

**MODELACIÓN DE LA CONTAMINACIÓN ATMOSFÉRICA MEDIANTE REGRESIÓN NO
PARAMÉTRICA TENIENDO EN CUENTA LA ESTRUCTURA DE CORRELACIÓN DE LOS
PROMEDIOS OCTOHORARIOS DEL MONÓXIDO DE CARBONO EN LA CIUDAD DE CALI EN
EL PERIODO 2003 - 2006**

PRESENTADO POR

**JENNY FERNANDA MINA VALENCIA
JOSÉ WARLEY MOSTACILLA HERRERA**

**UNIVERSIDAD DEL VALLE
FACULTAD DE INGENIERÍAS
ESCUELA DE INGENIERÍA INDUSTRIAL Y ESTADÍSTICA
PROGRAMA ACADÉMICO DE ESTADÍSTICA
SANTIAGO DE CALI
2012**

**MODELACIÓN DE LA CONTAMINACIÓN ATMOSFÉRICA MEDIANTE REGRESIÓN NO
PARAMÉTRICA TENIENDO EN CUENTA LA ESTRUCTURA DE CORRELACIÓN DE LOS
PROMEDIOS OCTOHORARIOS DEL MONÓXIDO DE CARBONO EN LA CIUDAD DE CALI EN
EL PERIODO 2003 - 2006**

Trabajo de grado presentado por

JENNY FERNANDA MINA VALENCIA y JOSÉ WARLEY MOSTACILLA HERRERA

**Para optar por el título profesional de
Estadístico**

**Director del Proyecto
ÁLVARO JOSÉ FLÓREZ
Estadístico**

**Codirector del Proyecto
RODRIGO RADA BARONA
Estadístico**

**UNIVERSIDAD DEL VALLE
FACULTAD DE INGENIERÍAS
ESCUELA DE INGENIERÍA INDUSTRIAL Y ESTADÍSTICA
PROGRAMA ACADÉMICO DE ESTADÍSTICA
SANTIAGO DE CALI
2011**

DEDICATORIA Y AGRADECIMIENTOS

Jenny Fernanda Mina Valencia

“A DIOS TODOPODEROSO por haberme dado la sabiduría, la fortaleza y el entendimiento para poder llegar al final de mi carrera y fuera posible alcanzar este triunfo. A mis padres Marino y Dora por su cariño, su apoyo incondicional, su dedicación, empeño y sacrificio que hicieron a lo largo de mi carrera, por ayudarme a ser una persona mejor cada día, por tanto esfuerzo para que yo alcanzara este triunfo. A mi hermano Jhojan por su cariño, apoyo, comprensión y ser mi ejemplo a seguir. A mi mamá Nery por ser mi guía espiritual, y por su confianza de siempre. A mi compañero de tesis Warley porque a pesar de todos los momentos difíciles que tuvimos pudimos salir adelante con nuestro trabajo de grado. A mis familiares por estar pendientes a lo largo de este proceso, brindado su apoyo incondicional. A mis amigos por todos los momentos compartidos a lo largo de la carrera, en especial a Yadira y Vanessa por ser mis hermanas y amigas. A la familia de odontología Univalle por su colaboración. A mis profesores de la universidad del Valle por los conocimientos compartidos y enseñados para mi desarrollo profesional, especialmente Álvaro Flores, Rodrigo Rada por ser más que profesores a amigos. A Luis Carlos por volverme a enseñar a creer que hay un futuro mejor. Y a todas las personas que contribuyeron a que esta meta se convierta en una realidad”

José Wvarley MostacillaHerrera

A Dios por entregarme todas las herramientas necesarias para cumplir este gran sueño. A mis padres Amparo y José por su apoyo incondicional, su empeño y sacrificio que hicieron a lo largo de mi carrera. A mi hermana carolina por ser la persona que me inspira para darle un gran ejemplo. A mi esposa Johana por su apoyo incondicional, entenderme y estar siempre a mi lado. A mi hija Ana Catalina por se mi motivo e inspiración para esforzarme todos los día. A Jenny mi compañera de tesis porque a pesar de todos los momentos difíciles que tuvimos pudimos salir adelante con nuestro trabajo de grado. A mis familiares por estar pendientes a lo largo de este proceso, brindado su apoyo incondicional. A mis amigos por todos los momentos compartidos a lo largo de la carrera, en especial a Yadira y Vanessa.

De manera muy especial expreso mi agradecimiento a los profesores Álvaro Flores quien fue nuestro director y guía en la elaboración de este trabajo, estando siempre atento para resolver nuestras dudas e inquietudes, al profesor Rodrigo Rada quien fue, pilar fundamental con sus consejos, sugerencias y recomendaciones para nuestro trabajo profesor Javier Olaya quien fue en principio quien nos brindó el privilegio de trabajar a su lado y además de manera muy generosa nos proporcionó la base de datos, al profesor Roberto Behar que con sus consejos y sugerencias nos dio una mano, al profesor y director del programa académico Rafael Klinger por su apoyo incondicional. Y en general a todos los profesores que en el transcurso de la carrera aportaron a mi formación como personal y profesional.

TABLA DE CONTENIDO

INTRODUCCION.....	9
1. PLANTEAMIENTO DEL PROBLEMA.....	11
1.1. Pregunta de investigación	13
2. JUSTIFICACIÓN DEL PROBLEMA.....	14
3. OBJETIVOS.....	16
3.1. Objetivo General.....	16
3.2. Objetivos Específicos	16
4. ESTADO DEL ARTE.....	17
5. MARCO TEORICO.....	20
5.1. Contaminación Atmosférica	20
5.1.1. Fuentes contaminantes del aire	20
5.2. Monóxido de Carbono (CO)	21
5.2.1. fuentes de emisión de CO	22
5.3. Normas de Calidad de Aire	22
5.3.1. generalidades de santiago de cali	23
5.3.2. Red de monitoreo de la calidad del aire en Cali	23
5.3.3. protocolo de monitoreo	24
5.3.4. Concentración octohoraria	26
5.4. El Modelo de Regresión	27
5.4.1. modelo de regresión no paramétrica con errores correlacionados...	27
5.4.2. suavización.....	28
5.4.3. estimador polinómico local con funciones kernel.....	30
5.4.4. Parámetro de suavización “h”	33
5.4.5. Selección del parámetro de suavizado <i>h</i>	35
5.4.6. Estimación de la varianza	39

5.4.7. Intervalos de confianza y bandas de variabilidad	40
6. METODOLOGÍA DE INVESTIGACIÓN.....	42
6.1. revisión bibliográfica	42
6.2. Recopilación de datos	42
6.3. Software estadístico.....	44
6.4. Análisis exploratorio	44
6.5. Estimación de curvas típicas	44
6.6. Validación de supuestos	46
6.7. Validación del comportamiento de los selectores de ancho de banda.....	46
7. RESULTADOS	47
7.1. ANÁLISIS EXPLORATORIO DE LAS BASES DE DATOS	47
7.1.1. Análisis descriptivo para la estación Calle 15 – Cl.15	47
7.1.2. Análisis descriptivo para la estación Escuela República de Argentina - ERA 49	
7.2. Estimación de curvas típicas	50
7.3. bandas de variabilidad	55
7.4. validación de supuestos.....	60
7.5. Validación del desempeño de los selectores de ancho de banda.....	60
8. CONCLUSIONES Y RECOMENDACIONES.....	64
Bibliografía	67
ANEXOS	71

LISTADO DE TABLAS

Tabla 1. Fuentes de contaminación de la atmósfera.....	21
Tabla 2. Norma de Calidad del Aire para el Monóxido de Carbono.....	22
Tabla 3. Estaciones y parámetros medidos en la RMCA (Cali).....	25

Tabla 4. Funciones Kernel de Segundo Orden.....	29
Tabla 5. Numero de Semanas (Tomadas – Eliminadas - Totales) por estación y año.	43
Tabla 6. Estadísticas descriptivas de las concentraciones promedio octohorarias de monóxido de carbono (8h-CO) para la estación calle 15 durante los años 2003 al 2006.	48
Tabla 7. Estadísticas descriptivas de las concentraciones promedio octohorarias de monóxido de carbono (8h-CO) para la estación ERA durante los años 2003 al 2006.....	49
Tabla 8. Anchos de banda obtenidos en la estimación de la curva de regresión no paramétrica para una semana típica utilizando diferentes métodos de selección.....	50
Tabla 9. Anchos de banda obtenidos al aplicar diferentes métodos de selección de ancho de banda en la estimación de la curva de regresión no paramétrica para una semana típica.....	63

LISTADO DE FIGURAS

Figura 1. Perfil fisiográfico de Santiago de Cali.....	23
Figura 2. Funciones Kernel de Segundo Orden	30
Figura 3. Diagrama de cajas y alambre de concentraciones octohorarias (8-h) para la estación calle 15 durante los años 2003 al 2006.....	48
Figura 4. Diagrama de cajas y alambre de concentraciones octohorarias (8-h) para la estación ERA durante los años 2003, 2004 y 2006.....	49
Figura 5. Funciones de autocorrelación estimadas de las suavizaciones preliminares de los residuales del modelo de regresión no paramétrico de las concentraciones octohorarias (8-h) para una semana típica en la estación Calle 15 en los años 2003, 2004, 2005 y 2006	51
Figura 6. Estimación de las curvas para las concentraciones octohorarias (8-h) de CO utilizando los selectores de ancho de banda (CV, DPI, GCV y ROT) para una semana típica en la estación Calle 15 en los años 2003 y 2004	52
Figura 7. Funciones de autocorrelación estimadas de las suavizaciones preliminares de los residuales del modelo de regresión no paramétrico de las concentraciones octohorarias (8-h) para una semana típica en la estación ERA en los años 2003, 2004 y 2006	54
Figura 8. Estimación de las curvas típicas para las concentraciones octohorarias (8-h) de CO utilizando los selectores de ancho de banda (CV, DPI, GCV y ROT) para una semana típica por año en la estación ERA.....	54
Figura 9. Bandas de Variabilidad para las curvas para las concentraciones octohorarias (8-h) estimadas mediante los selectores de ancho de banda GCVd y ROT para una semana típica del año 2003 en la estación Calle 15	55

Figura 10. Bandas de Variabilidad para las curvas para las concentraciones octohorarias (8-h) estimadas mediante los selectores de ancho de banda GCV y ROT para una semana típica del año 2006 en la estación Calle 15.....	56
Figura 11. Bandas de Variabilidad para las curvas para las concentraciones octohorarias (8-h) estimadas mediante los selectores de ancho de banda GCV y ROT para una semana típica del año 2004 en la estación ERA.....	57
Figura 12. Bandas de Variabilidad para las curvas para las concentraciones octohorarias (8-h) estimadas mediante los selectores de ancho de banda GCVd y ROT para una semana típica del año 2003 en la estación PDD	57
Figura 13. Bandas de Variabilidad para las curvas para las concentraciones octohorarias (8-h) estimadas mediante los selectores de ancho de banda GCVd y ROT para una semana típica del año 2003 en la estación HUV.....	58
Figura 14. Bandas de Variabilidad para las curvas para las concentraciones octohorarias (8-h) estimadas mediante los selectores de ancho de banda GCVd y ROT para una semana típica del año 2003 en la estación BA.....	59
Figura 15. Bandas de Variabilidad para las curvas para las concentraciones octohorarias (8-h) estimadas mediante los selectores de ancho de banda GCVd y ROT para una semana típica del año 2004 en la estación CDVA.....	59
Figura 16. Funciones de autocorrelación estimadas para los residuales de la suavización preliminar, con registros completos tomados durante una semana típica.	61
Figura 17. Funciones de autocorrelación estimadas para los residuales de la suavización preliminar, con registros incompletos tomados durante una semana típica.	62

LISTADO DE ANEXOS

Anexo A: Diagramas de cajas y alambres para las estaciones PDD, BA, CDVA y HUV de las concentraciones octohorarias para el monóxido de carbono.	71
Anexo B: Funciones de autocorrelación estimadas de las suavizaciones preliminares de los residuales del modelo de regresión no paramétrico de las concentraciones octohorarias (8-h) para una semana típica en las estaciones PDD, BA, CDVA y HUV.....	73
Anexo C: Estimación de las curvas típicas para las concentraciones octohorarias (8-h) de CO utilizando los selectores de ancho de banda (CV, DPI, GCV y ROT) para una semana típica por año en las estaciones PDD, BA, CDVA y HUV.....	74
Anexo D: Bandas de Variabilidad para las curvas para las concentraciones octohorarias (8-h) estimadas mediante los selectores de ancho de banda GCVd y ROT para una semana típica en las estaciones PDD, BA, CDVA y HUV.	76

Anexo E: Gráficos de Residuales para las concentraciones octohorarias para el monóxido de carbono en la ciudad de Cali para los años 2003 al 2006.....	78
Anexo F: Estimación de las curvas para diferentes selectores de ancho de banda de las estaciones PDD, BA, CDVA Y HUV de las concentraciones octohorarias para el monóxido de carbono en la ciudad de cali.....	82

INTRODUCCION

La contaminación del aire a diferencia de otros problemas ambientales puede percibirse con facilidad, especialmente en zonas donde se presenta gran concentración urbana. La calidad del aire en el municipio de Santiago de Cali, ha sido estudiada desde el año 1985 por la Secretaría de Salud y el Centro de Diagnóstico Automotor. En la actualidad la Corporación Autónoma Regional del Valle del Cauca (CVC) y el Departamento Administrativo de Gestión del Medio Ambiente (DAGMA) son los encargados de dirigir estudios encaminados a explorar la situación actual de la ciudad en materia ambiental¹.

El monóxido de carbono (CO) es un contaminante primario. Esto quiere decir que es emitido directamente de una fuente al aire. Su estabilidad en la atmósfera es significativa, lo cual indica que no reacciona con otros compuestos. En consecuencia, el CO está sujeto a procesos acumulativos en el aire como los que ocurren en el invierno durante situaciones de vientos débiles e inversión térmica² elevada, lo que constituye un riesgo para la población que se expone a respirar un aire más contaminado de lo normal (La Cisterna, 2005). Entre las principales fuentes de emisión de CO se encuentran los procesos industriales, las fuentes móviles (automóviles, buses, motos, etc.), aparatos de combustión (calentadores, calefactores, estufas, hornillas de la cocina o los calentadores a kerosene) y fuentes naturales como los incendios forestales (Reina, 2009).

Diversos estudios han revelado que los efectos tóxicos en personas expuestas a altos niveles de CO dependen de la duración a la exposición. Exposiciones prolongadas a altos niveles de CO traen como consecuencia, graves enfermedades y daños a la fisiología humana. Una medida estándar que utiliza la Agencia de Protección Ambiental de Estados Unidos (EPA, por sus siglas en inglés)³ para determinar los riesgos de la exposición prolongada al CO es el promedio del contaminante durante un periodo de ocho horas, el cual se denomina Promedio Octohorario (8-h)⁴. (Reina, 2009).

En Colombia, las autoridades ambientales han determinado que los niveles de prevención, alerta y emergencia dependen de la concentración y del tiempo de exposición a los Contaminantes. En el caso del monóxido de carbono (CO) se realizan mediciones de su concentración promedio durante periodos de una hora (1-h) y ocho horas (8-h) para determinar los riesgos que puede sufrir la población debido a la exposición prolongada a este

¹ Tomado de la página oficial de la alcaldía de Santiago de Cali. Recuperado el 27 de diciembre de 2010 en <http://www.cali.gov.co/sil/sil.php?id=1437&letra=E>

² Inversión Térmica: Fenómeno atmosférico natural que se produce cuando una capa de aire caliente queda aprisionada entre dos capas de aire frío, impidiendo la libre circulación atmosférica y permitiendo que se estacione a nivel del suelo. En estas condiciones los contaminantes atmosféricos se desplazan horizontalmente y no en sentido vertical lo cual hace que se estanquen y llegan a producir concentraciones altas en el aire, generando condiciones poco saludables para los habitantes de zonas urbanas. Tomado el 28 de febrero de 2010 de la página web www.peruecologico.com.pe/glosario_i.htm

³ Environmental Protection Agency, EPA.

⁴ Se entiende por concentración promedio octohoraria al promedio móvil simple de los datos de las concentraciones horarias de un contaminante en un periodo ocho horas.

contaminante. La metodología establecida para registrar la concentración promedio octohoraria, establece que su cálculo se debe realizar por medio del método de las medias móviles simples⁵, para ello se utilizan los registros de las concentraciones horarias de CO en un periodo ocho horas. Este procedimiento ha generado una interrogante sobre el efecto que puede llegar a tener la estructura de autocorrelación en los modelos utilizados para estimar el comportamiento de este contaminante.

De acuerdo a lo anterior es fundamental investigar a profundidad este fenómeno ambiental haciendo uso de los métodos estadísticos que conlleven a caracterizar y cuantificar la calidad del aire que se respira. (EPA, 2000).

En esta investigación se pretende modelar la contaminación del aire por Monóxido de Carbono (CO) en la ciudad de Santiago de Cali en el periodo 2003 – 2006, mediante un modelo de regresión no paramétrica que tenga en cuenta la estructura de correlación que presentan los promedios octohorarios de este contaminante, mas no estimar la función de correlación en sí. Para esto se empleará un método de selección del parámetro de suavizado (h) que incluya la estructura de autocorrelación de los errores que presenta el modelo. Puesto que en presencia de datos correlacionados, dichos métodos de selección se ven seriamente afectados a la hora de elegir un valor para (h) debido a que estos tienden a ser muy pequeños. Lo cual conlleva a que la estimación de la curva sea poco suave pues interpolaría los datos de acuerdo con lo expuesto por (Hart, 1991). Para esto se usarán los registros de los niveles de concentración horaria de monóxido de carbono, suministrados por la Red de Monitoreo de Calidad de Aire (RMCA) de la ciudad.

⁵ Media móvil: Se calcula del mismo modo que la media aritmética para un periodo de n datos y se va recalculando a medida que se agregan nuevos datos, partiendo del último dato agregado y manteniendo siempre el número de datos correspondiente al periodo definido. El periodo definido podrá ser un año, un día, ocho horas o tres horas. Tomado de la Resolución 0601 año 2006 Ministerio de Ambiente, Vivienda y Desarrollo Territorial

1. PLANTEAMIENTO DEL PROBLEMA

La contaminación del aire es actualmente uno de los problemas ambientales más severos a nivel mundial. Está presente en todas las sociedades, independiente al grado de desarrollo socioeconómico y constituye un fenómeno que tiene particular incidencia sobre la salud del hombre. En las últimas décadas, los países industrializados han generado y emitido millones de toneladas de contaminantes a la atmósfera, esta situación ha llevado a un deterioro paulatino en la calidad del aire que respiramos día a día. (Astromia.com, s.f.)

La combustión parcial de material orgánico produce monóxido de carbono (CO), el cual es considerado uno de los mayores contaminantes de la atmósfera terrestre. Las principales fuentes productoras de este contaminante son los vehículos automotores que utilizan como combustible gasolina o diesel y los procesos industriales. Para el año 1997 las fuentes móviles (automóviles) aportaban el 76% de las emisiones de CO, según lo registrado por Sainco Tráfico S.A.⁶, para el año 2006 de acuerdo con lo expresado por la Ministra de Ambiente, Vivienda y Desarrollo Territorial de Colombia, Sandra Suarez, esta cifra ascendió al 86%.

Actualmente las agencias internacionales como la EPA indican que los automotores que utilizan combustibles fósiles, emiten alrededor del 80% del total de emisiones de CO en el mundo. La incineración de materia orgánica, los incendios forestales y urbanos complementan el cuadro de las principales fuentes de propagación. Los efectos tóxicos que causa esta sustancia en el ser humano, han sido estudiados ampliamente. Sin embargo, sus potenciales efectos adversos a largo plazo son poco conocidos.

En los últimos años, los estudios de investigación experimental en animales y epidemiológicos en humanos, han evidenciado relación entre población expuesta en forma crónica a niveles medios y bajos de monóxido de carbono en el aire respirable, al tiempo que la aparición de efectos adversos en la salud humana, especialmente en órganos de alto consumo de oxígeno como cerebro y corazón.

Se han documentado problemas neuropsicológicos y efectos nocivos en el sistema cardiovascular en ser humano aun cuando las concentraciones de monóxido de carbono en aire son inferiores a 25 partes por millón. Las alteraciones cardiovasculares descritas abarcan patologías entre las que se incluyen hipertensión arterial, arritmias y signos electrocardiográficos de isquemia, déficit de memoria y atención; concentración y alteraciones del movimiento tipo parkinsonismo, asociados a exposición crónica a bajos niveles de monóxido de carbono y carboxihemoglobina.⁷ (Téllez, et al., 2006)

Como se puede apreciar, la contaminación atmosférica por monóxido de carbono es un fenómeno complejo, que puede llegar a ocasionar daños severos en el medio ambiente y en la

⁶Fuente: Sainco Tráfico S.A es una empresa europea encargada de integración de sistemas de control de información y de redes privadas e infraestructura para los mercados verticales de energía, medio ambiente, transporte y telecomunicaciones.

⁷La carboxihemoglobina es la unión de hemoglobina y monóxido de carbono el cual toma el lugar del oxígeno en los glóbulos rojos.

salud de las personas. Ante tal situación, no es suficiente sólo adoptar medidas correctivas orientadas al cumplimiento de las normas gubernamentales que garanticen la sostenibilidad ambiental. Se hace necesario implementar un mecanismo de seguimiento más detallado de aquellos contaminantes que pueden llegar a afectar la calidad del aire y por consiguiente la salud de la población, permitiendo de esta manera a las autoridades ambientales tomar medidas preventivas.

Para el año 2009, se modeló el comportamiento de los promedios de las concentraciones de monóxido de carbono en periodos de ocho horas (8-h) en la ciudad de Cali mediante la aplicación de técnicas de suavización no paramétricas. Este estudio reveló que durante el periodo 2003 - 2006 el centro de la ciudad de Santiago de Cali, presentó los niveles más altos de contaminación por CO, atribuyendo este fenómeno a las fuentes móviles. Sin embargo, las emisiones registradas en este periodo no superaron las normas ambientales vigentes en Colombia y aceptadas por la EPA. Se pudo establecer que los fines de semana presentaban niveles de concentración más bajos que en los días lunes y martes-viernes a partir de las 9:00 a.m. También encontró que a lo largo de un día ordinario, la mayor concentración 8-h de CO se da en horas de la tarde, aproximadamente entre las 3:00 p.m. y 4:00 p.m. Mientras que, la mayor concentración promedio hora (1-h) de CO se presentó cerca de las 9:00 a.m. (Reina, 2009). Gracias a esta investigación se identificó algunos parámetros importantes del comportamiento de las concentraciones octohorarias del monóxido de carbono para un día típico de la ciudad de Santiago de Cali.

De acuerdo con la metodología establecida por las agencias medioambientales en el mundo, el registro de la concentración promedio octohoraria se calcula por medio del método de las medias móviles simples, utilizando para ello los datos de las concentraciones horarias de un contaminante en un periodo de ocho horas. Este procedimiento ha generado un interrogante sobre el efecto que pueda llegar a tener la correlación de los errores en el modelo de regresión no paramétrica utilizado para estimar el comportamiento del CO.

Una revisión bibliográfica general de los estudios realizados por Opsomer, Wang, & Yang, (2001), Altman, (1990), (Fernández & Opsomer, 2005), De Brabanter, Suykens, & De Moor, (2011), entre otros, muestran que la presencia de correlación en los errores ocasiona un mal funcionamiento de los métodos de selección automáticos del parámetro de suavización comúnmente conocido como 'Ventana', el cual sirve para determinar la suavidad de la curva de regresión no paramétrica. Por esta razón se pretende implementar una técnica estadística que tenga en cuenta la correlación implícita de los datos y permita mitigar el impacto ocasionado en la estimación de la función media.

La consecución de información oportuna y clara del estado de la calidad del aire, brinda una herramienta muy útil a las autoridades ambientales y a los habitantes para establecer controles y desarrollar acciones preventivas encaminadas a la mitigación de impactos negativos de la contaminación en la población, pronosticar posibles episodios críticos y suministrar información ambiental a la población como es su derecho.

1.1. PREGUNTA DE INVESTIGACIÓN

Todos los aspectos referenciados permiten hacer la siguiente pregunta:

¿Qué influencia tiene la estructura de autocorrelación de los datos al modelar la contaminación atmosférica utilizando un modelo de regresión no paramétrica para los Promedios Octohorarios del Monóxido de Carbono en la Ciudad de Cali en el periodo 2003 - 2006?

2. JUSTIFICACIÓN DEL PROBLEMA

La contaminación atmosférica afecta la salud del ser humano e incluso puede provocar su muerte. Un estudio realizado por el Ministerio del Medio Ambiente en Colombia en el año 2009, demostró que anualmente se presentan alrededor de seis mil muertes asociadas a la contaminación del aire en las principales ciudades del país. Un informe realizado por el Instituto de Hidrología, Meteorología y Estudios Ambientales de Colombia sobre la calidad del aire, reveló que en el país los mayores grados de contaminación se encuentran en los barrios Tunal, Kennedy y puente Aranda de la ciudad de Bogotá, a estas zonas les sigue el centro de la ciudad de Medellín, el área metropolitana del Valle de Aburra y Guayabal en el departamento de Antioquia y la zona industrial entre las ciudades Cali y Yumbo. Otros lugares del país que superan los parámetros internacionales de contaminación son: Bucaramanga, Nemocón y Ráquira en los departamentos Cundinamarca y Boyacá respectivamente. (IDEAM, 2007)

De acuerdo con cifras del Departamento Administrativo de Planeación de Santiago de Cali, el parque automotor en el año 2003 era aproximadamente de 342 mil vehículos registrados, para el año 2006 alcanzaba 400 mil vehículos. Cifras reveladas en el último estudio en número de vehículos registrados en la ciudad es de aproximadamente 504 mil⁸. Para los años 2009 y 2010 en las calles de la ciudad circulaban cerca de 780 mil automotores diariamente (entre registrados y sin registrar)⁹.

La contaminación emitida por estos vehículos se origina en las siguientes causas: la antigüedad de la mayoría de autos, motocicletas, camiones y vehículos de transporte público, el uso de combustibles líquidos de origen fósil, en especial el amplio uso del diesel, intensificado por usuarios que optaron por transformar sus vehículos a este mecanismo (...) por ser más barato, la ignición incompleta en vehículos con motores de combustión interna; el mal estado técnico mecánico de los vehículos y la deficiente infraestructura vial que hace que el tráfico sea más pesado y por ende, los motores permanezcan más tiempo encendidos; además que por cada arranque se quema mayor cantidad de combustible. (DAGMA, 2009).

Uno de los principales contaminantes atmosférico es el monóxido de carbono, que abunda en las grandes ciudades industriales. De acuerdo con la Agencia de Protección Ambiental de Estados Unidos (EPA) este contaminante afecta la función neuronal de las personas en diversos grados, al tiempo que el sistema nervioso central provocando asfixia, dolor de cabeza, mareos, zumbido en los oídos, somnolencia y dificultad para respirar.

Debido a la importancia que tiene el tema de la contaminación a nivel mundial surge la necesidad de analizar y comprender el comportamiento de los contaminantes atmosféricos.

⁸Tomado de la página oficial del Departamento Administrativo de Planeación de Santiago de Cali. Recuperado el 15 de agosto de 2011 de <http://planeacion.cali.gov.co/dapweb/index.asp>

⁹Tomado de la página oficial de la Alcaldía de Santiago de Cali. Lineamiento Para Una Política Ambiental En Cali. Recuperado el 15 de agosto de 2011 de <http://www.cali.gov.co/descargar.php?id=27191>

Particularmente, se pretende modelar el comportamiento de las concentraciones octohorarias de las emisiones de CO producidas por las industrias, fenómenos naturales (incendios forestales o quemas controladas) y los vehículos automotores que utilizan combustibles fósiles en la ciudad de Santiago de Cali durante los años 2003-2006.

Para modelar el comportamiento de las concentraciones 8-h de este contaminante se podría considerar construir un modelo paramétrico. Sin embargo, la naturaleza de las mediciones no permite que esta opción sea viable, debido a que en muchas ocasiones, este tipo de modelos no cumplen con los supuestos acerca de la forma funcional de la distribución de probabilidad de las observaciones contenidas en la muestra, produciendo así estimaciones poco confiables. Además, este tipo de modelos también está sujeto al cumplimiento riguroso de una serie de supuestos basados en la distribución de los errores para poder tener validez en sus estimaciones. (Delicado, 2008:116)

Una alternativa para superar esta dificultad, es la construcción de un modelo que no realice suposiciones sobre la forma de la curva de regresión y que sea flexible en el cumplimiento de los supuestos de la distribución de los errores. Los modelos de regresión no paramétrica utilizan estimadores que poseen bondades de robustez a supuestos distribucionales, los cuales permiten estimar la curva de regresión sin que ésta se vea afectada por el incumplimiento de algunos los supuestos del modelo de regresión, de acuerdo a lo expuesto por Fan & Gijbels (1996), Wand & Jones (1995) y Fan and & Yao (2003). No obstante, autores como Altman (1990), Bowman & Azzalini (1997), De Brabanter, Suykens & De Moor (2011) y Opsomer, Wang & Yang (2001) demuestran que la correlación en los errores de un modelo de regresión no paramétrico afecta negativamente los métodos tradicionales de selección de ancho de banda tales como: el plug-in, la validación cruzada o la validación cruzada generalizada, los cuales arrojan valores muy pequeños para este ancho de banda, ocasionando a una sub suavización de la curva de regresión que casi interpola los datos. De acuerdo con Delicado (2008) esto ocurre por el hecho de que los métodos de selección de ancho de banda "perciben" toda la estructura de los datos que se debe a la función de media y al intento de incorporar esa información en su estimación de la tendencia.

La Norma Colombiana de Calidad del Aire impartida por el Ministerio de Ambiente, Vivienda y Desarrollo Territorial, prescribe que el cálculo de la concentración 8-h de CO se debe obtener a partir una medida móvil simple de la concentración promedio horaria de CO en un periodo de ocho horas. Este método utilizado por la RMCA de la ciudad de Santiago de Cali, ha motivado a estudiar los efectos que pueda tener la correlación de los errores en el modelo propuesto para estimar el comportamiento del CO.

De acuerdo a lo expuesto anteriormente, el propósito principal de esta investigación es estimar un Modelo de Regresión No Paramétrica con Errores Correlacionados, del comportamiento de los Promedios Octohorarios del Monóxido de Carbono en la Ciudad de Cali para una semana típica en lugares específicos durante los años 2003 a 2006. Y eventualmente, analizar las posibles dificultades y fallas que puede presentar dicha estimación.

3. OBJETIVOS

3.1. OBJETIVO GENERAL

Modelar el comportamiento de la contaminación atmosférica por Monóxido de Carbono en la Ciudad de Cali, mediante un modelo de regresión no paramétrica que emplee un método de selección de ancho de banda que incluya la estructura de autocorrelación de los errores del modelo, utilizando los promedios octohorarios para una semana típica en el periodo 2003 - 2006.

3.2. OBJETIVOS ESPECÍFICOS

- Determinar la existencia de una estructura de correlación serial de los errores del modelo de regresión no paramétrico.
- Comparar gráficamente el comportamiento de los métodos de selección de ancho de banda utilizados en la estimación del modelo de regresión no paramétrica con errores correlacionados, de los niveles de concentración octohoraria del monóxido de carbono.
- Determinar el mejor método para seleccionar el ancho de banda para la estimación de los niveles de concentración octohoraria del monóxido de carbono, por medio de un modelo de regresión no paramétrica con errores correlacionados.
- Ajustar un modelo de regresión no paramétrico con errores correlacionados, para el comportamiento de los niveles de concentración promedio octohoraria del monóxido de carbono (8-h) en ciudad de Santiago de Cali.

4. ESTADO DEL ARTE

En esta sección se presentan brevemente algunos trabajos científicos sobre contaminación atmosférica por CO y la regresión no paramétrica.

Olaya (2011) ilustra el uso de las técnicas de regresión no paramétrica para estimar curvas características de contaminación del aire de Cali debidas al Dióxido de Nitrógeno NO₂. La estimación en este caso enfrenta el hecho inusual de contar con varias mediciones para cada valor de X, lo que dificulta tanto la estimación de la función de regresión como la estimación de la varianza, además de la elaboración de inferencias. El autor indica que en general la literatura sobre estimación no paramétrica de la función de regresión omite este hecho y asume que se dispone de n pares (x_i, y_i) distintos. No obstante, explora dos soluciones propuestas por Bowman & Azzalini (1997) y Eubank (1999). La primera propone ajustar un modelo usando como única respuesta mediciones asociadas con un punto de diseño, en donde la función de regresión se estima de manera habitual usando un estimador lineal. La otra solución propuesta, se acerca más a la idea de considerar los datos como respuestas múltiples que como medidas repetidas usando un estimador spline. Por ultimo expone brevemente la posibilidad de utilizar las técnicas de estimación de la función de regresión con un modelo de regresión funcional como el propuesto por Ramsay & Silverman (2005).

Reina (2009) realizó un estudio para analizar el comportamiento de los promedios octohorarios (8-h) del Monóxido de Carbono (CO) en la atmósfera del municipio de Santiago de Cali, para esto tuvo en cuenta las diferentes zonas de monitoreo de la ciudad y el tipo de día durante el periodo 2003– 2006. Utilizando un Estimador Lineal Local, estimó las curvas de tendencia de las concentraciones horarias y octohorarias del CO. Llegando a la conclusión que a lo largo de un día ordinario, la mayor concentración 8-h de CO se presenta en las horas de la tarde, aproximadamente entre las 3:00 p.m. y 4:00 p.m. Mientras que, la mayor concentración promedio hora (1-h) de CO se presenta cerca de las 9:00 a.m. También determinó que los días sábado y domingo, presentaban los niveles de concentración (8-h) más bajos de toda la semana a partir de las 9:00 a.m. Del mismo modo logró verificar que la zona de la ciudad que presentaba los niveles más altos de contaminación por CO fue el centro de la ciudad, atribuyendo este fenómeno a las fuentes móviles. No obstante, esta investigación evidenció que las emisiones registradas durante el periodo de estudio, no superaron las normas ambientales vigentes en Colombia y aceptadas por la EPA, lo cual sugiere que los niveles de CO en Santiago de Cali durante el periodo 2003 -2006 no estuvieron por fuera del rango tolerable para la salud humana. La información utilizada en este proyecto fue suministrada por la Red de Monitoreo de Calidad de Aire (RMCA) de Cali.

Castro (2009) modeló el comportamiento del promedio horario de las emisiones de Monóxido de Carbono (CO) en la zona céntrica del municipio de Santiago de Cali en los años 2003 a 2006 mediante técnicas de regresión no paramétrica. Propuso una adaptación del estimador de Rice y el estimador Gasser, Sroka y Jennen-Steinmetz (GSJ) para la estimación de la varianza y pudo concluir 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. Los resultados de esta investigación permitieron establecer que en los días ordinarios, entre las 8 a.m. y 9 a.m. se presenta un incremento considerable de la concentración del monóxido de carbono en el centro de la ciudad. Que disminuye conforme cae la tarde y se incrementa nuevamente, pero en menor medida a partir de las 7 p.m. El día sábado, presentó el mayor nivel de concentración alrededor del mediodía. Mientras que para los días domingo y festivos, presentaron niveles por debajo de los 3 ppm., lo cual sugiere que estos niveles de concentración son muy bajos en comparación con los demás días de la semana.

Montoya y Morales (2004) modelaron y analizaron mediante técnicas de regresión no paramétrica, el comportamiento diario de las concentraciones horarias de los contaminantes NO_2 , SO_2 y CO en la ciudad de Santiago de Cali. De acuerdo a los objetivos planteados en la investigación, calcularon los índices de calidad del aire para el periodo 2003– 2006. Los resultados obtenidos de esta investigación permitieron caracterizar a las fuentes móviles como la principal fuente de contaminación en la ciudad. Encontraron evidencia que sugiere que los niveles de contaminación por CO en la Calle 15 y por SO_2 en la Base Aérea, varían consistentemente entre las temporadas lluviosa y seca. Y finalmente pudieron determinar que los contaminantes CO , SO_2 y NO_2 presentan una acumulación máxima matutina alrededor de las 10:00 a.m.

Niglio y Perna (2003) describieron y analizaron las diferentes metodologías para la selección de ancho de banda teniendo en cuenta la estructura de correlación de los errores. Exploraron el comportamiento de una serie de registros climáticos (tiempo y temperatura) que presentaba una fuerte estacionalidad. De acuerdo con resultados preliminares, al aplicar regresión Kernel y diversas versiones corregidas del criterio de Validación Cruzada Generalizada (GCV), a la base de datos con evidente dependencia, el procedimiento de selección de ancho de banda GCV incurrió en una sub-suavización de datos, lo cual condujo a la interpolación de las observaciones. Los datos fueron recogidos en el sur de Italia desde enero 1960 a diciembre 2000. Sin embargo, cuando aplicó el criterio de selección propuesto por Altman (1990), se evidenció una adecuada corrección de la dependencia entre los datos y se obtuvo una curva suave ajustada a los datos.

Opsomer J., Wang Y. & Yang Y. (2001) estudiaron las dificultades asociadas a la presencia de correlación en la regresión no paramétrica, en cuanto a la forma y suavidad de la función estimada que depende en gran medida del valor específico que se elija para el 'parámetro de suavizado'. Además, proporcionaron un recuento general de la literatura clásica que abarca este tema y finalmente contribuyeron con algunos desarrollos en esta área. Los resultados obtenidos por los autores se basaron en la comparación de los métodos de selección de ancho de banda; Cross Validation (CV, conocido como validación cruzada), Direct Plug-in (DPI) y su modificación Correlated Direct Plug-in (CDPI) la cual tiene en cuenta la presencia de correlación en los datos. En principio, demostraron que los métodos CV y DPI no se ven gravemente afectados por la correlación cuando dicha dependencia se presenta en periodos de tiempo cortos. No obstante, a medida que el periodo de dependencia entre los datos se

extiende, los anchos de banda seleccionados utilizando los métodos CV y DPI tienden a ser cada vez más pequeños, por lo cual, el ajuste de la curva de regresión no paramétrica a los datos es progresivamente menos suave. Mientras que al utilizar el método de selección CDPI, los anchos de banda seleccionados fueron mucho más estables.

Ruppert, Sheather & Wand (1995) propusieron tres métodos basados en la metodología Plug-in para seleccionar el ancho de banda en la regresión lineal local, mediante la adaptación de las ideas utilizadas en la estimación de la densidad kernel. El criterio Direct Plug-In (DPI), el cual destacan por su buen desempeño y que no requiere de minimización o búsqueda de raíces. El criterio Solve-The-Equation (STE) que parece funcionar tan bien como el DPI, pero tiene la complicación adicional de que requiere un paso para encontrar raíces. Y el criterio Rule of Thumb (ROT), un método rápido y sencillo de aplicar, el cual parece funcionar adecuadamente para una amplia gama de situaciones. Por medio de simulación Monte Carlo llegaron a la conclusión que los métodos DPI y STE funcionan muy bien en la práctica, aunque sugieren realizar un estudio más profundo comparándolos con otras propuestas. Los resultados obtenidos por ellos, sugieren que estos métodos pueden ser aplicados eficazmente en el ajuste de un polinomio local de grado impar (1 ó 3) o superiores, y se puede extender la estimación de derivadas y la regresión no paramétrica múltiple.

Altman (1990) realizó una demostración del mal funcionamiento de las técnicas estándar de selección de ancho de banda, tales como la validación cruzada y generalizada de validación cruzada cuando los errores están correlacionados. Determinó que cuando existe una estructura de correlación positiva en los datos la curva media de regresión estimada es sub-suavizada, mientras que cuando las correlaciones son negativas sobre-suaviza la curva media. Ella propuso dos métodos ajustar el criterio de la Validación Cruzada Generalizada (GCV) para corregir el efecto de la correlación, denominados "directo" e "indirecto" cuales involucran la estructura de correlaciones de los residuales transformándolos para así producir residuos menos correlacionados. Su investigación permitió concluir que al aplicar la metodología propuesta por ella, la regresión kernel funciona muy bien cuando los errores están correlacionados, siempre y cuando las correlaciones se presenten a corto plazo.

5. MARCO TEORICO

5.1. CONTAMINACIÓN ATMOSFÉRICA

De acuerdo con la Secretaria de Medio Ambiente y Recursos Naturales de México (SEMARNAT), la contaminación es la transmisión y difusión de humos o gases tóxicos a medios como la atmósfera y el agua, como también a la presencia de polvos y gérmenes microbianos provenientes de los desechos de la actividad del ser humano. La resolución 0601 de 2006 de la ley colombiana, define la contaminación atmosférica como la presencia de sustancias en la atmósfera en altas concentraciones en un tiempo determinado como resultado de actividades humanas o procesos naturales, que pueden ocasionar daños a la salud de las personas o al ambiente.

Inche (2004) argumenta que luego de la revolución industrial inglesa en el siglo XVIII, la industrialización comenzó a remplazar las actividades agrícolas y las poblaciones se desplazaron del campo a la ciudad. Las fábricas para producir requerían energía mediante la quema de combustibles fósiles, tales como el carbón y el petróleo. A mediados del siglo XIX e inicios del siglo XX el principal problema de contaminación del aire fue el humo y ceniza producidos por la quema de combustibles fósiles en las plantas estacionarias de energía. La situación empeoró con el creciente uso del automóvil. En las grandes ciudades de la actualidad, cuando los gases que contienen partículas sólidas capaces de mantenerse en suspensión como el monóxido de carbono (CO), son liberados a la atmósfera por motores que funcionan con combustibles fósiles, y estos alcanzan valores superiores a los permitidos, se perjudica la salud, tanto del ser humano como de animales y plantas.

Los efectos de la exposición a la contaminación atmosférica son múltiples y de diferente severidad, siendo los más afectados los sistemas respiratorio y cardiocirculatorio. Estos efectos mantienen un aumento tanto en la gravedad de sus consecuencias como en la población susceptible afectada. De acuerdo con Ballester (2005) la contaminación atmosférica procedente de las reacciones de los hidrocarburos y los óxidos de nitrógeno, estimuladas por la luz solar intensa y el incremento de la temperatura producen infecciones respiratorias, cardiopatías y cáncer de pulmón. Generalmente el ozono es considerado como el componente más tóxico de esta mezcla. Se forma por la acción de la radiación ultravioleta del sol sobre los óxidos de nitrógeno (NOx) y en presencia de compuestos orgánicos volátiles y otros contaminantes.

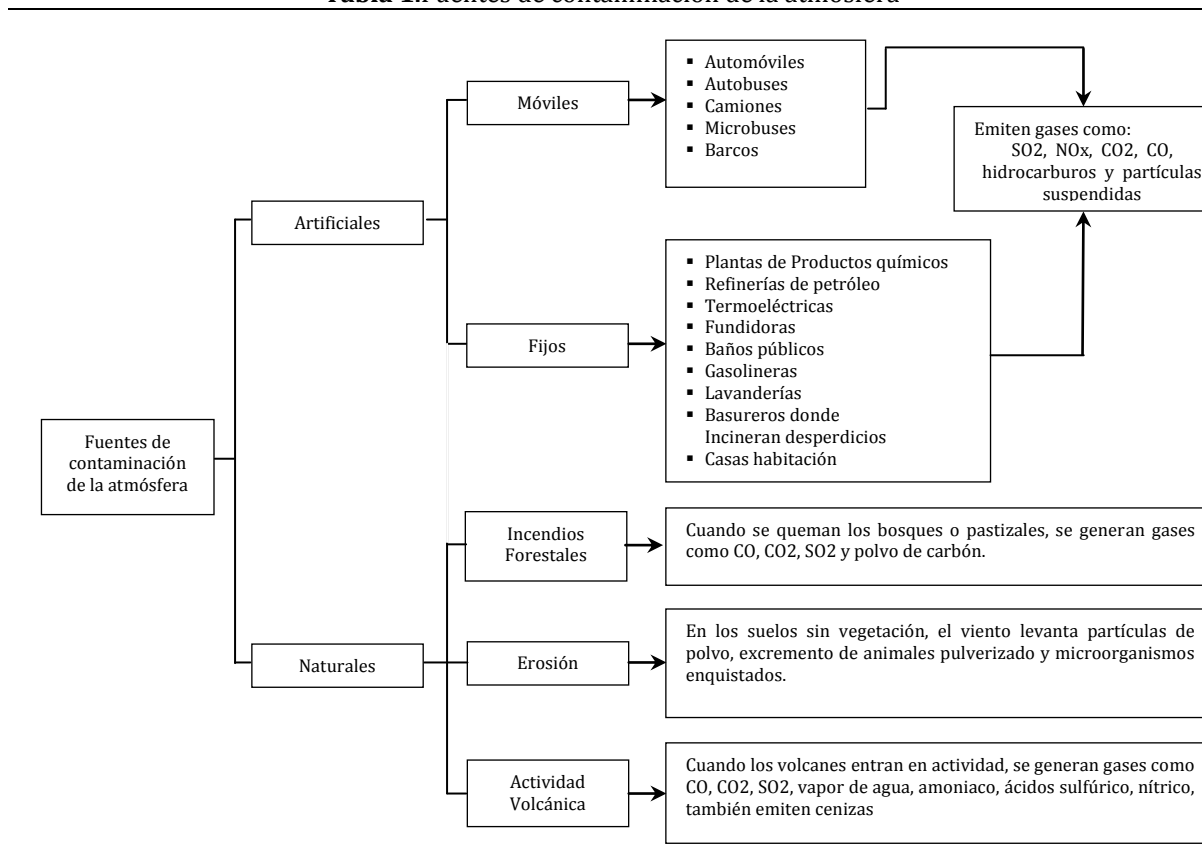
5.1.1. FUENTES CONTAMINANTES DEL AIRE

Conforme a lo manifestado por Inche (2004) los contaminantes presentes en la atmósfera proceden de dos tipos de fuentes emisoras; las naturales y las artificiales. Esta última se refiere a las actividades humanas. Mientras que, las emisiones naturales provienen fundamentalmente de los volcanes, incendios forestales y descomposición de la materia

orgánica en el suelo y en los océanos. En el tabla 1 se muestra detalladamente las fuentes de contaminación atmosférica.

Los contaminantes atmosféricos pueden ser compuestos gaseosos, aerosoles o material particulado. Entre los contaminantes gaseosos se encuentran el ozono, los óxidos de azufre y de nitrógeno, monóxido de carbono, dióxido de carbono y compuestos volátiles orgánicos e inorgánicos. El material particulado se caracteriza, a su vez, por partículas suspendidas totales, partículas suspendidas menores a diez micras y partículas suspendidas con diámetro menor a 2.5 micras. (Instituto Nacional de Ecología, 2007)

Tabla 1.Fuentes de contaminación de la atmósfera



5.2. MONÓXIDO DE CARBONO (CO)

La agencia de protección ambiental de EE.UU. define el monóxido de carbono como un gas tóxico inodoro, incoloro, e insípido. Se produce como consecuencia de la oxidación incompleta del carbón en la combustión. La exposición a niveles bajos de monóxido de carbono puede ocasionar en los seres humanos dolor de cabeza, mareos, confusión, náuseas y fatiga, síntomas similares a los de la gripe. Sin embargo, una exposición prolongada a niveles altos de este gas puede ser fatal. Los efectos pueden variar de persona a persona, dependiendo de la edad, las condiciones de salud, el nivel de concentración de monóxido de carbono al igual que el periodo de tiempo que se esté expuesto al mismo. (The Condensed Chemical Dictionary, 1977).

5.2.1. FUENTES DE EMISIÓN DE CO

Entre las diferentes fuentes de emisiones a la atmósfera se puede distinguir dos grandes tipos: las fuentes fijas y las móviles. Las fuentes más comunes de emisiones tóxicas son los vehículos automotores que utilizan como combustible gasolina o diesel, los procesos industriales, la incineración de materia orgánica, incendios forestales y urbanos, las estufas y chimeneas de leña, los sistemas centrales de calefacción con escapes, los calentadores portátiles o de espacio que utilizan querosén y no tienen ventiladores o respiraderos, los generadores, el humo de tabaco y otros enseres que usan gasolina o algún derivado del petróleo. (Téllez, et al., 2006)

5.3. NORMAS DE CALIDAD DE AIRE

La Agencia de Protección Ambiental (EPA) de los Estados Unidos, ha promulgado las Normas de Calidad de Aire Ambientales Nacionales (NAAQS) con base en recopilaciones modernas de conocimiento científico. En 1970, el primer documento de criterio de calidad del aire para el CO se emitió por la Administración de Control de Polución de Aire Nacional, un predecesor de la EPA. Aunque el criterio de calidad de aire ha cambiado durante las últimas dos décadas, los NAAQS para CO siguen siendo los mismos.

En Colombia, las Normas de Calidad del Aire para todo el territorio nacional empezaron a regir con la Ley 23 de 1973, Decreto Ley 2811 de 1974, Ley 9 de 1979, Decreto 02 del 11 de enero de 1982. Luego en el Decreto 948 del 5 de junio de 1995 se realizó una revisión y ajuste de la normatividad en materia de calidad del aire. Por último el 3 y 4 de Abril del 2006, el Ministerio de Ambiente, Vivienda y Desarrollo Territorial publicó la nueva norma de calidad del aire para Colombia: Decreto 979 de Abril 3 de 2006 y la resolución 601 de Abril 4 de 2006; por medio de la cual se modifican los artículos 7, 10, 93,94, y 108 del Decreto 948 de 1995. La Tabla 2 indica los niveles máximos permisibles establecidos en la anterior resolución.

Tabla 2. Norma de Calidad del Aire para el Monóxido de Carbono.

Nivel Máximo Permisible para el CO, artículo 4°de la resolución 601 de Abril 4 de 2006.					
Contaminante		Tiempo de exposición	Límite Máximo Permisible		
Monóxido de Carbono (CO)		1 hora	35 ppm (40 mg/m³)		
		8 horas	8.8 ppm (10 mg/m³)		
Niveles de Prevención, Alerta y Emergencia, artículo 10°de la resolución 601 de Abril 4 de 2006.					
Tiempo de exposición	Unidad de medición	Límite máximo permisible	Niveles de		
			Prevención	Alerta	Emergencia
8 horas	ppm (mg/m3)	8.8 (10)	14.9 (17)	29.7 (34)	40,2 (46)

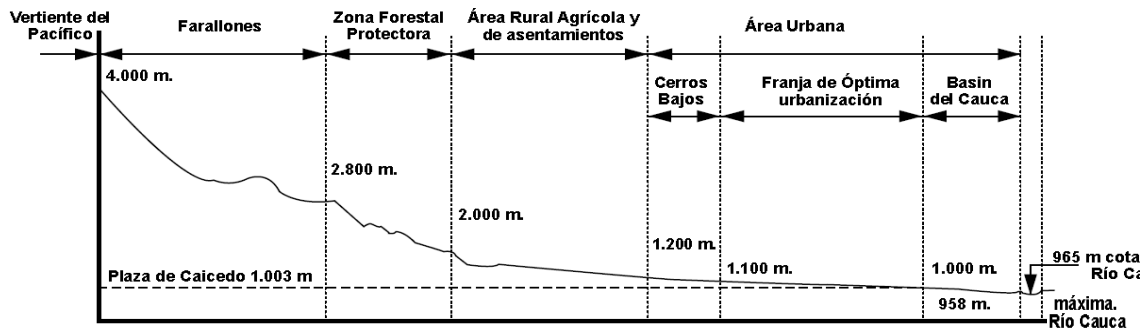
Nota: mg/m³ o µg/m³: a las condiciones de 298,15°K y 101,325 KPa. (25°C y 760 mm Hg.

5.3.1. GENERALIDADES DE SANTIAGO DE CALI

El municipio de Santiago de Cali es una de las principales ciudades del país con aproximadamente 2.963.000 habitantes según datos del censo del 2005 realizado por el Departamento Administrativo Nacional de Estadística (DANE). Se encuentra ubicado en el suroccidente colombiano y ocupa un área de 68.150,59 hectáreas, de las cuales el 82,2% correspondientes a 56.025,13 hectáreas, pertenecen a la zona urbana, mientras que el 17,8% restante, equivalente a 12.125,46 hectáreas, corresponden a la zona rural.

El perfil fisiográfico¹⁰ del municipio puede describirse en un corte orientado hacia el norte, como una línea que desciende de izquierda a derecha, con una zona de montaña entre las cotas 4000 y 2000 metros sobre el nivel del mar, una zona de ladera entre 2000 y 1100 metros sobre el nivel del mal, una zona plana entre 1100 y 965 msnm; finalmente se encuentra una zona de ribera definida como una franja paralela al jarillón del río Cauca y cuyo nivel está por debajo del nivel máximo de las aguas del río Cauca.

Figura 1. Perfil fisiográfico de Santiago de Cali



Fuente: Departamento Administrativo de Gestión de Medio Ambiente- DAGMA, 2009

Los períodos climáticos, duran aproximadamente tres meses cada uno e inician a mediados de mes. Los períodos secos van de diciembre a marzo y de junio a septiembre, y los lluviosos van de marzo a junio y de septiembre a diciembre. Los regímenes de precipitación se encuentran entre los 2600 y 1100 milímetros por año en su gradiente altitudinal, y entre 1400 y 1100 milímetros por año entre el sur y el norte de la zona urbana.

De acuerdo con el Departamento Administrativo de Planeación de Santiago de Cali, el parque automotor de la ciudad está constituido por más de 504,653 vehículos¹¹ registrados (servicio público, particular, oficial y motos) y más de 350 industrias adscritas al programa de control de impactos ambientales del Departamento de Gestión del Medio Ambiente (DAGMA).

¹⁰El perfil fisiográfico es una descripción de las características físicas de la Tierra y de los fenómenos de la naturaleza que en ella se originan, en particular de las características aparentes, conspicuas o superficiales de la superficie terrestre y la vegetación

¹¹Fuente: Secretaría de Tránsito y Transporte Municipal, Centro de Diagnóstico Automotor del Valle Ltda.

Debido a los problemas de movilidad presentados en la ciudad de Cali, el gobierno local decidió implementar en 2003 una restricción de circulación para vehículos particulares, ésta consistía en sacar de circulación aproximadamente al 20% del parque automotor diariamente durante las horas de mayor volumen de tráfico, la medida fue bautizada como “pico y placa” (Cabrera y Guerrero, 2005).

A mediados del mes de marzo de 2009, con el fin de mejorar la calidad, velocidad y seguridad del transporte público de la ciudad, se inauguró de manera formal el sistema de transporte masivo MIO. Uno de sus más destacados objetivos es reducir la contaminación ambiental (auditiva y atmosférica). Para esto el sistema cuenta con más de 81 buses articulados, que corresponden al 1,9% del total del parque automotor de la ciudad. También cuenta con 104 buses alimentadores, que corresponden al 2,5% del total de vehículos afiliados en la ciudad. Además de 186 buses especiales denominados padrones, que corresponden a 4,4% del total del parque automotor de la ciudad. Con la implementación de este sistema se espera que mejore la calidad del aire en la ciudad debido a la disminución del tráfico vehicular (salida de antiguo parque automotor rodante por las vías) y la inclusión al sistema de transporte público de la ciudad, de buses amigables con el medio ambiente que generan menos emisiones contaminantes al medio.¹² También se espera una disminución en los niveles de ruido debido a la implementación de una flota de buses nueva, con la garantía de un mantenimiento periódico. Adicionando un menor número de buses transitando por la zona.

5.3.2. RED DE MONITOREO DE LA CALIDAD DEL AIRE EN CALI

La calidad del aire de la ciudad de Santiago de Cali ha sido evaluada a través de la Red de Monitoreo de Calidad del aire (RMCA) constituida por 8 estaciones fijas de monitoreo y una unidad móvil. Inició operación a mediados de 1999. Sin embargo, el primer año correspondió a una etapa de pruebas y ajustes al sistema; en el año 2000 salió de operación hasta abril de 2003. El Departamento Administrativo de Gestión del Medio Ambiente (DAGMA) se ha encargado de analizar el comportamiento de los parámetros meteorológicos y de los contaminantes que afectan la calidad del aire en Santiago de Cali, presentando un análisis del comportamiento temporal y estacional de los contaminantes, la Tabla 3 muestra los parámetros medidos por estación.

5.3.3. PROTOCOLO DE MONITOREO

El IDEAM recomienda adoptar las recomendaciones establecidas en el Protocolo del Monitoreo y Seguimiento del Módulo Aire del Sistema de Información Ambiental, con el fin de garantizar la representatividad de los datos de calidad del aire, obtenidos a través de las

¹²De acuerdo con la empresa encargada del diseño, construcción y puesta en marcha del Sistema Integrado de Transporte Masivo (SITM)-Metrocali, se espera una reducción del 39% de las emisiones de monóxido de carbono, del 32% de las emisiones de óxido de nitrógeno y del 8% de la emisión de compuestos volátiles.

redesde monitoreo de calidad del aire. Este módulo se encuentra en la Norma de Calidad del Aire (2005) ¹³ y establece que:

- 1 Se considera válida la concentración de 8 horas, si al menos el 75% de los datos de concentración de 1 hora para un período de 8 horas se encuentran disponibles.
- 2 En el evento que se disponga de menos del 75% de los datos de concentración de 1 hora, la concentración de 8 horas será considerada, sólo para efectos de verificar el cumplimiento de la norma primaria de calidad de aire como concentración de 8 horas, si al remplazar por cero los datos que falten para completar el 75% requerido, la concentración de 8 horas es mayor o igual al nivel de la norma.
- 3 Si se dispone de datos de concentración de 1 hora para 6 ó 7 horas, la concentración de 8 horas se calculará como el promedio aritmético de los datos de concentración de 1 hora disponibles, utilizando como divisor 6 ó 7 según corresponda.
- 4 La concentración de ocho horas máxima diaria por un día dado del calendario es la más alta de las 24 concentraciones medias de ocho horas posibles computadas para ese día. Este proceso se repite, alcanzando una concentración media de ocho horas máxima diaria para cada día del calendario.

Para el cálculo de las concentraciones promedio 8-h de CO, solo se tuvo en cuenta los ítems 1 y 3 del protocolo de monitoreo de contaminación octohoraria en este proyecto de investigación. Como el objetivo de este proyecto no es verificar el cumplimiento de la norma primaria de calidad de aire, no se remplazó por ceros los datos faltantes para completar el 75% requerido.

Tabla 3.Estaciones y parámetros medidos en la RMCA (Cali).

	Estación	Abreviatura	Sector	Parámetros medidos				
				O ₃	CO	SO ₂	NO _x	PM ₁₀
1	Centro Diagnóstico Automotor del Valle	CDAV	Norte	*	*	*	*	*
2	Base Aérea Marco Fidel Suárez	BA	Nororiente	*	*	*	*	*
3	Hospital Universitario del Valle	HUV	Sur occidente		*	*		
4	Polideportivo el Diamante	PDD	Oriente	*	*	*	*	*
5	CVC Pance	CVC	Zona rural (Sur)	*				*
6	Escuela República de Argentina	ERA	Centro		*	*	*	*
7	Universidad del Valle	UV	Sur	*		*		*
8	Calle 15	CALLE 15	Centro		*		*	*
9	Unidad Móvil	UM		*	*	*	*	*

Fuente: DAGMA, 2005.

¹³Tomado del convenio de asociación No. 038/04 del IDEAM – normas de calidad del aire a nivel internacional – publicado en la página oficial del Ministerio de Ambiente, Vivienda y Desarrollo Territorial. Recuperado el 30 de enero de 2010 en www.minambiente.gov.co/documentos/1746_Anexo_3_info_nal_e_internal.pdf

5.3.4. CONCENTRACIÓN OCTOHORARIA

Se entiende por concentración promedio octohoraria al promedio móvil simple de los datos de las concentraciones horarias de un contaminante en un periodo ocho horas. Cuando una nueva hora es añadida, se desecha la primera hora, de esta manera siempre va a ser el promedio móvil de las últimas 8 horas. La sigla 8-h corresponde a la concentración promedio octohoraria de un contaminante. El primer periodo de cálculo para cualquier día dado será el periodo que comience a las 17:00 de la víspera y termine a las 0:00 de ese día; el último periodo de cálculo para cualquier día dado será el que transcurra entre las 16:00 y las 23:00 horas.

A partir de los registros de los niveles de concentración promedio horarios de CO suministrados por la RMCA y de acuerdo a la Norma Colombiana de Calidad del Aire Ministerio de Ambiente, Vivienda y Desarrollo Territorial, Decreto 979 del 3 de abril de 2006), el cálculo de la concentración 8-h de CO se obtiene a partir de un mínimo de 6 datos de concentración promedio horaria de CO en un periodo de ocho horas (al menos el 75% de los datos de concentración promedio horaria de CO en dicho periodo).

La contaminación 8-h que se reporta para una hora determinada corresponde al promedio de las ocho horas anteriores, incluyendo la hora del reporte. Por ejemplo, para estimar el promedio móvil de 8 horas de las 13:00 horas, se calcula el promedio con las concentraciones registradas desde las 06:00 hasta la 13:00 horas y así sucesivamente para cada hora.

La primera medición del promedio octohorario se obtiene a partir de la última hora (hora 7) del periodo de las primeras ocho horas (desde la hora 0 hasta la hora 7), por lo tanto no se tendrán promedios octohorarios en las primeras siete horas de ese día (desde la hora 0 hasta la hora 6). Si se cuenta con un total de 8 registros en un periodo de 8 horas, el cálculo de las concentraciones promedio octohorarias se realizará así:

$$\begin{array}{l} \text{Primer cálculo de concentración promedio 8-h} \\ \text{(registrado en la hora 7)} \end{array} \longrightarrow CO_{(8h_7)} = \frac{CO_{(1h_0)} + CO_{(1h_1)} + \dots + CO_{(1h_7)}}{8}$$

$$\begin{array}{l} \text{Segundo cálculo de concentración promedio 8-h} \\ \text{(registrado en la hora 8)} \end{array} \longrightarrow CO_{(8h_8)} = \frac{CO_{(1h_1)} + CO_{(1h_2)} + \dots + CO_{(1h_8)}}{8}$$

$$\begin{array}{l} \text{Tercer cálculo de concentración promedio 8-h} \\ \text{(registrado en la hora 9)} \end{array} \longrightarrow CO_{(8h_9)} = \frac{CO_{(1h_2)} + CO_{(1h_3)} + \dots + CO_{(1h_9)}}{8}$$

Y así sucesivamente. Si hay 6 ó 7 datos en el periodo de ocho horas, el promedio octohorario será el resultado de promediar los seis o siete datos así no estén consecutivos pero que se encuentren dentro de un periodo de ocho horas.

5.4. EL MODELO DE REGRESIÓN

Los modelos de regresión constituyen una herramienta muy útil y potente para los analistas estadísticos cuando su objetivo es relacionar una variable respuesta con una o más variables explicativas. A partir de ellos se pueden estimar parámetros, realizar inferencias y hacer predicciones. (Servy, et al., 2006).

Los **modelos de regresión paramétricos** suponen que los datos observados provienen de variables aleatorias cuya distribución es conocida, salvo por la presencia de algunos parámetros cuyo valor se desconoce. De acuerdo con Wand & Jones (1995) y Graybill (1976) la formulación de este modelo estadístico podría describirse así:

$$y_i = m(x; \theta) + e_i, \quad i = 1, \dots, n; \quad \theta \in \Theta \subseteq R^p \quad (1)$$

Donde y_i es una variable dependiente, comúnmente conocida como la variable repuesta; x es la variable explicativa o covariable; $m(x; \theta)$ es una función conocida de x cuyos parámetros θ se desconocen y se desean estimar; finalmente e_i representa los errores aleatorios del modelo incorrelacionado cuya distribución de probabilidad es $N(0, \sigma^2 > 0)$.

Sin embargo estos modelos tienen una estructura rígida (lineal, cuadrática, etc.) que a menudo no logra “captar” el comportamiento de los datos en todo el campo de variación de las variables explicativas. Por otro lado las inferencias realizadas sobre dichos modelos sólo son válidas bajo rigurosos supuestos distribucionales. Si la realidad no se ajusta a esta situación debe recurrirse a otras herramientas que permitan tanto representar adecuadamente los datos como flexibilizar los supuestos.

Los **modelos de Regresión No Paramétricos** sirven para ajustar una función de regresión cuando existe poco conocimiento a priori de su verdadera forma. La idea general de este modelo consiste en hallar la respuesta media que explique el comportamiento de los datos, a partir de un conjunto de n datos (x_i, y_i) por medio de la función de regresión $m(x_i)$ la cual no supone de antemano que las variables (explicativas y respuesta) estén relacionadas en una forma particular y cuyo resultado son curvas “suaves” (en el caso univariado) que representan adecuadamente los datos. De acuerdo con Wand & Jones (1995) la expresión matemática que resume este modelo es la siguiente:

$$y_i = m(x_i) + e_i, \quad i = 1, \dots, n. \quad (2)$$

En donde y_i y e_i son variables aleatorias definidas en la ecuación (1) mientras que $m(x_i)$ es la función o curva de regresión media de la cual se desconoce su forma funcional.

5.4.1. MODELO DE REGRESIÓN NO PARAMÉTRICA CON ERRORES CORRELACIONADOS

De acuerdo con Hart & Wehrly (1986) y Azzalini (1984) el modelo de regresión con errores correlacionados se compone de una función suave desconocida $m(\cdot)$ y de errores aleatorios

generados por un proceso estocástico estacionario con media 0. En otras palabras, dicho modelo se puede expresar como en la ecuación (2) excepto que ahora los términos de error (e_i) están serialmente correlacionados. La expresión matemática que resume este modelo es la siguiente:

$$y_i = m(x_i) + u_i \quad (3)$$

En donde, $E(u_i) = 0$, $Var(u_i) = \sigma^2$ y $Corr(u_i, u_j) = \rho_n(x_i - x_j)$. En este caso, ρ_n se supone continuo y satisface las condiciones de que $\rho(0) = 1$ y $|\rho(u)| \leq 1$ siempre y cuando $u \in [-1, 1]$. De acuerdo con González (2009) este supuesto sugiere que el efecto de los disturbios muy lejanos en el tiempo (pasado) se va diluyendo y el error de cada observación se ve más influenciado por los acontecimientos más recientes.

5.4.2. SUAVIZACIÓN

Es el procedimiento utilizado para estimar la función $m(x)$ en el marco del análisis de regresión no paramétrica. La idea básica según Eubank (1999), parte del supuesto que la función m es suave y que se desea estimar $m(x)$. Debido a la suavidad de m , se presume que las observaciones y_i obtenidas en los puntos x_i cercanos a x contengan cierta información acerca del valor de m en ese punto. Por tanto, debe ser posible utilizar algún tipo de promedio de las respuestas y_i en aquellos puntos x_i cercanos a x para estimar m . Estos promedios se conocen como suavizadores los cuales hacen uso de una función Kernel para 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.

Los estimadores Kernel, son estimadores lineales que expresan el valor del estimador en algún punto de diseño como una suma ponderada de la respuesta:

$$\hat{m}(x) = (nh)^{-1} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) y_i \quad (4)$$

Donde K es una colección de pesos (función) que dependen de una variable de diseño x que puede ser fija o aleatoria; X_i representa los puntos de interés y h es el parámetro de suavización o ancho de banda (bandwidth) el cual puede ser estimado a partir de la información contenida en los datos.

5.4.2.1. FUNCIONES KERNEL

Las funciones de pesos tipo Kernel, denotada como $K(u)$, donde $u = (x - X_i)/h$, son funciones de densidad simétricas alrededor de cero, que decrece comúnmente a cero a medida que se aleja del punto de diseño x . Su utilidad se ve reflejada en la asignación de pesos a los puntos contenidos en el ancho de banda.

Es habitual denotar de forma simplificada la función de pesos con respecto al parámetro de suavización h como:

$$K_h(x) = \left(\frac{1}{h}\right) K\left(\frac{x}{h}\right) \quad (5)$$

Luego, la ecuación (4) puede ser rescrita como:

$$\hat{m}(x) = n^{-1} \sum_{i=1}^n K_h(x - X_i) y_i \quad (6)$$

De acuerdo con Eubank, (1999) estas funciones de pesos no dependen de los datos que intervienen en la curva regresión. Las condiciones que estas funciones deben cumplir son:

$$(A1.) \quad \text{La suma de los pesos debe ser igual a uno: } \int_{-1}^1 K(u) du = 1$$

$$(A2.) \quad \text{Simetría alrededor de cero: } \int_{-1}^1 u K(u) du = 0$$

$$(A3.) \quad M_2 = \int_{-1}^1 u^2 K(u) du \neq 0 \quad |^{14}$$

$$(A4.) \quad V = \int_{-1}^1 K(u)^2 du < \infty$$

Existen numerosas funciones Kernel, las cuales tienen diferentes patrones en la asignación de pesos, las más usadas se expresan en la Tabla 4 y de forma gráfica en la Figura 2.

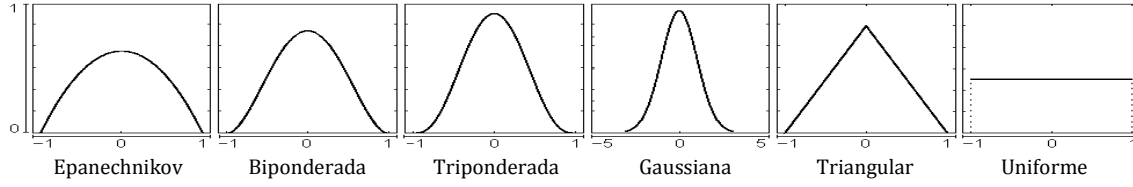
Tabla 4. Funciones Kernel de Segundo Orden

Kernel	Fórmula	Eficiencia = $\left\{ \frac{C(K^*)}{C(K)} \right\}^{5/4}$
Epanechnikov	$\frac{3}{4}(1 - u^2); \quad u < 1$	1
Biponderada	$\frac{15}{16}(1 - u^2)^2; \quad u < 1$	0.994
Triponderada	$\frac{35}{32}(1 - u^2)^3; \quad u < 1$	0.987
Gaussiana	$\frac{1}{\sqrt{2\pi}} e^{-(1/2)u^2}; \quad u < \infty$	0.951
Triangular	$1 - u ; \quad u < 1$	0.986
Uniforme	$\frac{1}{2}; \quad u < 1$	0.930

Como se aprecia en la Tabla 4, la función Epanechnikov es la función ponderadora óptima a la hora de asignación de pesos. No obstante, la pérdida de eficiencia de las demás funciones kernel con respecto a la función Epanechnikov, es muy pequeña. Simonoff (1996), demostró que usar otra función de ponderación fuera de la óptima, no afecta sensiblemente las curvas resultantes, y por ello a menudo se prefiere la función gaussiana debido a sus propiedades de diferenciabilidad.

¹⁴Estas propiedades hacen referencia a la variancia y la a curtosis de la función Kernel, que dan origen al Kernel de segundo orden o simplemente un Kernel, según su contexto. En el momento M en que la tercera condición sea diferente de cero este indicará el orden del Kernel.

Figura 2. Funciones Kernel de Segundo Orden



5.4.3. ESTIMADOR POLINÓMICO LOCAL CON FUNCIONES KERNEL

Como se mencionó anteriormente, la relación entre una variable explicativa y una respuesta puede expresarse a través del modelo (2) en donde la curva de regresión $m(x)$ es la esperanza condicional $E(Y|X = x)$.

Por definición,

$$m(x) = E(Y|X = x) = \int y f(y|x) dy = \int y \frac{f(x, y)}{f_x(x)} dy \quad (7)$$

En donde, $f_x(x)$ simboliza la función de densidad marginal de X , y $f(x, y)$ representa la función de densidad conjunta de X e Y . De acuerdo con Simonoff (1996) si cada una de estas funciones de densidad es estimada en forma no paramétrica a través del método de estimación Kernel, se obtiene el siguiente estimador, conocido como Nadaraya-Watson.

$$\hat{m}(x) = \frac{\sum_{i=1}^n K_h(x - x_i) y_i}{\sum_{i=1}^n K_h(x - x_i)} \equiv \sum_{i=1}^n w_i y_i \quad (8)$$

Donde, h es el ancho de banda; K es una función Kernel presentada en la sección 5.4.3.1 con ciertas propiedades que garantiza las estimaciones de las funciones de densidad $\hat{f}_x(x)$ y $\hat{f}(x, y)$.

Partiendo del hecho que (8) es una función lineal de y con pesos:

$$w_i = \frac{K_h(x - x_i) y_i}{\sum_{i=1}^n K_h(x - x_i)} \equiv \frac{K_h(x - x_i)}{f_x(x)} \quad (9)$$

Wand & Jones, (1995) aseguran que se puede llegar a deducir que esta función es también la solución por mínimos cuadrados ponderados de β_0 en una regresión donde la variable respuesta es explicada sólo por un intercepto. Esto quiere decir que la formulación de $\hat{\beta}_0$ que minimiza la expresión $\sum_{i=1}^n K_h(x - x_i) (y_i - \beta_0)^2$ es la ecuación (8). Como se puede apreciar a continuación:

$$\sum_{i=1}^n K_h(x - x_i) (y_i - \beta_0)^2 = \sum_{i=1}^n K_h(x - x_i) (y_i - \hat{m}(x))^2 \quad (10)$$

En donde β_0 es un polinomio de grado (p=0).

Resolviendo mediante la primera derivada con respecto a β_0 se tiene:

$$\sum_{i=1}^n K_h(x - x_i)(y_i - \beta_0) = 0 \quad (10.1)$$

$$\hat{\beta}_0(x) = \hat{m}(x) = \frac{1}{n} \sum_{i=1}^n w_i y_i \quad (10.2)$$

En consecuencia la ecuación (8) corresponde a la aproximación local de $m(x)$ por una constante. Los mayores valores de los pesos w_i corresponden a los x_i cercanos a x .

La función $m(\cdot)$ puede, sin embargo, ser aproximada por un polinomio de grado p . Basándose en el desarrollo de Taylor, la función $m(t)$ puede ser aproximada como:

$$m(t) \approx m'(x)(t - x) + \dots + m^{(p)}(x)(t - x)^p \frac{1}{p!} \quad (11)$$

Donde t , es un punto en la vecindad de x . Esta aproximación sugiere la posibilidad de una regresión polinómica local en el entorno de x . La función Kernel $K_{h(x-x_i)}$ puede ser incluida como peso en lugar de los pesos w_i del estimador Nadaraya-Watson, de manera que se ajuste un polinomio teniendo en cuenta la proximidad al punto x . Siguiendo el planteamiento hecho en (10) se sugiere, ajustar polinomios locales de mayor grado ya que una constante local sólo tendría sentido sobre un pequeño vecindario (ancho de ventana muy pequeño). Así, el problema queda planteado como:

$$\min_{\beta} \sum_{i=1}^n (y_i - \beta_0 - \dots - \beta_p(x - x_i)^p)^2 K_h(x - x_i) \quad (12)$$

Como puede advertirse en la expresión (12) la estimación dependerá del grado del polinomio, de la función de pesos K_h y del ancho de ventana h . El subíndice β representa el vector de coeficientes $(\beta_0, \beta_1, \dots, \beta_p)'$. La solución al problema de estimación de β puede ser obtenida aplicando la teoría de mínimos cuadrados ponderados, la cual conduce a la siguiente expresión:

$$\hat{\beta} = (X_x' W_x X_x)^{-1} X_x' W_x Y \quad (13)$$

En donde, Y es el vector de respuestas, X_x es la matriz de diseño y W_x es la matriz de los pesos. Wand & Jones, (1995) parten del hecho que $X_x' W_x X_x$ es invertible. Definiendo cada componente matricialmente así:

$$Y = (y_1, \dots, y_n)' \quad (14)$$

$$X_x = \begin{bmatrix} 1 & x - x_1 & \cdots & (x - x_1)^p \\ 1 & x - x_2 & \cdots & (x - x_2)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x - x_n & \cdots & (x - x_n)^p \end{bmatrix} \quad (15)$$

$$W_x = \frac{1}{h} \text{diag}[K_h(x - x_1), \dots, K_h(x - x_n)] \quad (16)$$

A partir de (13) se obtiene:

$$\hat{m}(x; p; h) = e_1' (X_x' W_x X_x)^{-1} X_x' W_x Y \quad (17)$$

Donde, e_1' es un vector fila con $(p + 1)$ columnas que contiene el valor 1 en la primera y 0 en las restantes.

Una vez realizada la estimación para todos los valores de x (incluso para otros valores que pertenezcan al campo de variación de la variable explicativa, para lograr una curva más suave), estas se unen formando la curva suavizada.

De acuerdo con Fan & Gijbels (1996) a menudo basta con elegir un polinomio de grado bajo y concentrarse en la elección del ancho de banda para obtener un ajuste satisfactorio. Una estimación lineal local por lo general produce un mejor ajuste, especialmente en las fronteras. Una estimación cuadrática local reduce el sesgo aún más, pero la varianza mayor puede ser un problema, especialmente en las fronteras. Una estimación cúbica y de orden superior rara vez produce un gran beneficio.

El estimador polinómico de grado $(p=1)$ se conoce comúnmente como el estimador lineal local queda definido de la siguiente manera:

$$\hat{m}(x; 1; h) = \sum_{i=1}^n \frac{\{\hat{s}_2(x; h) - \hat{s}_1(x; h)(x - x_i)\} K_h(x - x_i) y_i}{\hat{s}_2(x; h) \hat{s}_0(x; h) - \hat{s}_1(x; h)^2} \quad (18)$$

En donde, $\hat{s}_r(x; h) = n^{-1} \sum_{i=1}^n (x - x_i)^r K_h(x - x_i)$.

El estimador lineal local puede representarse matricialmente de la siguiente forma:

$$\hat{m}(x) = \mathbf{S} \vec{\mathbf{y}} \quad (19)$$

Donde $\hat{m}(x)$ denota el vector de valores ajustados en los puntos de estimación. La matriz de suavizamiento en el estimador lineal local está definida por:

$$\mathbf{S} = (\mathbf{S}_2 - \mathbf{D}_x \mathbf{S}_1) \times K_h / (\mathbf{S}_2 \times \mathbf{S}_0 - (\mathbf{S}_1)^2) \quad (20)$$

En este caso, Des la matriz de diferencias cuyo elemento (i, j) será $(x - x_i)$. Las matrices S_0, S_1 y S_2 pueden ser definidas como $S_r = (K_{h \times} \mathbf{D}^r)' (1_n \times 1_n')$.

De acuerdo con Bowman & Azzalini (1997) una característica particular de los estimadores polinómicos locales de grado cero (Nadaraya-Watson) y de grado uno (estimador lineal local) es que cuando los puntos de diseño están igualmente espaciados, las estimaciones realizadas por estos estimadores serán idénticas siempre y cuando se tengan la misma función de peso y un mismo ancho de banda.

Fan & Gijbels (1996) mencionan que el estimador lineal local ofrece algunas ventajas, como:

La aproximación lineal local se adapta tanto a diseños fijos como aleatorios; a diseños altamente densos como a diseños casi uniformes, a diferencia de lo que puede ofrecer la aproximación promedio local.

La componente del sesgo de la aproximación lineal local no depende del patrón de los puntos de diseño, por lo menos asintóticamente. Esto es útil a la hora de comparar curvas de regresión.

Una ventaja que tiene el estimador lineal local frente al estimador promedio local es cuando el valor del parámetro de suavización es grande, pues este último estimador arroja estimaciones más sesgadas en los bordes de la región de la covariable sobre el cual los datos fueron generados. Además, tiende a convergir a una línea recta en el promedio de los valores de y , mientras que el estimador lineal local converge a una línea de regresión por mínimos cuadrados ponderados.

Una desventaja de la aproximación lineal local, es que cuando el ancho de banda es pequeño y se tiene una región que presenta pocas observaciones, el estimador puede dar resultados indefinidos.

5.4.4. PARÁMETRO DE SUAVIZACIÓN “ h ”

Este parámetro es conocido como el ancho de banda del estimador no paramétrico, el cual se estima a partir de los datos. Se utiliza para establecer el tamaño de la vecindad o localidad en la función Kernel, es decir, establece el grado de importancia que tendrá la información alrededor del punto de estimación (mayor peso a los puntos cercanos al punto de interés y menor peso a los puntos que están alejados) por lo tanto las observaciones cercanas al punto de interés que estén dentro de la ventana¹⁵ establecida tendrán un peso mayor que cero y las observaciones lejanas al punto de interés tendrán pesos cercanos a cero.

¹⁵Se define ventana como un intervalo simple de la forma $(x-h, x+h)$ en donde h es el parámetro ancho de banda.

De acuerdo con Delicado (2008) la elección del parámetro de suavizado h tiene una importancia crucial en el aspecto y propiedades del estimador de la función de regresión. Puesto que este parámetro controla el equilibrio que el estimador no paramétrico de la función de regresión debe mantener entre el buen ajuste a los datos observados y la capacidad de predecir bien observaciones futuras. Por lo tanto, valores pequeños de h proporcionarían mucha flexibilidad al estimador y le permiten acercarse a todos los datos observados (cuando h tiende a 0 el estimador acaba por interpolar los datos), pero los errores de predicción asociados serán altos. Hay, por tanto, sobreajuste (overfitting). En el caso de que h tome un tamaño moderado no se ajustará tan bien a las observaciones (tampoco es necesario, dado que los datos pueden contener ruido aleatorio) pero predecirá mejor. En el otro extremo, si h es demasiado grande, tendremos falta de ajuste (underfitting), como puede ocurrir con los modelos paramétricos globales.

En consecuencia, para h pequeño el estimador es muy variable (aplicado a muestras distintas provenientes del mismo modelo da resultados muy distintos) y tiene poco sesgo (el promedio de los estimadores obtenidos para muestras distintas es aproximadamente la verdadera función de regresión). Si h es grande ocurre lo contrario.

De acuerdo con Opsomer, et al., (2001) existen dos formas de seleccionar el parámetro de suavizado en la práctica; Una de ellas se basa en el juicio del investigador el cual de manera manual compara los resultados obtenidos para distintos valores de h eligiendo el resultado más satisfactorio visualmente, o el más informativo (el que mejor resume la relación existente entre los datos). Esta forma de proceder está sujeta a la opinión subjetiva del usuario y no puede automatizarse, lo que la hace inviable cuando el número de estimaciones no paramétricas que se han de realizar es grande. La otra forma de selección comprende aquellos métodos que logran un balance entre la varianza y el cuadrado del sesgo, mediante la minimización de una medida global del error llamada Error Cuadrático Integrado Medio (MISE) el cual se define como:

$$MISE(h) = E\left(\int [\hat{m}(x) - m(x)]^2 dx\right) \quad (21)$$

$$= \int \text{sesgo}(\hat{m}(x))^2 dx + \int \text{var}(\hat{m}(x)) \quad (22)$$

Donde, $\hat{m}(x)$ es la estimación de la función que explica el comportamiento de los datos y $m(x)$ es la verdadera función desconocida de los datos. De acuerdo con Opsomer, et al., (2001) bajo los supuestos de que la segunda derivada función $m(x)$ es acotada y continua, $n \rightarrow \infty$, $h \rightarrow 0$, $nh \rightarrow \infty$ y los mencionados en (A1 – A4). El sesgo y la varianza de $m(x)$ puede aproximarse asintóticamente por:

$$E\{\hat{m}(x) - m(x)\} = h^2 \frac{\mu_2(k)}{2} m''(x) + O(h^2) \quad (22.a)$$

$$E\{\hat{m}(x) - m(x)\} = \frac{1}{nh} \mu_0(k^2) \sigma^2 (1 + 2R) + O\left(\frac{1}{nh}\right) \quad (22.b)$$

En donde, $\mu_r(G) = \int u^r G(u) du$ para cualquier función $G(\cdot)$ y R representa la suma de correlaciones.

Por lo tanto, cuando $\rho_I \neq 0$, es decir, existe presencia de correlación, ésta aporta un término adicional en el valor esperado del criterio GCV que depende de (h) y es del mismo orden que el componente de varianza de MISE que es probable que conduzca a elecciones inadecuadas de ancho de banda. (Niglio & Perna, 2003; Fernández & Opsomer, 2005; Liu 2001)

De acuerdo a lo anterior, el MISE puede expresarse de la siguiente forma:

$$E\{\hat{m}(x) - m(x)\} = h^4 \frac{\mu_2(k)^2}{4} m''(x) + \frac{1}{nh} \mu_0(k^2) \sigma^2 (1 + 2R) \quad (22.c)$$

La cual es minimizada por:

$$h_{opt} = n^{-1/5} \left\{ \frac{\mu_0(k^2) \sigma^2 (1 + 2R)}{n \mu_2(k)^2 f''(x)^2} \right\}^{1/5} \quad (23)$$

La expresión (23) conduce al parámetro de suavización óptimo para el caso del estimador lineal local. Como se puede observar, el valor óptimo del parámetro de suavización depende tanto de $m(x)$ como de R . En este sentido, si $R > 0$, esto implica que la correlación del error es positiva la varianza de $m(x)$ será mayor que en el caso incorrelacionado. En consecuencia, el valor del ancho de banda óptimo que minimiza al MISE será mayor que en el caso no correlacionado. Mientras que, si $R < 0$ el ancho de banda óptimo será menor que en el caso no correlacionado.

5.4.5. SELECCIÓN DEL PARÁMETRO DE SUAVIZADO h

La elección del valor que toma el parámetro de suavizamiento h es un paso muy importante en la estimación de curvas de regresión no paramétrica basada en Kernels, dado que éste controla la forma y suavidad de la función estimada $\hat{m}(x)$. La idea consiste en minimizar ciertos parámetros que miden el grado de ajuste de la curva por medio del MISE. La mayoría de los procedimientos propuestos de selección de ancho de banda utiliza un algoritmo de validación cruzada o un método plug-in. A continuación se describen algunos métodos.

5.4.5.1. RULES OF THUMB (ROT)

De acuerdo con Coakley et al. (2000). La Regla de Oro (Rule of Thumb) es considerada como la primera generación de los métodos de selección de ancho de banda Direct Plug-in (DPI), porque la idea consiste en “conectar” las estimaciones de las cantidades desconocidas dentro de la fórmula asintótica de ancho de banda óptimo.

La parte ambigua de este proceso es qué el método utiliza para estimar las cantidades desconocidas, la varianza, ciertas funciones derivadas, y la función de densidad. Fan Gijbels

(1996) sugiere ajustar un polinomio de cuarto grado por el método de mínimos cuadrados, y usar las derivadas de éste polinomio y la suma de cuadrados residuales estandarizados como las estimaciones de sus análogos desconocidos.

Ruppert et al. (1995) sugieren utilizar una técnica similar, excepto que su polinomio es ajustado en bloques. El número de bloques se elige basado en el estadístico C_p de Mallows (Mallows 1973).

$$\hat{h}_{ROT} = C_1(k) \left\{ \frac{\hat{\sigma}_Q^2(\hat{N})(b-a)}{\hat{\theta}_{22}^Q(\hat{N})n} \right\}^{1/5} \quad (24)$$

En dónde:

- $C_1(k) = \{1/(2\sqrt{\pi})\}^{1/5}$ es un número real que depende de un Kernel k .¹⁶
- $\hat{\theta}_{22}^Q(\hat{N}) = n^{-1} \sum_{i=1}^n \sum_{j=1}^{\hat{N}} (\hat{m}_j^Q(x_i)'')^2$
- $\hat{\sigma}_Q^2(\hat{N}) = (n - 5N)^{-1} \sum_{i=1}^n \sum_{j=1}^{\hat{N}} \{Y_i - \hat{m}_j^Q(x_i)\}^2$

La elección de $\hat{N}$ se hace dentro del conjunto $\{1, 2, \dots, N_{max}\}$ se minimiza la suma de los cuadrados residual $RSS(N)$ sobre la base de un ajuste bloqueado de cuarto grado sobre los N bloques.

$$N = \min_{arg} \left\{ C_p(N) = \frac{RSS(N)}{RSS(N_{max})/(n-5N_m)} - (n - 10N) \right\}$$

$$\text{Donde, } N_{max} = \max\{\min(\lfloor n/20 \rfloor, N^*), 1\}$$

Es importante tener en cuenta que el selector de ancho de banda por medio la regla de oro (Rule of Thumb) asume una varianza constante, debido a una estimación simple de ésta, y por lo tanto no están disponibles (sin modificaciones) en el tipo de ancho de banda variable.

5.4.5.2. DIRECT PLUG-IN

El método de selección de ancho de banda Direct Plug In (DPI) desarrolla, básicamente, la misma idea que los selectores ROT, es decir, sustituyen las estimaciones de las cantidades desconocidas dentro de la fórmula ancho de banda óptima basada en el Error Cuadrático Medio Integrado Asintótico AIMSE. La principal diferencia en los métodos de generación de DPI y los métodos de ROT es que los métodos DPI usan estimaciones funcionales suaves para calcular las incógnitas. El problema aquí es que estas estimaciones requieren algún tipo de ancho de banda de piloto. Ruppert et al. (1995) sugieren elegir el ancho de banda piloto por por medio del método ROT. La expresión de la ventana h_{AMSE} que minimiza el error cuadrático medio asintótico puntual (AMSE, por su siglas en inglés) del estimador lineal local $\hat{m}(x)$ como estimador de $m(x)$.

¹⁶ La Tabla 3.2 de Fan & Gijbels (1996) enumera algunos de los valores que toma $C_{j,p}(k)$

$$\hat{h}_{DPI} = C_1(k) \left\{ \frac{\hat{\sigma}_1^2(\lambda_{AMSE})(b-a)}{\hat{\theta}_{22}^{0.5}(\hat{g}_{AMSE})n} \right\}^{1/5} \quad (25)$$

Dónde:

- $C_1(k) = \{1/(2\sqrt{\pi})\}^{1/5}$ es un número real que depende de un Kernel k .
- $\hat{\lambda}_{AMSE} = C_3(k) \left\{ \frac{\hat{\sigma}_1^4(\hat{N})(b-a)}{\hat{\theta}_{22}^{0.5}(\hat{g}_{AMSE})^2 n^2} \right\}^{1/9}$
- $\hat{g}_{AMSE} = C_2(k) \left\{ \frac{\hat{\sigma}_Q^2(\hat{N})(b-a)}{|\hat{\theta}_{24}^Q(\hat{N})|n} \right\}^{1/7}$

5.4.5.3. CRITERIO DE LA VALIDACIÓN CRUZADA (CV)

Es un estimador casi insesgado del Error Cuadrático Integrado Medio 'MISE'(Eubank, 1999). Este método no involucra estimaciones de la varianza ni de la segunda derivada de la curva de regresión. El procedimiento que realiza es dejar por fuera una pareja de observaciones $(x_i; y_i)$, se construye un modelo sin esta pareja y se usa el modelo para predecir la observación y_i , utilizando x_i que han quedado por fuera. La función de validación cruzada está definida por:

$$CV(h) = n^{-1} \sum_{i=1}^n [y_i - \hat{m}_{-i}(x_i)]^2 \quad (26)$$

En donde, $\hat{m}_{-i}(x_i)$ denota una estimación suavizada cuando la pareja $(x_i; y_i)$ ha sido omitida del conjunto de datos. Se selecciona el ancho de banda que minimice esta medida.

La función CV es llamada validación cruzada porque valida la habilidad de predecir $\{y_i\}_{i=1}^n$ con las submuestras $(x_i; y_i)_{i \neq j}$. La idea está relacionada con la selección de variables en la regresión lineal.

No obstante, Hart (1991) demuestra que, cuando los datos están correlacionados positivamente, el criterio de CV elegirá h tal que la estimación del kernel muy cerca de una interpolación de los datos.

Para superar esta limitación, Altman(1990) propuso el método de la validación cruzada generalizada, para estimar el parámetro de suavización h . El cual tiene un uso muy importante cuando los datos de una serie de tiempo muestran regularidades fuertes en su comportamiento.

5.4.5.4. VALIDACIÓN CRUZADA GENERALIZADA (GCV).

Este procedimiento hace uso del estimador kernel (4) para una función suave $m(\cdot)$ asumiendo que los errores e_t provienen de un débil proceso estacionario con función de covarianza:

$$E[e_t, e_{t+k}] = \sigma^2 \rho_n(|k|) \quad t = 1, 2, \dots, n \quad (27)$$

Convarianza del error σ^2 y función de correlación $\rho_n(k)$ de tal manera que:

$$\lim_{n \rightarrow \infty} \sum_{k=1}^n |\rho_n(k)| < \infty \quad \lim_{n \rightarrow \infty} \sum_{k=1}^n k |\rho_n(k)| = 0 \quad (27.1)$$

En donde, la primera condición garantiza que observaciones suficientemente distantes no se correlacionan.

Fue planteado inicialmente por Craven & Wahba (1979) y tiene como propósito minimizar esta función para seleccionar el ancho de banda (h) óptimo, teniendo en cuenta los residuos cuadrados de (16), se expresa de la siguiente forma:

$$GCV(h) = \frac{RSS(h)}{(1 - n^{-1} \text{tr}(S))^2} \quad (28)$$

Donde,

- $RSS(h) = \sum_{i=1}^n (y_i - \widehat{m}_i)^2$,
- $\text{tr}(S)$ es la traza de la matriz de suavización de tamaño $n \times n$.

Fernández & Opsomer (2005), plantean que no se debe utilizar este criterio para seleccionar directamente el ancho de banda en la presencia de errores correlacionados, ya que su esperanza se ve seriamente afectada por la correlación.

En concreto, Liu (2001) demostró que la GCV es asintóticamente sesgada, en el sentido de que

$$E\{GCV(h)\} = \sigma^2 + MISE(h) - \frac{2\sigma^2 K(0)\rho_I}{nh} + O_p\left(\lambda_{max}^4(h) + \frac{1}{nh}\right) \quad (29)$$

Por lo tanto, cuando $\rho_I \neq 0$, es decir, existe presencia de correlación, ésta aporta un término adicional en el valor esperado del criterio GCV que depende de h y es del mismo orden que el componente de varianza de MISE que es probable que conduzca a elecciones inadecuadas de ancho de banda. (Niglio & Perna, 2003; Fernández & Opsomer, 2005; Liu 2001)

Altman (1990) proponen eliminar este efecto usando una modificación del criterio de Validación Cruzada Generalizada (GCV) al cual denomina Validación Cruzada Generalizada método “directo” (GCV_d). Este método corrige la GCV de tal manera que la función desconocida se construye utilizando los siguientes cuadrados de los residuos:

$$GCV(h)_d = \frac{RSS(h)}{(1 - n^{-1} \text{tr}(SR))^2} \quad (30)$$

Donde, S es la matriz cuadrada ($n \times n$) de pesos kernel, $tr(\cdot)$ es la traza de la matriz y R es la matriz de correlación de los residuales.

Fernández & Opsomer (2005) utilizan los resultados obtenidos por Liu (2001) para demostrar que la GCV_d es asintóticamente insesgada, en el sentido de que:

$$E\{GCV(h)\} = \sigma^2 + MISE(h) + O_p\left(\lambda_{max}^4(h) + \frac{1}{nh}\right) \quad (30.1)$$

Un segundo procedimiento denominado $GCV - "indirecto"$ propuesto por Altman (1990) transforma los residuales para limitar su relación lineal. En este caso $RSS(h)$ se convierte en $RSS^*(h) = R^{-1/2} \times RSS(h)$ llegando a la siguiente expresión:

$$GCV(h)_i = \frac{RSS^*(h)}{(1 - n^{-1} tr(SR))^2} \quad (31)$$

No obstante, el criterio (30) y (31) aún no son métodos de selección de ancho de banda prácticos, ya que requiere el conocimiento de la matriz de correlación de los residuales R . Es por esta razón que se debe suponer una forma paramétrica de la función de correlación de los residuales, expresada por $\rho_n(\cdot)$ para poder estimar dicha matriz. Bowman & Azzalini (1997) manifiestan que usar las covarianzas de la muestra para estimar esta matriz R podría producir resultados incoherentes cuando el rezago de los términos es comparable a n . Por lo cual, sugiere ajustar un proceso autorregresivo de orden inferior a los residuales de una estimación preliminar $m(\cdot)$ obtenida por medio de un parámetro de suavizado que sobre estime la curva de regresión de los datos y mediante ese ajuste calcular todos los términos restantes de la función de covarianza desde la función de covarianza teórica de la autorregresión estimada.

5.4.6. ESTIMACIÓN DE LA VARIANZA

De acuerdo con Bowman & Azzalini, (1997). En todo análisis estadístico es fundamental conocer el parámetro de varianza, no obstante, en la práctica se hace necesario su estimación ante su desconocimiento. En la regresión no paramétrica la estimación de la varianza es también uno de los principales objetivos, dado que además de brindar información sobre la dispersión del conjunto de datos, es necesaria para la construcción de intervalos de predicción; para esto suelen usarse estimadores que no dependen del parámetro de suavización, evitando de esta manera la influencia de los elementos determinísticos del modelo. Rice (1984)¹⁷ propone un estimador que se basa en las diferencias de la variable de respuesta para eliminar los principales efectos de la curva media subyacente o tendencia que puedan tener los datos. Este estimador es de la forma:

¹⁷ Citado por Bowman y Azzalini (1997)

$$\sigma^2 = \frac{1}{2(n-1)} \sum_{i=2}^n (y_i - y_{i-1})^2 \quad (32)$$

Se asume por notación que las observaciones $(x_i; y_i)$ se encuentran ordenadas por x_i . Una desventaja de este enfoque es que las diferencias $(y_i - y_{i-1})$ son influenciadas por la fluctuación en la función de regresión desconocida, por lo que la estimación de σ^2 podría incrementarse, no obstante, Bowman y Azzalini (1997) indican que el sesgo tiende a cero cuando se tienen grandes tamaños de muestra.

Gasser, Sroka, & Jennen-Steinmetz (1986) proponen eliminar gran parte de los inconvenientes generados por el método de Rice, empleando una forma de interpolación lineal mediante pseudo-residuales, los cuales miden la diferencia entre la respuesta y_i y la línea que une a sus dos vecinos inmediatos. Estos pseudo-residuales son definidos de la siguiente forma:

$$\tilde{\varepsilon}_i = a_i y_{i-1} + b_i y_{i+1} - y_i \quad (33)$$

En donde,

$$a_i = \frac{x_{i+1} - x_i}{x_{i+1} - x_{i-1}} \quad y \quad b_i = \frac{x_i - x_{i-1}}{x_{i+1} - x_{i-1}}$$

Por lo que el estimador de varianza GSJS se expresa de la siguiente forma:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n-1} c_i^2 \tilde{\varepsilon}_i^2 \quad (34)$$

En donde, $c_i^2 = a_i^2 (+b_i^2 + 1)^{-1}$.

Esta aproximación elimina en buena medida la influencia de la función de regresión subyacente. Existen otros estimadores propuestos por diferentes autores dentro de los cuales se puede consultar por ejemplo Hall & Marron (1990).

5.4.7. INTERVALOS DE CONFIANZA Y BANDAS DE VARIABILIDAD

La idea de los intervalos de confianza surge de la necesidad de conocer la precisión y confiabilidad de la estimación de interés. No obstante, en la regresión no paramétrica, el sesgo que involucra los estimadores de la curva de regresión no permite la construcción de intervalos de confianza como en regresión paramétrica, debido a que el sesgo involucra la segunda derivada de la curva de regresión lo cual resulta complejo de estimar. Estas bandas son una medida que indica 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 $\hat{m}(x)$. La banda superior indica dos errores estándar por encima y la banda inferior indica dos errores estándar por debajo de la estimación de la curva $\hat{m}(x)$.

$$\begin{aligned} \text{Bandas} &= \langle \text{LI}, \text{LS} \rangle \\ &= \langle \hat{m}(x) - 2 * \sqrt{\text{var}\{\hat{m}(x)\}} ; \hat{m}(x) + 2 * \sqrt{\text{var}\{\hat{m}(x)\}} \rangle \end{aligned} \quad (35)$$

El error estándar de la curva de regresión está dado por:

$$se\{\hat{m}(x)\} = \sqrt{\text{var}\{\hat{m}(x)\}} = \sqrt{\text{diag}(S * S^T) \hat{\sigma}^2} \quad (36)$$

En donde, S es la matriz de suavizamiento para el estimador lineal y $\hat{\sigma}^2$ es la estimación de varianza de los errores a partir de un estimador no paramétrico.

Las bandas de variabilidad asociadas a cada curva de regresión estimada, muestran la región en la cual se podría encontrar la verdadera curva, pero limitando esta clase de interpretaciones debido al sesgo involucrado en las estimaciones de funciones no paramétrica. Bandas de este tipo son fáciles de construir pero requieren una interpretación muy cuidadosa. Formalmente las bandas indican intervalos de confianza punto a punto para $E\{\hat{m}(x)\}$ en lugar de $m(x)$. Una estimación de la varianza de $\hat{m}(x)$ es todo lo que se requiere.

Otra forma de construir estas bandas de variabilidad, es a través de una metodología de remuestreo denominada Bootstrap, la cual consiste en tratar la muestra como si fuese la población y aplicar un muestreo con reposición para generar una estimación empírica de la distribución muestral del estadístico (en este caso $\hat{m}(x)$). Esta técnica no paramétrica, tiene la ventaja de que no precisa conocer la función de distribución teórica de los datos.

Las bandas de variabilidad se construyen seleccionando (repetidamente) muestras con remplazo de los datos originales y reajustando las curvas utilizando Regresión Polinómica Local de grado 1 para cada nueva muestra. Luego se calcula el percentil del 2.5% y del 97.5% a través de todas las muestras y se unen para todos los valores de x (Servy, et al., 2006).

Las curvas resultantes no corresponden a un intervalo de confianza del 95% ya que en su construcción se está ignorando el sesgo en la estimación de la curva. Unas “bandas” angostas (o cercanas a $\hat{m}(x)$) sugieren una mayor estabilidad en las estimaciones. Si se utiliza una función Kernel Gausiana, estas bandas se ensanchan en los extremos del campo de variación de x .

6. METODOLOGÍA DE INVESTIGACIÓN

6.1. REVISIÓN BIBLIOGRÁFICA

Para poder llevar a cabo los objetivos de esta investigación fue necesario consultar distintas fuentes de información. La primera fuente consultada fue Internet, de donde se obtuvo información general relacionada con la contaminación atmosférica por CO, la regresión No paramétrica y la estimación del ancho de banda en presencia de errores autocorrelacionados, entre otros. También se realizó una consulta exhaustiva en libros y artículos sobre la metodología a implementar, tomando como referencia trabajos de investigación realizados por egresados del programa académico de estadística. Las principales fuentes bibliográficas se relacionan con autores como el Ph.D en estadística Randall L. Eubank, actualmente es profesor del departamento de matemáticas y estadísticas de la universidad estatal de Arizona USA. El Ph.D en estadística Jean Opsomer, actualmente es profesor del departamento de estadísticas de la universidad estatal de Colorado USA. La Ph.D en estadística Naomi S. Altman, profesora asociada al Departamento de estadísticas de la universidad del estado de Pennsylvania USA y el profesor de la escuela de matemáticas y estadísticas de la universidad de Glasgow en Scotland UK, Adrián Bowman. Entre otros.

6.2. RECOPIACIÓN DE DATOS

La información para este proyecto proviene de la Red de Monitoreo de Calidad del Aire (RMCA) de Santiago de Cali, administrada por el DAGMA y constituida por las estaciones BA, ERA, PDD, HUV, CDAV y CALLE15. Dicha información contiene, entre otros datos, los registros de los niveles de concentración promedio horaria de CO (ppm) para el periodo abril de 2003 hasta abril de 2006.

Los registros utilizados para conformar la base de datos provienen de la depuración y validación de los datos que realizaron Reina García, J. y Castro, D. A. en el año 2009. Debido a que la red de monitoreo reanudó sus operaciones en el mes de abril del año 2003, la información recolectada durante ese periodo no se tendrá en cuenta, porque es considerada como poco confiable.

La base de datos para realizar el estudio contó con información de los registros de las estaciones BA, ERA y CALLE15 (Mayo a Diciembre), HUV (Mayo a Julio), PDD (Mayo a Diciembre sin Septiembre) en el año 2003. También contó con información de las estaciones de monitoreo CDAV (Marzo a Diciembre), BA y ERA (Enero a Diciembre) PDD (Enero a Marzo), y CALLE15 (Enero a Agosto) en el año 2004. Contó además con los registros de los meses de octubre a diciembre del año 2005 en la estación CALLE 15. Y con la información de las estaciones ERA y CALLE15 (Febrero y Marzo) en el año 2006.

Los registros de los niveles de concentración de monóxido de carbono son valores promediados cada hora, es decir, valores de concentración horaria (1-h) medidos en ppm. El sistema de monitoreo toma 360 datos por hora; información recolectada cada 10 segundos. Para efectos de análisis, el formato de la hora en la cual se capturó el registro de los valores de CO fue modificado a formato numérico en este trabajo, de tal manera que la hora 0 representa al periodo de tiempo transcurrido entre la 0 hora y la 1:00 a.m. y la hora 23 es la última hora del día a partir de las 11:00 p.m.

Trabajos previos como los de Montoya y Morales (2004), Castro (2009) y Reina J. (2009) han utilizado como variable independiente las horas de un día típico (la cual contiene 24 puntos de diseño) para modelar el comportamiento de ciertos contaminantes en la ciudad de Santiago de Cali.

Teniendo en cuenta que para periodos cortos de dependencia los estimadores Kernel no se ven afectados (Opsomer, et al., 2001). Se considera utilizar una variable independiente que cuente con un rango más amplio que el de la variable horas de un día típico, utilizada en los trabajos antes mencionados. Esto con el fin de evidenciar el efecto de la correlación en la estimación de la curva al emplear diferentes métodos de selección de ancho de banda los cuales contemplen o no la estructura de autocorrelación de los errores.

Por lo anterior, la variable independiente que se utilizará se define como Hora-Semana (HS), la cual consta de 168 puntos de diseño. Y la variable dependiente será denominada concentración promedio octohoraria de Monóxido de Carbono (8h).

Tabla 5. Numero de Semanas (Tomadas – Eliminadas - Totales) por estación y año.

Año	Estación																	
	CDAV			BA			HUV			PDD			ERA			CALLE15		
	TS	SE	ST	TS	SE	ST	TS	SE	ST	TS	SE	ST	TS	SE	ST	TS	SE	ST
2003				36	7 19%	29 81%	12	1 8%	11 92%	32	8 25%	24 75%	36	4 11%	32 89%	33	3 9%	30 91%
2004	40	4 10%	36 90%	46	10 22%	36 78%				11	2 18%	9 82%	53	3 6%	50 94%	33	4 12%	29 88%
2005	6	4 67%	2 33%										4	3 75%	1 25%	13	1 8%	12 92%
2006	8	3 38%	5 63%	1	1 100%	0 0%							9	1 11%	8 89%	10	2 20%	8 80%
TS: Total Semanas SE: Semanas Eliminadas ST: Semanas Tomadas																		

De acuerdo a lo expuesto anteriormente, se determinó eliminar de la base de datos las semanas que presentaban comportamientos extremadamente atípicos, o que no tuvieran al menos el 75% de los registros. La Tabla 5, muestra el total de semanas por estación que conformaban la base de datos original y el número de semanas que se eliminó.

6.3. SOFTWARE ESTADÍSTICO

Para dar cumplimiento a los objetivos planteados en este proyecto se utilizó el software estadístico R GUI versión 2.12.2 de libre distribución para realizar los cálculos de los diversos métodos estadísticos e implementar técnicas de suavización para obtener los resultados planteados en este proyecto. Se utilizó la función `locpoly` de la librería `KernSmooth`, desarrollada por Wand & Jones (1995), la cual estima una función de densidad de probabilidad o una función de regresión o sus derivados mediante polinomios locales para datos igualmente espaciados. Para la elección del ancho de banda se utilizaron las funciones: `dpill` de la librería `KernSmooth`, que desarrolla la metodología Direct Plug-in descrita por Ruppert, Sheather y Wand (1995); `thumbBw` de la librería `locpol`, la cual implementa la metodología propuesta por Fan and Gijbels(1996) denominada Rule of thumb; `regCVBwSelC` de la librería `locpol`, la cual realiza validación cruzada. Y finalmente la función `sm.regression.autocor` con la cual se obtiene una validación cruzada generalizada utilizando el método “*directo*” propuesto por Almtan (1990). El equipo computacional utilizado fue un procesador Intel® CORE i5 de 2.3 GHz y 4 GB de Memoria RAM.

6.4. ANÁLISIS EXPLORATORIO

Contando con la base de datos depurada y validada previamente, se calculan los promedios Octohorarios (8-h) que serán los datos utilizados en esta investigación. Luego se realiza un análisis inicial de la información mediante métodos gráficos y estadísticos para describir el comportamiento de estos datos.

6.5. ESTIMACIÓN DE CURVAS TÍPICAS

Para realizar la estimación de las curvas típicas de la concentración del Monóxido de Carbono en una semana típica en la ciudad de Santiago de Cali, se obtendrán las concentraciones octohorarias (8-h) de CO a partir de un mínimo de 6 datos de las concentraciones promedio horarias (1-h) de CO en un periodo de ocho horas. Dicha “*concentración octohoraria de monóxido de carbono (8-h de CO)*” será utilizada como la variable de respuesta medida en partes por millón (ppm) y la línea de tiempo “*hora-semana*” definida como 168 horas de una semana típica se utilizará como covariable al emplear la suavización Kernel la cual se ajusta de manera adecuada al conjunto de datos, dada

la falta de información a priori sobre la forma que define la estructura funcional de los datos, tal como lo menciona Ayala y Ojeda (2007).

Para realizar la suavización Kernel se utilizó el estimador lineal local de grado uno el cual posee algunas propiedades de la regresión lineal convencional, por ser un caso particular de esta y poseer un diseño adaptable. Tal como lo menciona Wand & Jones (1995)

La elección del ancho de banda es un paso fundamental en la regresión no paramétrica, ya que éste logra un balance óptimo entre el sesgo y la varianza del estimador a utilizar. Los seleccionadores de anchos de banda que se utilizarán son; Direct Plug-in (DPI) descrita por Ruppert, Sheather y Wand (1995); Rule of thumb (ROT) propuesta por Fan and Gijbels (1996); la validación Cruzada (CV) Y finalmente la Validación Cruzada Generalizada – método “*directo*” propuesta por Altman (1990).

Para estimar la matriz de autocorrelaciones de los residuales que utiliza el método de selección *directo*-GCV, se aplicará la metodología planteada por Bowman & Azzalini(1997) la cual consiste en ajustar los residuales de una sub-suavización preliminar de los datos a un proceso autorregresivo de orden 2. Para esto se utilizará un ancho de banda que sobre estime la curva de regresión inicial tal como lo sugiere Bowman & Azzalini(1997), en este caso se utilizó un h igual a 20.

Esta la metodología indica que una vez calculados los residuales el paso a seguir es calcular los 3 primeros valores de la función de autocovarianzas, para esto se utilizara la siguiente expresión $\hat{\gamma}_k = n^{-1} \sum_{k=0}^2 r_{n-k} r_{k+1}$ en donde r representa un valor residual. Seguido a este procedimiento se estiman los parámetros del proceso autoregresivo de orden 2 los cuales se definen así: $a_1 = \hat{\gamma}_1 \times (1 - \hat{\gamma}_2) / (1 - \hat{\gamma}_1)^2$ y $a_2 = (\hat{\gamma}_2 - \hat{\gamma}_1^2) / (1 - \hat{\gamma}_1)^2$. Cabe resaltar que los valores restantes de la función de autocovarianzas se obtendrán utilizando la siguiente expresión $\hat{\gamma}_k = a_1 \times r_k + a_2 \times r_{k-1}$ para $k = 1, 2, \dots, n$. La cual corresponde a la definición de la función de autocovarianzas para un proceso autorregresivo de orden 2.

Finalmente la matriz de autocorrelaciones se obtiene al ajustar en cada columna de esta matriz los rezagos de la función de autocorrelaciones, la cual se calcula aplicando la siguiente definición $\hat{\rho}_k = \hat{\gamma}_k / \hat{\gamma}_0$ para $k = 1, 2, \dots, n$. En donde $\hat{\gamma}_0$ representará la varianza de este proceso.

Para obtener las bandas de variabilidad, se aplicará, con la ayuda del lenguaje de programación estadística R, la metodología Bootstrap puesto que al ser una técnica no paramétrica, tiene la ventaja de que no precisa conocer la función de distribución teórica de los datos. (Servy, et al., 2006)

Para la estimación de la varianza inicialmente se consideró utilizar un estimador que contemplara la correlación de los errores y múltiples observaciones, en un mismo punto de diseño. Sin embargo, en la búsqueda bibliográfica realizada, no se encontró un método que

permitiera hacer dicha estimación. No obstante, Reina (2009) y Castro (2009) en sus trabajos de investigación, utilizan el método propuesto por Gasser, Sroka y Jennen Steinmetz (GSJ) para realizar la estimación de la varianza bajo el escenario de múltiples observaciones en un mismo punto de diseño. Este método remueve gran parte de las fluctuaciones abruptas en la función de regresión, teniendo en cuenta la discrepancia entre los puntos de diseño perdiendo, solamente la primera y la última observación.

Por lo anterior, para este proyecto, finalmente se consideró implementar el estimador GSJ para estimar la varianza de las observaciones. Aunque, este método no es el más adecuado para aplicar en las condiciones es que se presentan observaciones.

6.6. VALIDACIÓN DE SUPUESTOS

Luego de estimar las curvas de regresión, es importante validar el cumplimiento de los supuestos de los residuales¹⁸ (idénticamente distribuidos con media cero, varianza constante e independencia) en los que se basan estos modelos. Sin embargo, por ser métodos robustos se espera que si no se llega a cumplir alguno de estos supuestos, el modelo de regresión estimado no se vea afectado de manera significativa de acuerdo a lo expuesto por Fan & Gijbels (1996), Wand & Jones (1995) y Fan and & Yao (2003). De acuerdo a la revisión bibliográfica realizada previamente, se puede evidenciar que el método más común utilizado por algunos autores entre ellos Delicado, (2008) y Loader (1999) para realizar dicha verificación es graficar la dispersión de los residuales a lo largo del tiempo. Bowman y Azzalini (1997) proponen para evaluar el cumplimiento del supuesto de que los residuales tienen media cero, realizar una suavización de los residuales con el fin de observar si esta suavización se mantiene cercana a cero, si esto es así, indicaría que no se viola este supuesto.

6.7. VALIDACIÓN DEL COMPORTAMIENTO DE LOS SELECTORES DE ANCHO DE BANDA.

Dado que la base de datos presenta un panorama de múltiples respuestas por cada punto de diseño (mediciones repetidas con datos faltantes) se validará el comportamiento de los selectores de ancho de banda al utilizar una serie con datos (completos y otra incompletos) en el tiempo, en la estimación de la curva de regresión no paramétrica. Para esto se utilizarán las observaciones 8-h de CO registradas en una semana típica que será escogida al azar dentro de cada estación por año.

¹⁸Se conocen también como los errores del modelo de regresión.

7. RESULTADOS

A continuación, se presentan los resultados obtenidos al realizar el análisis descriptivo del comportamiento de las concentraciones promedio octohorarias de monóxido de carbono (8h-CO) para una semana típica en las estaciones de medición de calidad del aire. Así mismo, se presentan los resultados de la modelación del comportamiento de la contaminación atmosférica por Monóxido de Carbono en una semana típica mediante un modelo de regresión no paramétrica con errores correlacionados, utilizando para ello todos los registros de las concentraciones promedio octohorarias (por estación de medición) durante el periodo 2003 - 2006. Y finalmente, se muestran los resultados obtenidos al examinar el comportamiento de los selectores de ancho de banda cuando se aplican a las mediciones (completas e incompletas) realizadas en una semana.

7.1. ANÁLISIS EXPLORATORIO DE LAS BASES DE DATOS

El siguiente análisis descriptivo del comportamiento de las concentraciones promedio octohorarias de monóxido de carbono (8h-CO) se realizó para las estaciones de la calidad del aire de la ciudad que presentaron mayor información durante el periodo 2003 - 2006 para una semana típica (la estación Calle 15 y ERA).

7.1.1. ANÁLISIS DESCRIPTIVO PARA LA ESTACIÓN CALLE 15 – CLL.15

Como se puede observar en la Tabla 6, los mayores niveles de las concentraciones promedio octohorarias de monóxido de carbono (8h-CO) registrados en la estación Calle 15 durante los años 2003 al 2006 no superaron el límite máximo permisible según lo estipulado en la legislación colombiana (8.8 ppm).

La Figura 3 permite apreciar que para cada día de la semana, la variabilidad en los niveles máximos de las concentraciones 8h-CO registrados año a año presentan una tendencia a la

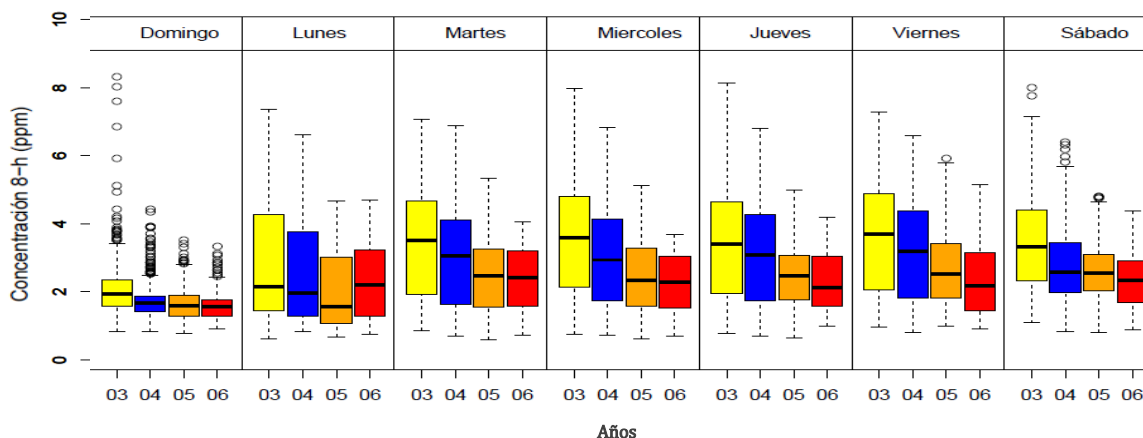
disminución. También se aprecia que las medianas de dichas concentraciones disminuyen anualmente en todos los días de la semana, salvo el día lunes del año 2006 en el cual se presenta un importante incremento en la mediana de las concentraciones 8h-CO. De igual forma se evidencia que día sábado, las concentraciones promedio octohorarias de CO empiezan a disminuir, alcanzando el día domingo, los niveles mas bajos de toda la semana.

Otro aspecto que se puede observar es la presencia de unos cuantos registros atípicos el día sábado y en mayor medida el día domingo, el cual registró para el año 2003, el máximo nivel de concentración (8h-CO) durante todo el periodo de estudio. En contraste, el nivel mas bajo fue registrado un día martes del año 2005. Se podría llegar a pensar que la reducción en la emisión de gases tóxicos al ambiente los días sábado y domingo está asociada a la reducción en la jornada laboral de las empresas, puesto que el sábado, en la mayoría de las empresas, se trabaja media jornada y el domingo se descansa todo el día, lo que implicaría una disminución en el flujo vehicular habitual en esta zona céntrica de la ciudad. También sería razonable suponer que se presente un aumento en de los niveles de las concentraciones 8h-CO a partir del día lunes debido a la reanudación de las actividades normales tanto en las empresas como en la circulación vehicular. Lo anterior da lugar para pensar que las distribuciones de las concentraciones promedio octohorarias de monóxido de carbono para los días martes, miércoles, jueves y viernes (en cada año) presentan un comportamiento muy similar. Por lo cual se podría utilizar, en estudios que así lo requieran, los registros de estos días para formar una sola variable que describa el comportamiento de los días ordinarios de la semana. Vale aclarar que existe la posibilidad de que variables externas que no hayan sido tenidas en cuenta en el estudio (no controladas) estén ocasionando este fenómeno.

Tabla 6. Estadísticas descriptivas de las concentraciones promedio octohorarias de monóxido de carbono (8h-CO) para la estación calle 15 durante los años 2003 al 2006.

Año	Min.	Q1.	Mediana	Media	Q3.	Max.
2003	0.550	1.768	2.860	3.162	4.445	8.410
2004	0.613	1.500	2.430	2.723	3.825	6.936
2005	0.513	1.436	2.188	2.304	3.024	5.948
2006	0.616	1.378	2.030	2.190	2.983	5.159

Figura 3.Diagrama de cajas y alambre de concentraciones octohorarias (8-h) para la estación calle 15 durante los años 2003 al 2006



7.1.2. ANÁLISIS DESCRIPTIVO PARA LA ESTACIÓN ESCUELA REPÚBLICA DE ARGENTINA - ERA

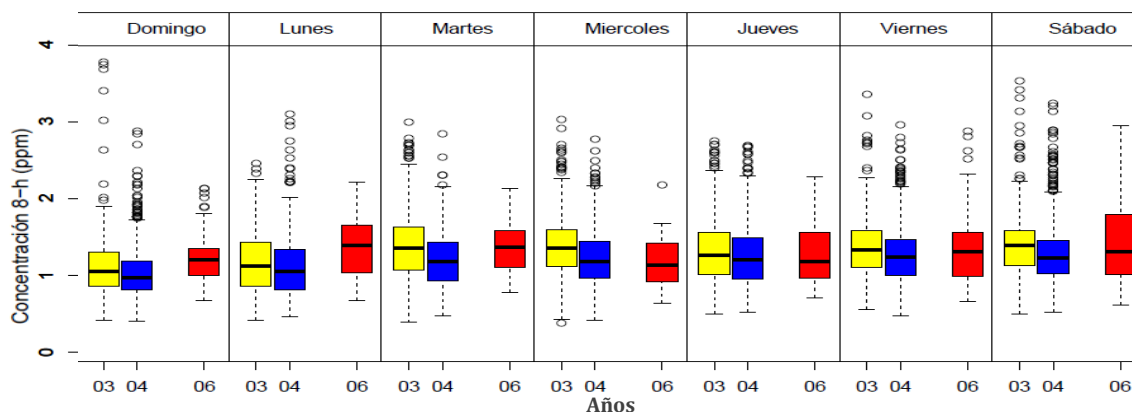
Al igual que en la estación la estación Calle 15, la estación de medición de calidad del aire ERA durante los años 2003, 2004 y 2006. No evidenció observaciones que superarán el límite máximo permisible (ver Tabla 7).

Al observar la Figura 4 se puede establecer que el nivel máximo de concentración 8h-CO registrado en esta estación se presentó un día domingo del año 2003, mientras que el mínimo nivel se dio un día jueves en el mismo año. Así mismo, se puede apreciar que para cada día de la semana, la variabilidad en los niveles máximos de las concentraciones 8h-CO registrados año a año presentan una tendencia a la disminución, exceptuando el día lunes del año 2004. Sin embargo, las medianas y promedios no se evidencian tendencia alguna, por el contrario los niveles que se perciben concentraciones bastante homogéneas. Por otro lado, y al igual que en la estación CALLE 15, las concentraciones promedio octohorarias de CO empiezan a disminuir en la estación ERA a partir del día sábado, alcanzando el día domingo, los niveles mas bajos de toda la semana. Se destaca el gran número de registros atípicos que presenta esta estación para los años 2003 y 2004 en todos los días de semana, específicamente los días sábado y domingo. Justamente en el año 2003, el día domingo registró el máximo y el mínimo nivel de concentración (8h-CO) durante todo el periodo de estudio.

Tabla 7. Estadísticas descriptivas de las concentraciones promedio octohorarias de monóxido de carbono (8h-CO) para la estación ERA durante los años 2003 al 2006.

Año	Min.	Q1.	Mediana	Media	Q3.	Max.
2003	0.383	1.095	1.399	1.428	1.713	4.261
2004	0.414	0.995	1.261	1.303	1.560	3.643
2006	0.651	1.082	1.401	1.444	1.711	3.319

Figura 4. Diagrama de cajas y alambre de concentraciones octohorarias (8-h) para la estación ERA durante los años 2003, 2004 y 2006



En el **ANEXO A.** se puede apreciar la distribución de los de las concentraciones octohorarias (8-h) para la estación PDD, BA, CDVA y HUV durante los años 2003 y 2004 por medio de los diagramas de cajas y alambres y las estadísticas descriptivas anuales contenidas en la Tabla 10. De allí, se puede afirmar que en el año 2003 presenta las mayores concentraciones de CO con respecto al año 2004. Además, se evidencia que los días martes a viernes muestran un comportamiento similar entre estaciones con la diferencia de que la distribución es diferente para cada año. Cabe resaltar que en ninguna de las estaciones de medición analizadas en este estudio se encontraron concentraciones 8h-CO que superaran el límite máximo permisible según lo estipulado en la legislación colombiana.

7.2. ESTIMACIÓN DE CURVAS TÍPICAS

A continuación, se presentan los resultados de la modelación del comportamiento de la contaminación atmosférica por monóxido de carbono en una semana típica de la ciudad de Santiago de Cali, por medio de regresión no paramétrica con errores correlacionados durante el periodo 2003 – 2006.

La Tabla 8 contiene los valores obtenidos para los diferentes selectores del ancho de banda (CV, DPI, ROT y CGVd) empleados en la estimación de las curvas típicas. En particular se puede apreciar que el método GCVd arrojó (en todas las estaciones) los valores más bajos para el parámetro de suavidad, este resultado indica que las curvas generadas por este criterio interpolarán los datos, lo cual se constituye como una falla del método de selección de acuerdo a lo expuesto por Opsomer, et al., (2001). Algo similar ocurre al utilizar los criterios DPI y CV los cuales no contemplan la presencia de autocorrelación de los residuales. No obstante, el criterio ROT aparece como una buena alternativa para elegir dicho parámetro, puesto que ofrece los valores más grandes frente a los demás selectores. Esto indica que este método de selección contribuirá a atenuar las distorsiones que produce la estructura de autocorrelación de los residuales en la generación de curvas suaves.

Tabla 8. Anchos de banda obtenidos en la estimación de la curva de regresión no paramétrica para una semana típica utilizando diferentes métodos de selección.

Estación y Año de medición													
Selector	CIL.15				ERA			PDD		BA		HUV	CDAV
	2003	2004	2005	2006	2003	2004	2006	2003	2004	2003	2004	2003	2004
GCVd	0,50	0,50	0,81	0,81	0,50	0,50	0,99	0,81	0,81	0,81	0,50	0,81	0,50
ROT	2.30	3.15	3.10	4.02	2.12	2.16	4.26	2.62	2.92	2.25	2.45	4.14	3.61
DPI	1,07	1,02	1,46	1,46	1,46	1,18	1,96	2,29	2,23	1,52	1,54	2,07	2,01
CV	1,14	1,09	1,59	1,54	1,21	1,03	1,74	1,87	1,49	1,17	1,09	1,71	1,30

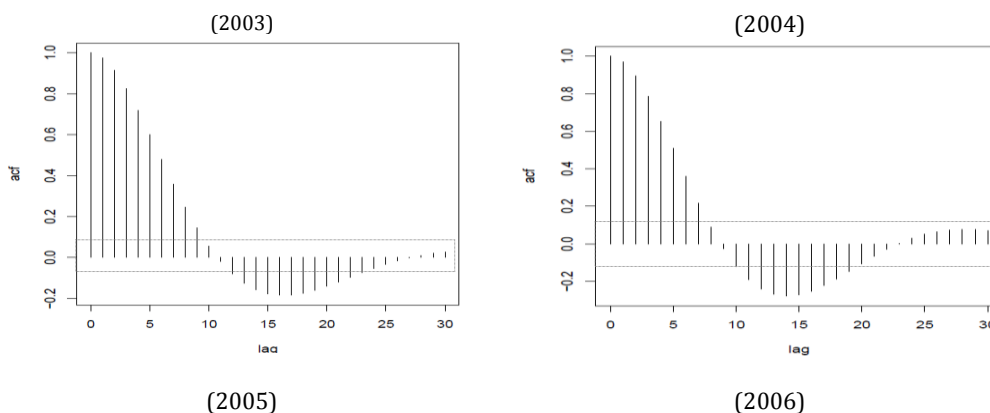
Para modelar el comportamiento de las concentraciones 8-h de Monóxido de Carbono en una semana típica en la ciudad de Santiago de Cali, utilizando regresión no paramétrica con errores correlacionados, en primer lugar se debe determinar la presencia de una estructura de correlación serial en los errores del modelo. Bowman & Azzalini(1997) mediante la metodología para estimar la matriz de autocorrelaciones descrita en el capítulo anterior, utiliza los residuales de la sub suavización preliminar de los datos ajustados a un proceso autorregresivo, para generar un correlograma con el cual evidencia la presencia o no de dicha estructura de correlación.

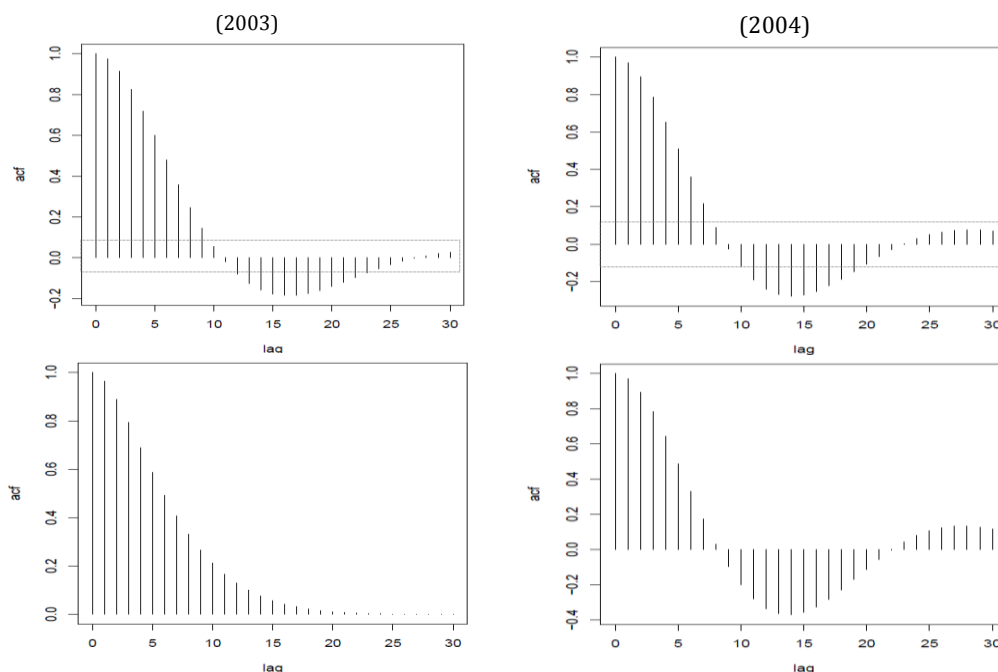
Con el fin de apreciar la influencia de la estructura de autocorrelación de los residuales en la estimación de las curvas típicas de las concentraciones promedio octohorarias de monóxido de carbono, se presenta a continuación, los resultados obtenidos en las estaciones de medición de la calidad del aire que presentaron mayor información durante el periodo 2003 – 2006.

De acuerdo con la Figura 5 se puede observar que, para cada año, los coeficientes de autocorrelación de residuales de las estimaciones preliminares de las curvas de regresión no paramétrica de las concentraciones 8-h de CO, presentan un decaimiento lento a cero y que además sobrepasan los límites de significancia, lo cual sugiere la presencia de una estructura de autocorrelación significativa en las mediciones de cada año tomadas en la estación Calle 15.

La Figura 6 Permite apreciar las curvas de regresión estimadas no paramétricamente de las concentraciones octohorarias (8-h) de CO registradas en la estación CALLE 15 durante el periodo 2003 -2006 usando diferentes selectores de anchos de banda, entre los cuales están la Validación Cruzada Generalizada directa (GCVd), Rule of Thumb (ROT), Direct plug-in (DPI) y Validación Cruzada (CV) los cuales presentan un comportamiento relativamente similar.

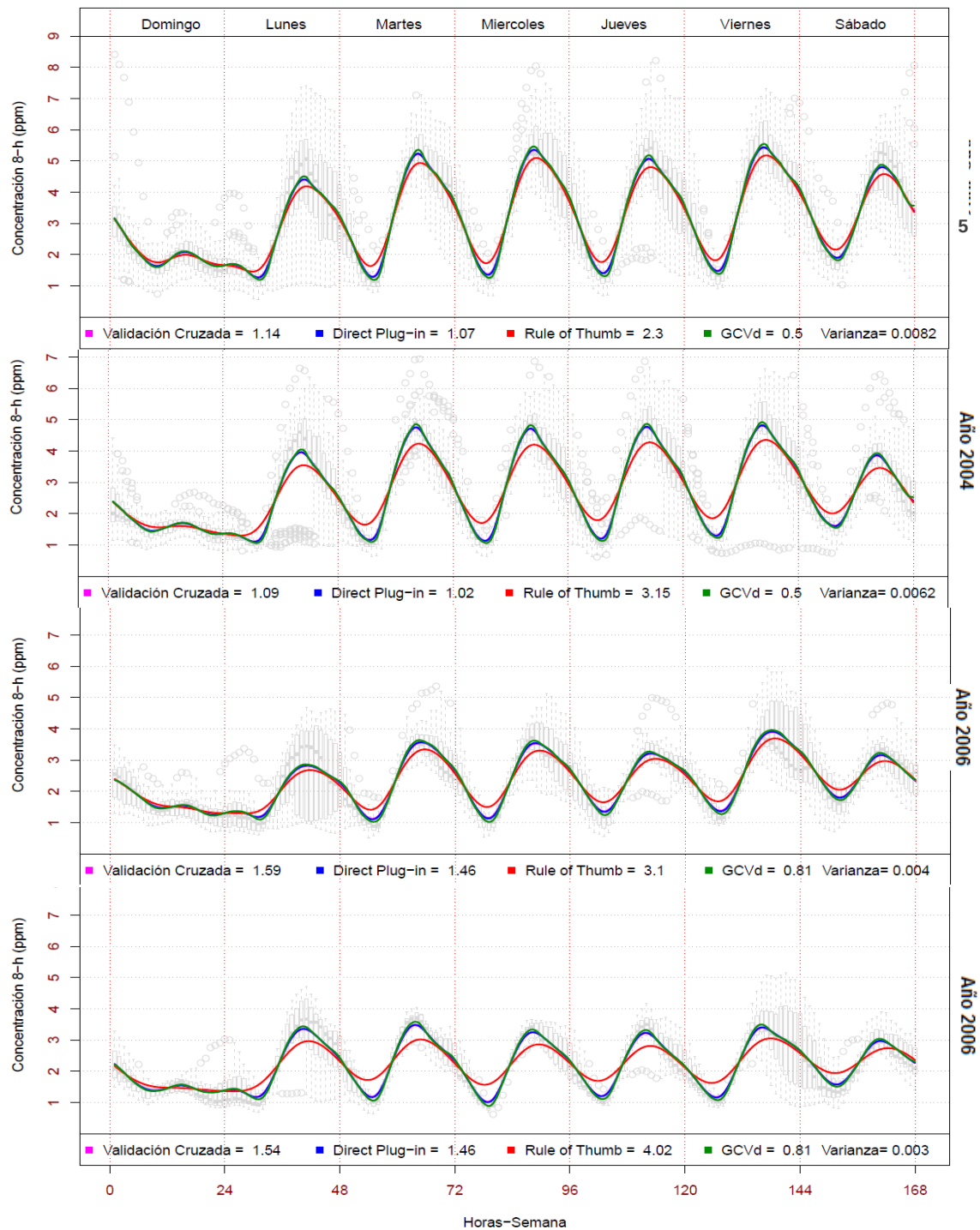
Figura 5. Funciones de autocorrelación estimadas de las suavizaciones preliminares de los residuales del modelo de regresión no paramétrico de las concentraciones octohorarias (8-h) para una semana típica en la estación Calle 15 en los años 2003, 2004, 2005 y 2006





De acuerdo a la teoría, se esperaba que los métodos CV y DPI arrojaran valores pequeños para el parámetro de suavidad, lo cual sucedió, y que el método GCVd arrojara un valor mucho mas grande para dicho parámetro, produciendo una curva más suave que mitigara el efecto de correlación que poseen los datos, pero los resultados obtenidos al aplicar este método fueron poco coherentes, ya que se presentó una interpolación de los datos. No obstante, la estimación de las curvas aplicando el método ROT evidencia una pequeña reducción de las fluctuaciones ocasionada por la correlación de los errores. Dado que produce un mayor ancho de banda el cual suaviza ligeramente las distorsiones producidas por estructura de correlación de los datos. La validación del comportamiento de los métodos de selección frente a varios escenarios realizada en el apartado anterior, evidenció que cada vez que el método GCVd arrojaba valores bajos el método ROT presentaba un comportamiento contrario, lo cual sugeriría utilizar este método cada vez que el GCVd presente dificultades con la estimación de dicho parámetro.

Figura 6. Estimación de las curvas para las concentraciones octohorarias (8-h) de CO utilizando los selectores de ancho de banda (CV, DPI, GCV y ROT) para una semana típica en la estación Calle 15 en los años 2003 y 2004



Cabe resaltar que durante la revisión bibliográfica no se encontraron antecedentes de estudios realizados con las condiciones en que se mide el fenómeno de las concentraciones octohorarias (series repetidas con datos faltantes) solamente se evidenció el caso de series univariadas en periodos relativamente cortos y sin repetición, abordado por Altman (1990), Hart(1991), Azzalini(1997), Bowman & Opsomer, et al.(2001) y Niglio & Perna, (2003), demostrando una corrección del efecto producido por la correlación (constantes fluctuaciones en la curva estimada).

Figura 7. Funciones de autocorrelación estimadas de las suavizaciones preliminares de los residuales del modelo de regresión no paramétrico de las concentraciones octohorarias (8-h) para una semana típica en la estación ERA en los años 2003, 2004 y 2006

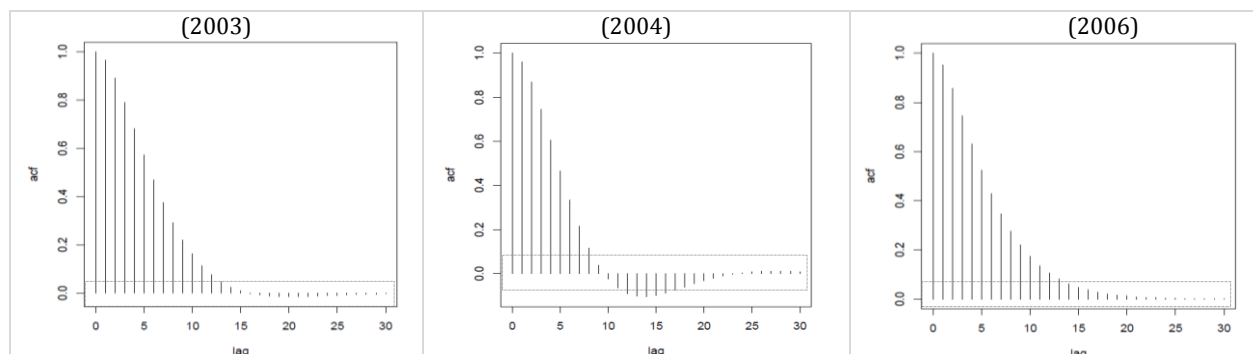
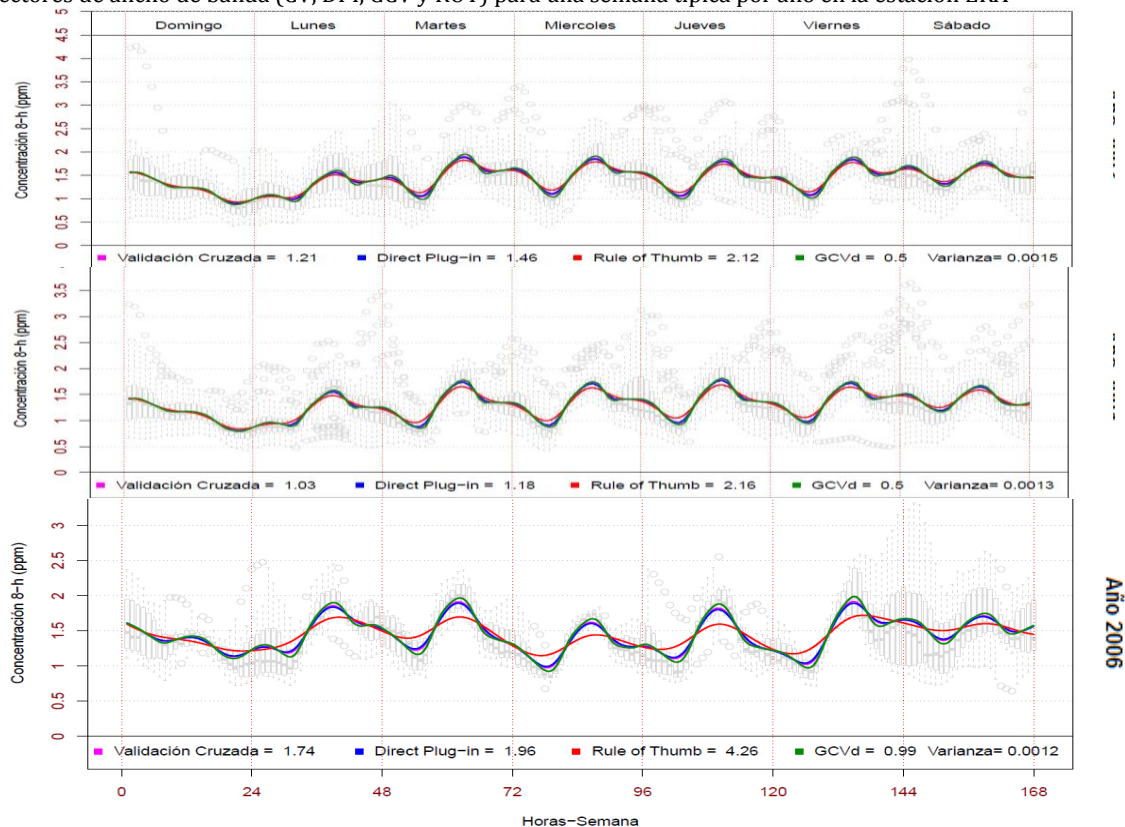


Figura 8. Estimación de las curvas típicas para las concentraciones octohorarias (8-h) de CO utilizando los selectores de ancho de banda (CV, DPI, GCV y ROT) para una semana típica por año en la estación ERA



La Figura 7 muestra las funciones de autocorrelación estimadas para los residuales las suavizaciones preliminares en la estación de medición ERA durante los años 2003 – 2006. En ella se puede apreciar que, para cada año, los coeficientes de autocorrelación de residuales de las estimaciones preliminares de las curvas de regresión no paramétrica de las concentraciones 8-h de CO, presentan un decaimiento lento a cero y que además sobrepasan los límites de significancia, lo cual sigue la presencia de una estructura de autocorrelación significativa en las mediciones de cada año tomadas en la estación ERA. La Figura 8 contiene

las curvas de regresión estimadas no paramétricamente de las concentraciones octohorarias (8-h) de CO registradas en la estación ERA durante el periodo 2003 -2006 usando diferentes selectores de anchos de banda, entre los cuales están la Validación Cruzada Generalizada directa (GCVd), Rule of Thumb (ROT), Direct plug-in (DPI) y Validación Cruzada (CV) los cuales presentan un comportamiento relativamente similar. No obstante se puede apreciar que para el año 2006 la curva generada utilizando como método de selección el ROT presenta un comportamiento más suave a comparación de las curvas generadas utilizando otros métodos. Esto indicaría una mitigación del efecto de la correlación de los residuales en la modelación de las concentraciones 8-h de CO.

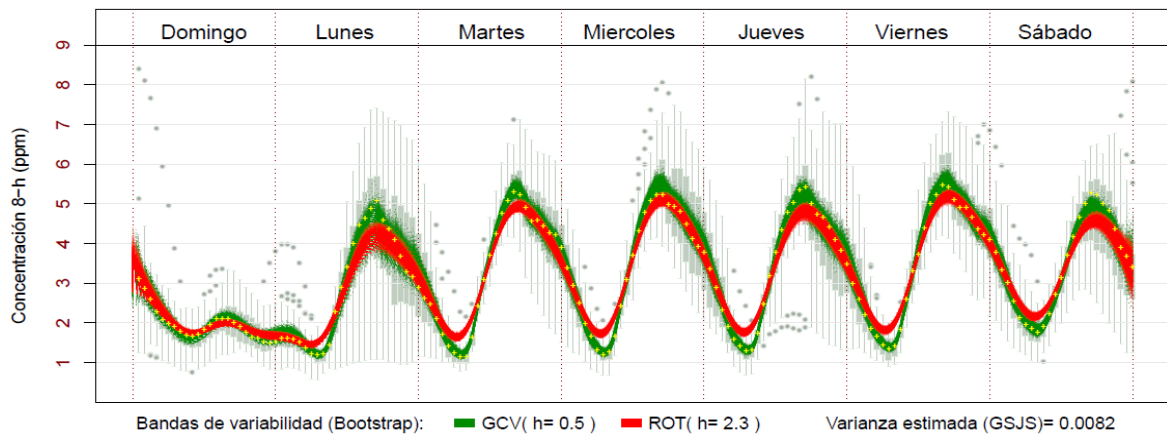
7.3. BANDAS DE VARIABILIDAD

Con el fin de proporcionar una idea de la precisión de las estimaciones de la curva típica al utilizar los selectores de ancho de bandas GCVd y ROT. A continuación se presentan los gráficos que contienen las bandas de variabilidad construidas utilizando el método de Bootstrap.

Calle 15 -CLL.15-

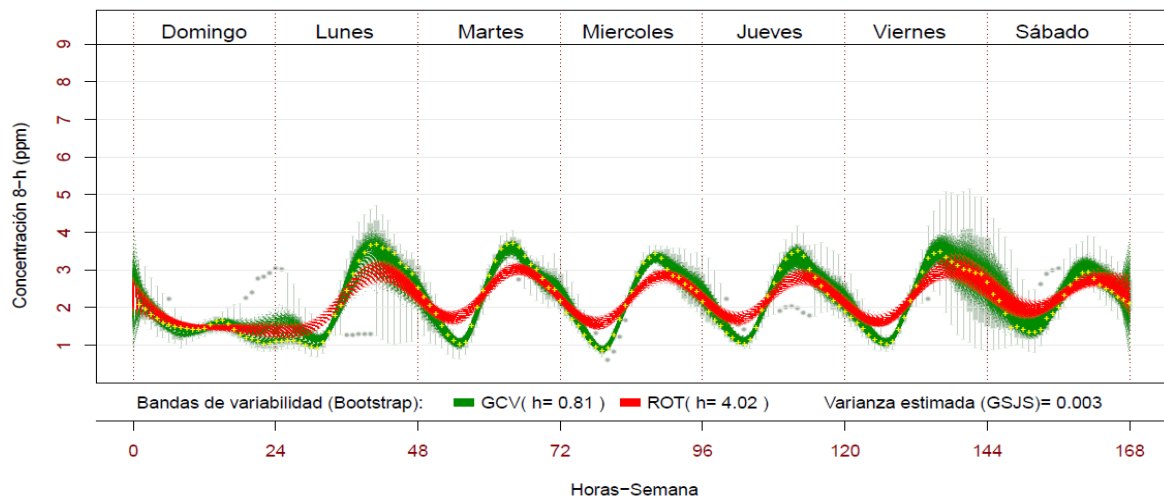
Las Figuras 9 y 10 presentan las bandas de variabilidad para las estimaciones de las curvas de las concentraciones octohorarias de CO para una semana típica en el año 2003 y 2006 para los registros de la estación calle 15 ubicada en el centro de la ciudad de Cali, mediante los seleccionadores GCVd y ROT. Como se puede apreciar las bandas de variabilidad para las curvas estimadas por los dos métodos muestran una tendencia similar aunque el método ROT presenta estimaciones más suaves frente a las realizadas utilizando el método GCVd. De acuerdo con lo anterior, las gráficas evidencian que los mayores niveles de contaminación por CO se presentan después del mediodía para durante los días lunes a viernes, descendiendo posteriormente en horas de la tarde. El nivel de concentración mínimo se obtiene en las primeras horas de las mañanas. Como se puede apreciar en las gráficas 9, aunque para el día domingo el comportamiento en las curvas sigue el mismo patrón de los días ordinarios, para este día los niveles de concentración de CO son los más bajos que se registran en la semana. En el día domingo, se presenta una reducción drástica de los niveles de CO por lo que su concentración octohoraria permanece en un nivel muy bajo comparado con resto de días. La razón de esta reducción se podría atribuir a la reducción del flujo vehicular y a que algunas empresas cesan sus actividades de producción al mínimo.

Figura 9. Bandas de Variabilidad para las curvas para las concentraciones octohorarias (8-h) estimadas mediante los selectores de ancho de banda GCVd y ROT para una semana típica del año 2003 en la estación Calle 15



Particularmente, se puede apreciar en la figura 9 que las concentraciones de 8-h CO eran más elevadas en el año 2003 en comparación con los niveles registrados en el año 2006 como lo muestra la figura 10. Sin embargo, el comportamiento es similar. Al utilizar los métodos ROT y GCVd en cada estimación se puede apreciar que para el año 2003 el comportamiento de las curvas es relativamente igual. Mientras que para el año 2006 la curva estimada por medio del método ROT presenta una suavización más profunda que la realizada utilizando el método GCVd. Esto indicaría que la curva estimada con el método ROT atenúa el comportamiento de las variaciones ocasionadas por la correlación de los errores. Además, contiene la varianza estimada de los errores empleada para la construcción de la bandas de variabilidad.

Figura 10. Bandas de Variabilidad para las curvas para las concentraciones octohorarias (8-h) estimadas mediante los selectores de ancho de banda GCV y ROT para una semana típica del año 2006 en la estación Calle 15

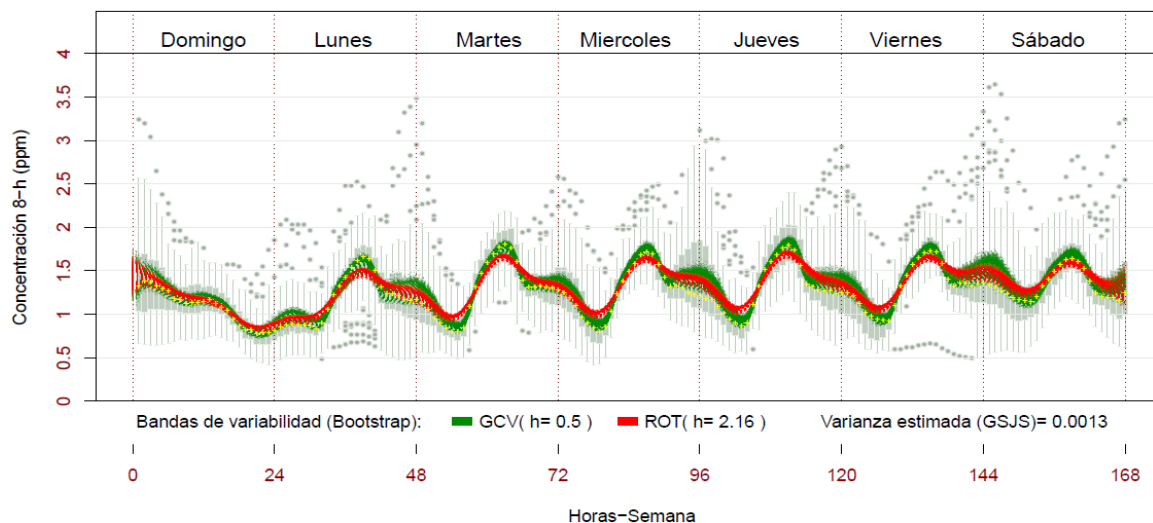


Escuela República de Argentina -ERA-

La Figura 11 presentan las bandas de variables para la estimación de la curva de la concentración octohorarias de CO para una semana típica para el año 2004 para los registros de la estación ERA ubicada en el centro de la ciudad de Cali, mediante los seleccionadores

GCVd y ROT. Se aprecia para estos dos métodos una tendencia similar aunque el método ROT presenta estimaciones más suaves que las realizadas utilizando el método GCVd.

Figura 11. Bandas de Variabilidad para las curvas para las concentraciones octohorarias (8-h) estimadas mediante los selectores de ancho de banda GCV y ROT para una semana típica del año 2004 en la estación ERA

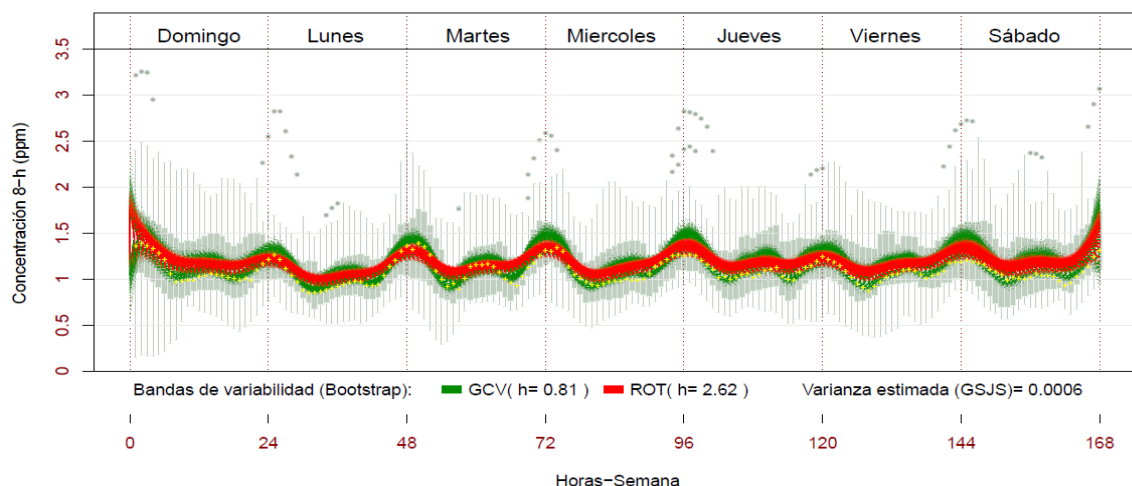


De acuerdo con la figura 11, los mayores niveles de contaminación por CO se presentan los días de martes a viernes, el cual muestra un aumento de la contaminación en las primeras horas de la madrugada, la cual decrece en las horas de la mañana y vuelve a crecer en las horas de la tarde alcanzando el máximo del contaminante. Además en la figura 11, el método ROT presenta una mejor suavización que la realizada utilizando el método GCVd. Esto indicaría que la curva estimada con el método ROT atenúa el comportamiento de las variaciones ocasionadas por la correlación de los errores. Además, contiene la varianza estimada de los errores empleada para la construcción de la bandas de variabilidad.

Polideportivo el Diamante -PDD-

Las Figura 12 se observan las bandas de variables para la estimación de la curva de la concentración octohorarias de CO para una semana típica para el año 2003 para los registros de la estación PDD ubicada en el oriente de Cali, mediante los seleccionadores GCVd y ROT. Se aprecia que la curva estimada por el método ROT presenta una suavización más fuerte que la realizada utilizando el método GCVd. Lo que indica que la curva estimada con el método ROT atenúa el comportamiento de las variaciones ocasionadas por la correlación de los errores. Además esta estación presenta la mayor contaminación de 8-h de CO en los días de martes a viernes, a diferencia de los días lunes, sábados y domingo en los cuales hay una disminución en los niveles de contaminación, esto se podría atribuir a la reducción del flujo vehicular y a que algunas empresas cesan sus actividades de producción al mínimo. También contiene la varianza estimada de los errores emplead para la construcción de la bandas de variabilidad.

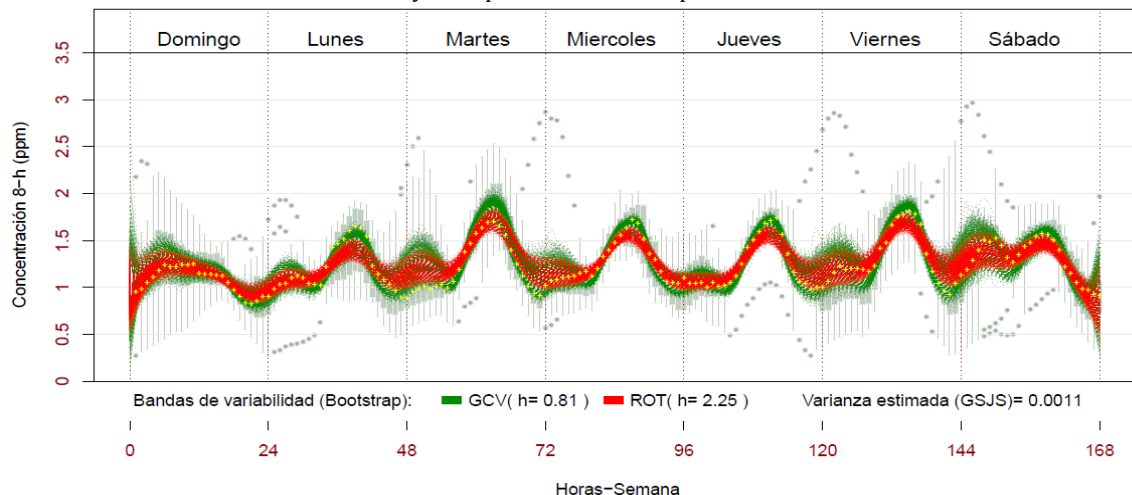
Figura 12. Bandas de Variabilidad para las curvas para las concentraciones octohorarias (8-h) estimadas mediante los selectores de ancho de banda GCVd y ROT para una semana típica del año 2003 en la estación PDD



Hospital Universitario del Valle -HUV-

Las Figura 13 se observan las bandas de variables para la estimación de la curva de la concentración octohorarias de CO para una semana típica para el año 2003 para los registros de la estación HUV ubicada en el Suroccidente de Cali, mediante los seleccionadores GCVd y ROT. Sé aprecia que la curva estimada por el método ROT presenta una suavización más fuerte que la realizada utilizando el método GCVd, la curva de la contaminación 8-h de CO para los días sábados presentó la mayor concentración de contaminación en las horas de la tarde, disminuyendo en las horas de la madrugada. Para los días domingo se presentó la menor contaminación en las horas de la mañana, esto se podría atribuir a la reducción del flujo vehicular y a que algunas empresas cesan sus actividades de producción al mínimo. La mayor contaminación la presentaron los días martes-viernes. Además, contiene la varianza estimada de los errores empleada para la construcción de la bandas de variabilidad.

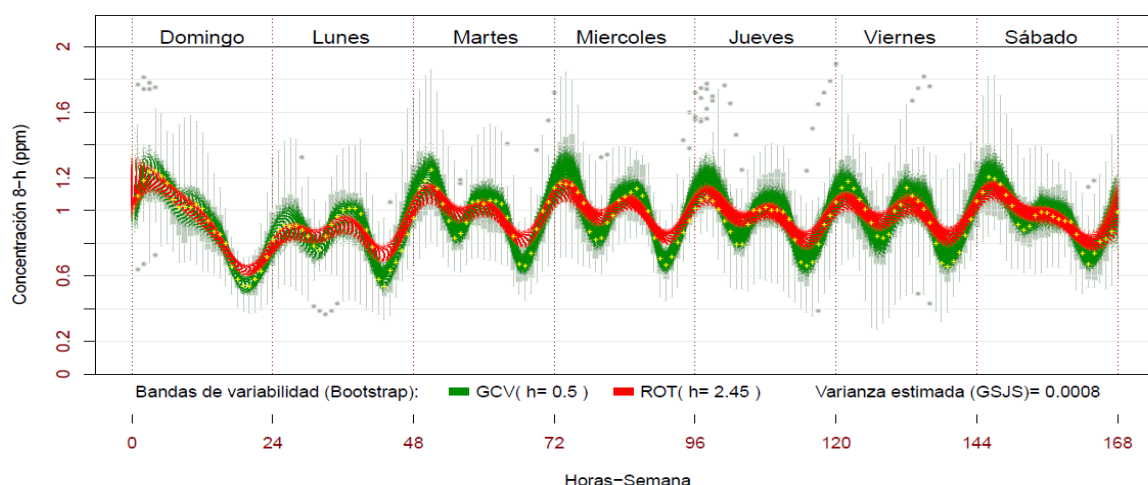
Figura 13. Bandas de Variabilidad para las curvas para las concentraciones octohorarias (8-h) estimadas mediante los selectores de ancho de banda GCVd y ROT para una semana típica del año 2003 en la estación HUV



Base Aérea -BA-

En la Figura 14 se observa las bandas de variables para la estimación de la curva de la concentración octohorarias de CO para una semana típica para el año 2003 para los registros de la estación BA ubicada en el Nororiente de la ciudad de Cali, mediante los seleccionadores GCVd y ROT. Sé aprecia para estos métodos una tendencia similar aunque el método ROT presenta estimaciones más suaves que las realizadas utilizando el método GCVd, la curva de la contaminación 8-h de CO para el día lunes se presenta los picos superiores 8-h de CO al medio día y parte de la tarde. Los días martes a viernes se presenta la mayor contaminación en las horas de la tardes. Para los días sábados y domingos se presenta los picos máximos en las primeras horas del día. Además, contiene la varianza estimada de los errores empleada para la construcción de la bandas de variabilidad.

Figura 14. Bandas de Variabilidad para las curvas para las concentraciones octohorarias (8-h) estimadas mediante los selectores de ancho de banda GCVd y ROT para una semana típica del año 2003 en la estación BA

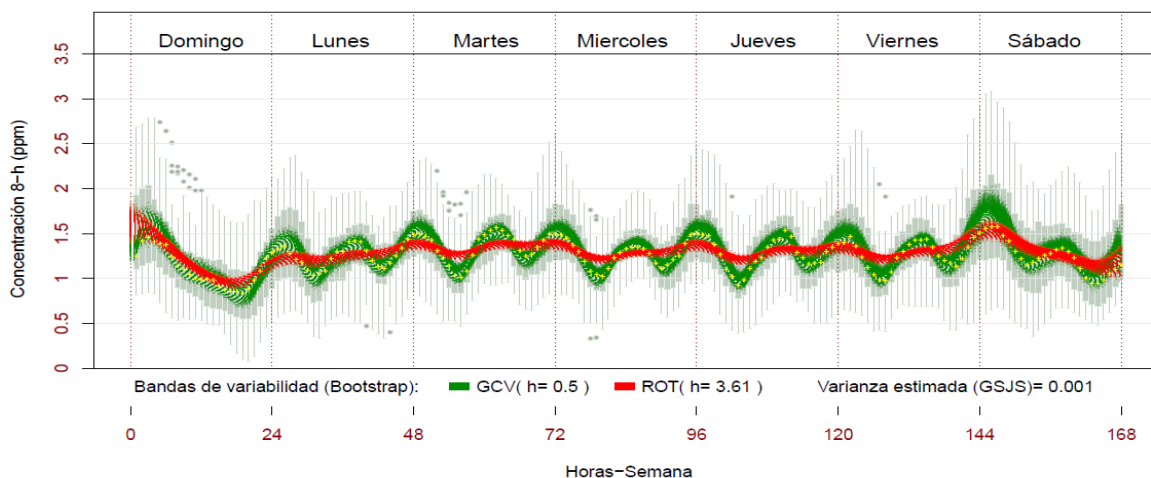


Centro de Diagnóstico Automotor del Valle –CDAV-

Las Figura 15 se observan las bandas de variables para la estimación de la curva de la concentración octohorarias de CO para una semana típica para el año 2004 para los registros de la estación CDAV ubicada en el norte de Cali, mediante los seleccionadores GCVd y ROT. Sé aprecia que la curva estimada por el método ROT presenta una suavización más fuerte que la realizada utilizando el método GCVd. Lo que indica que la curva estimada con el método ROT atenúa el comportamiento de las variaciones ocasionadas por la correlación de los errores.

La curva estimada para esta estación presenta la mayor contaminación de 8-h de CO en los días de martes a viernes al medio día, esto puede ser explicado por el alto flujo vehicular para estos días, a diferencia de los día lunes que presenta la contaminaciones superiores en las horas de la tarde y para los días sábados y domingo presenta una disminución en el contaminante CO debido a la disminución del flujo vehicular. Además, contiene la varianza estimada de los errores empleada para la construcción de la bandas de variabilidad.

Figura 15. Bandas de Variabilidad para las curvas para las concentraciones octohorarias (8-h) estimadas mediante los selectores de ancho de banda GCVd y ROT para una semana típica del año 2004 en la estación CDVA



En el **ANEXO C** se puede observar las gráficas de estimación de las curvas para las estaciones PDD, BA, CDVA y HUV de las concentraciones octohorarias para el monóxido de carbono en la ciudad de Cali. En estas graficas se evidencia que el método de Rule of Thumb (ROT) se comporta más suave que la validación cruzada generalidad- directa (GCVd)

7.4. VALIDACIÓN DE SUPUESTOS

Luego estimar las curvas de regresión, es importante validar los supuestos en el que se basan estos modelos, los cuales consisten en asumir que los errores son idénticamente distribuidos con media cero, varianza constante y deben ser independientes entre sí.

Los supuestos del modelo de regresión no paramétrica para la modelación de las emisiones octohorarias de monóxido de carbono fueron validados mediante herramientas gráficas. La verificación del supuesto de homogeneidad de varianzas sobre los errores, muestra que en general no se cumple, dado la naturaleza de donde provienen los datos.(Ver ANEXO E).

7.5. VALIDACIÓN DEL DESEMPEÑO DE LOS SELECTORES DE ANCHO DE BANDA.

Ante el comportamiento inesperado del método de selección de ancho de banda GCVd, a la hora estimar el parámetro óptimo de suavizado, este criterio no arrojó los valores esperados, por el contrario arrojó valores incluso más pequeños que los obtenidos por los métodos comunes CV y DPI. Esta situación no permitió evidenciar la corrección en las distorsiones subyacentes generadas por la estructura de autocorrelación de los residuales que expresa los autores Hart & Wehrly (1986), Altman (1990), Ruppert, et al. (1995), Bowman & Azzalini, (1997) y Fernández & Opsomer (2005).

Ante este panorama, se presenta a continuación una verificación del desempeño de los selectores de ancho de banda en la estimación de la curva de regresión no paramétrica con errores correlacionados cuando se utilizan mediciones de las concentraciones 8-h registradas

durante una la semana cada hora (una serie temporal de 168 datos). Y también, cuando se realiza la misma estimación pero con menos mediciones. Para hacer esto, se eliminará de manera aleatoria aproximadamente un 25% de los registros utilizados en la primera estimación. Los registros semanales que se analizaron pertenecen a las estaciones BA (año2003 y 2004), CDAV, HUV, CALLE15 (año 2006) y ERA (año 2006).

Las Figuras 16 y 17 presentan los correlogramas (representaciones graficas de la función de autocorrelación) de los residuales obtenidos a partir de la suavización preliminar de las concentraciones promedio octohorarias de CO de cada semana típica con registros completos e incompletos respectivamente. Algo similar se puede apreciar los gráficos (d), (e) y (f) dela Figura 17 (registros incompletos). No obstante los gráficos (a), (b) y (c) muestran un decaimiento a cero de los coeficientes de correlación más rápido en comparación con los demás gráficos. Sin embargo, se puede apreciar que algunos de estos coeficientes de autocorrelación se encuentran por fuera de los límites de significación, lo cual siguiere, la existencia de una estructura de autocorrelación en los residuales de la suavización de los niveles de las concentraciones 8h-CO. Por lo anterior, se esperaría que los valores para el parámetro de suavización arrojados por los métodos de selección convencionales (CV y DPI) presenten inconsistencias frente al método GCVd, el cuál, se presume arrojará un valor para el parámetro de suavización tal, que permita mitigar el efecto de la autocorrelación de los residuales en la suavización.

Figura 16. Funciones de autocorrelación estimadas para los residuales de la suavización preliminar, con registros completos tomados durante una semana típica.

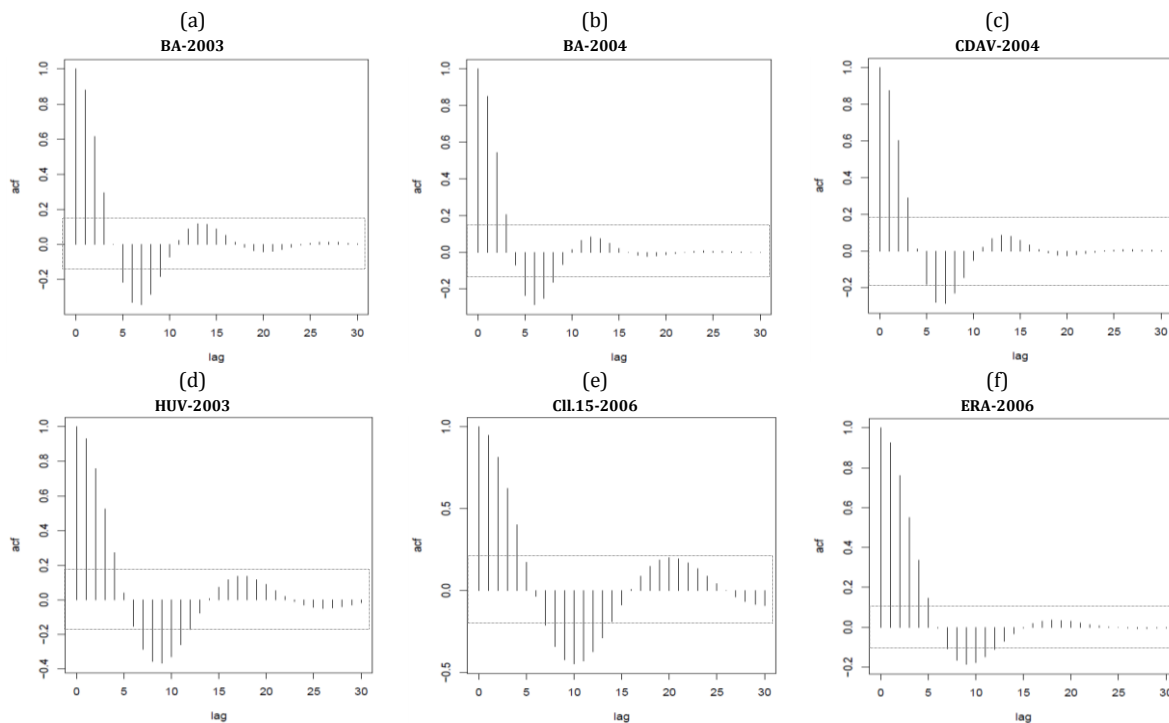
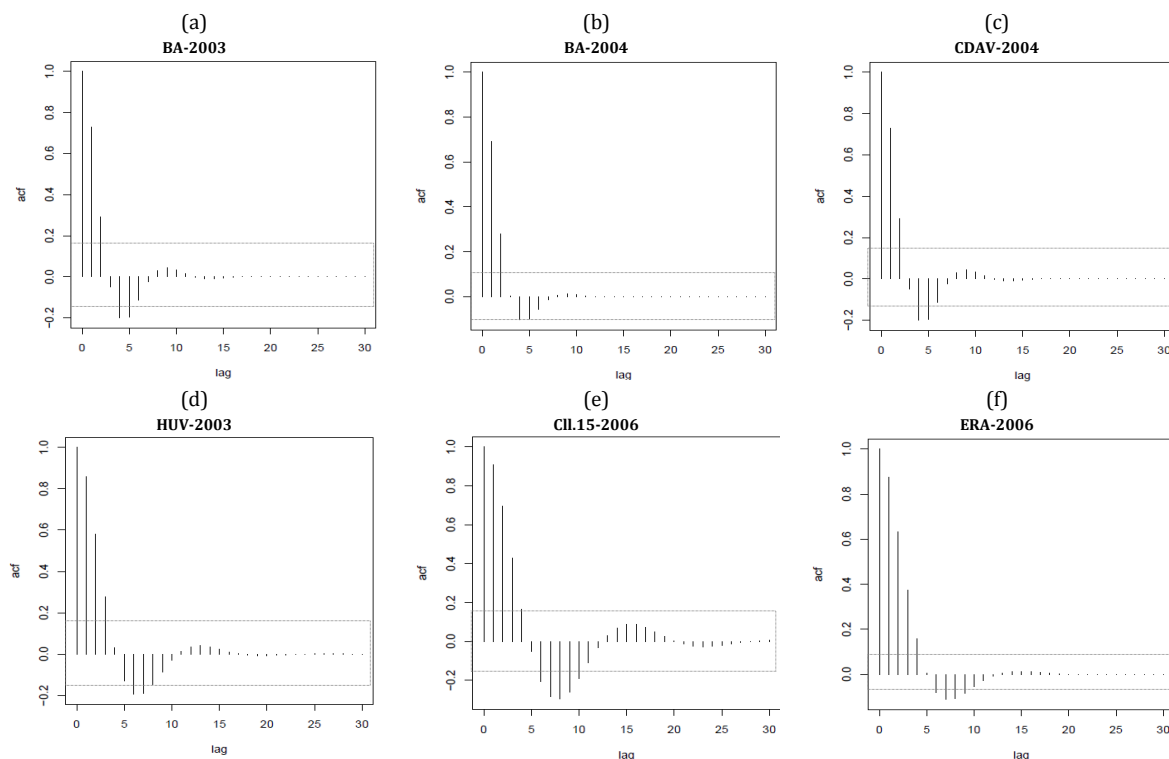


Figura 17. Funciones de autocorrelación estimadas para los residuales de la suavización preliminar, con registros incompletos tomados durante una semana típica.



La Tabla 9 contiene los valores obtenidos para el parámetro de suavizado h luego de aplicar cada método de selección de ancho de banda a los registros de cada semana escogida al azar por estación. En ella, se puede apreciar que cuando se tiene una muestra con registros completos o incompletos con observaciones dependientes, los métodos convencionales CV y DPI se ven seriamente afectados a la hora de arrojar un valor para el parámetro de suavidad dado que dichos valores son muy pequeños para h . En el contexto gráfico esto significaría que si se utiliza alguno de estos métodos en un estimador lineal local para modelar un fenómeno con observaciones medidas en el tiempo, las curvas generadas serían poco suaves debido a que éstas básicamente interpolarían los datos.

También se puede apreciar que cuando se tiene una muestra completa con observaciones dependientes, el método de selección de ancho de banda GCVd arroja valores ‘grandes’. Esto indica que al emplear dicho selector del parámetro de suavizado en un estimador lineal local, éste generará curvas mucho más suaves que si se utilizara cualquiera de los otros selectores.

En contraste, cuando se emplea este selector en una muestra que contiene registros en el tiempo incompletos, los valores arrojados son similares a los de los selectores CV y DPI. Por otro lado, el método ROT muestra un mejor comportamiento cuando se tiene una muestra incompleta con observaciones dependientes, ya que éste arroja los valores más grandes frente a los demás selectores.

Tabla 9. Anchos de banda obtenidos al aplicar diferentes métodos de selección de ancho de banda en la estimación de la curva de regresión no paramétrica para una semana típica.

Con registros completos							Con registros incompletos						
Selector	Estación año de medición						Selector	Estación año de medición					
	BA [2003]	BA [2004]	CDAV [2004]	HUV [2003]	CH.15 [2006]	ERA [2006]		BA [2003]	BA [2004]	CDAV [2004]	HUV [2003]	CH.15 [2006]	ERA [2006]
GCV _b	7,286	7,286	6,607	4,571	3,893	3,891	GCV _b	1,179	1,179	1,179	1,857	0,500	1,857
ROT	3,094	2,846	3,868	6,539	3,660	5,184	ROT	3,138	3,173	4,032	6,021	3,890	4,983
DPI	2,071	2,277	1,900	1,624	1,695	1,780	DPI	2,286	1,975	2,216	1,960	2,002	2,173
CV	0,223	0,242	0,282	0,260	0,278	0,257	CV	0,379	0,777	0,459	0,358	0,354	0,354

Las curvas típicas estimadas para los niveles de las concentraciones 8h-CO de los registros (completos e incompletos) medidos en una semana típica escogida al azar pueden ser apreciadas en las figuras del anexo B. En ellas se evidencian que los métodos convencionales CV y DPI presentan dificultades a la hora de arrojar un valor aceptable para el parámetro de suavización tanto en mediciones seriales completas como incompletas. Se puede apreciar que el criterio GVCd en condiciones óptimas (mediciones completas) funciona adecuadamente de acuerdo a lo planteado en la teoría. Sin embargo, cuando se aplica en mediciones incompletas, éste criterio presenta algunas fallas y tiende a comportarse igual que los métodos CV y DPI. No obstante, se destaca el desempeño del selector ROT en estas condiciones, puesto que arroja valores muy similares a los producidos por el criterio GCVd en condiciones normales.

8. CONCLUSIONES Y RECOMENDACIONES

- Al modelar la contaminación atmosférica mediante un modelo de regresión no paramétrica para los promedios octohorarios del monóxido de carbono de la ciudad de Cali en el periodo 2003-2006, se comprobó gráficamente por medio de correlogramas, la existencia de una estructura de autocorrelación en los residuales, la cual puede influir negativamente en el método de selección del parámetro de suavización. Por lo tanto no se recomienda utilizar selectores de ancho de banda comunes como el CV o el DPI, ya que al emplear alguno de estos selectores en el estimador lineal local, generaría una curva poco suave que prácticamente interpolaría los datos debido a que el valor del parámetro de suavidad sería muy pequeño.
- Después de comparar gráficamente los diferentes selectores de ancho de banda se pudo determinar que los métodos DPI, CV y GCVd cuando se emplean en la estimación de curvas típicas en las condiciones de mediciones repetidas de series temporales con datos faltantes, se pudo determinar que estos selectores tienen la tendencia arrojar valores pequeños para el parámetro de suavizado. No obstante, el criterio de selección ROT, aunque no es un criterio que tenga en cuenta la estructura de correlación de los residuales, presenta el comportamiento similar al esperado por el método directo de la validación cruzada generalizada de acuerdo a lo indicado Altman (1990), Opsomer J., Wang Y. & Yang Y. (2001), Bowman & Azzalini, (1997) entre otros. El cual arroja valores grandes para h , permitiendo así suavizar las fluctuaciones que se presentan debido a la estructura de correlación de los datos.
- De los diferentes selectores de anchos de bandas probados para la estimación de la curva de regresión. Gráficamente se pudo establecer que el criterio ROT es una buena alternativa para modelar el comportamiento de las concentraciones octohorarias, mediante un modelo de regresión no paramétrico con errores correlacionados, dado que brinda un resultado similar al esperado por el método Directo de la Validación Cruzada Generalizada que contempla la presencia de la autocorrelación de los residuales en condiciones comunes, es decir, en una serie de tiempo sin medidas repetidas.
- No se recomienda emplear el método de selección de ancho de banda GCVd en el escenario de mediciones repetidas de series temporales con datos faltantes, puesto que los resultados obtenidos al utilizar este criterio con los niveles de concentración de los Promedios Octohorarios de CO, no arrojó los resultados esperados (valores grandes para el parámetro de suavizado), el cual permitiera corregir el efecto de la interpolación de datos ocasionado por la autocorrelación en los residuales.
- Al ajustar los promedios octohorarios a un modelo de regresión no paramétrica con errores correlacionados, el selector de ancho de bandas que mejor comportamiento tuvo fue el método ROT, corrigiendo el efecto de la correlación en el

modelo estimando curvas suaves. Mientras que los métodos CV, DPI y GCVd presentan oscilaciones en la estimaciones en las curvas.

- Se pudo apreciar que los mayores niveles de concentración de monóxido de carbono se presentan en horas de la tarde y las mínimas concentraciones de este contaminante en horas de la mañana, por cada estación de monitoreo en Santiago de Cali para los años 2003 al 2006
- Las bandas de variabilidad permiten establecer que la contaminación por CO en un periodo de concentración de 8 horas presenta un comportamiento diferente dependiendo del tipo de día. Los días ordinarios lunes a viernes muestran un comportamiento similar, aunque el lunes por ser inicio de semana presenta un fenómeno de incremento de las concentraciones de CO pero en menor medida que de las presentadas por los días martes a viernes en los cuales se presenta los niveles máximos. El día sábado presenta el mismo fenómeno que el día lunes pero en lugar de aumentar se reduce las concentraciones 8-h de CO hasta llegar el domingo en el cual las concentraciones alcanzan su nivel más bajo.
- Con esta investigación se pudo constatar, de acuerdo a lo expresado por estudios anteriores, que no se violaron los límites máximos establecido por las normas ambientales vigentes en Colombia para el contaminante 8-h de CO en la ciudad de Santiago de Cali durante los años 2003 y 2006
- Al verificar el comportamiento de los selectores de ancho de banda utilizando los datos de una sola semana, se pudo establecer que cuando se utiliza el método de selección GCVd en una serie de datos (completa) la estimación de la curva de regresión no paramétrica del comportamiento de los promedios octohorarios de CO presenta un comportamiento suave, mitigando el efecto de interpolación de los datos producido por la correlación de los residuales en el modelo. Sin embargo, cuando se aplica a una serie de datos (incompletos) la estimación de la curva de regresión no paramétrica se ve afectada por el efecto de interpolación de los datos el cual es producido por la correlación de los residuales en el modelo.
- Al validar el comportamiento de los selectores de ancho de banda utilizando los datos de una sola semana, se pudo establecer que cuando se utiliza el método de selección ROT en una serie de datos (incompletos) la estimación de la curva de regresión no paramétrica del comportamiento de los promedios octohorarios de CO presenta un mejor comportamiento comparada con la que arrojan los demás métodos de selección, puesto que la curva generada suaviza el comportamiento de los 8-h, lo cual indica que esta estimación no se ve afectada por la correlación de los residuales en el modelo
- Se pudo establecer que utilizar los selectores de ancho de banda habituales CV y DPI para estimar la curva de regresión no paramétrica de las contracciones promedio

octohorarias de CO en cualquier escenario (medidas repetidas con datos faltantes, una serie de tiempo con datos completos o con datos incompletos) no es una buena opción, dado que las curvas generadas, se vieron notoriamente afectada por el efecto la correlación de los residuales en el modelo al interpolar los datos.

- El aporte en esta investigación fue probar el efecto de la estructura de correlación de los datos, al aplicar un modelo de regresión no paramétrica bajo condiciones de medidas repetidas y datos faltantes, utilizando los selectores de anchos de banda ROT, GCVd, DPI y CV existente en la literatura.
- Se recomienda realizar estudios de simulación, para evaluar los efectos y consecuencias que produce aplicar los diferentes selectores de ancho de banda que presenta la teoría, en el modelo de regresión no paramétrica con errores correlacionados
- se recomienda realizar más estudios de simulación en series de tiempo con datos faltantes y medidas repetidas para apreciar mejor el desempeño de los diferentes selectores que presenta la teoría de la regresión no paramétrica con errores correlacionados. Dado que a la fecha, no se conocen estudios sobre las consecuencias y efectos que produzcan aplicar métodos comunes de selección de ancho en estas condiciones.
- Se recomienda abordar por medio de simulación la manera de determinar el mejor método para seleccionar el ancho de banda, por medio de un modelo de regresión no paramétrica con errores correlacionado para la estimación de los niveles de concentración octohoraria del monóxido de carbono dado que tal objetivo no se pudo alcanzar en un 100% debido a limitaciones de tiempo del estudio.

BIBLIOGRAFÍA

Instituto Nacional de Ecología, 2007. *Tipos y fuentes de contaminantes atmosféricos*. [En línea]
Available at: <http://www2.ine.gob.mx/publicaciones/libros/396/tipos.html>
[Último acceso: 4 Ene. 2011].

Air Now, 2010. *About AIR Now, the Air Quality Index*. [En línea]
Available at: <http://www.airnow.gov/index.cfm?action=aqibasics.aqi>
[Último acceso: 27 dic 2010].

Altman, N. S., 1990. Kernel smoothing of data with correlated errors». *Journal of the American Statistical Association*, pp. 85, 749-759.

Astromia.com, s.f. *Contaminación atmosférica*. [En línea]
Available at: <http://www.astromia.com/tierraluna/contaminatm.htm>
[Último acceso: 30 diciembre 2010].

Ayala, O. D. & Ojeda, E. C., 2007. Evaluación de la Calidad del Aire en la Ciudad de Cali asociada al Monóxido de Carbono, Dióxido de Nitrógeno y Ozono a través de la Estimación y Comparación de Curvas Típicas Diarias No Paramétricas.. *Universidad del Valle, Facultad de Ingeniería*, p. 183.

Azzalini, A., 1984. Estimation and hypothesis testing of autoregressive time series. *Biometrika*, Volumen 71, pp. 85-90.

Ballester, F., 2005. *Contaminación atmosférica, cambio climático y salud*. *Rev. Esp. Salud Publica [online]*. [En línea]
Available at: http://scielo.isciii.es/scielo.php?pid=S1135-57272005000200005&script=sci_arttext
[Último acceso: 31 Enero 2011].

Bowman, A. & Azzalini, A., 1997. *Applied Smoothing Techniques for Data Analysis: the Kernel Approach With S-Plus Illustrations*. Oxford: Oxford University Press.

Cabrera, M. L. & Guerrero, J. M., 2005. *Evacuación de la efectividad de la medida del pico y placa en Bogotá D.C.*, Bogotá D.C.: s.n.

Caracol Radio, 2009. *Caracol Radio*. [En línea]
Available at: <http://www.caracol.com.co/nota.aspx?Id=846956>
[Último acceso: 31 1 2011].

Castro, D. A., 2009. *Propuesta de Estimación No Paramétrica de la Contaminación por Monóxido de Carbono en el Centro de la Ciudad de Cali, Colombia, Entre 2003 y 2006*, Cali - Colombia: s.n.

Coakley, C. W. y otros, 2000. Bandwidth Selection Concerns for Jump Point Discontinuity Preservation in the Regression Setting Using M-smoothers and the Extension to Hypothesis Testing. *CiteSeerX Scientific Literature Digital Library and Search Engine*.

Craven, P. & Wahba, G., 1979. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized crossvalidation. *Numerische Mathematik*, Volumen 31, pp. 377-403.

DAGMA, 2009. *Borrador de lineamientos para una Política Ambiental de Santiago de Cali*, Santiago de Cali: Alcaldía de Santiago de Cali.

De Brabanter, K., De Brabanter, J., Suykens, J. & De Moor, B., 2011. Kernel Regression in the Presence of Correlated Errors. *Journal of Machine Learning Research*, Volumen vol 12, pp. 1955-1976.

Delicado, P., 2008. Curso de Modelos no Paramétricos. 14 septiembre.

El Tiempo.com, 1998. *NORTE DE CALI, EL MÁS CONTAMINADO*. [En línea]
Available at: <http://www.eltiempo.com/archivo/documento/MAM-839164>
[Último acceso: 31 dic. 2010].

EPA, 2000. Air Quality Criteria for Carbon Monoxide. *U.S. Environmental Protection Agency*.

EPA, 2010. *Oficina de Aire y Radiación de la Agencia de Protección Ambiental*. [En línea]
Available at: http://www.epa.gov/iaq/co_sp.html
[Último acceso: 31 dic 2010].

Eubank, R., 1999. *Spline smoothing and nonparametric regression*. 2da ed. Dekke, New York: s.n.

Fan and, J. & Yao, Q., 2003. *Nonlinear Time Series: Nonparametric and Parametric Methods*. s.l.:Springer.

Fan, J. & Gijbels, I., 1996. Local polynomial modelling and its applications. *Chapman and Hall, London*.

Fernández, F. & Opsomer, J., 2005. Smoothing Parameter Selection Methods for Nonparametric Regression with Spatially Correlated Errors. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, Junio, Vol. 33(No. 2), pp. pp. 279-295.

Gasser, T., Sroka, L. & Jennen-Steinmetz, C., 1986. Residual Variance and Residual Pattern in Nonlinear Regression. *biometrika*, Volumen Vol 73, pp. 625-633.

González, M., 2009. Correlación en los errores del Modelo de Regresión Lineal. *Universidad Torcuato di Tella*.

Graybill, F. A., 1976. *Theory and applications of the linear model*, New York: Wadsworth & Brooks.

- Hall, P. & Marron, J., 1990. On Variance Estimation in Nonparametric Regression. *Biometrika*, Volumen 77, pp. 415-419.
- Hart, D. & Wehrly, T. E., 1986. Kernel regression estimation using repeated measurements data. *Journal of the American Statistical Association* 81: 1080-8, Volumen 81, pp. 1080-8.
- Hart, J. D., 1991. *Kernel regression estimation with time series errors*. S.I.: J. Roy. Statist. Soc. B.
- IDEAM, 2007. *Informe anual sobre el estado del medio ambiente y los recursos naturales renovables en Colombia: calidad del aire 2007*, Bogotá D.C.: s.n.
- Inche, J. L., 2004. *Gestión de la calidad del aire: causas, efectos y soluciones*. Lima: Instituto de Investigación de Ingeniería Industrial-UNMSM.
- Jaramillo, M., González, D. E. & Nuñez, M. E., 2009. Índice integrado de calidad del aire para ciudades colombianas. *Rev. fac. ing. univ*, Issue no.48, pp. p.97-106.
- La Cisterna S.A. - Sociedad Concesionaria Intermodal, 2005. *Declaración de impacto ambiental estación de intercambio modal quinta normal*. [En línea]
Available at: http://seia.sea.gob.cl/archivos/Anexo_G.pdf
[Último acceso: 28 Febrero 2011].
- Liu, F., 2001. *Environmental justice analysis : theories, methods, and practice*, Lewis Publishers.
- Loader, C., 1999. *Local Regression and Likelihood*. New York: Springer.
- Madrid, s.f.. *Calidad del Aire*. [En línea]
Available at:
<http://www.mambiente.munimadrid.es/opencms/opencms/cal aire/red/indice/descripcion.html>
[Último acceso: 27 dic 2010].
- Mallows, C. L., 1973. Some comments on Cp. *Technometrics* 15, pp. 661- 675..
- Montoya, R. & Morales, A., 2004. *Estimación no Paramétrica de curvas típicas diarias para los contaminantes CO, NO2 y SO2 en Santiago de Cali*, Cali - Colombia: s.n.
- Niglio, M. & Perna, C., 2003. Kernel smoothing for the analysis of climatic data. *Quaderni di Statistica*, p. Vol. 5.
- Opsomer, J., Wang, Y. & Yang, Y., 2001. Nonparametric Regression with Correlated Errors. *Statistical Science*, May, Vol. 16(No. 2), pp. pp. 134-153.
- Ramsay, J. O. & Silverman, B. W., 2005. *Functional data analysis*, Springer. second edn ed. New York: Springer.

Reina, J., 2009. *Aplicación de técnicas de suavización no paramétrica en la modelación de la contaminación promedio octohoraria del monóxido de carbono en cali durante los años 2003-2006*, Cali - Colombia: s.n.

Ruppert, D., Sheather, S. J. & Wand, M. P., 1995. An Effective Bandwidth Selector for Local Least Squares Regression. *Journal of the American Statistical Association*.

Secretaría de Medio Ambiente y Recursos Naturales de México - SEMARNAT, 2010.

semarnat.gob.mx. [En línea]

Available at:

<http://www.semarnat.gob.mx/educacionambiental/Paginas/temasprioritarios.aspx>

[Último acceso: 15 enero 2011].

Servy, E., Cuesta, C., Marí, G. & Armida, M. L., 2006. Utilización del paquete "KernSmooth" de r para construir suavizados loess y bandas de variabilidad a datos de la encuesta de ocupación hotelera. *Undécimas Jornadas "Investigaciones en la Facultad" de Ciencias Económicas y Estadística*, Noviembre.

Servy, E., Gracia, M. d. C. & Paccapelo, V., 2007. Regresión no Paramétrica: Una aplicación. *Instituto de Investigaciones Teóricas y Aplicadas, de la Escuela de Estadística*.

Simonoff, J., 1996. *Smoothing Methods in Statistics*. New York: Springer-Verlag.

Téllez, A. R. & A. F., 2006. CONTAMINACION POR MONÓXIDO DE CARBONO: UN PROBLEMA DE SALUD AMBIENTAL. *Revista de Salud Pública*, Marzo, Vol.8(001), pp. 108-117.

Téllez, J., Rodríguez, A. & Faja, Á., 2006. *estrucplan*. [En línea]

Available at: <http://www.estrucplan.com.ar/articulos/verarticulo.asp?IDArticulo=1334>

[Último acceso: 22 2011].

The Condensed Chemical Dictionary, 1977. *The Condensed Chemical Dictionary*. 9th ed. ed. NY: s.n.

Urbieto Lopez, J., 2006. *monografias.com*. [En línea]

Available at: <http://www.monografias.com/trabajos12/contatm/contatm.shtml>

[Último acceso: 29 dic 2010].

Wand, M. P. & Jones, M. C., 1995. *Kernel Smoothing*. Chapman and Hall, London.

ANEXOS

Anexo A: Diagramas de cajas y alambres para las estaciones PDD, BA, CDVA y HUV de las concentraciones octohorarias para el monóxido de carbono.

Tabla 10. Estadísticas descriptivas de las concentraciones promedio octohorarias de monóxido de carbono (8h-CO) para la estación calle 15 durante los años 2003 al 2006.

Estación	Año	Min.	Q1.	Mediana	Media	Q3.	Max.
PDD	2003	0.146	0.880	1.078	1.168	1.415	3.248
PDD	2004	0.384	0.797	0.890	0.893	0.995	3.130
BA	2003	0.274	0.753	0.928	0.947	1.116	1.891
BA	2004	0.024	0.488	0.645	0.662	0.815	1.870
HUV	2003	0.253	0.966	1.213	1.258	1.526	2.964
CDVA	2004	0.078	0.985	1.240	1.283	1.551	3.083

Figura 18. Diagrama de cajas y alambre de concentraciones (8-h) durante los años 2003 y 2004, para la estación PDD

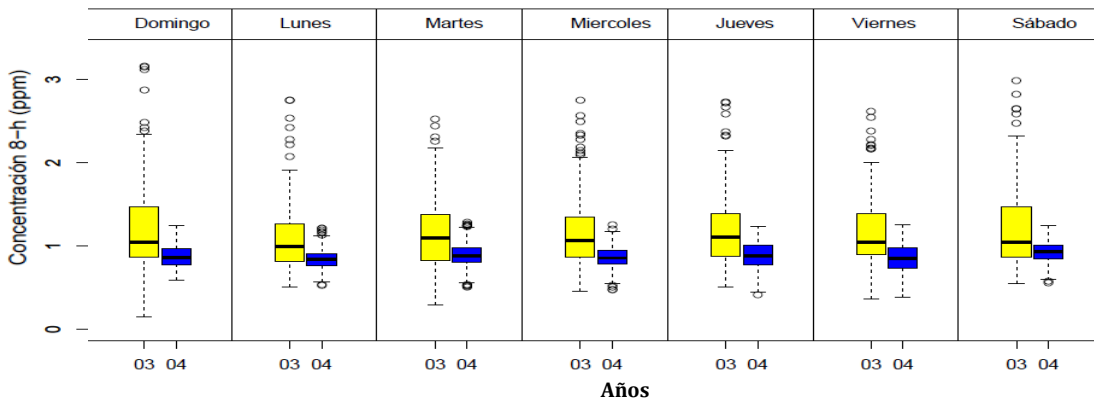


Figura 19. Diagrama de cajas y alambre de concentraciones (8-h) durante los años 2003 y 2004, para la estación BA

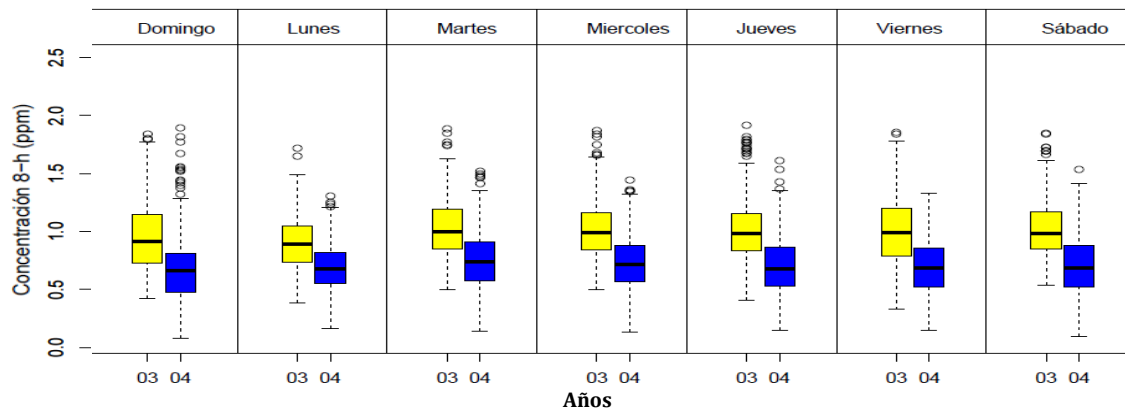


Figura 20. Diagrama de cajas y alambre de concentraciones (8-h) durante el año 2003, para la estación calle HUV

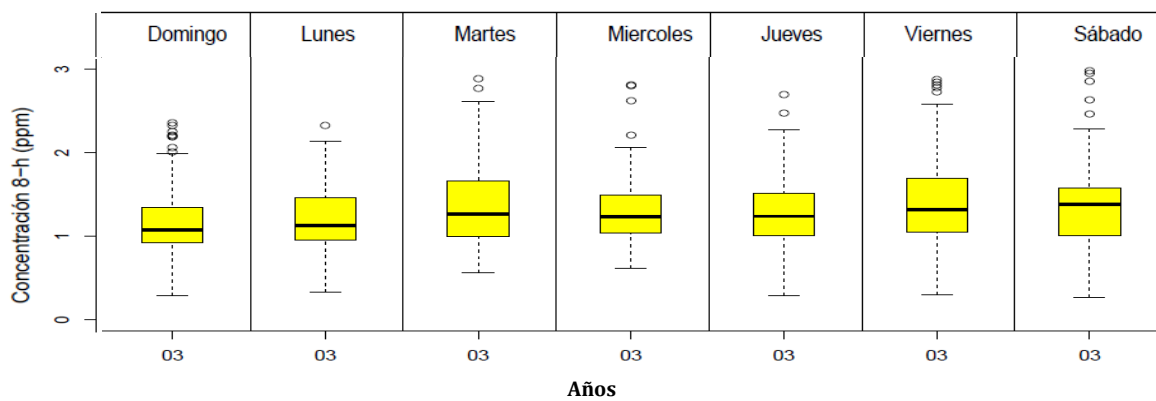
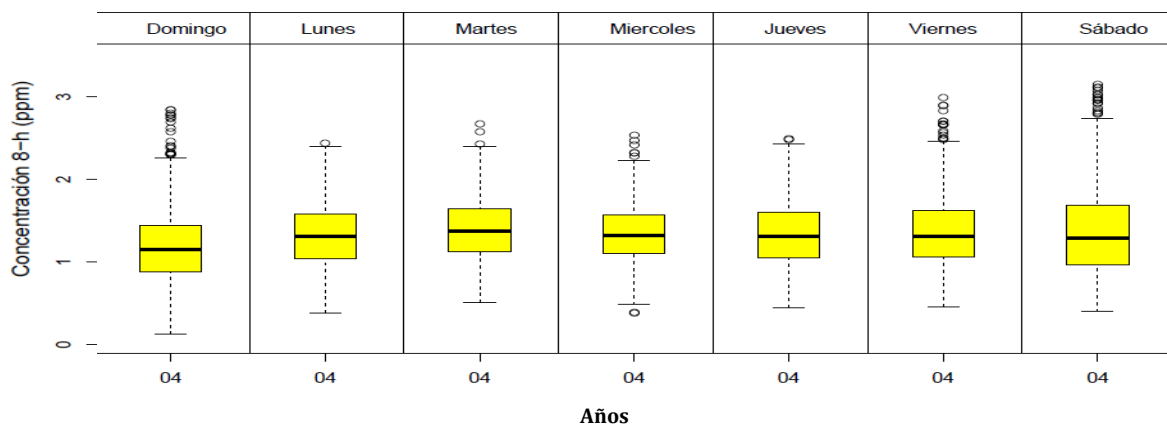


Figura 21. Diagrama de cajas y alambre de concentraciones (8-h) durante el año 2004, para la estación CDAV



Anexo B: Funciones de autocorrelación estimadas de las suavizaciones preliminares de los residuales del modelo de regresión no paramétrico de las concentraciones octohorarias (8-h) para una semana típica en las estaciones PDD, BA, CDVA y HUV.

Figura 22. Funciones de autocorrelación estimadas de las suavizaciones preliminares de los residuales del modelo de regresión no paramétrico de las concentraciones octohorarias (8-h) para una semana típica en la estación PDD en los años 2003 y 2004

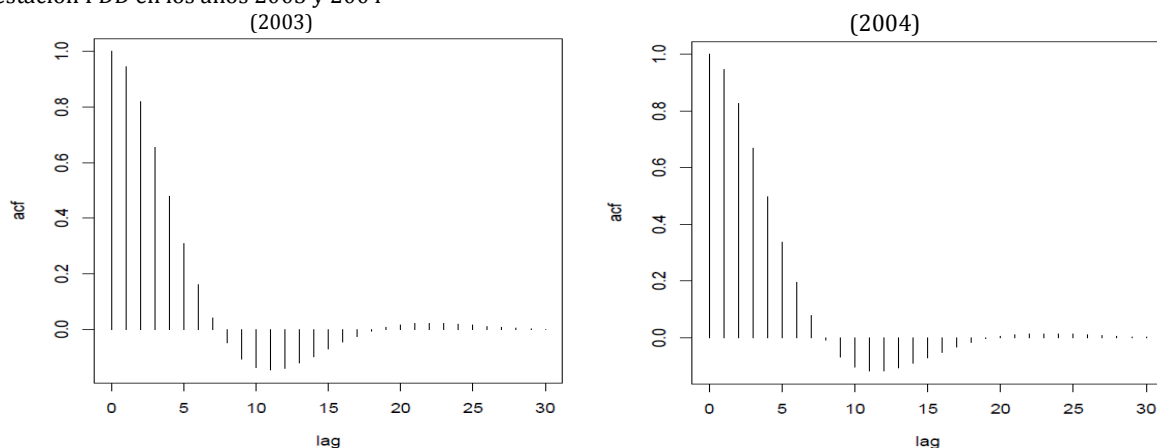


Figura 23. Funciones de autocorrelación estimadas de las suavizaciones preliminares de los residuales del modelo de regresión no paramétrico de las concentraciones octohorarias (8-h) para una semana típica en la estación BA en los años 2003 y 2004

(2003)

(2004)

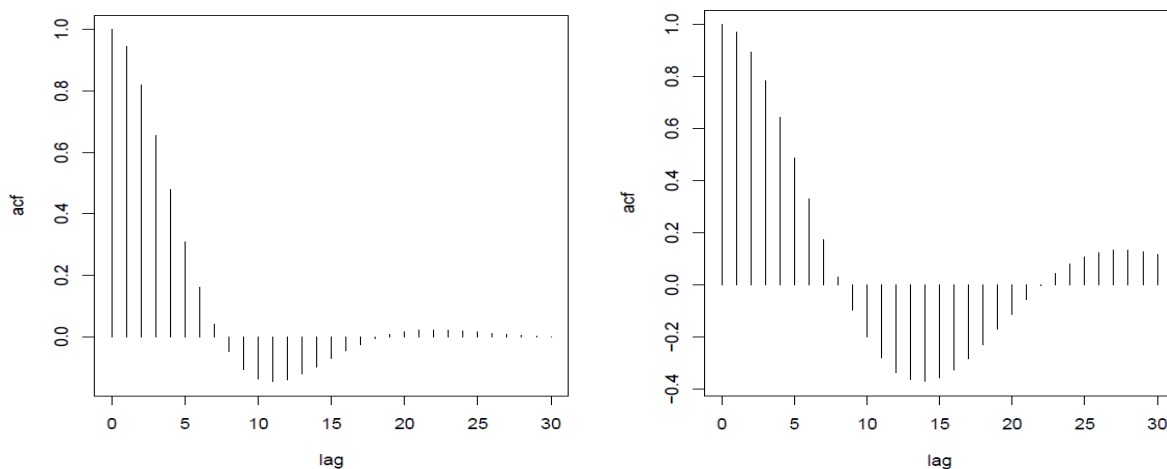
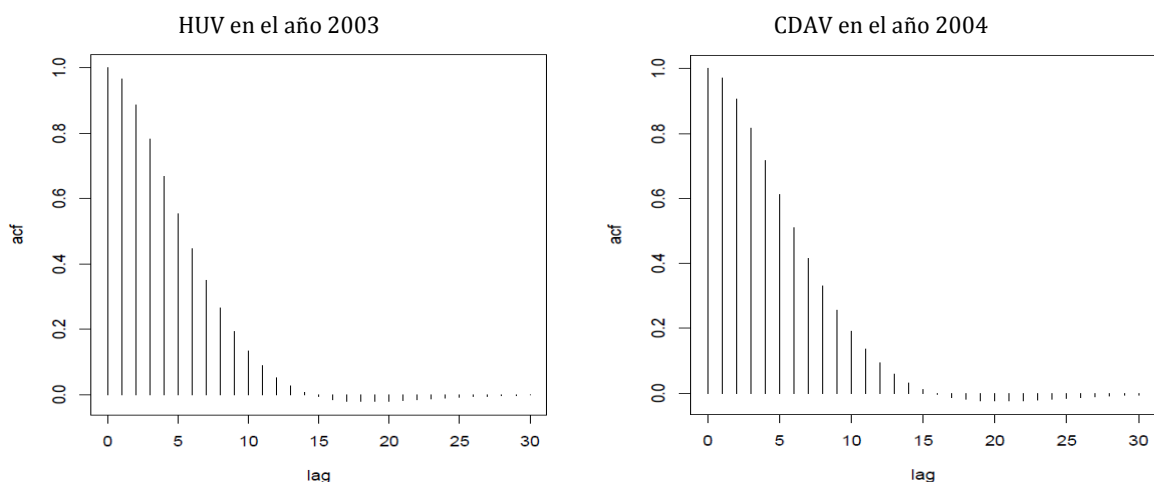


Figura 24. Funciones de autocorrelación estimadas de las suavizaciones preliminares de los residuales del modelo de regresión no paramétrico de las concentraciones octohorarias (8-h) para una semana típica en las estaciones



Anexo C: Estimación de las curvas típicas para las concentraciones octohorarias (8-h) de CO utilizando los selectores de ancho de banda (CV, DPI, GCV y ROT) para una semana típica por año en las estaciones PDD, BA, CDVA y HUV

Figura 25. Estimación de las curvas típicas para las concentraciones octohorarias (8-h) de CO utilizando los selectores de ancho de banda (CV, DPI, GCV y ROT) para una semana típica por año en la estación PDD

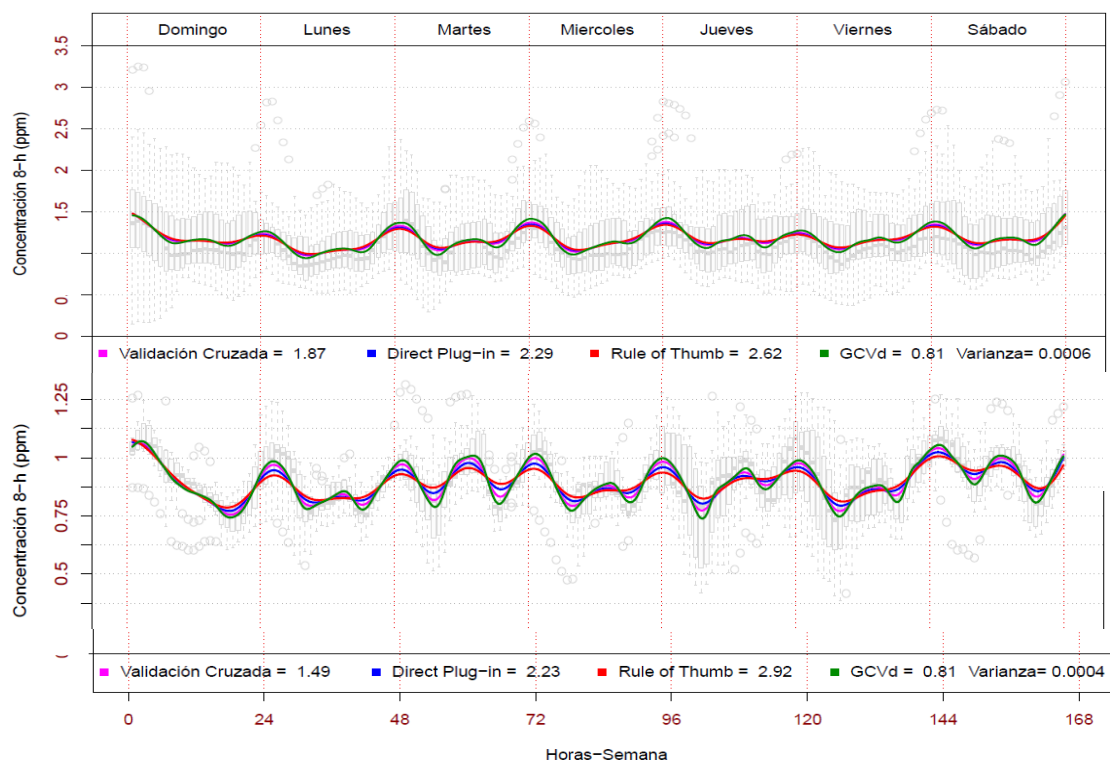


Figura 26. Estimación de las curvas de concentraciones octohorarias (8-h) de CO utilizando los selectores de ancho de banda (CV, DPI, GCV y ROT) para una semana típica por año en la estación BA

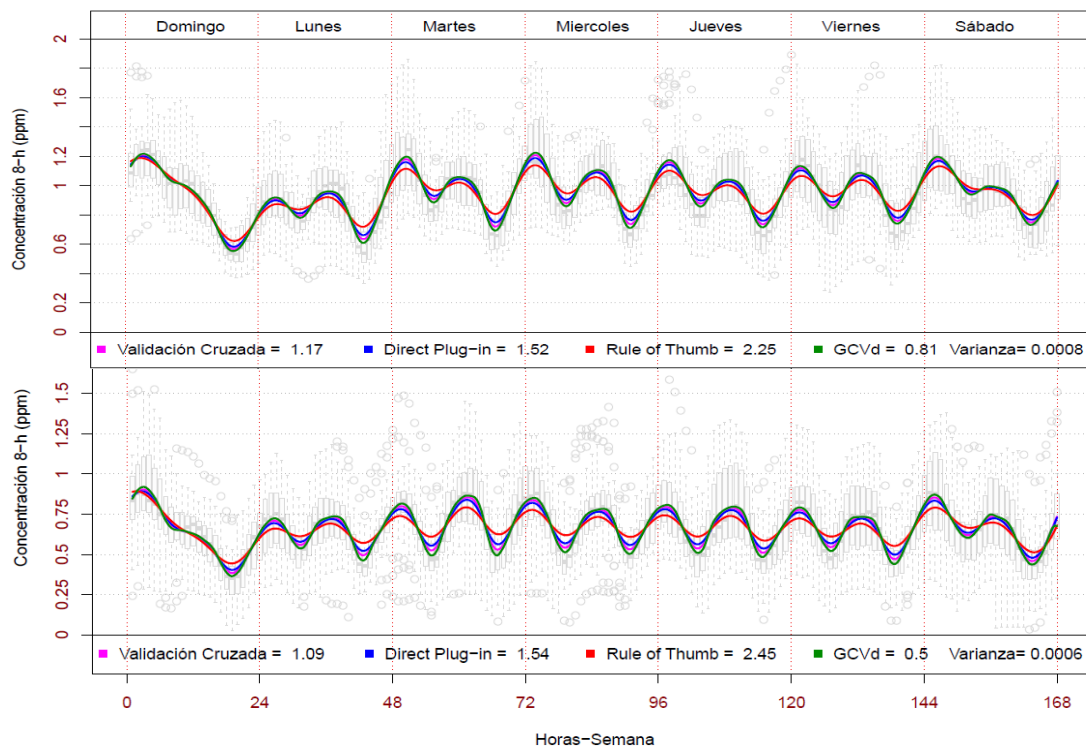
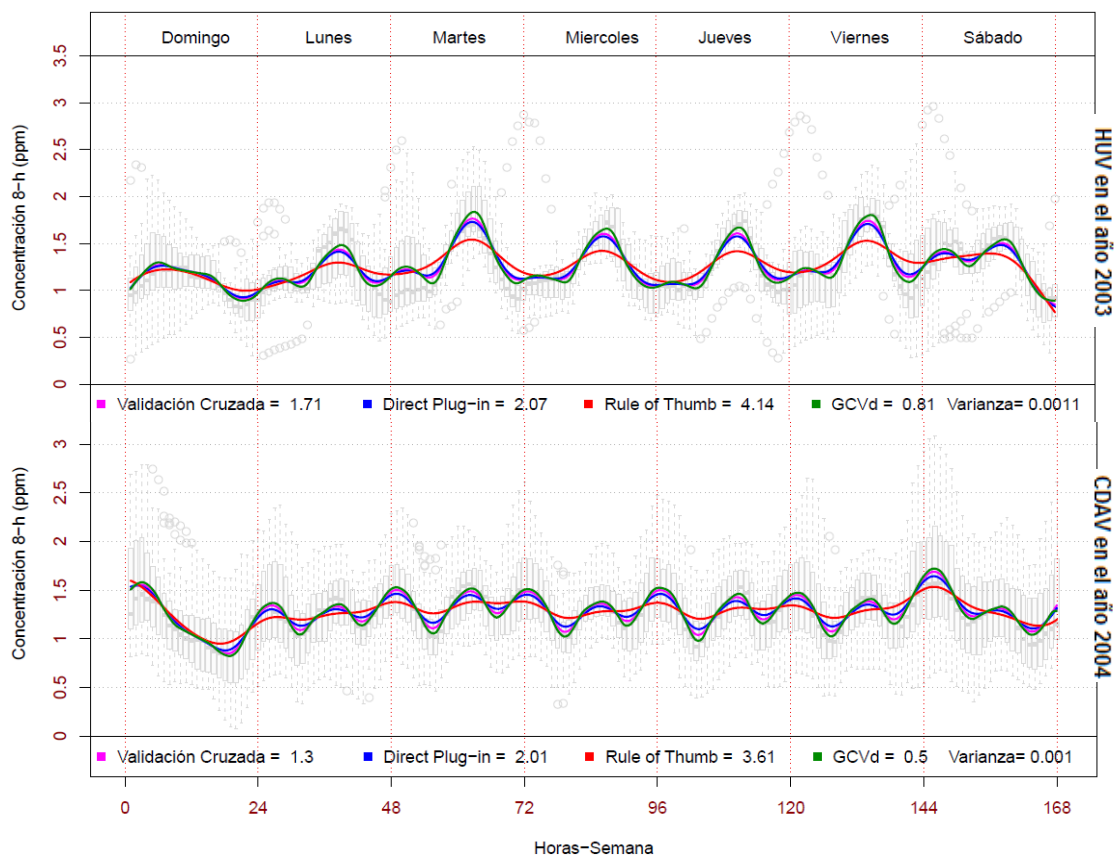


Figura 27. Estimación de las curvas de concentraciones octohorarias (8-h) de CO utilizando los selectores de ancho de banda (CV, DPI, GCV y ROT) para una semana típica por año en las estaciones



Anexo D: Bandas de Variabilidad para las curvas para las concentraciones octohorarias (8-h) estimadas mediante los selectores de ancho de banda GCVd y ROT para una semana típica en las estaciones PDD, BA, CDVA y HUV.

Figura 28. Bandas de Variabilidad para las curvas para las concentraciones octohorarias (8-h) estimadas mediante los selectores de ancho de banda GCVd y ROT para una semana típica del año 2004 en la estación Calle 15

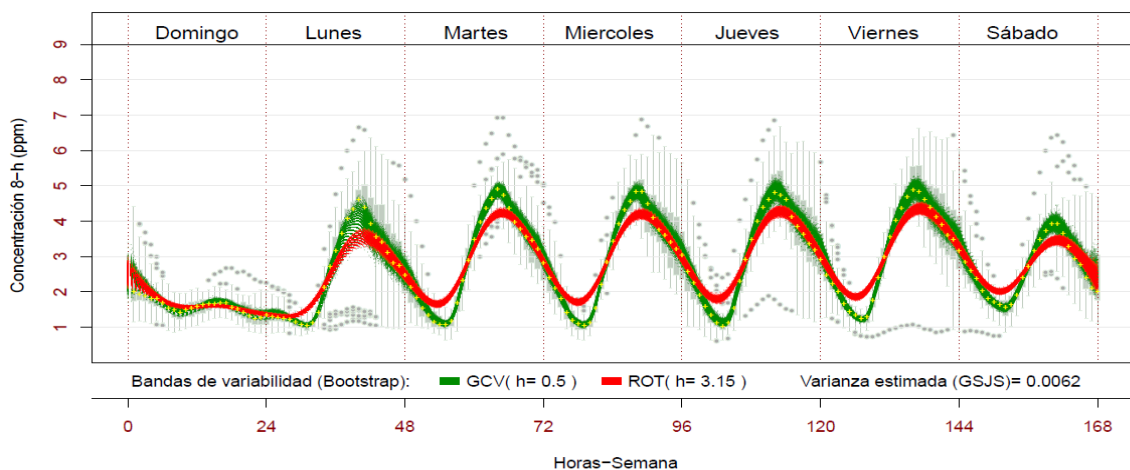


Figura 29. Bandas de Variabilidad para las curvas para las concentraciones octohorarias (8-h) estimadas mediante los selectores de ancho de banda GCVd y ROT para una semana típica del año 2005 en la estación Calle 15

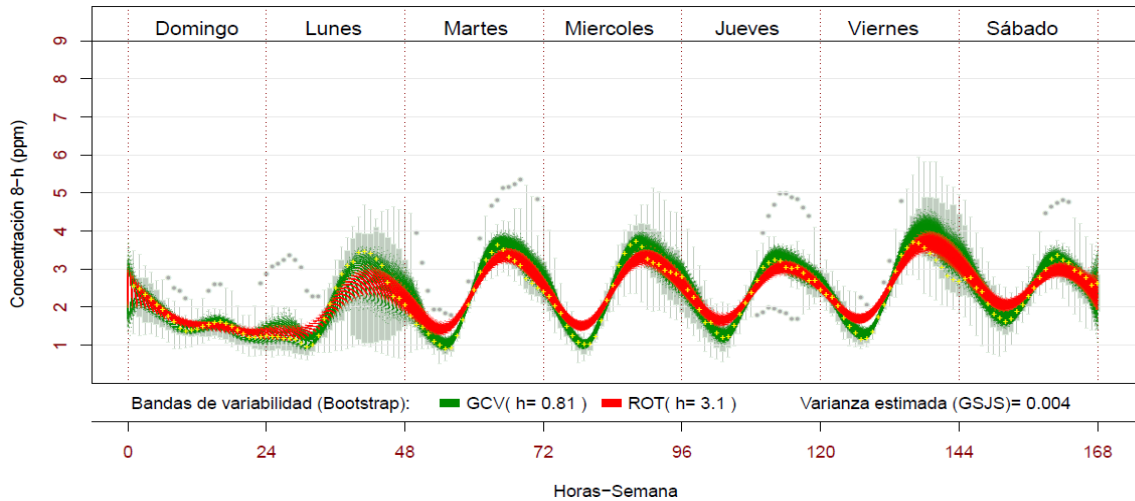


Figura 30. Bandas de Variabilidad para las curvas para las concentraciones octohorarias (8-h) estimadas mediante los selectores de ancho de banda GCVd y ROT para una semana típica del año 2003 en la estación ERA

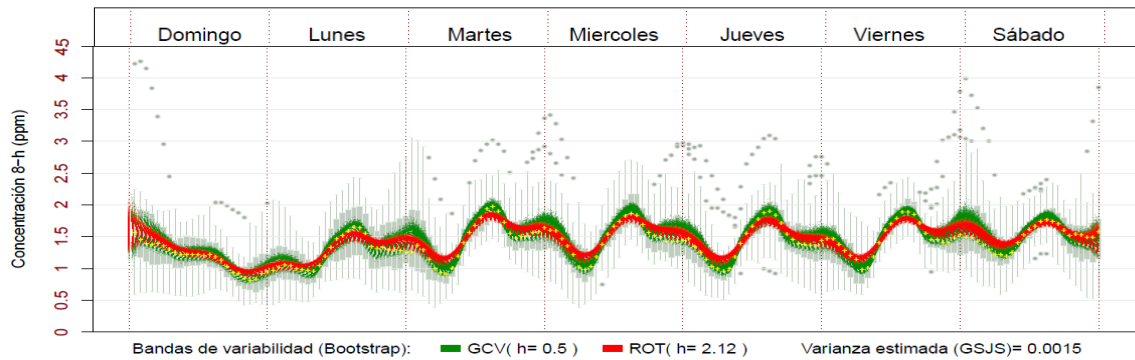


Figura 31. Bandas de Variabilidad para las curvas para las concentraciones octohorarias (8-h) estimadas mediante los selectores de ancho de banda GCV y ROT para una semana típica del año 2006 en la estación ERA

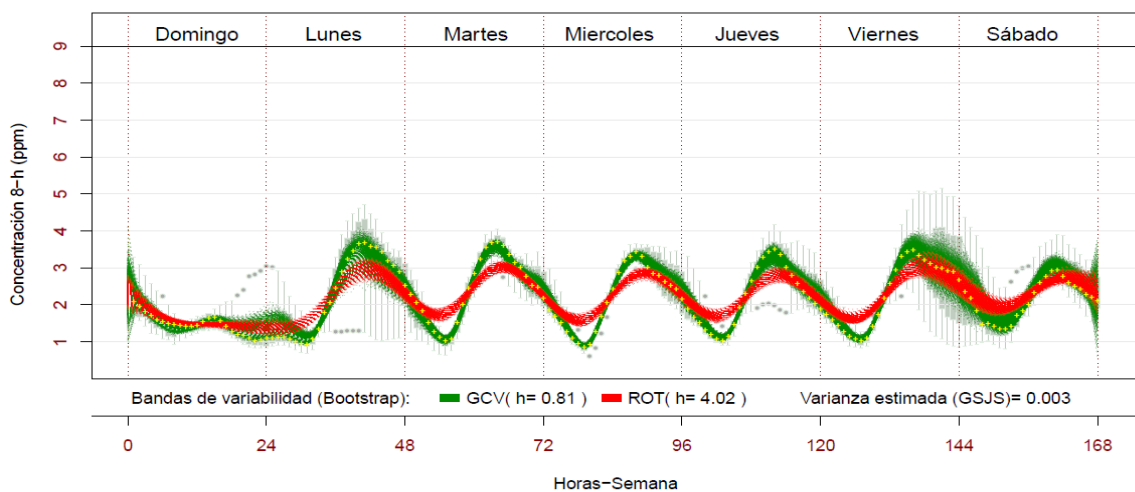


Figura 32. Bandas de Variabilidad para las curvas para las concentraciones octohorarias (8-h) estimadas mediante los selectores de ancho de banda GCVd y ROT para una semana típica del año 2004 en la estación PDD

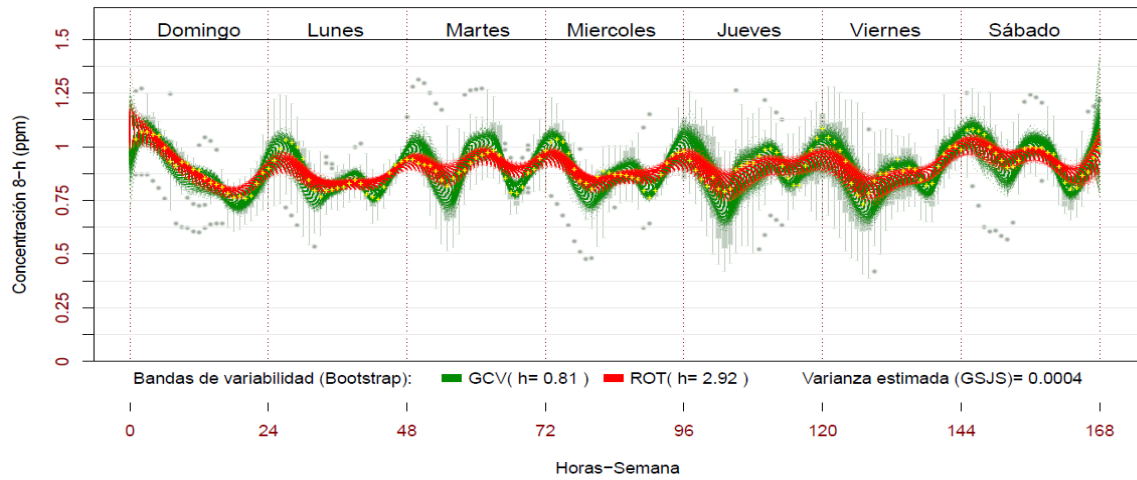
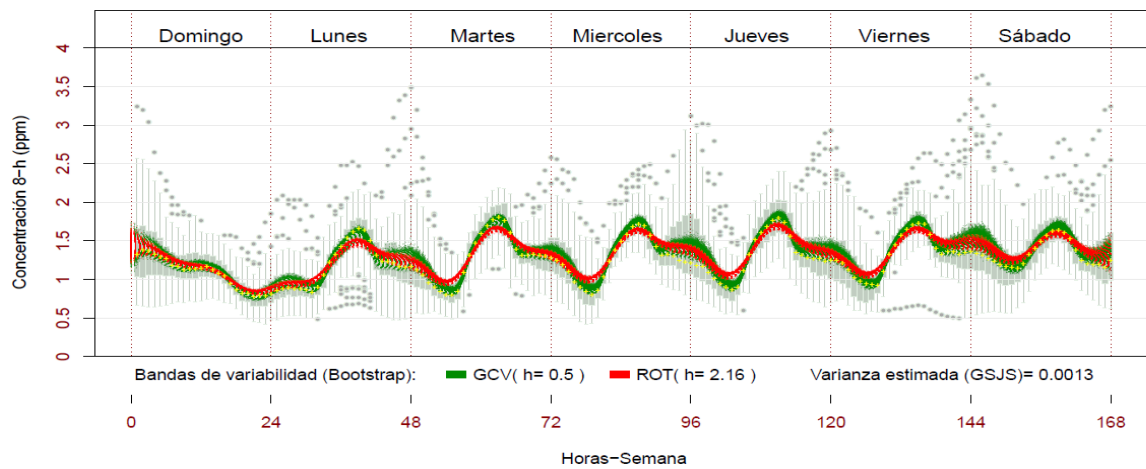


Figura 33. Bandas de Variabilidad para las curvas para las concentraciones octohorarias (8-h) estimadas mediante los selectores de ancho de banda GCVd y ROT para una semana típica del año 2004 en la estación BA



Anexo E: Gráficos de Residuales para las concentraciones octohorarias para el monóxido de carbono en la ciudad de Cali para los años 2003 al 2006.

Figura 34. Gráfico del comportamiento de los residuales de la estimación con GCVd de las curvas de las concentraciones de CO 8-h clasificados por Hora-semana, estación BA, año 2003.

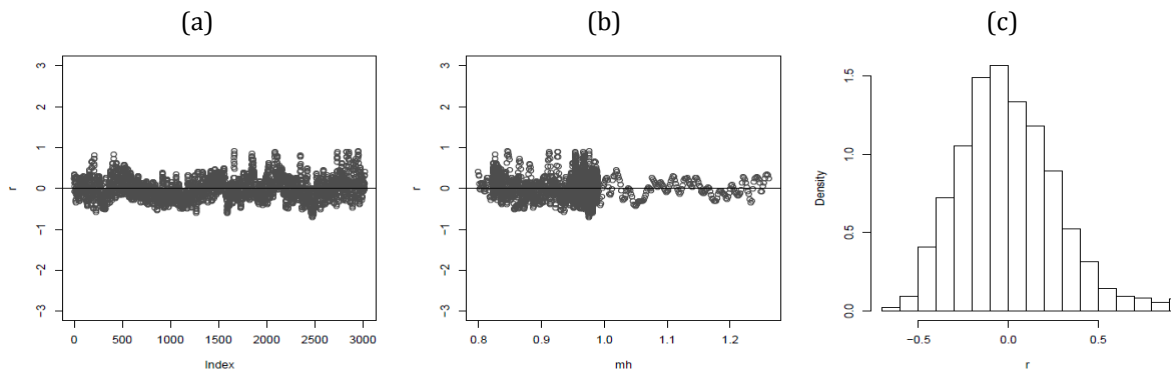


Figura 35. Gráfico del comportamiento de los residuales de la estimación con GCVd de las curvas de las concentraciones de CO 8-h clasificados por Hora-semana, estación BA, año 2004.

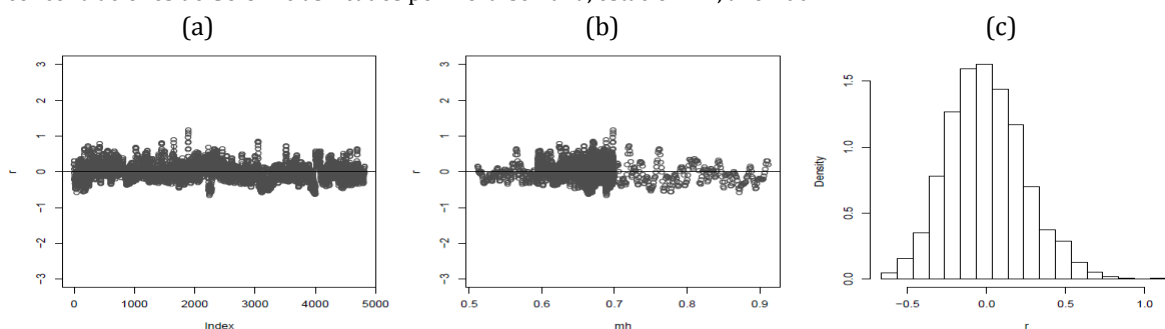


Figura 36. Gráfico del comportamiento de los residuales de la estimación con GCVd de las curvas de las concentraciones de CO 8-h clasificados por Hora-semana, estación CDAV, año 2004.

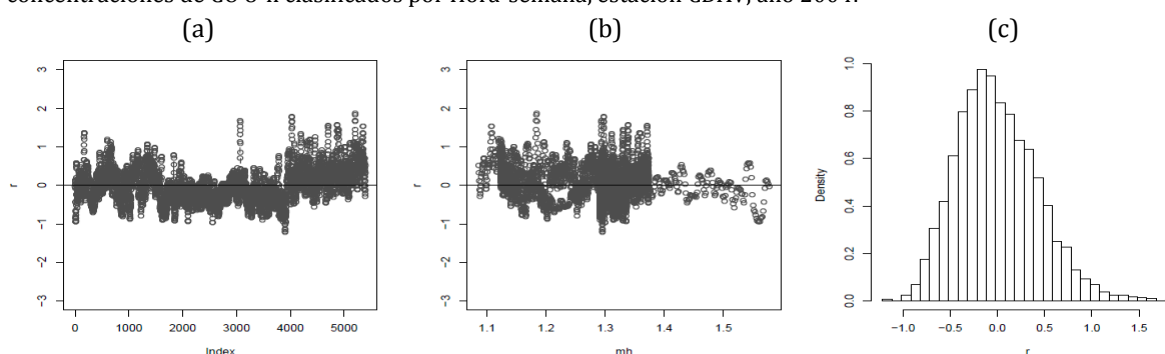


Figura 37. Gráfico del comportamiento de los residuales de la estimación con GCVd de las curvas de las concentraciones de CO 8-h clasificados por Hora-semana, estación Calle 15, año 2003.

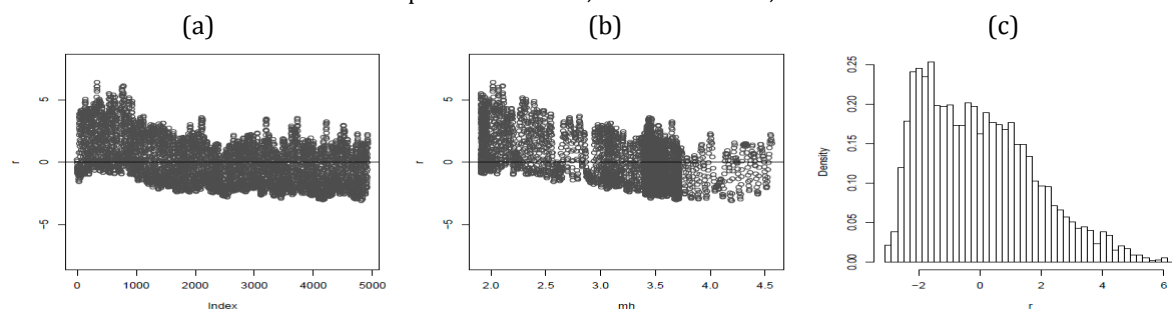


Figura 38. Gráfico del comportamiento de los residuales de la estimación con GCVd de las curvas de las concentraciones de CO 8-h clasificados por Hora-semana, estación Calle 15, año 2004.

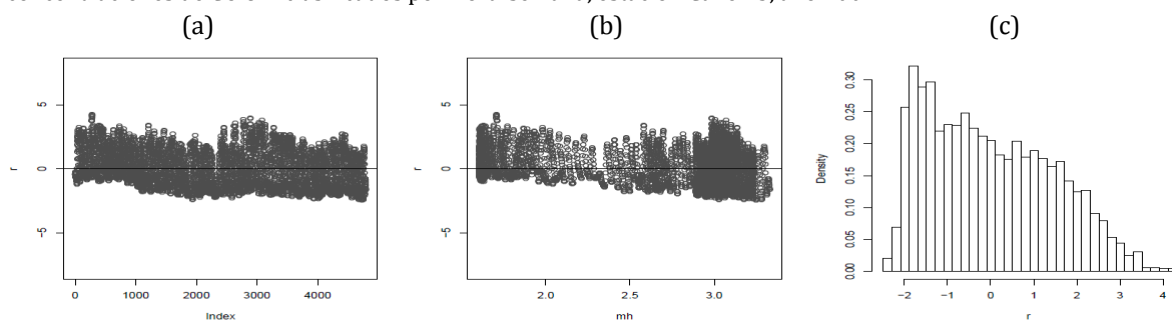


Figura 39. Gráfico del comportamiento de los residuales de la estimación con GCVd de las curvas de las concentraciones de CO 8-h clasificados por Hora-semana, estación Calle 15, año 2005.

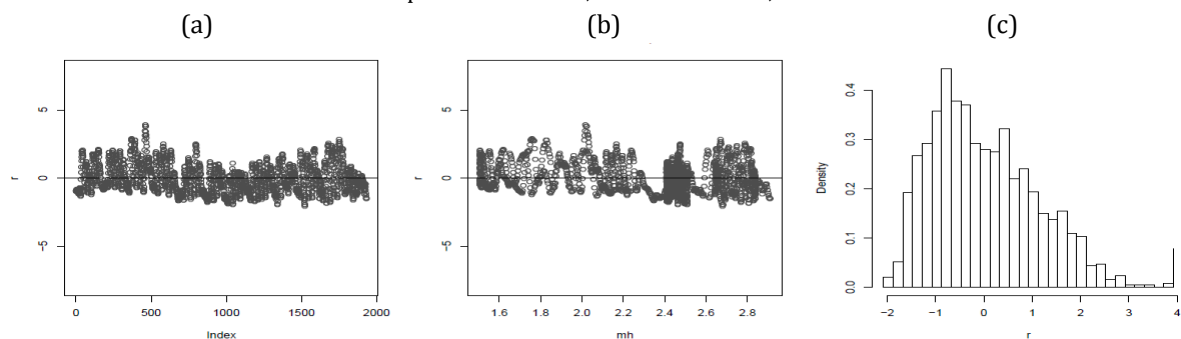


Figura 40. Gráfico del comportamiento de los residuales de la estimación con GCVd de las curvas de las concentraciones de CO 8-h clasificados por Hora-semana, estación Calle 15, año 2006.

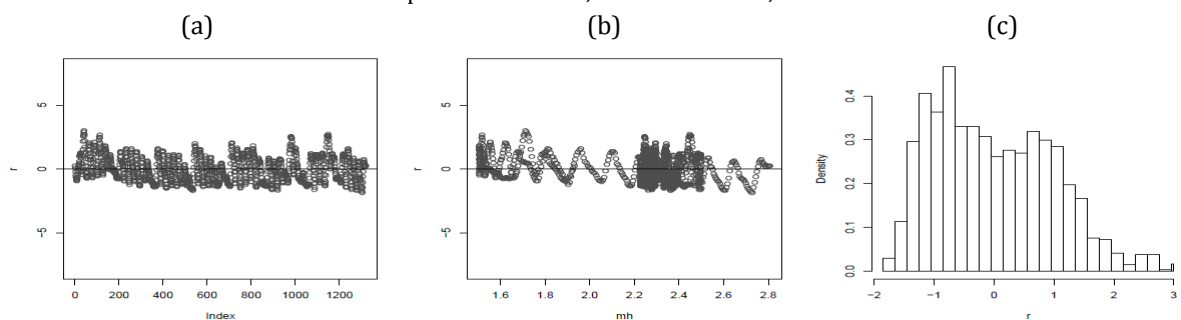


Figura 41. Gráfico del comportamiento de los residuales de la estimación con GCVd de las curvas de las concentraciones de CO 8-h clasificados por Hora-semana, estación ERA, año 2003.

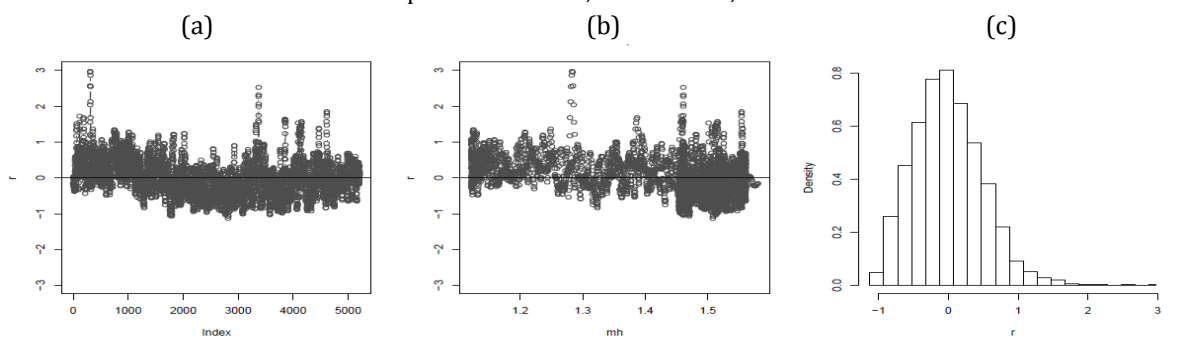


Figura 42. Gráfico del comportamiento de los residuales de la estimación con GCVd de las curvas de las concentraciones de CO 8-h clasificados por Hora-semana, estación ERA, año 2004.

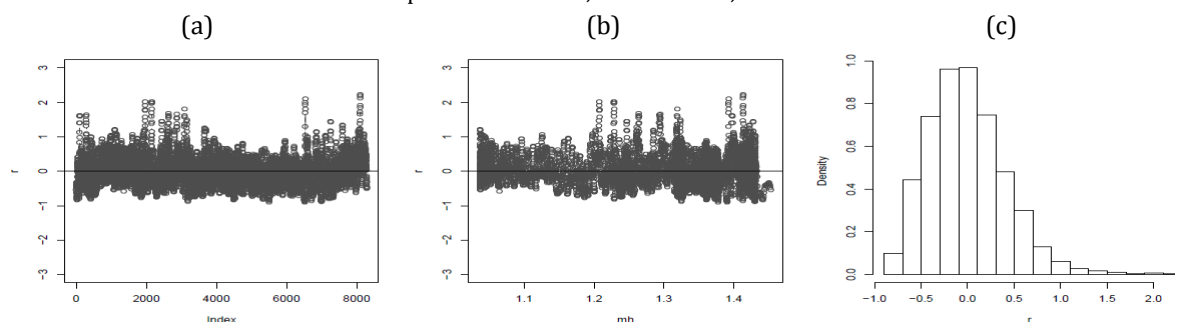


Figura 43. Gráfico del comportamiento de los residuales de la estimación con GCVd de las curvas de las concentraciones de CO 8-h clasificados por Hora-semana, estación ERA, año 2006.

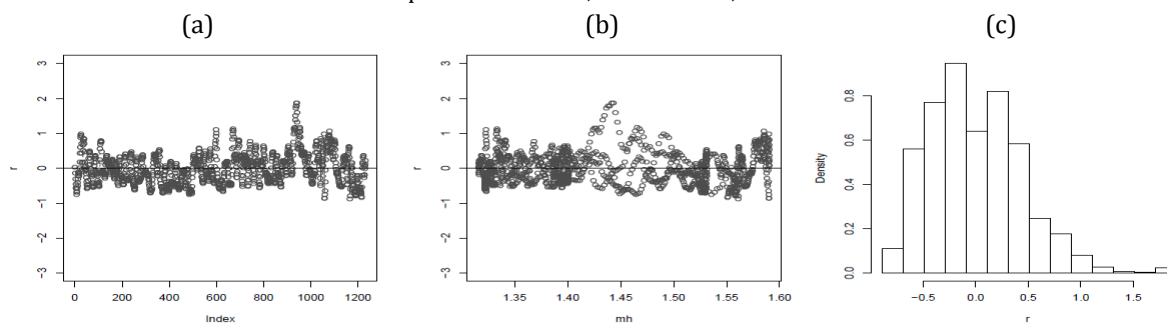


Figura 44. Gráfico del comportamiento de los residuales de la estimación con GCVd de las curvas de las concentraciones de CO 8-h clasificados por Hora-semana, estación HUV, año 2003.

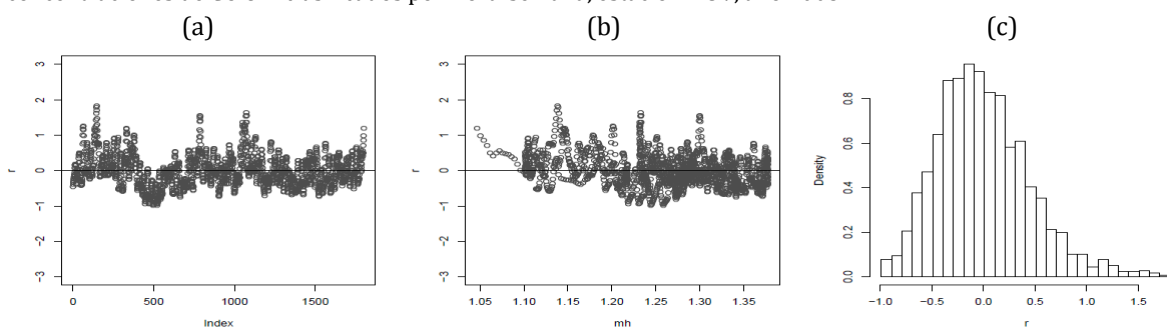


Figura 45. Gráfico del comportamiento de los residuales de la estimación con GCVd de las curvas de las concentraciones de CO 8-h clasificados por Hora-semana, estación PDD, año 2003.

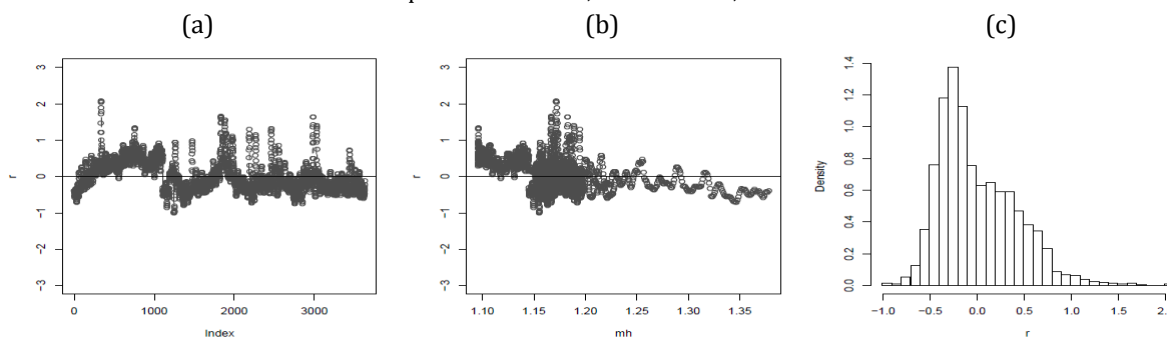
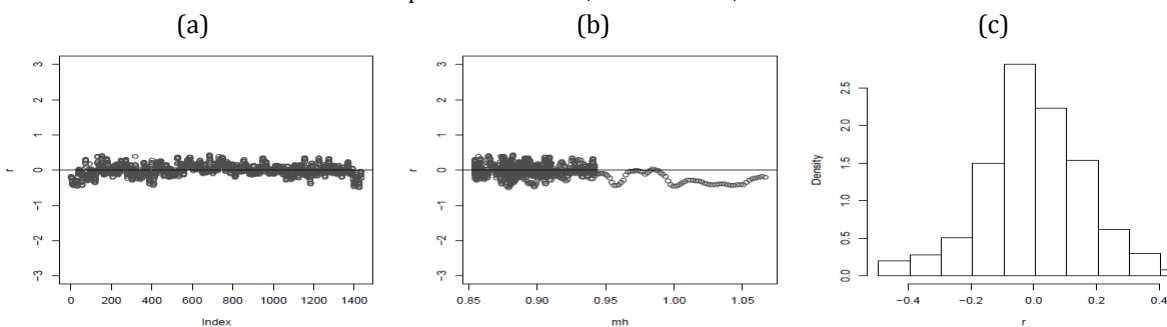


Figura 46. Gráfico del comportamiento de los residuales de la estimación con GCVd de las curvas de las concentraciones de CO 8-h clasificados por Hora-semana, estación PDD, año 2004.



Anexo F: Estimación de las curvas para diferentes selectores de ancho de banda de las estaciones PDD, BA, CDVA Y HUV de las concentraciones octohorarias para el monóxido de carbono en la ciudad de Cali.

Figura 47. Curvas de regresión no paramétrica estimadas para los niveles de concentración (8-h) de CO por criterio de selección de ancho de banda (CV, DPI, GCV y ROT) para mediciones completas en una semana.

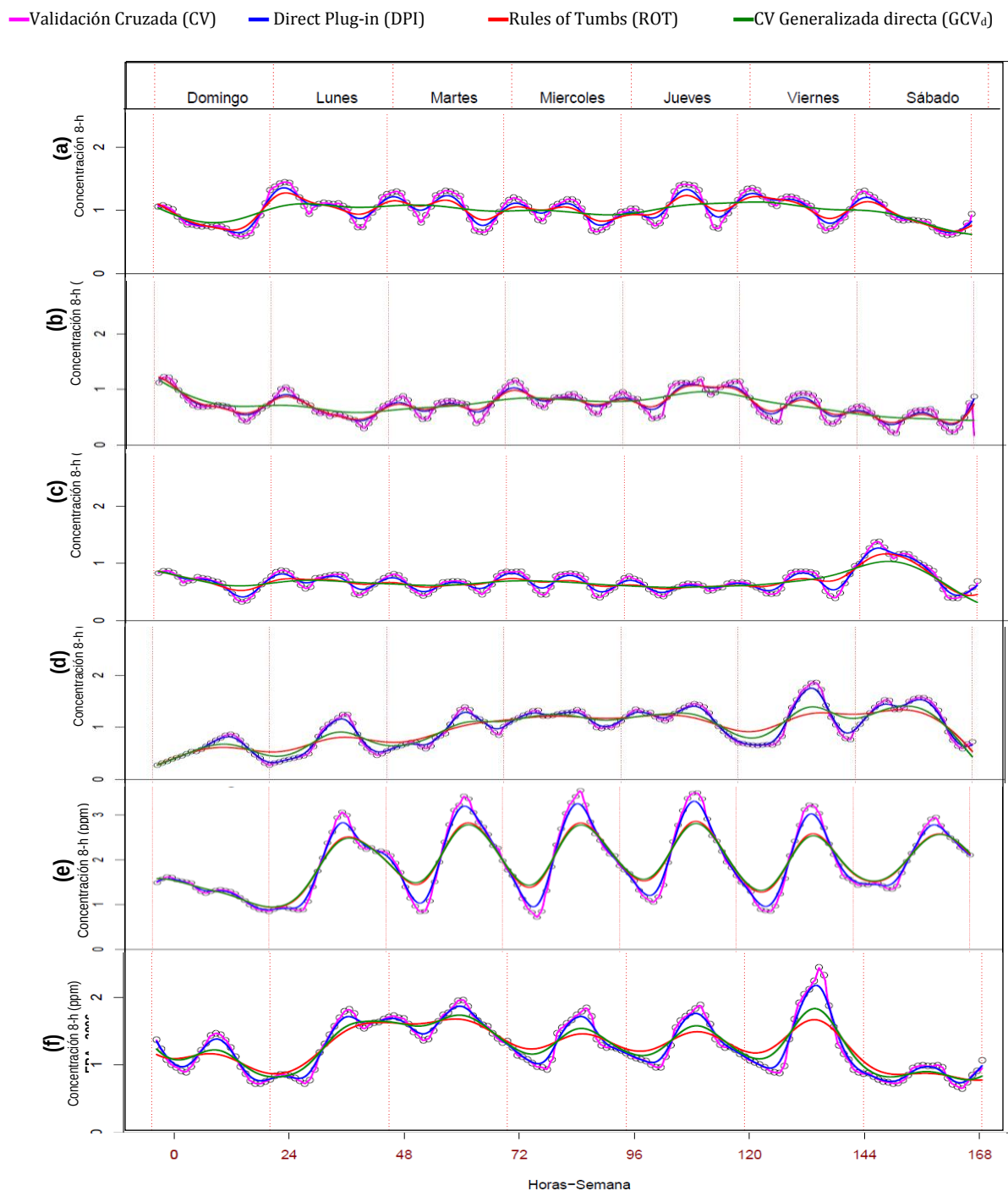


Figura 48. Curvas de regresión no paramétrica estimadas para los niveles de concentración (8-h) de CO por criterio de selección de ancho de banda (CV, DPI, GCV y ROT) para mediciones incompletas en una semana.

