

**MODELACIÓN DE LA CONTAMINACIÓN DEL AIRE POR
MATERIAL PARTICULADO (PM10) EN SANTIAGO DE CALI A
TRAVÉS DE REGRESIÓN NO PARAMETRICA DURANTE EL
PERÍODO 2003-2006**

**LUISA MARÍA MONTOYA QUESADA
SANDRA MILENA OROBIO BELALCAZAR**

**ESCUELA DE INGENIERÍA INDUSTRIAL Y ESTADÍSTICA
PROGRAMA ACADÉMICO DE ESTADÍSTICA
FACULTAD DE INGENIERÍA
UNIVERSIDAD DEL VALLE
SANTIAGO DE CALI**

2011

**MODELACIÓN DE LA CONTAMINACIÓN DEL AIRE POR
MATERIAL PARTICULADO (PM10) EN SANTIAGO DE CALI A
TRAVÉS DE REGRESIÓN NO PARAMETRICA DURANTE EL
PERÍODO 2003-2006**

**LUISA MARÍA MONTOYA QUESADA
SANDRA MILENA OROBIO BELALCAZAR**

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

**Director Proyecto de Grado
ÁLVARO JOSÉ FLÓREZ
Estadístico**

**ESCUELA DE INGENIERÍA INDUSTRIAL Y ESTADÍSTICA
PROGRAMA ACADÉMICO DE ESTADÍSTICA
FACULTAD DE INGENIERÍA
UNIVERSIDAD DEL VALLE
SANTIAGO DE CALI**

2011

DEDICATORIA

*“A Dios, a mis padres y a mi hermana
por su amor, comprensión, paciencia y
apoyo incondicional”*

Luisa María Montoya

*“Este éxito es dedicado a Dios, mi
madre, mi padre, familiares y seres
queridos que me acompañaron y apoya-
ron en la culminación de mis estudios”*

Sandra Milena Orobio

AGRADECIMIENTOS

Damos gracias principalmente a Dios por darnos la oportunidad de culminar con éxito esta etapa de nuestras vidas, por guiarnos en el camino y por darnos las fuerzas necesarias para seguir adelante cada día. A nuestras familias por brindarnos su amor, comprensión y apoyo incondicional a lo largo de esta carrera.

Gracias a las enseñanzas de todos los profesores que hicieron de nosotras unas profesionales, en especial a los profesores Álvaro Flórez y Javier Olaya por darnos la oportunidad de trabajar a su lado, por su orientación, sugerencias y apoyo en la culminación de este trabajo. A la joven investigadora Jhovana Reina por acompañarnos durante este proceso y quien a través de su asesoría en el tema, su ayuda y sugerencias hizo posible que este trabajo se desarrollara de manera exitosa. De la misma manera agradecemos a Adrian W. Bowman quien desde la lejanía estuvo dispuesto a ayudarnos, brindándonos sus conocimientos sobre el tema.

Agradecemos enormemente al Departamento Administrativo de Gestión del Medio Ambiente (DAGMA) por suministrarnos la base de datos sobre material particulado, al Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM) por proporcionarnos la información sobre los registros climatológicos para la ciudad de Cali y a la Vicerrectoría de Investigaciones de la Universidad del Valle por el apoyo financiero en la ejecución de este trabajo.

Por último agradecemos a nuestros amigos, con quienes compartimos tristezas, alegrías y logros durante el camino recorrido.

CONTENIDO

RESUMEN	8
ABSTRACT	9
INTRODUCCIÓN	10
1. PLANTEAMIENTO DEL PROBLEMA	11
2. ANTECEDENTES	12
2.1. MODELACIÓN DE PM10 BAJO DIFERENTES TÉCNICAS ESTADÍSTICAS .	12
2.2. REGRESIÓN NO PARAMÉTRICA EN LA MODELACIÓN DE CONTAMINANTES	13
3. OBJETIVOS	15
3.1. OBJETIVO GENERAL	15
3.2. OBJETIVOS ESPECÍFICOS	15
4. JUSTIFICACIÓN	16
5. MARCO TEÓRICO	18
5.1. ELEMENTOS DEL CLIMA	18
5.2. LA ATMÓSFERA	18
5.3. CONTAMINACIÓN ATMOSFÉRICA	19
5.3.1. FUENTES FIJAS O ESTACIONARIAS	19
5.3.2. FUENTES MÓVILES	19
5.4. CONTAMINANTES ATMOSFÉRICOS	19
5.4.1. PARTÍCULAS SUSPENDIDAS TOTALES (PST)	20
5.4.2. MATERIAL PARTICULADO (PM10)	20
5.5. MONITOREO ATMOSFÉRICO	20
5.5.1. MONITOREO DE LA CALIDAD DEL AIRE EN SANTIAGO DE CALI .	21
5.6. NORMAS DE CALIDAD DE AIRE	23
5.7. TEMPORALIDAD EN EL ANÁLISIS DE PM10	23
5.8. MODELO DE REGRESIÓN NO PARAMÉTRICA	24
5.9. SUAVIZACIÓN	24
5.10. FUNCIONES KERNEL	25
5.11. ESTIMADORES EN REGRESIÓN NO PARAMÉTRICA	26
5.11.1. ESTIMADOR NADARAYA-WATSON	26
5.11.2. ESTIMADOR GASSER-MÜLLER	26
5.11.3. ESTIMADOR LINEAL LOCAL	26
5.11.4. ESTIMADOR LOWESS	27

5.12. INFERENCIA EN EL MODELO NO PARAMÉTRICO	27
5.12.1. SELECCIÓN DEL ANCHO DE BANDA	28
5.12.2. ESTIMACIÓN DE LA VARIANZA	29
5.12.3. BANDAS DE VARIABILIDAD	31
5.12.4. COMPARACIÓN DE CURVAS	32
6. METODOLOGÍA	35
6.1. REVISIÓN DE LA LITERATURA	35
6.2. RECOLECCIÓN DE LA INFORMACIÓN	35
6.3. DEPURACIÓN Y VALIDACIÓN DE DATOS	36
6.4. HERRAMIENTA COMPUTACIONAL	38
6.5. ANÁLISIS ESTADÍSTICO	39
6.5.1. ANÁLISIS EXPLORATORIO DE DATOS	39
6.5.2. ESTIMACIÓN DE CURVAS TÍPICAS	39
6.5.3. ESTIMACIÓN DE LA VARIANZA	40
6.5.4. BANDAS DE VARIABILIDAD	40
6.5.5. COMPARACIÓN DE CURVAS	41
6.5.6. VALIDACIÓN DE SUPUESTOS	41
7. RESULTADOS	43
7.1. ANÁLISIS EXPLORATORIO DE DATOS	43
7.2. CURVAS DE SUAVIZACIÓN	46
7.3. VALIDACIÓN DE SUPUESTOS	50
8. PROPUESTA EPIDEMIOLÓGICA	52
9. CONCLUSIONES Y RECOMENDACIONES	54
BIBLIOGRAFÍA	56
ANEXOS	60

LISTA DE TABLAS

1.	Parámetros de calidad de aire medidos en las estaciones fijas	21
2.	Estadísticas descriptivas según día de la semana	60

LISTA DE FIGURAS

1.	Estaciones fijas de monitoreo de la calidad de aire en Santiago de Cali	22
2.	Concentración promedio (24 Horas) de PM10 por estación de monitoreo	43
3.	Concentración promedio (horaria) de PM10 por zona de monitoreo	45
4.	Curvas de suavización y bandas de variabilidad por zona de monitoreo	47
5.	Curvas de suavización y bandas de variabilidad según clima en la estación CDAV (2003 - 2004)	49
6.	Curvas de suavización y bandas de referencia para la estación CDAV	50
7.	Concentración promedio (horaria) de PM10 en la estación CDAV	60
8.	Concentración promedio (horaria) de PM10 en la estación CLL15	61
9.	Concentración promedio (horaria) de PM10 en la estación ERA	61
10.	Concentración promedio (horaria) de PM10 en la estación UV	62
11.	Concentración promedio (horaria) de PM10 en la estación BA	62
12.	Concentración promedio (horaria) de PM10 en la estación CVC	63
13.	Concentración promedio (horaria) de PM10 en la estación PPD	64
14.	Curvas de suavización y bandas de variabilidad para la estación CDAV	65
15.	Curvas de suavización y bandas de variabilidad para la estación CLL15	65
16.	Curvas de suavización y bandas de variabilidad para la estación ERA	66
17.	Curvas de suavización y bandas de variabilidad para la estación CVC	67
18.	Curvas de suavización y bandas de variabilidad para la estación UV	68
19.	Curvas de suavización y bandas de variabilidad para la estación BA	68
20.	Curvas de suavización y bandas de variabilidad para la estación PPD	69
21.	Curvas de suavización y bandas de variabilidad según clima en la estación CDAV (2005 - 2006)	70
22.	Residuales de las curvas de las concentraciones para la estación CDAV (2003) . . .	71
23.	Residuales de las curvas de las concentraciones para la estación CDAV (2004) . . .	71
24.	Residuales de las curvas de las concentraciones para la estación CDAV (2005) . . .	72
25.	Residuales de las curvas de las concentraciones para la estación CDAV (2006) . . .	72
26.	Residuales de las curvas de las concentraciones para la estación CLL15 (2003) . . .	73
27.	Residuales de las curvas de las concentraciones para la estación CLL15 (2004) . . .	73
28.	Residuales de las curvas de las concentraciones para la estación CLL15 (2005) . . .	74
29.	Residuales de las curvas de las concentraciones para la estación CLL15 (2006) . . .	74
30.	Residuales de las curvas de las concentraciones para la estación ERA (2003)	75
31.	Residuales de las curvas de las concentraciones para la estación ERA (2004)	75
32.	Residuales de las curvas de las concentraciones para la estación ERA (2005)	76
33.	Residuales de las curvas de las concentraciones para la estación ERA (2006)	76
34.	Residuales de las curvas de las concentraciones para la estación UV (2004)	77

35.	Residuales de las curvas de las concentraciones para la estación UV (2005)	77
36.	Residuales de las curvas de las concentraciones para la estación CVC (2003)	78
37.	Residuales de las curvas de las concentraciones para la estación CVC (2004)	78
38.	Residuales de las curvas de las concentraciones para la estación CVC (2005)	79
39.	Residuales de las curvas de las concentraciones para la estación BA (2003)	79
40.	Residuales de las curvas de las concentraciones para la estación BA (2004)	80
41.	Residuales de las curvas de las concentraciones para la estación PPD (2003)	80
42.	Residuales de las curvas de las concentraciones para la estación PPD (2004)	81
43.	Residuales de las curvas de las concentraciones para la estación PPD (2005)	81
44.	Residuales de las curvas de las concentraciones para la estación PPD (2006)	82

RESUMEN

Este trabajo se ocupa de la contaminación atmosférica de la ciudad de Santiago de Cali durante el período 2003-2006, la cual fue declarada por el Ministerio del Medio Ambiente como la segunda ciudad más contaminada de Colombia, después de Bogotá (CONPES 3344, 2005). Se modela el comportamiento de la concentración promedio horaria del material particulado con diámetro inferior a $10\ \mu m$ durante 24 horas por estación de monitoreo, tipo de día (Ordinarios y Domingos y Festivos) y año, mediante el Estimador Lineal Local propuesto por Cleveland (1979), identificando las horas del día en las cuales se presentan los niveles más altos del contaminante e incorporando la variable precipitación para la zona Norte de la ciudad. También se caracterizan las zonas incluidas en el estudio, según nivel de concentración del contaminante, las cuales muestran un comportamiento característico de acuerdo al tipo de día. Se propone el uso de bandas de variabilidad a partir del estimador de varianza de Gasser, Sroka y Jennen-Steinmetz (1986).

Palabras clave: Estimador de varianza, Material particulado (PM10), Precipitación, Regresión no paramétrica, Suavización.

ABSTRACT

This paper deals the pollution atmospheric of the city Santiago de Cali during the period 2003 - 2006. This city was declared by Environment Ministry as the second most polluted city, later of Bogotá city (CONPES 3344, 2005). It fitted a model the hourly average concentration of particulate matter with diameter less than $10\ \mu m$ (PM10) during 24 hours, per monitoring station, type of day (Ordinary day and sunday and holyday) and per year through a nonparametric smoothing method called Local Linear Estimator propose by Cleveland (1979). The model identifies the most polluted hours for PM10 and it analyses high levels of pollution according to precipitation variable in the CDAV station. Also, it characterizes different zones included in the study according to PM10 pollution level. Every zone shows a different behavior of PM10 concentration per type of day. Additionally, it proposes to use variability bands through of variance estimator of Gasser, Seifert and Wolf (1986).

Key words: Variance estimator, Particulate matter (MP10), Precipitation, Nonparametric Regression, Smoothing.

INTRODUCCIÓN

La contaminación del aire es un problema que afecta cada día a toda la población, ya que durante los últimos años se han generado grandes e importantes cambios atmosféricos debido a las intervenciones del ser humano en los ciclos biológicos del planeta. Dichas intervenciones han deteriorado el ambiente, ocasionando así graves daños a la capa de ozono, a la fauna, a la flora y principalmente a la salud humana. El municipio de Santiago de Cali, al ser una de las ciudades industriales más importantes de Colombia se ha visto afectada por la contaminación del aire, incurriendo en deterioro ambiental siendo muchos los factores que inciden en ello, tales como el humo de las fabricas, el alto número de vehículos que transitan por la ciudad y las emisiones de gases de los mismos (Granada et al., 2008).

Uno de los principales agentes contaminantes del aire son las denominadas partículas PM10, partículas que tienen un tamaño inferior a 10 micrómetros y que están formadas por compuestos inorgánicos y materiales pesados, entre otros. Tienen diferentes orígenes y variedad de efectos en la salud de los seres humanos, ya que debido a su facilidad de atravesar las vías respiratorias se consideran causantes de múltiples enfermedades que afectan la salud humana, tales como el asma y en general enfermedades pulmonares, con una prevalencia para la ciudad de Cali del 20.6 % y 25 % respectivamente¹.

Por esta razón es importante conocer y visualizar el comportamiento de las partículas suspendidas en el aire que respiran los habitantes de la ciudad de Cali, además de observar su evolución con el pasar de los años; esto con el fin de identificar las zonas más contaminadas, para así aportar elementos de juicio que permitan implementar planes de control y vigilancia para reducir los niveles de contaminación presentes.

Es por lo anterior, que particularmente en este trabajo se analizará el problema de la contaminación por dichas partículas para los años 2003 a 2006; para ello se hará uso de los registros de PM10 suministrados por la Red de Monitoreo de Calidad del Aire de Cali (RMCA) del Departamento Administrativo de Gestión del Medio Ambiente (DAGMA), la cual está conformada por una unidad móvil y ocho estaciones de monitoreo ubicadas en diferentes zonas de la ciudad. Dichas estaciones están dotadas de equipos de alto, bajo y medio volumen, los cuales son los tipos de muestreadores más comúnmente utilizados y provistas de instrumentación meteorológica para el análisis de temperatura, temperatura interna, humedad relativa, entre otros (Granada et al., 2007).

La modelación de las concentraciones de PM10 se realizará mediante curvas típicas de regresión no paramétrica, ya que tienen como ventaja el no hacer supuestos sobre la forma funcional del conjunto de datos, sino que permite la estimación de la posible forma de la curva de regresión aplicada a dichos datos, convirtiéndola de esta manera en un método flexible, además de servir como herramienta exploratoria. Para el desarrollo de este trabajo se contará con el apoyo del grupo de Investigación en Estadística Aplicada (INFERIR) de la Universidad del Valle.

¹Revista de Neumología. Prevalencia de asma y otras enfermedades alérgicas en Colombia.

1. PLANTEAMIENTO DEL PROBLEMA

El aire es una mezcla de gases, fundamental e indispensable para todos los seres del planeta Tierra, ya que, por ejemplo en el caso de personas y animales, éstos mediante la respiración, toman oxígeno del aire para poder vivir. Durante los últimos años se ha observado que la contaminación del aire se ha incrementado y se ha propagado por todo el mundo, esto debido a la polución de los vehículos automotores que transitan por las ciudades y también a las grandes industrias que operan día tras día para la creación y fabricación de nuevos productos. Una atmósfera contaminada puede dañar la salud de las personas y afectar el desarrollo de plantas y animales, además del deterioro ambiental, los cambios y fenómenos climáticos que se puedan presentar.

Entre los contaminantes del aire se encuentran las denominadas partículas suspendidas, las cuales se consideran como una amenaza para la salud humana, pues las personas constantemente se encuentran expuestas a la inhalación de este tipo de partículas, ya que sus fuentes de variación hacen parte del día a día de la sociedad.

Según el Ministerio del Medio Ambiente, Cali es una de las ciudades más contaminadas de Colombia, el aire que se respira en esta ciudad es catalogado como nocivo debido a que cada día se arrojan a la atmósfera aproximadamente 340 toneladas de material particulado². Hay zonas donde el nivel de contaminación por partículas PM10 resulta ser más alta dependiendo del flujo vehicular que exista en el momento, de las empresas productoras que se encuentren alrededor, así como de otros factores influyentes como el año, mes, día y hora del día.

Con el fin de identificar las zonas más contaminadas y así utilizar los resultados obtenidos a partir de la estadística como herramienta para la toma de decisiones en cuanto a la factibilidad de llevar a cabo campañas para que los ciudadanos tomen conciencia del problema de contaminación y así se establezcan políticas que busquen disminuir en gran parte la contaminación por material particulado, se plantean las siguientes preguntas de investigación:

¿Cómo se comporta la concentración promedio diaria de partículas suspendidas en el aire (PM10) en la ciudad de Cali durante el período 2003-2006?

¿En qué zonas de la ciudad, épocas del año, días de la semana y horas de cada día se presenta la mayor concentración promedio de PM10 en el aire que respiran los habitantes de la ciudad de Cali?

²Documento CONPES 3344 de 2005

2. ANTECEDENTES

Son numerosos los estudios que se han realizado sobre la contaminación del aire, ya sea por la concentración de bióxido de carbono presente o por residuos de materiales tales como el plomo, pero son muy pocos los estudios que se han realizado sobre la contaminación por partículas suspendidas en el aire PM10. Es por ello, que inicialmente se hará una presentación de los antecedentes que modelan este contaminante desde diversas técnicas estadísticas y seguidamente se presentarán diferentes aplicaciones de la regresión no paramétrica en la modelación de diversos casos para los contaminantes del aire.

2.1. MODELACIÓN DE PM10 BAJO DIFERENTES TÉCNICAS ESTADÍSTICAS

- **Echeverri C. & Maya G. (2008)**, investigaron sobre la relación entre las partículas finas (PM2.5) y respirables (PM10) en la ciudad de Medellín. En este estudio realizaron mediciones simultáneas de PM2.5 y PM10 durante el período febrero a octubre de 2007 en diferentes zonas de la ciudad. A través de modelos de regresión simple se evaluó la correlación por zona entre la variable respuesta nivel de contaminación por PM2.5 y la variable predictora que mide la contaminación por PM10. De igual forma mediante el cálculo de la relación promedio (PM2.5/PM10) se evaluó qué zonas cumplen con la norma anual establecida para PM10 y PM2.5, llegando a la conclusión que probablemente se cumple con los niveles promedio establecidos para el PM10 pero no para las partículas finas (PM2.5), por lo que se concluye que son el principal problema de contaminación en la ciudad de Medellín.
- **Machado A., Velázquez H., García N., García C., Acosta L., Córdova A., & Linares M^a. (2007)**, realizaron un estudio sobre los metales en PM10 y su dispersión en una zona de alto tráfico vehicular, en la ciudad de Maracaibo, estado de Zulia, Venezuela. El objetivo era evaluar las concentraciones de ocho metales (Plomo, Aluminio, Níquel, Zinc, Cobre, Cromo, Vanadio y Cadmio) presentes en el material particulado con el fin de evaluar la dispersión atmosférica de éstos. Haciendo uso de análisis descriptivos, del coeficiente de correlación lineal, de pruebas de medias normales con varianza y medias desconocidas *t student*, llegaron a la conclusión de que en la zona estudiada se superaban los niveles permisibles por la OMS sobre la concentración máxima de PM10 y de algunos de los metales que las componen; también se encontró una buena dispersión atmosférica de los contaminantes, principalmente Plomo, Cobre y Cromo.
- **Sanhueza P., Vargas C. & Mellado P. (2006)**, realizaron una investigación sobre el impacto de la contaminación del aire por partículas PM10 en la mortalidad diaria en Temuco, Chile. El principal objetivo era estudiar los efectos de dichas partículas en la mortalidad de esta ciudad. A través de la metodología APHEA (Air Pollution Health European Approach), la cual incluye la estimación de modelos de regresión Poisson, el análisis de series temporales y el cálculo del riesgo relativo (RR) como medida de efecto, se obtuvo como resultado asociación positiva entre el PM10 y tres causas de mortalidad: respiratoria total, cardiovascular total y cardiorrespiratorias en la ciudad de Temuco.

- **Jaramillo M., Núñez M^a. E., Ocampo W., Pérez D. & Portilla G. (2004)**, evaluaron las emisiones causadas directamente por actividades humanas en los compuestos orgánicos volátiles, material particulado PM10, monóxido de carbono, óxidos de nitrógeno y óxidos de azufre utilizando los factores de emisión determinados con el modelo MOBILE 6.0 proporcionado por la EPA. En el estudio se consideraron las fuentes móviles, de área y puntuales, haciendo referencia a la emisión de los contaminantes a través de vehículos, industrias con alto y bajo volumen de contaminación en la zona Cali-Yumbo respectivamente. Finalmente se llegó a la conclusión de que la mayor emisión de PM10 se da por el sector industrial, la industria papelera y artes gráficas en la zona analizada.
- **Ostro et al. (1995)**, realizaron una investigación donde se utilizó información sobre los niveles de contaminación por PM10 y la mortalidad en Santiago de Chile, encontrándose como resultado mediante el enfoque de la función dosis-respuesta y la estimación por mínimos cuadrados ordinarios, evidencia robusta sobre el efecto de contaminación por partículas suspendidas PM10 y la tasa de mortalidad en esta ciudad.
- **Taffur & Quintero (1994)**, realizaron un análisis para comparar la incidencia sobre la morbilidad en el área de influencia de una fábrica de productos derivados de azufre, donde encontraron que la contaminación atmosférica en Cali incide de manera negativa sobre una gran proporción de los habitantes de la ciudad.

2.2. REGRESIÓN NO PARAMÉTRICA EN LA MODELACIÓN DE CONTAMINANTES

El grupo de Investigación en Estadística Aplicada INFERIR ha trabajado en años recientes en varios proyectos que construyen modelos de regresión no paramétrica para la estimación y el pronóstico de variables económicas, financieras y ambientales. Entre los proyectos más recientes se destacan los siguientes:

- **Castro, Flórez & Olaya (2009)**, modelaron mediante técnicas de regresión no paramétrica el comportamiento del promedio horario de las emisiones de monóxido de carbono (CO) en la zona centro de Cali. El objetivo era observar cuál había sido la evolución, identificando patrones y tendencias que permitieran determinar la curva que mejor se ajustaba a las observaciones, teniendo en cuenta el año y el día de la semana. Se llegó a la conclusión de que la función de regresión para el día sábado y los primeros cuatro días de la semana era circular dado que empezaban y terminaban en el mismo punto. Además se encontró que el mejor modelo para estimar las curvas típicas diarias para el monóxido de carbono es el Estimador Lineal Local, debido a que se comporta mejor en los extremos de la curva.
- **Reina, Pereira & Olaya(2009)**, estudiaron el comportamiento de la contaminación octo-horaria del monóxido de carbono (CO) en Santiago de Cali. En dicho estudio se llevó a cabo un análisis exploratorio de datos y se implementaron técnicas estadísticas más estructuradas como la estimación de curvas típicas mediante el Estimador Lineal Local que ayudaron a analizar el comportamiento del contaminante, llegando a la conclusión de que las concentra-

ciones octohorarias del CO muestran un comportamiento característico de acuerdo a la zona de monitoreo y al tipo de día.

- **Torres, Paz & Olaya (2009)**, modelaron el comportamiento octohorario del ozono troposférico durante las 24 horas del día, en la ciudad de Santiago de Cali, durante el período 2003-2006. Los registros de la contaminación de O_3 en la ciudad provienen de la Red de Monitoreo de Calidad del Aire (RMCA), a cargo del DAGMA. En este estudio se implementó un modelo de regresión no paramétrico univariado, considerando como variable dependiente la concentración 8-h de O_3 y como variable independiente la hora del día. Los resultados obtenidos en este trabajo evidencian que el comportamiento típico de la concentración octohoraria del ozono de Santiago de Cali, se reflejó en una curva con forma sinusoidal. Los valores iniciales son bajos e incrementan desde las 10 de la mañana hasta las 5 de la tarde, hora en la cual descienden los niveles de concentración hasta finalizar el día.
- **Ayala, Ojeda & Olaya (2007)**, estimaron curvas típicas mediante la técnica de suavización del Estimador de Regresión Lineal Local para estudiar la evolución de los niveles típicos de contaminación a través de los años 2003 a 2005. Mediante un análisis de covarianza no paramétrico con el fin de identificar patrones y tendencias compararon los comportamientos típicos diarios de tres contaminantes: monóxido de carbono (CO), dióxido de nitrógeno (NO_2) y ozono (O_3), observando que existe un patrón diario típico del comportamiento de los contaminantes y que al parecer se presenta una reducción de los niveles a lo largo de los años.
- **Barrientos, González & Olaya (2007)**, implementaron modelos de regresión no paramétrica para la estimación de curvas típicas usando suavización *spline* con el fin de modelar la demanda diaria de energía en una región del suroccidente colombiano. Se concluyó que el usar modelos spline en la estimación de la demanda hace que sólo se necesite un modelo para realizar los pronósticos de un mes completo, por ello la calidad de los pronósticos obtenidos a partir de los modelos spline basados en el mes anterior indican que para realizar predicciones de la demanda de energía eléctrica del Valle del Cauca en el corto plazo es suficiente contar con información reciente.
- **Montoya, Morales & Olaya (2005)**. El principal objetivo de este trabajo fue estimar las curvas típicas diarias de los contaminantes monóxido de carbono (CO), dióxido de nitrógeno (NO_2) y dióxido de azufre (SO_2) utilizando los datos recolectados por las estaciones de monitoreo de la ciudad de Santiago de Cali. La técnica estadística utilizada para la modelación de los datos fue la regresión no paramétrica, específicamente el suavizador LOESS, el cual permite predecir el nivel de contaminación a cualquier hora del día. Los resultados muestran que los más altos niveles de contaminación se presentaron en la estación de la Calle 15, en el centro de la ciudad, producida por la emisión de gases vehiculares.

3. OBJETIVOS

3.1. OBJETIVO GENERAL

Caracterizar la concentración promedio diaria de las partículas suspendidas en el aire (PM10), en diferentes zonas de la ciudad de Cali durante el período 2003 - 2006.

3.2. OBJETIVOS ESPECÍFICOS

- Proponer una curva típica del comportamiento diario de los niveles de contaminación en el aire por PM10 en las zonas de monitoreo de la ciudad de Cali, que permita estimar los niveles a cualquier hora del día.
- Evaluar cómo se ve afectada la concentración promedio de partículas suspendidas en el aire por las zonas en donde son tomadas las mediciones.
- Analizar posibles cambios de la concentración promedio de las partículas suspendidas en el aire según la variación del clima.
- Identificar las zonas de la ciudad de Cali donde se presentan los niveles de mayor concentración de PM10.
- Producir una propuesta analítica para el estudio de este tipo de contaminación, que pueda incorporarse en estudios epidemiológicos.

4. JUSTIFICACIÓN

La contaminación del aire es un problema que cobra importancia cada día, debido principalmente a sus efectos sobre la salud en los seres vivos, desde irritaciones leves hasta el desarrollo de graves enfermedades. La ciudad de Santiago de Cali, fue declarada por el Ministerio del Medio Ambiente como la segunda ciudad más contaminada de Colombia por PM10, después de Bogotá D.C, al ser una de las ciudades más industrializadas³.

Diversos estudios se han realizado desde el ámbito nacional e internacional donde se ha encontrado asociación entre la contaminación por material particulado (PM10) y los efectos en la salud humana, tales como el daño que estas partículas causan al ADN⁴. En Colombia, el contaminante monitoreado de mayor interés, dados sus demostrados efectos nocivos sobre la salud humana es el Material Particulado (PST y PM10)⁵, el potencial de daño depende del tamaño de la partícula, pues en general, diámetros menores a $10 \mu\text{g}/\text{m}^3$ afectan la salud y las partículas con diámetro menor a $2.5 \mu\text{g}/\text{m}^3$ pueden llegar a ser alojadas en los alvéolos pulmonares.

Estudios aceptados internacionalmente sobre los efectos en la mortalidad por la exposición a largo plazo de Material Particulado han sido realizados para PM2.5, pero en la actualidad este parámetro no es monitoreado ni exigido por la legislación Colombiana. Sin embargo, en el Valle del Cauca, también se han realizado estudios sobre partículas suspendidas en diferentes ciudades llegando a la conclusión de que en Yumbo, la mayor emisión de PM10 se da por el sector de la industria⁶ y por la emisión de gases vehiculares en la calle 15 del centro de Cali⁷.

De acuerdo a los documentos CONPES 3343 de 2005 y 3550 de 2008, el número de muertes anuales por causas originadas en la calidad del aire en Colombia es de 6000, con un costo de 1.5 billones de pesos anuales, que representan el 0.8 del PIB (Producto Interno Bruto), esto medido en términos de mortalidad prematura por problemas cardiopulmonares. La ciudad de Santiago de Cali podría participar con más de 1000 muertes anuales de las sugeridas por los estudios que presentan los documentos CONPES 3344 y 3550, aunque en Cali no existen estudios epidemiológicos al respecto, los datos estadísticos de estudios de CEDETES-2005 de la Universidad del Valle, inducen a esta cifra.

La cantidad de material particulado que Cali emite por fuentes móviles es de aproximadamente 340 toneladas/día, constituyéndose en una de las ciudades más contaminadas, superando frecuentemente los estándares ambientales para las concentraciones de este contaminante estipulados en la regulación vigente⁸. El aire de esta ciudad es monitoreado por la Red de Monitoreo de Calidad

³Documento CONPES 3344, Tabla 1. Emisiones estimadas de contaminantes a partir de los consumos de combustible para el año 2002

⁴Comprueban que las partículas suspendidas en el aire dañan ADN. "La crónica de hoy ". México. Marzo 21 de 2009.

⁵Lozano, N. (2003). Air pollution in Bogotá Colombia: A concentration Response Approach. Department of Agricultural and Resource Economics. University of Maryland at College Park.

⁶Jaramillo M. et al. (2003)

⁷Montoya M., Morales A., Olaya J. (2005)

⁸Red de Monitoreo de la Calidad de Aire de Bogotá. Reporte Anual (2002).

del Aire (RMCA) constituida por ocho estaciones fijas y una unidad móvil. Sin embargo, durante el período analizado se presentaron dificultades durante su tiempo de operación, especialmente por la discontinuidad en el monitoreo y en el procesamiento de datos.

Una de las etapas críticas que se relaciona con el proceso de monitoreo de contaminantes atmosféricos es la del análisis de los datos obtenidos, es aquí donde la Estadística juega un papel fundamental mediante el análisis de grandes volúmenes de datos, contribuyendo a la obtención de resultados precisos y confiables para la toma de decisiones. Para ello la estadística ofrece dos alternativas de análisis: la primera es la modelación bajo el enfoque paramétrico, la cual tiene limitantes en cuanto a su aplicación, ya que presupone una serie de supuestos sobre los datos y un conocimiento a priori de la función de regresión, provocando errores en la estimación cuando estos supuestos no se cumplen. La segunda, se refiere al enfoque no paramétrico donde no se presupone un comportamiento del fenómeno a priori, sino que se concibe la forma de la curva a partir de los datos, lo que permite un amplio rango de acción con respecto a los mismos y por, tanto es una buena herramienta para describir su comportamiento gracias a la flexibilidad que ofrece.

Por estas razones, se sigue la línea de regresión no paramétrica en este trabajo a través del Estimador Lineal Local para la estimación de curvas típicas diarias de material particulado PM₁₀, con el fin de mejorar las estimaciones entre los años 2003 a 2006; se hace uso de esta alternativa ya que representa un método objetivo, bastante confiable gracias a su desarrollo teórico y práctico. Además se agrega la componente de comparación de curvas, la cual puede conducir a identificar tendencias en el comportamiento del contaminante de acuerdo a la variación del clima.

5. MARCO TEÓRICO

A continuación se presenta los principales elementos del clima, la definición, composición y características de la atmósfera, así como los tipos de contaminantes que se producen en ella y además se definen las técnicas de suavización más comunes en el marco del análisis de regresión no paramétrica, consideradas en la modelación del material particulado PM10.

5.1. ELEMENTOS DEL CLIMA

El clima es el resultado de numerosos factores que actúan conjuntamente, las montañas y los mares, influyen decisivamente en sus características. Dentro de los principales elementos que determinan las características del clima cabe mencionar: la presión atmosférica, temperatura, humedad, y precipitación.

- **Presión Atmosférica:** La presión atmosférica es el peso de la masa de aire por cada unidad de superficie⁹. Ésta no es siempre la misma en un punto determinado, sino que sufre variaciones, dependiendo de diversos factores, entre ellos la temperatura y la humedad.
- **Temperatura:** La temperatura atmosférica es el indicador de la cantidad de energía acumulada en el aire. Aunque existen tres escalas, la temperatura del aire se suele medir en grados centígrados ($^{\circ}C$). La temperatura depende de diversos factores, tales como la dirección y fuerza del viento, la latitud, la altura sobre el nivel del mar, entre otros.
- **Humedad:** La humedad indica la cantidad de vapor de agua presente en el aire, la cual puede variar con los cambios de temperatura. La humedad relativa se expresa en términos porcentuales de agua en el aire, mientras que la humedad absoluta se refiere a la cantidad de vapor de agua presente en una unidad de volumen de aire y se expresa en gramos por centímetro cúbico (gr/cm^3)¹⁰.
- **Precipitación:** Las precipitaciones atmosféricas son las modalidades del agua en estado líquido o sólido provenientes del vapor de agua atmosférico y que como consecuencia del proceso de condensación, tocan la superficie terrestre¹¹.

5.2. LA ATMÓSFERA

La atmósfera es la capa gaseosa que envuelve la tierra y el elemento necesario para que la vida sea posible. La atmósfera terrestre sirve como escudo contra la radiación solar, estabilizador de la temperatura de la superficie y provee la mezcla exacta de gases que permite la respiración. Está compuesta de gases y partículas sólidas y líquidas en suspensión atraídas por la gravedad terrestre. En ella se producen todos los fenómenos climáticos y meteorológicos que afectan de una u otra forma al planeta; regula la entrada y salida de energía de la tierra y es el principal medio de transferencia de calor.

⁹En <http://www.astromia.com/tierraluna/elemclima.htm>

¹⁰En <http://www.tecnun.es/asignaturas/Ecologia.htm>

¹¹En <http://www.educaplus.org/climatic/>

La atmósfera se encuentra compuesta básicamente por un 78.08 % de nitrógeno, 20.95 % de oxígeno y 0.93 % de argón; éstos suponen casi el 100 % del volumen del aire en la atmósfera hasta una altura de 80 Km sobre el nivel del mar. Los restantes componentes del aire están presentes en cantidades muy pequeñas y se expresan en partes por millón (ppm) o microgramos por metro cúbico ($\mu g/m^3$).

5.3. CONTAMINACIÓN ATMOSFÉRICA

Se entiende por contaminación atmosférica la presencia en el aire de sustancias y formas de energía que alteran la calidad del mismo, de modo que implique riesgos, daños o molestias graves para las personas y bienes de cualquier naturaleza. Todas las actividades humanas y los fenómenos culturales que se producen en la superficie o en el interior de la tierra van acompañados de emisiones de gases, vapores, polvos y aerosoles; éstos al difundirse a la atmósfera, se integran en los distintos ciclos biogeoquímicos que se desarrollan en la Tierra.

El proceso de contaminación en el aire inicia con la emisión de de contaminantes hacia la atmósfera a través de diversas fuentes de contaminación. Estas fuentes se clasifican como fijas y móviles.

5.3.1. FUENTES FIJAS O ESTACIONARIAS

Fuente de emisión situada en un lugar determinado e inamovible, aún cuando la descarga de contaminantes se produzca en forma dispersa. Por lo general, este tipo de fuentes hace referencia a las plantas de energía, industrias químicas, fábricas, imprentas, lavanderías y chimeneas residenciales que utilizan madera, entre otras. A su vez estas fuentes se subdividen en:

- **Puntuales:** Hace referencia a las emisiones provenientes de una sola chimenea o conducto.
- **De área:** En este tipo de fuente varios focos de emisión de varias fuentes fijas se dispersan en un área determinada.

5.3.2. FUENTES MÓVILES

Son las fuentes de emisión que, por razón de su uso o propósito, es susceptible de desplazarse, como los automotores o vehículos de transporte a motor de cualquier naturaleza. El automóvil es la principal fuente móvil de contaminación en el aire¹².

5.4. CONTAMINANTES ATMOSFÉRICOS

Entre los contaminantes atmosféricos más perjudiciales para la salud se encuentran las partículas suspendidas totales y el material particulado.

¹²Granada et al. (2007).

5.4.1. PARTÍCULAS SUSPENDIDAS TOTALES (PST)

Entre los diversos tipos de contaminantes que se transportan en el aire se encuentran las PST. Las partículas suspendidas en el aire tienen diferentes orígenes y variedad de efectos en la salud de los seres humanos al ser inhaladas mediante el proceso de respiración y represadas en el sistema respiratorio. Las PST se dividen en dos rangos de tamaños: PM2.5 y PM10.

Las partículas PM10 son consideradas como las más gruesas por el tamaño que presentan, son producto del “polvo u hollín” de las obras de construcción, el polvo levantado por los vehículos en las vías (principalmente las no pavimentadas), operaciones de pulido, entre otras fuentes. Mientras que el particulado fino menor a $2.5 \mu m$ (PM2.5), y que es parte del PM10 proviene del escape de los gases de combustión de los vehículos, humo del tabaco, smog fotoquímico, entre otros. Algunas de estas partículas son especialmente tóxicas para los humanos y, en la práctica, los principales riesgos para la salud humana por la contaminación del aire provienen de este tipo de polución, contribuyendo, por ejemplo, a aumentar las enfermedades respiratorias como la bronquitis.

5.4.2. MATERIAL PARTICULADO (PM10)

El Material Particulado (PM10) se define como la materia sólida o líquida dispersa en el aire, cuyo diámetro es inferior o igual a 10 micras, éste puede proceder de fuentes naturales como el viento, incendios forestales, erupciones volcánicas, la industria de la construcción, entre otras. Dependiendo de su tamaño pueden permanecer en suspensión en la atmósfera desde unos segundos a varios meses. Las partículas PM10 son respirables y se denominan contaminante criterio¹³ debido a la capacidad que tiene de alcanzar la zona inferior de los pulmones. La concentración de partículas PM10 es determinada dividiendo la masa particulada para el volumen del aire muestreado. La unidad de medida de este contaminante es $\mu g/m^3$.

5.5. MONITOREO ATMOSFÉRICO

El monitoreo atmosférico de calidad del aire, se refiere al conjunto de actividades diseñadas con el objetivo de medir, analizar y procesar en forma continua la información sobre las concentraciones de contaminantes y variables meteorológicas presentes en la atmósfera, en un lugar y tiempo específico. La importancia del monitoreo radica en que es base fundamental para la formulación de programas enmarcados dentro de la gestión de calidad del aire, que buscan principalmente evaluar la calidad de éste a través de la medición de los contaminantes presentes en la atmósfera.

Se conoce como red de monitoreo al conjunto de estaciones de muestreo, generalmente fijas y que operan de forma continua, que se establecen para medir los parámetros ambientales necesarios para cumplir con los objetivos fijados concernientes a la calidad del aire y que cubren toda la extensión de un área determinada.

¹³DAGMA, 2005.

5.5.1. MONITOREO DE LA CALIDAD DEL AIRE EN SANTIAGO DE CALI

El municipio de Santiago de Cali está ubicado en el sur occidente Colombiano, es una de las principales ciudades del país con más de dos millones de habitantes y cuenta con un parque automotor constituido por cerca de 530000 vehículos¹⁴.

El DAGMA fue creado por el acuerdo 18 del 30 de diciembre de 1994 y es la entidad encargada de velar por el cuidado de la ciudad. Esta entidad vigila el cumplimiento de las normas ambientales vigentes y procura el bienestar de la ciudadanía bajo el principio del desarrollo sostenible¹⁵.

En el año 1998, el DAGMA adquirió una Red de Monitoreo de Calidad del Aire (RMCA) que inició su operación a mediados del año 1999. Esta red consta de ocho estaciones fijas de monitoreo y una unidad móvil, ubicadas en diferentes puntos de la ciudad de acuerdo al estudio realizado por la Universidad Nacional de Colombia, Sede Medellín, en el cual se clasificaron las estaciones fijas en dos tipos:

- **Estaciones de tipo general:** Tienen como finalidad medir la calidad del aire en áreas típicas de la ciudad de Santiago de Cali.
- **Estaciones de tipo local:** Tienen como objetivo evaluar la calidad del aire en zonas que son de interés para la ciudadanía o que presentan altos niveles de contaminación.

Las estaciones de monitoreo que componen la RMCA, están dotadas de instrumentación para el análisis de los siguientes contaminantes atmosféricos: material particulado (PM_{10}), óxidos de azufre (SO_2), óxidos de nitrógeno (NO_x), ozono (O_3), monóxido de carbono (CO) e hidrocarburos. En la Tabla 1 se muestran los parámetros de calidad de aire que son medidos en cada una de las estaciones de monitoreo.

Tabla 1: Parámetros de calidad de aire medidos en las estaciones fijas

Estación	Abreviatura	PM_{10}	NO_x	SO_2	CO	O_3
Centro Diagnóstico Automotor del Valle	CDAV	x	x	x		x
Base Aérea Marco Fidel Suárez	BA	x	x	x	x	
Hospital Universitario del Valle	HUV	x		x	x	
Polideportivo el Diamante	PPD	x	x	x	x	x
CVC Pance	CVC	x				x
Escuela República de Argentina	ERA	x	x	x	x	
Universidad del Valle	UV	x		x		x
Calle 15	CLL15	x	x		x	

Fuente: DAGMA, 2005.

¹⁴Secretaría de Tránsito y Transporte Municipal. Centro de Diagnóstico Automotor del Valle Ltda.

¹⁵Tomado de “Propuesta de estimación no paramétrica de la contaminación horaria por monóxido de carbono en el centro de la ciudad de Cali, Colombia, entre 2003 y 2006”. Castro (2009).

Las estaciones fijas de monitoreo fueron ubicadas en los sitios que representaban interés para el monitoreo del estado de la calidad del aire, en zonas de la ciudad de Cali influenciadas por las emisiones provenientes tanto del sector industrial como del vehicular¹⁶. Los sitios de ubicación de las estaciones son (Figura 1):

- **Zona Centro:** Estaciones Calle 15 y Escuela República de Argentina
- **Zona Norte:** Estación Centro de Diagnóstico Automotor del Valle
- **Zona Sur:** Estación Universidad del Valle, CVC Pance
- **Zona Occidente:** Hospital Universitario del Valle
- **Zona Oriente:** Polideportivo El Diamante, Base Aérea

Figura 1: Estaciones fijas de monitoreo de la calidad de aire en Santiago de Cali



Fuente: DAGMA, 2005.

La red fue puesta en marcha en el año 1999 y tuvo una operación consecutiva hasta el año 2001; sin embargo, durante los años posteriores el sistema funcionó en forma interrumpida, obteniendo como últimos datos los correspondientes al año 2005 y mediados del 2006.

La RMCA se encontró fuera de funcionamiento por aproximadamente cuatro años, lo cual generó deterioro en el sistema y daños en algunos equipos debido a las largas interrupciones en su operación. La falta de funcionamiento del sistema no sólo ha ocasionado deterioro físico de las estaciones y sus componentes, sino que también, ha imposibilitado a la autoridad ambiental en

¹⁶DAGMA, 2005

el proceso de recolección de información suficiente, actualizada y veraz para tomar decisiones que permitan formular un Programa de Gestión Integral para la Prevención y Control de la Contaminación Atmosférica en Santiago de Cali (DAGMA, 2005). La RMCA reinició operaciones a partir del año 2010 sólo en tres estaciones de monitoreo (Estaciones ERA y Éxito de la Flora, Unidad Móvil), mientras el sistema se recupera paulatinamente.

5.6. NORMAS DE CALIDAD DE AIRE

En el contexto del marco legal de la contaminación de Colombia, la entidad encargada de regular y controlar los diferentes tipos de contaminantes que se encuentran dispersos en el aire es el Ministerio de Ambiente, Vivienda y Desarrollo Territorial. Este ente de control ambiental se encarga de expedir las resoluciones que enmarcan los máximos niveles permisibles para los contaminantes perjudiciales para la salud humana. En el caso del material particulado, la Resolución 601 contiene los valores de concentración promedio máximos permisibles en un período de 24 horas y anual. En este caso de aplicación se tomará como referencia la Resolución 601 expedida en el año 2006 ya que es una norma contemporánea con el período de análisis del presente estudio (2003-2006).

La Resolución 601 de 2006 estipula que para el material particulado (PM10), la máxima concentración promedio del contaminante para un período de 24 horas es equivalente a $150 \mu g/m^3$, y para un año se consideran permisibles hasta $70 \mu g/m^3$. Excedencias de los valores límites permisibles del material particulado pueden llegar a ser considerados altamente nocivos para la salud humana durante largos períodos de exposición.

La concentración promedio del material particulado (PM10) se determina a través de la media aritmética, según lo define el Ministerio de Medio Ambiente, Vivienda y Desarrollo Territorial en el Protocolo para el Monitoreo y Seguimiento de la Calidad del Aire. Las normas ambientales nacionales, toman como referencia los estándares estipulados por la Agencia de Protección Ambiental de Norte América (US EPA).

5.7. TEMPORALIDAD EN EL ANÁLISIS DE PM10

En este estudio se analizará la concentración promedio de PM10 teniendo en cuenta los registros tomados por hora y en un período de 24 horas (día), por ello se definirán los períodos de tiempo de acuerdo con los lineamientos de la Organización Meteorológica Mundial.

- **Hora:** Se define una “hora” como el período de sesenta minutos transcurridos “inmediatamente antes” de la hora reportada, es decir, los datos de concentraciones reportados a las 7 a.m. son los recolectados desde las 6:01 a.m. hasta las 7:00 a.m.
- **Día:** Se define un “día” como el período de 24 horas transcurrido entre las 00:01 y las 24:00, donde 00:01 es el primer minuto del día, después de la media noche. Por lo tanto en un día se obtendrán 24 promedios horarios desde la hora 0 que es la primera hora del día hasta la hora 23 que es la última.

- **Semana:** Se define una “semana” como el período que inicia en el minuto 00:01 del día lunes y termina en el minuto 24:00 del domingo.
- **Mes:** Se define un “mes” como el período de tiempo que se comprende en ocasiones por 28, 29, 30 o 31 días calendario.
- **Año:** Se define un “año” como el conjunto de 12 meses calendario.

5.8. MODELO DE REGRESIÓN NO PARAMÉTRICA

Las técnicas de regresión no paramétricas empiezan a usarse desde 1940, pero es a partir de 1980 que su uso empieza a expandirse; Härdle (1990), Fan y Gijbels (1996), Eubank (1999), entre otros. En regresión no paramétrica, al igual que en regresión paramétrica, la idea principal es estimar una curva de regresión poblacional a partir de las observaciones del fenómeno en estudio.

Los métodos de regresión no paramétrica tienen como fin estimar la forma de la función a través de técnicas de suavización, sin necesidad de especificar a priori su estructura, es decir, no se asume un comportamiento a priori de la distribución de los datos, por el contrario lo que se busca es *dejar que los datos hablen* (Olaya, 2002). La forma general de una función o curva de regresión es la siguiente (Bowman & Azzalini, 1997):

$$y = \mu(t) + \varepsilon \quad (1)$$

donde y es una variable continua dependiente, $\mu(t)$ es una forma funcional desconocida que explica el comportamiento de los datos en función del tiempo t y ε son los errores del modelo con media cero y una varianza constante ($\sigma^2 > 0$).

5.9. SUAVIZACIÓN

Se define suavización como el procedimiento para estimar la función de regresión μ , y suavizador a la herramienta que resume el comportamiento de la variable dependiente en función de una o más variables independientes. Eubank (1999) especifica que en regresión no paramétrica, se debe asumir que μ es suave, es decir, que si se desea estimar la función μ en un punto t , se espera que las observaciones y_i asociadas a los t_i cercanos a t , posean información de μ en t . Lo cual indica, que es posible promediar de alguna forma las respuestas y_i más cercanas al punto donde se estime $\mu(t)$.

Los suavizadores más comunes suelen tener una estructura lineal en la variable dependiente, es decir, se expresa como una función lineal de la variable respuesta. A ellos se les llama estimadores lineales de la función de regresión y se expresan de la siguiente forma:

$$\hat{\mu}(t) = \sum_{i=1}^n W(t, t_i; h) y_i \quad (2)$$

Donde $W(t, t_i; h)$ representa los pesos que dependen de la variable independiente y de un parámetro h no negativo que es llamado ancho de banda o parámetro de suavización. Si los pesos $W(t, t_i; h)$;

$i = 1, 2, \dots, n$ son tales que todos son positivos y que $\sum_{i=1}^n W(t, t_i; h) = 1$, entonces el suavizador (2) es un promedio ponderado de las respuestas y_i .

El ancho de banda o parámetro de suavización h delimita la zona en que las observaciones aportarán información de μ a un punto de interés t , entonces, las observaciones cercanas al punto de interés tendrán un peso alto y a medida que las observaciones se alejan de dicho punto, el peso asignado decrecerá hasta llegar a un peso cero en las observaciones que se encuentran fuera de la zona delimitada. Se espera, que un buen suavizador posea las siguientes características (Kafadar y Horn, 2002):

- Debe recrear μ tan precisamente como sea posible
- Debe capturar perfectamente tendencias lineales o superficies
- Debe poder manejar datos que no estén igualmente espaciados
- Su salida debe ser “suave”(excepto en puntos de quiebre bruscos)
- Outliers extremos deben ser correctamente ignorados y sobresalir claramente en los residuales

5.10. FUNCIONES KERNEL

Los suavizadores de una curva de regresión hacen uso de una función Kernel, la cual se encarga de ponderar las observaciones de la variable respuesta paralela a los valores de la variable de diseño que están lejanas o cercanas al punto de interés. Las funciones de pesos Kernel son funciones de densidad de probabilidad que poseen tres características importantes: es simétrica alrededor de cero, su pico ocurre en cero y decae comúnmente a cero a medida que se aleja del punto de estimación t . Entre las funciones Kernel más conocidas cabe mencionar la Gaussiana, la Triangular, la Uniforme y la Biponderada.

En el caso de la función Kernel Gaussiana, los pesos se asignan de acuerdo con una densidad normal estándar, la cual tiene soporte en $(-\infty, \infty)$; esta función de pesos es muy conocida y diferenciable en todo el recorrido, además el ancho de banda desempeña el papel de desviación, es decir, las observaciones sobre el rango de h en el eje de la covariable podrían contribuir a la estimación de la función, mientras que las observaciones fuera del rango se les asignará una ponderación cero.

Para poder ser usada, la función Kernel debe cumplir ciertas condiciones (Eubank, 1999). Dichas condiciones son:

- | | |
|------------------------------|--|
| 1. $\int_{-1}^1 K(u)du = 1$ | 3. $M_2 = \int_{-1}^1 U^2 K(u)du \neq 0$ |
| 2. $\int_{-1}^1 uK(u)du = 0$ | 4. $V = \int_{-1}^1 K(u)^2 du < \infty$ |

La primera condición se refiere a que la suma de los pesos debe ser igual a uno, la segunda es una condición de simetría alrededor de cero, la tercera se refiere a la varianza de la función Kernel y la cuarta a su curtosis.

5.11. ESTIMADORES EN REGRESIÓN NO PARAMÉTRICA

Los estimadores Kernel, son estimadores lineales en el sentido en que se puede expresar el valor del estimador en algún punto de diseño como una suma ponderada de la respuesta. A continuación se definen algunos tipos de estimadores usados en regresión no paramétrica:

5.11.1. ESTIMADOR NADARAYA-WATSON

La primera solución a la obtención de estimaciones no paramétricas fue presentada en 1964 por Nadaraya & Seckler y Watson, quienes introdujeron el estimador Nadaraya-Watson, donde t y y son variables aleatorias. Este estimador es construido a partir de un ajuste por mínimos cuadrados ponderados en un polinio de grado $d = 0$.

$$\min_{\alpha} \sum_{j=1}^n W(t_j - z_i; h) (y_j - \alpha)^2 \quad (3)$$

de esta manera se tiene la siguiente expresión, la cual es el cociente de dos funciones Kernel:

$$\hat{\mu}(t) = \frac{\sum_{i=1}^n K\left(\frac{t_j - z_j}{h}\right) y_i}{\sum_{i=1}^n K\left(\frac{t_j - z_j}{h}\right)} \quad (4)$$

donde $\hat{\mu}(t)$ es una estimación consistente de la curva de regresión $\mu(t)$ siempre y cuando $h \rightarrow 0$ y $nh \rightarrow \infty$.

5.11.2. ESTIMADOR GASSER-MÜLLER

Gasser y Müller (1979), propusieron una forma alternativa de un estimador Kernel, donde la cantidad de pesos asignada a cada observación localmente está determinada por los espacios entre el diseño de los puntos. El estimador está definido por:

$$\hat{m}_{GM}(t) = \sum_{i=1}^n y_{[i]} \int_{t_{i-1}}^{t_i} \frac{1}{h} W\left(\frac{t - y}{h}\right) dt \quad (5)$$

donde $t_0 = -\infty$, $t_1 = \frac{1}{2}(t_{[i]} + t_{[i+1]})$, $t_n = \infty$, $x_{[i]}$ denota el i -ésimo valor de la covariable observada y $y_{[i]}$ denota el correspondiente valor de la respuesta.

5.11.3. ESTIMADOR LINEAL LOCAL

Un estimador común de regresión no paramétrica el cual usa diferencias de vecinos más cercanos es el estimador de regresión local propuesto por Cleveland (1979), el cual recoge la solución a un problema mediante mínimos cuadrados ponderados de un polinomio $d = 1$ (Bowman y Azzalini, 1997).

$$\min_{\alpha, \beta} \sum_{i=1}^n (y_i - \alpha - \beta(t_i - t))^2 W(t, t_i; h) \quad (6)$$

donde $\hat{\alpha}$ es la estimación en t ya que define la posición de la línea de regresión local en el punto t , $W(t, t_i; h)$ es una función de pesos, que tiene como objetivo asegurar que las observaciones cercanas a t tengan el mayor peso en la determinación de las estimaciones. Aquí los pesos actúan para realizar un ajuste lineal localmente, donde la localidad está compuesta por las observaciones que posean pesos mayores a cero (Bowman y Azzalini, 1997).

El estimador lineal local se define matemáticamente como:

$$\hat{f}(t) = \frac{1}{n} \sum_{i=1}^n \frac{[S_2(t; h) - S_1(t; h)(t - t_i)]W(t_i - t; h)}{S_2(t; h)S_0(t; h) - S_1(t; h)^2} y_i \quad (7)$$

donde:

$$S_r(t; h) = \frac{1}{n} \left(\sum_{i=1}^n (t_i - t)^r W(t_i - t; h) \right) \quad (8)$$

Cuando el parámetro de suavización h es grande, las ponderaciones ligadas a cada observación mediante la función Kernel, se acercan cada vez más y por tanto, la curva estimada se aproxima a una línea de regresión por el método de mínimos cuadrados.

Una de las ventajas que tiene el estimador lineal local es que la componente del sesgo no depende del patrón de los puntos de diseño, por lo que es útil a la hora de realizar comparación de curvas de regresión. La desventaja que presenta es que cuando el ancho de banda es pequeño y se tiene una región con pocas observaciones, el estimador puede dar resultados indefinidos.

5.11.4. ESTIMADOR LOWESS

El estimador LOWESS “Regresión Robusta Localmente Ponderada” propuesto por Cleveland (1979) usa anchos de banda variables a partir del ajuste de una regresión polinomial ponderada de grado d .

$$\hat{\mu} = \min_{\beta_j} \sum_{j=1}^n W(t_j - z_r, h_r(t_j)) (y_j - \beta_0 - \beta_1(t_i - z_r) - \dots - \beta_d(t_i - z_r)^d)^2 \quad (9)$$

donde W depende de la covariable t y del parámetro de suavización $h_r(t_j)$.

La suavización está regulada por el span, el cual define la cantidad de vecinos más cercanos al punto que se usa para la estimación. Para encontrar el span es necesario calcular la máxima distancia de todas las distancias asociadas a los vecinos más cercanos del punto de interés. Este estimador se vuelve robusto mediante un ajuste de pesos asociados a los residuales de las estimaciones de los puntos de interés.

5.12. INFERENCIA EN EL MODELO NO PARAMÉTRICO

En el enfoque no paramétrico, se utilizan algunos criterios que permiten calificar un estimador como “bueno”. En ese sentido, para obtener un buen estimador se define el riesgo como una herramienta

que permite medir la bondad de los estimadores de la función $\hat{\mu}$, Olaya (2002).

$$R(\hat{\mu}) = E[\mu(t) - \hat{\mu}(t)]^2 \quad (10)$$

Lo deseable en este caso es obtener un estimador $\hat{\mu}$ de μ que tenga un riesgo tan bajo como sea posible.

Otra característica evaluada en un estimador de la función $\hat{\mu}$, es la varianza; en el enfoque no paramétrico anchos de banda pequeños hacen que los estimadores lineales tengan varianzas altas junto con sesgos pequeños, y viceversa. Este hecho impacta en alto grado las estimaciones debido a que anchos de banda pequeños producirán estimaciones muy variables, mientras que anchos de banda grandes producirán estimaciones más suaves pero muy sesgadas, Olaya (2002).

Para una covariable uni-dimensional las expresiones asintóticas son las siguientes (Bowman y Azalini, 1997):

$$E[\hat{\mu}(t)] \approx \mu(t) + \frac{h^2}{2} \sigma_w^2 \mu''(t) \quad (11)$$

$$var[\hat{\mu}(t)] \approx \frac{\sigma^2}{nh} \frac{\alpha(w)}{f(t)} \quad (12)$$

en donde σ_w^2 denota $\int_{-\infty}^{\infty} z^2 w(z) dz$ y $\alpha(w)$ denota $\int_{-\infty}^{\infty} w(z)^2 dz$. A fin de considerar expresiones asintóticas de este tipo, es necesario hacer algún supuesto acerca de cómo cambia el patrón de los puntos de diseño t_i cuando el tamaño de muestra se incrementa. Esto es expresado en la función de densidad $f(t)$. Cuando los valores de t son escogidos aleatoriamente de alguna distribución, entonces f representa la correspondiente función de densidad de probabilidad. Más aún, f puede caracterizar la densidad local de los puntos de diseño cuando el tamaño de muestra se incrementa.

Por lo tanto, para el Estimador Lineal Local estas expresiones asintóticas muestran que el sesgo aumenta con grandes parámetros de suavización y con altos grados de curvatura (μ''). También muestran que la varianza se incrementa con pequeños parámetros de suavización y en valores de t donde los puntos de diseño vecinos son escasos, dado que el factor $nhf(t)$ puede ser visto como controlador del tamaño de muestra local. Por lo tanto, se requiere un balance apropiado entre el sesgo y la varianza como una función del parámetro de suavización.

5.12.1. SELECCIÓN DEL ANCHO DE BANDA

En regresión no paramétrica el parámetro de suavización se conoce comúnmente como el ancho de banda, el cual es estimado a partir de los datos e indica el tamaño de la vecindad en la cual se realizan estimaciones sobre la curva de regresión ajustada. Su selección es de gran importancia ya que éste determina la suavidad de la curva, por ello el parámetro de suavización adecuado será aquel que logre un buen equilibrio entre el sesgo y la varianza del estimador a través de una

medida global llamada Error Cuadrático Medio Integrado (MISE) definido como:

$$MISE(h) = E \int \{\hat{\mu}(t) - \mu(t)\}^2 f(t) dt \quad (13)$$

Donde $\hat{\mu}(t)$ es la estimación de la función que explica el comportamiento de los datos, $\mu(t)$ es la verdadera función desconocida de los datos y $f(t)$ representa la densidad observada de los puntos de diseño. La minimización del MISE produce como resultado un ancho de banda asintóticamente óptimo.

Dentro de los criterios de selección del ancho de banda cabe mencionar los siguientes:

■ **CRITERIO DE VALIDACIÓN CRUZADA**

El criterio de validación cruzada es un estimador casi insesgado del MISE y es un método que no requiere de la estimación de la varianza (Eubank, 1999). La idea básica del método es dejar por fuera una pareja de observaciones (t_i, y_i) y construir un modelo sin esa observación y usar el modelo para predecir la observación que ha quedado por fuera. La función de validación está definida por:

$$CV(h) = n^{-1} \sum_{i=1}^n [y_i - \hat{\mu}_{-i}(t_i)]^2 \quad (14)$$

Donde $\hat{\mu}_{-i}(t_i)$ denota una estimación suavizada cuando una observación (t_i, y_i) ha sido omitida del conjunto de datos. Se selecciona el ancho de banda que minimice esta medida.

■ **MÉTODOS PLUG-IN**

Este tipo de métodos se basan principalmente en minimizar medidas del Error Cuadrático Medio (ECM) de la curva de regresión. Las técnicas Plug-In son bastante complejas, pues es necesario estimar la segunda derivada de la curva de regresión, además de requerir que σ^2 sea conocida o estimada.

5.12.2. ESTIMACIÓN DE LA VARIANZA

Debido a que los estimadores en regresión no paramétrica son sesgados, varios autores (Rice, 1984; Gasser, Sroka & Jennen-Steinmetz, 1986; entre otros) han propuesto estimadores de varianza basados en diferencias sucesivas.

- **Rice (1984)** propone utilizar la idea de diferenciación de primer orden para remover tendencia, estimando σ^2 de la siguiente manera:

$$\hat{\sigma}_R^2 = \frac{1}{2(n-1)} \sum_{i=2}^n (y_i - y_{i-1})^2 \quad (15)$$

El estimador de Rice puede presentar algunos problemas debido a que la diferencia $y_i - y_{i-1}$ puede ser influenciada por las fluctuaciones que pueda presentar la función de regresión μ y por lo tanto la estimación de σ^2 puede inflarse.

- **Gasser, Sroka y Jenner-Steinmetz (1986)** proponen una diferenciación de segundo orden removiendo efectos de tendencia local a partir de interpolación lineal:

$$\hat{\sigma}_{GSJ}^2 = \frac{1}{n-2} \sum_{i=2}^n c_i^2 \hat{\varepsilon}_i^2 \quad (16)$$

donde $\hat{\varepsilon}_i$ son los pseudo-residuales definidos como:

$$\hat{\varepsilon}_i = \left(\frac{t_{i+1} - t_i}{t_{i+1} - t_{i-1}} \right) y_{i-1} + \left(\frac{t_i - t_{i-1}}{t_{i+1} - t_{i-1}} \right) y_{i+1} - y_i = a_i y_{i-1} + b_i y_{i+1} - y_i \quad (17)$$

y $c_i^2 = (a_i^2 + b_i^2 + 1)^{-1}$ son los coeficientes escogidos de tal forma que $E(c_i^2 \hat{\varepsilon}_i^2) = \sigma^2$ para todo i donde μ sea lineal.

Si el diseño es equidistante, el estimador será:

$$\sigma_{GSJ}^2 = \frac{2}{3(n-2)} \sum_{i=2}^{n-1} \left(\frac{1}{2} y_{i-1} - y_i + \frac{1}{2} y_{i+1} \right)^2 \quad (18)$$

Gasser, Sroka y Jennen-Steinmetz (1986), en el artículo de publicación del estimador, hacen mención a los siguientes supuestos, los cuales son necesarios para obtener resultados asintóticos:

1. No existen medidad múltiples en algún punto de diseño, esto es $a = t_i < t_2 < \dots < t_n = b$; $a = 0$, $b = 1$ sin pérdida de generalidad
2. Se requiere que $\max |t_i - t_{i-1}| = O(1/n)$
3. Los errores ε_i son independientes e idénticamente distribuidos con $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$, $E(\varepsilon_i^4) < \infty$
4. La función μ es continua

- **Gasser, Seifert y Wolf (1993)**, propusieron una estimación no paramétrica de varianza residual que considera el efecto de la varianza entre y dentro de las muestras en k puntos de diseño (estimador mixto) de la siguiente forma:

Sean y_{ij} las observaciones para $j = 1, \dots, n_i \geq 1$ en diferentes puntos de diseño $t_1 < \dots < t_n$. Sea $\bar{y}_i$ la media de las observaciones por cada punto de diseño y s^2 el estimador insesgado de σ^2 en el análisis de varianza convencional. Para errores normales, s^2 es independiente y distribuida de algún estimador no paramétrico de la varianza $\hat{\sigma}^2$ en $\bar{y}_i$. Los pseudo-residuales respecto a $\bar{y}_i$ pueden ser calculados de manera convencional, siempre y cuando la condición $\sum_{j=1}^r d_{ij}^2 = 0$ sea reemplazada por $\sum_{j=1}^r \frac{d_{ij}^2}{n_{i+j}} = \frac{1}{n-r}$.

El estimador mixto propuesto por Gasser, Seifert y Wolf (1993) combina a s^2 y al estimador de Gasser $\hat{\sigma}_{GSJ}^2$ utilizando los promedios $\bar{y}_i$ en cada punto de diseño. Este estimador se basa en un diseño equidistante, con tamaño de muestra variable en cada punto de diseño y ajuste

mediante un estimador lineal local, el cual se encuentra dado por la siguiente expresión:

$$\hat{\sigma}_{MIX}^2 = \frac{\left(\frac{35}{18(k-2)} - \frac{1}{(k-2)^2}\right) * s^2 + \frac{1}{\sum_{i=1}^k n_i - k} * \hat{\sigma}_{GJS}}{\frac{1}{\sum_{i=1}^k n_i - k} + \frac{35}{18(k-2)} - \frac{1}{(k-2)^2}} \quad (19)$$

$$s^2 = \frac{1}{\sum_{i=1}^k n_i - k} \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \quad (20)$$

$$\hat{\sigma}_{GJS}^2 = \frac{1}{k - r} \sum_{i=1}^{k-r} \hat{\varepsilon}_i^2 \quad (21)$$

donde k es el número de puntos de diseño y $n_1 = \dots = n_k$ es el número de observaciones en cada uno de ellos. Los coeficientes d_{ij} de orden r en el estimador de Gasser et al., 1986 (20) definen los pseudo-residuales como:

$$\varepsilon_i = \sum_{j=1}^r d_{ij} \bar{y}_{i+j} \quad (22)$$

donde $i = 1, 2, \dots, k - r$.

5.12.3. BANDAS DE VARIABILIDAD

En los modelos de regresión paramétrica, los estimadores son insesgados, lo cual permite una construcción fácil de intervalos de confianza a partir de una cantidad pivotal. Sin embargo, en regresión no paramétrica, el sesgo que involucran los estimadores de la curva de regresión no permite la construcción de dichos intervalos, debido a que el sesgo involucra la segunda derivada de la curva de regresión, la cual es compleja de estimar.

Es por ello que en regresión no paramétrica se hace uso de las bandas de variabilidad, las cuales indican la variabilidad alrededor de la curva de regresión estimada, pero no puede ser usada para afirmar conclusiones acerca de la forma de la curva en regiones particulares, debido a la presencia del sesgo y a la naturaleza punto a punto de las bandas.

Bowman & Azzalini (1997) presentan una construcción de bandas de variabilidad a partir de dos errores estándar por encima y por debajo de la curva estimada $\hat{\mu}(t)$.

$$se[\hat{\mu}(x)] = \sqrt{var[\hat{\mu}(x)]} = \sqrt{diag(S * S^T) \hat{\sigma}^2} \quad (23)$$

Donde S es una matriz $n_i \times n_i$ de pesos y $\hat{\sigma}^2$ es la estimación de varianza de los errores a partir del estimador no paramétrico.

5.12.4. COMPARACIÓN DE CURVAS

El análisis de covarianza es una técnica muy útil para comparar los valores de una variable respuesta a través de varios grupos en la presencia del efecto de una covariable. El modelo se puede escribir de la siguiente forma general:

$$y_{ij} = \mu_i(t_{ij}) + \varepsilon_{ij} \quad (24)$$

Para $i = 1, 2, \dots, p$ y $j = 1, 2, \dots, n_i$, donde p es el número de grupos y n_i es el número de observaciones en el grupo i . El interés parte en identificar si las funciones de regresión son diferentes a través de los grupos.

Considerando el caso simple al probar la igualdad de dos curvas suaves, es decir, identificar si las curvas de regresión μ_1 y μ_2 son diferentes, se tiene que las hipótesis son:

$$\begin{aligned} H_0 : \mu_1(t) &= \mu_2(t) \\ H_1 : \mu_1(t) &\neq \mu_2(t) \end{aligned}$$

Cuando las curvas fundamentales son idénticas o simplemente constantes, los sesgos tienen la misma forma. Por lo tanto, el sesgo podría ser idéntico si se usa un parámetro de suavización común (Speckman, 1988).

Un estadístico de prueba para comparar curvas de regresión es:

$$\frac{\sum_{i=1}^p \sum_{j=1}^{n_i} \{\hat{\mu}_i(t_{ij}) - \hat{\mu}(t_{ij})\}^2}{\hat{\sigma}^2} \quad (25)$$

donde $\hat{\mu}_i$ denota la estimación de regresión del grupo i , $\hat{\mu}$ denota una estimación de regresión construida con todo el conjunto de datos, asumiendo que todos los grupos tienen idénticas funciones de regresión y $\hat{\sigma}^2$ denota una estimación de la varianza general del error σ^2 . Aquí las estimaciones de cada curva se contrastan con una estimación común bajo el supuesto de igualdad. Esta comparación se realiza en todos los puntos t_{ij} del espacio de diseño. A fin de remover los efectos de escala, una estimación de la varianza del error aparece en el denominador (Dette, 2001). Esta forma del estadístico brinda una manera de comparar un conjunto de curvas de regresión.

La distribución de este estadístico de prueba puede examinarse haciendo uso del hecho de que el numerador y el denominador son formas cuadráticas. En notación matricial el vector de valores estimados $\hat{\mu}_i$ para cada estimación de regresión individual puede ser escrito como $\hat{\mu}_i = S_i Y_i$, donde S_i es una matriz $n_i \times n_i$ de pesos y Y_i denota el vector de respuestas para el grupo i -ésimo. La colección completa de valores ajustados puede ser recogida de la forma $\hat{\mu}_i = S_d Y$, donde el subíndice d indica que la suavización es realizada bajo el modelo que asume que las curvas de regresión son diferentes. De igual manera, los valores ajustados bajo el modelo que asume curvas de regresión idénticas, pueden ser calculados al unir los datos y representarlos como $\hat{\mu} = S_s Y$ para una matriz de pesos conocidos S_s de tamaño $n \times n$. Por lo tanto, el numerador del estadístico de prueba es $Y^T (S_d - S_s)^T (S_d - S_s) Y$, el cual es una forma cuadrática. Por simplicidad, se puede

representar la matriz $Y^T(S_d - S_s)^T(S_d - S_s)Y$ como Q .

Un estimador aproximado, puede ser expresado como $\hat{\sigma}^2 = Y^T B Y$, la cual es una forma cuadrática en los datos para una matriz B de constantes. Finalmente el estadístico de prueba queda expresado de la siguiente forma (Dette, 2001):

$$F = \frac{Y^T Q Y}{Y^T B Y} \quad (26)$$

Desafortunadamente, los resultados estándar de modelos lineales no se aplican en este caso, porque las matrices Q y B no tienen las propiedades necesarias, tal como ser definidas positivas. Sin embargo, existen resultados (Jhonson & Kotz, 1970) que pueden ser utilizados para manejar estadísticos de este tipo bajo condiciones más generales. Estos resultados requieren solamente que las matrices de las formas cuadráticas sean simétricas, lo que corresponde a este caso.

Como un primer paso es útil reformular el problema y enfocarlo sobre la significancia del estadístico F según lo expresado en su valor p . Esto es:

$$p = P(F > F_{obs}) \quad (27)$$

donde F_{obs} denota el valor calculado de los datos observados. Esto puede ser re-escrito como:

$$p = P\left(\frac{Y^T Q Y}{Y^T B Y} > F_{obs}\right) = P(Y^T C Y > 0) \quad (28)$$

donde

$$C = (Q - F_{obs} B) \quad (29)$$

Jhonson & Kotz (1970) muestran resultados generales acerca de la distribución de una forma cuadrática en variables normales, tales como $Y^T C Y$ para alguna matriz simétrica C . Estos resultados se aplican con mayor facilidad cuando las variables aleatorias normales tienen media cero.

Cuando Y es reemplazado por $\mu + \varepsilon$ varios términos en la expansión del numerador de F desaparecen, al menos asintóticamente, como un resultado de las propiedades del sesgo del estimador lineal local, ya que los valores esperados para $\hat{\mu}_i(t)$ y $\hat{\mu}(t)$ son aproximadamente iguales. Así, la forma cuadrática $Y^T Q Y$ es equivalente a tener $\varepsilon^T Q \varepsilon$.

El valor esperado para el estimador de la varianza del error es $E(\hat{\sigma}^2) = E(\varepsilon^T B \varepsilon) + \mu^T B \mu$, donde B es una matriz simétrica $n \times n$. El último término es la suma de diferencias sucesivas al cuadrado de la verdadera función μ . Este término es relativamente pequeño comparado con el primer término en ε y así es fácil mostrar que, ignorando esta cantidad, se tiene el efecto de construcción de una prueba más conservadora. En otras palabras, el efecto del sesgo es inflar la estimación de σ^2 , lo cual conlleva a la subestimación de las diferencias entre las curvas, y aun procedimiento de prueba más cauteloso.

De acuerdo a estos resultados, la expresión del valor p para la prueba sería:

$$p = P\left(\frac{\varepsilon^T Q \varepsilon}{\varepsilon^T B \varepsilon} > F_{obs}\right) = P(\varepsilon^T C \varepsilon > 0) \quad (30)$$

Por lo tanto, ahora se tiene una forma cuadrática en variables normales del tipo $z^T A z$ donde $E(z) = 0$ y A es una matriz simétrica de tamaño $n \times n$. Los resultados de Johnson y Kotz (1970), permiten entonces calcular exactamente la probabilidad p , aunque de forma numérica en lugar de una forma analítica.

Las bandas de referencia pueden ser construidas a partir del error estándar (se) de la distribución de estimación, el cual se puede calcular como:

$$se = se[\hat{\mu}_1(t) - \hat{\mu}_2(t)] = \sqrt{se_1(t)^2 + se_2(t)^2} \quad (31)$$

donde $se_1(t)$ y $se_2(t)$ denotan los errores estándar de las curvas individuales.

Bajo el supuesto que las curvas de regresión son idénticas, las medias de las dos estimaciones son también idénticas y así la diferencia entre las dos estimaciones tiene media 0. En algún punto t de interés, será elevada donde las estimaciones están a más de dos errores estándar. Una banda de referencia de amplitud de dos errores estándar, centrados en el promedio de dos curvas de estimación pueden ser superpuestas sobre un gráfico para ayudar a la comparación de las curvas (Bowman y Azzalini, 1997).

6. METODOLOGÍA

A continuación se presenta la metodología desarrollada para dar cumplimiento a los objetivos planteados en la investigación.

6.1. REVISIÓN DE LA LITERATURA

Como primera medida se realizó la búsqueda de información en diversas fuentes de internet, concerniente a la problemática de la calidad del aire, considerando ciertos aspectos involucrados en la contaminación del mismo tales como las fuentes de producción, los efectos en la salud, las formas de medición y el control que se lleva a cabo a partir de la normatividad estipulada sobre contaminación ambiental tanto en el ámbito nacional como internacional. Posteriormente se recurrió a la revisión de libros y artículos tipo científicos en diversas bases de datos académicas sobre las diferentes técnicas de suavización empleadas en regresión no paramétrica para la estimación de curvas, teniendo en cuenta aspectos como la selección del ancho de banda, la estimación de la varianza, la construcción de bandas de variabilidad, entre otros de igual importancia. Los principales escritos consultados fueron los de Eubank (1999), Bowman & Azzalini (1997) y Cleveland (1979), los cuales proporcionaron elementos necesarios para el desarrollo de este proyecto de investigación.

6.2. RECOLECCIÓN DE LA INFORMACIÓN

Para este proyecto de investigación se dispone de información concerniente a los datos de contaminación por material particulado PM₁₀ suministrados por la Red de Monitoreo de Calidad de Aire (RMCA) del DAGMA entre los años 2003 y 2006. En el proceso de depuración de la base de datos se encontraron muchas inconsistencias, ocasionadas probablemente al reinicio de actividades de la red, después de haber estado detenida desde enero de 2001 hasta marzo de 2003.

Para el año 2003 se dispone de información para las estaciones CDAV (Mayo-Diciembre), CLL15 (Junio-Septiembre; Noviembre-Diciembre), ERA (Mayo-Diciembre), HUV (Abril-Julio), CVC (Mayo-Septiembre; Noviembre-Diciembre), UV (Noviembre-Diciembre), BA (Abril; Junio-Diciembre), PPD (Abril-Agosto; Noviembre). Debido a que para la estación Universidad del Valle (UV) se cuenta únicamente con información para dos meses, no se tendrá en cuenta en el análisis para este año, pues son pocos los registros de concentración que se tienen.

Para el año 2004 se cuenta con información para las estaciones CDAV (Enero-Diciembre), CLL15 (Enero-Agosto), ERA (Enero-Diciembre), CVC (Junio-Octubre), UV (Febrero-Diciembre), BA (Junio-Octubre; Diciembre), PPD (Junio-Diciembre).

Para el año 2005 solo se tuvo en cuenta los registros de los meses de Octubre a Diciembre en las estaciones CDAV, PPD, CLL15, CVC, ERA y UV, debido a que durante el período Enero a Septiembre se presentó una falla en la calibración de los instrumentos de medición, motivo por el cual no se tiene en cuenta la información de estos meses para el análisis.

Por último para el año 2006 se contó con información durante los meses de Enero a Abril en la estación PPD y durante el período Febrero a Marzo en las estaciones CDAV, CLL15 y UV. La red de monitoreo dejó de funcionar en el mes de abril de 2006.

Los registros de los niveles de concentración de material particulado PM10 se producen de manera discreta, uno cada 10 segundos y se reportan acumulados como datos horarios. Para cada hora se dispone de la medición mínima, máxima y promedio. Para efectos de análisis, el formato de la hora en la cual se capturó el registro de la concentración de PM10 fue modificado a formato numérico, de tal manera que la hora 0 representa el período transcurrido entre las 12:00 am y la 1:00 am, siendo la hora 23 la última hora del día (11:00 pm).

Los datos referentes a las condiciones climatológicas de la ciudad de Cali, fueron suministrados por el Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM) para la estación de monitoreo ubicada en el Aeropuerto Alfonso Bonilla Aragón durante el período 2003-2006. Se disponen de registros diarios de temperatura media, temperatura máxima y mínima, precipitación total, velocidad del viento, entre otros. Se cuenta con información completa para todas las variables en los meses de Enero a Diciembre para los años 2003 a 2006.

En el desarrollo de este trabajo, sólo se tendrá en cuenta para el análisis la variable precipitación, ya que el objetivo es determinar la influencia del clima (seco o lluvioso) en el comportamiento de la concentración de PM10. El término “*precipitación*” hace referencia a las formas de agua en estado líquido o sólido que caen directamente sobre la superficie terrestre, lo que incluye las lluvias, las lloviznas, granizo, nieve, etc¹⁷. El instrumento utilizado para medir la precipitación fue el pluviómetro, cuya instalación se realizó en lugares apropiados donde no se presentaran interferencias de edificaciones o árboles. La precipitación se mide en milímetros (*mm*).

6.3. DEPURACIÓN Y VALIDACIÓN DE DATOS

La validación de los datos es el proceso mediante el cual los datos son filtrados y aceptados bajo criterios específicos. La validación de datos involucra una serie de actividades encaminadas a la verificación del correcto funcionamiento del sistema de donde fue adquirida la información.

El proceso de validación realizado por el DAGMA que sigue los lineamientos de la agencia de protección ambiental de los Estados Unidos (EPA), define dos niveles de validación de datos: el nivel 1 se refiere a la validación operativa primaria y nivel 2 a la validación secundaria o estadística. El primer nivel consiste inicialmente en un seguimiento de todos los datos disponibles, identificando el estado que determina la calidad del dato recolectado y las alarmas (banderas), que indican una serie de eventos como fallas de potencia, instrumentos de baja calibración, actividades de mantenimiento, error de fecha/hora, dato perdido, entre otros. A continuación se relacionan las banderas y estados considerados en la validación de datos de los registros de PM10:

¹⁷Tomado de: <http://www.astromia.com/glosario/precipitacion.htm>

Estado: Esta variable determina la calidad del dato según las siguientes categorías:

- *Estado 0:* Normal
- *Estado 1:* Este estado no tiene significado establecido por el DAGMA
- *Estado 2:* Número de muestras pequeño
- *Estado 3:* Datos perturbados por calibración
- *Estado 4:* Hay canal de comunicación pero no hay analizador (En este caso todos los datos deben ser cero)
- *Estado 5:* Este estado no tiene significado establecido por el DAGMA
- *Estado 6:* Este estado no tiene significado establecido por el DAGMA
- *Estado 7:* Este estado no tiene significado establecido por el DAGMA
- *Estado 32:* Este estado no tiene significado establecido por el DAGMA
- *Estado 80:* Este estado no tiene significado establecido por el DAGMA
- *Estado 128:* El equipo ha sobrepasado el límite de la temperatura interna ($45^{\circ}C$)
- *Estado 160:* Este estado no tiene significado establecido por el DAGMA
- *Estado 1024:* Fallos del equipo o analizador

Bandera: Esta variable permite clasificar la muestra capturada como válida o no válida:

- *A:* Dato Atípico
- *C:* Calibración
- *E:* Dato que no concuerda con la distribución (horaria) del parámetro
- *F:* Falla o mantenimiento del analizador
- *K:* Dato faltante (No hubo monitoreo)
- *M:* Este estado no tiene significado establecido por el DAGMA
- *N:* Tamaño de muestra insuficiente (razón desconocida)
- *S:* Falla en el sistema de transmisión de datos
- *T:* Incremento en la temperatura interna
- *V:* Dato válido (revisión primaria)
- *Z:* Dato perturbado por calibración

Durante el proceso de validación de datos se eliminaron registros negativos e iguales a cero, ya que la red de monitoreo contempla un valor límite detectable positivo para la concentración de PM10, el cual corresponde a $0.05 \mu g/m^3$ ¹⁸. De la misma manera se eliminaron del análisis aquellas estaciones de monitoreo donde no se contaba con suficiente información de la concentración del contaminante, en este caso, no se tuvo en cuenta para el desarrollo del trabajo a la estación HUV, ya que únicamente dispone de información para cuatro meses del año 2003 (Abril a Julio), además la ubicación que ésta presentaba determinó un factor influyente en la decisión de eliminación, ya que se encontraba en la zona del parqueadero del Hospital Universitario del Valle lo que hacía se presentarán altos niveles de concentración por PM10 y por tanto los registros obtenidos no fueran confiables.

Se consideró como muestra válida aquellos registros que presenten un *Estado* = 0 y/o *Bandera* = V. Además, según las recomendaciones realizadas por el IDEAM a través del Protocolo del Monitoreo y Seguimiento del Módulo Aire del Sistema de Información Ambiental, para el desarrollo del análisis diario de PM10, se considerará válida la concentración si por lo menos el 75 % de los datos de contaminación para un período de 24 horas se encuentran disponibles, esto es, se tomaron en cuenta los días que tengan información para 18 horas o más.

Una vez aplicada la validación primaria, se procede a realizar la validación secundaria evaluando la consistencia de los datos mediante un análisis exploratorio de los mismos, el cual permite la descripción de importantes características como la dispersión y la identificación de valores que no se encuentran dentro de la distribución histórica del contaminante estudiado (puntos atípicos). La eliminación de dichos valores, se realizó en las estaciones CDAV, CLL15, ERA, PPD, CVC y UV.

De la base de datos de condiciones climatológicas proporcionada por el IDEAM, se consideró únicamente la variable precipitación (*pp*) la cual es medida en milímetros (*mm*), ya que el objetivo es identificar el comportamiento del contaminante en condiciones secas y lluviosas. Dado que la precipitación es cualquier cantidad de agua que cae desde el cielo a la superficie terrestre, se determinaron como días secos, aquellos días en los cuales la precipitación es igual a cero ($pp = 0 \text{ mm}$) y como días lluviosos, aquellos en los cuales se presenta algún nivel de precipitación ($pp > 0 \text{ mm}$).

6.4. HERRAMIENTA COMPUTACIONAL

La mayoría de los paquetes estadísticos disponibles en el mercado tienen altos costos de licencias, además de presentar limitaciones en cuanto a las herramientas estadísticas que ofrecen, entre ellas el ajuste modelos de regresión no paramétrica.

Para el desarrollo del proyecto se hizo uso de la versión 2.12.0 del software estadístico R para Windows, un software de libre distribución, el cual permite la elaboración de gráficas, la implementación de técnicas estadísticas complejas, además de la creación de funciones personalizadas mediante el lenguaje de programación.

¹⁸DAGMA, 2005

6.5. ANÁLISIS ESTADÍSTICO

La información analizada en este proyecto de investigación proviene de los promedios horarios de la concentración del material particulado PM10. Para abordar el análisis de los datos, se realizó en primera instancia un análisis descriptivo de la base de datos previamente validada; posteriormente se siguió con el proceso de estimación de las curvas típicas diarias, selección del parámetro de suavización, estimación de la varianza, construcción de bandas de variabilidad, comparación de las curvas de regresión y por último se realizó la validación del supuesto de homogeneidad de varianza.

6.5.1. ANÁLISIS EXPLORATORIO DE DATOS

Mediante métodos gráficos y estadísticos se exploró el comportamiento de la concentración promedio de material particulado PM10 en las zonas de monitoreo de la ciudad de Cali para los años 2003 a 2006. Para ello, se construyeron diagramas de cajas y alambres, además de gráficos de líneas con el fin de identificar si existe un comportamiento característico de la concentración promedio de PM10 durante los días de la semana.

Al realizar este análisis se encontró que las concentraciones promedio de PM10 en los días sábados presentaron un comportamiento similar al de los días ordinarios, mientras que los días domingos y festivos presentaban un comportamiento distinto al obtenido para los días Lunes a Sábado, en cuanto al nivel de concentración, reportando niveles inferiores de contaminación. Por esta razón, se realizó el análisis en todas las estaciones de monitoreo según el tipo de día, clasificándolos de la siguiente manera: Días Ordinarios (Lunes a Sábado) y Domingos y Festivos.

6.5.2. ESTIMACIÓN DE CURVAS TÍPICAS

Para la estimación de las curvas típicas diarias, se empleó la suavización Kernel debido al desconocimiento de la estructura estadística que modela el comportamiento horario del material particulado. Esta técnica estima una función de regresión a un conjunto de datos, donde no se parte de una función conocida sino de la forma funcional dada por los datos. Se utilizó como suavizador el Estimador Lineal Local propuesto por Cleveland (1979), el cual posee algunas propiedades de la regresión lineal convencional, por ser un caso particular de ella y por poseer un diseño adaptable, además de que la estimación de las curvas es mejor en los extremos. Para la asignación de pesos a los datos, se utilizó la función Kernel Gaussiana, la cual asigna los pesos de acuerdo a una función de densidad normal.

Las curvas típicas del comportamiento horario de PM10 se construyeron a partir del siguiente modelo:

$$y_i = \mu(t_i) + \varepsilon_i$$

donde y_i representa el nivel de concentración de material particulado de acuerdo a la clasificación de los días, t_i es la hora del día asociada al registro de las concentraciones horarias de PM10 ($0 \leq t \leq 23$) y $\mu(t_i)$ es la estimación de la curva típica diaria en la hora t_i .

En esta estimación de curvas típicas por regresión local, se considera un parámetro de suavización o ancho de banda (h) fijo al tener puntos de diseño (t_i) uniformemente densos (Bowman & Azzalini, 1997). El ancho de banda adecuado será aquel que logre un balance óptimo entre el sesgo y la varianza del estimador utilizado; para la selección de éste se utilizó el método de validación cruzada, que no requiere conocimiento previo de la varianza.

Para la suavización de las curvas típicas en el software estadístico *R*, se utilizó la función *sm.regression* de la librería *sm* desarrollada por Bowman y Azzalini (1997), la cual realiza estimaciones de densidades y curvas de regresión no paramétrica. Los resultados asociados a la función *sm.regression* muestran la estimación del valor óptimo del ancho de banda mediante el método de validación cruzada, el cual logra un balance óptimo entre el sesgo y la varianza.

6.5.3. ESTIMACIÓN DE LA VARIANZA

Para la estimación de la varianza se hace uso del estimador propuesto por Gasser, Sroka & Jennen-Steinmetz (1986), considerando los pseudo-residuales expresados en (17). Dette et al (1998), sugieren utilizar este estimador de segundo orden ($r = 2$) basado en diferencias ordinarias, ya que es considerado una buena elección en la estimación de la varianza en el caso de obtener el ajuste de una función de regresión periódica, es decir, una función de regresión con efectos de estacionalidad; por el contrario, no es recomendable utilizar el estimador de varianza de Rice (1984), debido a que no siempre se comporta bien en la estimación de varianza en curvas de regresión periódicas.

Para el cálculo del estimador de Gasser, fue necesario crear una variable auxiliar que considerara el orden de las observaciones horarias de PM10 de acuerdo a la fecha de obtención de las mediciones, esto con el fin de reducir el efecto de las posibles fluctuaciones de la función de regresión y considerar el hecho de poseer 24 puntos de diseño con múltiples medidas. Para ello, se organizaron las observaciones de acuerdo a cada estación de monitoreo, año y tipo de día (Ordinarios y Domingos y Festivos).

6.5.4. BANDAS DE VARIABILIDAD

Las bandas de variabilidad se construyeron a partir del error estándar de $\hat{\mu}(t)$ de la siguiente forma:

$$LI = \hat{\mu}(t) - 2 * (\sqrt{\text{var}[\hat{\mu}(t)]}) \quad (32)$$

$$LS = \hat{\mu}(t) + 2 * (\sqrt{\text{var}[\hat{\mu}(t)]}) \quad (33)$$

donde la banda *LS* indica dos errores estándar por encima y *LI* dos errores estándar por debajo de la estimación de la curva $\hat{\mu}(t)$. El error estándar de la curva se expresa en (23).

Las bandas de variabilidad asociadas a cada curva de regresión estimada, muestran la región en la cual se podría encontrar la verdadera curva, pero se limita toda clase de interpretaciones debido al sesgo involucrado en la estimación no paramétrica.

6.5.5. COMPARACIÓN DE CURVAS

Para la comparación de curvas típicas diarias se incluyó en el análisis la variable precipitación para la zona Norte de la ciudad, específicamente para la estación CDAV, debido a su cercanía con el Aeropuerto Alfonso Bonilla Aragón, del cual se tienen los registros climatológicos para los años 2003 a 2006. El objetivo de realizar la comparación de curvas, es observar si existe alguna diferencia en el comportamiento del material particulado PM10 para días secos y lluviosos dependiendo del tipo de día (Ordinarios y Domingos y Festivos). Dado que el análisis de las curvas típicas diarias mostró que la concentración de PM10 para domingos y festivos presenta un comportamiento similar tanto en días secos como lluviosos, el análisis de comparación se realizará únicamente para días ordinarios. Debido a las restricciones de información en cuanto a los datos climatológicos para la ciudad de Cali, el análisis para días secos y lluviosos sólo se realizará para la estación CDAV.

En el proceso de comparación de curvas típicas diarias se empleó la técnica de análisis de covarianza en el escenario no paramétrico, con la cual se pueden comparar los niveles de concentración de PM10 (y) según la condición climatológica (i) en la presencia del efecto de una covariable, en este caso las horas del día (t). Este análisis se utilizó debido a que se desarrolla a partir de la estimación de curvas de regresión no paramétricas con un estimador lineal local.

Teniendo en cuenta lo anterior, el modelo asociado a este análisis es el siguiente:

$$y_{ij} = \mu(t_{ij}) + \varepsilon_{ij} \text{ para } i = 1, 2 \text{ y } j = 1, 2, \dots, n_i$$

donde n_i es el número de observaciones en el año i . Aquí, la estimación de la curva de regresión μ se realiza a través del Estimador Lineal Local. Dado que el interés es identificar si las funciones de regresión μ_1 y μ_2 son diferentes, se formularon las siguientes hipótesis:

$$H_0 : \mu_1(t) = \mu_2(t)$$

$$H_1 : \mu_1(t) \neq \mu_2(t)$$

En este caso, la hipótesis nula indica que existe una única curva típica diaria de concentración para material particulado, es decir, que los niveles típicos de contaminación no se ven afectados según la condición del clima, mientras la hipótesis alterna propone que sí existen diferencias entre las curvas estimadas. Para contrastar esta hipótesis se realizó un análisis gráfico mediante las bandas de referencia, el cual considera que si alguna de las curvas estimadas queda por fuera de las bandas, entonces, los niveles de concentración de PM10 presentan diferencias significativas entre días secos y lluviosos, en caso contrario, cuando las curvas estimadas se encuentren dentro de las bandas de referencia, no se tienen diferencias significativas en cuanto a la contaminación por material particulado para días secos y lluviosos.

6.5.6. VALIDACIÓN DE SUPUESTOS

Un paso fundamental que sigue a la suavización no paramétrica es la validación de los supuestos que se hacen acerca de los errores en el modelo propuesto. Estos supuestos asumen que los errores

son independientes entre sí, con media cero y varianza constante.

Gasser, Sroka y Jennen-Steinmetz (1986) proponen un método alternativo para el cálculo de los errores del modelo, usando una forma de interpolación lineal, teniendo en cuenta la distancia que hay entre los puntos de diseño y las fluctuaciones en la función de regresión (pseudo-residuales).

Una vez obtenidos los pseudo-residuales, se procedió a verificar cada uno de los supuestos mediante herramientas graficas, entre ellas graficar los pseudo-residuales vs la hora en la que se realizó el monitoreo del material particulado, esto con el fin de determinar si se cumple el supuesto de homogeneidad de varianza; para ello se espera que los pseudo-residuales muestren una variabilidad constante.

Es necesario tener en cuenta que los estimadores en regresión no paramétrica no se ven afectados por la violación de alguno de los supuestos gracias a la robustez con la que fueron construidos (Olaya, 2002).

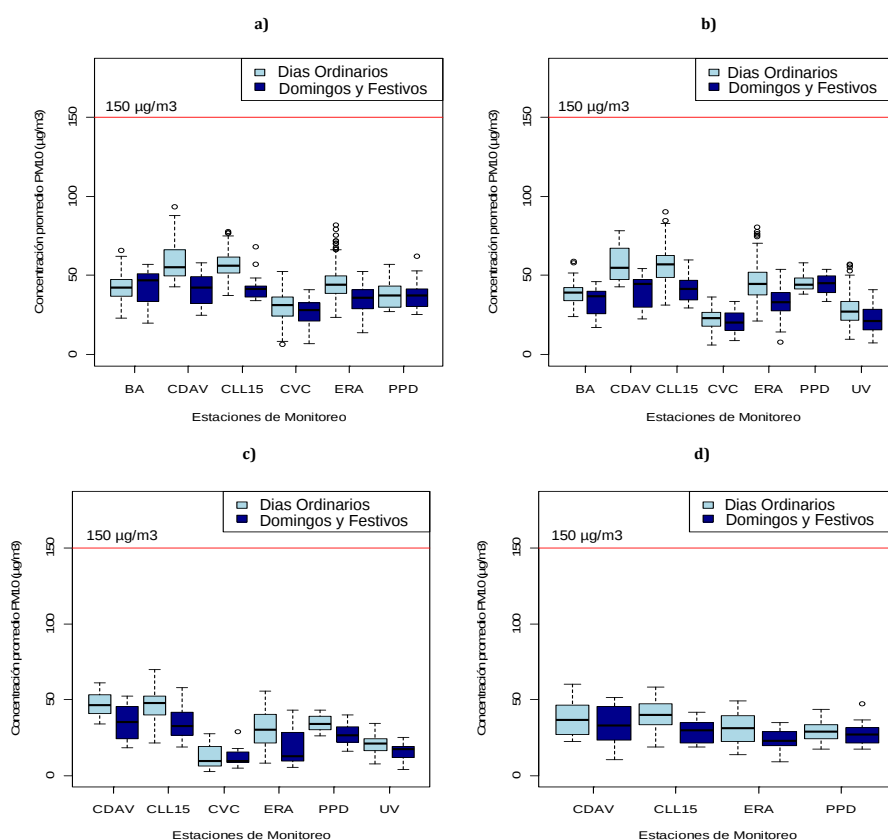
7. RESULTADOS

En esta sección se muestran los resultados obtenidos en el desarrollo de este proyecto.

7.1. ANÁLISIS EXPLORATORIO DE DATOS

A continuación se presentarán los resultados del análisis exploratorio de datos del contaminante PM₁₀ durante el período 2003-2006, para cada una de las estaciones de la red de monitoreo de calidad de aire del DAGMA (RMCA) por tipo de día.

Figura 2: Concentración promedio (24 Horas) de PM₁₀ por estación de monitoreo. (a) Año 2003, (b) Año 2004, (c) Año 2005, (d) Año 2006



Durante el año 2004 (Figura 2b), el comportamiento de la concentración promedio (24 horas) de PM₁₀ fue similar al observado en el año 2003 (Figura 2a). En general para todas las estaciones, las mayores concentraciones promedio se presentaron durante los días ordinarios, siendo las estaciones CLL15 y CDAV las más contaminadas; por el contrario, las estaciones con menores niveles de concentración promedio fueron CVC y UV. Los años 2005 (Figura 2c) y 2006 (Figura 2d) presentaron las menores concentraciones con respecto a los años anteriores para los diferentes tipos de día.

Durante el año 2005, las estaciones CLL15 y CDAV siguen reportando las mayores concentraciones promedio de PM10 durante días ordinarios y domingos y festivos. Las estaciones CVC y UV reportan los menores niveles de concentración independiente del tipo de día. La contaminación por material particulado durante el año 2006 fue similar en todas las estaciones monitoreadas tanto para días ordinarios como para domingos y festivos.

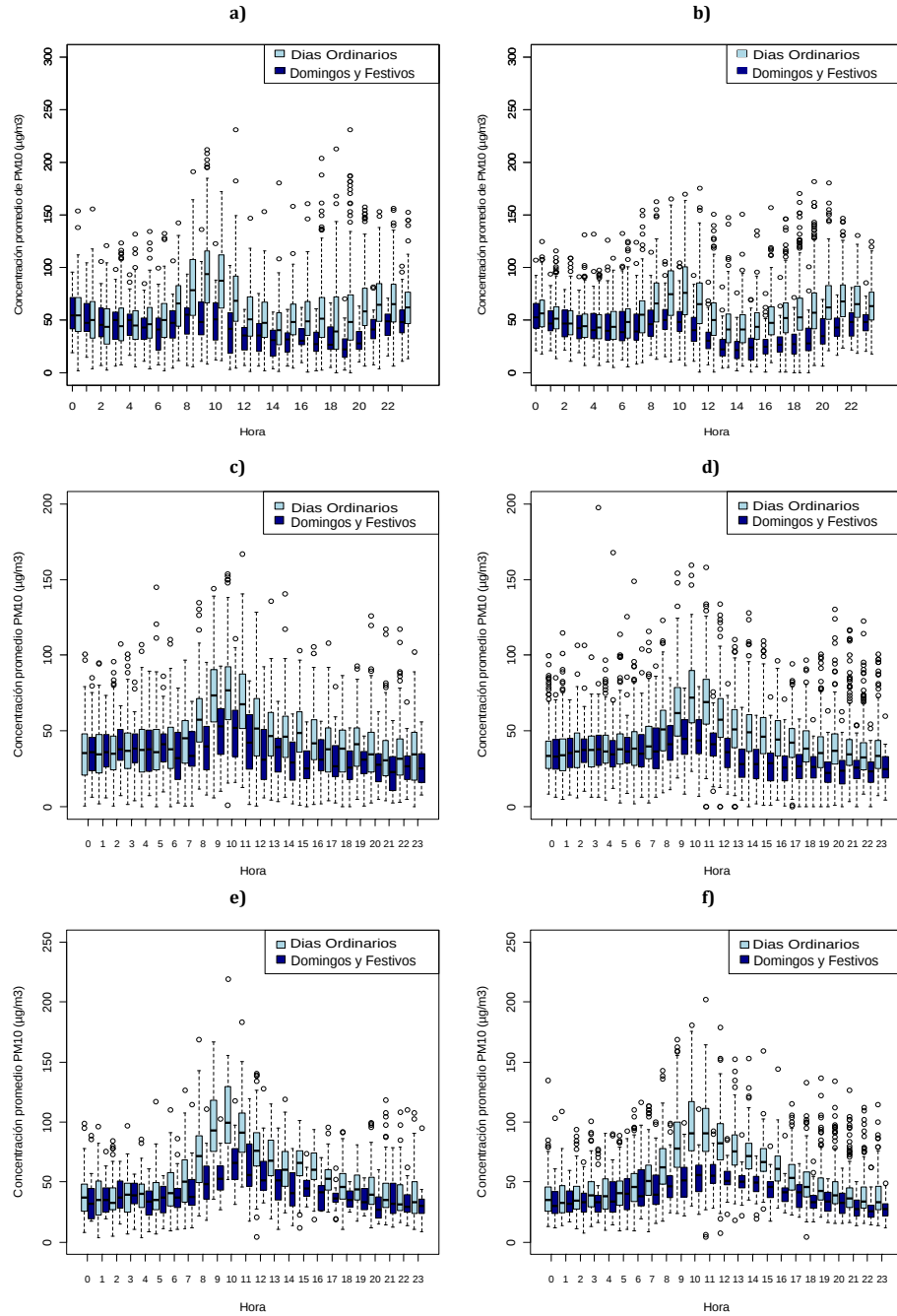
De forma general se tiene que las máximas concentraciones promedio de PM10 se presentaron durante los días ordinarios para las diferentes zonas de monitoreo de la RMCA durante el período de estudio; además en ninguna de las zonas analizadas se presentaron excedencias en el valor permisible para la contaminación por PM10 estipulado en la norma de calidad del aire (Resolución 601 de 2006).

En el Anexo 2, se puede observar el comportamiento del material particulado PM10 durante los años 2005 y 2006 en las zonas norte (CDAV) y centro de la ciudad de Cali (CLL15 y ERA). La tendencia del contaminante en este período de análisis no presentó ninguna variación durante las 24 horas de monitoreo, además de observarse una disminución de la concentración promedio del contaminante con respecto a los años 2003 y 2004.

Por el contrario, en las zonas rural y sur de la ciudad, se presentaron los menores niveles de contaminación por PM10. De la misma manera en el Anexo 2, se encuentran los resultados del comportamiento de la concentración promedio (horario) de PM10 para la estación CVC en el período 2003-2005 y UV en los años 2003 y 2004. En la estación CVC, los resultados del año 2003 reflejan que la máxima concentración promedio se presentó alrededor de las 10:00 am en días ordinarios; el período 2004 y 2005 de la zona rural de la ciudad evidenció una disminución en los niveles de contaminación por PM10 y no se observan notables diferencias entre la concentración promedio del particulado durante días ordinarios, domingos y festivos. En la estación UV durante el período de estudio se refleja un comportamiento del PM10 similar al observado en la zona rural de la ciudad.

La zona oriente de la ciudad se caracterizó según el comportamiento de la concentración promedio de PM10 para la estación PPD y BA durante el período 2003-2006 (Anexo 2). Para la estación BA se analizaron los registros de los años 2003-2004. Los resultados obtenidos reportaron altos niveles de contaminación en esta zona durante días ordinarios del año 2003, dicha contaminación disminuyó durante el año 2004. En ambos años el pico de contaminación más alto se presentó alrededor de las 10:00 am de días ordinarios y domingos y festivos. La estación PPD presentó los mayores niveles de contaminación por PM10 en los años 2003 y 2004, aunque en esta estación no se reflejaron marcadas diferencias entre la concentración promedio del material particulado para días ordinarios y domingos y festivos. Al igual que en las otras estaciones, la máxima concentración promedio de PM10 se observó alrededor de las 10:00 am y en los años 2005-2006 se evidenció una disminución en los niveles de contaminación del material particulado.

Figura 3: Concentración promedio (horaria) de PM10 por zona de monitoreo. (a) CDAV-2003, (b) CDAV-2004, (c) CLL15-2003, (d) CLL15-2004, (e) ERA-2003, (f) ERA-2004



En la Figura 3, se muestran los resultados de la concentración promedio (horario) para las estaciones de monitoreo que evidenciaron los mayores niveles de contaminación por PM10 en las diferentes zonas y años de estudio.

En general, el análisis horario del comportamiento de la concentración promedio de PM10 muestra que los máximos niveles de contaminación se presentaron alrededor de las 10 de la mañana, principalmente en días ordinarios de los años 2003 y 2004 en las zonas norte y centro de la ciudad. Los registros horarios de contaminación por PM10 para la estación CDAV en los años 2003 (Figura 3a) y 2004 (Figura 3b) evidenciaron altos niveles de concentración para esta zona. De forma similar, en la zona centro de la ciudad se observaron elevados niveles de contaminación para la estación CLL15 en los años 2003 (Figura 3c) y 2004 (Figura 3d); la estación ERA fue otra de las estaciones que reportó altos niveles de concentración durante el período 2003 (Figura 3e) y 2004 (Figura 3f). Para cada una de las zonas y períodos analizados, se obtuvo que los mayores niveles de contaminación por PM10 se presentaron durante días ordinarios.

7.2. CURVAS DE SUAVIZACIÓN

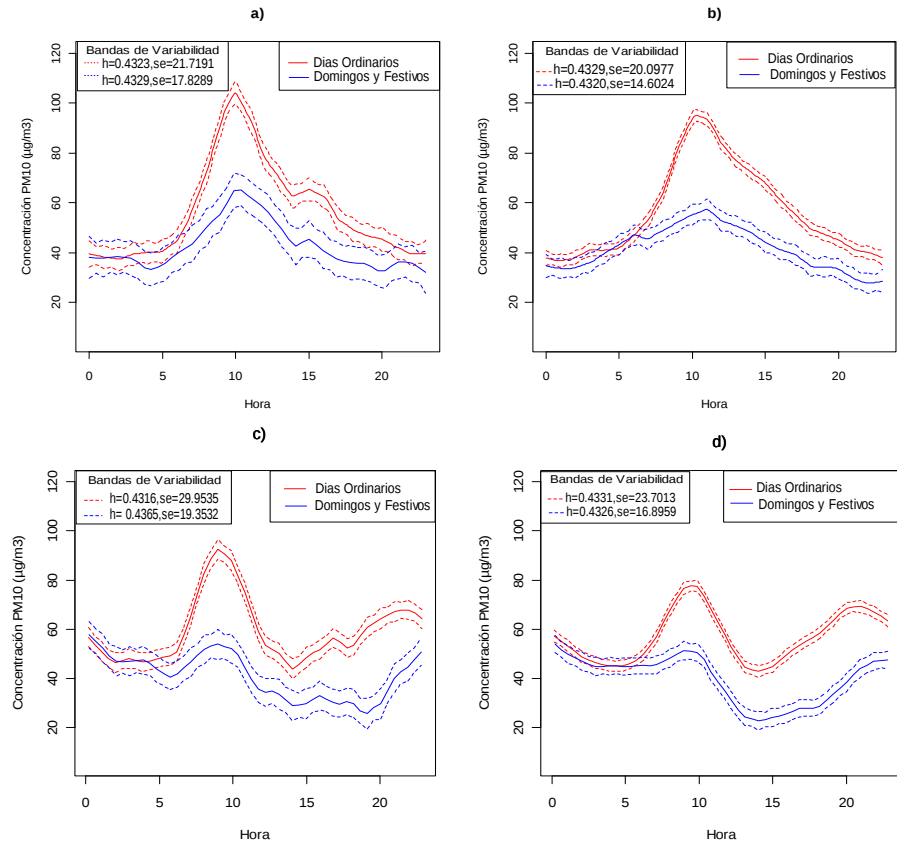
La estimación de las curvas típicas para este estudio se ajustaron para cada una de las estaciones de monitoreo de la RMCA en los respectivos períodos de análisis para una concentración promedio (horario) de PM10 en días ordinarios y domingos y festivos a través del estimador lineal local. A continuación se enseñarán las curvas típicas que evidenciaron los mayores niveles de contaminación por PM10 para una concentración horaria en las zonas de estudio.

En la Figura 4, se observan las zonas más contaminadas por PM10 según la estimación de curvas de suavización a través de la regresión lineal local.

La estimación de la concentración promedio horaria de material particulado a través de las curvas de suavización, permitieron identificar las zonas con los mayores niveles de contaminación por PM10 según el día de la semana. En la Figura 4a, se observa que el centro de la ciudad fue una de las zonas más contaminadas, debido a que los registros de la estación CLL15 para el año 2003 reportaron los mayores niveles de concentración promedio por PM10 durante 24 horas; es evidente que los niveles de contaminación por material particulado PM10 en la zona centro se incrementan a partir de las 8:00 am hasta alcanzar el pico de máxima concentración alrededor de las 10:00 am y después dicha contaminación decrece hasta las últimas horas de la noche (11:00 pm) tanto en días ordinarios como domingos y festivos. Sin embargo, se obtuvo que durante las 24 horas analizadas, los días ordinarios fueron los más contaminados por PM10 en todas las zonas y años de monitoreo. La Figura 4b refleja un comportamiento similar al presentado en el año 2003, con una leve disminución en la concentración promedio del contaminante y variación en la tendencia de éste durante las últimas horas del día. La zona norte se presenta en las Figuras 4c y 4d que corresponden a los registros de monitoreo del material particulado en la estación CDAV para los años 2003 y 2004, respectivamente. Los resultados muestran un comportamiento análogo al presentado en la estación CLL15, pues los días ordinarios presentaron los mayores niveles de contaminación después de las

8:00 am, alcanzando su punto de máxima concentración alrededor de las 10:00 am. La principal diferencia radica en que la zona norte reporta un incremento en los niveles de concentración de PM10 durante las últimas horas del día hasta las 11:00 pm, caso opuesto al presentado en la zona centro.

Figura 4: Curvas de suavización y bandas de variabilidad por zona de monitoreo. (a) CLL15-2003, (b) CLL15-2004, (c) CDAV-2003, (d) CDAV-2004



En el Anexo 3, se muestran los resultados de las curvas de suavización de la zona norte (CDAV) y centro (CLL15) de la ciudad para los años 2005 y 2006. Durante este período de estudio, el nivel máximo de contaminación por PM10 se presentó a las 10:00 am, los días ordinarios reportaron mayores niveles de concentración promedio de material particulado con respecto al observado en domingos y festivos. Para estos años de monitoreo, los resultados evidencian una reducción en la concentración promedio respecto a la contaminación presentada en años anteriores. Un resultado particular de esta tendencia se relaciona con el hecho de que al disminuir los niveles de contaminación por PM10, no es muy notable la diferencia entre la concentración promedio de este contaminante en domingos y festivos. Las bandas de variabilidad estimadas reflejan que existe una mayor precisión en la estimación de la curva de suavización de la contaminación por PM10 en días ordinarios con respecto a la curva de los domingos y festivos, ya que las bandas de variabilidad

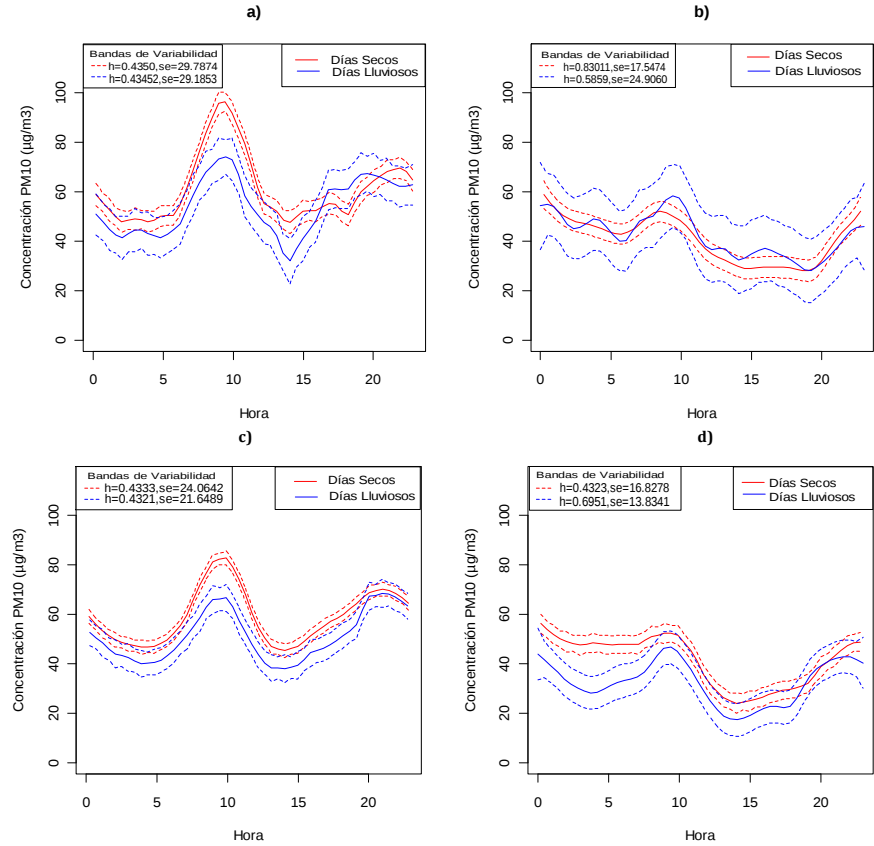
construidas para estos días presentan una mayor amplitud que las bandas de días ordinarios. Las estaciones CVC y UV ubicadas en la zona rural y sur de la ciudad respectivamente, reportaron las más bajas y constantes concentraciones promedio de material particulado con respecto a las demás estaciones de monitoreo analizadas en cada período de estudio. Las curvas típicas no reflejan diferencia entre los días de monitoreo (ordinarios y domingos y festivos) en la zona rural y sur de Cali (Anexo 3).

De igual manera los resultados correspondientes a las estaciones del oriente de Cali (PPD, BA y ERA) se muestran en el Anexo 3. Las curvas típicas evidencian que la contaminación en la estación PPD en el período 2003-2006 no reflejó diferencias en la concentración promedio de material particulado para días ordinarios y domingos y festivos, presentándose alrededor de las 10:00 am los niveles máximos de contaminación. Durante los años 2005 y 2006 se disminuyó la concentración promedio del contaminante tanto en días ordinarios como domingos y festivos. Un comportamiento particular de la estación BA en el período 2003-2004, se observó después de las 6:00 pm, en la que se evidencia un incremento para la concentración de PM10 en días ordinarios y domingos y festivos hasta las 11:00 pm. La curva típica estimada de la estación ERA, ubicada en el centro de la ciudad reporta un nivel de contaminación por PM10 inferior al observado en el análisis exploratorio de datos para cada uno de los años evaluados, en general, el comportamiento observado en esta estación no difiere del presentado en las otras zonas de monitoreo, es decir, los mayores niveles de contaminación por PM10 se presentaron durante días ordinarios, la máxima concentración promedio del contaminante se observó alrededor de las 10:00 am y los años que reportaron mayor contaminación por PM10 en la zona de la estación ERA fueron 2003 y 2004.

En la Figura 5, se presentan los resultados de la estimación de las curvas de suavización a través de regresión lineal local en la estación CDAV ubicada al norte de la ciudad, esto con el fin de evaluar el comportamiento de la concentración promedio de PM10 en días ordinarios y domingos y festivos bajo las condiciones de un clima seco o lluvioso en el período 2003 y 2004, el cual reportó un comportamiento característico del contaminante.

Los resultados de la concentración promedio para el contaminante PM10 en la estación CDAV durante días ordinarios en los años 2003 (Figura 5a) y 2004 (Figura 5c) reflejan que en general, los mayores niveles de contaminación se presentaron durante días secos. Las curvas de suavización estimadas permiten observar que durante las primeras horas de los días ordinarios y domingos y festivos se registraron bajos niveles de concentración; a las 10:00 am se reportó el pico de máxima contaminación por PM10, después de esta hora la concentración promedio disminuyó hasta alcanzar un mínimo nivel de contaminación alrededor de las 2:00 pm; luego a partir de las 3:00 pm se incrementa nuevamente la concentración de PM10 hasta las 11:00 pm independientemente del tipo de día. En los días domingos y festivos del año 2003 (Figura 5b) y 2004 (Figura 5d) se evidenció un comportamiento diferente del contaminante, pues sus niveles fueron inferiores a los presentados durante días ordinarios y no se observan aparentes diferencias entre la concentración promedio (horario) de PM10 para días secos y lluviosos. En el Anexo 4 se muestran los resultados obtenidos para esta estación durante días secos y lluviosos en los años 2005 y 2006.

Figura 5: Curvas de suavización y bandas de variabilidad según clima en la estación CDAV. (a) Ordinarios-2003, (b) Domingos y festivos-2003, (c) Ordinarios-2004, (d) Domingos y festivos-2004

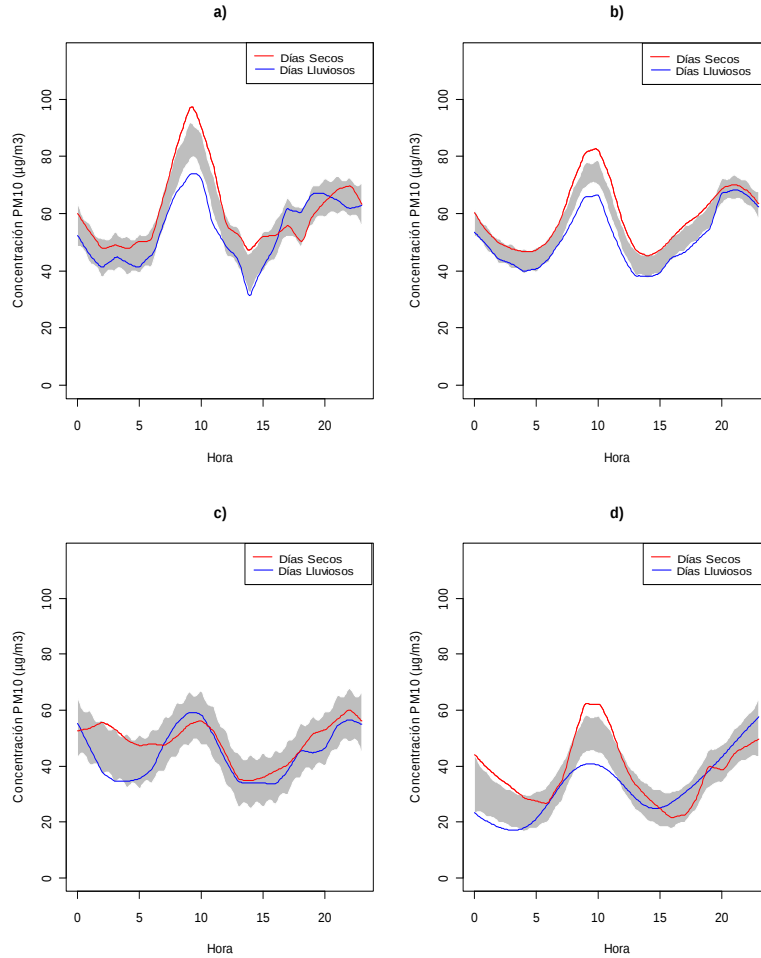


En la Figura 6, se observa la comparación de las curvas de la estación CDAV en el año 2003 según el clima seco o lluvioso en días ordinarios. La comparación de curvas de regresión no paramétrica se realizó mediante el análisis gráfico de Bandas de Referencia. El comportamiento del contaminante según la condición climática en la estación CDAV para los años 2004, 2005 y 2006 en días ordinarios presentó variaciones con respecto al observado en el año 2003, pues se evidencia que existen diferencias entre las curvas típicas en los horarios en que éstas no se entrelazan.

Las bandas de referencia de los años 2003 (Figura 6a) y 2004 (Figura 6b) reflejan que existen diferencias en los niveles de concentración promedio (horario) de PM10 entre días secos y lluviosos alrededor de las 10:00 am, el año 2005 (Figura 6c) no evidenció diferencias representativas entre los niveles de concentración promedio de días ordinarios en días secos y lluviosos; los resultados del año 2006 (Figura 6d), reportaron diferencias significativas, en los niveles de concentración de PM10 en la estación CDAV en clima seco y lluvioso durante las primeras horas del día y a las 10:00 am, hora en la que se presentaron los máximos picos de contaminación. Los resultados indican que en días secos a las 10:00 am, la contaminación por PM10 en días ordinarios fue superior a la

presentada en días lluviosos en la zona norte de la ciudad, en los años 2003, 2004 y 2006.

Figura 6: Curvas de suavización y bandas de referencia para la estación CDAV. (a) 2003, (b) 2004, (c) 2005, (d) 2006



7.3. VALIDACIÓN DE SUPUESTOS

En la regresión no paramétrica el paso final al ajuste de las curvas de suavización a través del método de regresión lineal local es la validación de los supuestos sobre los residuales. Al considerar la covariable tiempo, la cual contiene múltiples observaciones de concentración de material particulado por hora, se considera para la estimación de curvas típicas la definición de pseudoresiduales propuesta por Gasser, Sroka & Jennen-Steinmetz (1986). Los pseudoresiduales de los modelos

de regresión no paramétricos ajustados a través del estimador lineal local tienen una varianza homogénea y su valor esperado es igual a cero. En ese sentido, se muestra en el Anexo 5 que para cada una de las curvas estimadas según zona y año de monitoreo se satisfacen los supuestos validados para los pseudoresiduales de los modelos ajustados.

8. PROPUESTA EPIDEMIOLÓGICA

Según el documento CONPES 3344 de 2005, la contaminación del aire en las ciudades es un problema sobre cuya importancia existe una amplia conciencia social en Colombia. De hecho, es el problema ambiental de mayor preocupación para los Colombianos (Lemoine, 2004) y el generador de los mayores costos sociales después de la contaminación del agua y de los desastres naturales. Estos costos han sido estimados en 1.5 billones de pesos anuales (Bjorn, 2004) y están relacionados con efectos sobre la salud pública, mortalidad y morbilidad.

El material particulado PM10 y PM2.5 es conocido como “respirable”, ya que puede penetrar los mecanismos de defensa del sistema respiratorio y llegar hasta los bronquios o incluso el alvéolo pulmonar, como es el caso de las “partículas ultrafinas” que están por debajo de $2.5 \mu m$ y pueden llegar a medir incluso menos de $0.1 \mu m$ de diámetro (Harrison et al, 2000). Entre los efectos negativos de la contaminación del aire sobre la salud y la productividad de las personas están: cáncer, asma, bronquitis crónica y desórdenes respiratorios (Shah J. et al 1997).

En el parágrafo segundo de la Resolución 601 de 2006 se invita a las autoridades ambientales competentes a la investigación de probables afectaciones a la salud humana por la exposición a partículas suspendidas totales (PST) entre ellas PM10 y PM 2.5. Es por esa razón que luego de caracterizar la contaminación por material particulado PM10 en las principales zonas de la ciudad de Cali, durante el período 2003-2006 e identificar que los mayores niveles de contaminación por PM10 se presentaron en la zona norte y centro de la ciudad, se propone analizar datos de monitoreo de la calidad del aire de PM10 y PM2.5 recientes con el objeto de observar si estas zonas aún mantienen elevadas concentraciones de material particulado e investigar sobre la asociación del desarrollo de enfermedades respiratorias en los habitantes de las zonas más contaminadas por PM10 y PM2.5 debido a largos períodos de exposición al mismo.

El presente estudio aporta de forma retrospectiva en la identificación de las zonas, horas y días de la ciudad de Cali que reportaron altos niveles de contaminación por material particulado PM10 en el período 2003-2006.

En ese sentido, se propone realizar un estudio epidemiológico prospectivo seleccionando una cohorte de habitantes de las zonas centro y norte de la ciudad de Cali que según los resultados de este estudio reportaron ser las zonas de mayor contaminación por material particulado. Para ello es fundamental involucrar en dicho estudio la variable concentración promedio de PM10 y PM2.5 y además considerar como variable respuesta la presencia o ausencia de enfermedades respiratorias en las personas que se seleccionen en la muestra, y a partir de esto ajustar una regresión logística no paramétrica que permita estimar si la exposición al material particulado se asocia como factor de riesgo en el desarrollo de las afecciones respiratorias en las personas. En este estudio se propone utilizar la regresión logística como metodología apropiada al considerar una variable respuesta dicotómica (presencia ó ausencia de enfermedades respiratorias).

Adicionalmente, se propone a los epidemiólogos indagar en la incidencia de crisis asmáticas en los habitantes de las zonas centro y norte de la ciudad durante las horas en que se presentaron las máximas concentraciones de material particulado, en especial a las 10:00 am, hora que reportó el máximo nivel de contaminación por PM10 en las diferentes zonas de monitoreo y alrededor de la media noche (entre las 11:00 pm y 12:00 am). También se recomienda comparar la frecuencia de complicaciones respiratorias entre días ordinarios, domingos y festivos con el objetivo de analizar si en los días que se evidenciaron mayores niveles de contaminación por PM10 (días ordinarios) es mayor la incidencia de enfermedades respiratorias.

El llevar a cabo este análisis le permitiría a las entidades de salud pública tomar posibles medidas de control y prevención para los habitantes de las zonas que reportaron elevados niveles de concentración por material particulado PM10 y PM2.5 en la ciudad de Cali, con respecto a la exposición al aire contaminado durante las horas y días en que se identificaron características de alta concentración promedio de material particulado.

9. CONCLUSIONES Y RECOMENDACIONES

- El ajuste de curvas de suavización Kernel a través del Estimador Lineal Local de regresión no paramétrica en la modelación del comportamiento del material particulado PM10, permitió analizar la contaminación en las principales zonas de la ciudad de Santiago de Cali, tomando como referencia la ubicación de las estaciones de monitoreo del DAGMA, en el período de estudio 2003-2006. Los resultados reflejaron que el año que presentó una mayor concentración promedio (horario) de PM10 en la ciudad de Cali fue el 2003; los días más contaminados fueron los ordinarios (lunes, martes, miércoles, jueves, viernes y sábado) y a las 10:00 am se reflejó el pico de máxima concentración en las diferentes zona de la ciudad y para cada año de estudio. Un resultado particular de este contaminante, es que durante el día sábado la concentración promedio se comportó de forma similar a un día ordinario.
- Una posible explicación al comportamiento sinusoidal del material particulado PM10 en las zonas que presentaron elevados niveles de contaminación durante el período de estudio, se puede asociar con el hecho de que en el año 2003 (primer año de monitoreo del contaminante en la ciudad) se implementó la medida de control de flujo vehicular “pico y placa”(Decreto No. 0722 de Octubre 28 de 2005) para vehículos particulares entre 7:00-9:00 am y las 5:30-7:00 pm. Es evidente que a las 10:00 am se presentaron los máximos niveles de contaminación por PM10, hora en la que se ha normalizado el flujo vehicular en la ciudad después del pico y placa diurno, en el caso de la noche la concentración de material particulado en todas las zonas y años de análisis fue bajo en el horario de pico y placa nocturno (5:30 a 7:00 pm).
- Las zonas centro y norte de Cali, fueron los sectores más contaminados por material particulado PM10 en días ordinarios. La contaminación en la zona centro se diferenció de la presentada en la zona norte, en el comportamiento del contaminante durante las horas de la noche, pues en el centro de la ciudad se evidenció que después de las 10:00 am la concentración promedio del contaminante decreció hasta las 11:00 pm, caso contrario al presentado en la zona norte, ya que en este sector se observó que en la noche se vuelven a incrementar los niveles de contaminación. En ese sentido, se considera la hipótesis de que las fuentes de contaminación características de cada zona de monitoreo juegan un papel determinístico en el comportamiento de la concentración por material particulado PM10, pues la zona norte de la ciudad es un sector industrial de alto flujo vehicular, mientras que la zona centro de la ciudad, es un sector comercial de alto tráfico vehicular activo durante las horas de la mañana y tarde.
- Al analizar la concentración según la condición climática del día, se tomó como referencia la estación CDAV ubicada en la zona norte de la ciudad al ser una de las más contaminadas. Se evaluó el comportamiento de la concentración promedio de PM10 en días ordinarios que presentaron clima seco o lluvioso. Los resultados evidenciaron diferencias significativas en los niveles de contaminación por PM10 en días ordinarios para clima seco y lluvioso a las 10:00 am, es decir, los días ordinarios con clima seco reflejaron un mayor grado de contaminación por PM10, lo cual representa un comportamiento consistente pues comúnmente se conoce el

material particulado con el nombre de polvo u hollín y es de esperar que cuando hay presencia de precipitación, el contaminante disminuye sus niveles de concentración en el aire.

- Una recomendación para estudios posteriores se realiza en el sentido de que futuros investigadores indaguen sobre el comportamiento del material particulado PM10 según las condiciones climáticas en la zona centro, sur y oriente de la ciudad, debido a que en este estudio por limitación de información de la variable meteorológica “precipitación” no fue posible llevar a cabo el análisis del contaminante para esas zonas de la ciudad.
- Adicionalmente, se recomienda investigar sobre el comportamiento del material particulado PM2.5 en la ciudad de Cali considerado aún más perjudicial que el PM10 para la salud humana, pues el DAGMA a partir del año 2010 y hasta la actualidad monitorea este contaminante en la ciudad. Finalmente, se considera necesario que las entidades de control ambiental nacional determinen a través de la investigación, un valor máximo permisible para la concentración promedio de PM10 y PM 2.5 horario, con el objeto de evaluar en que estado se encuentra la ciudad de Cali en contaminación horaria por material particulado.

BIBLIOGRAFÍA

1. AYALA D., OJEDA C. & OLAYA J. (2007). “*Evaluación de la calidad del aire en la ciudad de Cali asociada al monóxido de carbón, dióxido nitrógeno y ozono a través de la estimación y comparación de curvas típicas diarias no paramétricas*”. Universidad del Valle.
2. BARRIENTOS A., GONZÁLEZ V. & OLAYA J. (2007). “*Un modelo spline para el pronóstico de la demanda de energía eléctrica*”. Revista Colombiana de Estadística, Vol. 31, pp. 187-202.
3. BJORN L. (2004). “*Cost of environmental damage: A Socio-Economic and Environmental Health Risk Assessment*”.
4. BOWMAN A. & AZZALINI A. (1997). “*Applied smoothing techniques for data analysis. The kernel approach with S-Plus illustrations*”. Oxford University Press Inc., 193 p.
5. CASTRO D., FLÓREZ A. & OLAYA J. (2009). “*Propuesta de estimación no paramétrica de la contaminación horaria por Monóxido de Carbono en el centro de la ciudad de Cali, Colombia, entre 2003 y 2006*”. VI Encuentro Colombia-Venezuela de Estadística, Valencia, Venezuela.
6. CLEVELAND W. (1979). “*Robust Locally Weighted Regression and Smoothing Scatterplots*”. Journal of the American Statistical Association, Volumen 74, Número 368, 829-836.
7. CLEVELAND W. & DEVLIN S. (1988). “*Locally weighted regression: An approach to regression analysis by local fitting*”. Journal of the American Statistical Association, Volumen 83, 596-610.
8. CONPES 3344 (2005). “*Lineamientos para la formulación de la política de prevención y control de la contaminación del aire*”. Colombia: s.n.
9. DAGMA (Departamento Administrativo de Gestión Ambiental Municipal). Consultada en enero 23 de 2010. <http://www.cali.gov.co/Dagma.php>
10. DENNIS R. et al. (1999). “*Prevalencia de asma y otras enfermedades alérgicas en Colombia: Resultados preliminares en Santa Fe de Bogotá*”. Revista Colombiana de Neumología, Volumen 11, 13-23.
11. DETTE H. & NEUMEYER N. (2001). “*Nonparametric analysis of covariance*”. The Annals of Statistics. Volumen 29, Número 5, 1361-1400.
12. DETTE H. & NEUMEYER N. (2001). “*Nonparametric comparison of regression curves - an empirical process approach*”. Bochum, 33p.
13. DETTE H., MUNK A. & WAGNER T. (1998). “*Estimating the Variance in Nonparametric Regression-What is a Reasonable Choice?*”. Journal of the Royal Statistical Society. Series B (Statistical Methodology), Volumen 60, Número 4, 51-764.
14. ECHEVERRI C. & MAYA G. (2008). “*Relación entre las partículas finas (PM_{2.5}) y respirables (PM₁₀) en la ciudad de Medellín*”. Revista Ingenierías Universidad de Medellín, Volumen 6 Número 12, 23-42.

15. EUBANK R. (1999). *"Nonparametric Regression and Spline Smoothing"*. 2 Ed. New York: Marcel Dekker, 338 pag.
16. FLÓREZ, A. & OLAYA, J. (2007). *"Comparación por simulación de varios estimadores de va-rianza basados en diferencias sucesivas en el modelo de regresión no-paramétrica"*. Universidad del Valle.
17. FOX J. (2000). *"Nonparametric Simple Regression: Smoothing Scatterplots"*. Sage, Thousand Oaks CA.
18. FOX J. (2005). *"Introduction to Nonparametric Regression"*. Hamilton. 12 p.
19. GASSER T., SROKA L., & JENNEN-STEINMETZ C. (1986). *"Residual Variance and Residu-al Pattern in Nonlinear Regression"*. Biometrika, Volumen 73, 625-633.
20. GRANADA L. et al (2008). *"Aplicación del software Predicción de la Calidad del Aire (PCA) en la gestión de la calidad del aire en Cali - Colombia"*. XIV Convención Científica en Ingeniería y Arquitectura. Instituto Politécnico Superior José Antonio Echeverría CUJAE. Ciudad de la Habana, 113-114.
21. GRANADA L. et al (2007). *"Relación entre enfermedades respiratorias agudas y contaminación atmosférica por fuentes móviles en Cali - Colombia"*, Libreprensa 4, 76-77.
22. HARRISON R. & YIN J. (2000). "Particulate matter in the atmosphere: which particle properties are important for its effects on health?". The Science of the Total Environment, Número 249, 85-101.
23. IDEAM (Instituto de Hidrología, Meteorología y Estudios Ambientales de Colombia). Consultada en marzo 11 de 2010. <http://www.ideam.gov.co>.
24. JARAMILLO M. et al. (2004). *"Inventario de emisiones de contaminantes atmosféricos convencionales en la zona de Cali-Yumbo"*. Revista Facultad de Ingeniería, Número 31, 38-48.
25. JOHNSON N. & KOTZ S. (1970). *"Distributions in statistics: Continuous univariate distribution"*. Volumen 2. New York: Wiley.
26. KAFADAR K. & HORN P. (2002). *"Smoothing"*. Encyclopedia of Envirometrics, Volumen 4, pp 2014-2020.
27. LEMOINE C. (2004). *"Prioridades Ambientales Colombianas"*. Informe de Consultoría. Banco Mundial.
28. LOZANO N. (2003). *"Air pollution in Bogotá, Colombia: A concentration response approach"*. Department of Agricultural and Resource Economics. University of Maryland of Collage Park.
29. MACHADO A. et al. (2007). *"Metales en PM10 y su dispersión en una zona de alto tráfico vehicular"*. INTERCIENCIA, Volumen 32 Número 5, 312-317.

30. MONTOYA M., MORALES A. & OLAYA J. (2005). "*Estimación no paramétrica de curvas típicas diarias para los contaminantes CO, NO₂ y SO₂ en Santiago de Cali, Colombia*". Revista EIDENAR, 23-27.
31. MÜLLER H. & STADMÜLLER U. (1987). "*Estimation of heterocedasticity in regression analysis*". The Annals of Statistics. Número 15, 610-635.
32. OLAYA, J. (2002). "*Suavización y regresión no-paramétrica: una alternativa de modelación estadística*".
33. OSTRO B. et al. (1995). "*Air Pollution and Mortality: Results from a Study in Santiago, Chile*". The World Bank, Policy Research Working Paper No. 1453, Public Economic Division.
34. PROGRAMA DE LAS NACIONES UNIDAS PARA EL MEDIO AMBIENTE (PNUMA) (2005) Capa de Ozono, Proyecto Ciudadanía Ambiental Global. Consultado en enero 11 de 2010. <http://www.parlatino.org/downloads/ozono.pdf>.
35. PAZ M., PEREIRA L. & OLAYA J. (2009). "*Estimadores de varianza en regresión no paramétrica basados en sucesiones de diferencias*". Heurística, Número 15, 5-12.
36. REINA J., PEREIRA L. & OLAYA J. (2009). "*Modelación no paramétrica del promedio octohorario del ozono troposférico en la ciudad de Cali, Colombia*". VI Encuentro Colombia-Venezuela de Estadística, Valencia, Venezuela.
37. RICE J. (1984). "*Bandwidth choice for nonparametric regression. The Annals os Statistics*". Número 12, 1215 - 1230.
38. SÁNCHEZ Y. & CUÉLLAR C. (2009). "*Comprueban que las partículas suspendidas en el aire dañan ADN*". La crónica de hoy. México. Marzo 21 de 2009.
39. SANHUEZA P., VARGAS C. & MELLADO P. (2006). "*Impacto de la contaminación del aire por PM10 sobre la mortalidad diaria en Temuco*". Revista Médica de Chile, Volumen 134, 754-761.
40. SECRETARIA DE TRANSITO Y TRANSPORTE MUNICIPAL. Consultada en febrero 17 de 2011. <http://www.cali.gov.co/transito/>
41. SEIFERT B., GASSER T. & WOLF A. (1993). "*Nonparametric estimation of residual variance revisited*". Biometrika. Volumen 80, Número 2, 373 - 383.
42. SHAH J., NAGPAL T. & BRANDON C. (1997). "*Urban Air Quality Management Strategy in Asia*". The World Bank. Washington.
43. SIMONOFF J. (1996). "*Smoothing Methods in Statistics*". Springer, New York.
44. SPECKMAN P. (1988). "*Kernel smoothing in partial linear models*". Journal of the Royal Statistical Society. Series B. Número 50, 413-436.
45. TAFFUR L. & QUINTERO J. (1994). "*Contaminación Ambiental y Enfermedad Respiratoria en Cali*", Revista Avances en Medicina Social, Volumen 4, Número 1.

46. TORRES L., PAZ M. & OLAYA J. (2009). “*Aplicación de Técnicas de suavización no paramétricas en la modelación de la contaminación promedio octohoraria del Monóxido de Carbono en Cali*”, Colombia. VI Encuentro Colombia-Venezuela de Estadística, Valencia, Venezuela.
47. YOUNG S & BOWMAN A. (1995). “*Nonparametric analysis of covariance*”. Biometrics. Volumen 51, Número 3, 920-931.

Páginas Web:

http://www.metrogas.cl/metrogas_industrial/asesoria/ambiental/ambiental_contamina.asp
<http://emmaq.corpaire.org/paginas/articulos/efectos.pdf>
http://whqlibdoc.who.int/hq/2006/WHO_SDE_PHE_OEH_06.02_spa.pdf
<http://socioeconomia.univalle.edu.co/nuevo/>
http://www.cepchile.cl/dms/archivo_1630_326/rev69_ostro.pdf
<http://www.cepis.ops-oms.org/bvsci/e/fulltext/intromon/cap7.pdf>
http://www.cronica.com.mx/especial.php?id_tema=1232&id_nota=421683
<http://www.epa.gov/airscience/quick-finder/particulate-matter.htm>
http://revistaeidenar.univalle.edu.co/edicion3/revista3_4a.phtml
http://www.drgdiaz.com/eco/salud/certi_cadogases3.jpg
http://www.ohdsa.com/calidad_de_aire_ambiental_pts_pm10_pm25.html

ANEXOS

ANEXO 1

Tabla 2: Estadísticas descriptivas según día de la semana

Est	Año	PROMEDIO							MÁXIMO						
		Dom	Lun	Mar	Mie	Jue	Vie	Sab	Dom	Lun	Mar	Mie	Jue	Vie	Sab
CDAV	2003	40.18	48.95	61.58	61.05	61.19	63.52	53.18	48.67	57.47	69.51	71.89	72.11	77.10	65.12
CDAV	2004	40.89	52.37	57.6	56.93	56.50	57.43	55.55	59.12	69.70	81.61	78.46	92.25	85.45	80.05
CDAV	2005	27.75	25.90	30.25	32.03	33.40	37.90	41.45	36.49	32.67	46.06	48.85	53.79	60.51	61.50
CDAV	2006	30.41	31.99	34.08	28.57	34.35	35.78	42.23	32.05	32.74	36.30	29.63	35.99	38.14	43.07
ERA	2003	33.38	37.69	43.99	44.61	47.41	47.43	44.60	39.85	45.69	55.13	52.23	60.75	63.80	58.33
ERA	2004	29.92	33.87	39.13	41.34	42.18	42.36	40.24	41.09	46.36	56.78	70.34	61.21	59.59	60.53
ERA	2005	22.84	15.12	31.85	34.68	32.15	29.47	36.11	39.36	16.51	32.38	51.21	42.34	38.87	48.28
ERA	2006	23.12	29.55	30.37	28.34	32.20	33.56	32.69	27.63	29.86	30.66	32.34	32.36	35.26	39.75
CLL15	2003	46.24	51.07	75.09	62.52	58.01	60.40	59.14	101.85	113.22	243.85	145.16	129.00	141.22	130.84
CLL15	2004	41.69	48.48	54.68	56.38	55.01	56.86	57.31	51.94	56.70	66.16	73.12	63.99	65.94	66.17
CLL15	2005	32.72	37.50	42.79	40.27	42.84	44.05	44.31	42.96	50.55	50.34	49.55	53.65	53.91	55.03
CLL15	2006	30.35	35.49	43.90	38.61	42.66	41.10	38.51	34.04	38.94	44.70	38.81	44.55	42.27	43.05
CVC	2003	26.54	27.62	32.32	33.97	34.79	31.51	30.95	35.18	34.21	36.53	42.07	41.76	42.31	41.53
CVC	2004	20.63	20.83	20.34	20.56	21.93	23.48	21.57	24.97	24.69	27.05	27.05	26.56	28.23	24.51
CVC	2005	12.88	8.73	9.96	9.24	17.34	10.04	15.19	21.51	11.31	18.23	11.61	20.55	20.49	26.66
UV	2004	22.73	23.95	28.47	29.19	28.97	29.74	28.58	32.61	28.93	46.70	50.41	40.69	45.06	41.58
UV	2005	14.08	14.07	16.41	20.66	21.71	21.65	21.26	18.77	16.35	18.22	22.04	25.48	22.12	26.59
BA	2003	36.25	45.08	63.95	73.27	70.94	48.88	48.31	40.31	123.11	235.5	230.62	166.87	68.92	61.74
BA	2004	30.87	33.67	33.20	37.41	39.40	40.74	38.28	36.38	40.87	45.85	45.21	46.84	49.41	43.44
PPD	2003	31.83	32.65	33.37	36.05	44.80	38.75	28.32	39.15	43.33	43.16	54.54	51.15	50.76	42.85
PPD	2004	82.15	108.33	74.15	46.89	46.73	48.69	43.51	237.76	283.19	225.84	75.75	57.40	54.90	55.95
PPD	2005	29.73	23.30	37.22	34.20	37.67	33.85	37.06	38.27	32.10	44.99	39.97	48.87	43.56	46.64
PPD	2006	26.78	24.83	27.69	25.98	27.76	33.65	26.90	33.09	32.02	33.94	36.26	29.20	39.09	29.47

ANEXO 2

Figura 7: Concentración promedio (horaria) de PM10 en la estación CDAV. (a) Año 2005, (b) Año 2006

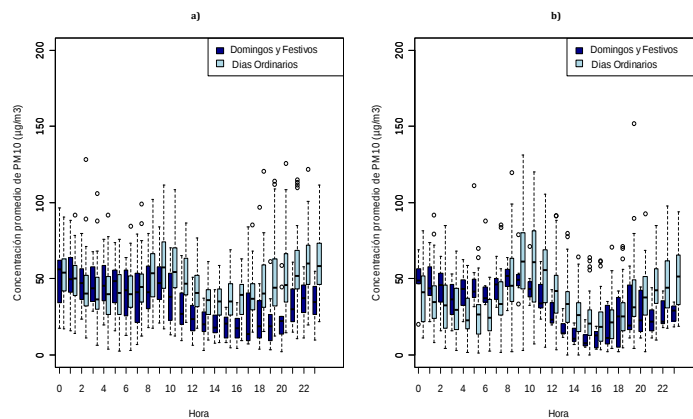


Figura 8: Concentración promedio (horaria) de PM10 en la estación CLL15. (a) Año 2005, (b) Año 2006

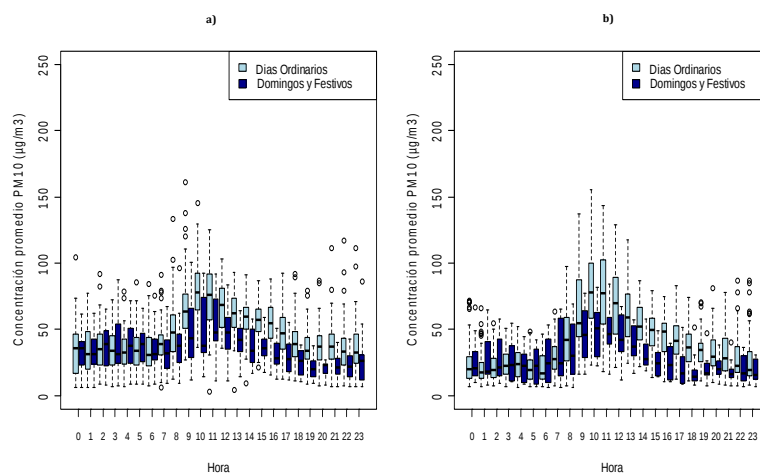


Figura 9: Concentración promedio (horaria) de PM10 en la estación ERA. (a) Año 2005, (b) Año 2006

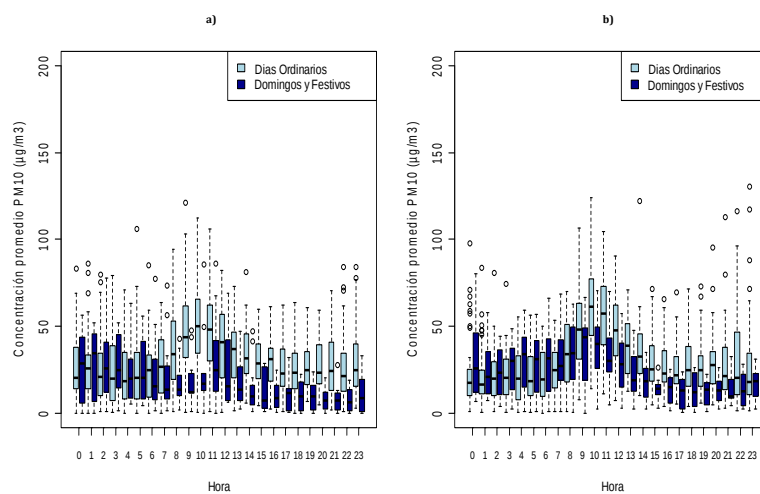


Figura 10: Concentración promedio (horaria) de PM10 en la estación UV. (a) Año 2004, (b) Año 2005

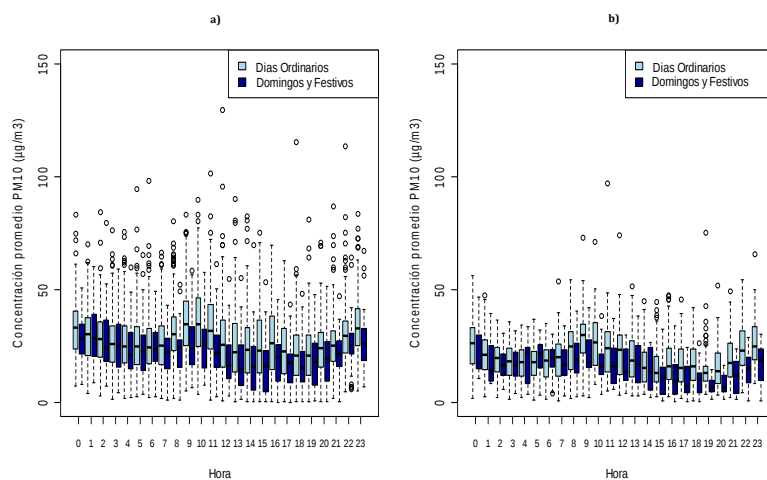


Figura 11: Concentración promedio (horaria) de PM10 en la estación BA. (a) Año 2003, (b) Año 2004

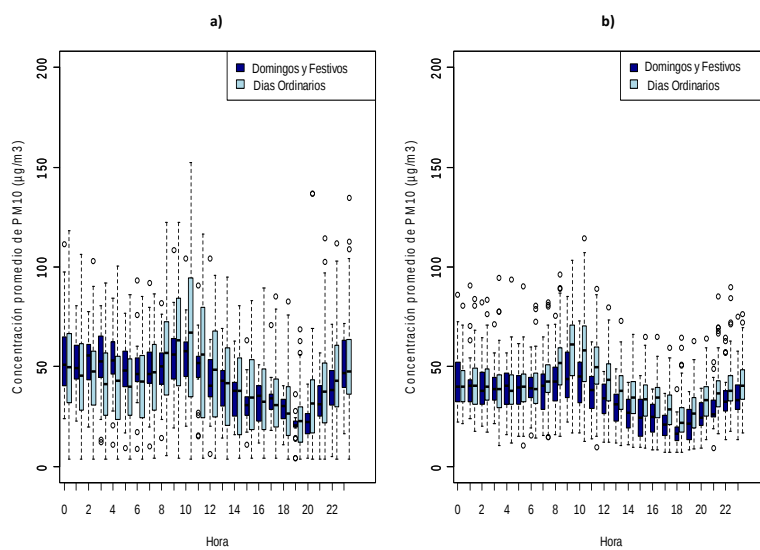


Figura 12: Concentración promedio (horaria) de PM10 en la estación CVC. (a) Año 2003, (b) Año 2004, (c) Año 2005

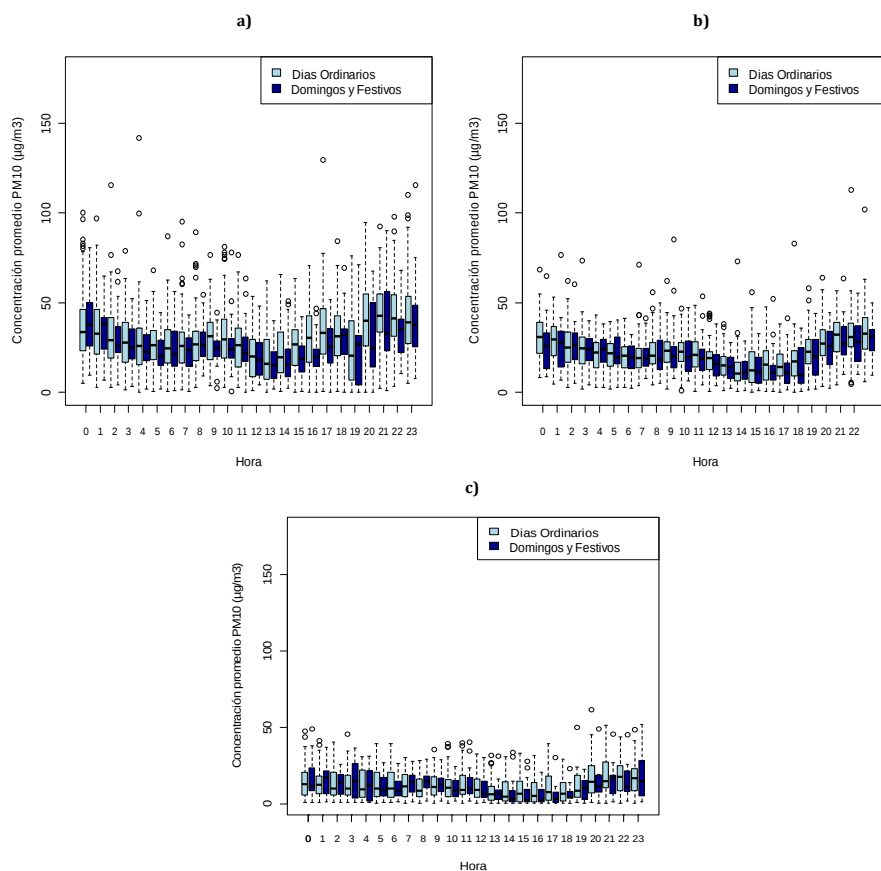
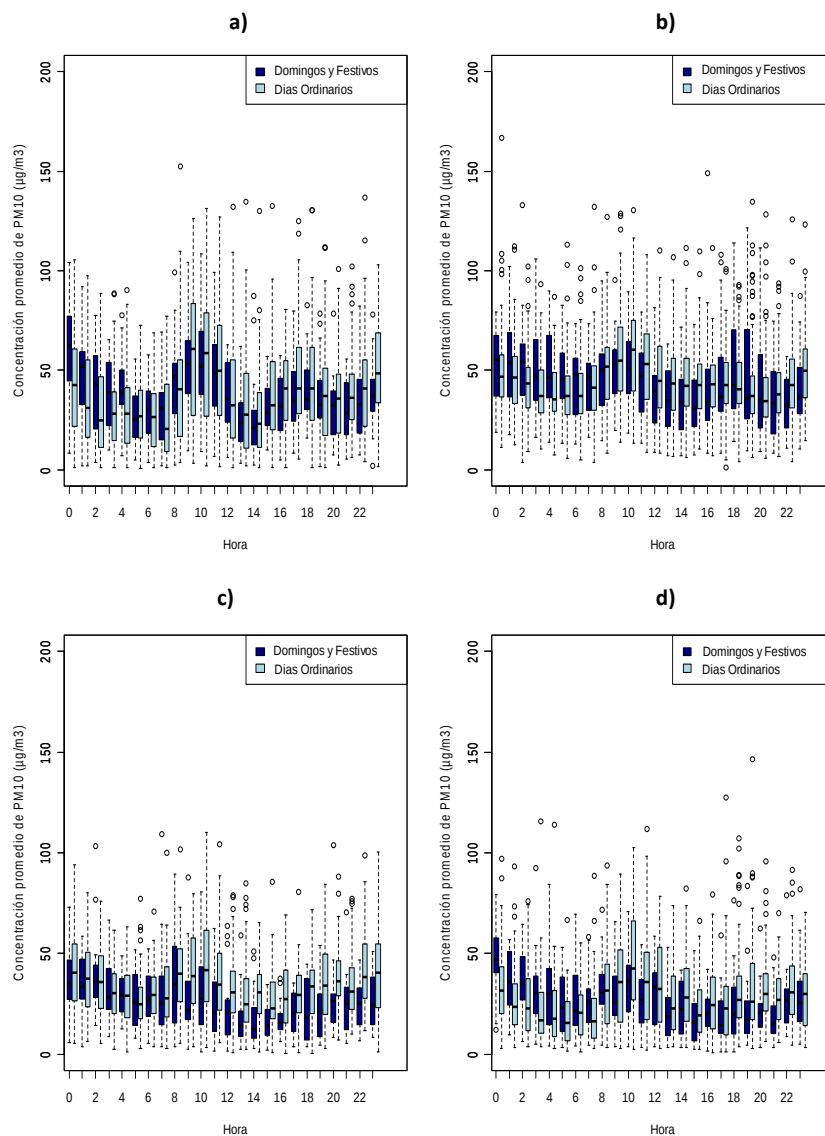


Figura 13: Concentración promedio (horaria) de PM10 en la estación PPD. (a) Año 2003, (b) Año 2004, (c) Año 2005, (d) Año 2006



ANEXO 3

Figura 14: Curvas de suavización y bandas de variabilidad para la estación CDAV. (a) Año 2005, (b) Año 2006

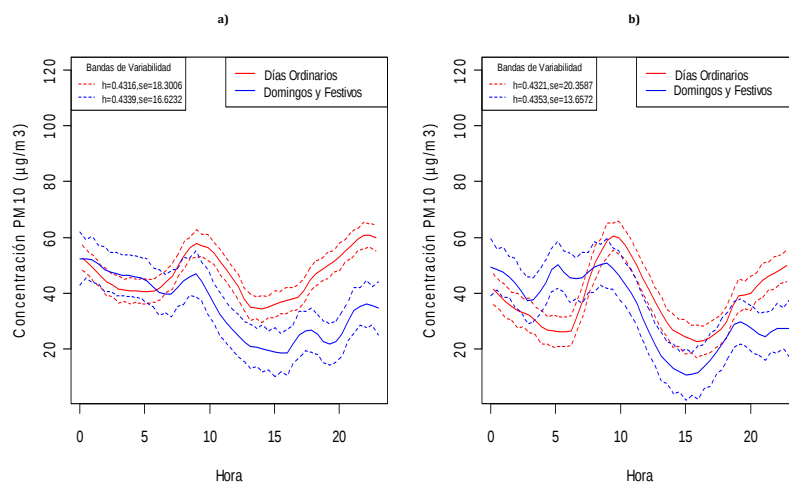


Figura 15: Curvas de suavización y bandas de variabilidad para la estación CLL15. (a) Año 2005, (b) Año 2006

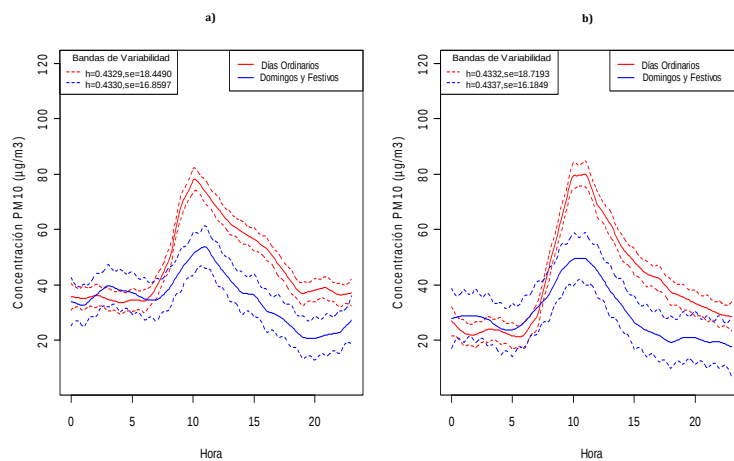


Figura 16: Curvas de suavización y bandas de variabilidad para la estación ERA. (a) Año 2003, (b) Año 2004, (c) Año 2005, (d) Año 2006

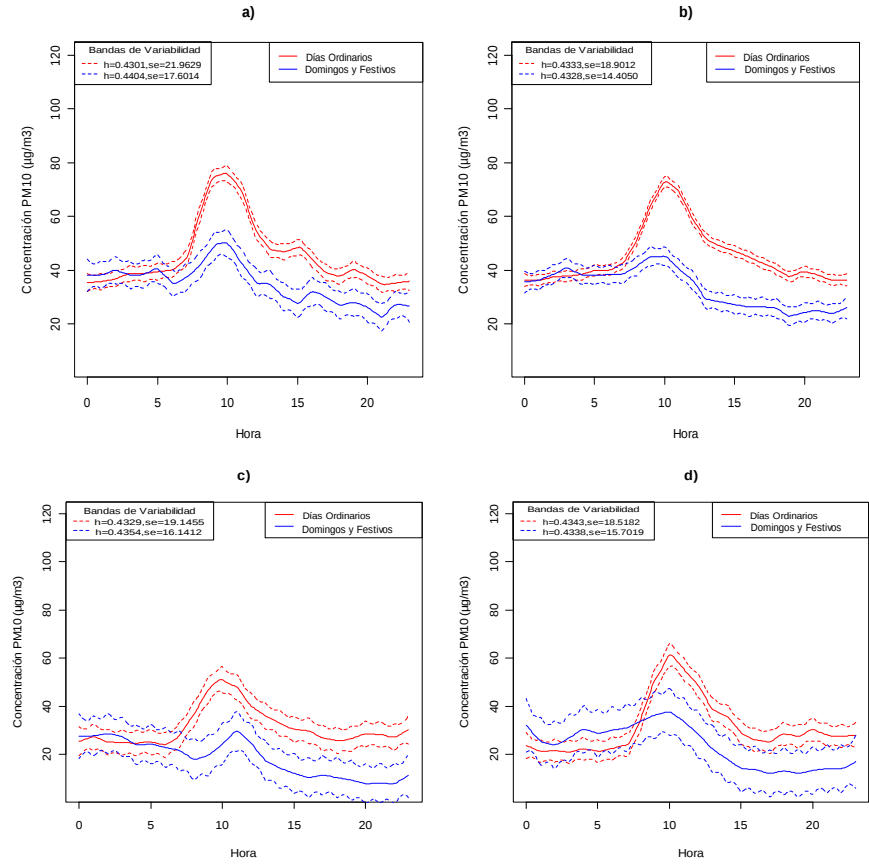


Figura 17: Curvas de suavización y bandas de variabilidad para la estación CVC. (a) Año 2003, (b) Año 2004, (c) 2005

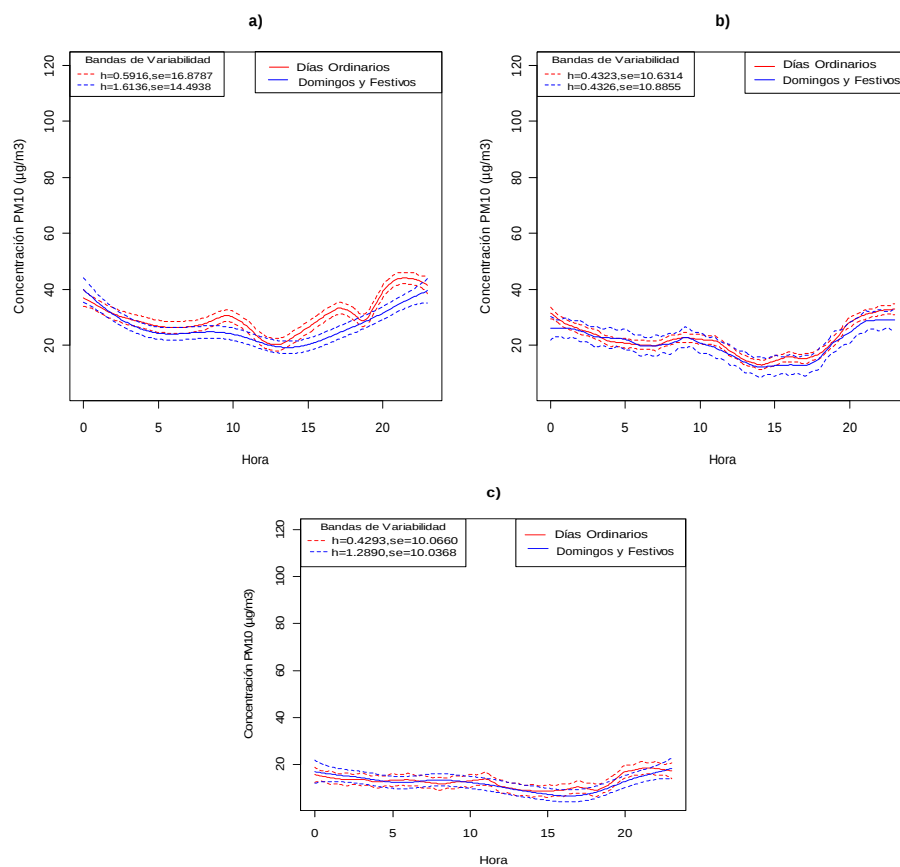


Figura 18: Curvas de suavización y bandas de variabilidad para la estación UV. (a) Año 2004, (b) Año 2005

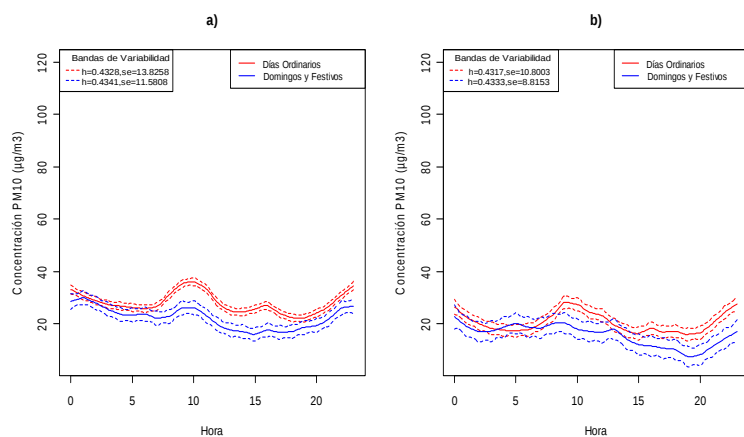


Figura 19: Curvas de suavización y bandas de variabilidad para la estación BA. (a) Año 2003, (b) Año 2004

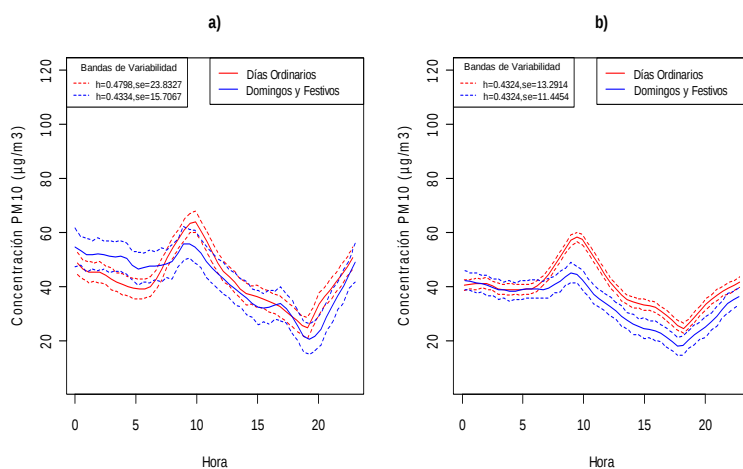
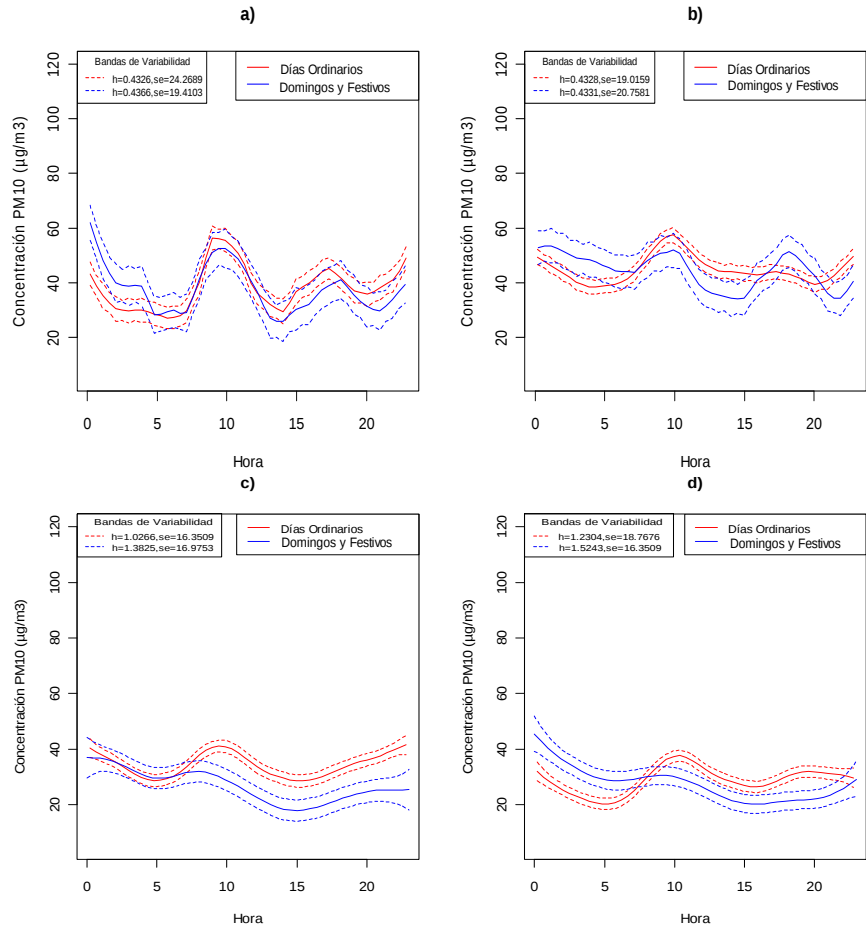
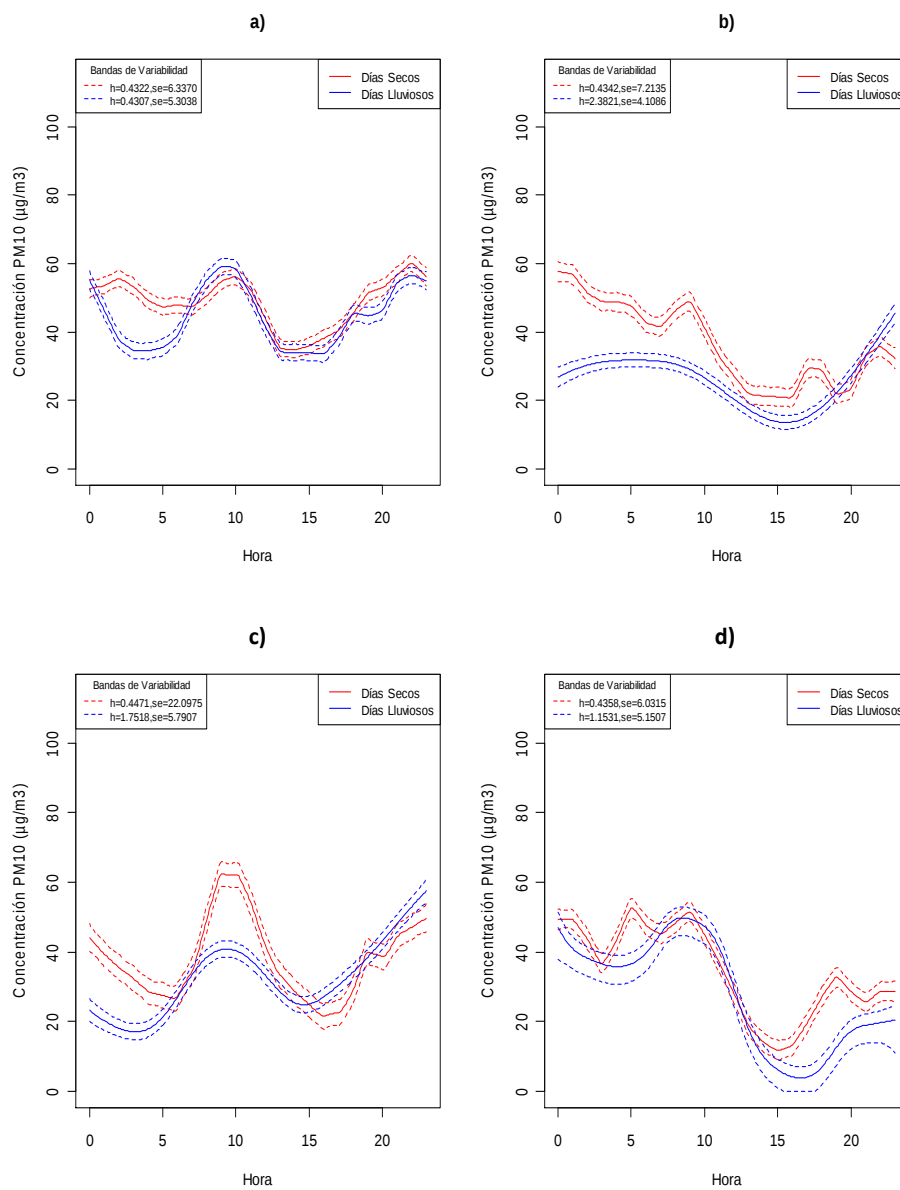


Figura 20: Curvas de suavización y bandas de variabilidad para la estación PPD. (a) Año 2003, (b) Año 2004, (c) Año 2005, (d) Año 2006



ANEXO 4

Figura 21: Curvas de suavización y bandas de variabilidad según clima en la estación CDAV. (a) Ordinarios-2005, (b) Domingos y festivos-2005, (c) Ordinarios-2006, (d) Domingos y festivos-2006



ANEXO 5

Figura 22: Residuales de las curvas de las concentraciones para la estación CDAV (2003). (a) Días Ordinarios, (b) Domingos y Festivos

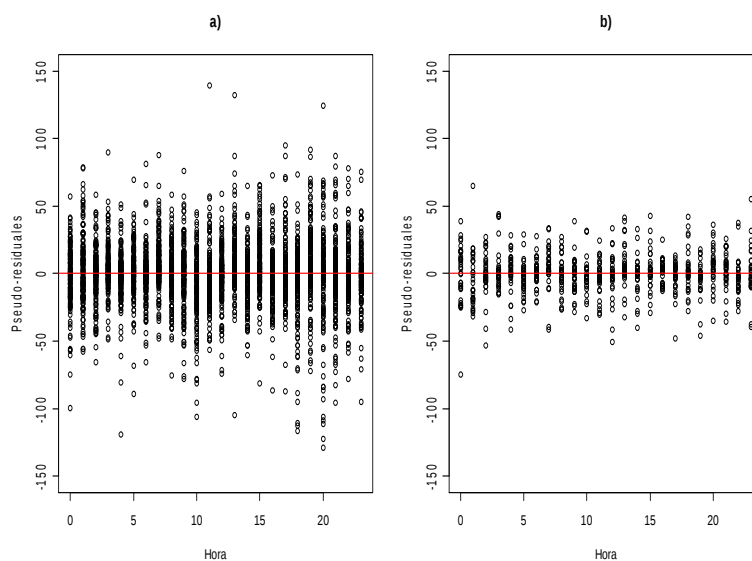


Figura 23: Residuales de las curvas de las concentraciones para la estación CDAV (2004). (a) Días Ordinarios, (b) Domingos y Festivos

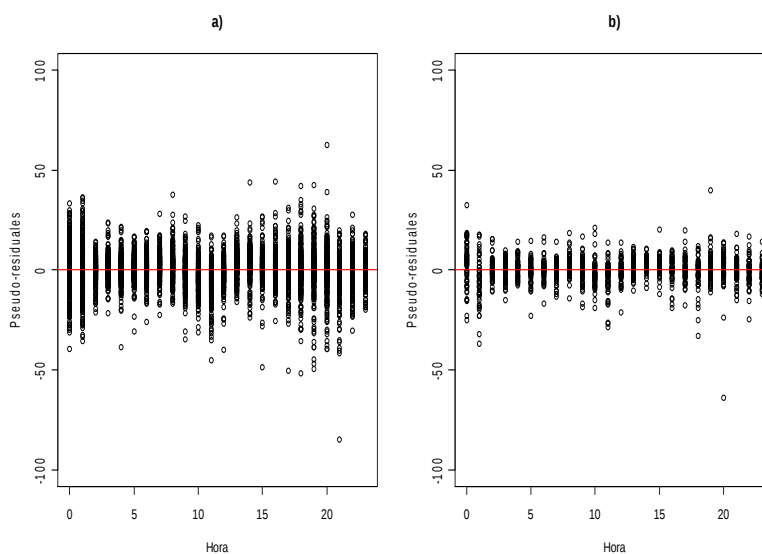


Figura 24: Residuales de las curvas de las concentraciones para la estación CDAV (2005). (a) Días Ordinarios, (b) Domingos y Festivos

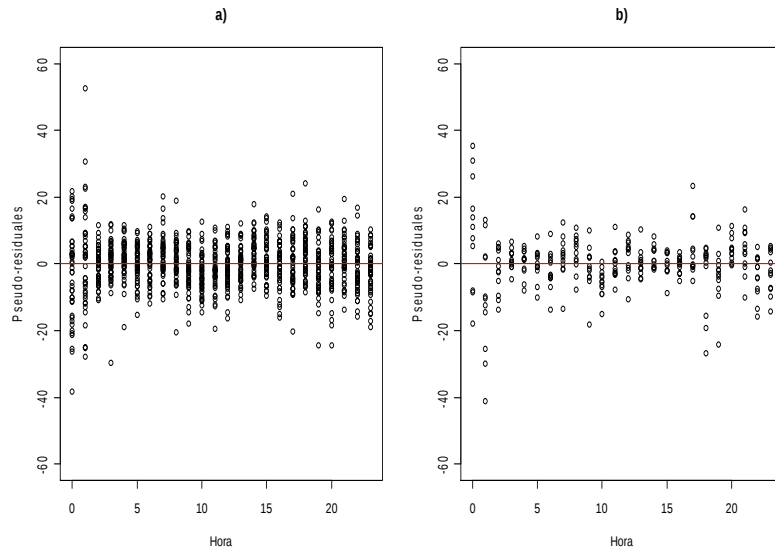


Figura 25: Residuales de las curvas de las concentraciones para la estación CDAV (2006). (a) Días Ordinarios, (b) Domingos y Festivos

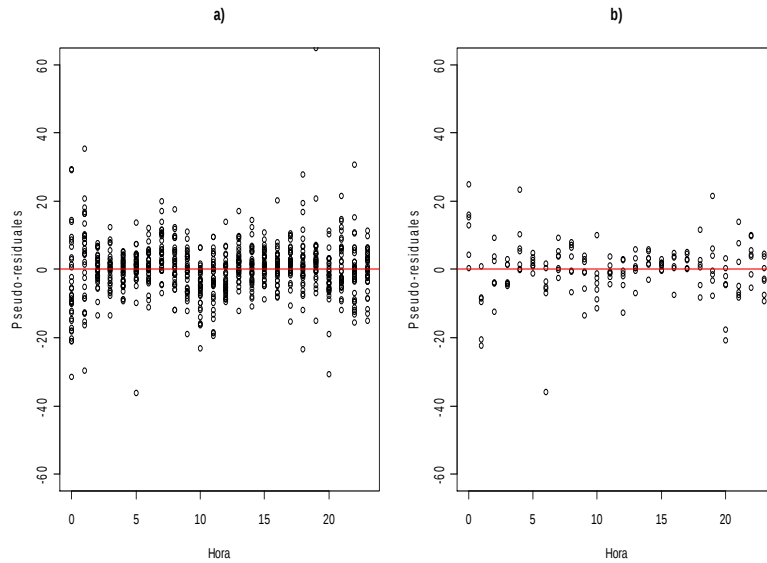


Figura 26: Residuales de las curvas de las concentraciones para la estación CLL15 (2003). (a) Días Ordinarios, (b) Domingos y Festivos

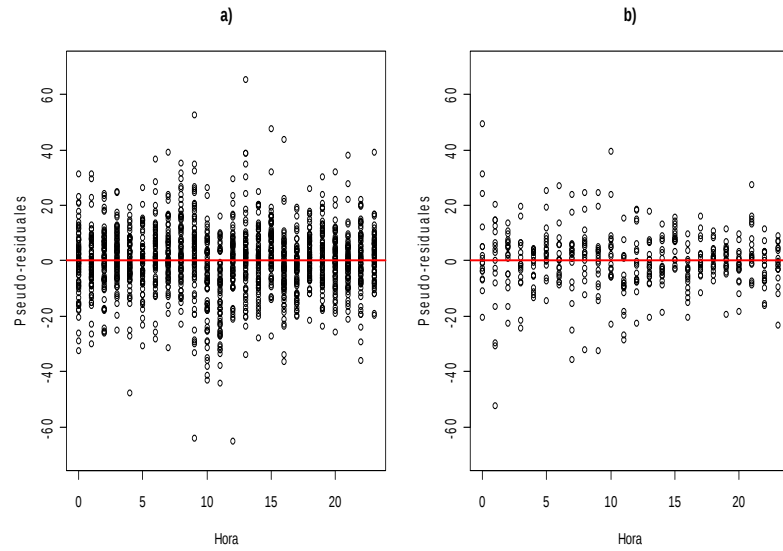


Figura 27: Residuales de las curvas de las concentraciones para la estación CLL15 (2004). (a) Días Ordinarios, (b) Domingos y Festivos

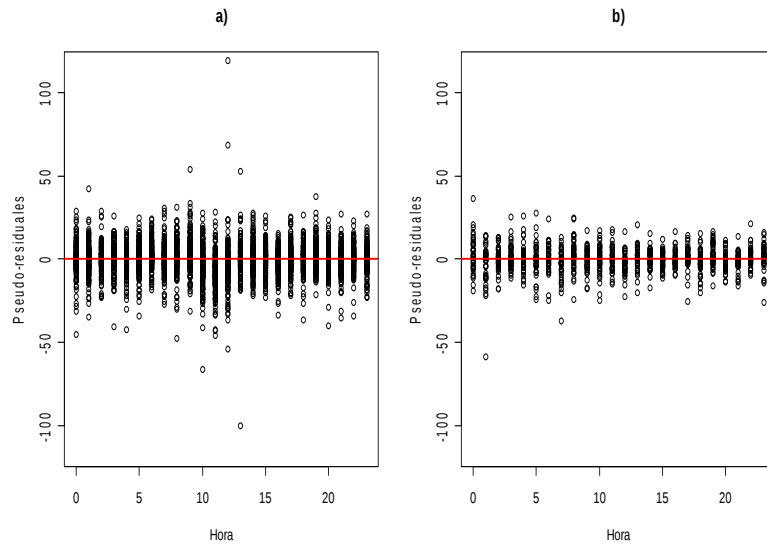


Figura 28: Residuales de las curvas de las concentraciones para la estación CLL15 (2005). (a) Días Ordinarios, (b) Domingos y Festivos

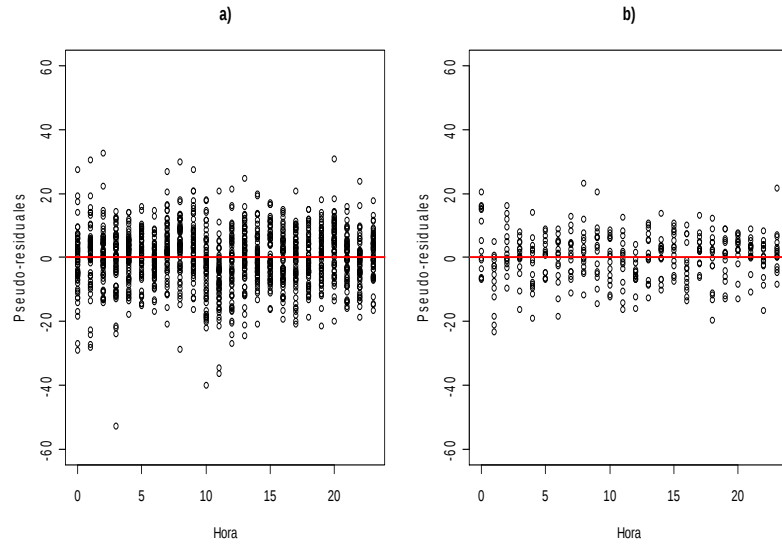


Figura 29: Residuales de las curvas de las concentraciones para la estación CLL15 (2006). (a) Días Ordinarios, (b) Domingos y Festivos

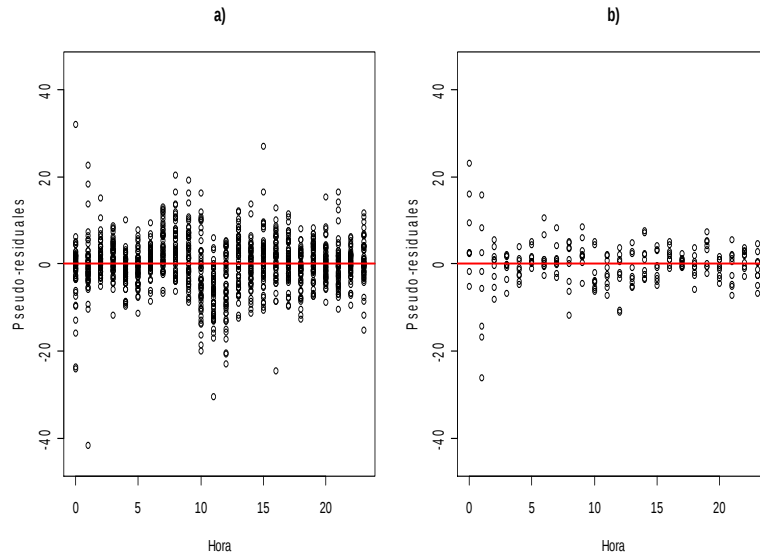


Figura 30: Residuales de las curvas de las concentraciones para la estación ERA (2003). (a) Días Ordinarios, (b) Domingos y Festivos

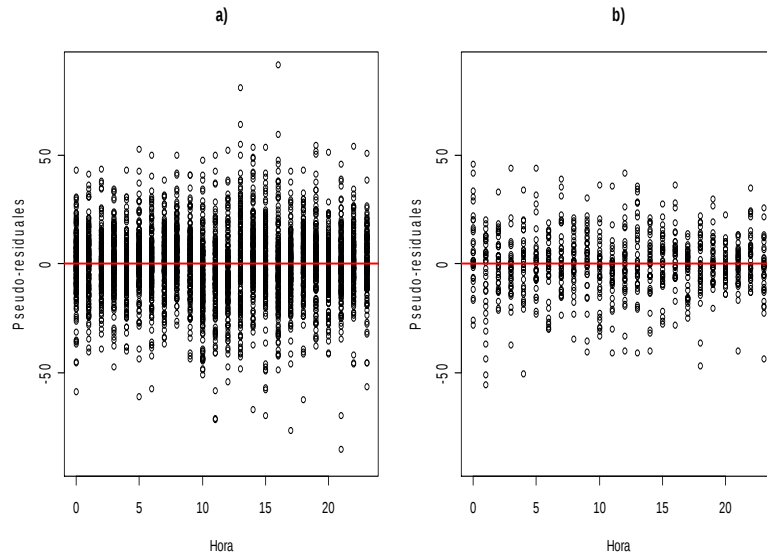


Figura 31: Residuales de las curvas de las concentraciones para la estación ERA (2004). (a) Días Ordinarios, (b) Domingos y Festivos

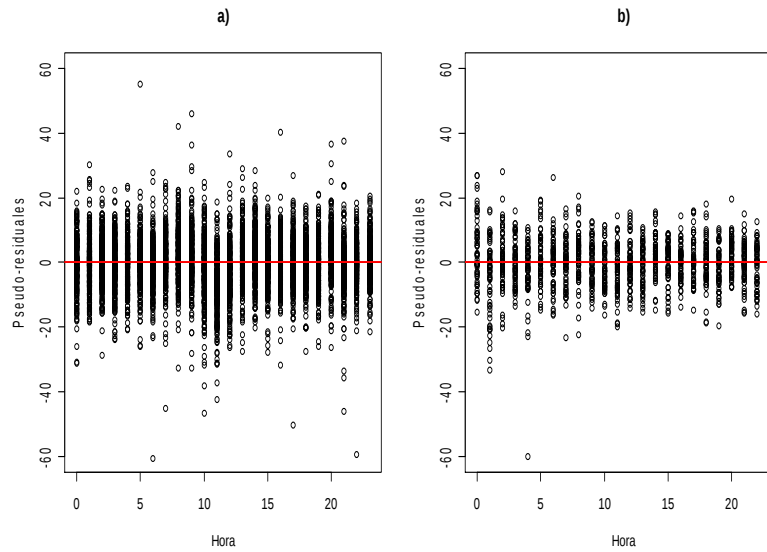


Figura 32: Residuales de las curvas de las concentraciones para la estación ERA (2005). (a) Días Ordinarios, (b) Domingos y Festivos

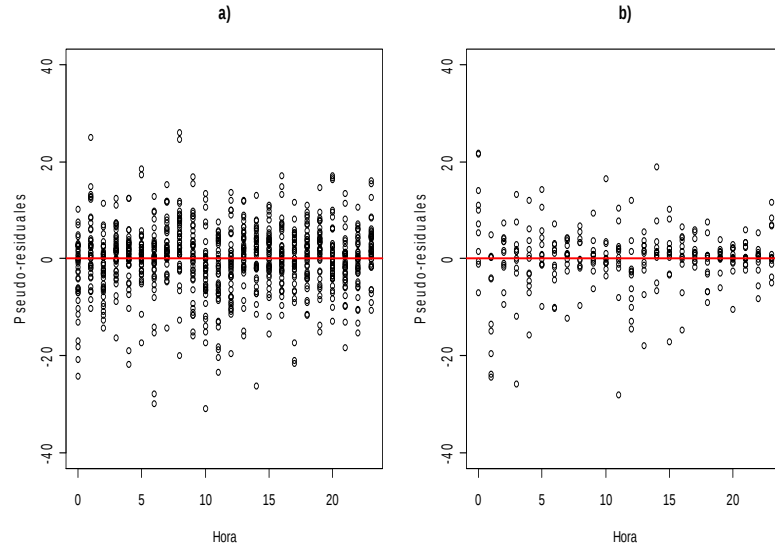


Figura 33: Residuales de las curvas de las concentraciones para la estación ERA (2006). (a) Días Ordinarios, (b) Domingos y Festivos

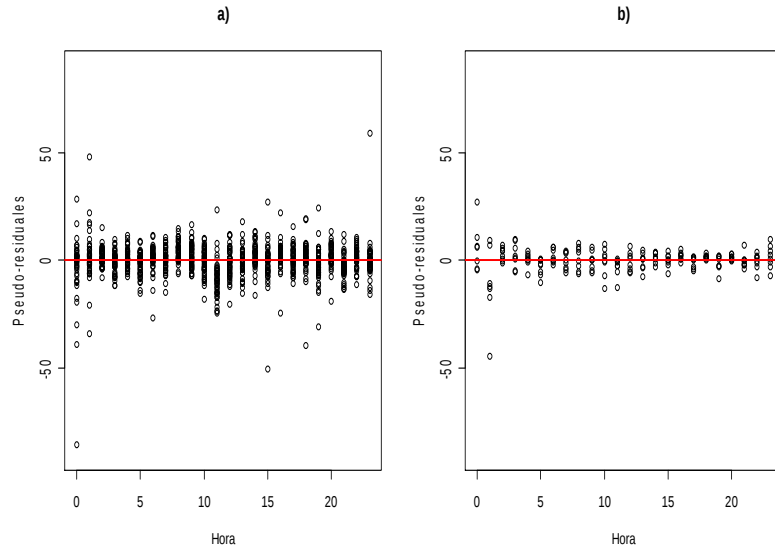


Figura 34: Residuales de las curvas de las concentraciones para la estación UV (2004). (a) Días Ordinarios, (b) Domingos y Festivos

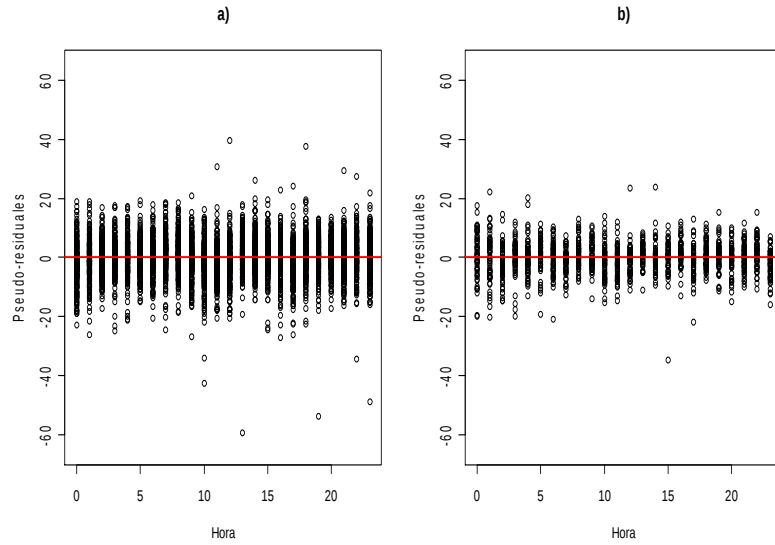


Figura 35: Residuales de las curvas de las concentraciones para la estación UV (2005). (a) Días Ordinarios, (b) Domingos y Festivos

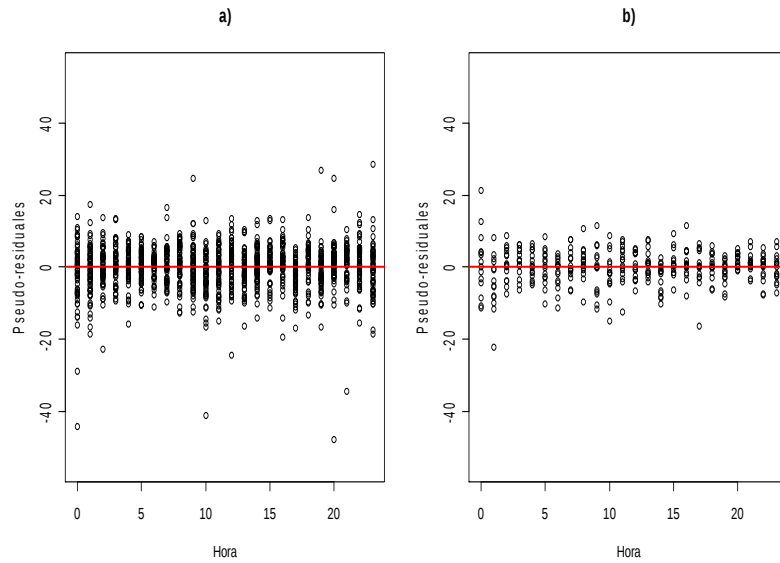


Figura 36: Residuales de las curvas de las concentraciones para la estación CVC (2003). (a) Días Ordinarios, (b) Domingos y Festivos

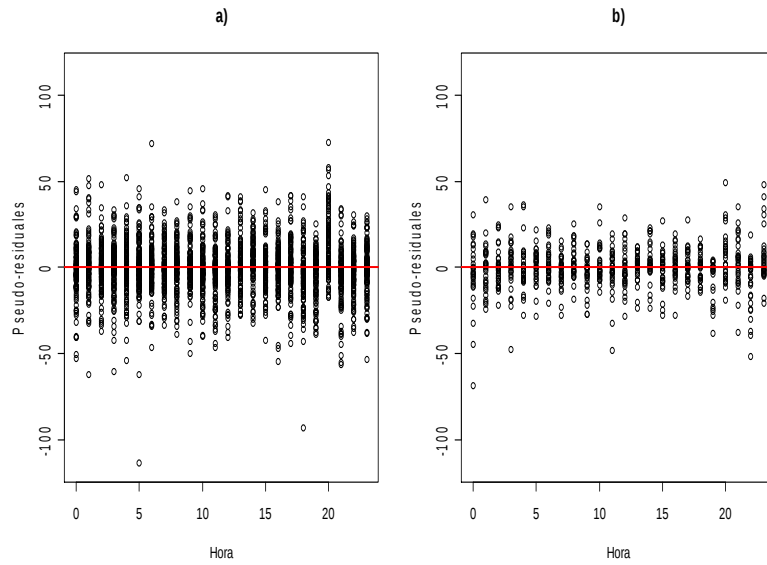


Figura 37: Residuales de las curvas de las concentraciones para la estación CVC (2004). (a) Días Ordinarios, (b) Domingos y Festivos

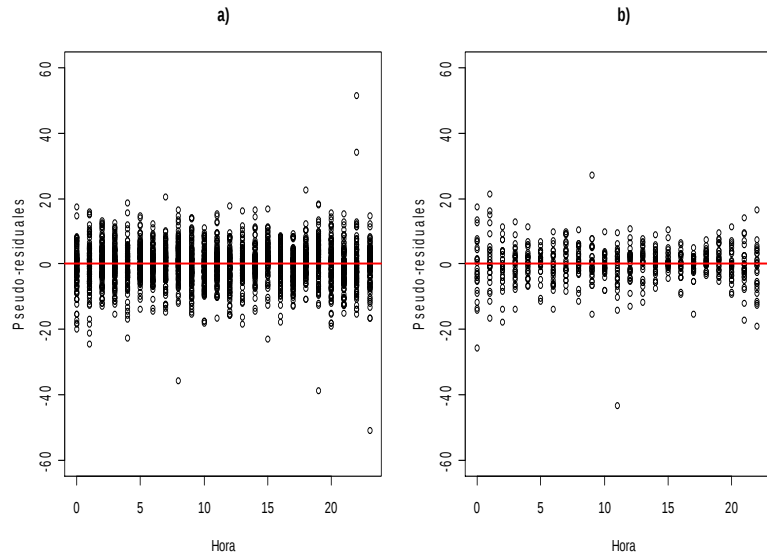


Figura 38: Residuales de las curvas de las concentraciones para la estación CVC (2005). (a) Días Ordinarios, (b) Domingos y Festivos

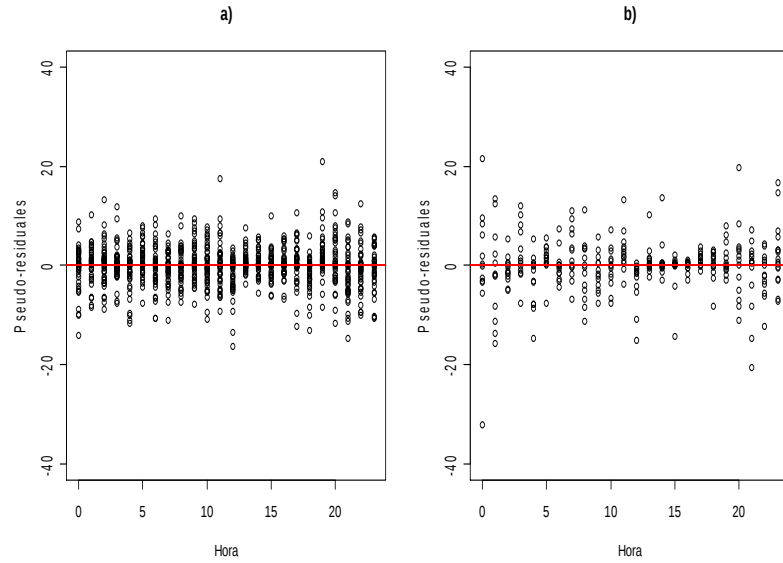


Figura 39: Residuales de las curvas de las concentraciones para la estación BA (2003). (a) Días Ordinarios, (b) Domingos y Festivos

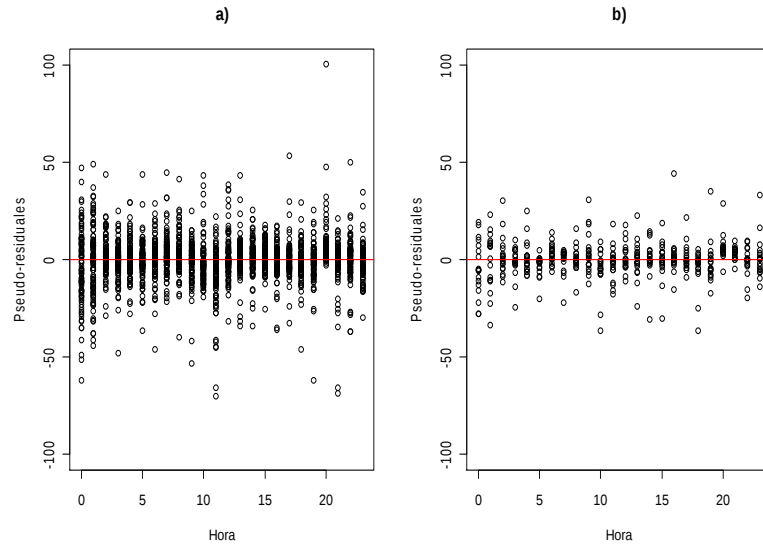


Figura 40: Residuales de las curvas de las concentraciones para la estación BA (2004). (a) Días Ordinarios, (b) Domingos y Festivos

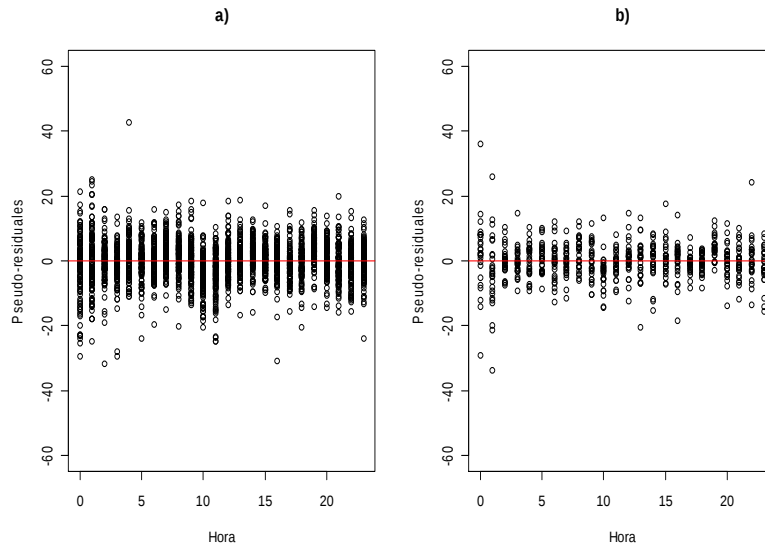


Figura 41: Residuales de las curvas de las concentraciones para la estación PPD (2003). (a) Días Ordinarios, (b) Domingos y Festivos

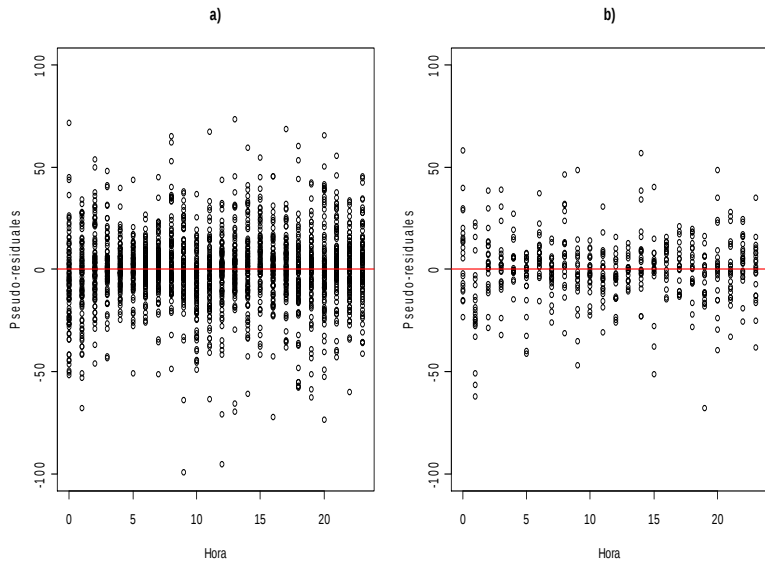


Figura 42: Residuales de las curvas de las concentraciones para la estación PPD (2004). (a) Días Ordinarios, (b) Domingos y Festivos

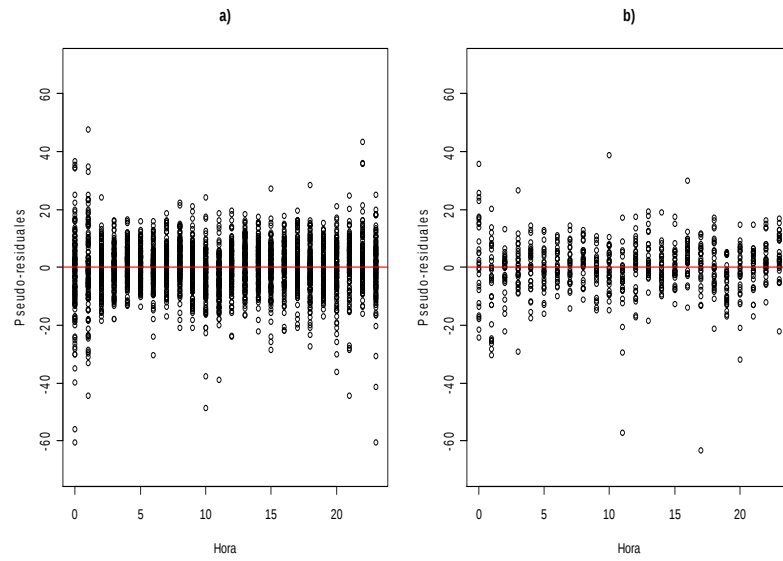


Figura 43: Residuales de las curvas de las concentraciones para la estación PPD (2005). (a) Días Ordinarios, (b) Domingos y Festivos

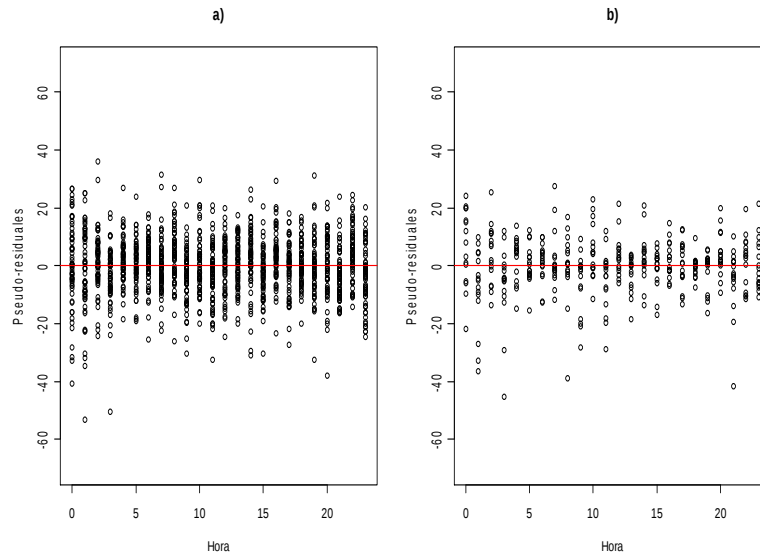


Figura 44: Residuales de las curvas de las concentraciones para la estación PPD (2006). (a) Días Ordinarios, (b) Domingos y Festivos

