Institución Universitaria

Metodología de predicción de calidad del aire con base en Vecinos más Cercanos Vagamente Cuantificados optimizados por Enjambre de Partículas

Juan Pablo Murillo Escobar

Instituto Tecnológico Metropolitano
Facultad de Ingeniería
Medellín, Colombia
2017

Metodología de predicción con base en vecinos más cercanos vagamente cuantificados optimizados por enjambre de partículas con aplicación en calidad del aire

Juan Pablo Murillo Escobar

Tesis presentada como requisito parcial para optar al título de:
Magister en Automatización y Control Industrial

Director:
MSc. Diana Alexandra Orrego Metaute
Co-director:
MSc. Miguel Alberto Becerra Botero

Línea de Investigación:
Inteligencia computacional
Grupo de Investigación:
Grupo de Investigación e Innovación Biomédica

Instituto Tecnológico Metropolitano
Facultad de Ingenierías
Medellín, Colombia
2017

Resumen

El uso de los sistemas de predicción ha sido ampliamente explorado en el estudio de la polución del aire, debido a ser reconocido globalmente como uno de los mayores factores de riesgo para la salud humana. Vecinos más cercanos vagamente cuantificados (*Vaguely Quantified Nearest Neighbor*-VQNN) es un modelo de predicción con base en la teoría de conjuntos difusos aproximados, la cual permite modelar la ambigüedad e incertidumbre existente en los datos por medio de cuantificadores difusos. Por otro lado, los datos asociados al monitoreo de la calidad del aire se caracterizan por su alta no linealidad, así como por su susceptibilidad al ruido e incertidumbre. En este contexto, VQNN es un método óptimo para la predicción de concentración de contaminantes en aplicaciones a la calidad del aire, no obstante, una mala sintonización de los cuantificadores difusos ocasiona una alta sensibilidad al ruido con baja generalidad y aumenta la incertidumbre e imprecisión de las predicciones realizadas. Este trabajo tiene como objetivo, proponer una metodología de predicción utilizando Vecinos más Cercanos Vagamente Cuantificados optimizados por Enjambre de Partículas (VQNN-PSO) para aplicaciones en la estimación de la concentración de contaminantes del aire en el Valle de Aburrá. La metodología propuesta hace uso de una función objetivo basado en el Error Absoluto Porcentual Medio (MAPE) y utiliza un algoritmo de enjambre de partículas modificado para trabajar con las restricciones inherentes a los parámetros del cuantificador. La metodología fue validada por medio de 24 bases de datos de referencia y fue comparado contra 10 algoritmos de predicción. Finalmente, se estudió el desempeño de la metodología en una aplicación real basada en la estimación de la calidad del aire en el valle de Aburrá. Los resultados muestran que VQNN-PSO tiene un desempeño superior a la mayoría de métodos evaluados a excepción de Regresión por Vectores de Soporte optimizados con enjambre de partículas, donde exhiben un rendimiento similar. De otro modo, en la predicción de calidad del aire, VQNN-PSO se mostró como un sistema robusto ante cambios climáticos y estable a lo largo de las estaciones de monitoreo en el valle de Aburrá.

Palabras clave: Vagamente Cuantificados Vecinos más Cercanos, Optimización por Enjambre de Partículas, Regresión por Vectores de Soporte, Predicción, Conjuntos difusos aproximados.

Abstract

Air pollution is known globally as one major risk factor for human health, in this regard, prediction models have been widely used to developed prevention systems. Vaguely Quantified Nearest Neighbor (VQNN) is a prediction model based on fuzzy rough set theory, it is able to model the ambiguity and uncertainty in the data using fuzzy quantifiers. On the other hand, data associated with air quality monitoring are characterized for their high non-linearity and are very sensitive to noise and uncertainty. In this context, VQNN is an ideal method to

forecast the pollutant concentration in air quality applications, nevertheless, if fuzzy quantifiers are not properly tuning lead to a high sensitive to noise with a poor generalization capability and increase the uncertainty and imprecision of the predictions. This work aims to propose a prediction methodology using Vaguely Quantified Nearest Neighbor optimized by particle swarm Optimization technique (VQNN-PSO) to estimate pollutants concentration in the Aburrá Valley. The proposed methodology uses a loss function based on the Mean Absolute Percentage Error and employ a modified particle swarm optimization algorithm to handle the imposed restrictions in the fuzzy quantifiers parameters. The methodology was evaluated through 24 benchmark databases and it was compared with 10 prediction algorithms. Finally, the methodology performance was studied in a real application based on air quality estimation in the Aburrá Valley. Results shows that VQNN-PSO overperform most of the evaluation methods, except for Support Vector Regression optimized by particle swarm. Finally, in air quality prediction task, VQNN-PSO shows to be a robust strategy against climatic changes and stable a long of the monitoring stations in the aburra valley.

Keywords: Vaquely Quantified Nearest Neighbors, Particle Swarm Opimization, Support Vector Regresion, Forecasting, Fuzzy Rough Sets

Contenido

Lista de Figuras

Lista de Tablas

1 Introducción

1.1. Planteamiento del problema

La predicción de eventos y variables es una tarea de gran importancia en diferentes áreas del conocimiento como la economía, medicina, geoinformática entre otras, y se consolida como una herramienta clave para el desarrollo de sistemas de control, sistemas de alerta temprana y la toma de decisiones. Un gran número de metodologías para la predicción y su clasificación son reportados en la literatura [82, 72, 81, 49]. Sin embargo, para el desarrollo de este trabajo dichos métodos fueron clasificados en determinísticos, estadísticos, aprendizaje de máquina e híbridos.

El rendimiento de los sistemas de predicción determinísticos está ligado a su complejidad, es decir que para obtener un sistema con un alto desempeño es preciso contar con una gran cantidad de ecuaciones que permitan establecer la relación entre las variables involucradas, generando así un aumento excesivo en el costo computacional y limitando su aplicación a sistemas con respuesta en tiempo real [40, 50]. Métodos estadísticos como los modelos autoregresivos y sus derivaciones presentan una alta capacidad de universalidad y un desempeño eficiente en predicciones a corto plazo, sin embargo el desempeño de los sistemas es inferior comparado con sistemas basados en aprendizaje de máquina [31, 45]. No obstante, los sistemas de predicción de aprendizaje de máquina requieren datos con bajos niveles de ruido e incertidumbre, para poder establecer una dinámica entre las variables del fenómeno que se está estudiando en ambientes de prueba controlados, donde se requiere un protocolo estricto para la recolección o medición de los datos . Esto sugiere una dificultad ya que en ambientes reales los datos están expuestos a múltiples fuentes de incertidumbre, producto del mal etiquetado, interferencia electromagnética entre otros, generando lo que se conoce como conocimiento imperfecto [14, 26, 43, 49, 85].

En aras de superar las dificultades que supone el conocimiento imperfecto, han emergido diversas alternativas entre las que destacan las teorías de conjuntos aproximados y conjuntos difusos, ambos métodos permiten mejorar el desempeño de los sistemas de predicción frente a datos ruidosos, no obstante estas alternativas se alejan de dar una solución definitiva. En la teoría de Conjuntos Aproximados la vaguedad existente en la definición de las aproximaciones altas y la inflexibilidad que supone la definición de las aproximaciones bajas, provoca un alto grado de dicotomía, generando un comportamiento similar a la teoría de conjuntos

clásica lo que derivan en una alta sensibilidad frente imprecisiones en la información [56, 86].

Para mejorar el cálculo de las aproximaciones altas y bajas se ha utilizado el concepto de función de membresía tomado de la teoría de conjuntos difusos, a esta combinación se le conoce como teoría de conjuntos difusos aproximados (*Fuzzy Rough Sets* FRS) empleado en el desarrollo de sistemas de predicción [22]. Una derivación de FRS es Vecinos más Cercanos Difusos Aproximados (*Fuzzy Rough Nearest Neighbor* FRNN), el cual usa la información promedio disponible de la aproximaciones altas y bajas para realizar una predicción, sin embargo ésta sólo es influenciada por el objeto con mayor similitud, por lo cual el sistema no utiliza la información disponible en el vecindario y cuando se trabaja con datos ruidosos e imprecisos es posible que las predicciones puedan ser realizadas con base en información errónea [26, 89]. Vecinos más Cercanos Vagamente Cuantificados (*Vaguely Quantified Nearest Neighbor* VQNN) realiza las predicciones basado en la sumatoria de similitudes en el vecindario, de este modo VQNN hace uso de toda la información disponible en el vecindario para aumentar su capacidad de trabajo cuando se tienen datos ruidosos, además, VQNN permite disminuir la incertidumbre e imprecisión asociada con datos imperfectos, haciendo uso de cuantificadores difusos para reducir la sensibilidad del sistema al ruido [26, 22, 56, 138].

La elección de los parámetros de los cuantificadores es una tarea compleja debido a la dependencia con el tipo y grado de ruido en lo datos, la variación entre diferentes aplicaciones y la incertidumbre asociada a los procesos de predicción [37, 138]. Así, una mala sintonización de los cuantificadores difusos ocasiona una alta sensibilidad al ruido del sistema con una baja generalidad y aumento en la imprecisión e incertidumbre de las predicciones realizadas. En la revisión de la literatura realizada, no se reportó una metodología para la elección de los parámetros de los cuantificadores difusos, en cambio se muestra que la sintonización de estos se realiza de forma empírica, lo que no garantiza un desempeño óptimo, de otro modo, realizar una búsqueda exhaustiva de los parámetros representa una tarea de alta complejidad, debido a la cantidad de parámetros y restricciones. Por lo anterior, cuando se utilizan sistemas de predicción con base en VQNN en aplicaciones reales y no se garantiza un correcto funcionamiento de los cuantificadores difusos, la incertidumbre e imprecisión de las predicciones pueden causar diagnósticos erróneos, planificación y toma decisiones inadecuadas, entre otras.

1.2. Justificación

El uso de los sistemas de predicción ha sido ampliamente explorado en el estudio de la polución del aire, debido su reconocimiento global como uno de los mayores factores de riesgo para la salud humana. La exposición a contaminantes es un contribuyente importante a las tasas de muerte por enfermedad cardiorrespiratoria de acuerdo con la Organización Mundial

de Salud. Muertes prematuras causadas por isquemia cardíaca (72 %), enfermedad pulmonar obstructiva crónica (14 %) y Cáncer de pulmón (14 %) son reportadas a nivel mundial, aumentando en regiones de bajo desarrollo tecnológico y económico, como en el pacífico occidental y el sur este asiático [137, 48].

En Colombia, los mayores costos ambientales y sociales son generados por la contaminación del agua, los desastres naturales y la contaminación del aire [122], siendo esta ultima el problema ambiental más grave que presenta el país, según las estadísticas del Banco Mundial[97]. En el Valle de Aburrá se ha evidenciado en los últimos años un aumento significativo de la concentración de contaminantes, generando un aumento en la morbilidad por enfermedades asociadas al tracto respiratorio, sistema cardiovascular y en las capacidades reproductivas, estimando 46 defunciones por cada 100.000 habitantes en la ciudad de Medellín asociadas con la calidad del aire [7].

Las fuentes de emisión de contaminantes del aire están clasificadas en estacionarias y móviles, las primeras están compuestas por la producción agrícola, la minería, las canteras, zonas industriales, plantas químicas y de generación de energía, calefacción de edificios e incineradores de residuos. Por otra parte, las fuentes móviles las componen los vehículos con motores de combustión interna [76] como los automóviles, motocicletas y camiones. En áreas urbanas las fuentes móviles contribuyen alrededor del 14 % de los gases de efecto invernadero, generando fuertes variaciones en el clima global y alterando las condiciones naturales de los ecosistemas [55, 141].

Desde otro perspectiva, el comportamiento meteorológico también influye en la dispersión de los contaminantes atmosféricos generando cambios en su concentración, factores el viento, estabilidad atmosférica, radiación solar, precipitación y topografía generan condicione propicias para su transporte y dispersión [66]. En Colombia, se dan fenómenos climatológicos como el ciclo del Niño y la Niña - Oscilación del Sur, debido a la ubicación geográfica del país y a la circulación de los vientos alisios entre los trópicos, estos fenómenos ademas de generar impactos socioeconómicos como inundaciones por altas precipitaciones y sequías por la carencia de lluvias, afectan la circulación atmosférica, generando así cambios abruptos en los apoltrones de dispersión de los contaminantes.

Diferentes herramientas computacionales permiten predecir la concentración de contaminantes del aire, generando aplicaciones móviles y servicios web, que brindan a las personas la oportunidad de evaluar el estado de la calidad del aire y así planear sus actividades diarias, evitando la exposición a niveles insalubres de contaminación. Además, estas herramientas son de gran aplicación en los organismos gubernamentales, ya que son sistemas de apoyo a la toma de decisiones para evaluar políticas y mitigar situaciones potencialmente perjudiciales para la salud pública y de los ecosistemas. Numerosos modelos computacionales de predic-

ción de calidad del aire han sido desarrollado en torno a diversas estrategias matemáticas, de los cuales los modelos determinísticos no logran modelar la dinámica de las fuentes de emisión (principalmente las móviles) en zonas urbanas. De otro modo, los modelos estadísticos y de aprendizaje de máquina no requieren modelar las fuentes de emisión, en cambio utilizan información histórica de la concentración de los contaminantes recolectada por los sistemas de monitoreo instalados en las diferentes ciudades, susceptible a diferentes fuentes de incertidumbre asociadas con los equipos de medición y factores humanos, así como a las variaciones repentinas en la meteorología.

VQNN es un modelo de predicción no lineal, de alta tolerancia al ruido e incertidumbre, lo cual es ideal para el trabajo con datos del calidad del aire, dada su naturaleza imperfecta y alta no linealidad, estos factores son más severos en el Valle de Aburrá dada sus condiciones topográficas y climáticas.

1.3. Objetivo

1.3.1. General

Proponer una metodología de predicción de calidad del aire con base en Vecinos más Cercanos Vagamente Cuantificados optimizados por Enjambre de Partículas.

1.3.2. Específicos

Determinar una función objetivo para la optimización los cuantificadores difusos de un sistema de predicción basado en conjuntos difusos aproximados vagamente cuantificados y medidas de desempeño tipo *wrapper*.

Desarrollar un sistema para predicción de la concentración de contaminantes usando el algoritmo Vecinos más Cercanos Vagamente Cuantificados optimizados por enjambre de partículas.

Validar el desempeño de la metodologia propuesta respecto a algoritmos de predicción convencionales con datos reales derivados de aplicaciones biomédicas.

2 Estado del arte

2.1. Métodos de predicción

Recientemente la predicción de variables y eventos ha tomado gran interés en diversas áreas del conocimiento, gracias a que permiten el desarrollo de sistemas de alerta temprana, apoyan los procesos de toma de decisiones y permiten generar sistemas de medición y monitoreo. La literatura reporta un gran número de metodologías utilizadas para realizar predicciones, las cuales pueden ser clasificadas de diferentes formas según la estructura del modelo, tipo de datos o información requerida para la formación del sistema. En [82], se clasifican los sistemas de predicción en asociación de reglas, agrupamiento, arboles de decisión, vecinos más cercanos, redes neuronales, análisis de conexiones, regresión y métodos heurísticos. Lotte y col en [72], proponen una clasificación en sistemas lineales, redes neuronales, sistemas no lineales, vecinos más cercanos y sistemas híbridos. En [81], los sistemas de predicción se clasifican como sistemas expertos, algoritmos genéticos, redes neuronales, sistemas basados en conocimiento, sistemas de soporte de decisión, sistemas basados en lógica difusa y sistemas híbridos. Otros autores reconocen dos enfoques en el desarrollo de sistemas de predicción basado en la información y conocimientos necesarios para formular un modelo, estos son los métodos determinísticos y los métodos empíricos [49]. De este modo en este trabajo los métodos de predicción se clasificaron en determinísticos, estadísticos, aprendizaje de máquina e híbridos, esto con el fin de agrupar los métodos tanto por su arquitectura como por el tipo de información necesaria para su constitución (Figura **2-1**).

Los métodos determinísticos implican un conocimiento profundo del fenómeno que se está estudiando, con el fin de establecer un modelo matemático entre las variables involucradas y su dinámica, por su parte los métodos estadísticos y de aprendizaje de máquina no requieren de un conocimiento extenso del fenómeno, sin embargo es necesario contar con una gran cantidad de datos que reflejen la dinámica del sistema para así determinar un modelo [49]. Finalmente la combinación de dos o más métodos bien sean determinísticos, estadísticos o de aprendizaje de máquina, se conocen como métodos híbridos, esto permite aprovechar las bondades de las técnicas que se están fusionando, no obstante en ocasiones esto conlleva a un aumento en la complejidad de los sistemas provocando comportamientos no deseados [110, 120].

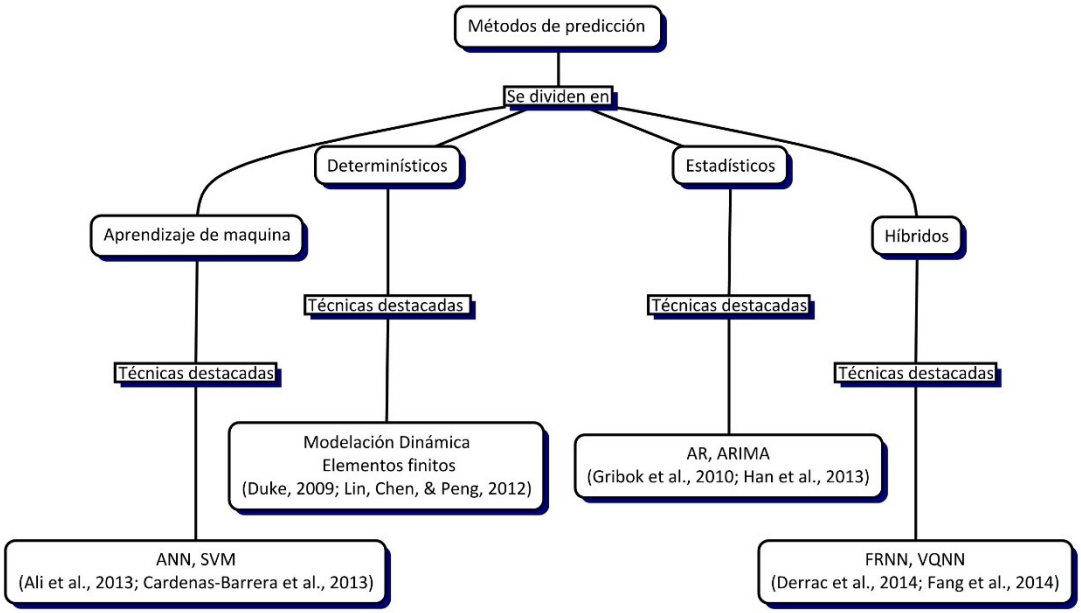


Figura 2-1: Clasificación de métodos de predicción

2.1.1. Métodos determinísticos

Entre las técnicas deterministas más utilizadas para generar modelos de predicción son la Modelación Dinámica, Elementos Finitos, Modelación Difusa y Modelación Numérica. En el área de geoinformática se han empleado para el desarrollo de sistemas de alerta y prevención de desastres naturales como inundaciones, tsunamis y terremotos [32, 69, 99]. De otro modo, en el área de salud humana se reportan modelos que permiten predecir los efectos hemodinámicos de procedimientos quirúrgicos [50], la respuesta inmune ante infecciones pulmonares [108], el riesgo de función retardada en trasplantes de riñón [53], los niveles de glucosa en sangre [29], la fuerza y el consumo energético de un músculo [124], evidenciando comportamientos adecuados en predicciones a corto plazo, capacidad de funcionamiento en condiciones variables y una precisión alta cuando el modelo se encuentra bien calibrado.

2.1.2. Métodos estadísticos

En este campo diversas técnicas se reportan en la literatura, entre las cuales se destacan los modelos Auto Regresivos (AR) y sus derivados, lo cuales han sido utilizados en el desarrollo de sistemas de predicción de la temperatura central en humanos [45], la polución del aire [31] y el precio de la energía eléctrica [64]. Los modelos Autoregresivos Integrados de Media Móvil (*Autoregressive Integrated Moving Average*-ARIMA) han sido usados para predecir la presión intracraneal en pacientes con traumas cerebrales, la producción de trigo y la inci-

dencia de fiebre hemorrágica con síndrome renal, presentando un mejor rendimiento frente a los AR al trabajar con series de tiempo no estacionarias [47, 88].

Sistemas basados en derivaciones de los modelos AR como los Modelos Autoregresivos Exógenos (*Autoregressive Exogenous*-ARX) y los Modelos Autoregresivos de Media Móvil Exógenos (*Autoregressive Integrated Moving Average Exogenous*-ARIMAX), se han utilizado para predecir la concentración de glucosa subcutánea, el tiempo, la demanda de electricidad, entre otros [23, 34, 65]. Los modelos AR y sus derivaciones presentan una alta capacidad de universalidad y un desempeño eficiente en predicciones a corto plazo, sin embargo cuando los datos utilizados para entrenar el modelo son de baja calidad el desempeño del sistema se ve reducido comparado con sistemas basados en Aprendizaje de Máquina.

2.1.3. Métodos de aprendizaje de máquina

Diversos métodos de aprendizaje de máquina para tareas predictivas han sido usados para aplicaciones en las ciencias sociales, ambientales, biomédicas, entre otras. Dentro del amplio espectro de técnicas asociadas al aprendizaje de máquina, las Redes Neuronales Artificiales (*Artificial Neural Network*-ANN) y la Regresión por Soporte de Vectores (*Support Vector Regression*-SVR), se destacan como técnicas ampliamente utilizadas a lo largo de diferentes aplicaciones , debido a la velocidad de entrenamiento, flexibilidad y desempeño [31, 36].

Por medio de ANN se han desarrollado sistemas para predecir la velocidad de viento, la radiación solar, el flujo másico de refrigerante y el consumo de combustible de vehículos entre otros, encontrando que en comparación con los métodos deterministas, las ANN presentan un desempeño superior en términos de precisión, de igual manera, en comparación con ARIMA se reportan incrementos de hasta un 50 % en el rendimiento [14, 21, 90, 104, 121].

La técnica SVR se ha empleado en el desarrollo de sistemas para predecir la extensión del hielo en el mar, precio de acciones, precio de la energía eléctrica, precio del petróleo, niveles de glucosa y nivel de agua en represas, encontrando una gran capacidad en la captura de la dinámica dominante del sistema[1, 13, 25, 43, 91, 119].

Wohlfarth y otros en [136], presentan la predicción de la evolución del precio de viajes aéreos usando la técnica de agrupamiento *Marked Point Process of Return*, concluyendo que el sistema posee una gran flexibilidad para predecir el precio de un pasaje, incluso en periodos superiores a siete días. En [105] se presenta un método con base en evolución gramatical para la extracción y selección de características en la predicción de la carga eléctrica con aplicaciones en procesos de control de generación de energía, el método propuesto genera características que luego alimentan una máquina de aprendizaje para realizar la predicción, el sistema obtuvo errores de entre 1.48 % y el 2.16 %.

El método k Vecinos más Cercanos (k *Nearest Neighbor* k-NN) en su forma tradicional es uno de los algoritmos de clasificación más usados debido a su simplicidad, además por medio de pequeñas modificaciones puede utilizarse para realizar la predicción de una variable continua, a este algoritmo se le conoce como Regresión k-NN el cual ha sido empleado para predecir la edad basado en imágenes faciales, el consumo eléctrico de servidores, fallas en la manufactura de semiconductores, el tempo de la música, energía eólica y fallas en software de interacción, mostrando un desempeño aceptable. Sin embargo, su desempeño es superado por métodos como SVR [16, 17, 36, 38, 44, 73].

2.1.4. Métodos híbridos

Como se aprecia en la Figura **2-2**, en los métodos híbridos se presentan diferentes combinaciones entre métodos estadísticos, de Aprendizaje de Máquina y Determinísticos, los cuales tienen como objetivo aprovechar las bondades de cada método en pro de una solución al problema. Las Máquinas de Soporte Vectorial de Mínimos Cuadrados (*Least Squares Support Vector Machines* LS-SVM) son métodos híbridos usados en la predicción de series de tiempo univariantes y multivariantes, mostrando resultados consistentes conforme se aumenta el horizonte predictivo [18]. Un método híbrido usando LS-SVM y enjambre de partículas con base en Recocido Simulado (*Particle Swarm Opmization Simulated Annealing*-PSOSA), se ha empleado en la predicción de la velocidad del viento, obteniendo desempeños superiores en un 2 % comparado con métodos estadísticos como ARIMA [128], para la misma tarea se propone el uso de un método basado en SVR, Índice de Ajuste Estacional y Redes Neuronales Recurrentes Elman (*Elman Recurrent Neural Networks*-ERNN) con el fin de eliminar los datos atípicos y mejorar en las predicciones a mediano plazo [130].

Por medio de Máquinas de Aprendizaje Extremas (*Extreme Learning Machines* ELM), el Ljung-Box Q-test (LBQ) y un Modelo Estacional Autoregresivo de Media Móvil (*Self Autoregressive Integrated Moving Average* SARIMA), se desarrolló un sistema para predecir la velocidad del viento, el modelo híbrido fue comparado con métodos estadísticos y de aprendizaje de máquina ARIMA, SARIMA, ANN y ELM obteniendo mejores resultados para predicciones a corto y largo plazo [103]. Modelos de predicción híbridos entre Filtros de Kalman (*Kalman Filter* KF) y ANN han sido formulados para predecir la velocidad del viento, mostrando un desempeño superior al de cada uno de sus componentes por separado [129].
Técnicas de predicción híbridas basadas en agrupamiento se han desarrollado para predecir la velocidad del viento y completar información pérdida en el censado del tráfico vehicular, entre ellas se destacan los híbridos entre información granular y agrupamiento difuso, C-means difusos y Algoritmos Genéticos y por último agrupamiento espectral y las redes de estado de eco, mostrando predicciones precisas y desempeños superiores a modelos existentes

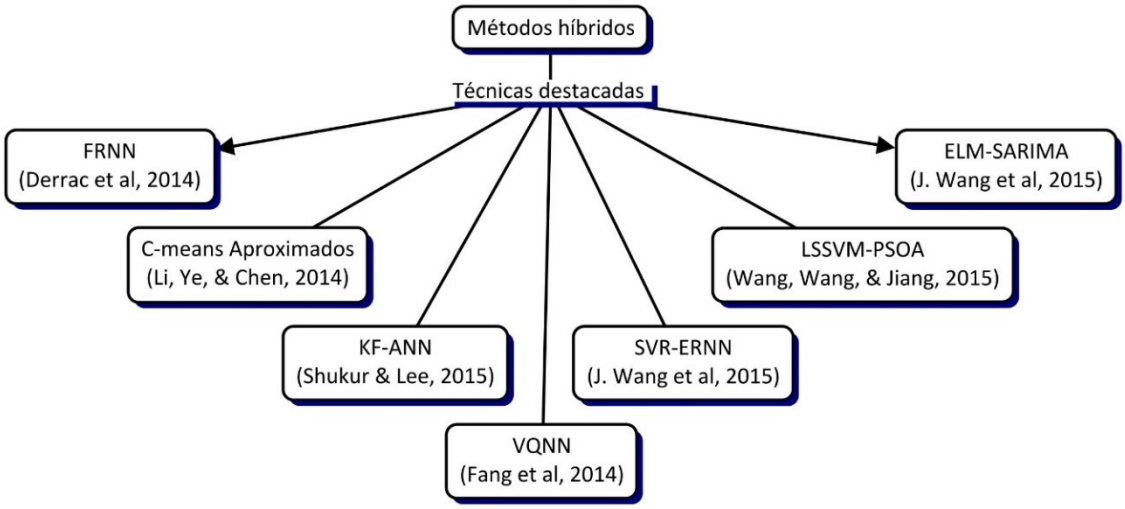


Figura 2-2: Métodos de predicción híbridos

basado en ARIMA y ANN [71, 118, 132].

Con base en la teoría de conjuntos aproximados, se han desarrollado métodos de predicción como *Rough C-means* que no incorporan información acerca del vecindario, lo que puede generar soluciones inadecuadas, por ello se proponen métodos híbridos que utilizan los modelos de decisión teórica aproximados, para calcular una función de costo que permita incluir información del vecindario [66]. Los sistemas basados en conjuntos aproximados han sido combinados con sistemas difusos para aplicaciones de regresión mostrando un desempeño superior que SVR y Regresión k-NN [3, 96].

k-NN y SVR se han empleado para predecir la generación de energía eólica y la edad basada en imágenes faciales mostrando en términos del error cuadrático medio una mejoría de hasta un 5 % respecto a sistemas basados en Modelos Persistentes [16, 123]. Otros métodos híbridos como vecinos más cercanos difusos (Fuzzy-NN), k-NN probabilístico, k-NN intusionista, vecinos más cercanos difusos aproximados (*Fuzzy Rough Nearest Neighbor*-FRNN) y vecinos más cercanos vagamente cuantificados (*Vaguely Quantified Nearest Neighbor*-VQNN), han sido explorados en la literatura en su mayoría para tareas de clasificación [26], en este sentido sistemas basado en FRNN se han utilizado para el análisis de imágenes marcianas y reconocimiento de patrones en bases de datos de alta dimensionalidad [101, 117]. Por su parte, los algoritmos basados en VQNN se han usado en el reconocimiento de expresiones faciales e identificación de personas con una precisión superior al 80 % [37, 68, 138]. De igual forma, en la literatura se reportan metodologías enfocadas en tareas predictivas, por ejemplo Fuzzy-NN se ha empleado para predecir la demanda de gas natural [92], de otro modo en

[56] proponen un método para utilizar los algoritmos de FRNN y VQNN en tareas regresivas conservando la simplicidad algorítmica de las estrategias basadas en k-NN, obteniendo un desempeño superior a métodos como Fuzzy-NN y un desempeño comparable con SVR.

2.2. Predicción de la calidad del aire en zonas urbanas

En el contexto de calidad del aire se han empleado múltiples algoritmos de regresión para predecir la concentración temporal y espacial de diversos contaminantes, bajo múltiples condiciones experimentales [100]. La predicción espacial consiste en determinar la concentración de un contaminante en un punto geográfico, utilizando la información disponible en otros lugares. De otro modo, la predicción temporal hace referencia a determinar la concentración futura de un contaminante, utilizando la información actual y/o pasada disponible en el mismo punto [100]. En este trabajo, es de interés identificar los métodos de regresión más empleados en la predicción temporal de contaminantes del aire y/o índices de calidad del aire.

2.2.1. Predicción espacial

La predicción espacial de contaminantes del aire pretende determinar la concentración de uno o varios contaminantes a lo largo de una superficie, en este contexto se han desarrollado diversas estrategias basados en métodos de interpolación estadística y geoestadística como *Inverse Distance Weigthing* (IDW), Vecino más Cercano, *Splines* y *Kriging*, sin embargo estos métodos solo muestran buenos resultados trabajando en escalas regionales o globales, por lo cual no exhiben buenos resultados trabajando en zonas urbanas, debido a la complejidad dada la heterogeneidad de las fuentes de emisión [100].

Para predecir la concentración de Material Particulado menor a 2.5 μm ($PM_{2,5}$) en la superficie de diversas ciudades en el norte de China se usaron ANN, mostrando un aumento en el desempeño del 22 % en comparación con la Regresión Lineal Múltiple [70], resultados similares se encontraron para predecir Óxidos Nitrosos (NO_x) en Londres [102]. Sistemas Adaptativos de Inferencia Neuro-Difusos (*Adaptive Neuro Fuzzy Inference System*-ANFIS) se usaron para predecir la polución de dióxido de sulfuro (SO_2), alrededor de una mina de cobre en Bor (Serbia), encontrando resultados poco precisos debido a la falta de datos [75].

2.2.2. Predicción temporal

La predicción temporal ha sido el enfoque más estudiado en la literatura, de este modo podemos encontrar múltiples estudios que se centran en la elección de las variables de en-

trada usando métodos como transformada wavelet, características caóticas y el análisis de correlación con variables ambientales y otros contaminantes [9, 94, 98]. Además, se ha evaluado la predicción en múltiples horizontes predictivos partiendo desde una hora hasta semanas.

Las ANN es uno de los métodos más recurrente en la literatura para el modelado de series de tiempo de contaminantes del aire. Trabajos recientes muestran su uso en predecir dióxido de nitrógeno (NO_2), ozono (O_3), SO_2, material particulado suspendido (*Suspended Particle Matter*-SPM) y Material Particulado Menor a 10 μm (PM_{10}), usando horizontes de predicción de 24 y 48 horas, encontrando que estas tiene una buena capacidad de generalización (alrededor del 90 %), además que puede ser fácilmente implementadas en sistemas en tiempo real, sin embargo también reportan que estos sistemas pueden ser extremadamente sensibles a valores atípicos en los datos, lo cual es común en los sistemas de monitoreo de calidad, debido a problemas con los sensores [94, 30, 106, 24, 27, 2].

Sistemas basados en ANFIS son frecuentes en la literatura, dos trabajos recientes reportan su uso en predicción de PM_{10}, $PM_{2,5}$ y O_3, en áreas urbanas de Rumanía y en la mega ciudad de Nueva Delhi (India), encontrado que si bien ANFIS mejora el desempeño de ANN y regresión lineal múltiple (*Multiple Linear Regresion*-MLR), es incapaz de predecir concentraciones altas, además reportan que la topología de la red y el método de optimización son dependientes de la locación geográfica [78, 79]. De forma menos frecuente, aparecen modelos basados en Modelos Ocultos de Markov (*Hidden Markov Models*-HMM) y Arboles de Decisión, mostrando buenos resultados para predecir O_3 y $PM_{2,5}$, no obstante las concentraciones elevadas de los contaminantes modifican la distribución estadística de los datos (alarga la cola) alterando el desempeño de las HMM [30, 114, 115].

Los métodos basados en SVR han tomado relevancia y se han desarrollado múltiples sistemas para predecir Monóxido de Carbono (CO), $PM_{2,5}$, PM_{10}, NO, NO_2, SO_2 y O_3, encontrando resultados prometedores debido a las capacidades no lineales de las SVR, además se evidencia del uso de métodos de optimización para la selección de los parámetros de la SVR dada la complejidad de la tarea [15, 113, 67, 41, 139, 112, 5, 42].

Recientemente se ha evidenciado el desarrollo de sistemas basados en la combinación de diferentes regresores, por ejemplo en [107], se combinan tres árboles de decisión diferentes (*Single Decision Tree, Desicion Three Forest* y *Desicion Treeboost*) para predecir el Índice de Calidad del Aire (*Air Quality Index*-AQI), obteniendo un desempeño superior a un SVR convencional. Combinación aditiva de diferentes métodos se han usado para predecir NO_2, SO_2 y PM_{10}, encontrando resultados superiores al uso de los métodos por separado [27, 98, 135, 131].

3 Marco conceptual

3.1. Conjuntos aproximados

La teoría de conjuntos aproximados ha sido desarrollada en el marco de las ciencias computacionales específicamente en el área de inteligencia artificial con el fin de mitigar los problemas asociados al conocimiento imperfecto. Propuesta en 1982 por el matemático Polaco Zdzislaw Pawlak, la teoría de conjuntos aproximados es una extensión de la teoría de conjuntos clásica, en la cual se asumen que disponemos de información adicional sobre los elementos que componen un conjunto (sistema de información) [84].

Un sistema de información es una matriz en la que cada fila se asocia a un objeto por ejemplo, paciente, muestra o evento, cada columna representa una característica o atributo del objeto. De manera formal un sistema de información se representa como $B = (U, D)$, donde U es un conjunto finito no-vacío de objetos denominado Universo. Por su parte D es un conjunto finito no vacío de características o atributos de los objetos [116]. La teoría de conjuntos aproximados utiliza el concepto de indiscernibilidad para mitigar el impacto del conocimiento imperfecto, en el desempeño de sistemas de clasificación y regresión. Dos objetos (x, y) son indiscernibles si son reflexivos (xRy), simétricos (si xRy, entonces yRx) y transitivos (si xRy y zRy entonces xRz), esta relación binaria es conocida como relación de equivalencia y se representa como $[x]_D$ o $R(x, y)$ [62].

Una relación de equivalencia genera particiones del universo, obteniendo nuevos subconjuntos, generalmente dichos subconjuntos están relacionados con un característica de decisión. Debido a problemas de conocimiento imperfecto es posible que un objeto no pueda ser descrito de forma precisa frente a su argumento de decisión, sin embargo podemos determinar algunos objetos que ciertamente pertenecen a una característica de decisión, aquellos que no pertenecen a dicha característica y finalmente un último subconjunto de objetos que se encuentran en medio de pertenecer o no a la característica de decisión, si este último conjunto es no vacío se dice que el conjunto es aproximado [127].

Si $B = (U, D)$ es sistema de información, de modo tal que $A \subseteq D$ y $X \subseteq U$, es posible aproximar X por medio de la información disponible en A, por medio de las aproximaciones altas ($\overline{A}X$) y aproximaciones bajas de ($\underline{A}X$), la información disponible en A para cualquier elemento $y \in X$ se puede cuantificar por medio de una relación de indiscernibilidad R:

$$y \in \underline{A} \quad \Leftrightarrow \quad (\forall x \in X)((x,y) \in R \Rightarrow x \in A), \tag{3-1}$$

$$y \in \overline{A} \quad \Leftrightarrow \quad (\exists x \in X)((x,y) \in R \wedge x \in A). \tag{3-2}$$

La información disponible en A se cuantifica por medio de la relación de indiscernibilidad. Los objetos agrupados en $\underline{A}X$, son aquellos que pueden ser clasificados con certeza en como miembros de X, por su parte $\overline{A}X$ es el conjunto de los objetos que posiblemente pueden ser clasificados como miembros de X. El conjunto de los elementos que no podrán ser clasificados de manera decisiva como miembros de X, se denomina región de frontera de X y se expresa como $BN_A = \overline{A}X - \underline{A}X$. Finalmente aquellos objetos que con certeza no pueden ser clasificados como miembros de X, se denominan región de afuera y se expresa como $A - \overline{A}X$, se dice que el conjunto es aproximado si BN_A es un conjunto no vació.

3.2. Conjuntos difusos aproximados

Propuesta en 1965 por Lotfi Zadeh, la teoría de conjuntos difusos plantea una generalización a la teoría de conjuntos clásica con el fin de proporcionar un marco de trabajo robusto frente a situaciones de conocimiento imperfecto en aplicaciones de reconocimiento de patrones y procesamiento de la información. En la teoría de conjuntos clásica un objeto pertenece o no a una clase mientras que en la teoría de conjuntos difusos la pertenencia a una clase es medida por medio de una función de membresía, la cual tiene diferentes grados de pertenencia [46].

Ahora cuando A es un conjunto difuso y R es una relación difusa, las aproximaciones altas y bajas de X en función de la información disponible en A se pueden calcular como:

$$\underline{A}(y) = \inf_{y \in X} \mathcal{I}\left(R\left(x,y\right), A\left(y\right)\right), \tag{3-3}$$

$$\overline{A}(y) = \sup_{y \in X} \mathcal{T}\left(R\left(x,y\right), A\left(y\right)\right), \tag{3-4}$$

donde $\mathcal{I}$ y $\mathcal{T}$ representan un Implicador difuso y una T-norma respectivamente.

La "*fuzzificación*"de las aproximaciones altas y bajas permite suavizar las fronteras respecto a sus contrapartes basadas únicamente en conjuntos aproximados, no obstante el uso de los operadores inf y sup hacen a las aproximaciones sensibles frente a datos ruidosos [22].

3.3. Conjuntos difusos aproximados vagamente cuantificados

Conjuntos Difusos-Aproximados Vagamente Cuantificados (*Vaguely Quantified Fuzzy Rough Sets* VQFRS) propone reemplazar los operadores *sup* e *inf* por cuantificadores difusos para construir las aproximaciones altas y bajas, en este caso para construir el conjunto de aproximaciones bajas se utilizaría el cuantificador "La mayoría" y para las aproximaciones altas se utilizaría el cuantificador "Algunos" [56]. De esta forma dado un par de cuantificadores difusos (Q_u, Q_l), las aproximaciones altas y bajas de un sistema se puede expresar como:

$$_{Q_l}\overline{A}(y) = Q_l\left(\frac{|R_y \cap A|}{|R_y|}\right) = Q_l\left(\frac{\sum_{x\in X} \text{mín}\left(R(x,y), A(y)\right)}{\sum_{x\in X} R(x,y)}\right), \tag{3-5}$$

$$_{Q_u}\underline{A}(y) = Q_u\left(\frac{|R_y \cap A|}{|R_y|}\right) = Q_u\left(\frac{\sum_{x\in X} \text{mín}\left(R(x,y), A(y)\right)}{\sum_{x\in X} R(x,y)}\right). \tag{3-6}$$

Los cuantificadores difusos pueden ser generados utilizando la siguiente función [22]:

$$Q_{(\alpha,\beta)}(x) = \begin{cases} 0, & x \leq \alpha \\ \frac{2(x-\alpha)^2}{(\beta-\alpha)^2}, & \alpha \leq x \leq \frac{\alpha+\beta}{2} \\ 1 - \frac{2(x-\beta)^2}{(\beta-\alpha)^2}, & \frac{\alpha+\beta}{2} \leq x \leq \beta \\ 1, & \beta \leq x \end{cases} \tag{3-7}$$

donde $0 < \alpha < \beta < 1$.

3.4. Métodos de regresión basados en vecinos más cercanos

3.4.1. Regresión k-NN

El algoritmo de los k vecinos más cercanos (k-NN) es uno de los algoritmos más usados en tareas de reconocimiento de patrones, si bien su aplicación más común es tareas de clasificación, este puede ser fácilmente modificado para usar como un regresor (ver algoritmo 1).

3.4.2. Regresión Fuzzy-NN

El algoritmo Fuzzy-NN fue propuesto para aumentar la robustez del sistema frente a datos imprecisos, en este sentido Fuzzy-NN utiliza una Relación de similitud difusa en lugar de una medida de distancia, en este trabajo se utilizó la siguiente medida de similitud:

$$R_A(x,y) = \frac{1}{|A|}\sum_{a\in A}\frac{R_a(x,y)}{a_{max}-a_{min}}, \tag{3-8}$$

donde A es el conjunto de características y $R_a(x,y) = 1 - |a(x) - a(y)|$.

De este modo el algoritmo 2 establece el procedimiento empleado en Fuzzy-NN.

3.4.3. Vecinos más Cercanos Vagamente Cuantificados (VQNN)

VQNN es una modificación del algoritmo de Fuzzy-NN, que permite brindar una mayor robustez al hacer uso de la información de las aproximaciones altas y bajas definidas en VQFRS (ver algoritmo 3).

En el algoritmo 3, R_yz es un relación difusa respecto a la variable de salida calculada como $R_yz = R_y(x,z) = \frac{1-|y(x)-y(z)|}{y_{max}-y_{min}}$. En este contexto las aproximaciones altas y bajas se calculan como:

$${}_{Q_l}\overline{R_yz}(\mathbf{x}_i) = Q_l\left(\frac{\sum_{\mathbf{x}\in NN}\text{mín}\left(R(\mathbf{x},\mathbf{x}_i), R_y(x,z)\right)}{\sum_{\mathbf{x}\in X}R(\mathbf{x},\mathbf{x}_i)}\right), \tag{3-9}$$

Algoritmo 1: Regresión k-NN

Entradas: X_{ent}, Y_{ent}, k y $\mathbf{x}_i$
Salida: $\widehat{y}_i$
$j \leftarrow 0$
para *cada* $x \in X_{ent}$ *Calcular la distancia con respecto a* x_i **hacer**
  $j \leftarrow j + 1$
  $d(j) \leftarrow ||\mathbf{x} - \mathbf{x}_i||$
NN $\leftarrow$ Obtener los k vecinos más cercanos con base en la distancia d
$t \leftarrow 0$
para *cada* $z \in NN$ **hacer**
  $t \leftarrow t + y_z$
$\widehat{y}_i \leftarrow t/k$

Algoritmo 2: Regresión Fuzzy-NN

Entradas: X_{ent}, Y_{ent}, k y $\mathbf{x}_i$
Salida: $\widehat{y_i}$
$j \leftarrow 0$
para *cada* $x \in X_{ent}$ **hacer**
  $j \leftarrow j + 1$
  $R_j(\mathbf{x}, \mathbf{x}_i)$
NN $\leftarrow$ Obtener los k vecinos más cercanos con base en R
$t_1 \leftarrow 0$; $t_2 \leftarrow 0$
para *cada* $z \in NN$ **hacer**
  $t_1 \leftarrow t_1 + y_z * R(\mathbf{x}_i, \mathbf{x}_z)$
  $t_2 \leftarrow t_2 + R(\mathbf{x}_i, \mathbf{x}_z)$
$\widehat{y_i} \leftarrow t_1/t_2$

$$ {}_{Q_u}\underline{R_y z}(\mathbf{x}_i) = Q_u \left(\frac{\sum_{\mathbf{x} \in NN} \text{mín}\left(R(\mathbf{x}, \mathbf{x}_i), R_y(\mathbf{x}, z)\right)}{\sum_{\mathbf{x} \in X} R(\mathbf{x}, \mathbf{x}_i)} \right). \tag{3-10} $$

Los parámetros de los cuantificadores difusos corresponden a los reportados en [22, 57, 37], es decir $\overline{Q_{l(0,1,0,6)}}$ y $\underline{Q_u}_{(0,2,1)}$.

Algoritmo 3: Regresión VQNN

Entradas: X_{ent}, Y_{ent}, k, Q_u, Q_l y $\mathbf{x}_i$
Salida: $\widehat{y_i}$
$j \leftarrow 0$
para *cada* $\mathbf{x} \in X_{ent}$ **hacer**
  $j \leftarrow j + 1$
  $R_j(\mathbf{x}, \mathbf{x}_i)$
NN $\leftarrow$ Obtener los k vecinos más cercanos con base en R
$t_1 \leftarrow 0$; $t_2 \leftarrow 0$
para $z \in NN$ **hacer**
  $M \leftarrow \frac{Q_l \overline{R_y Z}(\mathbf{x}_i) + Q_u \underline{R_y Z}(\mathbf{x}_i)}{2}$
  $t_1 \leftarrow t_1 + M * y_z$
  $t_2 \leftarrow t_2 + M$
$\widehat{y_i} = t_1/t_2$

3.5. Regresión por Vectores de Soporte Epsilon (ϵ-SVR)

Para un conjunto de datos de entrenamiento $X = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots (\mathbf{x}_n, y_n)\}$, donde y_i representa el valor objetivo para un vector de entrada $\mathbf{x}_i$, el objetivo de SVR es encontrar

una función que satisfaga $f(\mathbf{x}_i) \approx y_i$, para todas las muestras en el conjunto de datos de entrenamiento. El algoritmo más usado para este propósito es el denominado ϵ-SVR, el cual utiliza una función de perdida que le permite a $f(\mathbf{x})$ presentar errores menores que ϵ [19]. Para el caso de funciones lineales f puede representarse como:

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}_i + b, \tag{3-11}$$

donde $\mathbf{w}$ es un vector que contiene coeficientes de ponderación para cada una de las características y b es el término de tendencia, ambos parámetros pueden encontrarse resolviendo el siguiente problema de optimización [109]:

$$\begin{aligned} &\min \tfrac{1}{2}\left\|\mathbf{w}\right\|^2 \\ &\text{Sujeto a: } \begin{cases} y_i - (\mathbf{w}^T\mathbf{x}_i + b) \leqslant \epsilon \\ (\mathbf{w}^T\mathbf{x}_i + b) - y_i \leqslant \epsilon \end{cases} \end{aligned} \tag{3-12}$$

Sin embargo en la mayoría de ocasiones esto representa un problema de optimización inviable, debido a que no es posible aproximar todos los valores de entrenamiento con una precisión ϵ, por lo cual es necesario agregar variables de holgura ξ_i, ξ_i^* y un parámetro de costo ($C > 0$) para controlar la relación entre la complejidad del sistema y la cantidad de error mayor que ϵ que es tolerado, de este modo el problema de optimización queda como [6]:

$$\begin{aligned} &\min \tfrac{1}{2}\left\|\mathbf{w}\right\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ &\text{Sujeto a: } \begin{cases} y_i - (\mathbf{w}^T\mathbf{x}_i + b) \leqslant \epsilon + \xi_i \\ (\mathbf{w}^T\mathbf{x}_i + b) - y_i \leqslant \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \end{aligned} \tag{3-13}$$

La formulación dual del problema de optimización ofrece la clave para extender la capacidad de SVR a funciones no lineales, para ello se emplean multiplicadores de Lagrange (α_i y α_i^*) y el problema de optimización puede expresare como:

$$\begin{aligned} &\max \ -\tfrac{1}{2}\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\mathbf{x}_j^T\mathbf{x}_i - \epsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{n} y_i(\alpha_i + \alpha_i^*) \\ &\text{Sujeto a: } \begin{cases} \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0 \\ \alpha_i, \alpha_i^* \in [0, C] \end{cases} \end{aligned} \tag{3-14}$$

De este modo $f(\mathbf{x})$ puede representarse como $\sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\mathbf{x}_i^T \cdot \mathbf{x} + b$, lo que se conoce como la expansión en vectores de soporte. Sustituyendo en la ecuación 3-14 $\mathbf{x}_j^T\mathbf{x}_i$, por una función Kernel $K(\mathbf{x}_j, \mathbf{x}_i)$, se puede extender el funcionamiento de las SVR a funciones no lineales.

En las ϵ-SVR es necesario la elección del parámetro de costo C y el ϵ de la función de perdida, además de un función kernel y sus parámetros inherentes. En este contexto [20] propone la elección del parámetro de costo C como el rango de la variable de salida, no obstante este enfoque puede ser sensible al ruido por lo que los autores sugieren encontrar C como:

$$C = \text{máx}(|\overline{y} + 3\sigma_y|, |\overline{y} - 3\sigma_y|) \tag{3-15}$$

donde $\overline{y}$ es la media de la variable de salida del conjunto de entrenamiento y σ_y su varianza.

En [20] se propone encontrar el parámetro ϵ en función del ruido de los datos y el número de muestras como:

$$\epsilon = 3\sigma_r \sqrt{\ln(n)/n} \tag{3-16}$$

Donde n es el número de muestras y σ_r representa la desviación estándar del ruido en los datos.

Dado que en la mayoría de aplicaciones se desconoce el nivel de ruido, es necesario estimar σ_r directamente de los datos. Para ello utilizamos una regresión de módulo mínimo, es decir se entrena una SVR donde todos los datos sean vectores de soporte (esto se logra haciendo $\epsilon = 0$), posteriormente se realiza un predicción con la función obtenida para los datos de entrenamiento, finalmente σ_r es la desviación estándar de las diferencias entre los datos reales y las estimaciones.

El estado del arte muestra que el Kernel Gaussiano de la forma $K(\mathbf{x}, \mathbf{y}) = e^{-||\mathbf{x}-\mathbf{y}||^2/2\sigma^2}$ tiene buenos resultados con su uso, no obstante el parámetro de amplitud σ tiene un gran impacto en el desempeño del regresor por ello su elección debe hacerse de forma cuidadosa. Para ello se siguió la metodología propuesta en [133], la cual consiste en un método adaptativo inspirado en representación multi-escala del sistema visual (ver algoritmo 4).

3.6. Regresión por Vectores de Soporte Nu (ν-SVR)

Durante el entrenamiento de las ϵ-SVR, es necesario establecer de forma preliminar el valor de ϵ, lo cual resulta en una tarea compleja debido a que este parámetro tiene una gran influencia en el desempeño del algoritmo, este problema ha sido solucionado de forma parcial por el denominado ν-SVR [6]. El cual propone solucionar el siguiente problema de optimización:

Algoritmo 4: Método adaptativo para la selección del parámetro de amplitud σ

inicio
- $i = 0$
- $\sigma_i = 0{,}001$
- Entrenar SVR
- Obtener estimación $\widehat{y}_i$
- $MAPE_i = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \widehat{y}_i|}{y_i}$
- **repetir**
  - $i = i + 1$
  - $\sigma_i = 1{,}029\sigma_{i-1}$
  - Entrenar SVR
  - Obtener estimación $\widehat{y}_i$
  - $MAPE_i = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \widehat{y}_i|}{y_i}$
- **hasta que** $MAPE_i > MAPE_{i-1}$;
- **devolver** σ_{i-1}

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

$$\text{Sujeto a: } \begin{cases} y_i - (\mathbf{w}^T\mathbf{x}_i + b) \leqslant \epsilon + \xi_i^* \\ (\mathbf{w}^T\mathbf{x}_i + b) - y_i \leqslant \epsilon + \xi_i \\ \xi_i, \xi_i^* \geq 0 \end{cases} \tag{3-17}$$

Nuevamente si expresamos el problema de optimización en su forma dual y utilizamos una función Kernel $K(\mathbf{x}_i, \mathbf{x}_j)$, podemos dar solución únicamente en términos de los multiplicadores de Lagrange:

$$\max \sum_{i=1}^{n}(\alpha_i^* + \alpha_i)y_i - \tfrac{1}{2}\sum_{i,j=1}^{n}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)K(\mathbf{x}_i, \mathbf{x}_j)$$

$$\text{Sujeto a: } \begin{cases} \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0 \\ \alpha_i, \alpha_i^* \in [0, C/n] \\ \sum_{i=1}^{n}(\alpha_i + \alpha_i^*) \leq C\nu \end{cases} \tag{3-18}$$

Finalmente podemos expresar la función de regresión como una expansión por medio de vectores de soportes como:

$$f(\mathbf{x}) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)k(\mathbf{x}_i, \mathbf{x}) + b \tag{3-19}$$

Las ν-SVR se utilizaron con un Kernel Gaussiano y seleccionando sus parámetros con el algoritmo 4, la elección del parámetro de costo C se realizó de la misma manera que en las ϵ-SVR. Finalmente, la elección de ν puede realizar por medio de una búsqueda exhaustiva que optimice alguna medida de desempeño.

3.7. Optimización por enjambre de partículas

PSO es un algoritmo metaheurístico inspirado en el comportamiento de los enjambres de aves o peces introducido para optimizar funciones continuas no lineales [60]. Este algoritmo realiza la búsqueda en el espacio de los parámetros de la función objetivo ajustando la trayectoria de las partículas.

La matriz P de dimensiones $n \times d$, contiene la posición de las n partículas en cada una de las d dimensiones, el vector $\mathbf{v}$ contiene la velocidad de las n partículas, el movimiento del enjambre de partículas está determinado por dos componentes: pensamiento individual y pensamiento colectivo. Cada partícula es atraída a moverse a la mejor posición $\mathbf{p}_i^*$ donde se ha encontrado, a este comportamiento se le denomina pensamiento individual, por su parte el pensamiento colectivo atrae a las partículas a la mejor posición encontrada entre todas las partículas $\mathbf{g}^*$. El pensamiento individual le brinda al algoritmo la capacidad de exploración mientras que el pensamiento colectivo garantiza la convergencia del algoritmo.

La actualización de las posiciones de las partículas en t+1, se realiza por medio de la siguiente ecuación:

$$\mathbf{p}_i^{(t+1)} = \mathbf{p}_i^{(t)} + \mathbf{v}_i^{(t+1)} \tag{3-20}$$

Por su parte la actualización de la velocidad se realiza por medio de:

$$\mathbf{v}_i^{(t+1)} = \mathbf{v}_i^{(t)} + \alpha\epsilon_1 \cdot [\mathbf{g}^* - \mathbf{p}_i^{(t)}] + \beta\epsilon_2 \cdot [\mathbf{p}_i^* - \mathbf{p}_i^{(t)}] \tag{3-21}$$

Donde α y β son constantes de aceleración, de forma típica $\alpha \approx \beta \approx 2$, por su parte ϵ_1 y ϵ_2 son vectores aleatorios, con valores entre $[0, 1]$ que permiten controlar la influencia que tiene el pensamiento individual y colectivo sobre el cambio en la velocidad de la partícula.

3.8. Contaminantes del aire

La contaminación del aire se define como la presencia de sustancias contaminantes en el aire que interfieren con la salud o el bienestar humana, o que producen efectos medio ambientales

dañinos [125]. En este contexto, *The Clean Air Act* de 1970 establece los seis contaminantes del aire críticos:

- Material Particulado (*Particulate matter* PM)
- Ozono (O_3)
- Monoxido de carbono (CO)
- Dióxido de sulfuro (SO_2)
- Dióxido de nitrógeno (NO_2)
- Plomo (Pb)

En el desarrollo de este trabajo es de interés el estudio de PM_{10}, $PM_{2,5}$, NO, NO_2 y O_3.

3.8.1. Material particulado

El aire está compuesto de una serie de gases y vapores (Nitrógeno, oxígeno, agua, argón, dióxido de carbono, neón, helio, metano, NO, hidrógeno, xenón y vapores orgánicos), los cuales están presentes como moléculas individuales. No obstante, en el aire siempre está presente una cantidad PM suspendido, el cual se compone de diversas moléculas. El PM alcanza el aire gracias a rutas naturales como la condensación, reacciones químicas, partículas de sal, polen, etc., además, la actividad humana genera PM debido a la quema de combustibles fósiles, actividades mineras, construcción y producción agrícola [126].

El PM es usualmente clasificado de acuerdo a su tamaño, ya que este define su comportamiento en la atmósfera, en este sentido podemos clasificarlo en Material Particulado menor a 10 μm (PM_{10}) y Material Particulado menor a 2.5 μm ($PM_{2,5}$). El PM_{10} puede ser encontrado en el humo proveniente del exosto de los autos e industrias, por su parte el $PM_{2,5}$ aparece solo de forma indirecta , es decir solo se detectan por la forma en la que difuminan, absorben y reflejan la luz. Es importante notar que el $PM_{2,5}$ tiene la capacidad de estar suspendida en el aire por mayor tiempo debido a su poca masa, mientras que las partículas de mayor dimensión tienden a removerse del aire en un proceso llamado sedimentación, el cual es favorecido por factores como la lluvia, nevadas, granizo y niebla[126].

La literatura ha establecido una clara relación entre el PM y la tasa morbilidad y mortalidad, demostrando que un incremento en la concentraciones del PM tiene efectos negativos en la salud [33]. EL PM está relacionado con enfermedades cardiovasculares, pulmonares, mortalidad diaria, función pulmonar, tos y cáncer de pulmón [33, 4, 93]. Se ha identificado que polimorfismo genéticos, enfermedades cardiovasculares y respiratorias pre-existentes, y

el estatus socio económico, son factores que agudizan la susceptibilidad de la población a los efectos adversos de la exposición a niveles de PM insalubres [95].

3.8.2. Óxidos nitrosos y ozono troposférico

Los óxidos nitrosos (NO_x), se conforman en su mayoría de NO y NO_2, estos son generados por los motores de combustión interna. Su mayor impacto esta asociado con la generación O_3, cuando reacciona con la luz solar [59]. Sin embargo, se ha demostrado que el NO_x por si solo acarrea una serie de impactos en la salud humana a nivel del sistema cardiovascular, respiratorio y nervioso [10, 59].

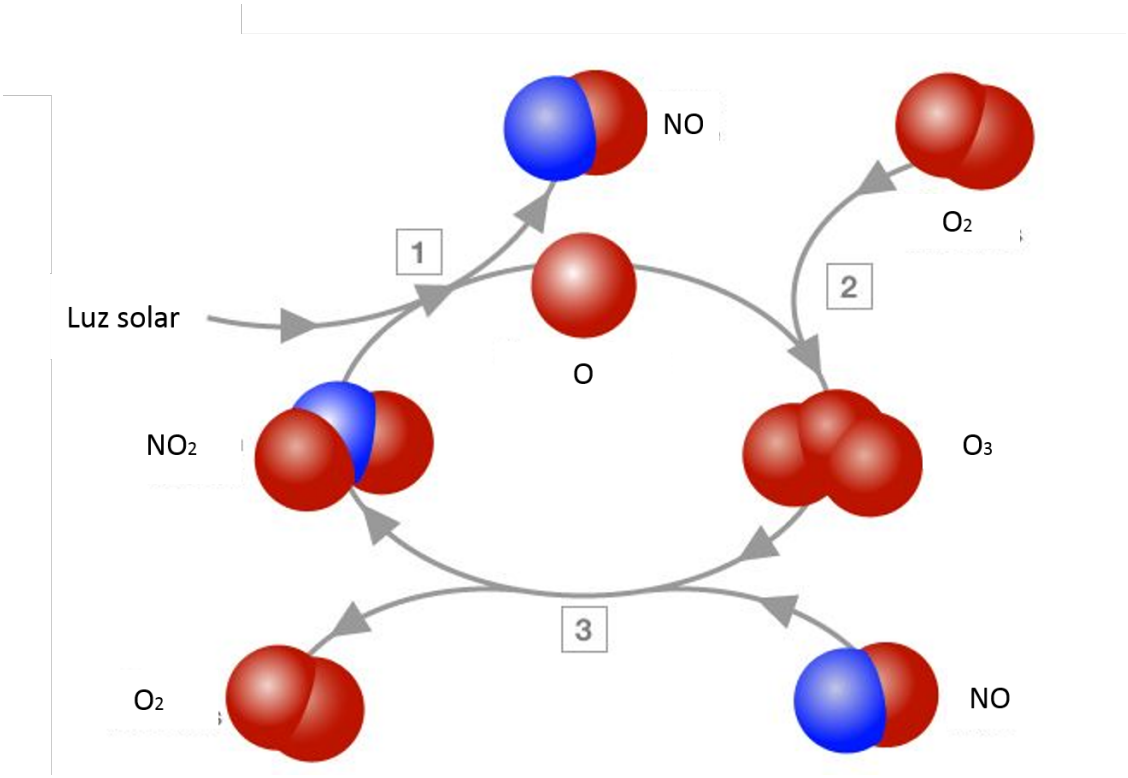


Figura 3-1: Ciclo del ozono troposférico

Si bien el O_3 es un gas que se encuentra de forma natural en el aire, en la estratosfera (50 Km de altitud con respecto al nivel del mar) el O_3 forma la denominada capa de ozono, la cual tiene como tarea repeler la radiación UV emitida por el sol [39]. De otro modo, el O_3 troposférico (a menos de 10 Km sobre el nivel del mar) tiene un impacto negativo sobre la salud huma y los ecosistemas. El O_3 en troposfera, se produce como resultado de reacciones químicas entre gases, la fuente más común de ozono se asocia con los NO_x, por medio del ciclo denominado ciclo del ozono (ver Figura **3-1**), no obstante se reconocen otras fuentes como

los compuestos volátiles orgánicos [28, 35]. Referente a los impactos a la salud humana, se ha demostrado que el O_3 irrita las vías respiratorias, está relacionado con infartos al miocardio y empeora el asma infantil [111, 33, 80].

4 Sistema de predicción basado en VQNN y cuantificadores difusos optimizados por enjambre de partículas

El desarrollo de sistemas de regresión es una tarea ampliamente abordada en la literatura, donde se han propuesto múltiples metodologías con el fin de abordar los diferentes problemas asociados a la regresión como complejidad computacional, costo computacional y conocimiento imperfecto [134]. El conocimiento imperfecto hace referencia a la incertidumbre existente en los datos e información a partir de los cuales se pretende establecer un modelo, la incertidumbre se asocia comúnmente con errores humanos en recolección de los datos y a la imprecisión en los sistemas de medición. En este orden de ideas la teoría de conjuntos difusos [140] y la teoría de conjuntos aproximados [83], ofrecen un marco conceptual que permite manejar algunos de los desafíos que supone establecer modelos en ambientes de conocimiento imperfecto.

Los métodos de regresión basados en la hibridación de la teorías de conjuntos difusos y conjuntos aproximados han permitido abordar algunas de las dificultades que suponen cada uno de las teorías de forma individual (ver Sección 3.2). Vecinos más Cercanos Vagamente Cuantificados (VQNN) propuesto en [57], emplea cuantificadores difusos lingüísticos "alguno" y "la mayoría" con el fin de hacer más flexible la definición de las aproximaciones bajas y aumentar la rigidez en las aproximaciones altas, en este sentido los cuantificadores difusos permiten modelar la vaguedad existente en los datos y aumenta la robustez del sistema frente a datos corruptos con ruido o mal etiquetado, no obstante, la función de dichos cuantificadores requiere la definición de parámetros que garanticen su correcto funcionamiento, sin embargo, la literatura no reporta una metodología para la elección de estos parámetros en función del ruido existente en los datos, en cambio reporta que la elección se realiza de forma empírica.

En este trabajo se desarrolla una metodología tipo *wrapper* para la selección de cuantificadores difusos con base en optimización por enjambre de partículas (PSO), que permite la optimización de los cuantificadores difusos de un sistema de regresión basado en VQNN,

con el fin de garantizar el desempeño del sistema en diferentes condiciones de ruido y ante diversos tipos de datos.

Este capítulo está organizado en tres secciones: Marco metodológico para la optimización de cuantificadores difusos, Marco experimental y Resultados. En el Marco metodológico se expone la formulación del problema de optimización para la selección de parámetros de los cuantificadores, en esta se aborda la elección de la función objetivo, las restricciones del problema de optimización y la implementación del selector de parámetros por medio de un algoritmo híbrido VQNN-PSO. En el marco experimental se describen los experimentos realizados y el análisis del desempeño de la metodología propuesta. Finalmente la sección de resultados y discusión expone los resultados obtenidos en los experimentos realizados.

4.1. Marco metodológico para la optimización de cuantificadores difusos

4.1.1. Función objetivo

Con el fin de optimizar los cuantificadores difusos en el marco de una metodología tipo *wrapper*, fue necesario utilizar como función objetivo una medida que permita estimar el desempeño del sistema de regresión, en este sentido cualquiera de las medidas de desempeño numeradas en la tabla **4-1** puede ser utilizada como función objetivo, no obstante en el desarrollo de este trabajo se utilizó el Porcentaje del Error Absoluto Medio (MAPE), dado que esta medida es insensible a la escala de los datos, permitiendo establecer una comparación objetiva entre diferentes bases de datos [52], además es una de las medidas más empleadas en la evaluación de sistemas de regresión aplicados en [18, 25, 43, 51, 61, 63, 64].

Tabla 4-1: Medida de desempeño

Sigla	Nombre	Fórmula
MSE	Error Cuadrático Medio	$\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\widehat{y_i}\right)^2$
MBE	Error Parcial Medio	$\frac{1}{n}\sum_{i=1}^{n}(y_i-\widehat{y_i})$
RMSE	Raíz del Error Cuadrático Medio	$\frac{1}{n}\sqrt{\sum_{i=1}^{n}\left(y_i-\widehat{y_i}\right)^2}$
MSM	Medida de Similaridad Media	$\frac{1}{n}\sum_{i=1}^{n}\frac{\min(y_i\widehat{y_i})}{\max(y_i\widehat{y_i})}$
MAE	Error Absoluto Medio	$\frac{1}{n}\sum_{i=1}^{n}\left\|y_i-\widehat{y_i}\right\|$
MAPE	Porcentaje del Error Absoluto Medio	$\frac{1}{n}\sum_{i=1}^{n}\frac{\left\|y_i-\widehat{y_i}\right\|}{y_i}$
IA	Índice de Concordancia	$1-\frac{(\widehat{y_i}-y_i)^2}{(\left\|\widehat{y_i}-\overline{y}\right\|+\left\|y_i-\overline{y}\right\|)^2}$

y_i valor real, $\widehat{y_i}$ valor estimado, $\overline{y}$ media valores reales, $\overline{\widehat{y_i}}$ media valores estimados, n número de muestras

Dado que las estimaciones son obtenidas utilizado VQNN por medio del algoritmo expuesto en la sección 3.4.3, este proceso puede ser definido como una función la cual depende de los datos de entrenamiento, el vector de entrada, el número de vecinos más cercanos y los parámetros del cuantificador, de este modo una estimación de VQNN puede ser expresado como:

$$\widehat{y_i} = \text{VQNN}(X_{train}, Y_{train}, \mathbf{x}_i, \alpha_l, \beta_l, \alpha_u, \beta_u, k), \tag{4-1}$$

donde X_{train} y Y_{train}, representan los datos de entrenamiento, $\mathbf{x}_i$ representa el vector de entrada, α_l y β_l representan los parámetros del cuantificador difuso "alguno" usado para el cálculo de las aproximaciones altas, mientras que α_u y β_u son los parámetros del cuantificador difusos "la mayoría" usado para la definición de las aproximaciones bajas, finalmente k representa el número de vecinos más cercanos.

De acuerdo a lo anterior, el problema de optimización puede ser expresado como:

$$\min_{\alpha_l,\beta_l,\alpha_u,\beta_u,k} \qquad \text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \widehat{y_i}|}{y_i} \tag{4-2}$$

En el problema de optimización se incluyó el número de vecinos cercanos k, con el fin de garantizar que el desempeño del sistema de regresión, no dependa de la elección de parámetros por parte del usuario. Además, esto implica una disminución en el costo computacional, al no ser necesario repetir la optimización de los cuantificadores si k es modificado.

4.1.2. Restricciones

Teniendo en cuenta la definición matemática de los cuantificadores difusos (Ecuación 3-7), los parámetros agregan una restricción en la solución del problema de optimización ya que $0 \leq \alpha < \beta \leq 1$. Ahora en el contexto de modelar los cuantificadores difusos "alguno" y "la mayoría" utilizados en el cálculo de las aproximaciones altas y bajas respectivamente, a los que nos referiremos como $\overline{Q_u}_{(\alpha_l,\beta_l)}$ y $\underline{Q_l}_{(\alpha_u,\beta_u)}$, fue necesario modificar las restricción de los parámetros del cuantificador con el fin de preservar su sentido lingüístico, de modo que para $\underline{Q_l}$ se debe cumplir que $0{,}4 \leq \alpha_u < \beta_u \leq 1$ y para $\overline{Q_u}$ se tiene que $0 \leq \alpha_l < \beta_l \leq 0{,}5$, además, se debe garantizar que $\alpha_l < \alpha_u$ y $\beta_l < \beta_u$.

El parámetro k debe pertenecer al conjunto de los reales positivos ($k \in \mathbb{Z}^+$), por lo cual el límite inferior para k es 1, el límite superior, está definido por el número de muestras en el conjunto de datos de entrenamiento, al cual nos referiremos como n, no obstante en bases de datos grandes este límite puede ser inapropiado, por lo cual decidimos limitar k a el mínimo

entre n y 100.

Ahora teniendo en cuenta las restricciones inherentes a cada parámetro el problema de optimización definido en la ecuación 4-2 debe reescribirse como:

$$\begin{aligned} \min_{\alpha_l,\beta_l,\alpha_u,\beta_u,k,m} \quad & \text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \widehat{y_i}|}{y_i} \\ \text{sujeto a:} \quad & 0 \leq \alpha_l < \beta_l \leq 0{,}5 \\ & 0{,}4 \leq \alpha_u < \beta_u \leq 1 \\ & \alpha_l \leq \alpha_u \\ & \beta_l \leq \beta_u \\ & 1 \leq k \leq \text{ínf}(n, 100) \\ & k \in \mathbb{Z}^+ \end{aligned} \tag{4-3}$$

4.1.3. VQNN-PSO

Para la solución del problema de optimización formulado en la ecuación (4-3), se propone un sistema híbrido basado en VQNN y PSO (VQNN-PSO), el cual integra un selector de parámetros tipo *wrapper* basado en el algoritmo de optimización metaheurístico PSO.

Como se aprecia en la Figura **4-1**, VQNN-PSO solo tiene como entrada una base de datos, compuesta por la matriz de características X_{ent} y el vector de decisión Y_{ent}, de este modo no es necesario que el usuario ingrese ningún parámetro adicional. El algoritmo inicia generando de forma aleatoria una serie de soluciones (una solución hace referencia a un conjunto de parámetros α_u, β_u, α_l, β_l y k) que cumplan con las restricciones establecidas, para así por medio de un procedimiento de validación *leave one out* y usando el algoritmo VQNN obtener una estimación del vector de decisión $\widehat{Y}_{ent}$, de este modo se puede evaluar la función objetivo (Ecuación 4-2). PSO es un algoritmo iterativo, por lo cual el procedimiento descrito previamente se repite hasta satisfacer un criterio de parada, en este trabajo se usaron dos criterios de parada, el primero consiste en el número de iteraciones, mientras que el segundo corresponde a evaluar la convergencia de las partículas.

Una vez alcanzado el criterio de parada de PSO, la solución con el MAPE más bajo es entregado al VQNN para conformar el módulo de regresión definitivo (VQNN optimizado), este módulo también recibe el conjunto de datos de entrenamiento (X_{ent} y Y_{ent}), de este modo el sistema está listo para el recibir un vector de entrada x_i para generar una estimación $\widehat{y}_i$.

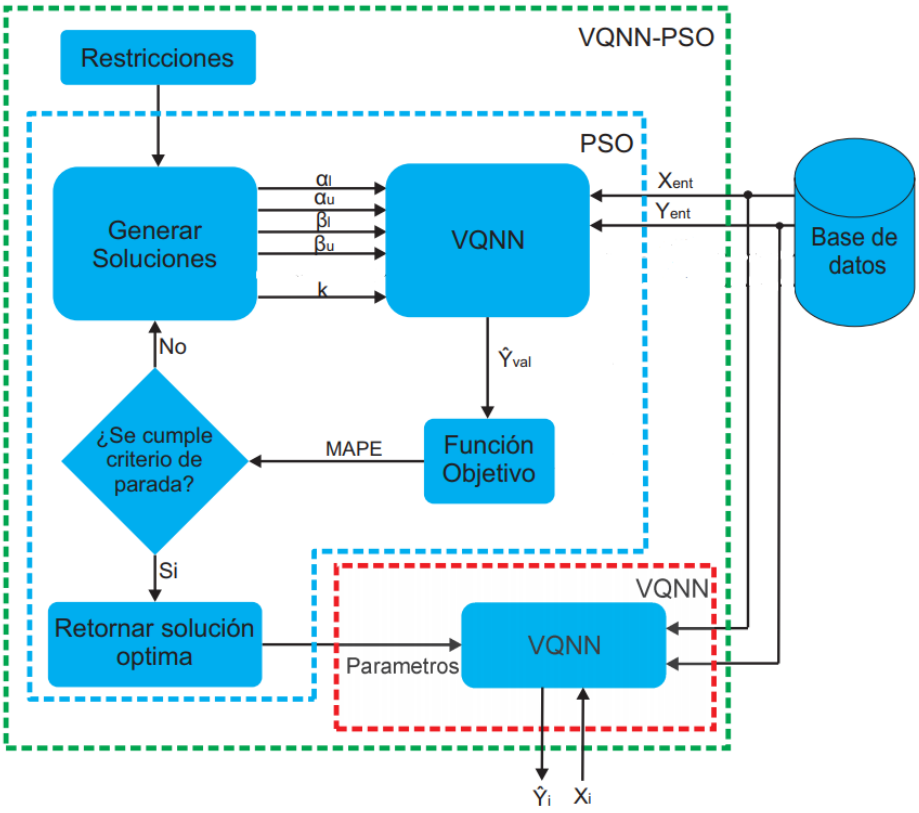
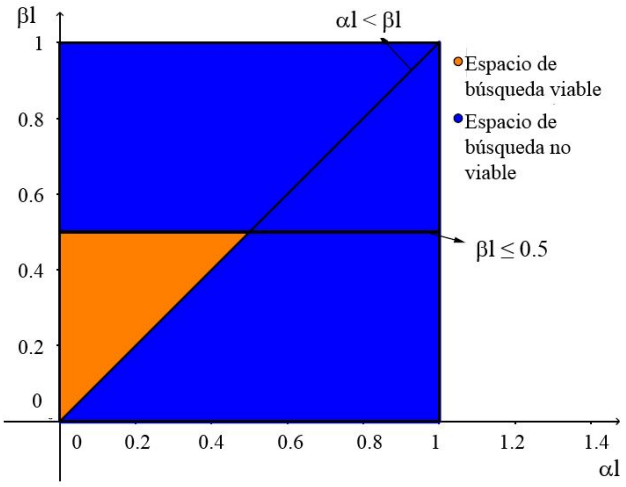


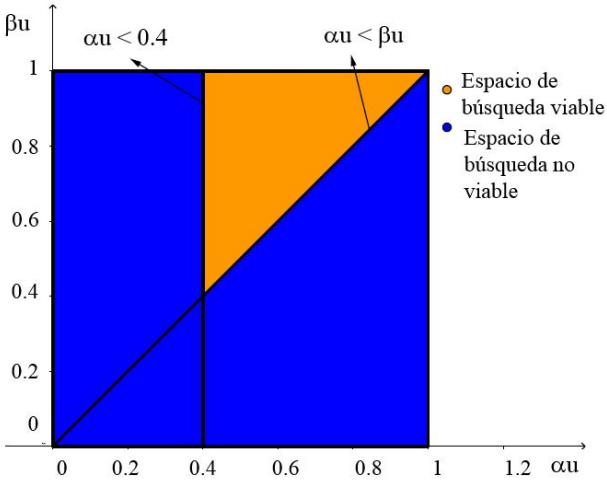
Figura 4-1: Arquitectura VQNN-PSO

Manejo de restricciones

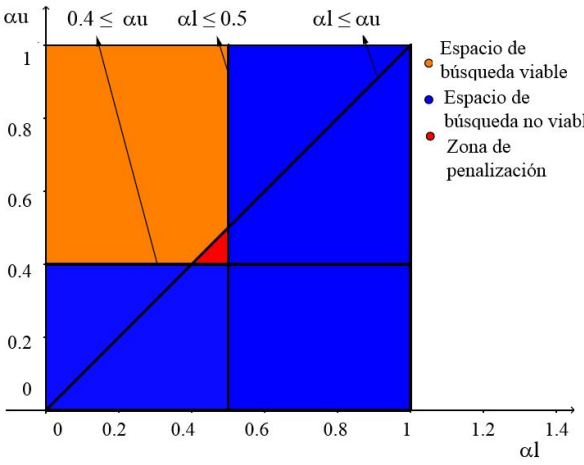
La Figura **4-2**, muestra el espacio de búsqueda viable para los parámetros de los cuantificadores teniendo en cuenta las restricciones establecidas.



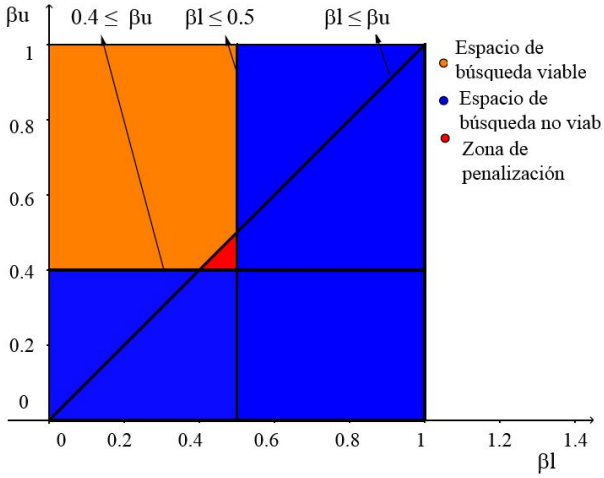
(a) Restricciones cuantificador "algunos"



(b) Restriciones cuantificador "la mayoría"



(c) Restriciones en α



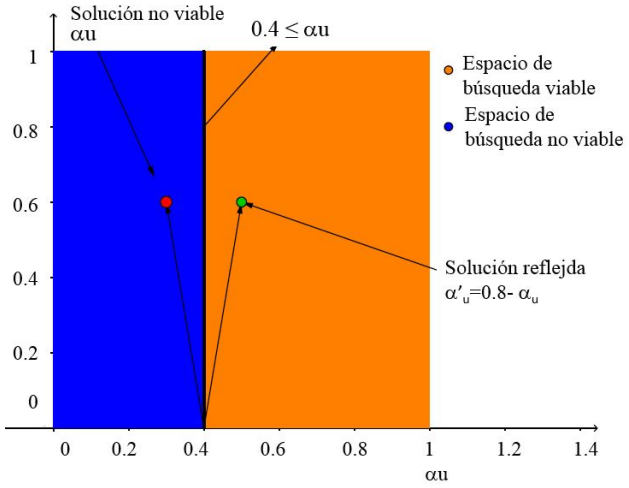
(d) Restriciones en β

Figura 4-2: Espacio de búsqueda para los parámetros de los cuantificador

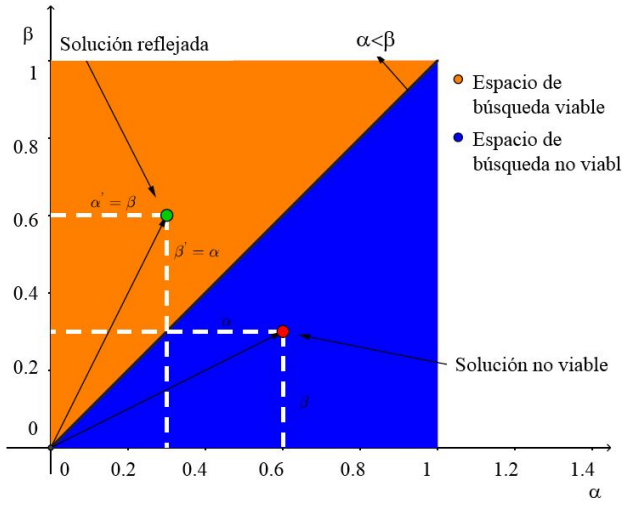
Fue necesario definir el mecanismo por el cual se garantice que las soluciones generados por PSO se encuentren dentro del espacio de búsqueda viable de los parámetros. La literatura muestra varias estrategias para el manejo de las restricciones, el mecanismo más común es el propuesto por Kennedy en [60], el cual se conoce como penalización. En este se asigna una valor muy alto (infinito) a la función objetivo de soluciones que no cumplen con las restricciones, de este modo se garantiza que estas nunca serán tomadas como el mejor local o global ($\mathbf{p}_i^*$ y $\mathbf{g}^*$), sin embargo este método genera espacios de búsqueda planos en las áreas de restricción, lo que genera dificultades en la convergencia de las partículas. De otro modo en [87], se analizan múltiples estrategias para trabajar con restricciones, sus resultados sugieren que el método *Reflejar-Absorber*, el cual consiste en reflejar al espacio viable las partículas que incumplan las restricciones [8] y simultáneamente hacer cero la velocidad de la partícula

en la dimensión donde se incumplieron las restricciones. Este método muestra una exploración equitativa a lo largo del espacio de búsqueda, al igual que un buen desempeño en diferentes funciones de prueba.

El método Reflejar fue propuesto para manejar restricciones concernientes a los límites del espacio de búsqueda (ver figura **4-3**.a), es decir para una partícula con posición en una de sus dimensiones (p_i), se limita el espacio de búsqueda entre p_{max} y p_{min}, esta restricción se puede expresar como $p_{min} \leq p_i \leq p_{max}$. De este modo, para una restricción violada, el método Reflejar aplica la siguiente corrección:



(a) Método Reflejar



(b) Método Reflejar Diagonal

Figura 4-3: Estrategias para el manejo de restricciones

$$\begin{array}{ll} \text{si} & p_i < p_{min} \quad entonces \quad p_i^{'} = p_{min} + (p_{min} - p_i), \\ \text{si} & p_i > p_{max} \quad entonces \quad p_i^{'} = p_{max} - (p_i - p_{max}). \end{array} \tag{4-4}$$

El método Reflejar fue utilizado para manejar las restricciones concernientes a los límites establecidos en el espacio de búsqueda de los cuantificadores ($0 \leq \alpha_l < \beta_l \leq 0{,}5$ y $0{,}4 \leq \alpha_u < \beta_u \leq 1$), note que si $\beta_l \leq 0{,}5$ entonces $\alpha_l \leq 0{,}5$, de igual forma para $0 \leq \alpha_l$ entonces $0 < \beta_l$. Para el cuantificador "algunos" se tiene que si $0{,}4 \leq \alpha_u$ entonces $0{,}4 < \beta_u$ y si $\beta_u \leq 1$ entonces $\alpha_u < 1$.

Es importante tener en cuenta que el método Reflejar fue propuesto solo para trabajar con restricciones que suponen un acotamiento horizontal o vertical del espacio de búsqueda, es por ello que las restricciones asociadas a $\alpha_l < \beta_l$ y $\alpha_u < \beta_u$ no pueden ser tratadas con este método, debido a que generan acotamientos diagonales en el espacio de búsqueda. Para ello se propuso una modificación del método reflejar, con el fin de permitir que la partícula se refleje al espacio viable utilizado la diagonal como eje de simetría, a este método nos referiremos como Reflejar-Diag.

Para logar una reflexión al espacio viable utilizando $\alpha_u = \beta_u$ y $\alpha_l = \beta_l$ como eje de simetría, basta con invertir las coordenadas de las partículas (ver Figura **4-3**.b), es decir que para ambos cuantificadores se tiene que:

$$\begin{array}{ll} \text{si} & \alpha_l > \beta_l \quad entonces \quad \alpha_l^{'} = \beta_l \quad \beta_l^{'} = \alpha_l, \\ \text{si} & \alpha_u > \beta_u \quad entonces \quad \alpha_u^{'} = \beta_u \quad \beta_u^{'} = \alpha_u. \end{array} \tag{4-5}$$

Si bien las restricciones $\alpha_l < \alpha_u$ y $\beta_l < \beta_u$ suponen también un acotamiento diagonal en el espacio de búsqueda (ver Figura **4-2**.c-d), no se abordaron utilizando el método Reflejar-diag, debido a que si una solución incumple estas restricciones, pero satisface las del cuantificador ($\alpha_l < \beta_l$ y $\alpha_u < \beta_u$), al reflejar la partícula se podrían incumplir las condiciones del cuantificador, es por ello que estas restricciones se abordaron utilizando el método Penalización.

Es importante destacar de la Figura **4-2**.c que la zona donde se impondrá la restricción de penalización es muy pequeña y las dificultades de espacios de búsqueda planos se mitigan, no obstante es necesario tener en cuenta que en el algoritmo original de PSO el mínimo local y global ($\mathbf{p}_i^*$ y $\mathbf{g}^*$), solo se actualizan si la nueva solución es menor a los valores actuales, por esta razón cuando se trabaja en espacios de búsqueda planos se genera una convergencia prematura del algoritmo. Para permitir la actualización de $\mathbf{p}_i^*$ y $\mathbf{g}^*$ cuando una nueva solución sean menor o igual, no obstante, esto se realiza solo el 50 % de las ocasiones, con el fin de no interferir demasiado con la dinámica normal del PSO [87].

La restricción referente a k, limita la solución únicamente a números reales positivos, lo cual genera un acotamiento drástico en el espacio de búsqueda, ocasionando mayor dificultad en obtener una solución, por ello esta restricción fue abordada aproximando las soluciones al entero mas cercano, para ello fue necesario agregar una corrección en los límites del espacio de búsqueda con el fin de garantizar la igualdad de probabilidades a lo largo del espacio de búsqueda, por lo cual el espacio de búsqueda queda limitado a $0{,}6 \leq k \leq \text{mín}(n, 100)+0{,}4$.

La Tabla **4-2** resume el manejo del algoritmo con cada una de las restricciones consideradas para la optimización de los cuantificadores difusos, el orden de aparición en la tabla corresponde también al orden en que se aplican las correcciones sobre las soluciones.

Tabla 4-2: Manejo de restricciones

Restricción	**Método**	**Corrección**
$\beta_l < 0{,}5$	Reflejar/absorber	$\beta_l^{'} = 1 - \beta_l$; $V_{i,\beta_l}^{'} = 0$
$\alpha_l < 0{,}5$	Reflejar/absorber	$\alpha_l^{'} = 1 - \alpha_l$; $V_{i,\alpha_l}^{'} = 0$
$\beta_l > 0$	Reflejar/absorber	$\beta_l^{'} = -\beta_l$; $V_{i,\beta_l}^{'} = 0$
$\alpha_l > 0$	Reflejar/absorber	$\alpha_l^{'} = -\alpha_l$; $V_{i,\alpha_l}^{'} = 0$
$\beta_u < 1$	Reflejar/absorber	$\beta_u^{'} = 2 - \beta_u$; $V_{i,\beta_u}^{'} = 0$
$\alpha_u < 1$	Reflejar/absorber	$\alpha_u^{'} = 2 - \alpha_u$; $V_{i,\alpha_u}^{'} = 0$
$\beta_u > 0{,}4$	Reflejar/absorber	$\beta_u^{'} = 0{,}8 - \beta_u$; $V_{i,\beta_u}^{'} = 0$
$\alpha_u > 0{,}4$	Reflejar/absorber	$\alpha_u^{'} = 0{,}8 - \alpha_u$; $V_{i,\alpha_u}^{'} = 0$
$\alpha_l < \beta_l$	Reflejar-Diag/absorber	$\alpha_l^{'} = \beta_l$; $\beta_l^{'} = \alpha_l$; $V_{i,\alpha_l,\beta_l}^{'} = 0$
$\alpha_u < \beta_u$	Reflejar-Diag/absorber	$\alpha_u^{'} = \beta_u$; $\beta_u^{'} = \alpha_u$; $V_{i,\alpha_u,\beta_u}^{'} = 0$
$\alpha_l < \alpha_u$	Penalización	$MAPE = \text{ínf}$
$\alpha_l < \alpha_u$	Penalización	$MAPE = \text{ínf}$
$0{,}5 < k$	Reflejar/absorber	$k^{'} = 1 - k$; $V_{i,k}^{'} = 0$
$k < \text{mín}(100, n) + 0{,}5$	Reflejar/absorber	$k^{'} = 2\,\text{mín}(100, n) + 1 - k$; $V_{i,k}^{'} = 0$
$k \in Z^+$	Redondear	$k^{'} = round(k)$

4.2. Marco experimental

Con el fin de evaluar la metodología propuesta, se realizó un análisis comparativo respecto a diez métodos de regresión, para ello se utilizaron 24 bases de datos derivadas de diferentes aplicaciones.

4.2.1. Descripción bases de datos

En el desarrollo de este estudio se emplearon 24 bases de datos derivadas de diferentes aplicaciones (ver Tabla **4-3**), las bases de datos utilizadas tienen entre 15 y 2178 muestras y cuentan con entre 3 y 17 características. Las bases de datos se clasificaron según el tipo de variable de entrada en Continuas, Enteras y Mixtas, además se clasificaron según el tipo de variable de salida en Enteras y Continuas. En este contexto, según la entrada se tienen 13 bases de datos Continuas, dos Enteras y 9 Mixtas, según la salida se tienen 5 bases de datos Enteras y 19 Continuas. Las bases de datos usadas son públicas y están disponibles en las páginas web http://www.keel.es/ y https://www.cs.waikato.ac.nz/ml/weka/.

Tabla 4-3: Descripción bases de datos

Nombre	**número de instancias**	**número de características**		
		Totales	**Continuas**	**Enteras**
Longley	15	7	2	5*
Schlvote	36	6	3	3*
Pollution	59	16	16*	0
Baskball	95	4	4*	0
CPU	118	7	7*	0
Fruitfly	124	5	3*	2
Veteran	136	8	0	8*
AutoPrice	158	16	16*	0
MachineCPU	209	7	0	7*
Bodyfat	251	15	15*	0
Baseball	337	17	2	15*
Dee	365	7	7*	0
AutoMPG6	392	6	3*	3
AutoMPG8	392	8	3*	5
Ele-1	495	3	2*	1
Stock	950	10	10*	0
Laser	993	5	5*	0
Concrete	1030	9	8*	1
Treasury	1049	16	16*	0
Mortgage	1049	16	16*	0
Ele-2	1056	5	5*	0
Friedman	1200	6	6*	0
Plastic	1650	3	3*	0
Quake	2178	4	3*	1

* representa el tipo de la variable de salida de la base de datos

4.2.2. Análisis comparativo

Con el fin de evaluar el desempeño del método propuesto (VQNN-PSO) se realizó un análisis comparativo entre diferentes sistemas de regresión donde se incluyeron métodos basados en estadística, vecinos más cercanos, vectores de soporte y árboles de decisión.

Métodos basados en vecinos más cercanos

Los métodos de regresión k-NN y Fuzzy-NN se incluyeron para establecer una comparación que permita comprobar que el uso de la teoría de conjuntos difusos-aproximados mejora el desempeño de los sistemas de regresión. De otro modo se incluyó el algoritmo de VQNN utilizando cuantificadores difusos con parámetros fijos para analizar si la optimización de los cuantificadores difusos por medio de VQNN-PSO mejora el desempeño.

La elección del número de vecinos más cercanos en estos algoritmos es una tarea muy importante para garantizar un buen desempeño y debido a que VQNN-PSO incluye la elección del número de vecinos más cercanos durante su proceso de entrenamiento, es necesario garantizar que la diferencia en el desempeño no se deba a una mala elección del número de vecinos, por ello en los algoritmos de k-NN, Fuzzy-NN y VQNN se seleccionó el número de vecinos que minimice el MAPE de una validación *leave one out*, este valor se encontró por medio de una búsqueda exhaustiva entre uno y el número de muestras en el conjunto de entrenamiento sin superar 100 vecinos.

Métodos basados en vectores de soporte

La Regresión por Vectores de Soporte (SVR) es una de las técnicas más exploradas en el estado del arte y presenta algunos de los mejores resultados, es por esto que incluimos esta técnica en el análisis comparativo. Según el planteamiento del problema de optimización podemos encontrar dos tipos de SVR las ϵ-SVR y las ν-SVR, ambos métodos fueron utilizados siguiendo el procedimiento descrito en las secciones 3.5 y 3.6 respectivamente.

- **SVR-PSO**

Con el fin de garantizar igualdad en el proceso de selección de parámetros se decidió incluir un ν-SVR optimizada por medio de PSO, de este modo el problema de optimización utilizado para seleccionar los parámetros fue:

$$\begin{aligned} \min_{C,\nu,\sigma} \quad & \text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \widehat{y_i}|}{y_i} \\ \text{sujeto a:} \quad & 0 < C < 1000 \\ & 0 < \nu \leq 1 \\ & 0 < \sigma \leq 100 \end{aligned} \tag{4-6}$$

La implementación del algoritmo sigue el mismo esquema metodológico propuesto para VQNN-PSO descrito en la figura **4-1**, en este caso los bloques correspondientes a VQNN

fueron reemplazados por SVR, se modificaron las restricciones y los parámetros. Se utilizó el método Reflejar/Absorber para manejar las restricciones del espacio de búsqueda.

Otros métodos

Diferentes análisis comparativos entre métodos de regresión para aplicaciones específicas [74, 54, 77, 58], muestran que los métodos de árboles de decisión CART y Regresión Lineal (LinearLMS) y Polinómica (PolLMS) presentan un desempeño adecuado, por lo cual se incluyeron en el análisis de este trabajo.

4.3. Resultados y discusión

La Tabla **4-4** reporta el MAPE promedio durante el proceso de validación cruzada con 10 particiones, obtenido por cada uno de los métodos evaluados, a lo largo de las 24 bases de datos. Observando el promedio del desempeño de cada método se aprecia que SVR-PSO presenta el error más bajo, mientras que VQNN-PSO se encuentra en el segundo lugar.

Tabla 4-4: Error de los métodos de regresión evaluados

Base de datos	k-NN	Fuzzy-NN	VQNN	VQNN-PSO	SVR-PSO	ν-SVR	ϵ-SVR	Cart	linearLMS	PolLMS
Baskball	**15,99**	16,39	16,29	16,45	16,63	22,04	22,18	26,35	17,85	18,48
bodyFat	18,03	18,90	18,91	18,24	**1,44**	1,90	4,11	5,26	3,62	6,25
cpu	31,36	28,89	28,92	**27,73**	30,79	56,01	1511,53	65,56	77,82	52,07
ele-2	**5,79**	5,83	5,83	5,83	10,49	55,89	560,73	33,97	17,37	14,49
autoMPG6	7,70	7,53	7,53	7,33	**7,33**	9,74	8,99	11,12	12,46	10,36
laser	14,01	14,01	12,85	12,54	**6,12**	43,11	133,33	12,12	68,38	29,20
stock	1,04	1,02	1,02	1,02	**0,98**	2,22	1,91	1,66	3,99	1,96
autoPrice	10,54	10,23	10,25	**10,04**	14,23	22,31	114,84	14,65	16,11	40,10
dee	10,96	10,83	10,71	10,55	**9,79**	12,40	12,58	15,02	11,88	11,54
machineCPU	33,41	29,76	29,77	**29,18**	31,67	51,27	1.510,27	63,59	77,91	51,25
Schlvote	47,65	**40,91**	41,61	52,90	73,67	51,19	644,15	57,32	128,76	569,04
Veteran	185,20	182,31	173,97	164,29	**162,83**	223,68	2337,66	491,36	420,35	511,40
Longley	1,19	1,17	1,17	1,13	1,98	3,57	3,65	1,97	**0,52**	1,10
autoMPG8	8,00	7,60	7,61	7,37	**7,22**	10,51	11,87	11,53	11,80	8,85
pollution	3,37	3,28	3,26	3,07	**2,55**	4,75	5,04	4,68	4,00	7,95
baseball	59,25	57,12	56,93	**55,20**	63,91	88,47	729,64	83,32	100,66	102,69
concrete	23,56	24,73	24,70	24,40	**11,47**	23,92	48,20	23,75	31,51	21,51
ele-1	25,80	25,12	25,23	**24,64**	25,08	78,76	290,72	35,02	27,05	26,44
mortgage	1,04	1,00	1,00	0,97	**0,42**	0,92	0,96	1,60	1,30	1,28
fruitFly	89,81	90,23	87,44	86,66	**84,02**	88,29	222,39	109,20	102,80	112,05
quake	2,46	2,45	2,42	2,36	2,36	**2,29**	2,30	2,56	2,45	2,45
plastic	8,75	8,78	8,71	8,42	**8,11**	8,75	8,83	18,69	8,84	8,79
treasury	1,22	1,17	1,17	**1,14**	1,16	1,71	1,69	1,91	2,11	2,12
friedman	12,76	12,42	12,44	12,26	**7,51**	15,96	14,76	19,61	18,23	11,32
Promedio	25,79	25,07	24,57	24,32	**24,24**	36,65	341,76	46,33	48,66	67,61

Realizando una prueba estadística no paramétrica (Test de Friedman), para comprobar si existe una diferencia estadística entre los desempeños obtenidos por los métodos evaluados, se obtuvo un $p = 9{,}86 \times 10^{-19}$, lo que indica que existe una clara diferencia estadística en el desempeño de los métodos evaluados. De otro modo el Ranking de medias del test de Friedman muestra que VQNN-PSO es el método que brinda el mejor desempeño, seguido por SVR-PSO (ver Tabla **4-5**). VQNN, Fuzzy-NN y k-NN siguen los resultados mostrando que la teoría de conjuntos difusos aproximados mejora el desempeño de Fuzzy-NN y a su vez este mejora el enfoque clásico del k-NN.

Tabla 4-5: Ranking de medias test de friedman

Base de datos	**ranking**
VQNN-PSO	**2.67**
SVR-PSO	2.71
VQNN	3.94
Fuzzy-NN	4.04
k-NN	4.81
ν-SVR	6.46
PolLMS	6.83
Cart	7.58
linearLMS	7.63
ϵ-SVR	8.33

Además un análisis de comparaciones múltiples por medio del método de diferencia mínima significativa, muestra que si bien VQNN-PSO presenta el mejor desempeño (Según el ranking de medias de Friedman), esta diferencia no es estadísticamente significativa frente a k-NN, Fuzzy-NN, VQNN y SVR-PSO (Tabla **5-4**).

Analizando el comportamiento de los algoritmos de acuerdo al rendimiento en las bases de datos (ver tabla **4-7**), en esta se observa que VQNN-PSO presentó el mejor desempeño en el 29.2 % de las bases de datos, mientras que SVR-PSO en el 50 % de los casos.

Analizando los resultados de acuerdo al tipo de variable de entrada, se encontró que para bases de datos con entradas mixtas SVR-PSO y VQNN-PSO presentaron el mejor desempeño en el mismo número de bases de datos, es decir en 33.3 % cada uno, mientras que para bases de datos con entradas continuas SVR-PSO fue el mejor en el 61.5 % y VQNN-PSO en el 27.3 %, solo se contaban con dos bases de datos cuyas entradas fueran enteras los resultados muestran que en una de ellas VQNN-PSO tuvo el mejor desempeño y en la restante fue SVR-PSO.

Tabla 4-6: Análisis de comparaciones múltiples de VQNN-PSO contra los demás métodos evaluados

Método	**p-Valor**
SVR-PSO	1
Fuzzy-NN	$0,86$
VQNN	$0,91$
k-NN	0,29
ν-SVR	$6{,}02 \times 10^{-4}$
PolLMS	$8,09 \times 10^{-5}$
Cart	$9,32 \times 10^{-7}$
LinearLMS	$7,35 \times 10^{-7}$
ϵ-SVR	$1,30 \times 10^{-7}$

Tabla 4-7: Número de bases de datos en las que cada método se desempeñó mejor

Metodo	**Tipo de entrada**			**Tipo de salida**	
	Mixta	**Continua**	**Entera**	**Continua**	**Entera**
k-NN	0	2	0	2	0
Fuzzy-NN	1	0	0	0	1
VQNN-PSO	3	3	1	5	2
SVR-PSO	3	8	1	11	1
ν-SVR	1	0	0	1	0
LinearLMS	1	0	0	0	1

Finalmente, respecto al tipo de variable de salida se evidencia que variables continuas SVR-PSO fue el mejor método, presentando el mejor desempeño en 57.9 % de las bases de datos mientras que VQNN-PSO fue el mejor en el 26.3 %. Para salidas enteras se aprecia que VQNN-PSO fue el mejor enfoque en el 40 % de las bases mientras que SVR-PSO en el 20 %.

5 Caso de estudio: Predicción de la calidad del aire en el Valle de Aburrá

5.1. Materiales y métodos

5.1.1. Base de datos

La orografía colombiana cuenta con valles interandinos o depresiones tectónicas que separan las cordilleras, los cuales suelen estar acompañados de ríos que los recorren y a su vez generan. En el departamento de Antioquia, Colombia, se encuentra ubicado el Valle de Aburrá y según lo prescribió la Gobernación de Antioquia, es una subregión ubicada en la cordillera central de los Andes, el cual comprende diez municipios: Barbosa, Girardota, Copacabana, Bello, Medellín, Envigado, Itagüí, Sabaneta, La Estrella y Caldas.

La topografía que está contenida en esta zona se caracteriza por ser irregular y pendiente. El valle de aburra tiene un ancho promedio de 30 km, sin embargo en su parte más ancha puede alcanzar 80 o 90 km (ver **5-1**); así mismo, el valle es recorrido de forma longitudinal por el río Medellín. La orografía, topografía, nivel de humedad y temperatura en la región genera que dificultades en la circulación de la masa de aire en el valle e impide la dispersión de los contaminantes.

Para este estudio, se utilizó los datos suministrados por el Laboratorio de Calidad del Aire de la Universidad Nacional de Colombia, el cual es el encargado de operar la red de monitoreo de calidad propiedad del Área Metropolitana de Valle de Aburrá. Los datos usados corresponden a cinco estaciones de monitoreo: Museo de Antioquia (MED-MANT), Universidad nacional – Núcleo el Volador (MED-UNNV), Universidad San Buenaventura (BEL-USBV), Casa de Justicia municipio de Itagüí (ITA - CJUS) y Liceo Concejo de Itagüí (ITA-CONC). Los datos corresponden a mediciones obtenidas entre el primero de enero de 2013 y el 31 de diciembre del mismo año. Los datos incluyen mediciones hora a hora de seis variables meteorológicas y seis contaminantes del aire, es importante aclarar que estos contaminantes no se miden de forma simultánea en todas las estaciones, su medición se realiza en combinaciones de tres y cuatro contaminantes, permitiendo obtener entre espacios de representación entre 10 y 9 variables por estación y un total de 8760 muestras.

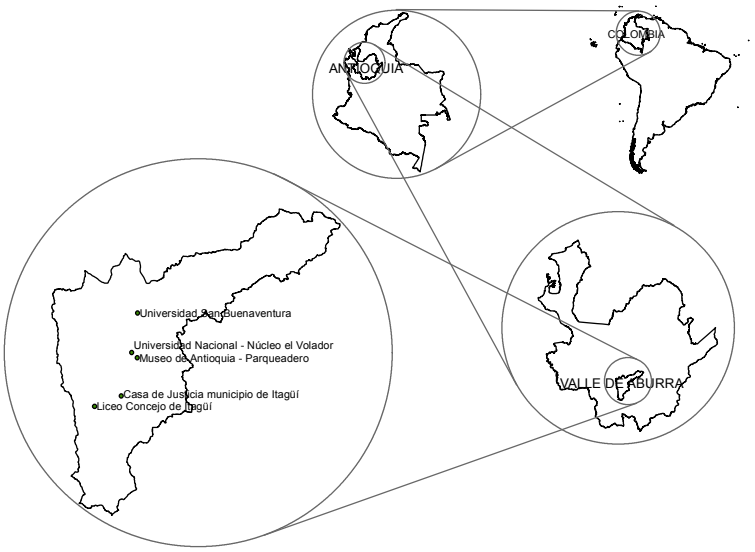


Figura 5-1: Valle de Aburrá

Descripción de la estaciones de monitoreo

El municipio de Bello, cuenta con una estación de monitoreo de la calidad del aire, la cual se encuentra instalada en la Universidad San Buenaventura (BEL-USBV), en las coordenadas 6,330° latitud norte y 75,568° longitud oeste. En esta estación se realizan mediciones de contaminantes como NO_2, O_3 y PM_{10}. Está clasificada según el Área Metropolitana del Valle de Aburrá como de Fondo Urbano, cuyo objetivo es conferir información acerca de los contaminantes generados en núcleos urbanos pero alejados de vías, alto flujo vehicular o asentamientos industriales.

La ciudad de Medellín, cuenta con una estación de monitoreo de la calidad del aire llamada Universidad Nacional de Colombia, Núcleo el volador (MED-UNNV), la cual se encuentra ubicada en las coordenadas 6,263° latitud norte y 75,577° longitud oeste. En esta estación, se realizan mediciones a contaminantes como $PM_{2,5}$, NO_2 y O_3. La estación corresponde según la clasificación de las estaciones de la Red de Calidad del Aire del Valle de Aburrá a una estación de Fondo Urbano, con el objetivo de arrojar información acerca de los contaminantes generados en núcleos urbanos pero alejados de vías con alto flujo vehicular o asentamientos industriales.

La estación Museo de Antioquia (MED-MANT), se encuentra ubicada en el centro de la ciudad de Medellín en las coordenadas 6,252° latitud norte y 75,569 longitud oeste. En esta estación se realizan mediciones de $PM_{2,5}$, NO_2, O_3 y PM_{10}. Según la clasificación de las estaciones definida por el Área Metropolitana del Valle de Aburrá, corresponde a categoría urbana, la cual tiene como objetivo hacer seguimiento en áreas con altas concentraciones de

emisiones vehiculares, por ser un sitio donde confluyen rutas de servicio público colectivo intermunicipal.

En el municipio de Itagüí, se encuentra ubicada la estación de monitoreo de la calidad del aire Casa de Justicia (ITA-CJUS) en las coordenadas 6,185° latitud norte y 75,597° longitud oeste. En esta estación se realizan mediciones de material $PM_{2,5}$, NO y NO_2. Esta estación según Área Metropolitana del Valle de Aburrá está categorizada como urbana, con el objetivo de hacer seguimiento en áreas con altas concentraciones de emisiones vehiculares, por ser un sitio donde confluyen rutas de servicio público colectivo intermunicipal. De igual modo, realiza mediciones meteorológicas de presión atmosférica, temperatura, velocidad y dirección del viento, radiación solar y humedad relativa.

La estación Colegio Concejo de Itagüí (ITA-CONC) definida por la Red de Calidad del Aire, se encuentra ubicada en las coordenadas l6,168° latitud norte y 75,644° longitud oeste, donde se realiza medición de emisiones de $PM_{2,5}$, O_3 y PM_{10}. Esta estación tiene categoría suburbana, el objetivo es conocer los niveles de contaminación que se presentan en las zonas de ladera del valle, entornos con particularidad en circulación de los vientos y comportamiento como sumidero de contaminación.

Todas las estaciones se encuentran dotadas con una estación meteorológica para medir variables como presión atmosférica, temperatura ambiente, precipitación, radiación solar, humedad relativa, velocidad y dirección del viento. La altura de toma de muestra de los contaminantes y de las variables meteorológicas se realiza a una altura aproximada de 10 metros sobre el nivel del suelo. Los equipos operan con métodos aprobados por la Agencia de Protección Ambiental de los Estados Unidos (EPA), de tecnología automática que garantiza una frecuencia de toma de datos horaria.

5.1.2. Caracterización

Inicialmente los datos de cada estación de monitoreo fueron tratados para detectar mediciones ausentes y valores atípicos. Las bases de datos utilizadas contienen una gran cantidad de mediciones ausentes causadas por mantenimiento y daños de los sensores en las estaciones de monitoreo. En este contexto decidimos no reemplazar estos valores debido a que este proceso genera un aumento en la incertidumbre de los datos. De otro modo los datos atípicos aparecen de forma repentina o son generados por eventos excepcionales como incendios, tormentas de polvo días festivos entre otros, esto puede afectar o deteriorar el desempeño de sistemas de predicción basados en estadística o modelos matemáticos [18]. En este trabajo, la detección de mediciones atípicos se realizó por medio de un análisis de diagramas de cajas de cada una de las variables, la longitud de los bigotes fue de 1.5 veces el rango intercuartil,

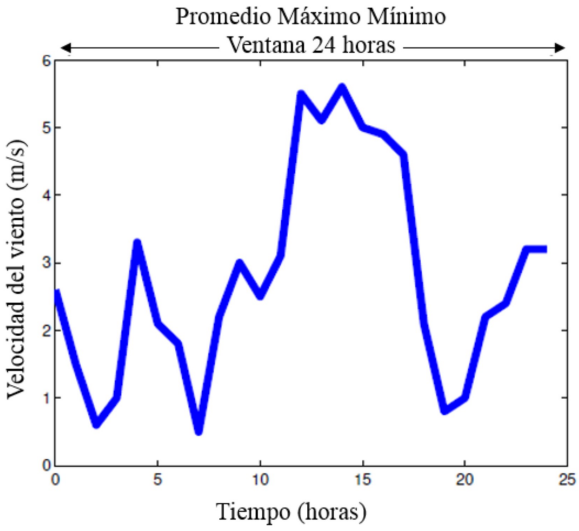
los datos por fuera los bigotes fueron tratados como mediciones ausentes.

Con el fin de representar de forma efectiva el comportamiento de series temporales de concentración de contaminantes, para predecir su concentración con 24 horas de anticipación en diferentes puntos del Valle de Aburrá, se usaron características derivadas del análisis temporal de algunas variables meteorológicas y de la concentración de los contaminantes. Las características obtenidas pueden dividirse en cuatro grupos: Variables Meteorológicas Horarias (VMH), Variables Meteorológicas Temporales (VMT), Concentración Horaria de Contaminantes (CHC) y Concentración Temporal de Contaminantes (CTC).

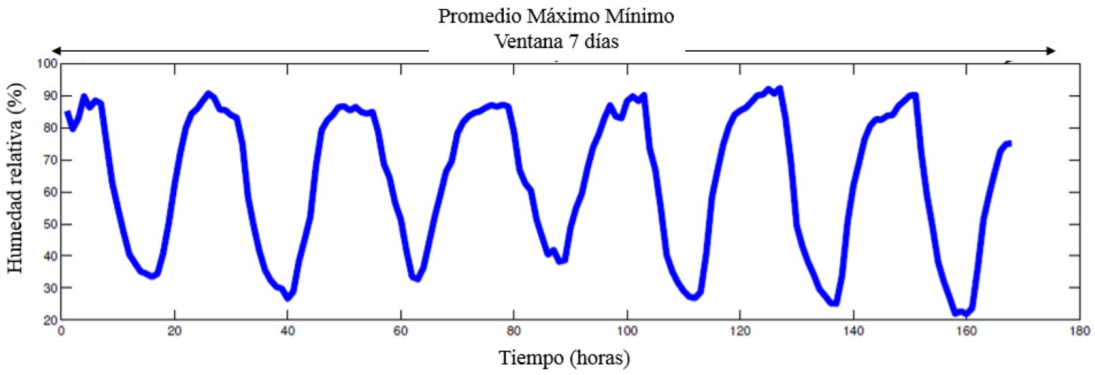
El grupo VMH, corresponde a las últimas mediciones de las variables meteorológicas en las estaciones de monitoreo, las cuales corresponde a la Velocidad del Viento (VV), Dirección del Viento (DV), Temperatura del Aire (TA), Humedad Relativa (HR) y Radiación Solar (RS). En este orden de ideas el grupo CHC responde a las últimas mediciones de las concentración de los contaminantes estudiados, que corresponde a NO, NO_2, $PM_{2,5}$, PM_{10} y O_3.

Variables Meteorológicas Temporales (VMT)

El comportamiento de las variables meteorológicas responde a los patrones climáticos de las regiones, por lo cual están relacionadas con el momento del año y los microclimas [11], estas variables contribuyen de manera significativa con episodios de polución del aire [2]. Por lo anterior, fue necesario incluir variables que le permitan al sistema evaluar el estado meteorológico en diferentes intervalos de tiempo previos al momento de la predicción, para ello, empleamos dos intervalos de tiempo 24 horas y 7 días. En cada uno de estos intervalos calculamos el máximo, mínimo y promedio de cada una de las variables meteorológicas (ver Figura **5-2**). Es importante aclarar que durante la noche la RS es cero, como resultado el mínimo de esta variable durante los intervalos estudiados será siempre cero, por lo cual para variable solo se calculó el máximo y el promedio.



(a) Velocidad del viento durante 24 horas



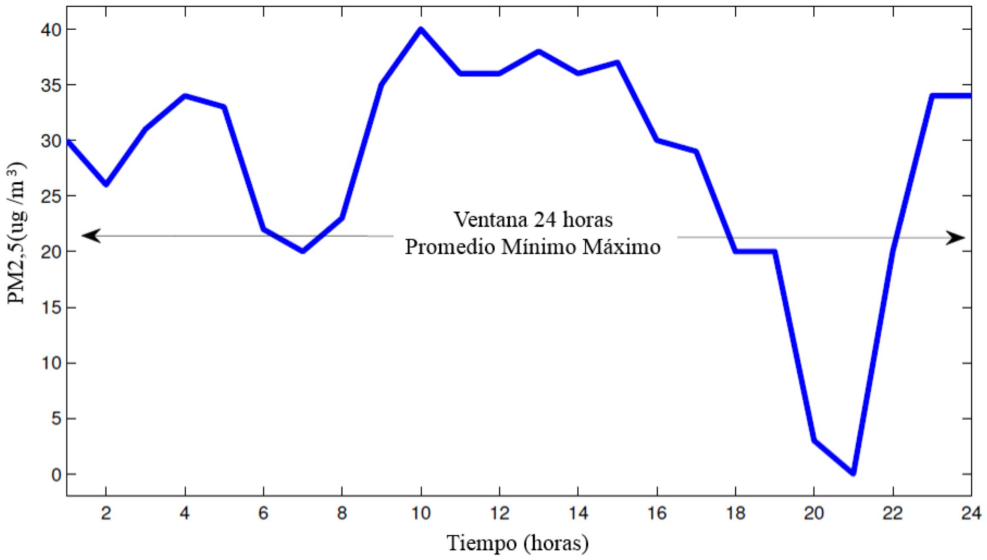
(b) Humedad relativa durante 7 semanas

Figura 5-2: Caracterización de Variables Meteorológicas Temporales (VMT)

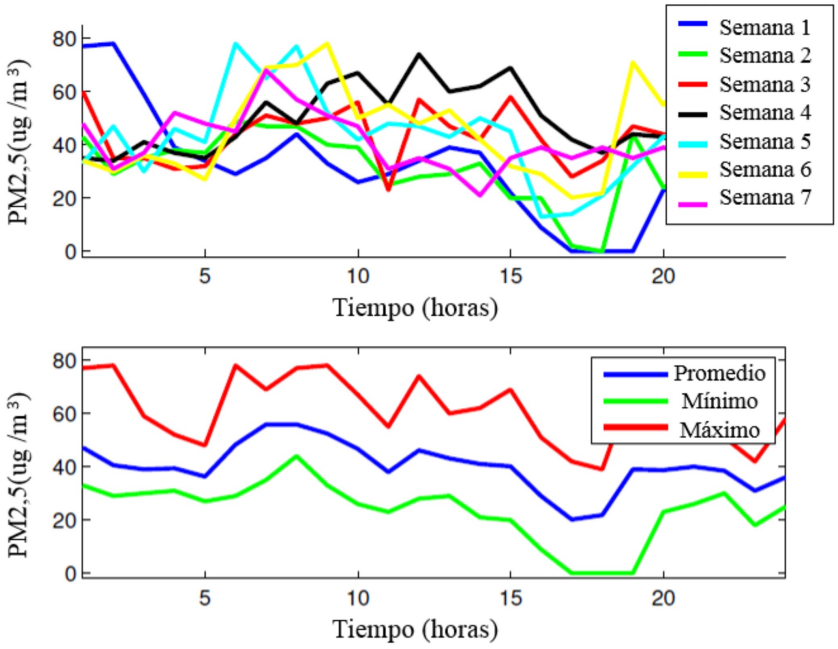
Concentración Temporal de Contaminantes (CTC)

La concentración de contaminantes del aire es afectada principalmente por la actividad humana y la meteorología [12, 76], por esta razón dividimos el análisis temporal de los contaminantes en dos bloques. Con el fin de analizar las tendencias recientes en la concentración de los contaminantes y así poder capturar el efecto de la meteorología, el primer bloque de características consiste en el máximo, mínimo y promedio de la concentración de cada uno de los contaminantes durante las 24 horas previas al momento de realizar la predicción. El segundo bloque está compuesto de la concentración promedio, máxima y mínima de cada contaminantes durante el mismo día y hora en las últimas 7 semanas (e.g. para predecir la concentración de PM_{10} un lunes a las 2 p.m, nosotros calculamos la concentración promedio, máxima y mínima de cada contaminante los lunes a las 2 p.m durante las

7 semanas previas), este grupo de variables pretende capturar el impacto de la actividad humana sobre la concentración de contaminantes de acuerdo con el día de la semana y la hora del día, debido a factores como el tráfico, días no laborales, entre otras (ver Figura **5-3**).



(a) 24 horas de concentración de $PM_{2,5}$



(b) Concentración de $PM_{2,5}$ el mismo día durante 7 semanas

Figura 5-3: Caracterización temporal de la concentración de contaminantes

La Figura **5-4** muestra el proceso de caracterización a partir de las mediciones realizadas por la estación de monitoreo, en total se obtuvieron 68, es importante aclarar que no todas están disponibles de forma simultánea en las cinco estaciones de monitoreo, debido a que los contaminantes estudiados no se miden en todas las estaciones, su medición se realiza en combinaciones de tres y cuatro contaminantes. En la estación Bello-USBV solo hay disponible 61 variables, debido a que solo se mide NO, NO_2, PM_{10} y O_3, en MED-UNNV también hay 61 características, ya que se miden cuatro contaminantes ($PM_{2,5}$, NO, NO_2 y O_3), en las estaciones restantes se miden 3 contaminantes, por lo cual se cuenta con 54 variables.

El proceso de caracterización propuesto puede ser sensible a mediciones ausentes debido a que estas se propagaran a los intervalos de tiempo donde se calcularan las características. Para ello, decidimos calcular las características máximos, mínimos y promedios usando la información disponible en los intervalos de tiempo analizados, este proceso se realizó tanto para las variables meteorológicas como para los contaminantes.

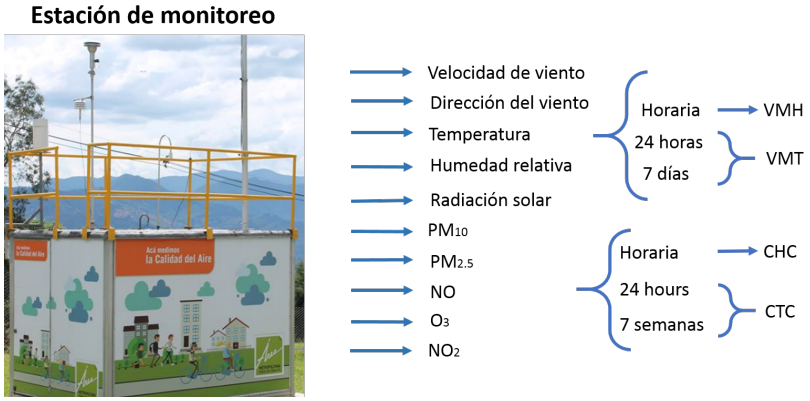


Figura 5-4: Proceso de caracterización a partir de las mediciones arrojadas por las estaciones de monitoreo

5.1.3. Selección de variables y evaluación de predicción

En este trabajo realizamos la predicción de los contaminantes por medio de VQNN-PSO, usando como entrada los grupos de variables propuestos, en este sentido, fue necesario evaluar cual grupo de características ofrece las mejores capacidades predictivas.

Inicialmente creamos 11 bases de datos usando todas las posibles combinaciones entre los grupos de variables propuestos:

- VMT+CHC

- CHC+CTC
- CHC+VMT
- VMH+CTC
- VMH+VMT
- VMH+CHC
- CHC+CTC+CTC
- VMH+VMT+CTC
- VMH+CHC+CTC
- VMH+CHC+VMT
- VMH+CHC+VMT+CTC

Posteriormente, las bases de datos fueron dividas de forma aleatoria en dos conjuntos, entrenamiento (70 %) y validación (30 %). Realizamos el entrenamiento de cada uno de los métodos de predicción (SVR-PSO y VQNN) usando el conjunto de entrenamiento, luego se predijo la concentración usando los datos de prueba y se calculó la Raíz del Error Cuadrático Medio (RMSE) entre la concentración predicha y la concentración medida. Luego usamos la prueba de Friedman para determinar cuál método ofrece el mejor desempeño y posteriormente, por medio de un análisis de comparaciones múltiples (Test de la Diferencia Mínima Significativa) se evalúo con el grupo de variables cual presenta las mejores capacidades predictivas.

En el Valle de Aburrá los patrones climáticos pueden dividirse en cuatro trimestres: Enero-Marzo, Abril-Junio, Julio-Septiembre y Octubre-Diciembre. Los periodos de Enero-Marzo y Julio-Septiembre corresponden a periodos de verano, mientras Abril-Junio y Octubre-Diciembre responde a un periodo de lluvias. Además, las vacaciones principales en Colombia tienen lugar entre Diciembre y Enero. El segundo periodo de vacaciones escolares se realiza durante Junio y Julio, durante este tiempo la mayoría de estudiantes de colegios y universidades no asisten a clases. Lo anterior evidencia la necesidad de evaluar el desempeño del sistema de forma individual en cada trimestre dada sus particularidades tanto en el comportamiento climático como en las actividades humanas.

De este modo la base de datos fue dividida en dos conjuntos Entrenamiento y Validación, pero a diferencia de la sección previa, la división no fue completamente aleatoria, en cambio seleccionamos de forma aleatoria 250 horas continuas en cada uno de los trimestres, para

conformar el conjunto de Prueba y los datos restantes corresponden al conjunto de Entrenamiento. Entrenamos un modelo para predecir la concentración de cada contaminante a lo largo de las cinco estaciones de monitoreo, la evaluación del desempeño se realizó con base en el Error Escalado Absoluto Medio (MASE) y RMSE (ver tabla **4-1**).

5.1.4. Análisis comparativo

Finalmente, se comparó el desempeño de VQNN-PSO frente a otros algoritmos de predicción destacados en la literatura para la predicción de contaminantes del aire, en este sentido ANN y SVR destacan como uno de los métodos más usados y con mejores resultados en esta tarea. Se decidió emplear el método SVR-PSO debido a los buenos resultados obtenidos en el Capítulo 4, usando la misma formulación de la ecuación 4-6. Por su parte, para las ANN se usó un perceptrón multicapa con 30 capas ocultas y se empleó el algoritmo de Levenberg–Marquardt para su entrenamiento. En ambos métodos se usó como entrada el mejor subconjunto de característica encontrado. La comparación se realizó con base en el RMSE y usando un protocolo de validación cruzada 70/30.

5.2. Resultados y discusión

5.2.1. Caracterización

Con el fin de generar un sistema capaz de predecir la concentración de contaminantes del aire en el Valle de Aburrá, propusimos un esquema de caracterización basado en el comportamiento temporal de variables meteorológicas y de la concentración de contaminantes en diferentes intervalos de tiempo, en total se calcularon cuatro grupos de variables VMH, CHC, VMT y CTC.

Durante el pre-procesamiento se detectó que en las 5 estaciones de monitoreo en promedio el 37.6 % de los datos presentaban mediciones ausentes, después de calcular los cuatro grupos de variables el porcentaje de mediciones ausentes aumentó a un 47.9 %, alrededor de un 10.3 % (Tabla **5-1**). Este incremento se debe a que en este estudio, solo se utilizó los datos del 2013 y teniendo en cuenta que para calcular las variables CTC es necesario contar con información previa de 7 semanas, solo se pudieron realizar predicciones a partir de la cuarta semana de febrero, por lo cual este proceso genera por si solo una pérdida de mediciones de alrededor del 13.4 %, bajo este punto de vista, se esperaba un aumento en la pérdida de mediciones alrededor de este valor, pero los resultados evidencian una pérdida inferior, esto se debe a que durante las primeras siete semanas del año existen mediciones ausentes y además la estrategia de caracterización propuesta aumentó la robustez a estos valores.

Tabla 5-1: Porcentaje de mediciones perdidas durante el proceso de caracterización

Estación	Antes (%)	Despues (%)
BEL-USBV	35	44.6
MED-MANT	37.2	44.9
MED-UNNV	36.3	46.7
ITA-CJUS	33.2	45.4
ITA-CONC	46	57.7

5.2.2. Selección de variables y evaluación de predicción

Para definir el mejor grupo de variables, procedimos a realizar la predicción de los cinco contaminantes estudiados ($PM_{2,5}$, PM_{10}, NO, NO_2 y O_3) en las cinco estaciones de monitorio evaluadas (BEL-USBV, MED-MANT, MED-UNNV, ITA-CJUS y ITA-CONC), usando como entrada todas combinaciones posibles entre los cuatro grupos de variables propuestas. Usamos una estrategia de validación cruzada 70/30 y el desempeño del sistema se evaluó respecto al RMSE.

La Tabla **5-2** muestra el RMSE de la predicción obtenido por VQNN-PSO. En este contexto, aplicamos el test de Shapiro-Wilik, con el fin determinar la normalidad de los datos ($P = 0{,}041$), mostrando que los datos no son normales, por esta razón decidimos utilizar la prueba estadística no paramétrica de Friedman para determinar si el grupo de características usado afecta el desempeño de VQNN-PSO, un p-valor de $6{,}04^{-12}$, muestra que las variables de entrada tiene un efecto considerable en el desempeño de VQNN-PSO. Revisando el ranking de medias resultante del test de Friedman (Tabla **5-3**), se aprecia que el grupo de características que presenta el menor error es **CHC+VMT+CTC**.

Tabla 5-3: Ranking de medias test de Friedman

Variable	Mean
CHC+VMT+CTC	2.91
VMH+VMT+CTC	3.59
VMH+CHC+VMT+CTC	4
VMT+CTC	5.03
CHC+CTC	5.06
VMH+CHC+CTC	5.79
VMH+CTC	6.85
VMH+CHC+VMT	6.91
CHC+VMT	7.5
VMH+VMT	7.97
CHC+VMH	10.38

Tabla 5-2: RMSE de predicción usando VQNN-PSO

Estación	Contaminante	VMT+CTC	VMT+CTC VMH+CHC	CHC+CTC	VMH+CHC CTC	CHC+VMT CTC	VMH+CTC	VMH+VMT CTC
BEL-USBV	**PM_{10}**	**7,39**	7,69	8,35	8,54	7,62	8,35	7,60
	NO	3,03	2,78	**2,54**	2,68	2,63	**2,54**	2,81
	NO_2	7,25	**6,59**	7,14	7,32	6,62	7,14	7,05
	O_3	11,27	**10,02**	11,05	10,51	13,11	11,05	10,48
ITA-CJUS	**$PM_{2,5}$**	8,27	8,51	8,47	9,20	8,17	8,47	**8,12**
	NO	8,19	**8,04**	8,08	8,23	7,46	8,08	8,30
	NO_2	10,42	10,34	10,91	10,53	**9,80**	10,91	9,89
ITA-CONC	**PM_{10}**	12,75	12,35	13,41	13,11	12,60	13,41	**12,31**
	$PM_{2,5}$	7,73	8,02	8,35	8,53	7,73	8,35	**7,65**
	O_3	15,68	15,59	15,41	15,80	14,90	15,41	**14,63**
MED-MANT	**$PM_{2,5}$**	9,30	9,62	9,33	9,61	**9,22**	9,33	9,67
	NO	16,22	15,90	16,09	16,33	**15,54**	16,09	17,31
	NO_2	8,66	8,98	9,44	10,02	**8,62**	9,44	8,65
MED-UNNV	**$PM_{2,5}$**	6,94	7,07	7,24	7,35	**6,72**	7,24	7,03
	NO	14,75	14,08	**13,28**	13,39	13,40	**13,28**	14,73
	NO_2	9,83	9,53	10,07	9,62	9,72	10,07	**9,39**
	O_3	11,83	9,86	9,67	**9,39**	10,08	9,67	9,78

Se omite los resultados de CHC+VMT, CHC+VMH, CHC+VMH+VMT y VMH+VMT, dado que presentaban errores muy altos

Además, conducimos una prueba de comparaciones múltiples usando el método de Mínima Diferencia Significativa (LSD) entre CHC+VMT+CTC y el resto de combinaciones (Tabla **5-4**), los resultados muestran VQNN-PSO usando CHC+VMT+CTC tiene un desempeño superior que usando el resto de combinaciones de variables, sin embargo en la prueba LSD revela que no existe una diferencia estadísticamente significativa con respecto a VMH+VMT+CTC, VMH+CHC+VMT+CTC, VMT+CTC y CHC+CTC. El grupo de características CTC, esta presenta simultáneamente en las 5 combinaciones con el mejor desempeño, lo que muestra

que el esquema de caracterización temporal propuesto, refleja de forma efectiva el comportamiento de los contaminantes bajo estudio, por el contrario, usando solo las características horarias (CHC y VMH) el sistema alcanza el desempeño más bajo.

Tabla 5-4: Resultados de la prueba LSD comparando CHC+VMT+CTC vs el resto de combinaciones

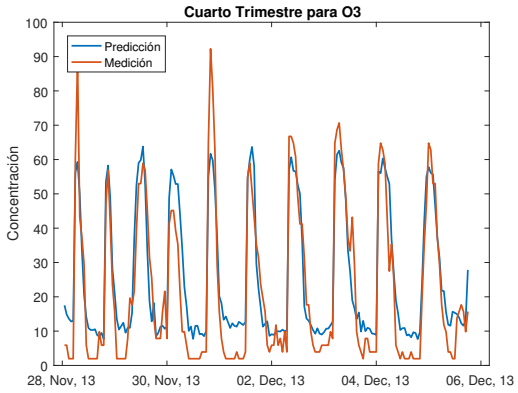
Variables	**p-valor**
VMH+VMT+CTC	1
VMH+CHC+VMT+CTC	1
VMT+CTC	0.74
CHC+CTC	0.73
VMH+CHC+CTC	0.28
VMH+CTC	0.02
VMH+CHC+VMT	0.02
CHC+VMT	0
VMH+VMT	0
CHC+VMH	0

Los resultados obtenidos sugieren que usar como entrada características que evalúan las tendencias actuales y pasadas de la concentración de contaminantes durante las 24 horas previas y durante las últimas 7 semanas a la misma hora y día que se pretende realizar la predicción, le permite identificar la influencia de la actividad humana, así como posibles condiciones atmosféricas que influencia la concentración o dispersión de contaminantes, de otro modo, las variables VMT agregan información sobre la meteorología actual, a través de una ventana de 24 horas, la ventana de 7 semanas le permite al sistema la capacidad de detectar transiciones entre periodos climáticos y brinda mayor robustez ante eventos meteorológicos atípicos.

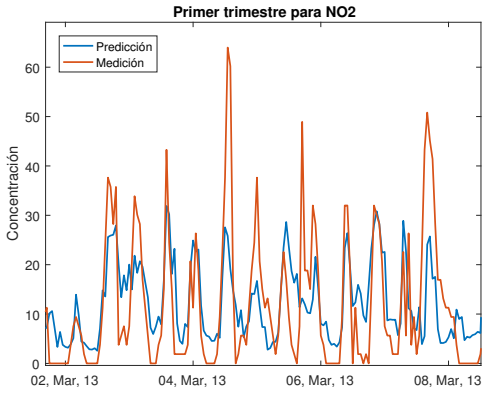
Tabla 5-5: Desempeño de VQNN-PSO en la predicción por trimestres

Estación	Pollutant	MASE				RMSE			
		T1	**T2**	**T3**	**T4**	**T1**	**T2**	**T3**	**T4**
BEL-USBV	PM_{10}	1,27	1,38	1,25	1,01	0,38	3,10	1,81	3,31
	NO	0,96	0,93	0,91	0,97	0,88	0,93	0,54	1,13
	NO_2	1,38	1,04	1,07	1,16	0,63	4,12	1,27	0,52
	O_3	0,88	0,83	1,03	0,88	1,51	0,90	0,37	2,57
ITA-CJUS	$PM_{2,5}$	1,26	1,19	1,25	1,14	1,94	4,44	1,29	1,61
	NO	1,01	0,82	1,02	1,08	0,83	2,71	1,82	5,26
	NO_2	1,19	1,24	1,47	1,13	2,10	7,92	3,18	2,25
ITA-CONS	PM_{10}	1,10	0,89	1,00	1,00	1,74	1,45	1,68	2,74
	$PM_{2,5}$	1,21	0,94	1,00	0,98	0,91	2,40	0,35	2,65
	O_3	1,15	1,05	1,10	1,07	1,92	3,26	1,30	1,41
MED-MANT	$PM_{2,5}$	0,91	0,99	1,20	1,14	1,42	1,29	2,57	1,29
	NO	1,79	0,91	0,95	1,07	0,74	5,55	1,41	5,04
	NO_2	1,04	1,17	1,29	0,96	1,39	0,32	4,34	1,30
MED-UNNV	$PM_{2,5}$	1,14	1,01	1,08	1,02	3,76	2,43	0,26	1,93
	NO	0,84	1,00	1,01	1,25	4,02	7,63	5,11	10,48
	NO_2	1,16	1,03	1,01	0,97	5,69	0,57	1,42	2,55
	O_3	0,84	0,93	0,88	1,30	1,13	0,58	0,63	3,58

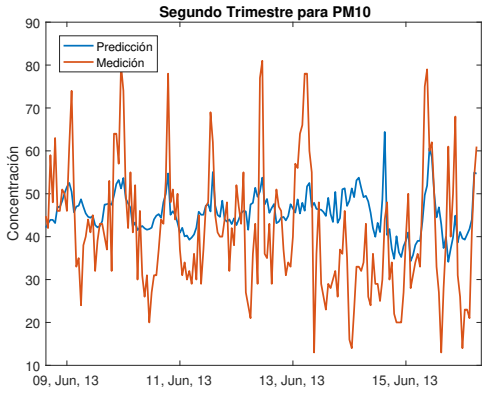
En la figura **5-5**, se muestran los gráficos de tendencia, donde se puede comparar el desempeño de las predicciones de VQNN-PSO con respecto a los valores medidos y además permite contrarrestar el desempeño del sistema en cada trimestre del año, usando como entrada el mejor grupo de características (CHC+VMT+CTC). De las gráficas podemos observar que el sistema tiene un buen desempeño, el cual es similar en cada trimestre del año. Para confirmar esta observación se usó el test de Friedman, con el fin de determinar si existe una diferencia estadística en el desempeño del método por trimestre, usando como medidas de desempeño el MASE y el RMSE (Tabla **5-5**), el test arrojo un a p-valor de 0.1461 para RMSE y 0.0835 MASE, mostrando que no existe una diferencia entre trimestres, lo que sugiere que el método tiene el mismo desempeño en estaciones húmedas y secas.



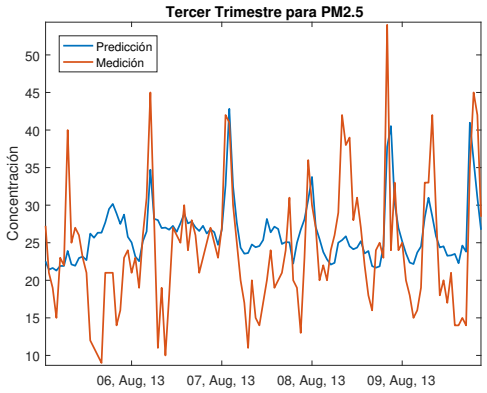
(a) O_3



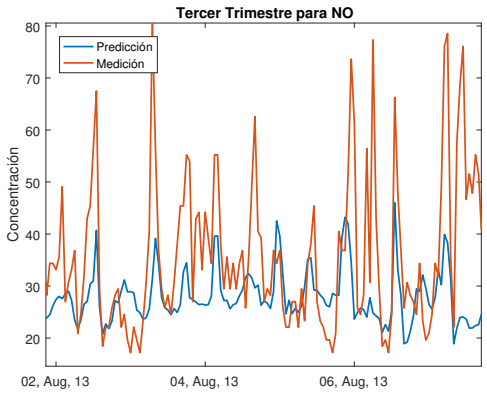
(b) NO_2



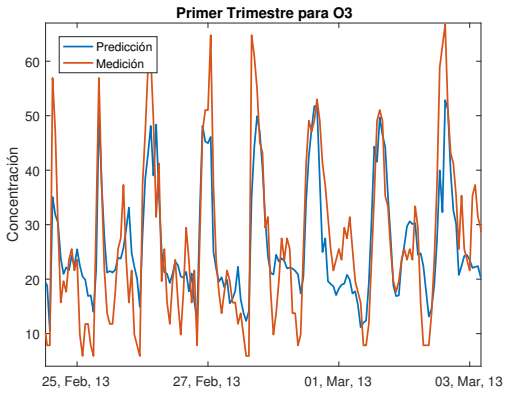
(c) PM_{10}



(d) $PM_{2,5}$



(e) NO



(f) O_3

Figura 5-5: Tendencia de concentración horaria de varios contaminantes del aire en diferentes estaciones de monitores en el Valle de Aburrá: **(a)** Octubre-Diciembre en BEL-USBV para O_3 , **(b)** Enero-Marzo en ITA-CJUS para NO_2, **(c)** Abril-Junio en ITA-CONC para PM_{10}, **(d)** Agosto-Septiembre en MED-MANT para $PM_{2,5}$, **(e)** Agosto-Septiembre en MED-UNNV para NO, **(f)**Enero-Marzo en MED-UNNV para O_3

En la figura **5-5** se observa que el modelo tiende a subestimar las concentraciones de los contaminantes en los puntos de mayor concentración medida. Al igual se observa que el sistema tiene dificultades para seguir cambios abruptos en la concentración, sin embargo se observa una buena capacidad de seguir la tendencia. Para confirmar esto podemos observar los gráficos de dispersión (Figure **5-6**) donde se observa que la mayoría de los puntos están por debajo de la función identidad, sin embargo se puede observar las excelentes capacidades predictivas del sistema para O_3, además se observa un buen ajuste para NO_2 y NO.

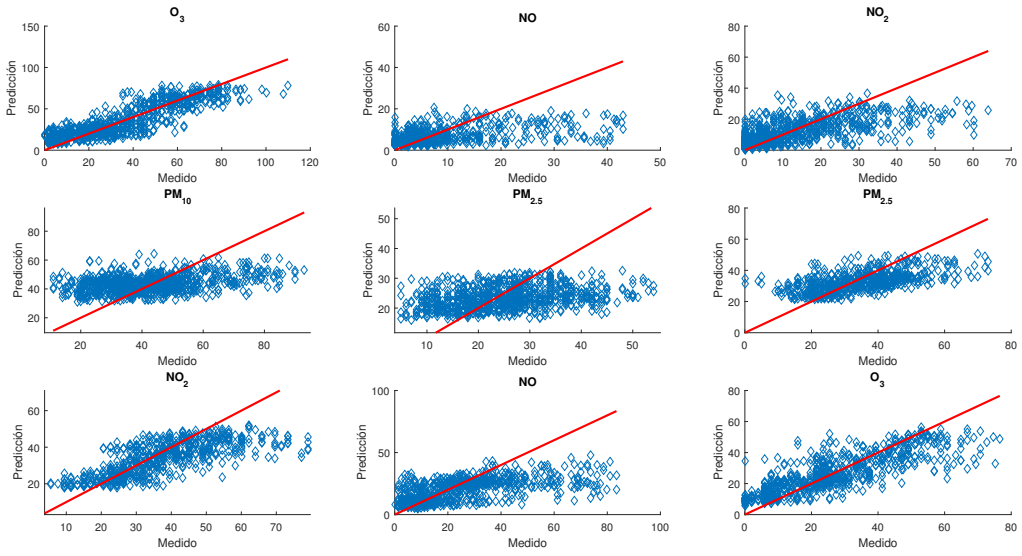


Figura 5-6: Comparación entre niveles de concentración predecidos y medidos

5.2.3. Análisis comparativo

Con el fin de comparar VQNN-PSO con respecto a SVR-PSO y ANN, debido a su relevancia en la literatura para la predicción de la calidad del aire. La Tabla **5-6**, presenta el RMSE obtenido por cada uno de los métodos comparados, usando como variable de entrada CHC+VMT+CTC, ya que fue el grupo de característica que presentó la mejor capacidad predictiva, de la tabla se puede observar que SVR-PSO presenta el mejor desempeño en la mayoría de contaminantes.

Una prueba estadística no paramétrica (test de Friedman), muestra que la diferencia en el desempeño de los métodos es estadísticamente significativo (p-valor=$5{,}37 \times 10^{-6}$), el ranking de media del test muestra que SVR-PSO es el método que presenta un menor error (μ=1.18), seguido por VQNN-PSO (μ=1.97). Finalmente, ANN fue quien presento el error más alto (μ=2.85). Además, se realizó una prueba de comparaciones múltiples (Test de diferencia mínima significativa), la cual muestra que la diferencia en el desempeño entre VQNN-PSO y SVR-PSO no es estadísticamente significativa (p-valor=0,0514), por su parte el desempeño de

Tabla 5-6: RMSE obtenido por los tres métodos comparados usando como variables de entrada CHC+VMT+CTC

Estación	Contaminante	VQNN-PSO	SVR-PSO	ANN
BEL-USBV	PM_{10}	7,60	9,56	**9,44**
	NO	2,81	**2,46**	2,7
	NO_2	7,05	**6,46**	8,32
	O_3	10,48	**9,37**	11,65
ITA-CJUS	$PM_{2,5}$	8,12	**7,79**	9,75
	NO	8,30	**6,86**	9,31
	NO_2	9,89	**9,28**	12,08
ITA-CONC	PM_{10}	12,31	**11,98**	14,61
	$PM_{2,5}$	**7,65**	7,68	9,25
	O_3	14,63	**11,99**	16,90
MED-MANT	$PM_{2,5}$	9,67	**8,56**	10,22
	NO	17,31	**13,72**	19,07
	NO_2	8,65	**8**	10,26
MED-UNNV	$PM_{2,5}$	7,03	**6,40**	8,11
	NO	14,73	**11,78**	14,73
	NO_2	9,39	**8,84**	11,86
	O_3	9,78	**8,74**	9,93

ANN fue estadísticamente inferior a VQNN-PSO y SVR-PSO (p-valor=0,0259 y $2{,}53 \times 10^{-6}$, respectivamente).

6 Conclusiones y recomendaciones

6.1. Conclusiones

Vecinos más Cercanos Vagamente Cuantificados es un método con altas capacidades predictivas, comparable con otros métodos de gran impacto en el área de la inteligencia computacional. En este trabajo se desarrolla una metodología de predicción con base en VQNN-PSO, para la optimización de los cuantificadores difusos en función de los datos utilizados, garantizando una mayor generalización en comparación con la mayoría de estrategias de predicción a lo largo de diferentes aplicaciones y tipos de datos tanto de entrada como de salida. Dicha metodología permitió además desarrollar un sistema robusto para la predicción de contaminantes en el Valle de Aburrá, que podría usarse como una herramienta de pronóstico para aplicaciones móviles y servicios web que informen a la comunidad sobre estados futuros de la calidad del aire, además de permitir a las instituciones gubernamentales evaluar normativas para mitigar impactos negativos sobre la salud publica.

La función objetivo planteada en la ecuación 4-3 usando el Error Absoluto Porcentual Medio (MAPE), permitió optimizar los parámetros de los cuantificadores difusos en función de los datos usados para el entrenamiento, además facilito obtener una comparación directa entre diferentes bases de datos, para evaluar objetivamente el desempeño del método en diferentes aplicaciones. No obstante el MAPE tiende a infinito cuando el valor a estimar tiende a cero, por lo cual es necesario aplicar una transformación lineal a los datos para evaluar correctamente el desempeño.

La metodología de predicción con base en Vecinos más Cercanos Vagamente Cuantificados optimizados por enjambre de partículas (VQNN-PSO) ofrece un marco metodológico para la optimización de los cuantificadores difusos en función del ruido e incertidumbre presente en los datos, mostrando un desempeño superior a los métodos basados en vecinos más cercanos evaluados (Regresión k-NN, Fuzzy-NN y VQNN), también superó a métodos destacados en el estado del arte como ϵ-SVR, ν-SVR y CART, sin embargo, respecto a SVR-PSO no se encontró evidencias de una mejora significativa, esto se puede explicar, gracias a que SVR-PSO maneja un menor número de parámetros libres respecto a VQNN-PSO, al igual tiene menos restricciones. El procedimiento propuesto para abordar las restricciones inherentes a los parámetros de los cuantificadores difusos, permitió al algoritmo PSO conservar su capacidad de exploración en espacios de búsqueda planos.

VQNN-PSO permitió obtener un sistema de predicción de calidad del aire, capaz de operar de manera robusta a lo largo de los periodos climáticos en el Valle de Aburrá, así como en las diferentes estaciones evaluadas. Además, el sistema muestra una gran capacidad para predecir ozono troposférico (O_3) y una tendencia a subestimar las concentraciones altas de los óxidos nitrosos y el material particulado, al igual que los métodos ANN y SVR-PSO.

6.2. Recomendaciones

En este trabajo se evaluó el uso de VQNN-PSO en tareas predictivas (regresión), es de interés evaluar el desempeño de VQNN-PSO en otras tareas como clasificación o selección de características, lo que acarrea el estudio de nuevas funciones objetivos dentro del mismo marco metodológico. Además, es de interés comparar el desempeño de la metodología tipo *wrapper* propuesta, respecto a una metodología tipo filtro, ya que estas ofrecen una mayor eficiencia computacional.

Bibliografía

[1] Ali, Maqbool ; Qamar, Ali M. ; Ali, Bilal: Data analysis, discharge classifications, and predictions of hydrological parameters for the management of Rawal Dam in Pakistan. En: *Proceedings - 2013 12th International Conference on Machine Learning and Applications, ICMLA 2013* Vol. 1, 2013. – ISBN 9780769551449, p. 382–385

[2] Alkasassbeh, Mouhammd ; Sheta, Alaa F. ; Faris, Hossam ; Turabieh, Hamza: Prediction of PM10 and TSP Air Pollution Parameters Using Artificial Neural Network Autoregressive , External Input Models : A Case Study in Salt , Jordan. En: *Middle-East Journal of Scientific Research* 14 (2013), Nr. 7, p. 999–1009. – ISBN 1990–9233

[3] An, Shuang ; Shi, Hong ; Hu, Qinghua ; Li, Xiaoqi ; Dang, Jianwu: Fuzzy rough regression with application to wind speed prediction. En: *Information Sciences* 282 (2014), p. 388–400. – ISSN 00200255

[4] Anderson, H R. ; Atkinson, Richard W. ; Peacock, Janet L. ; Sweeting, Michael J. ; Marston, Louise: Ambient Particulate Matter and Health Effects. En: *Epidemiology* 16 (2005), Nr. 2, p. 155–163. – ISBN 0000152528

[5] Arampongsanuwat, S ; Meesad, P: PM10 Prediction Model by Support Vector Regression Based on Particle Swarm Optimization. En: *Advanced Materials Research* 403-408 (2012), p. 3693–3698. – ISSN ISSN 1022–6680

[6] Basak, Debasish ; Pal, Srimanta ; Patranabis, Dipak C.: Support Vector Regression. En: *Neuronal Information Processing - Letters and Reviews* 11 (2007), Nr. 10, p. 203–224. – ISSN 2093–369X

[7] Bedoya, Julian ; Martinez, Elkin: Calidad Del Aire En El Valle De Aburrá Antioquia -Colombia Air Quality in the Aburrá Valley Antioquia-Colombia. En: *DYNA* 76 (2009), Nr. 158, p. 7–15

[8] Bochenek, B.: Problems of structural optimization for post-buckling behaviour. En: *Structural and Multidisciplinary Optimization* 25 (2003), Nr. 5-6, p. 423–435. – ISBN 1615–147X

[9] Bravo, Mercedes A. ; Fuentes, Montserrat ; Zhang, Yang ; Burr, Michael J. ; Bell, Michelle L.: Comparison of exposure estimation methods for air pollutants:

Ambient monitoring data and regional air quality simulation. En: *Environmental Research* 116 (2012), p. 1–10. – ISBN 1096–0953 (Electronic)\r0013–9351 (Linking)

[10] Brunekreef, Bert ; Holgate, Stephen T.: Air pollution and health. En: *Lancet* 360 (2002), Nr. 9341, p. 1233–1242. – ISBN 0140–6736 (Print)\n0140–6736 (Linking)

[11] Brunner, Dominik ; Savage, Nicholas ; Jorba, Oriol ; Eder, Brian ; Giordano, Lea ; Bianconi, Roberto ; Chemel, Charles ; Badia, Alba ; Balzarini, Alessandra ; Bar, Rocío ; Hirtl, Marcus ; Hodzic, Alma ; Curci, Gabriele ; Forkel, Renate ; Jim, Pedro ; Honzak, Luka ; Im, Ulas ; Knote, Christoph ; Makar, Paul ; Manders-groot, Astrid ; Pirovano, Guido ; San, Roberto ; Meijgaard, Erik V. ; Neal, Lucy ; Juan, L P. ; Tuccella, Paolo ; Werhahn, Johannes ; Wolke, Ralf ; Yahya, Khairunnisa ; Zabkar, Rahela ; Zhang, Yang ; Hogrefe, Christian ; Galmarini, Stefano: Comparative analysis of meteorological performance of coupled chemistry-meteorology models in the context of AQMEII phase 2 Wolfram Schr o. 115 (2015)

[12] Buczyńska, Anna J. ; Krata, Agnieszka ; Van Grieken, Rene ; Brown, Andrew ; Polezer, Gabriela ; De Wael, Karolien ; Potgieter-Vermaak, Sanja: Composition of PM2.5 and PM1 on high and low pollution event days and its relation to indoor air quality in a home for the elderly. En: *Science of the Total Environment* 490 (2014), p. 134–143. – ISBN 0048–9697 1879–1026

[13] Bunescu, Razvan ; Struble, Nigel ; Marling, Cindy ; Shubrook, Jay ; Schwartz, Frank: Blood Glucose Level Prediction Using Physiological Models and Support Vector Regression. En: *2013 12th International Conference on Machine Learning and Applications*, Ieee, 12 2013. – ISBN 978–0–7695–5144–9, p. 135–140

[14] Cardenas-Barrera, Julian L. ; Meng, Julian ; Castillo-Guerra, Eduardo ; Chang, Liuchen: A Neural Network Approach to Multi-step-ahead, Short-Term Wind Speed Forecasting. En: *2013 12th International Conference on Machine Learning and Applications* (2013), 12, p. 243–248. ISBN 978–0–7695–5144–9

[15] Chaiyakhan, Kedkarn ; Chujai, Pasapitch ; Kerdprasop, Nittaya ; Kerdprasop, Kittisak: Hourly Ground - level Ozone Concentration Prediction using Support Vector Regression. En: *International MultiConference of Engineers and Computer Scientists* Vol. I, 2017. – ISBN 9789881404732

[16] Chao, Wei L. ; Liu, Jun Z. ; Ding, Jian J.: Facial age estimation based on label-sensitive learning and age-oriented regression. En: *Pattern Recognition* 46 (2013), Nr. 3, p. 628–641. – ISBN 9781467300469

[17] Chao, Wei L. ; Liu, Jun Z. ; Ding, Jian J.: Facial age estimation based on label-sensitive learning and age-oriented regression. En: *Pattern Recognition* 46 (2013), Nr. 3, p. 628–641. – ISBN 9781467300469

[18] Chen, Thao-Tsen ; Lee, Shie-Jue: A weighted LS-SVM based learning system for time series forecasting. En: *Information Sciences* 299 (2015), p. 99–116. – ISSN 00200255

[19] Cherkassky, V.: The Nature Of Statistical Learning Theory˜. En: *IEEE Transactions on Neural Networks* 8 (1997), Nr. 6, p. 1564–1564. – ISBN 978–1–4419–3160–3

[20] Cherkassky, Vladimir ; Ma, Yunqian: Selection of Meta-parameters for Support Vector Regression. En: *Artificial Neural Networks* (2002), p. 687–693. – ISBN 978–3–540–44074–1 978–3–540–46084–8

[21] Chu, Yinghao ; Pedro, Hugo T. ; Li, Mengying ; Coimbra, Carlos F.: Real-time forecasting of solar irradiance ramps with smart image processing. En: *Solar Energy* 114 (2015), p. 91–104. – ISSN 0038092X

[22] Cornelis, Chris ; Cock, Martine D. ; Radzikowska, Anna M.: Vaguely Quantified Rough Sets. En: *Lecture Notes Artificial Intelligent* 281 (2007), p. 87–94

[23] Cui, Herui ; Peng, Xu: Short-Term City Electric Load Forecasting with Considering Temperature Effects : An Improved ARIMAX Model. En: *Mathematical Problems in Engineering* 2015 (2015)

[24] De Gennaro, Gianluigi ; Trizio, Livia ; Di Gilio, Alessia ; Pey, Jorge ; P??rez, Noemi ; Cusack, Michael ; Alastuey, Andr??s ; Querol, Xavier: Neural network model for the prediction of PM10 daily concentrations in two sites in the Western Mediterranean. En: *Science of the Total Environment* 463-464 (2013), p. 875–883. – ISBN 0048–9697

[25] Deng, S. ; Mitsubuchi, T. ; Shioda, K. ; Shimada, T. ; Sakurai, a.: Multiple Kernel Learning on Time Series Data and Social Networks for Stock Price Prediction. En: *2011 10th International Conference on Machine Learning and Applications and Workshops* (2011), 12, p. 228–234. ISBN 978–1–4577–2134–2

[26] Derrac, Joaquín ; García, Salvador ; Herrera, Francisco: Fuzzy nearest neighbor algorithms: Taxonomy, experimental analysis and prospects. En: *Information Sciences* 260 (2014), p. 98–119. – ISBN 0020–0255

[27] Donnelly, Aoife ; Misstear, Bruce ; Broderick, Brian: Real time air quality forecasting using integrated parametric and non-parametric regression techniques. En: *Atmospheric Environment* 103 (2015), Nr. 2, p. 53–65. – ISBN 1352–2310

[28] Duan, Jingchun ; Tan, Jihua ; Yang, Liu ; Wu, Shan ; Hao, Jimin: Concentration, sources and ozone formation potential of volatile organic compounds (VOCs) during ozone episode in Beijing. En: *Atmospheric Research* 88 (2008), Nr. 1, p. 25–35. – ISBN 01698095 (ISSN)

[29] Duke, David l.: *Intelligent diabetes assistant: A telemedicine system for modeling and managing blood glucose*, Carnegie Mellon University, Tesis de Grado, 2009

[30] Durão, Rita M. ; Mendes, Manuel T. ; João Pereira, M.: Forecasting O3 levels in industrial area surroundings up to 24 h in advance, combining classification trees and MLP models. En: *Atmospheric Pollution Research* 7 (2016), Nr. 6, p. 961–970. – ISSN 13091042

[31] D'Urso, Pierpaolo ; De Giovanni, Livia ; Massari, Riccardo: Time series clustering by a robust autoregressive metric with application to air pollution. En: *Chemometrics and Intelligent Laboratory Systems* 141 (2015), p. 107–124. – ISSN 01697439

[32] Dutykh, Denys ; Poncet, Raphaël ; Dias, Frédéric: The VOLNA code for the numerical modeling of tsunami waves: Generation, propagation and inundation. En: *European Journal of Mechanics - B/Fluids* 30 (2011), Nr. 6, p. 598–615. – ISSN 09977546

[33] Ebi, Kristie L. ; McGregor, Glenn: Climate change, tropospheric ozone and particulate matter, and health impacts. En: *Environmental Health Perspectives* 116 (2008), Nr. 11, p. 1449–1455. – ISBN 0091–6765

[34] Edward M. Ishiyama, D. Ian W.: Modeling and simulation of the polymeric nanocapsule formation process. En: *AIChE Journal* 57 (2011), Nr. 11, p. 3199–3209. – ISBN 9783902661548

[35] EPA: Volatile Organic Compounds Emissions. 2015. – Informe de Investigación

[36] Eronen, Antti J. ; Klapuri, Anssi P.: Music tempo estimation with k-NN regression. En: *IEEE Transactions on Audio, Speech, and Language Processing* 18 (2010), Nr. 1, p. 50–57. – ISSN 1063–6676

[37] Fang, Hui ; Mac Parthaláin, Neil ; Aubrey, Andrew J. ; Tam, Gary K L. ; Borgo, Rita ; Rosin, Paul L. ; Grant, Philip W. ; Marshall, David ; Chen, Min: Facial expression recognition in dynamic sequences: An integrated approach. En: *Pattern Recognition* 47 (2014), Nr. 3, p. 1271–1281. – ISSN 00313203

[38] Farahnakian, Fahimeh ; Pahikkala, Tapio ; Liljeberg, Pasi ; Plosila, Juha: Energy Aware Consolidation Algorithm Based on K-Nearest Neighbor Regression for Cloud Data Centers. En: *2013 IEEE/ACM 6th International Conference on Utility and Cloud Computing* (2013), p. 256–259. ISBN 978–0–7695–5152–4

[39] Finlayson-Pitts, B. J.: Tropospheric Air Pollution: Ozone, Airborne Toxics, Polycyclic Aromatic Hydrocarbons, and Particles. En: *Science* 276 (1997), Nr. 5315, p. 1045–1051. – ISBN 0036–8075

[40] Fishman, G. I. ; Chugh, S. S. ; DiMarco, J. P. ; Albert, C. M. ; Anderson, M. E. ; Bonow, R. O. ; Buxton, a. E. ; Chen, P.-S. ; Estes, M. ; Jouven, X. ; Kwong, R. ; Lathrop, D. a. ; Mascette, a. M. ; Nerbonne, J. M. ; O'Rourke, B. ; Page, R. L. ; Roden, D. M. ; Rosenbaum, D. S. ; Sotoodehnia, N. ; Trayanova, N. a. ; Zheng, Z.-J.: Sudden Cardiac Death Prediction and Prevention: Report From a National Heart, Lung, and Blood Institute and Heart Rhythm Society Workshop. En: *Circulation* 122 (2010), Nr. 22, p. 2335–2348. – ISSN 0009–7322

[41] García Nieto, P. J. ; Combarro, E. F. ; Del Coz Díaz, J. J. ; Montañés, E.: A SVM-based regression model to study the air quality at local scale in Oviedo urban area (Northern Spain): A case study. En: *Applied Mathematics and Computation* 219 (2013), Nr. 17, p. 8923–8937. – ISSN 00963003

[42] García Nieto, P. J. ; García-Gonzalo, E. ; Bernardo Sánchez, A. ; Rodríguez Miranda, A. A.: Air Quality Modeling Using the PSO-SVM-Based Approach, MLP Neural Network, and M5 Model Tree in the Metropolitan Area of Oviedo (Northern Spain). En: *Environmental Modeling & Assessment* (2017). – ISSN 1420–2026

[43] Geidel, C. ; Zareipour, H.: Price Forecasting in the Spanish Day-Ahead Electricity Market Using Preconditioned Wind Power Information. En: *2013 12th International Conference on Machine Learning and Applications* (2013), 12, p. 203–210. ISBN 978–0–7695–5144–9

[44] Goyal, Rinkaj ; Chandra, Pravin ; Singh, Yogesh: Suitability of KNN Regression in the Development of Interaction based Software Fault Prediction Models. En: *IERI Procedia* 6 (2014), p. 15–21. – ISSN 22126678

[45] Gribok, Andrei V. ; Buller, Mark J. ; Hoyt, Reed W. ; Reifman, Jaques: A real-time algorithm for predicting core temperature in humans. En: *IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society* 14 (2010), 7, Nr. 4, p. 1039–45. – ISSN 1558–0032

[46] H. J. Zimmermann: *Fuzzy set theory and its applications*. 1993. – ISBN 9789401038706

[47] Han, Bin ; Muma, Michael ; Feng, Mengling ; Zoubir, Abdelhak M.: An online approach for intracranial pressure forecasting based on signal decomposition and robust statistics. En: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (2013), 5, p. 6239–6243. ISBN 978–1–4799–0356–6

[48] Health Effects Institute: STATE OF GLOBAL AIR 2017 / Health Effects Institute. Boston, 2017. – Informe de Investigación

[49] Hrust, Lovro ; Klaić, Zvjezdana B. ; Križan, Josip ; Antonić, Oleg ; Hercog, Predrag: Neural network forecasting of air pollutants hourly concentrations using optimised temporal averages of meteorological variables and pollutant concentrations. En: *Atmospheric Environment* 43 (2009), 11, Nr. 35, p. 5588–5596. – ISSN 13522310

[50] Hsia, Tain-Yen Y. ; Cosentino, Daria ; Corsini, Chiara ; Pennati, Giancarlo ; Dubini, Gabriele ; Migliavacca, Francesco: Use of mathematical modeling to compare and predict hemodynamic effects between hybrid and surgical Norwood palliations for hypoplastic left heart syndrome. En: *Circulation* 124 (2011), Nr. 11 Suppl, p. 204–10. – ISBN 1524–4539 (Electronic)\r0009–7322 (Linking)

[51] Hu, Zhongyi ; Bao, Yukun ; Xiong, Tao ; Chiong, Raymond: Hybrid filter–wrapper feature selection for short-term load forecasting. En: *Engineering Applications of Artificial Intelligence* 40 (2015), p. 17–27. – ISSN 09521976

[52] Hyndman, Rob J. ; Koehler, Anne B.: Another look at measures of forecast accuracy. En: *International Journal of Forecasting* 22 (2006), 10, Nr. 4, p. 679–688. – ISBN 0169–2070

[53] Irish, W. D. ; Ilsley, J. N. ; Schnitzler, M. a. ; Feng, S. ; Brennan, D. C.: A Risk Prediction Model for Delayed Graft Function in the Current Era of Deceased Donor Renal Transplantation. En: *American Journal of Transplantation* 10 (2010), Nr. 10, p. 2279–2286. – ISSN 16006135

[54] Jawaid, Faizan ; NazirJunejo, Khurum: Predicting daily mean solar power using machine learning regression techniques. En: *2016 Sixth International Conference on Innovative Computing Technology (INTECH)* (2016), Nr. September, p. 355–360. ISBN 9781509020003

[55] Jayaratne, E. R. ; Ristovski, Z. D. ; Morawska, L. ; Meyer, N. K.: Carbon dioxide emissions from diesel and compressed natural gas buses during acceleration. En: *Transportation Research Part D* 15 (2010), Nr. 5, p. 247–253. – ISBN 1361–9209

[56] Jensen, Richard ; Cornelis, Chris: Fuzzy-rough nearest neighbour classification and prediction. En: *Theoretical Computer Science* 412 (2011), Nr. 42, p. 5871–5884. – ISBN 978–3–642–18301–0

[57] Jensen, Richard ; Cornelis, Chris: Fuzzy-rough nearest neighbour classification and prediction. En: *Theoretical Computer Science* 412 (2011), Nr. 42, p. 5871–5884. – ISBN 978–3–642–18301–0

[58] Jing, Wenlong ; Yang, Yaping ; Yue, Xiafang ; Zhao, Xiaodan: A Comparison of Different Regression Algorithms for Downscaling Monthly Satellite-Based Precipitation over North China. En: *Remote Sensing* 8 (2016), Nr. 10, p. 1–17. – ISBN 2072–4292

[59] Kampa, Marilena ; Castanas, Elias: Human health effects of air pollution. En: *Environmental Pollution* 151 (2008), Nr. 2, p. 362–367. – ISBN 0269–7491

[60] Kennedy, J ; Eberhart, R: Particle swarm optimization. En: *Neural Networks, 1995. Proceedings., IEEE International Conference on* 4 (1995), p. 1942–1948. – ISBN VO – 4

[61] Khan, Gul M. ; Zafari, Faheem ; Mahmud, S. A.: Very Short Term Load Forecasting Using Cartesian Genetic Programming Evolved Recurrent Neural Networks (CGPRNN). En: *2013 12th International Conference on Machine Learning and Applications* (2013), 12, p. 152–155. ISBN 978–0–7695–5144–9

[62] Komorowski, Jan ; Polkowski, Lech ; Skowron, Andrzej: Rough sets: A tutorial. En: *Rough fuzzy ...* (1999), p. 2–8. – ISSN 08247935

[63] Kou, Peng: Sparse Heteroscedastic Gaussian Process for Short- term Wind Speed Forecasting. En: *IEEE Joint Conference on Neural Networks* (2012), p. 10–15

[64] Kristiansen, Tarjei: Forecasting Nord Pool day-ahead prices with an autoregressive model. En: *Energy Policy* 49 (2012), p. 328–332. – ISSN 03014215

[65] Lazos, Dimitris ; Sproul, Alistair B. ; Kay, Merlinde: Development of hybrid numerical and statistical short term horizon weather prediction models for building energy management optimisation. En: *Building and Environment* 90 (2015), p. 82–95. – ISSN 03601323

[66] Li, Fan ; Ye, Mao ; Chen, Xudong: An extension to Rough c-means clustering based on decision-theoretic Rough Sets model. En: *International Journal of Approximate Reasoning* 55 (2014), Nr. 1, p. 116–129. – ISSN 0888613X

[67] Li, W D. ; Kong, D M. ; Wu, J R.: A New Hybrid Model FPA-SVM Considering Cointegration for Particular Matter Concentration Forecasting: A Case Study of Kunming and Yuxi, China. En: *Computational Intelligence and Neuroscience* 2017 (2017). – ISSN 1687–5265

[68] Liew, Siaw H. ; Choo, Yun H. ; Low, Yin F.: Fuzzy-rough nearest neighbour classifier for person authentication using EEG signals. En: *iFUZZY 2013 - 2013 International Conference on Fuzzy Theory and Its Applications* (2013), p. 316–321. ISBN 9781479903863

[69] Lin, Jeng-Wen ; Chen, Cheng-Wu ; Peng, Cheng-Yi: Potential hazard analysis and risk assessment of debris flow by fuzzy modeling. En: *Natural Hazards* 64 (2012), Nr. 1, p. 273–282. – ISSN 0921–030X

[70] Ling, Yao ; Ning, Lu ; Sheng, Jiang: Artificial Neural Network (ANN) for Multi-source PM2.5 Estimation Using Surface, MODIS, and Meteorological Data. En: *Biomedical Engineering and Biotechnology (iCBEB), 2012 International Conference on* (2012), Nr. Figure 1, p. 1228–1231. ISBN 978–0–7695–4706–0

[71] Liu, Da ; Wang, Jilong ; Wang, Hui: Short-term wind speed forecasting based on spectral clustering and optimised echo state networks. En: *Renewable Energy* 78 (2015), p. 599–608. – ISSN 09601481

[72] Lotte, F ; Congedo, M ; Lécuyer, a ; Lamarche, F ; Arnaldi, B: A review of classification algorithms for EEG-based brain-computer interfaces. En: *J Neural Eng* 4 (2007), Nr. 2, p. R1–R13. – ISBN 1741–2552

[73] Luo, Ming ; Zheng, Ying ; Liu, Shujie: Data-based fault-tolerant control of the semiconductor manufacturing process based on K-nearest neighbor nonparametric regression. En: *Proceedings of the 10th World Congress on Intelligent Control and Automation* (2012), p. 3008–3012. ISBN 978–1–4673–1398–8

[74] Mannodi-Kanakkithodi, Arun ; Pilania, Ghanshyam ; Ramprasad, Rampi: Critical assessment of regression-based machine learning methods for polymer dielectrics. En: *Computational Materials Science* 125 (2016), p. 123–135. – ISSN 09270256

[75] Marija, Savic ; Ivan, Mihajlovic ; Zivan, Zivkovic: An ANFIS - based air quality model for prediction of SO2 concentration in urban area. En: *Serbian Journal of Management* 8 (2013), Nr. 1, p. 25–38. – ISSN 1452–4864

[76] Masiol, Mauro ; Agostinelli, Claudio ; Formenton, Gianni ; Tarabotti, Enzo ; Pavoni, Bruno: Thirteen years of air pollution hourly monitoring in a large city: Potential sources, trends, cycles and effects of car-free days. En: *Science of the Total Environment* 494 (2014), p. 84–96. – ISBN 0048–9697

[77] Mathur, N ; Glesk, I ; Buis, a: Skin Temperature Prediction in Lower Limb Prostheses. En: *Ieee Journal of Biomedical and Health Informatics* 20 (2016), Nr. 1, p. 158–165. – ISBN 2168–2194

[78] Mihalache, Sanda F. ; Popescu, Marian ; Oprea, Mihaela: Particulate Matter Prediction using ANFIS Modelling Techniques. En: *19th International Conference on System Theory, Control and Computing (ICSTCC)*, 2015. – ISBN 9781479984817, p. 895–900

[79] Mishra, Dhirendra ; Goyal, P: Neuro-Fuzzy approach to forecasting Ozone Episodes over the urban area of Delhi, India. En: *Environmental Technology & Innovation* 5 (2016), p. 83–94. – ISSN 2352–1864

[80] Mustafic, H ; Jabre, P ; Caussin, C ; Murad, M H. ; Escolano, S ; Tafflet, M ; Périer, M C. ; Marijon, E ; Vernerey, D ; Empana, J P. ; Jouven, X: Main air pollutants and myocardial infarction: A systematic review and meta-analysis. En: *JAMA: Journal of the American Medical Association* 307 (2012), Nr. 7, p. 713–721. – ISBN ISSN 0098–7484

[81] Ngai, E. W T. ; Peng, S. ; Alexander, Paul ; Moon, Karen K L.: Decision support and intelligent systems in the textile and apparel supply chain: An academic review of research articles. En: *Expert Systems with Applications* 41 (2014), Nr. 1, p. 81–91. – ISBN 09574174

[82] Park, Deuk H. ; Kim, Hyea K. ; Choi, Il Y. ; Kim, Jae K.: A literature review and classification of recommender systems research. En: *Expert Systems with Applications* 39 (2012), Nr. 11, p. 10059–10072. – ISBN 09574174

[83] Pawlak, Z: Rough sets. En: *International Journal of Computer & Information ...* (1982), p. 1–51. – ISBN 978–94–010–5564–2

[84] Pawlak, Zdzisław: Rough Sets. En: *International Journal of Computer and Information Science* 11 (1982), Nr. 5, p. 341–356. – ISSN 1558–0032

[85] Pawlak, Zdzisław ; Skowron, Andrzej: Rudiments of rough sets. En: *Information Sciences* 177 (2007), Nr. 1, p. 3–27. – ISBN 0020–0255

[86] Pawlak, Zdzisław ; Skowron, Andrzej: Rudiments of rough sets. En: *Information Sciences* 177 (2007), Nr. 1, p. 3–27. – ISBN 0020–0255

[87] Ping Tian, Dong: A Review of Convergence Analysis of Particle Swarm Optimization. En: *International Journal of Grid and Distributed Computing* 6 (2013), Nr. 6, p. 117–128. – ISSN 2005–4262

[88] Prabakaran, K ; Sivapragasam, C ; Jeevapriya, C ; Narmatha, A: Forecasting Cultivated Areas And Production Of Wheat In India Using ARIMA Model. En: *Golden Research Thoughts* 3 (2013), Nr. 3, p. 1–8. – ISSN 2231–5063

[89] Qu, Yanpeng ; Shen, Qiang ; Parthaláin, Neil M. ; Shang, Changjing ; Wu, Wei: Fuzzy similarity-based nearest-neighbour classification as alternatives to their fuzzy-rough parallels. En: *International Journal of Approximate Reasoning* 54 (2013), Nr. 1, p. 184–195. – ISBN 0888–613X

[90] Rawat, Rohit ; Vora, Kunal ; Manry, Michael ; Eapi, Gautam: Multi-variable Neural Network Forecasting Using Two Stage Feature Selection. En: *2014 13th International Conference on Machine Learning and Applications* (2014), 12, Nr. 4, p. 243–250. ISBN 978–1–4799–7415–3

[91] Reid, Tyler G. ; Tarantino, Paul M.: Arctic Sea Ice Extent Forecasting Using Support Vector Regression. En: *2014 13th International Conference on Machine Learning and Applications* (2014), 12, p. 1–6. ISBN 978–1–4799–7415–3

[92] Rodger, James a.: A fuzzy nearest neighbor neural network statistical model for predicting demand for natural gas and energy cost savings in public buildings. En: *Expert Systems with Applications* 41 (2014), Nr. 4 PART 2, p. 1813–1829. – ISBN 0957–4174

[93] Russell, B.: A focus on particule matter and health. En: *Environmental, Science and Technology* 43 (2009), Nr. 13, p. 4620–4625

[94] Russo, Ana ; Raischel, Frank ; Lind, Pedro G.: Air quality prediction using optimal neural networks with stochastic variables. En: *Atmospheric Environment* 79 (2013), p. 822–830. – ISBN 1352–2310

[95] Sacks, Jason D. ; Stanek, Lindsay W. ; Luben, Thomas J. ; Johns, Douglas O. ; Buckley, Barbara J. ; Brown, James S. ; Ross, Mary: Particulate matter-induced health effects: Who is susceptible? En: *Environmental Health Perspectives* 119 (2011), Nr. 4, p. 446–454. – ISBN 0091–6765

[96] Saha, Indrajit ; Prasad, Jnanendra ; Maulik, Ujjwal: Knowledge-Based Systems Ensemble based rough fuzzy clustering for categorical data. En: *Knowledge-Based Systems* 77 (2015), p. 114–127. – ISSN 0950–7051

[97] Sánchez Triana, Ernesto ; Ahmed, Kulsum ; Awe, Yewande ; Banco Mundial (Ed.): *Prioridades ambientales para la reducción de la pobreza en Colombia. Un análisis ambiental del país para Colombia*. Washington, DC : The World Bank, 2007. – 522 p.. – ISBN 978–958–8307–10–7

[98] Sestelo, Marta ; Roca-Pardiñas, Javier ; Ordóñez, Celestino: Predicting SO2 pollution incidents by means of additive models with optimum variable selection. En: *Atmospheric Environment* 95 (2014), p. 151–157. – ISBN 1352–2310

[99] Sevim, Barış ; Bayraktar, Alemdar ; Altunişik, Ahmet C. ; Atamtürktür, Sezer ; Birinci, Fatma: Finite element model calibration effects on the earthquake response of masonry arch bridges. En: *Finite Elements in Analysis and Design* 47 (2011), Nr. 7, p. 621–634. – ISSN 0168874X

[100] Shahraiyni, Hamid T. ; Sodoudi, Sahar: Statistical modeling approaches for pm10 prediction in urban areas; A review of 21st-century studies. En: *Atmosphere* 7 (2016), Nr. 2, p. 10–13. – ISBN 4930838711

[101] Shang, Changjing ; Barnes, Dave: Fuzzy-rough feature selection aided support vector machines for Mars image classification. En: *Computer Vision and Image Understanding* 117 (2013), Nr. 3, p. 202–213. – ISSN 10773142

[102] Shi, Ji P. ; Harrison, Roy M.: Regression modelling of hourly NO(x) and NO2 concentrations in urban air in London. En: *Atmospheric Environment* 31 (1997), Nr. 24, p. 4081–4094. – ISBN 1352–2310

[103] Shukur, Osamah B. ; Lee, Muhammad H.: Daily wind speed forecasting through hybrid KF-ANN model based on ARIMA. En: *Renewable Energy* 76 (2015), p. 637–647. – ISSN 09601481

[104] Siami-Irdemoosa, Elnaz ; Dindarloo, Saeid R.: Prediction of fuel consumption of mining dump trucks: A neural networks approach. En: *Applied Energy* 151 (2015), p. 77–84. – ISSN 03062619

[105] Silva, Anthony Mihirana D. ; Noorian, Farzad ; Davis, Richard I. ; Leong, Philip H.: A Hybrid Feature Selection and Generation Algorithm for Electricity Load Prediction Using Grammatical Evolution. En: *2013 12th International Conference on Machine Learning and Applications* (2013), 12, p. 211–217. ISBN 978–0–7695–5144–9

[106] Singh, Kunwar P. ; Gupta, Shikha ; Kumar, Atulesh ; Shukla, Sheo P.: Linear and nonlinear modeling approaches for urban air quality prediction. En: *Science of the Total Environment* 426 (2012), p. 244–255. – ISBN 1879–1026 (Electronic)\r0048–9697 (Linking)

[107] Singh, Kunwar P. ; Gupta, Shikha ; Rai, Premanjali: Identifying pollution sources and predicting urban air quality using ensemble learning methods. En: *Atmospheric Environment* 80 (2013), p. 426–437. – ISBN 1352–2310

[108] Smith, Amber M. ; McCullers, Jonathan a. ; Adler, Frederick R.: Mathematical model of a three-stage innate immune response to a pneumococcal lung infection. En: *Journal of Theoretical Biology* 276 (2011), Nr. 1, p. 106–116. – ISSN 00225193

[109] Smola, Alexander J. ; Schölkopf, Bernhard: A Tutorial on Support Vector Regression. En: *Statistics and Computing* 14 (2004), Nr. 3, p. 199–222. – ISBN 0960–3174

[110] Soman, Saurabh S. ; Zareipour, Hamidreza ; Malik, Om ; Mandal, Paras: A review of wind power and wind speed forecasting methods with different time horizons. En: *North American Power Symposium 2010*, 2010. – ISBN 978–1–4244–8046–3, p. 1–8

[111] Sousa, S. I. ; Alvim-Ferraz, M. C. ; Martins, F. G.: Health effects of ozone focusing on childhood asthma: What is now known - a review from an epidemiological point of view. En: *Chemosphere* 90 (2013), Nr. 7, p. 2051–2058. – ISBN 0045–6535

[112] Suárez Sánchez, A. ; García Nieto, P. J. ; Iglesias-Rodríguez, F. J. ; Vilán Vilán, J. A.: Nonlinear air quality modeling using support vector machines in Gijón urban area (Northern Spain) at local scale. En: *International Journal of Nonlinear Sciences and Numerical Simulation* 14 (2013), Nr. 5, p. 291–305. – ISSN 15651339

[113] Suárez Sánchez, A. ; García Nieto, P. J. ; Riesgo Fernández, P. ; del Coz Díaz, J. J. ; Iglesias-Rodríguez, F. J.: Application of an SVM-based regression model to the air quality study at local scale in the Avilés urban area (Spain). En: *Mathematical and Computer Modelling* 54 (2011), Nr. 5-6, p. 1453–1466. – ISBN 9781612093420

[114] Sun, Wei ; Zhang, Hao ; Palazoglu, Ahmet: Prediction of 8 h-average ozone concentration using a supervised hidden Markov model combined with generalized linear models. En: *Atmospheric Environment* 81 (2013), p. 199–208. – ISSN 13522310

[115] Sun, Wei ; Zhang, Hao ; Palazoglu, Ahmet ; Singh, Angadh ; Zhang, Weidong ; Liu, Shiwei: Prediction of 24-hour-average PM2.5 concentrations using a hidden Markov model with different emission distributions in Northern California. En: *Science of the Total Environment* 443 (2013), p. 93–103. – ISBN 0048–9697

[116] Suraj, Zbigniew: An Introduction to Rough Set Theory and Its Applications. En: *ICENCO, Cairo, Egypt* (2004), Nr. July

[117] Suresh Babu, V. ; Viswanath, P.: Rough-fuzzy weighted k-nearest leader classifier for large data sets. En: *Pattern Recognition* 42 (2009), Nr. 9, p. 1719–1731. – ISBN 0031–3203

[118] Tang, Jinjun ; Zhang, Guohui ; Wang, Yinhai ; Wang, Hua ; Liu, Fang: A hybrid approach to integrate fuzzy C-means based imputation method with genetic algorithm for missing traffic volume data estimation. En: *Transportation Research Part C* 51 (2015), p. 29–40. – ISSN 0968–090X

[119] Tebyanian, Ardalan ; Hedayati, Fares: Intelligent Crude Oil Price Forecaster. En: *2014 13th International Conference on Machine Learning and Applications* (2014), 12, p. 453–455. ISBN 978–1–4799–7415–3

[120] Tian, Z. ; Gu, B. ; Yang, L. ; Lu, Y.: Hybrid ANN-PLS approach to scroll compressor thermodynamic performance prediction. En: *Applied Thermal Engineering* 77 (2015), p. 113–120. – ISSN 13594311

[121] Tian, Zhen ; Gu, Bo ; Qian, Cheng ; Yang, Lin ; Liu, Fen: Electronic expansion valve mass flow rate prediction based on dimensionless correlation and ANN model. En: *International Journal of Refrigeration* 57 (2015), p. 1–10. – ISSN 01407007

[122] Toro, Victoria ; Molina, Eliana ; Pati, Josshual S. ; Fern, Marcelo ; Ram, Gloria E.: Plan de descontaminación del aire en la región metropolitana del Valle de Aburrá. En: *Producción+Limpia* (2010), p. 10–26. – ISSN 19090455

[123] Treiber, Nils A. ; Heinermann, Justin ; Kramer, Oliver: Aggregation of Features for Wind Energy Prediction with Support Vector Regression and Nearest Neighbors. En: *Proc. of the Joint ECML/PKDD 2013 Workshops* (2013), Nr. SEPTEMBER

[124] Tsianos, George a. ; Rustin, Cedric ; Loeb, Gerald E.: Mammalian muscle model for predicting force and energetics during physiological behaviors. En: *IEEE transactions on neural systems and rehabilitation engineering : a publication of the IEEE Engineering in Medicine and Biology Society* 20 (2012), 3, Nr. 2, p. 117–33. – ISSN 1558–0210

[125] U.S. Environmental Protection Agency. *Terms of environment: Glossary, abbreviations, and Acronyms.* 1992

[126] Vallero, Daniel: *Fundamentals of air Pollution.* 4ta. Elsevier, 2008. – ISBN 9780123736154

[127] Walczak, B. ; Massart, D.L.: Rough sets theory. En: *Chemometrics and Intelligent Laboratory Systems* 47 (1999), Nr. 1, p. 1–16. – ISSN 01697439

[128] Wang, Jian-Zhou ; Wang, Yun ; Jiang, Ping: The study and application of a novel hybrid forecasting model – A case study of wind speed forecasting in China. En: *Applied Energy* 143 (2015), p. 472–488. – ISSN 03062619

[129] Wang, Jianzhou ; Hu, Jianming ; Ma, Kailiang ; Zhang, Yixin: A self-adaptive hybrid approach for wind speed forecasting. En: *Renewable Energy* 78 (2015), p. 374–385. – ISSN 0960–1481

[130] Wang, Jianzhou ; Qin, Shanshan ; Zhou, Qingping ; Jiang, Haiyan: Medium-term wind speeds forecasting utilizing hybrid models for three different sites in Xinjiang, China. En: *Renewable Energy* 76 (2015), p. 91–101. – ISSN 09601481

[131] Wang, Ping ; Liu, Yong ; Qin, Zuodong ; Zhang, Guisheng: A novel hybrid forecasting model for PM10 and SO2 daily concentrations. En: *Science of the Total Environment* 505 (2015), p. 1202–1212. – ISSN 18791026

[132] Wang, Weina ; Pedrycz, Witold ; Liu, Xiaodong: Time series long-term forecasting model based on information granules and fuzzy clustering. En: *Engineering Applications of Artificial Intelligence* 41 (2015), p. 17–24. – ISSN 09521976

[133] Wang, Wenjian ; Xu, Zongben ; Lu, Weizhen ; Zhang, Xiaoyun: Determination of the spread parameter in the Gaussian kernel for classification and regression. En: *Neurocomputing* 55 (2003), Nr. 3-4, p. 643–663. – ISBN 0925–2312

[134] Weron, Rafa: Electricity price forecasting: A review of the state-of-the-art with a look into the future. En: *International Journal of Forecasting* 30 (2014), Nr. 4, p. 1030–1081. – ISBN 01692070

[135] Westerlund, Joakim ; Urbain, Jean P. ; Bonilla, Jorge: Application of air quality combination forecasting to Bogota. En: *Atmospheric Environment* 89 (2014), p. 22–28. – ISBN 1352–2310

[136] Wohlfarth, Till ; Clemencon, Stephan ; Roueff, Francois ; Casellato, Xavier: A Data-Mining Approach to Travel Price Forecasting. En: *2011 10th International Conference on Machine Learning and Applications and Workshops* (2011), 12, Nr. M, p. 84–89. ISBN 978–1–4577–2134–2

[137] World Health Organization: Ambient Air Pollution: A global assessment of exposure and burden of disease. En: *World Health Organization* (2016), p. 1–131. ISBN 978 92 4 151135 3

[138] Yanpeng, Qu ; Changjing, Shang ; Qiang, Shen ; Mac Parthalain, N ; Wei, Wu: Kernel-based Fuzzy-rough Nearest Neighbour Classification. En: *Proceedings 2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011)* (2011), p. 1523–1529. ISBN 9781424473175

[139] Yeganeh, B. ; Motlagh, M. Shafie P. ; Rashidi, Y. ; Kamalan, H.: Prediction of CO concentrations based on a hybrid Partial Least Square and Support Vector Machine model. En: *Atmospheric Environment* 55 (2012), p. 357–365. – ISSN 13522310

[140] Zadeh, L.a.: Fuzzy sets. En: *Information and Control* 8 (1965), Nr. 3, p. 338–353. – ISBN 0019–9958

[141] Zhang, Yang ; Bocquet, Marc ; Mallet, Vivien ; Seigneur, Christian ; Baklanov, Alexander: Real-time air quality forecasting, Part II: State of the science, current research needs, and future prospects. En: *Atmospheric Environment* 60 (2012), p. 656–676. – ISBN 1352–2310