Revista EIA

UNIVERSIDAD EIA
Ser, Saber y Servir

Análisis de características tiempo-frecuencia para la predicción de series temporales de Material Particulado usando Regresión por Vectores de Soporte y Optimización por Enjambre de Partículas

Norbey Yovany Alzate Zuluaga[1]
Juan Pablo Sepúlveda Suescún[1]
Juan Pablo Murillo Escobar[1]
Diana Alexandra Orrego Metaute[1]
Mauricio Andrés Correa Ochoa[2]

1. Instituto Tecnológico Metropolitano ITM
2. Universidad de Antioquia

Revista EIA, ISSN 1794-1237 /
e-ISSN 2463-0950
Año XVII/ Volumen 17/ Edición N.34
Julio-Diciembre de 2020
Reia34015 pág 1-15

Publicación científica semestral
Universidad EIA, Envigado, Colombia

Para citar este artículo / To reference this article /
Alzate Zuluaga, N.Y.; Sepúlveda Suescún, J.P.; Murillo Escobar, J.P.; Orrego Metaute, D.A.,; Correa Ochoa, M.A. (2020). Análisis de características tiempo-frecuencia para la predicción de series temporales de Material Particulado usando Regresión por Vectores de Soporte y Optimización por Enjambre de Partículas. Revista EIA, 17(34), Julio-Diciembre, Reia34015. https://doi.org/10.24050/reia.v17i34.1347

Autor de correspondencia:

Sepúlveda Suescún, J.P. (Juan Pablo): Instituto Tecnológico Metropolitano ITM, Calle 73 No. 76A - 354, Medellín, Colombia.Teléfono: 3105231576. Correo electrónico: juansepulveda146730@correo.itm.edu.co

Recibido: 25-07-2019
Aceptado: 18-06-2020
Disponible online: 25-10-2020

Resumen

La contaminación atmosférica por Material Particulado (PM) es un problema claramente reconocido a nivel mundial como uno de los factores de riesgo más importantes para la salud humana, en los últimos años han surgido diferentes modelos basados en inteligencia artificial para predecir la concentración de PM, con el fin de generar sistemas de alerta temprana que eviten la exposición de las personas. En este trabajo, se analizó un esquema de caracterización en el dominio tiempo-frecuencia usando la transformada *Wavelet* para la predicción de series temporales de PM_{10} y $PM_{2.5}$ usando un algoritmo de Regresión por Vectores de Soporte optimizado por Enjambre de Partículas (SVR-PSO), además, se evaluó el efecto de la imputación de datos sobre las estimaciones. Los resultados obtenidos mostraron que, empleando características temporales, más las características tiempo-frecuencia propuestas, se obtiene el mejor desempeño de la SVR-PSO, además se encontró que el uso de la imputación de datos no afecta el desempeño de la SVR-PSO. El sistema propuesto en este trabajo permite disminuir el error de las estimaciones de concentración de PM_{10} y $PM_{2.5}$ haciendo uso de características tiempo-frecuencia y es capaz de operar de forma robusta contra datos perdidos, aumentando su viabilidad de ser implementado en escenarios reales.

Palabras Clave: SVR, PSO, Transformada Wavelet, Imputación de datos, Predicción, Regresión.

Time-Frequency characteristics analysis for forecasting time series of particulate matter using Support Vector Regression and Particle Swarm Optimization

Abstract

Atmospheric pollution by particulate matter is a problem recognized worldwide as a major risk factor for human health, over last years different models based on artificial intelligence has been proposed to forecast particulate matter concentration with the purpose of generate early warning systems that avoid people exposition. This paper analyzed a characterization scheme in time-frequency domain using the Wavelet to predict time series of PM_{10} and $PM_{2.5}$ using the Support Vector Regression optimized with Particle Swarm Optimization (SVR-PSO). This paper also evaluated the effect of data imputation over estimations. Results showed that using time characteristics along with time-frequency characteristics SVR-PSO reach its best performance, also, it was found that use of data imputation does not affect SVR-PSO performance. The system proposed in this paper allow to estimate PM_{10} and $PM_{2.5}$ concentrations with less error through time-frequency characteristics, in addition, it is capable to operate robustly against missing data, which improve its viability to be implemented in real scenarios.

***Keywords*:** SVR, PSO, Wavelet Transform, Data imputation, Prediction, Regression.

1. Introducción

La contaminación del aire hoy en día es un problema de salud pública reconocido a nivel mundial, según la Organización Mundial de la Salud (OMS), se estima que 1 de cada 8 muertes en todo el mundo son atribuibles a la contaminación del aire (Brugha, Edmondson and Davies, 2018; Khaniabadi *et al.*, 2018). Aunque se sabe que la contaminación del aire es dañina para los pulmones y las vías respiratorias, también puede alterar otros órganos del cuerpo. Se calcula que alrededor de 500,000 muertes por cáncer de pulmón y 1,6 millones de muertes por enfermedad pulmonar obstructiva crónica (EPOC) pueden atribuirse a la contaminación del aire, pero está también puede representar el 19% de todas las muertes cardiovasculares y el 21% de todas las muertes por accidente cerebrovascular (Delpont *et al.*, 2018; Schraufnagel *et al.*, 2018). En Colombia, según el Departamento Nacional de Planeación, en el 2015 se reportaron 10.527 fallecimientos relacionados con esta problemática (Betancur Alarcon, 2017).

El Valle de Aburrá es una subregión ubicada en el centro-sur del departamento de Antioquia constituido por 10 municipios, dentro de los cuales se encuentra la ciudad de Medellín. El panorama de contaminación dentro del Valle de Aburrá es aún más complejo y peligroso, debido a que es un valle estrecho rodeado por una cadena de montañas, lo cual provoca que el régimen de vientos sólo permita el desplazamiento de los contaminantes, más no su dispersión hacia las capas más altas de la atmósfera (Muñoz, Quiroz and Paz, 2006). Otro factor que impide la dispersión de estos es la baja altura de las nubes, ya que el aire y los contaminantes no ascienden lo suficiente, desencadenando el aumento de las concentraciones de los mismos (Siata, 2017). Lo anterior no sólo representa una problemática sanitaria, sino también una problemática

Revista EIA https://doi.org/10.24050/reia.v17i34.1347

económica, al generarse gastos en el sistema de salud equivalentes al 5% del producto interno bruto (PIB) de Medellín, la capital de Antioquia (Betancur Alarcon, 2017).

La contaminación del aire tiene una relación directa con las actividades cotidianas de las personas, como el uso de automotores y el trabajo en industrias, por lo que ejercer control en estas tareas es la mejor forma de reducir los picos de contaminación y por ende la calidad de vida de los habitantes. En (Baklanov *et al.*, 2007), el desarrollo de herramientas para predecir la calidad de aire como mínimo con 24 o 48 horas de anticipación constituyen una herramienta fundamental para desarrollar planes de contingencia, tanto a nivel industrial como vehicular, logrando evitar el aumento de picos de contaminación y la exposición de la población a estos (Shahraiyni and Sodoudi, 2016).

Actualmente, diferentes investigaciones en aprendizaje de máquina han logrado predecir las concentraciones de los contaminantes del aire utilizando una variedad de técnicas de predicción, en (De Gennaro *et al.*, 2013; Prasad, Gorai and Goyal, 2016; García Nieto *et al.*, 2017) utilizaron Redes Neuronales Artificiales (ANN), Regresión por Vectores de Soporte (SVR) y Sistemas Adaptativos de Inferencia Neurodifusa (ANFIS), sin embargo, aunque usaron las técnicas más utilizadas recientemente, el desempeño predictivo reportado sigue siendo bajo (Zhang *et al.*, 2012; Bai *et al.*, 2018). Además, en (Sun *et al.*, 2013; Donnelly, Misstear and Broderick, 2015) usaron características temporales simples, como mínimos, máximos y promedios de variables meteorológicas, así como de la concentración de contaminantes, las cuales tienen un buen desempeño reportado para predecir concentraciones de NO_2, pero no tan sobresaliente para PM. A razón de esto, se han desarrollado diferentes estrategias de generación de espacios de características. En (De Gennaro *et al.*, 2013) se logró predecir la concentración del Material Particulado utilizando características como: velocidad y dirección del viento, lluvias, masas aéreas, temperatura y la concentración de Material Particulado, se determinó la concentración del PM en un promedio entre 8 y 24 horas del día utilizando ANN, esto se realizó con el fin de observar los cambios en la predicción diaria de acuerdo al tipo de tráfico, si era alto o si era nulo, los resultados de la predicción fueron reportados como altos, con un R^2 de 0,80, pero este valor corresponde a predecir las concentraciones promedio de PM_{10} en 24 horas y al ser un valor promedio, no refleja si el sistema es capaz de llegar a los puntos críticos de contaminación. En (Donnelly, Misstear and Broderick, 2015) para predecir NO_2 utilizaron Regresión Lineal Múltiple cuya entrada tiene en cuenta los datos horarios de temperatura, humedad relativa, presión atmosférica, concentración de NO_2, promedio y máximos diarios de concentración de NO_2 y O_3, además de otros factores temporales como el año, el día y la hora a la que corresponde cada medición. De igual forma en la literatura se reporta la exploración de otros espacios de características como lo hicieron en (Chen *et al.*, 2013), en donde se utilizó la descomposición *Wavelet* para convertir señales no estacionarias en estacionarias y regulares, logrando un RMSE de 15,06 al predecir PM_{10} con 12 horas de anticipación. En (Feng *et al.*, 2015), utilizaron la transformada *Wavelet* para descomponer la serie temporal concentración de $PM_{2.5}$ en sub-series con menor variabilidad en un modelo de Redes Neuronales Artificiales, con el fin de predecir la concentración a los dos días siguientes, además de esto, utilizaron las características temporales descritas anteriormente en (Donnelly, Misstear and Broderick, 2015), logrando un RMSE de 21,67.

En este artículo se presenta una novedosa forma de caracterización de series temporales de diferentes variables meteorológicas, como la velocidad y dirección del viento, la temperatura del aire, la humedad relativa y la concentración de contaminantes del aire NO, NO_2, PM_{10}, $PM_{2.5}$ y O_3, utilizando la transformada *Wavelet* como técnica principal de la caracterización para desarrollar un sistema de predicción de la concentración de PM en el Valle de Aburrá utilizando la Regresión por Vectores de

Soporte optimizada con Enjambre de Partículas (*Support Vector Regresion-Particle Swarm Optimization* SVR-PSO), adicionalmente, se evaluó el efecto de la imputación de datos faltantes en problemas de regresión utilizando una metodología con base en vecinos más cercanos (*k-Nearest Neighbors k*-NN).

2. Materiales y Métodos

2.1. k Vecinos más cercanos(*k*-NN)

El método de los *k* vecinos más cercanos (*k-Nearest Neighbors k*-NN) es uno de los métodos más utilizados para problemas de clasificación, clustering y regresión (Ertu\ugrul and Ta\ugluk, 2017), la regla de los *k* vecinos más cercanos indica que la clase asignada a un nuevo caso será la clase más votada entre sus *k* vecinos más próximos del conjunto de entrenamiento, la letra *k* indica el número de vecinos a utilizar y cuando se utilizan varios vecinos se aprovecha de forma más eficiente la información que se puede extraer del conjunto de entrenamiento (Gallego *et al.*, 2018).

Debido a que los procesos de generación de datos a menudo producen patrones repetidos de comportamiento es posible aplicar el *k*-NN para regresión (*k*-NNR), debido a que estos patrones pueden proporcionar información valiosa para predecir el comportamiento de los datos en el futuro (LINDSAY and NORMAN, 1977). El principio de *k*-NNR consiste estimar la respuesta del punto de prueba t como un promedio ponderado de las respuestas de los puntos *k* de entrenamiento más cercanos, x_1, x_2,..., x_k, en el vecindario de x_t (Hu *et al.*, 2014).

2.2. Transformada Wavelet

La transformada *Wavelet* puede ser utilizada para analizar características en tiempo-frecuencia de cualquier tipo de señal o series temporales. A medida que la *Wavelet* madre (Función de ventana flexible) se mueve a través de la señal durante el proceso de Transformada *Wavelet* (*Wavelet Transform* WT), ésta genera varios coeficientes que representan la similitud entre la señal y la *Wavelet* madre (en cualquier escala específica) (Araghi *et al.*, 2015). El término madre da a entender que las funciones con diferentes regiones de actuación que se usan en el proceso de transformación provienen de una función principal (Martínez and Castro, 2002).

Hay dos tipos de transformada *Wavelet*: Continua y discreta. El uso de la transformada continua de *Wavelet* (*Continuos Wavelet Transform* CWT) puede generar un gran número de coeficientes, haciendo su uso e interpretación más complicado, la transformada discreta de *Wavelet* (*Discrete Wavelet Transform* DWT) simplifica el proceso de transformación al tiempo que proporciona un análisis muy efectivo y preciso (Partal and Küçük, 2006).

Para una serie de tiempo discreta x_i, donde *x* ocurre en un tiempo discreto *i*, Los coeficientes de DWT pueden ser calculados por la siguiente ecuación (Partal and Küçük, 2006).

$$W_{m,n} = \frac{1}{2^{m/2}} \sum_{t=0}^{N-1} x_i \, \varphi\left(\frac{i}{m^2} - n\right)$$

Donde *m* y *n* son enteros que controlan, respectivamente la dilatación *Wavelet* (escala) y la translación (tiempo), $W_{m,n}$ es el coeficiente *Wavelet* para la *Wavelet* discreta, $2^{m/2}$ es la ubicación de la DWT.

Aplicando la DWT a una señal, ésta se descompone en dos componentes llamados Coeficientes de Aproximación (*Aproximation Coeficients* ApCo) y Coeficientes de

Revista EIA https://doi.org/10.24050/reia.v17i34.1347

Detalle (*Detail Coeficients* DeCo). ApCo comprende el componente de la señal a gran escala y baja frecuencia, mientras que el DeCo representa el componente de alta frecuencia y pequeña escala. En general, ApCo muestra las características más importantes de la señal, especialmente en el caso del análisis de variaciones a largo plazo y lo que es más importante para estudios de análisis de tendencias. El proceso de descomposición puede utilizarse como un proceso iterativo, en el cual, ApCo de la primera descomposición se desglosa en nuevos ApCo y DeCo (Araghi *et al.*, 2015).

2.3. Regresión por Vectores de Soporte

La idea básica del algoritmo de Regresión por Vectores de Soporte (*Support Vector Regresion* SVR) es utilizar un mapeo no lineal para llevar los datos a un espacio de características de alta dimensión (Smola and Scholkopf, 2004). Para un conjunto de entrenamiento dado $X=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$,, donde y_i es el valor objetivo de un valor de entrada x_i, el objetivo de la SVR es encontrar una función $f(x_i)$ que tenga como máximo una desviación ε de los valores objetivos reales de y_i y al mismo tiempo sea lo más plana posible. Lo que permite aceptar datos que tengan errores menores que ε y rechazar lo que tengan un error mayor (Smola and Scholkopf, 2004). El caso de una función lineal *f* puede describirse como:

$$f(x) = w^T x + b$$

Donde, *w* es un vector de coeficientes de pesos y *b* es un termino de ajuste, ambos pueden ser encontrados resolviendo el problema de optimización (Smola and Scholkopf, 2004):

$$\min \frac{1}{2}\|w\|$$

$$sujeto\ a \begin{cases} y_i - (w^T x_i + b) \leq \varepsilon \\ (w^T x_i + b) - y_i \leq \varepsilon \end{cases}$$

El problema de optimización convexa presentado anteriormente sólo es viable cuando se tiene un *f* existente que aproxime todos los pares (x_i, y_i) con ε de precisión. Encontrar un conjunto de datos que permitan esto es realmente difícil, por esto se permiten algunos errores, al agregar variables de holgura, ξ_i, ξ_i^* haciendo que el problema de optimización sea replanteado como (Kazem *et al.*, 2013):

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

$$sujeto\ a \begin{cases} y_i - (w^T x_i + b) \leq \varepsilon + \xi_i \\ (w^T x_i + b) - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases}$$

Donde $C > 0$ es la constante de costo, cuya función es encontrar un punto de equilibrio entre la complejidad del modelo y el valor más alto de (Smola and Scholkopf, 2004). Usando la formulación dual a través de multiplicadores de Lagrange, la SVR puede extenderse para funciones no lineales, en este sentido el problema de optimización se define sólo en términos del multiplicador de Lagrange α_i, α_i^*. Esto es posible porque la función Kernel $\varphi(x_i, x_j)$ devuelve el producto punto entre los pares en un espacio dimensional de orden superior, sin un mapa explícito de los datos (Kazem *et al.*, 2013).

$$\max -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)\varphi(x_i,x_j)\ -\varepsilon\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)+\sum_{i=1}^{l}y_i(\alpha_i-\alpha_i^*)$$

$$sujeto\ a\ \begin{cases}\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)=0\\ \alpha_i,\alpha_i^*\ \epsilon\ [0,C]\end{cases}$$

La función de predicción se puede formular en términos de los multiplicadores de Lagrange y la función del kernel como:

$$f(x)=\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)\varphi(x_i,x)+b$$

En este artículo, se emplea el Kernel de Base Radial (*Radial Basis Kernel* RBK), ya que el estado de la técnica muestra buenos resultados a lo largo de una gran variedad de aplicaciones, la función RBK es $\varphi(x_i, x_j)=e^{-\gamma\|x_i-x_j\|^2}$, donde γ es un parámetro libre que controla la amplitud de la función.

2.4. Optimización por Enjambre de Partículas

Optimización por Enjambre de Partículas (*Particle Swarm Optimization* PSO), es una técnica de optimización metaheurística inspirada en el comportamiento colectivo de los animales sociales (Marini and Walczak, 2015). Muchas partículas se distribuyen aleatoriamente en el espacio de búsqueda de dimensión *n*. El estado actual de cada partícula se describe con posición y velocidad. La partícula actualiza constantemente su posición y velocidad de acuerdo con dos indicadores en cada proceso de iteración. Un indicador es la solución óptima individual de partículas y la otra es la solución óptima global. Esas dos soluciones óptimas se actualizan nuevamente después de cada proceso de iteración. Todas las partículas se agregan a la solución óptima global. Cuando la solución óptima global tiende a ser estable finalmente, el resultado es la solución óptima de la partícula en el espacio variable dimensional N (Hu, Dong and Yu, 2016).

La posición x_i y velocidad v_i de cada partícula es actualizada de acuerdo con las siguientes fórmulas respectivamente:

$$\boldsymbol{x}_i^{t+1}=\boldsymbol{x}_i^t+\boldsymbol{v}_i^{t+1}$$

$$\boldsymbol{v}_i^{t+1}=\ \boldsymbol{v}_i^t+\alpha\epsilon_1\cdot[\boldsymbol{g}^*-\boldsymbol{x}_i^t]+\ \beta\epsilon_2\cdot[\boldsymbol{x}_i^*-\boldsymbol{x}_i^t]$$

Donde ϵ_1 y ϵ_2 son vectores aleatorios, con valores entre 0 y 1. Los términos α y β son parámetros de aprendizaje, donde ambos pueden tomar valores iguales a 2, finalmente $x_i^{*(t)}$ es la mejor partícula actual *i* y $\boldsymbol{g}^* \approx \min f(x_i)(\forall i=1,2,\cdots,n)$ es el mejor global actual.

2.5. Base de datos

Para este estudio, se utilizaron los datos suministrados por el Sistema de Alerta Temprana de Medellín y el Valle de Aburrá (SIATA). Cada estación realiza mediciones horarias de diferentes variables meteorológicas (velocidad y dirección del viento, temperatura y humedad del aire y radiación solar) y contaminantes del aire (PM_{10}, $PM_{2.5}$, NO, NO_2 y O_3). Los datos usados comprenden las mediciones obtenidas entre el 1 de enero de 2013 y el 31 de diciembre del 2016 de dos estaciones de monitoreo ITA-CJUS (Itagüí – Casa de Justicia) y BEL-USBV (Bello – Universidad de San Buenaventura). Se seleccionaron estas dos estaciones, ya que estas realizan mediciones de los contaminantes PM_{10} y $PM_{2.5}$ con menor cantidad de datos faltantes, permitiendo

Revista EIA https://doi.org/10.24050/reia.v17i34.1347

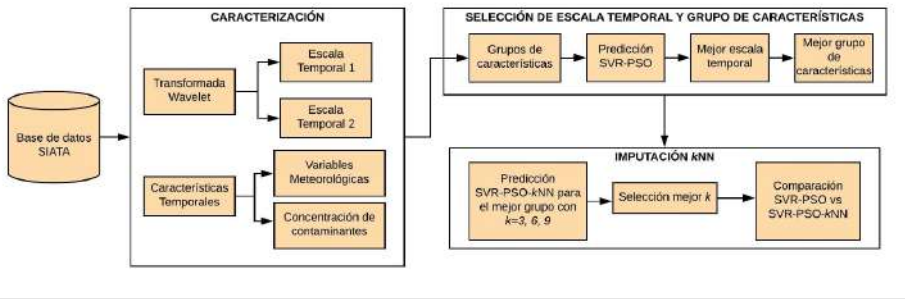
obtener espacios de representación compuestos por 35040 filas para ITA-CJUS y BEL-USBV por 9 columnas para ITA-CJUS y 10 columnas para BEL-USBV.

3. Metodología Propuesta

3.1. Caracterización

Con el fin de obtener un espacio de representación para realizar una predicción con 24 horas de anticipación de la concentración de PM_{10} y $PM_{2.5}$ utilizando una SVR-PSO (**Figura 1**), se llevó a cabo una caracterización en tiempo y tiempo-frecuencia de la concentración de los contaminantes y de las variables meteorológicas medidas por el SIATA, la caracterización en tiempo, ha sido con base al trabajo desarrollado previamente en (Murillo-Escobar *et al.*, 2019).

Figura 1. Metodología Propuesta



Las características temporales utilizadas se dividieron en dos grupos, estos son: promedios, máximos y mínimos de las últimas 24 horas de las variables meteorológicas y de las concentraciones de contaminantes, este grupo de características fue denominado Características Temporales (CT).

Con el objetivo de convertir series temporales de la concentración de PM_{10} y $PM_{2.5}$ en señales estacionarias y regulares con menores fluctuaciones, se utilizó la transformada *Wavelet* estacionaria, lo que permite conservar la misma cantidad de puntos iniciales. Las *Wavelet* madres más comúnmente utilizadas son Daubechies (db) y Morlet (Chen *et al.*, 2013; Kalteh, 2015) , para este artículo se utilizó Daubechies 5 con 4 niveles de descomposición. Debido a que este tipo de caracterización requiere un mínimo de muestras, antes de aplicar la transformada *Wavelet* se realizó una interpolación con la función Spline Cubic para obtener una mayor resolución temporal y así dar más argumentos a la transformada *Wavelet*.

Se utilizaron diferentes escalas temporales para caracterizar la señal de Material Particulado y obtener el mejor espacio de representación, información necesaria para poder predecir la concentración de este en determinado momento del día. En este contexto, se analizaron dos escalas temporales, la primera permite el análisis de la tendencia previa del contaminante durante más tiempo, con el fin de obtener la dinámica subyacente de este, mientras la segunda, pretende obtener información en una menor escala de tiempo, con el fin de detectar la influencia de posibles fenómenos climáticos o sociales.

Escala temporal 1 (E1)

Esta escala temporal reúne información sobre el comportamiento del contaminante durante las últimas 7 semanas en la hora y día que se realizará la estimación, de este modo teniendo como objetivo predecir la concentración de PM_{10} o $PM_{2.5}$ a la hora *n* (e.g. lunes 14:00), se toma el valor de la concentración del contaminante 30

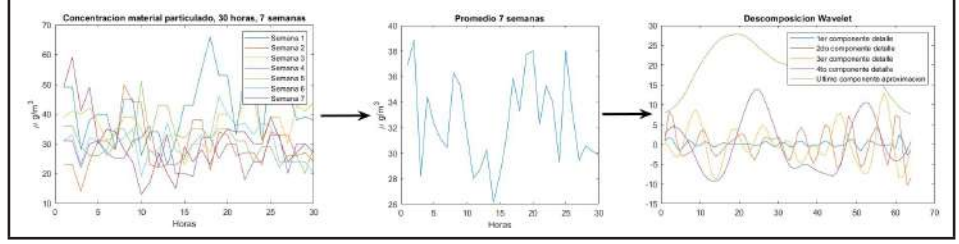
horas previas de la semana anterior *n*-30 (e.g. domingo 8:00), además, se obtiene una muestra de las 30 horas previas a la hora objetivo durante las 7 semanas anteriores, posterior a ello, se deriva una serie de tiempo promedio a partir de las 7 series temporales, para luego aplicar la TW con 4 niveles de descomposición. Finalmente se calcula la energía y la media de cada uno de los DeCo y del ultimo ApCo de cada una de las bandas, estas características fueron almacenadas en dos grupos denominados Energías Wavelet (WaEn) y Medias Wavelet (WaMe) (**Figura 2**).

Escala temporal 2 (E2)

Esta escala temporal pretende obtener información más reciente sobre el comportamiento del contaminante, por ello sólo se utilizaron las 30 horas previas de la semana anterior a la hora objetivo en la que se realiza la predicción, de esta escala se obtuvieron de igual manera la energía y media de los coeficientes *Wavelet* después de 4 niveles de descomposición.

Adicionalmente, se creó un tercer grupo de características, la cual consistió en combinar las variables obtenidas en la Escala 1 y en la Escala 2 (E1+E2).

Figura 2. Caracterización y Escala Temporal 1



3.2. Selección de características y escala temporal

Con el fin de determinar la escala temporal y el grupo de características que ofrece las mejores capacidades predictivas del Material Particulado, se realizaron todas las combinaciones posibles de los tres grupos de características propuestos (CT, WaEn y WaMe), dando como resultado un total de cuatro combinaciones.

En la construcción del modelo predictivo de PM_{10} y $PM_{2.5}$ con un horizonte predictivo de 24 horas utilizando SVR-PSO se tomó como conjunto de entrenamiento los datos del año 2013 los cuales fueron divididos aleatoriamente en 70/30 y se realizó la validación en los años 2014, 2015 y 2016. El PSO buscó maximizar el R^2 y como parámetros se determinaron 100 iteraciones y un total de 20 partículas.

Para determinar la mejor escala temporal se aplica inicialmente el test de normalidad Shapiro Wilk, posterior a ello se aplica el test Friedman para determinar si existe una diferencia estadísticamente significativa al realizar la predicción del Material Particulado al usar cada una de las escalas temporales propuestas. Luego de encontrar en cuál subconjunto de las escalas temporales se logra el mejor desempeño se procede a realizar la misma estrategia, pero con el fin de determinar cuál de las cuatro combinaciones de características presenta el mejor desempeño.

3.3. Imputación

Las bases de datos que se utilizaron contienen una gran cantidad de mediciones ausentes debido a que las diferentes estaciones de monitoreo pueden sufrir daños o ser sometidas a mantenimiento y calibración, desencadenando que sean registrados datos faltantes en las bases de datos, afectando el desempeño del sistema de predicción (Chen *et al.*, 2015), a la vez que limita su aplicación en tiempo real.

Revista EIA https://doi.org/10.24050/reia.v17i34.1347

Por esta razón se utilizó una imputación de datos faltantes con base a la técnica de los *k*-NNR (Ahmat Zainuri, Aziz Jemain and Muda, 2015; Zhang *et al.*, 2017), con el fin de evaluar el efecto de los nuevos valores generados utilizando *k*-NNR al momento de la predicción, cabe resaltar que sólo se imputaron valores perdidos en las muestras que no superaban 5 variables ausentes, con el fin de no agregar demasiada incertidumbre al sistema. Se evaluó el desempeño usando la mejor escala temporal y el mejor grupo de características para diferentes valores de *k* (*k*=3, 6 y 9). Con el fin de encontrar el mejor *k* para la imputación se utilizó el test de Friedman y un test de comparaciones múltiples. Finalmente, se realiza la comparación entre SVR-PSO con y sin imputación por medio del test de Mann-Whitney.

4. Resultados y Discusión

4.1. Caracterización

Luego de la descomposición *Wavelet* estacionaria de 4 niveles, se obtuvieron 4 componentes de detalle y de aproximación, de los cuales se generaron 2 grupos de características provenientes de dicha descomposición, la energía de los coeficientes de detalle, el último de aproximación (WaEn) y las medias de las mismas(WaMe), obteniendo así un total de 10 características en la matriz proveniente de la caracterización *Wavelet*, por otro lado, se obtuvieron las características temporales hora a hora, en donde se agruparon la concentración de contaminantes (PM_{10}, NO, NO_2 y O_3 para BEL-USBV y $PM_{2.5}$, NO y NO_2 para ITA-CJUS) y las características meteorológicas (temperatura, dirección y velocidad del viento, humedad del aire y radiación solar) de ambas estaciones.

4.2. Selección de características y escala temporal

Para seleccionar el mejor grupo de características, así como la mejor escala temporal, se realizó la predicción de $PM_{2.5}$ y PM_{10} durante el periodo 2014-2016 usando SVR-PSO, donde obtuvo el RMSE por mes.

En la **Figura 3** y **Figura 4** se pueden observar los errores de predicción de PM10 en BEL-USBV y PM2.5 en ITA-CJUS respectivamente usando SVR-PSO a partir de los 3 grupos de características con las diferentes escalas temporales. El test Shapiro Wilk arrojó un P-Valor < 0,05, indicando que los datos se comportan de manera no normal.

Figura 3. Desempeños predicción SVR-PSO BEL-USBV

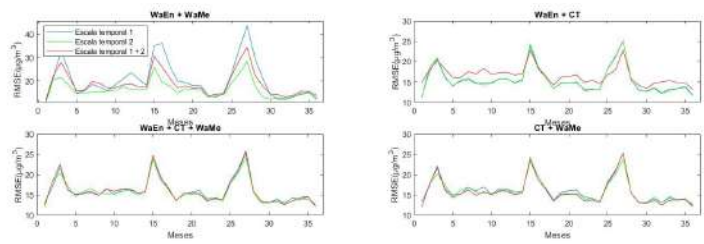
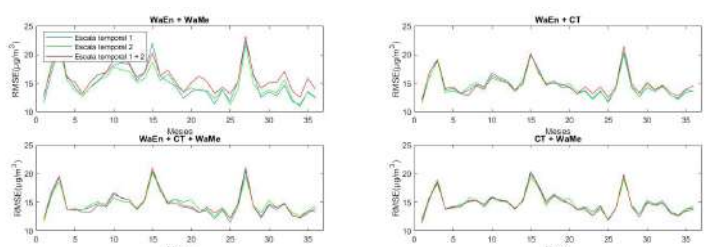


Figura 4. Desempeños predicción SVR-PSO ITA-CJUS



Posteriormente, el test no paramétrico de Friedman arrojó un P-Valor < 0,05, comprobando que existe una diferencia estadísticamente significativa entre las escalas temporales analizadas E1, E2, E1+E2, tanto en BEL-USBV como ITA-CJUS. El ranking de medias (**Tabla 1**) del test de Friedman evidencia un mejor desempeño en E2, tanto para BEL-USBV como para ITA-CJUS, además, un test de comparaciones múltiples (**Tabla 2**) muestra que existe una diferencia estadística entre E2 y las demás escalas para BEL-USBV. En ITA-CJUS no existe una diferencia estadística entre E1 y E2, teniendo en cuenta que el ranking de medias para ITA-CJUS E2 presenta el menor error y para BEL-USBV E2 supera estadísticamente el desempeño de los demás, este fue seleccionado como la mejor alternativa de trabajo.

TABLA 1. RANKING DE MEDIAS ESCALAS TEMPORALES

Escara Temporal	BEL-USBV	ITA-CJUS
E1	2,1182	1,9583
E2	**1,6389**	**1,7778**
E1+E2	2,2431	2,2639

TABLA 2. COMPARACIONES MÚLTIPLES ESCALAS TEMPORALES

Comparación	BEL-USBV	ITA-CJUS
E1 vs E2	<0,05	0,2759
E1 vs E3	0,5386	<0,05
E2 vs E3	<0,05	<0,05

Una vez determinada la mejor escala temporal, definimos cuál de los 3 grupos de características es el que mejor desempeño brinda, obteniendo un P-Valor $<$ 0,05 en Friedman para BEL-USBV e ITA-CJUS. En la **Tabla 3** se observa el ranking de medias y destaca el grupo WaEn+CT con respecto a los demás. Para comprobar que WaEn+CT presenta una diferencia con respecto a los demás se realizó un test de comparaciones múltiples (**Tabla 4**) en donde se observa que en todas sus comparaciones se obtiene un P-Valor < 0,05 en BEL-USBV. Para ITA-CJUS, sólo se obtuvo significancia estadística respecto a CT+WaMe, no obstante, el ranking de medias en esta estación lo muestra como el mejor grupo de características.

TABLA 3. RANKING DE MEDIAS GRUPOS

Grupos	**BEL-USBV**	**ITA-CJUS**
WaEn+WaMe	3,11369	2,6111
WaEn+CT	**1,6389**	**1,83331**
WaEn+WaMe+CT	2,5278	2,5833
CT+WaMe	2,6944	2,9722

TABLA 4. COMPARACIONES MÚLTIPLES GRUPOS

WaEn+CT vs	**BEL-USBV**	**ITA-CJUS**
WaEn+WaMe	**<0,05**	0,0517
WaEn+WaMe+CT	**<0,05**	0,0655
CT+WaMe	**<0,05**	**<0,05**

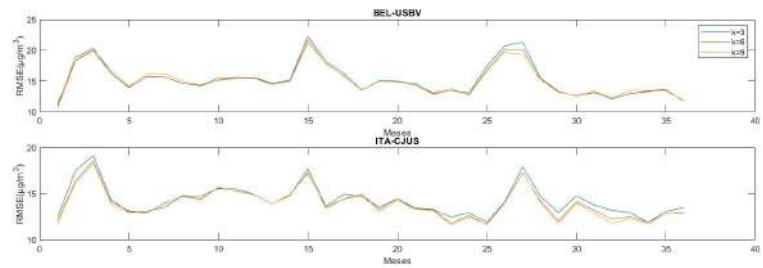
Revista EIA https://doi.org/10.24050/reia.v17i34.1347

Esta etapa arroja como resultado que la escala temporal E2 usando los grupos de características WaEn+CT ofrece las mejores capacidades predictivas. De la **Tabla 3** se puede notar que usar sólo las características *Wavelet* (WaEn+WaMe) genera el desempeño más bajo, no obstante, al combinar CT con los grupos de características *Wavelet* da lugar a una disminución significativa en el error, lo cual concuerda con (Feng *et al.*, 2015), en donde obtuvieron los mejores resultados al realizar un modelo que combina características temporales y características provenientes de la transformada *Wavelet* luego de descomponer la señal en 5 niveles y utilizar una red neuronal para cada señal de descomposición, lo cual supone un método altamente complejo y costoso a nivel computacional, ya que, a pesar de usar redes neuronales, no se define claramente un método de optimización como sí se hizo en este estudio. Además de esto, utilizaron características temporales con una baja resolución, ya que sólo se estimó la concentración promedio diaria, a diferencia del estudio actual que realiza la estimación hora a hora.

4.3. Imputación

Se realizó la predicción con SVR-PSO imputando la base de datos con *k*-NNR usando diferentes valores de *k* (*k*=3, 6 y 9), posteriormente se siguió con la estructura de predicción propuesta anteriormente utilizando el mejor grupo de características. Finalmente, al comparar el desempeño obtenido con los diferentes k (**Figura 5**) se obtuvo un P-valor < 0,05 en el test de Friedman para BEL-USBV e ITA-CJUS, lo cual indica que el valor de *k* sí afecta el rendimiento al momento de realizar la predicción, lo cual hace necesario utilizar un test de comparaciones múltiples para identificar el mejor *k* para cada una de las bases de datos.

Figura 5. Desempeño diferentes valores de k



En la **Tabla 5** se observa que para BEL-USBV el mejor k, en términos de desempeño es k =6, por el contrario, para ITA-CJUS, el mejor es *k* =9, teniendo en cuenta que no hay diferencia estadísticamente significativa entre *k*=6 y *k*=9 para BEL-USBV (**Tabla 6**), se concluye que *k*=9 es el más adecuado para imputar ambas bases de datos.

El porcentaje de datos perdidos antes de la imputación era de 8,78% para BEL-USBV y 7,22% para ITA-CJUS, luego de llevar a cabo la imputación con *k*=9 se redujo a 3,32% y 2,08% respectivamente.

TABLA 5. RANKING DE MEDIAS NÚMERO DE VECINOS (K)

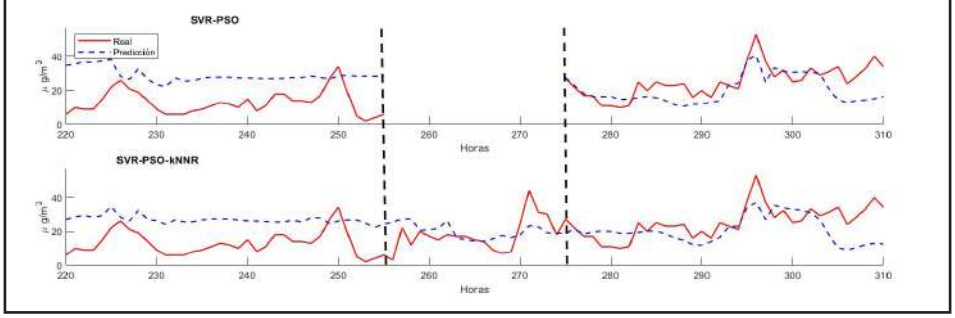
Comparación	BEL-USBV	ITA-CJUS
3	2,4167	2,7222
6	1,6111	1,9360
9	1,9722	1,3472

TABLA 6. COMPARACIONES MÚLTIPLES NÚMERO DE VECINOS (K)

Comparación	BEL-USBV	ITA-CJUS
3 vs 6	<0,05	<0,05
6 vs 9	0,1428	<0,05
3 vs 9	0,2759	<0,05

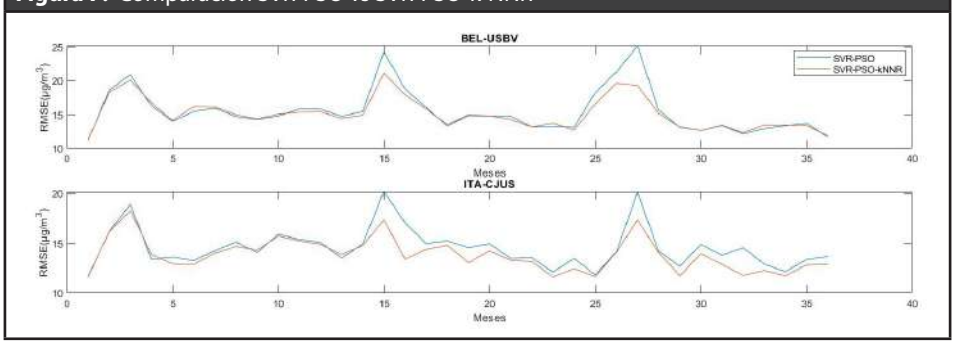
En la **Figura 6** se pueden observar los resultados obtenidos luego de la imputación, en donde se evidencia el espacio perdido que hay entre determinadas horas, además se observa que luego de realizar la imputación con *k*-NNR se logran recuperar estos datos perdidos, generando cierta confiabilidad ya que la figura muestra cómo estos datos siguen la tendencia de la señal y no muestran un comportamiento errático a simple vista, además, se pueden observar las líneas de predicción, las cuales continúan comportándose de manera similar luego de imputar el sistema, convirtiendo la imputación en un elemento de gran importancia porque permite que la predicción siga funcionando ante la presencia de datos perdidos, lo cual es de vital importancia en la implementación real de este tipo de metodologías.

Figura 6. Resultado de la imputación ITA-CJUS



En la **Figura 7** se observa la comparación entre los desempeños obtenidos luego de la predicción SVR-PSO y SVR-PSO-*k*-NNR. Se aplicó el test Mann-Whitney, el cuál arrojó un P-valor de 0,9416 para BEL-USBV y 0.0990 para ITA-CJUS, lo cual indica que no existe una diferencia estadísticamente significativa luego de realizar la imputación, lo que sugiere que este proceso no afectó el desempeño de la predicción como se constata en la **Tabla 7** en donde se muestra que los errores en ambos sistemas SVR-PSO y SVR-PSO-*k*-NNR se conservan.

Figura 7. Comparación SVR-PSO vs SVR-PSO-k-NNR



Revista EIA https://doi.org/10.24050/reia.v17i34.1347

TABLA 7. PROMEDIO MÉTRICAS DE DESEMPEÑO SVR-PSO

Métodos	Métricas	Estación	
		BEL-USBV	ITA-CJUS
SVR-PSO	RMSE	15,5373	14,577
	MAPE	35,373	56,0443
SVR-PSO-k-NNR	RMSE	15,1337	13,8408
	MAPE	34,8817	52,5728

Estos hallazgos son importantes ya que a diferencia de otros estudios, el trabajo actual permito predecir brechas de hasta cuatro a diferencia de otros trabajos como (Qin *et al.*, 2014) y (Feng *et al.*, 2015), Además el manejo valores no tuvo un impacto negativo sobre el desempeño del predictor como en (Shen, Huang and Yan, 2016) .

5. Conclusiones

En este documento se propone un sistema para pronosticar concentraciones de los contaminantes $PM_{2.5}$ y PM_{10}, con base en un enfoque de caracterización en el dominio de Tiempo-Frecuencia utilizando la transformada *Wavelet* en conjunto de una estrategia de imputación de datos utilizando *k*-NNR.

La predicción SVR-PSO utilizando solamente la caracterización *Wavelet* no ofrece buenos resultados predictivos, sin embargo, al combinar las características *Wavelet* con características temporales, los resultados mejoran notablemente, ofreciendo así, mejores capacidades predictivas.

El método propuesto a pesar de tener una estrategia de caracterización de una alta complejidad computacional presentó una velocidad de procesamiento razonable, lo que facilita su implementación en sistemas de operación en tiempo real. De igual forma el método propuesto fue capaz de operar con el mismo rendimiento al agregar los nuevos datos provenientes de la imputación, lo cual permite que el sistema de predicción tenga un funcionamiento a lo largo del tiempo sin importar que las estaciones de monitoreo estén en procesos de mantenimiento.

Referencias

Ahmat Zainuri, N., Aziz Jemain, A. and Muda, N. (2015) 'A Comparison of Various Imputation Methods for Missing Values in Air Quality Data (Perbandingan Pelbagai Kaedah Imputasi bagi Data Lenyap untuk Data Kualiti Udara)', *Sains Malaysiana*, 44(3), pp. 449–456. Available at: http://www.ukm.edu.my/jsm/pdf_files/SM-PDF-44-3-2015/17 NuryAzmin.pdf.

Araghi, A. *et al.* (2015) 'Using wavelet transforms to estimate surface temperature trends and dominant periodicities in Iran based on gridded reanalysis data', *Atmospheric Research*. Elsevier B.V., 155, pp. 52–72. doi: 10.1016/j.atmosres.2014.11.016.

Bai, L. *et al.* (2018) 'Air pollution forecasts: An overview', *International Journal of Environmental Research and Public Health*, 15(4), pp. 1–44. doi: 10.3390/ijerph15040780.

Baklanov, A. *et al.* (2007) 'Integrated systems for forecasting urban meteorology, air pollution and population exposure', *Atmospheric Chemistry and Physics*, 7(3), pp. 855–874. doi: 10.5194/acp-7-855-2007.

Betancur Alarcon, L. (2017) 'Atencion de males por calidad del aire cuesta 1,6 billones al año', *El Tiempo*, May.

Brugha, R., Edmondson, C. and Davies, J. C. (2018) 'Outdoor air pollution and cystic fibrosis', *Paediatric Respiratory Reviews*, 28, pp. 80–86. doi: https://doi.org/10.1016/j.prrv.2018.03.005.

Chen, M. *et al.* (2015) 'A clustering algorithm for sample data based on environmental pollution characteristics', *Atmospheric Environment*. Elsevier Ltd, 107, pp. 194–203. doi: 10.1016/j.atmosenv.2015.02.042.

Chen, Y. *et al.* (2013) 'Ensemble and enhanced PM10 concentration forecast model based on stepwise regression and wavelet analysis', *Atmospheric Environment*. Elsevier Ltd, 74, pp. 346–359. doi: 10.1016/j.atmosenv.2013.04.002.

Delpont, B. *et al.* (2018) 'Environmental Air Pollution: An Emerging Risk Factor for Stroke', in Vasan, R. S. and Sawyer, D. B. (eds) *Encyclopedia of Cardiovascular Research and Medicine*. Oxford: Elsevier, pp. 231–237. doi: https://doi.org/10.1016/B978-0-12-809657-4.99588-7.

Donnelly, A., Misstear, B. and Broderick, B. (2015) 'Real time air quality forecasting using integrated parametric and non-parametric regression techniques', *Atmospheric Environment*. Elsevier Ltd, 103(2), pp. 53–65. doi: 10.1016/j.atmosenv.2014.12.011.

Ertu\ugrul, Ö. F. and Ta\ugluk, M. E. (2017) 'A novel version of k nearest neighbor: Dependent nearest neighbor', *Applied Soft Computing Journal*, 55, pp. 480–490. doi: 10.1016/j.asoc.2017.02.020.

Feng, X. *et al.* (2015) 'Artificial neural networks forecasting of PM2.5 pollution using air mass trajectory based geographic model and wavelet transformation', *Atmospheric Environment*, 107, pp. 118–128. doi: 10.1016/j.atmosenv.2015.02.030.

Gallego, A. J. *et al.* (2018) 'Clustering-based k-nearest neighbor classification for large-scale data with neural codes representation', *Pattern Recognition*. Elsevier Ltd, 74, pp. 531–543. doi: 10.1016/j.patcog.2017.09.038.

García Nieto, P. J. *et al.* (2017) 'Air Quality Modeling Using the PSO-SVM-Based Approach, MLP Neural Network, and M5 Model Tree in the Metropolitan Area of Oviedo (Northern Spain)', *Environmental Modeling & Assessment*. doi: 10.1007/s10666-017-9578-y.

De Gennaro, G. *et al.* (2013) 'Neural network model for the prediction of PM10 daily concentrations in two sites in the Western Mediterranean', *Science of the Total Environment*. Elsevier B.V., 463–464, pp. 875–883. doi: 10.1016/j.scitotenv.2013.06.093.

Hu, C. *et al.* (2014) 'Data-driven method based on particle swarm optimization and k-nearest neighbor regression for estimating capacity of lithium-ion battery', *Applied Energy*. Elsevier Ltd, 129, pp. 49–55. doi: 10.1016/j.apenergy.2014.04.077.

Hu, X. P., Dong, X. D. and Yu, B. H. (2016) 'Method of Optimal Design with SVR-PSO for Ultrasonic Cutter Assembly', *Procedia CIRP*, 50, pp. 779–783. doi: 10.1016/j.procir.2016.04.180.

Kalteh, A. M. (2015) 'Wavelet Genetic Algorithm-Support Vector Regression (Wavelet GA-SVR) for Monthly Flow Forecasting', *Water Resources Management*, 29(4), pp. 1283–1293. doi: 10.1007/s11269-014-0873-y.

Kazem, A. *et al.* (2013) 'Support vector regression with chaos-based firefly algorithm for stock market price forecasting', in *Applied Soft Computing*. Elsevier B.V., pp. 947–958. doi: 10.1016/j.asoc.2012.09.024.

Khaniabadi, Y. O. *et al.* (2018) 'Mortality and morbidity due to ambient air pollution in Iran', *Clinical Epidemiology and Global Health*. doi: https://doi.org/10.1016/j.cegh.2018.06.006.

LINDSAY, P. H. and NORMAN, D. A. (1977) 'Neural information processing', *Human Information Processing*, 8226(November), pp. 190–254. doi: 10.1016/B978-0-12-450960-3.50010-5.

Marini, F. and Walczak, B. (2015) 'Particle swarm optimization (PSO). A tutorial', *Chemometrics and Intelligent Laboratory Systems*. Elsevier B.V., 149, pp. 153–165. doi: 10.1016/j.chemolab.2015.08.020.

Martínez, J. and Castro, R. (2002) 'Análisis de la teoría ondículas orientada a las aplicaciones en ingeniería eléctrica:Fundamentos', *E.T.D.I. Industriales Dpt. de ingeniería eléctrica*, p. 161.

Muñoz, A., Quiroz, C. and Paz, J. (2006) *Efectos de la contaminación atmosférica sobre la salud en adultos*. Universidad de Antioquia.

Murillo-Escobar, J. *et al.* (2019) 'Forecasting concentrations of air pollutants using support vector regression improved with particle swarm optimization: Case study in Aburrá Valley, Colombia', *Urban Climate*. Elsevier, 29(March), p. 100473. doi: 10.1016/j.uclim.2019.100473.

Partal, T. and Küçük, M. (2006) 'Long-term trend analysis using discrete wavelet components of annual precipitations measurements in Marmara region (Turkey)', *Physics and Chemistry of the Earth*, 31(18), pp. 1189–1200. doi: 10.1016/j.pce.2006.04.043.

Prasad, K., Gorai, A. K. and Goyal, P. (2016) 'Development of ANFIS models for air quality forecasting and input optimization for reducing the computational cost and time', *Atmospheric Environment*. Elsevier Ltd, 128, pp. 246–262. doi: 10.1016/j.atmosenv.2016.01.007.

Qin, S. *et al.* (2014) 'Analysis and forecasting of the particulate matter (PM) concentration levels over four major cities of China using hybrid models', *Atmospheric Environment*. Elsevier Ltd, 98, pp. 665–675. doi: 10.1016/j.atmosenv.2014.09.046.

Schraufnagel, D. E. *et al.* (2018) 'Air Pollution and Noncommunicable Diseases: A Review by the Forum of International Respiratory Societies' Environmental Committee, Part 2: Air Pollution and Organ Systems', *Chest*. doi: https://doi.org/10.1016/j.chest.2018.10.041.

Shahraiyni, H. T. and Sodoudi, S. (2016) 'Statistical modeling approaches for pm10 prediction in urban areas; A review of 21st-century studies', *Atmosphere*, 7(2), pp. 10–13. doi: 10.3390/atmos7020015.

Shen, C. H., Huang, Y. and Yan, Y. N. (2016) 'An analysis of multifractal characteristics of API time series in Nanjing, China', *Physica A: Statistical Mechanics and its Applications*. Elsevier B.V., 451(June 2000), pp. 171–179. doi: 10.1016/j.physa.2016.01.061.

Siata (2017) *Estabilidad atmosférica en el Valle de Áburra*. Colombia.

Smola, a J. and Scholkopf, B. (2004) 'A tutorial on support vector regression', *Statistics and Computing*, 14(3), pp. 199–222. doi: Doi 10.1023/B:Stco.0000035301.49549.88.

Sun, W. *et al.* (2013) 'Prediction of 24-hour-average PM2.5 concentrations using a hidden Markov model with different emission distributions in Northern California', *Science of the Total Environment*. Elsevier B.V., 443, pp. 93–103. doi: 10.1016/j.scitotenv.2012.10.070.

Zhang, Y. *et al.* (2012) 'Real-time air quality forecasting, Part II: State of the science, current research needs, and future prospects', *Atmospheric Environment*. Elsevier Ltd, 60, pp. 656–676. doi: 10.1016/j.atmosenv.2012.02.041.

Zhang, Z. *et al.* (2017) 'Application of a novel hybrid method for spatiotemporal data imputation: A case study of the Minqin County groundwater level', *Journal of Hydrology*. Elsevier B.V., 553, pp. 384–397. doi: 10.1016/j.jhydrol.2017.07.053.