Pronostico y Prevención de la contaminación por PM_{10} en la red de monitoreo de calidad de aire de Bogotá (RMCAB)

Nicolás Mejía Martínez[1]

Asesor: Jorge Bonilla

Resumen

La contaminación del aire en Bogotá, especialmente producto del PM_{10} se ha vuelto una creciente preocupación para las autoridades locales debido a los efectos adversos que tiene en materia de salud. De esta manera, un modelo de pronóstico es beneficioso para la ciudad pues permite, a la población, tomar medidas preventivas y minimizar los efectos en salud de la contaminación. Adicionalmente, un análisis en cuanto a la implementación de un sistema de alertas puede probar ser útil para las autoridades y la población como una herramienta adicional para mitigar los efectos del material particulado. En este artículo se propone una metodología estructurada para el desarrollo de modelos de pronostico basados en series de tiempo, para cada una de las horas pico de los días de la semana (de 6:00 am a 6:00 pm), de las 8 estaciones de la red de monitoreo que contaban con la mayor cantidad de datos válidos. Una vez se desarrollan los modelos se realiza un análisis de la viabilidad en la implementación de un sistema de alerta temprana basado en primer lugar del sistema ya diseñado en la ciudad de Medellín, seguido de la cuantificación de los costos económicos por atención hospitalaria, que ha presentado para la ciudad, el hecho de no poseer un sistema de alertas; equivalentes a aproximadamente $57.885.106.339 COP.

Palabras Clave: PM_{10}, Contaminación atmosférica, Series de tiempo, RMCAB, Bogotá

Clasificación JEL: C22, C52, C53

[1] Nicolás Mejía Martínez. Correo: n.mejia10@uniandes.edu.co. Código: 201314597

1. Introducción

La calidad del aire se ha convertido en una de las principales problemáticas a nivel mundial, en el tema de salud pública. Según informes de organización mundial de la salud, la exposición a material particulado (PM) desencadena en diversas afecciones como enfermedades respiratorias, obstrucción pulmonar crónica, cáncer de pulmón, derrames cerebrales y enfermedades cardiacas (World Health Organization, 2016). Anualmente, cerca de 6.5 millones de muertes están relacionadas con la contaminación del aire, lo cual equivale a aproximadamente un 10% del total de muertes globales (Cozzi & et.al, 2016).

Las alarmantes cifras y el aumento en la concientización acerca de los efectos negativos que tiene la contaminación atmosférica sobre la salud de las personas, ha llevado al adelanto y desarrollo de políticas públicas enfocadas a disminuir los peligros que vienen con la exposición al material particulado, como los son las restricciones a contaminantes durante periodos definidos de tiempo, como lo es el caso de la resolución 610 de 2010 en Bogotá, que determina los límites aceptables de emisiones de PM_{10}. Adicional a esto, los gobiernos han buscado desarrollar distintos mecanismos de acceso a la información con el fin de facilitar su revisión por parte de la población, llegando a tomar medidas reactivas o preventivas de acuerdo a la situación de contaminación del momento.

Los mecanismos reactivos resultan un tanto ineficientes debido a que la exposición al contaminante, por parte de la población, ocurre en los momentos de alerta. Por otro lado, los mecanismos preventivos ayudan a la población a protegerse de antemano contaminación permitiéndoles tomar las medidas y precauciones necesarias de acuerdo al nivel de contaminación.

Bogotá, es la sexta ciudad con más densidad poblacional de Latinoamérica, teniendo alrededor de 10 millones de habitantes. En la ciudad las preocupaciones van aumentando debido a la relación entre las distintas enfermedades y la contaminación atmosférica. El observatorio ambiental menciona que según los datos del sistema nacional de vigilancia en salud pública, los reportes de admisiones a salas de urgencia por enfermedades respiratorias han ido en aumento desde el 2009 (Secretaria Distrital del Ambiente, 2016). Los registros de emisiones a lo largo de la red muestran que la mayoría de veces se sobrepasa el límite

máximo sugerido por la organización mundial de la salud, que son 50 $\mu g/m^3$ y en ciertas localidades como lo son Kennedy y Carvajal, el limite local, que es el doble del recomendado por la WHO, 100 $\mu g/m^3$. Las autoridades locales están en la transición a la implementación de una plataforma web en la que se incorpora el estado actual de la contaminación, junto con un pronóstico del día siguiente, todo esto discriminado por localidades. Adicional a esto, la ciudad cuenta con un sistema de monitoreo en tiempo real, no solo de distintos contaminantes, sino también de variables climáticas y atmosféricas, así como un índice de colores que muestra el estado del ambiente con respecto a la acumulación del contaminante.

Con esto en mente, y utilizando la información disponible se utilizarán modelos de series de tiempo con el fin de generar un modelo robusto de pronóstico para cada una de las horas de mayor afluencia poblacional en las calles de la ciudad (denominadas horas pico), en cada una de las estaciones de la red. En este sentido, el artículo propone no solo un modelo que pronostique el promedio diario, sino modelos diferenciados para cada franja horaria que involucren la información relevante disponible para cada una de las estaciones y las franjas horarias, algo que hasta donde se tiene conocimiento no ha sido elaborado previamente.

Los modelos diferenciados resultan de gran ayuda para la creación de políticas y planes de prevención discriminando no solo por la estación donde se da la exposición al contaminante, sino por la hora en la que se produce, la cual teniendo en cuenta la estacionalidad de las series, cambian constantemente a lo largo de un mismo día. Es importante resaltar que, para el desarrollo empírico de la metodología se hará uso de la base de datos obtenida de la RMCAB la cual contiene información de las variables a estudiar de manera horaria, por un periodo de tiempo que comprende desde el año 1998 hasta el año 2016.

En este sentido, esta investigación tiene como finalidad, en primer lugar, resolver los interrogantes ¿Cuáles son los determinantes en la acumulación de PM_{10} durante las horas pico, entendidas como aquellos momentos del día con más afluencia poblacional en las calles, en cada una de las estaciones de la RMCAB? Para el cual se busca desarrollar un modelo de pronósticos que involucre toda la información disponible obtenida de las estaciones de monitoreo. Así como, indagar sobre ¿Cuáles han sido los impactos económicos y sociales que han traído el hecho de no tener desarrollado un sistema de alerta temprana? y apoyados por el modelo de pronostico ¿Cuál es la mejor manera de implementarlo? De manera que se

facilite la toma de medidas preventivas por parte de la población en escenarios de alta polución.

2. Revisión de Literatura

El pronóstico y la predicción de la calidad del aire, entendida como la cantidad de un contaminante particular, es un campo que se ha venido desarrollando en los últimos años debido a la importancia que ha venido tomando los temas de preservación ambiental, la polución y los efectos diversos que puede tener no solo sobre los seres humanos, sino sobre el planeta mismo. En virtud de ello, en la literatura existen distintas formas de abordar el tema, desde la utilización de modelos "clásicos" como lo son las series temporales o la regresión lineal, hasta modelos más modernos, basados en el aprendizaje estadístico y la minería de datos. Cabe resaltar que en la literatura se encuentra que, sin importar el método predictivo, se tiene el consenso de la predicción de únicamente el promedio diario del perfil de contaminación.

Dentro de estos últimos se encuentran el estudio realizado por Siwek y Osowski (2016) en el cual se adentran en la creación de modelos basados en *Random Forest*, *Support Vector Machines* y Redes neuronales, con el fin de realizar una predicción de la contaminación atmosférica por material particulado, ozono, dióxido de nitrógeno y dióxido de azufre, aplicado al caso de la ciudad polaca de Varsovia. Con el fin de encontrar el mejor input para los modelos, hacen uso de dos metodologías para la selección de predictores; uno de ellos es la aplicación de un algoritmo de optimización genética de manera global y el otro se basa en un método linear de selección (*Stepwise Regression*). El artículo encuentra que llevar a cabo una selección estructurada de predictores, de manera previa a la construcción de modelos, contribuye en gran medida a la mejora en la calidad predictiva de los modelos.

Para el caso de Bogotá, Mejía y Montes (2017) desarrollan modelos para la predicción de la contaminación por material particulado haciendo uso de técnicas de aprendizaje estadístico como lo son los modelos de regresión logística y modelos basados en arboles como *CART* y *Random Forest.* Cabe resaltar que las predicciones generadas por estos modelos no son numéricas, sino ordinales; es decir, los modelos se utilizaron con el fin de encontrar si en un día en particular se daría una contaminación por debajo (o por encima) del límite recomendado por la organización mundial de la salud. En este artículo, los autores combinan

estos métodos de predicción con métodos de selección de variables, para generar 9 modelos por cada una de las estaciones que más datos validos poseían en la red de monitoreo de la ciudad, basado en la selección de variables por parte de expertos y los métodos de selección automática *Forward y Backwards Regression.* Los modelos fueron construidos y comparados con el fin de encontrar el que mejor comportamiento tuviera para cada una de las estaciones, para poder ser analizado en cuanto a la información que provee acerca de las dinámicas que existen entre las distintas variables y como estas pueden variar dependiendo del sitio de medición. Para esto se utilizaron variables atmosféricas y climáticas como predictores de los niveles diarios de PM_{10}.

Por otro lado, el artículo de Karatzas, Pappadopulus y Slini (2002) hace uso de un modelo de regresión lineal para pronosticar el promedio de la concentración de ozono en la ciudad de Atenas. Para la predicción considera distintas variables atmosféricas que de acuerdo a la literatura poseen un papel importante en la dispersión o acumulación de contaminante como lo son la radiación solar, la temperatura y la velocidad del viento, todas rezagadas con el fin de que fueran útiles para la predicción. Él artículo concluye que el modelo de regresión no logra un buen poder predictivo debido a que, por su simpleza, no logra capturar el comportamiento de la contaminación.

Por último, con respecto a los modelos de predicción se encuentran trabajos como el de Cortes (2010) y Quiñones (2015) en los cuales hacen uso de modelos dinámicos, basados en series temporales, con el fin de predecir el nivel de contaminación promedio de un día particular, haciendo uso de variables meteorológicas para mejorar los resultados. Ambos estudios se basan en las estaciones de la ciudad de Bogotá que mayor contaminación presentan (Kennedy y Carvajal en el artículo de Quiñones y Puente Aranda en el artículo de Cortes), dejando de lado las demás estaciones.

Para el caso de análisis de sistema preventivos de alerta, se encuentran los artículos de Cifuentes, Troncoso y de Grange (2012) y el ya mencionado artículo de Karatzas, Pappadopulus y Slini (2002).

Cifuentes, Troncoso y de Grange (2012) llevan a cabo un análisis de impacto con el fin de analizar el número de casos reportados de enfermedades respiratorias, relacionadas con la contaminación, antes y después de la implementación del sistema de alerta temprana en la

ciudad de Santiago de Chile. Para analizar el impacto del sistema, hacen uso de la red de datos proveniente de las estaciones de monitoreo de la ciudad, un sistema similar al que se encuentra en Bogotá. Se analizan las restricciones que se impusieron en los estados críticos de contaminación por la acumulación de material particulado, ozono, monóxido de carbono, dióxido de azufre y óxidos nítricos. Los autores encuentran que la imposición de las alertas genera una disminución significativa en los casos respiratorios de la ciudad, resaltando la utilidad de este tipo de alertas para la prevención y disminución de efectos nocivos sobre la salud de la población.

Por último, se encuentra la investigación de Karatzas, Pappadopulus y Slini (2002). En la cual, acto seguido a la selección del modelo de pronósticos que mejor se comporta en cuanto a las medidas de error clásicas, se realiza una propuesta y un análisis de un sistema de alerta, el cual se pone a prueba con el modelo seleccionado. Cada uno de los escenarios se evalúa con respecto a que tan bien el modelo es capaz de predecir los días en que hay y no hay alertas (Predicciones correctas) y comparándolos con aquellos escenarios en los que el modelo se equivoca por completo (falsos positivos y falsos negativos).

3. Marco Teórico

3.1. Modelos de pronóstico

Basado en la naturaleza de los datos se consideró que los modelos más apropiados para el modelaje de la dinámica de la acumulación de contaminarte serian modelos de regresión dinámica y series de tiempo multivariadas; estos son modelos que no solo consideran a la variable dependiente rezagada en el tiempo, sino que utilizan información de distintas variables, también rezagadas en el tiempo que presenten una correlación con la variable dependiente y puedan ser utilizadas para mejorar su pronóstico.

Entre este tipo de modelos destacan los modelos autorregresivos de rezagos distribuidos (ARDL por su acrónimo en inglés). Estos son modelos dinámicos en los cuales el efecto de una variable regresora (X) sobre la variable dependiente (Y) ocurren a lo largo del tiempo, en lugar de solo un efecto a la vez. (Parker, 2018). En su forma más simple, de una variable explicativa y una relación lineal, el modelo se puede escribir de la siguiente forma:

$$\phi(L)y = \alpha + \theta(L)x + u_t \quad (1)$$

$$y_t = \alpha + \sum_{q=1}^{\infty} \theta_q x_{t-q} + \sum_{p=1}^{\infty} \phi_p y_{t-p} + u_t \quad (2)$$

Donde u_t es el termino estacionario de error, los coeficientes θ representan la manera como la variable x afecta a y a través del tiempo y los coeficientes ϕ representan la manera en que la variable dependiente y se relaciona con sus valores pasados.

El análisis de los modelos ARDL se asemeja aquel de los procesos ARMA univariados; con la diferencia de que la estructura de los rezagos se aplica para la variable explicativa x, y no a los términos de ruido blanco ε. Así, como en los modelos ARMA, los coeficientes de orden q afectan solo los primeros q rezagos del efecto dinámico de la variable x sobre la variable y. El comportamiento de la "cola" de los rezagos, después de q depende enteramente de la estructura autoregresiva del polinomio en $\phi(L)$.

3.2. Material Particulado

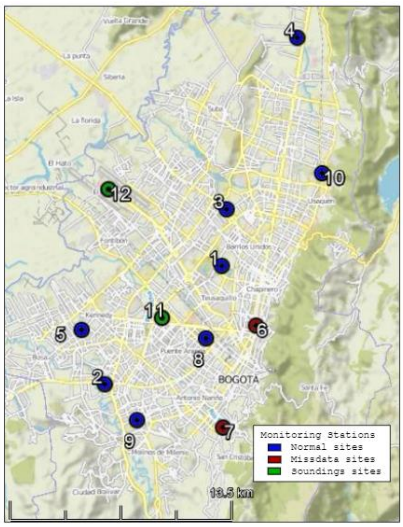
El material particulado es uno de los contaminantes atmosféricos más comunes y estudiados en el mundo. Se define como el conjunto de partículas sólidas y/o liquidas que están presentes de forma suspendida en la atmosfera. (Mészáros, 1999). Estas partículas encuentran su origen a partir de una gran variedad de fuentes naturales o antropogénica y poseen un amplio rango de propiedades de tipo morfológico, físico y químicas. (Arciniegas, 2011).

En la literatura se registran los efectos del material particulado, a la vegetación, los materiales y el hombre, entre los cuales destacan la disminución visual en la atmosfera, producto de la absorción y dispersión de luz por parte del contaminante, además de estar relacionado con el incremento en el riesgo de muerte por enfermedades respiratorias en pacientes adultos. (Pope, 2004)

3.3. Red de monitoreo

La red de monitoreo de la calidad del aire de Bogotá, fue establecida en 1997 y es administrada por la secretaria de ambiente. La red está compuesta por 13 estaciones automatizadas, distribuidas a lo largo de la ciudad, que registran de manera horaria los niveles de contaminación, así como diversas variables meteorológicas. En la figura 1 muestra la ubicación de cada una de las estaciones, junto con su clasificación. De las 13 estaciones que

se muestran en el mapa, en las primeras 11 se toman medidas de las concentraciones de material particulados de 2.5 y 10 micras, junto con las variables climáticas de la temperatura, precipitación, dirección y velocidad del viento. Adicionalmente, de manera intermitente se encuentran, en algunas de estas estaciones, medidas de variables como la humedad relativa, radiación solar, presión barométrica y otros contaminantes como el ozono (O_3), el monóxido de carbono (CO), oxido de nitrógeno (NO), entre otros.



Code	Sitename
1	CAR
2	Carvajal
3	Las Ferias
4	Guaymaral
5	Kennedy
6	Ministry of Environment
7	San Cristobal
8	Suba
9	Tunal
10	Usaquen
11	Puente Aranda
12	El Dorado Airport

Figura 1: Distribución de las estaciones de la RMCAB

En las dos últimas estaciones, correspondientes a Puente Aranda y el aeropuerto el dorado, son los sitios en los cuales, además de registrar variables mencionadas anteriormente, se hacen sondeos atmosféricos todos los días calendario a las 7:00 am. En estos sondeos, se registran mediciones de variables atmosféricas, entre las cuales destacan: el agua precipitable ($PWAT$), la altura de la capa planetaria ($PBLH$), la inversión superficial ($dHInv$). Vale la pena mencionar que estas mediciones se realizan únicamente en las dos estaciones descritas anteriormente, pero se consideran válidas para la ciudad en su totalidad debido a que las condiciones atmosféricas, varían en medida insignificante para un mismo territorio. La base de datos que contiene las mediciones atmosféricas es administrada por la universidad de Wyoming y pueden ser obtenidos de su página web[2].

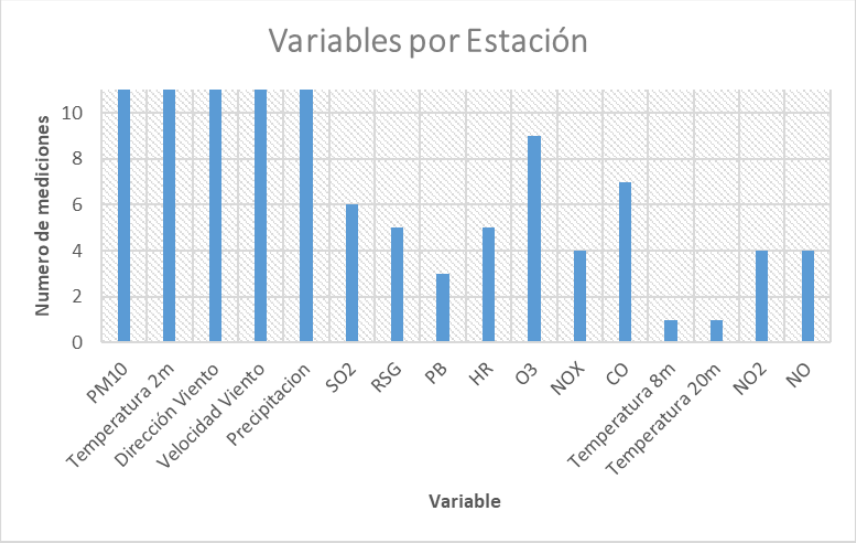
4. Marco Empírico

[2] http://weather.uwyo.edu/upperair/sounding.html

4.1. Análisis de los Datos

El estudio cuenta con una fuente principal de información. Los datos a utilizar, se obtuvieron de los reportes anuales de la red de monitoreo. Estas bases de datos se encuentra la información general para cada uno de los contaminantes y variables climáticas para todas las estaciones de la red, que han existido, y en algunos casos dejado de funcionar, es decir actualmente no son operativas, desde la inauguración de la red. Los registros fueron obtenidos directamente del ente controlador de la red, la secretaria de ambiente de Bogotá. Una de las limitaciones principales de las bases de datos es la inestabilidad que se tiene en los registros. Con esto se refiere al hecho de que por diversas fallas y malfuncionamientos de los equipos en las estaciones de monitoreo existe una cantidad considerable de datos faltantes o inválidos a lo largo de la red. Con el fin de tratar estos problemas se utilizarán metodologías de limpieza (*Data Tidying*) para eliminar todos estos registros no válidos y metodologías de imputación de datos para poder trabajar con series de datos completas.

Un análisis preliminar de los datos muestra como las variables que se registran, no son constante en toda la red. Es decir, si bien existen variables que se miden en todas las estaciones, como el PM_{10}, la temperatura o la precipitación, se encuentran otras que solo se miden en ciertas estaciones, como la radiación solar o la presión barométrica. La grafica 1 muestra cada una de las variables de la base, junto con el número de estaciones en las cuales se mide. Por esta razón se decide, para el pronóstico de cada estación, utilizar todas las variables disponibles que resulten relevantes.



Grafica 1: Medición de variables en la RMCAB

Una vez se tienen los datos, se propone una metodología estructurada para la realización de los modelos de pronóstico de contaminación y análisis de alertas. La metodología puede ser dividida en dos partes: La primera parte considera la manipulación inicial de los datos, en la cual se define el horizonte de tiempo del estudio, así como las estaciones a utilizar. Una vez se definen estos dos parámetros se procede a la construcción individual de las bases de datos para cada una de las estaciones. Cada una de estas bases contiene los registros de todas las variables medidas en cada sitio. La figura 2 muestra un resumen de la parte inicial de la metodología.

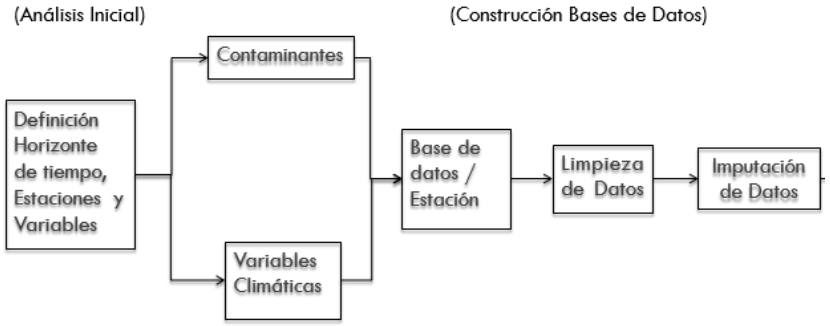
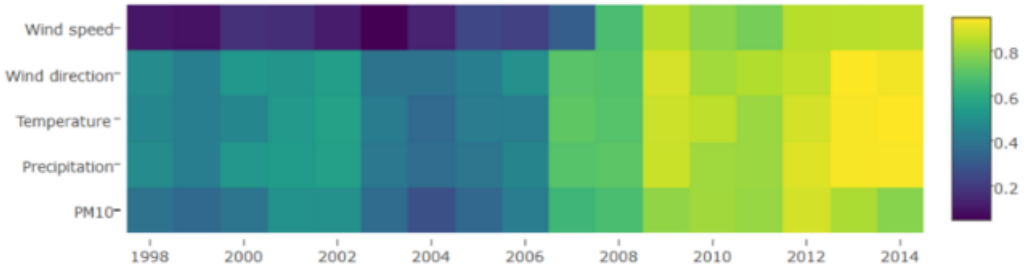


Figura 2: Metodología para la selección del horizonte de estudio y creación de bases.

Como se muestra en la figura 2, el primer paso en la metodología es la definición del horizonte de estudio, entendido como los años y lugares a analizar. Para la selección del horizonte de tiempo del estudio se realizó un análisis de la densidad de datos validos a la cual se tiene acceso según los datos obtenidos. La definición del horizonte de tiempo es un paso vital en el estudio pues, con el fin de poder obtener series de datos validas a la hora de realizar imputaciones, se requiere que la serie original posea la mayor información posible, esto es en general una densidad de datos mayor al 70%.

En primer lugar, se limpió las bases de datos mediante el uso de limpieza, definidas por Bernal y Melo (2016). Las reglas consideran la eliminación de todos los valores inválidos, que quedan registrados como cadenas de texto, así como mediciones por fuera de los rangos, que curren por malfuncionamiento en los equipos de medición. Esto considera casos de: “Sin Data”, “< Muestra”, “Apagado”, “Cero”, “Fail Tech” presentes en los registros de las bases de datos.

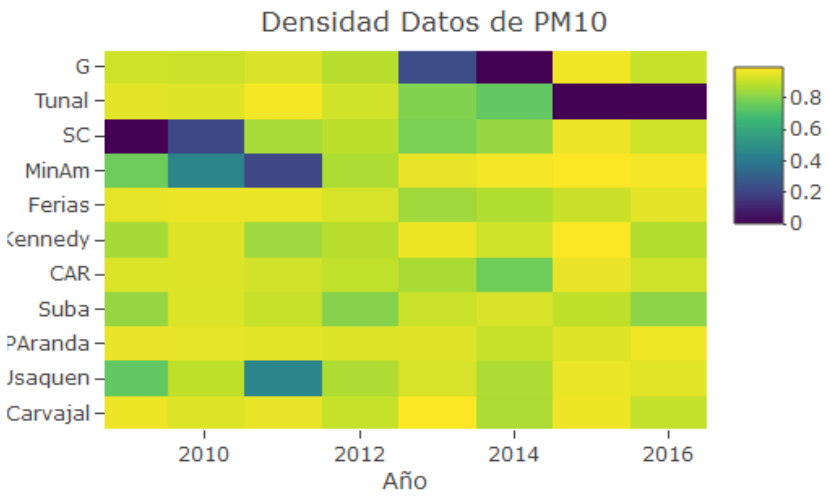
Una vez se limpia la base, se construyó un gráfico de calor que muestra la distribución de la densidad de los datos, medida como la cantidad de datos validos sobre el total de datos que se debería tener para cada año. Este análisis se realizó para las variables de las cuales mas registros se tienen: siendo el PM_{10}, la temperatura a 2m, la precipitación y la dirección y velocidad del viento, las cuales son medidas en todas las estaciones.



Gráfica 2: Densidad de datos validos 1998-2016

Como puede observarse en la gráfica 2, los primeros años de establecida la red de monitoreo se tenía una muy mala calidad de los datos. Este comportamiento ha ido mejorando con el pasar del tiempo, lo cual obedece a las mejoras y constantes mantenimientos que se han realizado a los equipos de registro. Adicionalmente, se puede apreciar que aproximadamente desde el año 2009 en adelante se logra el objetivo de un porcentaje de datos validos superior al 70-75%. Así se define como el horizonte de tiempo de estudio los años desde el 2009 hasta el 2016.

Una vez se ha definido el horizonte temporal del trabajo, se procede a analizar las estaciones de la base para seleccionar aquellas con la mejor calidad de datos. De manera similar, se construye un mapa de calor para los datos de material particulado en cada una de las estaciones de la red, con el fin de discriminar las estaciones con una mayor cantidad de los mismos. Una vez más, se utiliza como criterio de decisión, un porcentaje de datos validos superior al 70%



Gráfica 3: Densidad de datos de PM_{10} en las estaciones de la RMCAB

En la Gráfica 3 se puede observar que en general todas las estaciones poseen una densidad aceptable de datos desde el año 2009. Estaciones como lo son Tunal, San Cristóbal y Ministerio de ambiente parecen saltar a la vista, debido a su inconsistencia en los datos. La estación de tunal presenta los dos últimos años, totalmente vacíos, es decir no se presenta ningún dato valido para cada año; así mismo los años anteriores (2013 y 204) poseen densidades ligeramente inferiores al 70%. La estación ubicada en el ministerio de ambiente, presenta la peor densidad de datos de todas las estaciones, no solo en material particulado, sino también en las variables meteorológicas; presenta 3 años de datos inutilizables. Por último, la estación de San Cristóbal, es más inconsistente en los datos inválidos; presenta los años 2009 y 2010 con densidades de datos menores al 20%, seguido de los años 2013 y 2014 con densidades de datos menores al 70%.

La baja calidad de los datos registrados en estas estaciones se presenta como un obstáculo para el estudio, ya que, en caso de ser utilizada cualquiera de estas series, dificulta la generación de una serie completa valida, generando sesgos y finalmente invalidando el análisis y las conclusiones que de este se puedan derivar.

Así, después del análisis, se seleccionan únicamente 8 de las estaciones de la RMCAB, que son, todas aquellas que cumplían los criterios definidos de densidad de datos válidos. La tabla 1 muestra las principales estadísticas descriptivas de las concentraciones de material particulado para cada una de las estaciones seleccionadas, donde se puede observar en la

tabla que todas las estaciones cumplen con el criterio de densidad de datos, teniendo todos valores superiores al 70%

	Media	Desviación	Mínimo	Máximo	Datos Válidos	Datos Faltantes	% Faltantes
Carvajal	84,89	42,58	1	448	65941	4165	5,94%
Usaquén	38,13	23,58	0	598	58683	11423	16,29%
Puente Aranda	54,15	34,08	0	339	66574	3532	5,04%
Kennedy	73,68	38,59	0,7	471	64034	6072	8,66%
Ferias	39,6	25,97	0	257	65014	5092	7,26%
Guaymaral	34,83	20,88	0	381,5	50719	19387	27,65%
Suba	54,03	27,24	1	701	61939	8167	11,65%
CAR	35,7	26,96	0	275	63473	6633	9,46%

Tabla 1: Estadísticas Descriptivas de la concentración de PM_{10} por cada estación

De acuerdo a las estadísticas se puede ver la gran disparidad que existe entre las mediciones de la red. Pueden diferenciarse 3 tipos de estaciones: Las estaciones con contaminación alta, que en este caso corresponden a Kennedy y Carvajal donde incluso el promedio sobre los años de estudio sobrepasa los límites sugeridos. Las estaciones con contaminación media, en los cuales se sobrepasan los límites, pero en un margen muy inferior al de las estaciones altas; se consideran las estaciones de Puente Aranda y Suba. Y por último las estaciones de contaminación baja, donde en promedio se cumplen los limites, con concentraciones de alrededor de 34 $\mu g/m^3$.

Una vez se definen las estaciones se procede a la consolidación de las bases de datos para cada una de las estaciones, conteniendo cada una de las variables que se miden dentro de cada una de ellas. Vale la pena resaltar, que previo a la imputación y elaboración de modelos de pronósticos, se realizara un análisis similar para la selección de posibles predictores basado en la densidad de datos.

4.2. Imputación de Datos Faltantes

Teniendo las bases de datos construidas se procede al último paso de la primera parte de la metodología propuesta: la imputación de datos. Para el relleno de las series se utilizó el *Site Dependant Effect Method (SDEM)* propuesto por Plaia y Bondi (2006). El método es ampliamente utilizado pues tiene en cuenta la correlación temporal y espacial de los datos, además de superar métodos de imputación singular y múltiple (Westerlund, Urbain, & Bonilla, 2014). El método, realiza la imputación considerando el promedio de la contaminación en la semana, día de la semana y hora. Adicionalmente, considera la

diferencia del sitio analizado con el promedio de un vecindario, compuesto por aquellas estaciones que presentan una mayor correlación. La imputación se realiza siguiendo la ecuación:

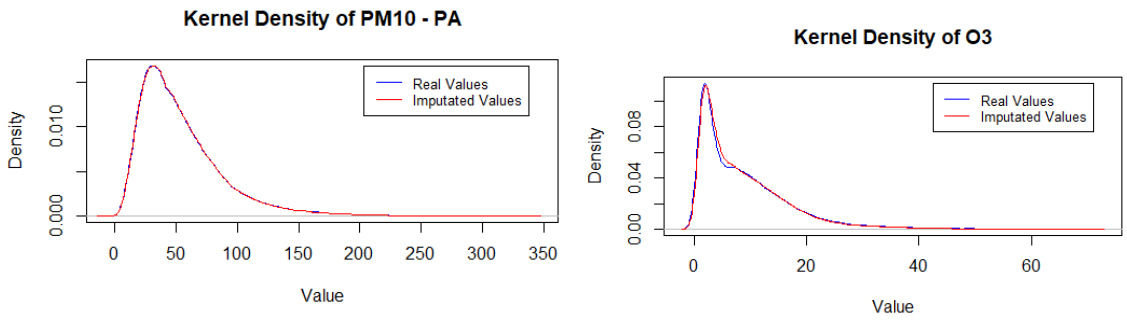
$$\hat{x}_{sdwh} = \bar{x}_{.wdh} + \frac{1}{2}\left(\bar{x}_{sw..} - \sum_{s=1}^{S}\frac{\bar{x}_{sw..}}{S}\right) + \frac{1}{2}\left(\bar{x}_{s.d.} - \sum_{s=1}^{S}\frac{\bar{x}_{s.d.}}{S}\right) + \frac{1}{2}\left(\bar{x}_{s..h} - \sum_{s=1}^{S}\frac{\bar{x}_{s..h}}{S}\right) \quad (3)$$

De manera complementaria, para los datos e instancias que no pudieron ser imputadas por el método *SDEM*, se utilizó el método de cicloestacionariedad, el cual consiste en reemplazar el dato faltante por el promedio de los datos correspondientes a la misma hora, dio semana y trimestre del año, en la estación analizada. Ambos métodos descritos se utilizaron para todas las variables que pasaban el limite necesario de densidad para las estaciones seleccionadas. La tabla 2 muestra las variables resultantes en cada estación.

Kennedy	Suba	CAR	Puente Aranda	Ferias	Carvajal	Guaymaral	Usaquén
Temperatura	Temperatura	Temperatura	Temperatura	Temperatura	Temperatura	Dirección del Viento	Temperatura
Dirección del Viento	Dirección del Viento	Dirección del Viento	Dirección del Viento	Dirección del Viento	Dirección del Viento	Velocidad del Viento	Dirección del Viento
Velocidad del Viento	Velocidad del Viento	Velocidad del Viento	Velocidad del Viento	Velocidad del Viento	Velocidad del Viento	Precipitacion	Velocidad del Viento
Precipitación	Precipitación	Precipitación	Precipitación	Precipitación	Precipitación	Precipitación	Precipitación
Radiación Solar	SO2	SO2	O3	Humedad Relativa	SO2	O3	O3
Presión Barométrica	O3	O3	NOX	Presión Barométrica	O3	NOX	CO
Humedad Relativa		Radiación Solar		O3		Temperatura	
CO		Humedad Relativa		SO2		Temperatura 8m	
						Temperatura 20m	
						Humedad Relativa	
						Presión Barométrica	

Tabla 2: Variables finales por cada estación

Con el fin de analizar la validez de las imputaciones realizadas, se construyeron gráficos de densidad (Kernel) de las series de datos antes y después de su imputación con el fin de verificar que la distribución de los datos se mantuviera igual. Las gráficas 5 y 6 muestran un ejemplo para la estación de Puente Aranda. En estas se muestra la densidad real de la variable analizada, junto con la densidad de la variable con los valores imputados.



Graficas 5 y 6: Densidad Kernel de los datos originales e imputados del PM_{10} y el O_3 de la estación de Puente Aranda.

Con las bases de datos completas, se ha terminado la primera parte de la metodología. En este sentido, la segunda parte consiste en la construcción, comparación y validación de los modelos de pronostico ARDL, junto con el análisis del sistema de alertas. La figura 3 muestra las etapas correspondientes.

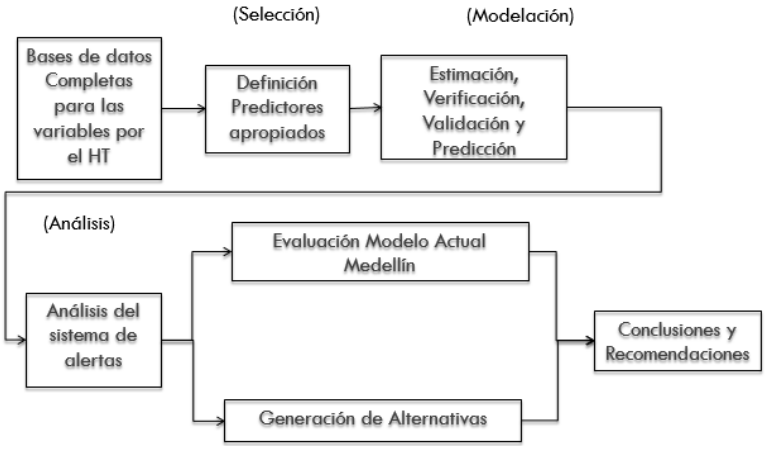


Figura 3: Metodología para la construcción, validación y análisis basado en modelos de pronóstico.

4.3. Modelos de Pronóstico

Como primer paso en la construcción de los modelos, se debe validar que cada una de las variables con las cuales se va a trabajar sean estacionarias, de modo que se cumplan los supuestos del modelo ARDL y que este resulte significativo y útil a la hora del ser utilizado para el pronóstico. Para comprobar la estacionariedad de las variables se hace uso de la prueba de Dickey-Fuller aumentada la cual se basa en el modelo:

$$\Delta Y_t = \alpha_0 + \gamma Y_{t-1} + \sum_{j=1}^{p} \beta_j Y_{t-j} + u_t \qquad (4)$$

En el cual, se prueba si $\gamma = \phi - 1 = 0$, es decir $\phi = 1$, lo cual implica que el proceso tiene raíz unitaria y no es estacionario.

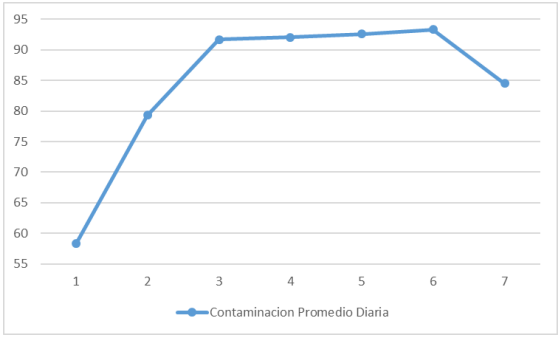
En este sentido, se aplicó la prueba de raíz unitaria a los datos de cada una de las estaciones, agrupados por hora del día. Es decir, la prueba se realizó 24 veces para las variables de cada una de las estaciones, una por cada hora.

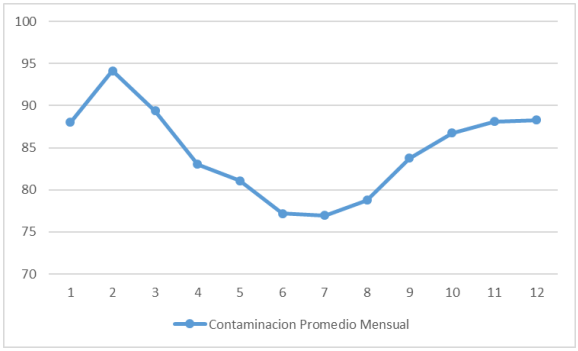
En cada una de las pruebas se revisó el p-valor para concluir. No hubo ningún caso en el cual diera un p-valor mayor a una significancia del 10%. Se encontró que en todos los casos se rechaza la hipótesis nula de raíz unitaria. Por esta razón se concluye que todas las variables del estudio son estacionarias y no es necesario aplicar diferenciación a las mismas. La tabla 3 muestra un resumen de los resultados para la estación de Usaquén.

	Usaquén						
Hora	**CO**	**Dirección_Viento**	**Precipitación**	**Temperatura**	**Velocidad_Viento**	**O3**	**PM10**
0:00	-7,74	-10,42	-13,02	-7,77	-9,06	-7,65	-7,27
1:00	-8,45	-11,25	-12,07	-7,82	-9,15	-7,78	-7,04
2:00	-8,02	-11,55	-11,50	-8,09	-9,05	-7,63	-6,68
3:00	-8,30	-10,66	-13,30	-8,33	-9,40	-7,55	-6,96
4:00	-8,44	-11,83	-13,81	-8,59	-9,21	-7,83	-6,70
5:00	-8,48	-11,28	-14,13	-8,75	-9,52	-7,90	-6,81
6:00	-9,11	-12,05	-13,94	-8,59	-9,33	-8,13	-6,71
7:00	-8,68	-12,55	-14,11	-7,82	-8,54	-8,94	-7,15
8:00	-9,17	-13,18	-13,48	-6,97	-8,78	-8,54	-7,97
9:00	-9,99	-13,14	-13,43	-7,39	-8,66	-7,70	-8,06
10:00	-10,23	-13,35	-12,60	-7,30	-8,63	-6,71	-8,26
11:00	-10,84	-12,65	-13,86	-6,59	-8,67	-6,13	-7,90
12:00	-11,18	-12,30	-13,64	-6,21	-8,60	-6,06	-7,49
13:00	-10,97	-11,54	-13,32	-6,50	-8,77	-6,34	-7,34
14:00	-10,13	-11,03	-10,81	-7,21	-9,59	-6,83	-7,46
15:00	-9,26	-10,41	-12,28	-8,70	-9,20	-6,98	-7,83
16:00	-8,31	-11,24	-13,72	-9,50	-10,11	-6,88	-7,61
17:00	-7,96	-11,68	-13,55	-9,18	-9,85	-7,33	-7,46
18:00	-7,71	-10,84	-13,48	-8,16	-9,42	-7,55	-7,35
19:00	-7,38	-10,87	-14,08	-7,11	-9,96	-7,91	-7,38
20:00	-7,89	-10,70	-13,91	-6,98	-9,56	-8,05	-7,38
21:00	-7,69	-11,32	-13,34	-7,04	-9,29	-8,06	-7,39
22:00	-7,16	-10,10	-13,80	-7,19	-8,92	-7,98	-7,36
23:00	-7,73	-10,22	-14,12	-7,48	-8,97	-7,93	-7,26

Tabla 3: Estadístico Tau, prueba de Dickey-Fuller para la estación de Usaquén

Una vez se comprobó la estacionariedad de las series de datos, se generan rezagos de las variables climáticas y de los contaminantes. Se decide tomar como un número máximo 7 rezagos debido a la estacionalidad semanal que presentan las series de material particulado. De manera adicional, se crea un sistema de variables dummy para el día de la semana y el mes del año, esto, respondiendo al hecho de que las series de contaminantes presentan días y meses con comportamientos muy diferentes, siendo los domingos los días menos contaminados, lo cual va aumentando conforme pasa la semana, para volver a caer los fines de semana.





Graficas 7 y 8: Contaminación promedio diaria y mensual

Un comportamiento similar se puede ver en los meses, con los meses de vacaciones, siendo los menos contaminados debido a que la mayoría de la gente sale de la ciudad por esas fechas, reduciéndose el flujo vehicular y permitiendo una disminución en la cantidad de material particulado.

Con los rezagos y nuevas variables generadas, se plantea la forma general del modelo a construir:

$$PM_{10_{it}} = \beta_o + \sum_{j=1}^{n} \beta_j \left(PM_{10_{it-j}}\right) + \sum_{k=1}^{m} \gamma_k (Env_{it-k}) + \sum_{l=1}^{o} \theta_l (Env_{it-l}^2) + Dia_t + Mes_t + \varepsilon_t \quad (5)$$

El modelo pone la cantidad de PM_{10} registrado en la estación i a la hora t, en función del material particulado de dicha hora y estación, rezagado en el tiempo; variables atmosféricas (Env) , también rezagadas en el tiempo; se introducen variables atmosféricas al cuadrado con el fin de recoger cualquier tipo de relación no lineal que tenga el PM_{10} con estas; terminando por las variables binarias que representan el día de la semana y el mes del año.

Debido a la gran cantidad de variables que resultan para cada estación, se buscó reducir la dimensionalidad del problema, Esto es encontrar aquellas variables más relevantes y con mayor incidencia en la cantidad de PM_{10} a una hora específica. Por ejemplo, para Usaquén, se tendrían: 7 rezagos para cada variable inicial (42 en total), cuadrados de las variables climáticas (4), variables binarias de día de la semana (6) y mes del año (11), para un total de 63 variables, lo cual representa un aumento en tiempo computacional a la hora de estimar los parámetros y adicionalmente representa un problema en la parsimonia e interpretabilidad del modelo resultante.

Para este fin, se estudió la posibilidad de implementar métodos multivariados como componentes principales y métodos de selección automática. Componentes principales fue descartado debido al hecho de que, para disminuir la cantidad de variables del problema, genera una combinación lineal de todas las variables presentes, resultando en componentes sin interpretación conceptual, imposibilitando poder investigar las dinámicas de contaminación en cada estación de la red. Por el lado de los métodos de selección automática se optó por utilizar la metodología General to Specific (GETS) la cual combina el método de Backwards Elimination, pruebas de hipótesis simples y múltiples, estadísticos de diagnóstico y bondad de ajuste. Estos ingredientes se combinan con el fin de encontrar una distinta variedad de modelos, de modo que puedan ser comparados entre sí, para terminar con un modelo parsimoniosos y estadísticamente valido, que provea el mejor ajuste a los datos que están siendo analizados (Pretis, Reade, & Sucarrat, 2016).

Con la metodología GETS se construyeron los modelos para cada estación, partiendo del número máximo de variables, respectivamente. El uso de GETS permite obtener modelos con un número máximo de 23 variables, sin tener que probar todas las combinaciones posibles de las variables totales. Adicionalmente, debido a la naturaleza del algoritmo permite obtener modelos que ya pasaron la etapa de validación en cuanto a la correlación serial del error, sus residuos son ruido blanco, y la homocedasticidad.

Hora	Variables Seleccionadas por Modelo				
7:00 a. m.	AR1	AR2	AR4	AR5	AR7
	PreciL1	TempL1	VVL1	O3L1	COL2
	COL3	PreciL3	TempL3	O3L2	O3L3
	PreciCuad	TempCuad	VVCuad	Lunes	Martes
	Miercoles	Jueves	Viernes	Sabado	Septiembre
12:00 a. m.	AR1	AR2	AR3	AR6	DVL1
	TempL1	VVL1	O3L1	COL2	COL3
	DVL2	O3L2	O3L3	DVCuad	TempCuad
	VVCuad	Martes	Miercoles	Jueves	Viernes
	Sabado				
5:00 p. m.	AR1	AR2	AR3	AR4	AR6
	AR7	COL1	DVL1	VVL1	O3L1
	COL2	COL3	DVL2	DVL3	TempL2
	VVl2	VVL3	O3L2	O3L3	DVCuad
	TempCuad	VVCuad	Lunes	Martes	Miercoles
	Jueves	Viernes	Febrero		

Hora	Coeficientes Variables Seleccionadas por Modelo				
7:00 a. m.	0,805	0,187	0,033	0,031	0,048
	-2,465	7,220	-1,750	-0,247	8,720
	-10,221	-2,675	-5,313	-0,327	0,288
	0,112	-0,153	0,438	4,663	3,677
	4,600	5,779	5,317	2,536	3,902
12:00 a. m.	0,613	0,033	0,041	0,040	-0,039
	0,718	-4,186	0,340	16,907	-9,610
	0,009	-0,787	0,567	0,000	-0,024
	0,701	1,781	3,574	2,662	4,000
	1,720				
5:00 p. m.	0,384	0,218	0,043	0,037	0,037
	0,031	-3,665	-0,055	-3,610	-0,180
	14,605	-8,292	0,033	-0,017	1,002
	-2,731	0,949	0,496	-0,204	0,000
	-0,034	0,621	1,594	3,896	5,279
	4,766	6,009	2,699		

Tabla 4[3] y 5: Variables y coeficientes resultantes estación referencia Usaquén.

La tabla 4 muestra los modelos resultantes en 3 franjas horarias, distribuidas a lo largo del día en la estación de Usaquén, junto con las variables que resultaron después de la eliminación iterada. A primera vista, resulta interesante el darse cuenta la heterogeneidad que existe de las variables que resultan significativas en cada una de las horas del día. La precipitación, así como sus rezagos, son significativos únicamente para los modelos de la mañana. De la misma manera, los modelos de la tarde resultan siendo los únicos que no consideran relevante la información proveniente de la temperatura. Otro resultado interesante es la inclusión de todas las variables dicótomas de días lo cual soporta el hecho de que existe una estacionalidad relevante en la dinámica de acumulación del material particulado, haciendo que esta sea diferente cada uno de los días de la semana. Por el contrario, las variables dicótomas de los meses del año parecen no ser relevantes para el estudio; únicamente son consideradas de a una, en los modelos de inicio y final del día, indicando que en su mayoría los meses no poseen comportamientos diferentes en acumulaciones de PM_{10} al ser separados por horas.

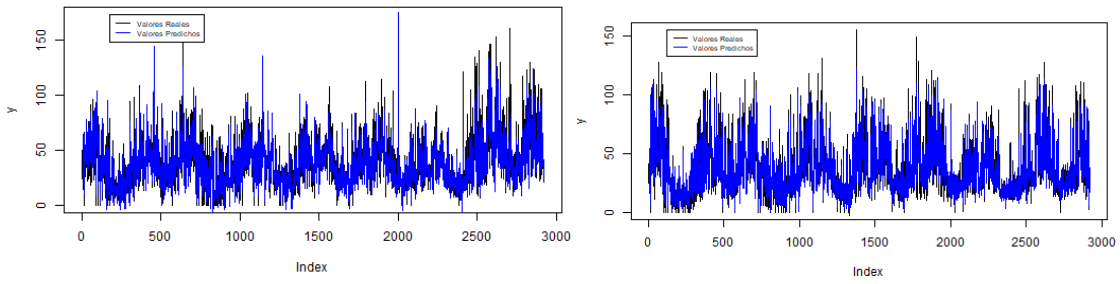
Se puede observar la importancia que toman los distintos contaminantes, como el O3 o el CO en la acumulación de material particulado, indicando una relación estrecha entre estos, cuyo

[3] Las variables AR hacen referencia a el valor del material particulado rezagado en el tiempo. Algunas variables poseen un LX al lado del nombre, lo cual indica que están rezagadas un número X de periodos. VV se refiere a la velocidad del viento. DV se refiere a la dirección del viento. Temp se refiere a temperatura. O3 se refiere a ozono. CO se refiere a monóxido de carbono. Las variables con Cuad, se refieren al cuadrado de dicha variable.

comportamiento puede ayudar a explicar los cambios en PM_{10} debido a la relación que existe y se busca explorar entre el material particulado y la emisión de carbono por parte de vehículos motorizados en el tráfico de las distintas horas. Esto nos indica que el problema de contaminación no está únicamente limitado al PM_{10} sino constituye una variedad de contaminantes que igualmente son altamente nocivos para la salud de la población.

Por otro lado, variables como la precipitación y la velocidad del viento aparecen con coeficientes negativos, actuando como agentes dispersores de material particulado, mientras que la temperatura presenta un coeficiente positivo presentándose como un agente que permite y aumenta la acumulación de contaminación. La dirección de los coeficientes es coherente en vista a la literatura, lo cual aporta una mayor robustez y validez al modelo creado.

Por último, las gráficas 9 y 10 muestran los valores reales (en negro), junto a los predichos por los modelos desarrollados (en azul) para la estación de Usaquén en las horas 7:00 am y 5:00 pm, respectivamente. Como se puede observar, los modelos presentan un buen ajuste gráfico, el hecho de utilizar modelos autorregresivos permite a las predicciones imitar y seguir el comportamiento de la serie original, permitiéndolo reaccionar ante los súbitos cambios que presenta en subidas y bajadas.



Graficas 8 y 9: Pronósticos vs Reales para Usaquén a las 7:00 am y 5:00 pm, respectivamente.

Una vez se tienen los modelos, estos deben ser validados y verificados con el fin de determinar que sean estadísticamente válidos y significativos para la explicación de la contaminación. Como se mencionó anteriormente, debido a la naturaleza del algoritmo GETS, utilizado para la construcción de los modelos, el resultante resulta valido estadísticamente en el sentido que cumple los supuestos del modelo. Para realizar la

verificación, se mide la calidad del pronóstico en términos de la raíz del error cuadrado medio:

$$RMSE = \sqrt[2]{\sum \frac{\left(Y_i - \widehat{Y}_i\right)^2}{n}} \quad (6)$$

Esta medida nos permite ver la distancia promedio que existe entre los valores reales y los valores generados utilizando el modelo. En esta medida se busca el menor valor posible, lo cual sería un indicativo del buen ajuste del pronóstico. El RMSE es relativo a la variable que se esté midiendo, es decir es un valor en unidades de la variable sobre el cual no existe un punto de referencia o de corte para juzgar un buen modelo. Por esta razón se propone la utilización del coeficiente de Theil:

$$U = \frac{\left(\sqrt[2]{\sum \frac{\left(Y_i - \widehat{Y}_i\right)^2}{n}}\right)}{\sqrt[2]{\sum \frac{(Y_i)^2}{n}} + \sqrt[2]{\sum \frac{\left(\widehat{Y}_i\right)^2}{n}}} \quad (7)$$

El cual también mide la distancia o desigualdad que existe entre los valores reales y los predichos, pero los normaliza a una escala entre 0 y 1; siendo 1 la máxima desigualdad y 0 un pronóstico perfecto. De este modo teniendo un valor de referencia bajo el cual evaluar la calidad de los modelos construidos. La tabla 6 muestra los resultados de la evaluación de todos los modelos de pronóstico de la estación de Usaquén, con respecto a las dos medidas mencionadas.

Hora	RMSE	Coef Theil	Hora	RMSE	Coef Theil
0	8,776	0,116	**12**	11,245	0,137
1	8,683	0,122	**13**	10,902	0,141
2	8,392	0,123	**14**	12,861	0,167
3	8,293	0,126	**15**	13,585	0,161
4	7,771	0,120	**16**	15,527	0,176
5	7,495	0,117	**17**	13,271	0,154
6	8,808	0,128	**18**	11,509	0,132
7	11,501	0,131	**19**	11,781	0,139
8	16,364	0,139	**20**	11,035	0,133
9	15,804	0,133	**21**	10,419	0,126
10	14,370	0,133	**22**	9,922	0,122
11	12,313	0,135	**23**	9,600	0,121

Tabla 6: RMSE y U de Theil para los modelos de la Estación de Usaquén

En la tabla se puede apreciar que los modelos poseen valores bajos en ambas medidas. Los valores del RMSE yendo desde 7.4 como el valor mínimo a 16.36 como el valor máximo. Por su parte los valores del coeficiente de Theil son cercanos a 0. Unos valores bajos del RMSE corresponden a valores bajo del coeficiente de Theil, dándole consistencia a los resultados. Se puede ver que los modelos con los valores más bajos corresponden a las horas del final/inicio del día, como desde las 10:00 p.m. hasta las 3:00 a.m. Esto tienen sentido debido a que dichas horas, son en las que menos afluencia de gente y trafico existe por lo tanto deben ser, aquellas horas donde los niveles de contaminación se presentan más estables, lo cual facilita su predicción. De igual manera, las horas de mayor interés como las 6 y 7 a.m. o las 5,6 y 7 p.m. también presentan valores bajos, mostrando un buen ajuste para dichos modelos, haciéndolos no solo útiles sino relevantes para la predicción del contaminante en las horas de mayor afluencia poblacional, permitiendo la toma de iniciativas preventivas que ayuden a disminuir el impacto que la contaminación tiene en el ambiente de la ciudad y su población.

4.4. Análisis de Alertas

La RMCAB tiene una deficiencia en la falta de un apropiado sistema de alertas temprana. Los más parecido que se encuentra en la ciudad es el índice de calidad de aire, el cual mide la concentración de diversos contaminantes, los pondera y entrega un resultado que clasifica

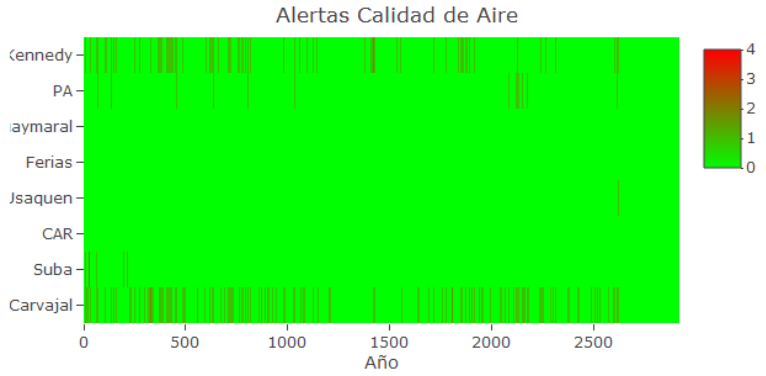
en 5 categorías, el índice actualmente cumple una función meramente informativa y cuenta con recomendaciones para la ciudadanía (Alcaldía Mayor de Bogotá D.C., 2015).

En este sentido, se utilizará como caso de referencia el único sistema de alerta temprana en el país, del cual se tiene conocimiento. El Plan Operacional para Enfrentar Episodios Críticos de Contaminación Atmosférica, que está en operación en la ciudad de Medellín desde el 28 de noviembre de 2016 (Area Metropolitana Valle de Aburrá, 2016). La reglamentación del POECA para el caso específico del material particulado de 10 micras se presenta en la tabla a continuación.

Contaminante	Tiempo de Exposición	Unidad	Niveles de Contingencia en los valores del ICA y en concentraciones $\mu g/m^3$			
			Alerta Naranja	Alerta Roja Fase I	Alerta Roja Fase II	Emergencia
PM10	24 Horas	Adimensional	101-150	151-177	178-200	$\geq$ 201
		$\mu g/m^3$	155-254	255-308	309-354	$\geq$ 355

Tabla 7: Reglamentación POECA Medellín

Para poder aplicar el escenario de alertas presentado en Medellín, se obtuvieron los promedios diarios de los niveles de contaminantes registrados por cada una de las estaciones de la red. Una vez se tiene estos valores, se genera una variable para cada uno de los días que indica en cuál de los cinco escenarios se encontró la contaminación. Se consideran 5 escenarios en total, un escenario bueno y los 4 escenarios de alerta contemplados en el acuerdo. Los resultados se pueden observar en la gráfica 10.



Grafica 10: Análisis de alertas por PM_{10} en la RMCAB

En la gráfica se puede ver como en la mayoría del horizonte de tiempo analizado, la red presenta mediciones aceptables por debajo de 155 $\mu g/m^3$, lo que corresponde al color verde en la gráfica, indicando que el 94.5% de los casos analizados corresponden al mejor escenario. Aun así, se puede observar tintes de rojo a lo largo de la gráfica, estos corresponden en su mayoría a alertas naranjas, con un 5.38% de ocurrencias, estos corresponden en un 92.44% a las estaciones de Carvajal y Kennedy, las cuales son históricamente las estaciones con el peor desempeño en la calidad del aire.

Escenario	% Ocurrencia
Bueno	94,5053%
Alerta Naranja	5,3834%
Alerta Roja Fase I	0,0856%
Alerta Roja Fase II	0,0171%
Emergencia	0,0086%

Tabla 8: Porcentaje de ocurrencia de cada escenario en la RMCAB

En estas dos estaciones se presentan los casos más preocupantes; en Carvajal en 760 días (26.7%) debió lanzarse una alerta y tomado medidas preventivas para evitar la subsecuente acumulación, mientras que en la estacione de Kennedy en 423 días (14.4%) debió tomarse medidas. Por ultimo en toda la red, de los 2921 días analizados, en 859 (29%) debió haberse lanzado una alerta de prevención, lo cual habría permitido a la población tomar medidas que evitaran su contacto y exposición al material particulado y prevenido la generación de casos de enfermedad respiratoria aguda.

4.5. Capacidad Predictiva de los Modelos ARDL

Con el fin de complementar ambas partes del estudio, los modelos de pronostico y el sistema de alertas, se busca medir la capacidad que tienen los modelos desarrollados en cada estación para pronosticar la necesidad de una alerta dependiendo de la contaminación en la atmosfera de la región. Para esto, se obtienen los pronósticos de todos los modelos y se obtienen promedio diarios. Con los datos diarios, se generan los escenarios adecuados para cada estación y estos se comparan contra los escenarios reales que se presentaron. Esto nos permite medir la calidad del modelo en términos de la matriz de predicciones correctas, más específicamente de la sensibilidad del modelo, entendida como la capacidad de que el modelo

prediga la alerta, dado que efectivamente se produjo una alerta. En la tabla 9 se puede ver la calidad diagnostica del modelo para cada uno de los escenarios de alerta

No Alerta		Predicho	
		1	0
Real	1	2133	22
	0	138	621

Alerta Naranja		Predicho	
		1	0
Real	1	608	138
	0	32	2136

Alerta Roja I		Predicho	
		1	0
Real	1	1	10
	0	1	2902

Alerta Roja II		Predicho	
		1	0
Real	1	1	0
	0	1	2912

Emergencia		Predicho	
		1	0
Real	1	1	0
	0	0	2913

Tabla 9: Matrices de Confusión para cada escenario de contaminación Carvajal

En las matrices de confusión el 1 representa que ocurre el evento en cuestión y 0 es cualquier otro evento. En este sentido, el promedio de los modelos de la estación de Carvajal es muy útiles para la predicción de todo tipo de alertas (sensibilidades por encima del 80%) a excepción de los escenarios de alerta roja fase 1, en los cuales solo se acierta el 9% de las veces. Estos resultados resultan interesantes, debido a la viabilidad que presentan los modelos como instrumento de política ambiental, diagnosticando de alertas y habilitando su uso para la identificación y formulación de alertas tempranas que permitan contrarrestar o, al menos, disminuir la contaminación y la exposición poblacional.

4.6. Análisis de Costos Incurridos

Para calcular el costo de oportunidad que se ha incurrido por la falta en el desarrollo e implementación de un sistema de alerta temprana, se busca calcular el monto económico que le ha costado a la ciudad los escenarios de alerta en términos del número de consultas que se generan por el PM_{10}.

Con este fin, se toma como referencia el trabajo de Hernández (2017), en el cual se documenta una relación directa entre el nivel de material particulado y el número de consultas respiratorias atendidas, en la localidad de Carvajal. En el estudio, utilizan un modelo log-log con el fin de medir los efectos porcentuales que un aumento en el PM_{10} tiene sobre el

promedio de consultas al día. En este sentido se encuentra que un aumento de 10% en los niveles de contaminación lleva a un aumento de 7.6% en la cantidad de consultas atendidas al día, en salas ERA (Hernández, 2017).

Para obtener el costo que una consulta representa a la entidad prestadora del servicio de salud nos remitimos al estudio realizado por Pérez et, al (2007) donde encuentra que los costos directos que asume el prestador del servicio se dividen en: diagnostico, manejo ambulatorio, cirugía, cuidados intensivos y hospitalización. Así, ponderando las frecuencias relativas de los casos leves, moderados o graves de enfermedades respiratorio, se encuentra que los costos asociados a estos diagnósticos son el promedio $848,1 USD del año 2007, por caso atendido. (Pérez, Murillo, Pinzon, & Hernández, 2007).

Una vez se tienen los insumos mencionados anteriormente, se aplican al caso de la RMCAB. Para esto se calculan los promedios de contaminación diaria en toda la red, así como los promedios diarios de consultas por casos respiratorios en toda la ciudad (Observatorio Ambiental, 2017) con el fin de usarlos de referencia para aplicar las relaciones encontradas por Hernández (2017).

Consultas Atendidas/Dia	Año
102,934	2009
93,299	2010
69,781	2011
86,381	2012
114,534	2013
130,359	2014
135,595	2015
142,932	2016

Contaminación Promedio Diaria	Año
91,912	2009
90,431	2010
85,725	2011
75,435	2012
80,525	2013
90,391	2014
86,097	2015
76,059	2016

Tablas 12 y 13: Valores base contaminación y consultas atendidas/día

Con los valores base, se utiliza la metodología de trasferencia de beneficios para calcular los costos de los escenarios de alerta, en los que ha incurrido la ciudad. De esta manera, se utilizan los datos de inflación reportados por el banco de la república, con el fin de hacer cálculos siempre en unidades monetarias equivalentes. Así, para cada día de alerta se calcula el aumento porcentual en la contaminación de acuerdo al promedio diario respectivo. Con el porcentaje de aumento en el material particulado, se calcula el aumento en los casos respiratorios atenidos, utilizando la relación encontrada por Hernández (2017). Finalmente,

estos nuevos casos son multiplicados por el costo promedio de atención, dando como resultado los costos aproximados que el PM_{10} trae consigo en términos de salud para la población.

Costos en Pesos de los Episodios de Alerta									
Año	**Carvajal**	**Suba**	**CAR**	**Usaquen**	**Ferias**	**Guaymaral**	**PA**	**Kennedy**	**Total**
2009	$ 4.880.393.744	$ 961.976.965	$ -	$ -	$ -	$ -	$ 121.443.761	$ 2.769.556.954	$ 8.733.371.423
2010	$ 3.622.656.358	$ 19.950.808	$ -	$ -	$ 29.355.883	$ -	$ 209.225.928	$ 3.537.936.476	$ 7.419.125.452
2011	$ 2.804.585.259	$ -	$ -	$ -	$ -	$ -	$ 95.931.681	$ 1.275.360.236	$ 4.175.877.175
2012	$ 1.919.918.853	$ -	$ -	$ -	$ 36.477.470	$ -	$ 36.231.187	$ 1.495.656.726	$ 3.488.284.235
2013	$ 3.314.026.263	$ 83.209.074	$ -	$ -	$ -	$ -	$ -	$ 1.608.581.743	$ 5.005.817.079
2014	$ 5.289.507.209	$ 117.287.062	$ 150.710.646	$ 61.542.527	$ 20.749.512	$ 69.650.816	$ 1.374.734.070	$ 2.114.412.788	$ 9.198.594.629
2015	$ 7.774.334.805	$ -	$ -	$ -	$ -	$ -	$ 123.940.726	$ 1.905.069.932	$ 9.803.345.463
2016	$ 6.420.631.250	$ 96.792.225	$ -	$ 193.288.881	$ 91.442.133	$ -	$ 771.752.267	$ 2.456.784.127	$ 10.030.690.882
								Total	**$ 57.855.106.339**

Tabla 14: Costos Incurridos en escenarios de alerta por concepto de consultas atendidas

La tabla 14 muestra en unidades monetarias los costos que han incurrido las entidades prestadoras del servicio de salud, únicamente por concepto de consultas relacionadas a enfermedades respiratorias. Vale la pena aclarar que la tabla presenta todos los valores en pesos del 2016.

Se puede ver que, las estaciones de peor comportamiento, mencionadas anteriormente siendo Kennedy y Carvajal son las únicas en las cuales existe un costo en todos los años analizados. Por otra parte, estaciones como el Centro de Alto Rendimiento, Usaquén, Guaymaral presentan un número mínimo de años en los cuales se presentan escenarios de alerta y costos asociados. Esto nos da un indicio de la distribución de la contaminación, resaltando las estaciones más contaminadas y las que menos presentan días críticos.

En síntesis, los escenarios de alertas que se han presentado en la ciudad de Bogotá, acerca de los cuales no se ha tenido ningún sistema de prevención, han representado $57.885.106.339 COP en gastos a la ciudad y al sistema de salud. La existencia de medidas preventivas y reactivas en la RMCAB se presenta como una excelente herramienta en la disminución de costos y de casos perjudiciales en la salud de la población.

5. Conclusiones

La contaminación del aire en Bogotá, especialmente producto del PM_{10} se ha vuelto una creciente preocupación para las autoridades locales debido a los efectos adversos que tiene en materia de salud. El monitoreo de la contaminación es llevado a cabo de manera sectorizada por la secretaria de ambiente, mediante la red de monitores (RMCAB). Cada

estación de la red registra de manera horaria los niveles de contaminación, así como de variables atmosféricas como la temperatura y la precipitación. Esta sectorización, permite ver las grandes diferencias que existen en los comportamientos de cada una de las estaciones.

A pesar de que, en algunas estaciones, se cumplen las normas nacionales de contaminación diaria, existen puntos críticos en la ciudad donde lo normal es que se sobrepasen los límites recomendables. Estaciones como Carvajal y Kennedy, presentan las mayores contaminaciones en toda la ciudad llegando a niveles críticos por encima de 250 $\mu g/m^3$, siendo estos cuatro veces lo recomendado por la organización mundial de la salud (50 $\mu g/m^3$) y más del doble que lo establecido en la ley nacional (100 $\mu g/m^3$). Estas cifras son preocupantes considerando el daño ambiental que se produce y más importante aún la cantidad de gente que habita, trabaja o se desplaza por estas zonas de manera recurrente, ya que el estar expuestos a cantidades tan altas de contaminantes, puede desencadenar en varios problemas respiratorios como asma, EPOC o cáncer de pulmón.

Este trabajo encontró que las estaciones con mayor contaminación promedio, presentan de manera sistemática concentraciones de material particulado que, bajo un sistema preventivo, clasificarían como escenarios de alerta y emergencia. Del análisis de toda la ciudad se encontró que, de los escenarios de alerta, el 92% pertenecen a las estaciones de Kennedy y Carvajal en conjunto.

La identificación de este tipo de comportamientos, permite a los hacedores de política encontrar los puntos críticos en los cuales concentrar sus esfuerzos, de manera que en el corto plazo se vea un verdadero cambio en las condiciones ambientales de la ciudad.

Por otro lado, este trabajo presenta un primer paso en el desarrollo de sistema de pronóstico de material particulado, discriminado por cada una de las horas del día. Esta forma de abordar el problema permitió encontrar que existen grandes disparidades, no solo a lo largo de las estaciones de la ciudad, sino dentro de la misma estación a lo largo del día, en cuanto a las dinámicas de dispersión y acumulación de PM_{10}. Esto se ve representado por las diferencias existentes en las variables que resultan representativas en cada uno de los modelos realizados para la predicción de la cantidad de contaminación atmosférica en la estación. Así mismo, encontramos que las variables varían su impacto en la contaminación dependiendo de la hora que se esté analizando. La metodología de selección de variables resulto satisfactoria pues

no solo entrego modelos con un buen comportamiento, sino que a su vez son parsimoniosos y de fácil interpretación.

Al ser evaluados, los modelos presentaron ajustes satisfactorios con respecto a los datos originales y desempeños sobresalientes en cuento a las medidas de calidad utilizadas. Adicionalmente, al analizar su capacidad predictiva de los escenarios de interés, se presentó que en su mayoría reportan valores bastante buenos de sensibilidad, acertando en un 80% de los casos. Esto implica, que los modelos presentan una gran herramienta para el planteamiento de estrategias de prevención de la contaminación atmosférica, permitiendo el pronóstico anticipado de escenarios de alerta, lo cual permite a las autoridades adoptar medidas preventivas y reactivas que se enfoquen en la disminución del impacto ambiental y la mitigación del impacto en salubridad que las emisiones de PM tienen sobre la población.

En último lugar, se estimó de manera aproximada el costo de la ausencia de un sistema de alerta estructurado para la ciudad desde el año 2009 hasta el año 2016. Utilizando como base el artículo de Hernández (2017) se tomó una relación porcentual entre la contaminación y los casos atendidos por enfermedades respiratorias (10% vs 7.6%), así como un valor estimado por consulta de $841.8 USD, tomando como referencia el trabajo de Pérez, et. al (2007). Con estos valores, y tomando como casos bases las contaminaciones y consultas promedio diarias, se estimó un impacto económico de $57.885.106.339 COP (total acumulado de los últimos 7 años) que recaen sobre el sistema de salud.

La cuantificación en términos económicos, provee un marco de referencia sobre el cual actuar de manera que se desarrollen políticas encaminadas a la mitigación de los daños ambientales y sociales que causa el PM_{10} en la ciudad, proporcionando una aproximación de los ahorros que pueden tenerse al desarrollar e implementar un plan de medidas preventivas enfocadas a disminuir los impactos de la contaminación del aire, sobre el sistema de salud.

Este trabajo, se presenta como evidencia de la necesidad en la implementación de un sistema estructurado de acción frente a los distintos escenarios de contaminación que pueden darse en Bogotá, así como la difusión de la información de manera que permita tanto a las autoridades como a los habitantes tomar medidas efectivas antes de la exposición, disminuyendo los efectos nocivos y maximizando el bienestar que proporcionan.

Es también importante resaltar las limitaciones que posee el estudio. El más notorio es la disponibilidad de datos, los cuales en muchas ocasiones faltan o hacen referencia a cadenas de texto o datos no válidos. En este sentido, el trabajo aplico metodologías de limpieza de datos e imputación simple que produjeron resultados satisfactorios y un buen ajuste a los datos reales, proporcionando validez y robustez a el análisis posterior. Por otro lado, está la estimación de los costos, los cuales se utiliza una base del 2007 la cual sería apropiado que pudiese actualizarse para reflejar valores más cercanos a los costos de atención actuales.

Como trabajo futuro, se propone la implementación de la metodología propuesta en esta memoria, para los años más recientes 2017 y 2018, así como la incorporación de la información completa de los sondeos atmosféricos que se realizan de manera diaria en el aeropuerto el Dorado, pues las variables de las que se toma medida, pueden resultar relevantes a la hora de pronosticar los niveles de material particulado. Así mismo resulta relevante el análisis de distintas formas de pronostico con el fin de hacer comparaciones y recomendar la opción más adecuada para cada una de las horas, de cada una de las estaciones de la red. Por último, se recomienda incrementar el alcance del estudio de costos, desde la estimación de costos directos e indirectos causados como resultado de la exposición al aire, hasta el cálculo estimativo de los costos totales de los últimos años.

Referencias

Alcaldía Mayor de Bogotá D.C. (2015). Resolucion 2410.

Arciniegas, C. (2011). Diagnóstico y control de material particulado: partículas suspendidas totales y fracción respirable pm10. *Luna Azul*, 195-213.

Area Metropolitana Valle de Aburrá. (28 de Noviembre de 2016). Acuerdo Metropolitano N 15. Medellín.

Cortes, S. (2010). Analisis del comportamiento de la concentración de material particulado menor a 10 micras en la localidad de puente aranda a partir de un modelo de regresión dinámica. Bogotá.

Cozzi, L., & et.al. (2016). Energy and Air Pollution.

Enders, W. (2014). *Applied Econometric Time Series.* Wiley.

Hernández, J. (2017). Valor económico de la calidad del aire en la localidad de kennedy medido desde la salud de los habitantes. Bogotá: Uniandes.

Karatzas, K., Papadopoulos, A., & Slini, T. (2002). Regression Analysis and Urban Air Quality Forecasting: An Application for the city of Athens. *Global Nest*, 153-162.

Mejia, N., & Montes, M. (2017). Application of Statistical Modeling Techniques for PM10 levels Forecast in Bogota. Bogotá.

Mészáros, E. (1999). Fundamentals of Atmospheric Aerosol Chemistry. *Akadémiai Kiado*.

Parker, J. A. (2018). Distributed- Lag Models. En J. A. Parker, *Theory and Practice of Econometrics* (págs. 35-54).

Pérez, N., Murillo, R., Pinzon, C., & Hernández, G. (2007). Costos de la Atención Médica del Cáncer de Pulmón, la EPOC y el IAM atribuibles al consumo de tabaco en Colombia. *Revista Colombia Cancerol*.

Pope. (2004). Air Pollution and Health - Good News and Bad. *New England Journal of medicine*, 1132-1134.

Pretis, F., Reade, J., & Sucarrat, G. (10 de Marzo de 2016). General-to-Specific Modelling and Indicator Saturation with the R Package gets.

Quiñones, L. (2015). Predicción de contaminación por PM10 en las estaciones de Kennedy y Carvajal . Bogotá.

Secretaria Distrital del Ambiente. (2016). Datos e Indicadores para medir la calidad del ambiente en Bogotá. Bogotá D.C, Colombia.

Siwek, K., & Osowski, S. (2016). Data Minig Methods for Prediction of Air Pollution. *International Journal of Applied Math and Computer Sciences*, 467-478.

Wei, W. (2005). *Time Series Analysis, Univariate and Multivariate Methods.* Addison-Wesley.

Westerlund, J., Urbain, J., & Bonilla, J. (2014). Application of air quality combination forecasting to Bogota. *Atmospheric Enviroment*, 22-28.

World Health Organization. (2016). Fact Sheet Ambient Outdoor Air Quality and Health.