificación del modelo adecuado y proporcionando así predicciones poco acertadas.

Debido a los problemas de sensibilidad de la transformación Box-Cox en presencia de observaciones atípicas y cambios estructurales, el método planteado de elección del λ en el proyecto, presento buenos resultados al solucionar la asimetría de la serie ayudando a corregir los inconvenientes del supuesto de normalidad que presentaban los residuales del modelo estimado por la metodología Box-Jenkins.

Las imputaciones obtenidas por medio de la regresión por Mínimos Cuadrados Ponderados y la metodología propuesta por Castaño (2007) presentaron estimaciones plausibles acorde con el comportamiento esperado del material particulado menor a 2.5 micras, permitiendo emplear metodologías estadísticas que requieren tener un conjunto de datos completos.

Al realizar la comparación de las concentraciones diarias de $PM_{2.5}$ en presencia de días con lluvia y sin lluvia se encontró una diferencia poco significativa, contradiciendo otros estudios realizados sobre el material particulado fino, donde demuestran que este contaminante está fuertemente correlacionado con la precipitación al funcionar esta variable meteorológica como agente de limpieza en el medio ambiente. Una posible razón para no encontrar los resultados esperados, es el nivel de dato tomado en este estudio (Promedio diario), debido a que los datos reportados por el IDEAM muestran que la precipitación en Cali se da por lo general pocas horas al día (Periodos de lluvia menor a 3 horas), por tal motivo se sugiere realizar el análisis del $PM_{2.5}$ con esta variable a un nivel de dato horario.

Los pronósticos del modelo siguen el comportamiento de la serie en cuanto a nivel, no obstante, al estar compuesto por un solo parámetro de media móvil su capacidad predictiva es muy limitada, es decir, al intentar realizar varias predicciones el valor del pronóstico va a tender al promedio del modelo. Es recomendable incluir en el modelo variables que puedan aportar en el ajuste y estimación de pronósticos, como es el caso del Material Particulado menor a 10 micras o algunas variables climatológicas de interés que influyan en el $PM_{2.5}$. Por otro lado, las estimaciones tienden a sobreestimar y subestimar los valores reales, aumentando su error cuadrático a medida que los pronósticos aumentan.

Capítulo 9

Recomendaciones

Al utilizar el promedio diario del $PM_{2.5}$ se disminuye la cantidad de registros faltantes para llevar a cabo la identificación de un modelo por medio de series de tiempo, sin embargo las propiedades de usar los datos en su mínima expresión (Horarios) enriquece más el estudio proporcionando información valiosa del comportamiento del contaminante en una escala de tiempo menor. Teniendo como ventaja igualmente el poder construir modelos horarios que permitan aumentar el número de predicciones.

El método de imputación de datos por Mínimos Cuadrados Ponderados presentó resultados plausibles para completar la serie, conservando la estructura de autocorrelación del $PM_{2.5}$ al tomar como referencia un contaminante altamente relacionado con este. No obstante, se pueden tener resultados más satisfactorios al tener los dos contaminantes (PM10 y PM2.5) en el mismo lugar de medición, aumentando así la correlación entre ellos y por ende mejorando sus estimaciones.

La transformación Box-Cox es sensible a la presencia de datos atípicos, generando estimaciones de λ poco confiables, por tal razón se recomienda evaluar otros métodos de transformación robustos, como por ejemplo los métodos no paramétricos propuesto por Castaño.

La alta variabilidad que presenta el $PM_{2.5}$ es una característica recurrente de la mayoría de contaminantes atmosféricos, por tal motivo y en caso que se demuestre presencia de heterocedasticidad, se sugiere tener en cuenta este factor a la hora de realizar la modelación de estas variables. Autores como Engle (1942) recomiendan utilizar modelos de heterocedasticidad condicional como los procesos ARCH, GARCH, entre otros, los cuales logran explicar ciertas propiedades que no son identificadas por los modelos ARIMA.

Anexo A

Relación contaminantes y variables meteorológicas

Tabla A.1: Supuestos Serie Filtrada Completa

Paquete	Uso	Referencia
xts	Crea un objeto de serie temporales y facilita la inclusión de fechas	Jeffrey & Ulrich (2013)
tsoutliers	Identifica, crea y reemplaza valores atípicos y cambios estructurales	Javier Lopez de Lacalle (2015)
TSA	Elimina el problema de la primer correlación de la FAC y FACP.	Chan & B (2012)
forecast	Realiza predicciones de los modelos ARIMA	Hyndman (2006)
nortest	Pruebas de normalidad (Jarque Bera en este proyecto)	Gross & Ligges (2015)
lmtest	Test de modelos de regresión lineal. Breusch Pagan en el caso de este proyecto	Zeileis & Hothorn (2002)

A.1. Rosa de viento de la estación Pance - Nivel de polución del PM10

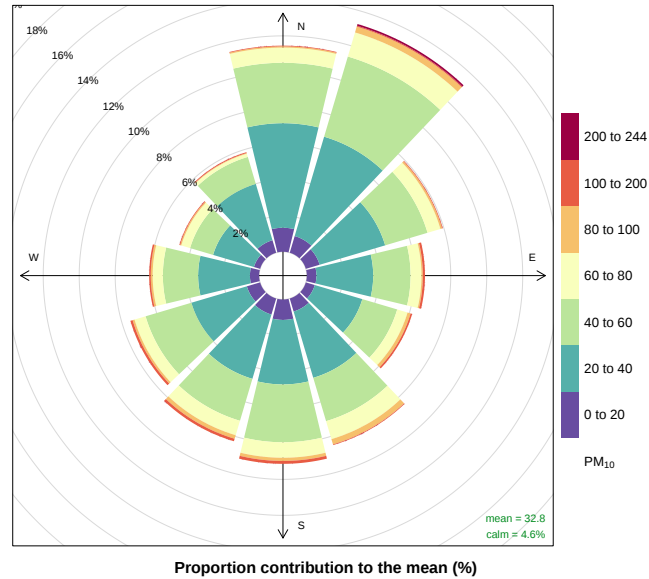


Figura A.1: Rosa de viento de la estación Pance

A.2. Correlación del PM 2.5 con las variables meteorológicas

Tabla A.2: Correlación entre contaminante y variables meteorológicas

	PM2.5	Vel Viento	Temperatura	Humedad	Rad Solar
Vel Viento	-0.160				
Temperatura	-0.056	0.489			
Humedad	0.030	-0.479	-0.966		
Radiacion	-0.004	0.134	0.563	-0.517	
Lluvia	-0.076	0.019	-0.102	0.115	-0.064

A.3. Dirección del viento estación Univalle y estación de Pance

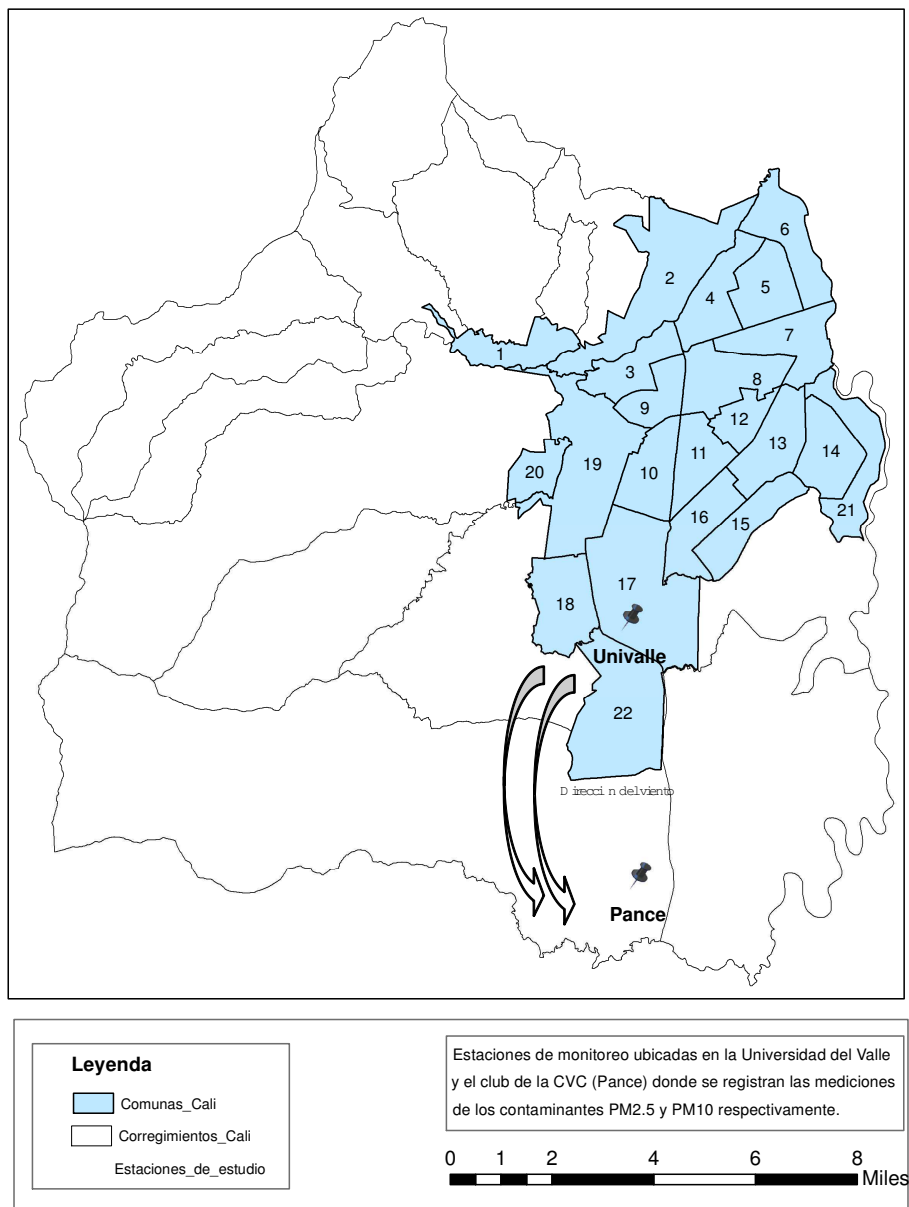


Figura A.2: Mapa ubicación de las estaciones y dirección del viento

Anexo B

Análisis de los residuales de los tramos de la serie

B.1. FAC, FACP y evaluación de los supuestos del Tramo 1

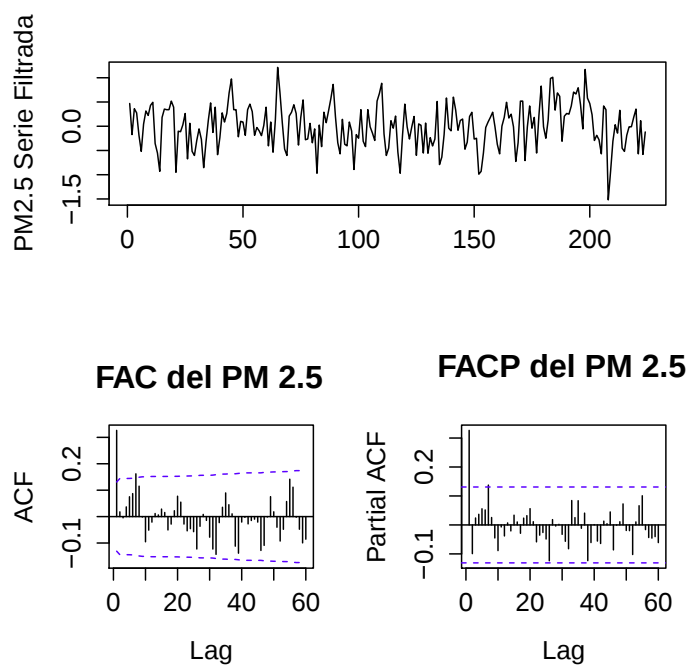


Figura B.1: Serie Filtrada, FAC y FACP Tramo 1

Tabla B.1: Supuestos Tramo 1

	AIC	Ljung-Box	Jarque Bera	Breusch Pagan
ARMA(0,0,1)(0,0,1)[7]	250.45	0.9848	0.06583	0.2706
MA(1)	252.66	0.9289	0.06583	0.2856

B.2. FAC, FACP y evaluación de los supuestos del Tramo 2

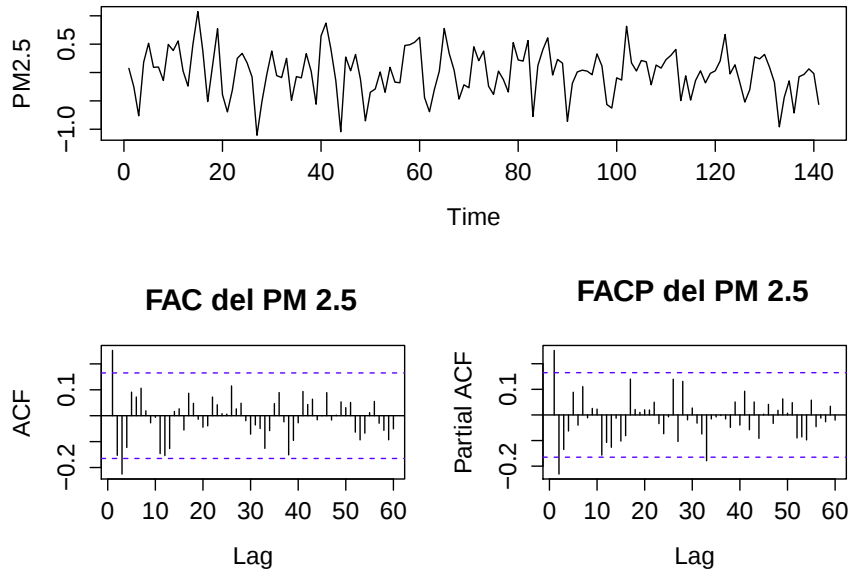


Figura B.2: Serie Filtrada, FAC y FACP Tramo 2

Tabla B.2: Supuestos Tramo 2

	AIC	Ljung-Box	Jarque Bera	Breusch Pagan
AR(2)	135.63	0.6939	0.5472	0.2179
ARMA(1,0,2)	135.64	0.7953	0.6028	0.1979
ARMA(2,0,1)	135.22	0.75	0.759	0.164

B.3. FAC, FACP y evaluación de la Subserie

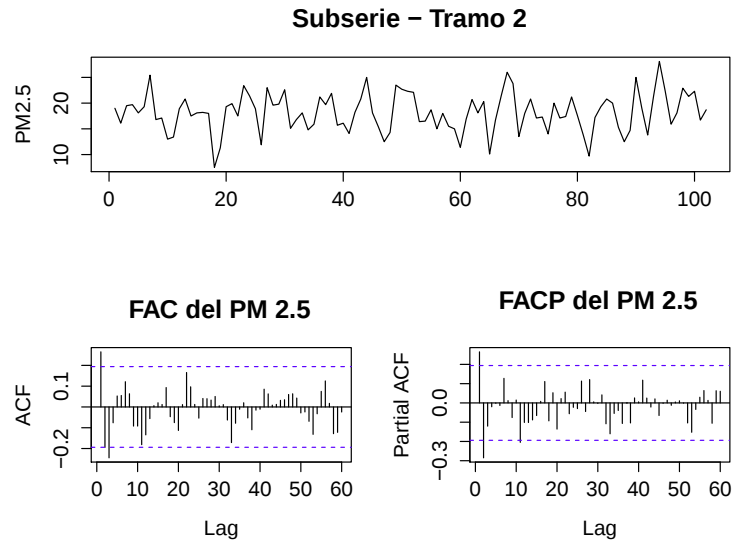


Figura B.3: FAC y FACP Subserie

Tabla B.3: Supuestos Series Subtramo

	AIC	Ljung-Box	Jarque Bera	Breusch Pagan
MA(1)	552.38	0.8602	0.6664	0.6845
ARMA(1,0,2)	548.09	0.8744	0.8067	0.6845
ARMA(2,0,1)	553.3	0.9085	0.6553	0.6553

Tabla B.4: Pronósticos modelo MA(1) para la imputación de 2 datos faltantes

Dia	Fecha	Pronostico	Intervalos	
			Lo 95	Hi 95
Domingo (Faltante)	18/05/2014	18.74047	11.83971	25.64123
Sabado (Real)	17/05/2014	22.5		
Viernes (Faltante)	16/05/2014	18.174	10.82552	25.52386

Anexo C

Análisis de los residuales del modelo filtrado completo

C.1. FAC, FACP y evaluación de los supuestos de la serie completa

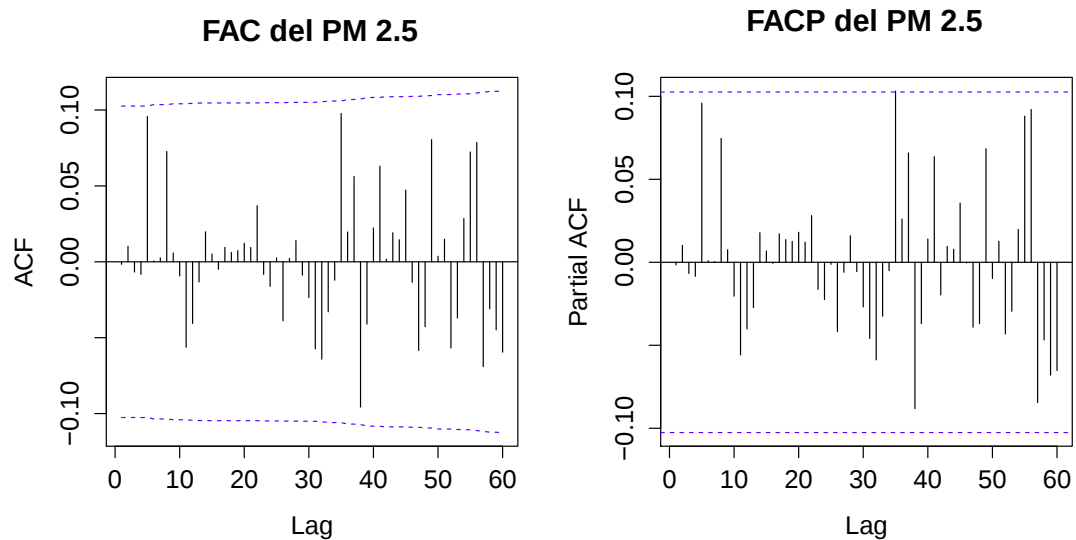


Figura C.1: FAC y FACP residuales modelo $\text{ARMA}(0,1)(0,1)[7]$

Tabla C.1: Supuestos Serie Filtrada Completa

	AIC	Ljung-Box	Jarque Bera	Breusch Pagan
MA(1)	374.51	0.8465	0.2291	0.395
ARMA(0,1)(0,1)[7]	371.10	0.9826	0.5881	0.4259
ARMA(1,2)(0,1)[7]	372.86	0.9825	0.4628	0.4255

C.2. Intervalos para el ajuste de la serie de tiempo

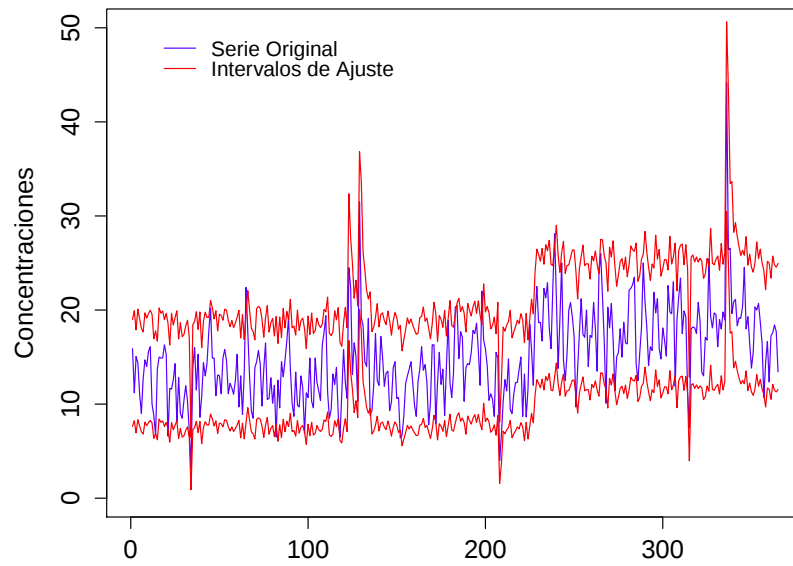


Figura C.2: Intervalos de ajuste Serie Completa

Bibliografía

- Agudelo, D., Teixeira, E. & Pereira, F. (2014), ‘Time-series analysis of surface ozone and nitrogen oxides concentrations in an urban area at brazil’, *Atmospheric Pollution Research* **5**(3), 411 – 420.
- Arce, R. & Mahía, R. (2000), ‘Modelos arima’, Nota Universitaria. Metodología Box y Jenkins.
- Bickel, P. & Doksum (1981), ‘An analysis of transformations revisited’, *Journal of the American Statistical Association* **76**(296-311).
- Box, G. & Cox, D. (1964), ‘An analysis of transformations’, *Journal of the Royal Statistical Society* **26**, 211–252.
- Box, G. & Tiao, G. (1975), ‘Intervention analysis with applications to economic and environmental problems’, *Journal of the American Statistical Association* **70**(349), 70–79.
- Buchinsky, M. (1995), ‘Quantile regression, box-cox transformation model, and the u.s. wage structura, 1963-1987’, *Journal of Econometrics* **65**(100-154).
- Cadena, A. & García, C. (2014), ‘Modelación del contaminante dióxido de azufre (so2) en la ciudad de cali entre los años 2010 y 2011’.
- Cali Como Vamos (2013), ‘Informe de Calidad de Vida en Cali 2012’.
- Carroll & Raymond (1980), ‘A robust method for testing transformation to achieve normality’, *Journal of the Royal Statistical Society* **42**(71-78).
- Castaño, E (2007), ‘Reconstrucción de datos de series de tiempo: una aplicación a la demanda horaria de la electricidad’, *Revista Colombiana de Estadística* **30**(2), 247–263.
- Castaño, E. (2011), ‘Una estimación no paramétrica y robusta de la transformación box-cox para el modelo de regresión’, *Lecturas de Economía Universidad de Antioquia* .

- Cavazos, G. & Rivas, S. (2009), ‘Relación entre la inflación y tasas de interés en México y Estados Unidos’, *Scielo: Problemas del desarrollo* **40**(157).
- Chamberlain, G. (1994), ‘Quantile regression, censoring, and the structure of wages’, *Econometric Society Monograph* **1**.
- Chan, K. & B, R. (2012), ‘Time series analysis’, [urlhttps://cran.r-project.org/web/packages/TSA/TSA.pdf](https://cran.r-project.org/web/packages/TSA/TSA.pdf).
- Chen, C. & Liu, L. (1993), ‘Joint estimation of model parameters and outlier effects time series’, *Journal of the American Statistical Association* **88**(421), 284–297.
- Chen, C. & Liu, L.-M. (1990), ‘Join estimation of model parameters and outlier effects in time series’, *Scientific Computing Associates* **88**(421).
- DAGMA (2013), Informe de Gestión Ambiental 2013, Technical report, Departamento Administrativo de la Gestión del Medio Ambiente (DAGMA).
- Echeverri, L. & Maya, G. (2008), ‘Relación entre las partículas finas (pm2.5) y respirables (pm10) en la ciudad de Medellín’, *Revista de ingeniería ambiental, Universidad de Medellín* **7**(12), 23–42.
- ElPaís (2014), Por qué el sur de Cali fue cubierto por un manto de niebla en la mañana de hoy, Technical report, Diario el País.
- García, J. & Gonzalez, T. (2014), ‘Modelación de la concentración máxima diaria de ozono en la zona norte de la ciudad de Cali, periodo 2010 a 2011: Estación meteorológica exito de la flora’.
- Gross, J. & Ligges, U. (2015), ‘Tests for normality’, [urlhttps://cran.r-project.org/web/packages/nortest/nortest.pdf](https://cran.r-project.org/web/packages/nortest/nortest.pdf).
- Guerrero, V. (2003), *Análisis estadístico de series de tiempo económicas*, segunda edición edn.
- Gujarati, D. (2010), *Econometria*, Vol. 5, Mc Graw Hill.
- Hyndman, R. (2006), ‘Forecasting functions for time series and linear models’, [urlhttp://github.com/robjhyndman/forecast](http://github.com/robjhyndman/forecast).
- IARC (2013), ‘Monografías IARC: La contaminación del aire es una de las principales causas de muerte por cáncer’, *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans* **109**.

- Iglesias, P., Jorquera, H. & Palma, W. (2006), ‘Data analysis using regression models with missing observations and long-memory an application study’, *Computational Statistics and Data Analysis* (50), 2028–2043.
- Javier Lopez de Lacalle (2015), *Package tsoutliers*, Foundation for Statistical Computing.
- Jeffrey, A. & Ulrich, J. (2013), ‘Create or test for an xts time-series object’, [urlhttps://cran.r-project.org/web/packages/xts/xts.pdf](https://cran.r-project.org/web/packages/xts/xts.pdf).
- Krasnov, H., Katra, I. & Friger, M. (2015), ‘Increase in dust storm related pm10 concentrations: A time series analysis of 2001 - 2015’, *Environmental Pollution* **213**, 36–42.
- Kumar, U. & Ridder, K. (2014), ‘Modeling and forecasting daily average pm10 concentrations by a seasonal long-memory model with volatility’, *Environmental Modelling and Software* **51**, 286 – 295.
- McAleer, M. & Chan, F. (2006), ‘Modelling trends and volatility in atmospheric carbon dioxide concentrations’, *Environmental Modelling and Software* **21**(9), 1273 – 1279.
- Ministerio de Ambiente Vivienda y Desarrollo Territorial (2010), ‘Resolución 610 de 2010: Por la cual se modifica la resolución 601 del 4 de abril de 2006’.
- Ministerio del Medio Ambiente (2002), ‘Gestión de calidad del aire’.
- Montoya, L. & Orobio, S. (2009), ‘Modelacion de la contaminacion del aire por material particulado PM 10 en Santiago de Cali durante el periodo 2003 a 2006 a traves de regresion no parametrica’, *Universidad del Valle* .
- Murcia (2010), ‘Consejería de sanidad de la región: Material particulado (pm10 y pm2,5)’, [urlhttp://www.murciasalud.es/pagina.php](http://www.murciasalud.es/pagina.php).
- OMS (2005), ‘Guías de calidad del aire de la oms relativas al material particulado, el ozono, el dióxido de nitrógeno y el dióxido de azufre’.
- OMS (2011), ‘Calidad del aire en exteriores y efectos en la salud’, [urlhttp://www.who.int/mediacentre/factsheets/fs313/es](http://www.who.int/mediacentre/factsheets/fs313/es).
- Peña, D. (2005), *Análisis de series temporales*, primera edición edn.
- Powell, J. (1991), ‘Estimation of monotonic regression models under quantile restrictions’, *Cambridge University Press* .
- Rojano, R., Angulo, L. & Restrepo, G. (2013), ‘Niveles de particulas suspendidas totales (pst), pm10 y pm2.5 y su relación en lugares públicos de la ciudad de rioacha, caribe colombiano’, *Scielo* **24**(2), 37–46.

- Rojas, N. & Galvis, B. (2005), ‘Relación entre pm2.5 y pm10 en la ciudad de bogotá’, *Universidad de los Andes: Revista de ingeniería* **22**, 54–60.
- Romero, M., Román, R., Limón, J. M. G. & Praena, M. (2015), ‘Effects of fine particles on childrens hospital admissions for respiratory health in seville’, *Journal of the Air and Waste Management Association* **65**, number = 4, pages = 436-444,.
- U.S. Environmental Protection Agency EPA (2009), ‘A Guide to Air Quality and Your Health’.
- Valdes, J. & Torres, W. (2009), ‘Modelación de los precios diarios en bolsa de energía eléctrica en colombia’.
- Wang, W., Van Gelder, P., J, V. & Ma, J. (2005), ‘Testing and modelling autoregressive conditional heteroskedasticity of streamflow processes’, *European Geosciences Union: Nonlinear Processes in Geophysics* **12**, 55–66.
- Wei, W. (2006), *Time Series Analysis*, second edition edn.
- Yaglom, A. (1985), ‘Correlation theory of processes with random stationary nth increments’, *American Mathematical Society* **2**(8), 87 – 141.
- Zeileis, A. & Hothorn, T. (2002), ‘Diagnostic checking in regression relationships’, *R News* **2**(3), 7–10.