Universidad del Azuay

Facultad de Ciencias de la Administración

Escuela de Ingeniería de Sistemas y Telemática

IMPUTACIÓN DE DATOS A PARTIR DE LA BÚSQUEDA DE PATRONES EN UN CONJUNTO DE DATOS DE CONTAMINANTES ATMOSFÉRICOS

Trabajo de graduación previo a la obtención del título de Ingeniero en Sistemas y Telemática

Autor:

Miguel Angel Calle Beltrán

Director:

Ing. Patricia Margarita Ortega Chasi, PhD

Cuenca – Ecuador

2020

DEDICATORIA

Todo el esfuerzo y sacrificio realizando en estos años, plasmado de alguna manera en este trabajo de titulación quiero dedicar a mi familia y principalmente a mis padres Teodoro y Alexandra. Sin el apoyo incondicional de ellos nada de esto hubiese sido posible, permitiéndome crecer de manera personal como profesional, sus palabras de aliento que me fortalecieron en cada momento difícil y me permitieron afrontar las dificultades de la mejor manera posible.

Esperando que se sientan orgullosos del hijo que tienen, esto es para ustedes mis viejitos.

AGRADECIMIENTO

Seguramente me faltaría espacio para agradecer a todos quienes me ayudaron de una u otra manera a estar en este lugar y haberme realizado de manera profesional, sin embargo quiero agradecer principalmente a Dios, que me ha brindado la fuerza necesaria para culminar mis estudios y me dio a los mejores padres del mundo, Teodoro Calle y Alexandra Beltrán, quienes con su apoyo y consejos me permitieron seguir una carrera universitaria y culminarla, también agradecer a mis hermanos Christian y Diego que en todo momento están conmigo, esperando ser un buen ejemplo para ellos en todos los ámbitos.

Un agradecimiento a la Universidad del Azuay, que me abrió las puertas para realizar mis estudios, a mi directora de tesis Ing. Patricia Ortega y Co-director Ing. Marcos Orellana, quienes me han sabido orientar para llevar a cabo de la mejor manera el proyecto de titulación y a todos los profesores quienes me guiaron para ampliar el conocimiento, que además de ser profesores, también se convirtieron en amigos. Sin más que decir, muchas gracias a todos quienes formaron parte de esta etapa.

RESUMEN:

Los sensores de contaminantes atmosféricos capturan gran cantidad de datos, parte de esta información se pierde por diversas causas incluyendo errores en los sensores y errores humanos. Este trabajo plantea una solución a este inconveniente con la imputación de datos perdidos a través de una red neuronal NARX implementada en Matlab para el relleno de estos datos. El pre-procesamiento de los datos incluyó la estandarización de las variables de entrada, y la eliminación los valores atípicos. Posteriormente se calculó el valor del ángulo entre los niveles de las variables de entradas considerando intervalos de 10 minutos. La red neuronal utiliza como variables de entrada los contaminantes O3, CO, NO2, SO2, PM2_5 y Temperatura de acuerdo al análisis previo de interacciones entre contaminantes.

El O3 y NO2 muestran los mejores resultados con valores de R = 0.85 y R=0.73 respectivamente, valores obtenidos con el conjunto de pruebas.

Palabras clave: Redes Neuronales, Minería de datos, Imputación de datos, Relleno de datos faltantes, Red NARX.

ABSTRACT

Air pollutant sensors capture a large amount of data. Some of this information is lost due to various causes, including sensor errors and human errors. This work poses a solution to this problem with the imputation of lost data through a NARX neural network, implemented in Matlab for the filling of this data. The pre-processing of the data included the standardization of the input variables and the elimination of outliers. Subsequently, the value of the angle between the levels of the input variables was calculated considering 10 minute intervals. The neural network uses the contaminants O3, CO, NO2, SO2, PM2_5 and Temperature as input variables according to the previous analysis of interactions between pollutants. The O3 and NO2 show the best results with values of R=0.85 and R=0.73 respectively, values obtained with the set of tests.

Keywords: Neural Networks, Data Mining, Data Allocation, Filling of missing data, NARX Network.

Ing. Patricia Margarita Ortega Chasi

Thesis Director

UNIVERSIDAD DEL AZUAY
Dpto. Idiomas

Translated by

Ing. Paúl Arpi

ÍNDICE

Tablas

Figuras

Anexos

1. INTRODUCCIÓN

El crecimiento exponencial del suelo urbano y la industrialización en los últimos años ha aumentado significativamente la contaminación atmosférica afectando principalmente el ámbito de la salud (Alvarez et al., 2017). El problema de la contaminación ha afectado la salud humana con las exposiciones tanto de material particulado (PM), como de monóxido de carbono, dióxido de azufre y otros gases (Cevallos, Díaz, & Sirois, 2017). Los gases tóxicos existentes en el aire son realmente preocupantes, debido a que causan problemas a la salud de todos los habitantes, sin embargo los más vulnerables son los niños, esto debido al tamaño pequeño de sus órganos respiratorios y bajas defensas en sus mecanismos (Estrella, Sempértegui, Franco, Cepeda, & Naumova, 2019). Esta contaminación es algo que está causando problemas, tanto en la salud como al medio ambiente, siendo la causante de lluvias ácidas, destrucción de la capa de ozono, y además del calentamiento global (Kingsy et al., 2017).

Data *mining* o minería de datos es una disciplina que ha permitido extraer conocimiento de gran cantidad de datos relacionados con la contaminación del medio ambiente, de esta manera estudiar los patrones y su comportamiento (Gore & Deshpande, 2017). Una de las herramientas para minería de datos son las redes neuronales, usadas en muchos proyectos para la predicción del comportamiento de contaminantes en distintas partes del mundo (Azid et al., 2014; Kurt, Gulbagci, Karaca, & Alagha, 2008; Lo, Lu, & Fan, 2003). Las redes neuronales han sido utilizadas en su gran mayoría debido a la gran certeza en comparación con otras herramientas que realizan la misma función (Viotti, Liuti, & Di Genova, 2002).

Las redes neuronales han sido utilizadas en diferentes ámbitos en minería de datos donde han sido aplicadas por ejemplo en la predicción del comportamiento meteorológico para la toma de decisiones en las actividades agrícolas (Fuentes, Campos, & García-Loyola, 2018). En ambientes urbanos contaminados como por ejemplo la ciudad de Perugia se realizó un análisis de predicción de contaminantes utilizando variables monitorizadas como dióxido de azufre, óxidos de nitrógeno, PM10, benceno, monóxido de carbono, ozono, velocidad del viento horizontal, humedad, presión, temperatura, radiación solar total (Viotti et al., 2002), sin embargo las predicciones del

comportamiento meteorológico son difíciles de realizar, debido a la alta cantidad de variación de parámetros y comportamiento que se deben considerar (Geetha & Nasira, 2015). Desde este punto de vista, las redes neuronales son una herramienta muy utilizada, ya que una vez entrenadas incorporan un análisis más rápido para poder predecir series de tiempo con alta exactitud en sus predicciones (Kukkonen et al., 2003).

Varios estudios se han centrado en el análisis de los contaminantes atmosféricos, dichos estudios se han realizado utilizando diferentes herramientas como son: redes neuronales, árboles de decisiones, lógica difusa, algoritmos genéticos, entre otros (Geetha & Nasira, 2015), sin embargo algo que no se ha podido evidenciar en trabajos anteriores es una técnica que permita buscar patrones dentro del comportamiento de los contaminantes utilizando el ángulo que se forma entre los niveles de un contaminante en un determinado intervalo de tiempo. Permitiendo a partir del ángulo buscar patrones, para posteriormente realizar predicciones de una mejor manera.

Desde un punto de vista tecnológico, los datos que se tienen acerca de los contaminantes, exposición, comportamiento, factores ambientales y salud pública es cada vez más relevante. La abundante cantidad de datos permite tener un gran potencial para procesar información y observar sus comportamientos, sin embargo los métodos tradicionales existentes son poco adecuados para los grandes volúmenes de datos que se obtienen. Esto ha dado como resultado la búsqueda de nuevas formas de análisis para avanzar con la comprensión de datos (Bellinger, Mohomed Jabbar, Zaïane, & Osornio-Vargas, 2017). Por ello, se puede describir que la minería de datos es una disciplina híbrida, en donde se permite la integración de tecnologías como bases de datos, estadísticas, aprendizaje automático, computación de alto rendimiento, etc. (Li & Shue, 2004). Los algoritmos de minería de datos, y entre ellos las redes neuronales, son cada vez más utilizados en la investigación de temas relacionados a la contaminación del aire (Bellinger et al., 2017).

Dentro de la ciudad de Cuenca, y específicamente en la Universidad del Azuay, se han realizado investigaciones conjuntamente con el Instituto de Estudios de Régimen Seccional del Ecuador (IERSE), departamento que cuenta con datos de los contaminantes atmosféricos de la ciudad. Estos datos han sido combinados con algoritmos de minería de datos en proyectos de investigación anteriores. En una primera fase se realizó el trabajo

de titulación “Búsqueda de patrones a partir del comportamiento de los contaminantes atmosféricos” (Ortega, 2018), y posteriormente el trabajo de titulación “Relación de los contaminantes con datos de tipo meteorológico” (Andrade, 2018), sin embargo estos estudios tuvieron una cantidad considerable de datos perdidos al realizar su análisis, debido a que el sensor en muchas ocasiones tiene inconvenientes y no obtiene toda la información. Es por ello, que es necesario direccionar la imputación de datos perdidos. En este trabajo se parte de la premisa que las redes neuronales, combinadas con una técnica angular desarrollada en el marco de este trabajo, la cual consiste en obtener el valor del ángulo del contaminante en función del nivel en el que se encuentra en un intervalo de tiempo, permitiendo obtener resultados satisfactorios en cuanto a la imputación de datos.

1.1 Marco Teórico y Estado del Arte

En esta sección se establecen los conceptos teóricos que permiten entender de mejor manera la investigación que se realizó. Estos conceptos incluyen una descripción de los datos con los cuales se realizó el estudio, metodología utilizada, herramientas que permitieron desarrollar el estudio y aspectos de la disciplina en la cual está enfocada el trabajo de titulación.

Esta sección también incluye la revisión de estudios relacionados al tema, evidenciando que ha sido realizado, y contrastando dichos trabajos con el del estudio.

De esta manera se permite al lector tener un poco de conocimiento previo, ayudando a entender de una manera más fácil el documento, evitando cualquier tipo de confusión.

1.1.1 Minería de datos

La minería de datos es la ciencia que por medio de procesos de cálculos permite analizar grandes conjuntos de datos, descubriendo el comportamiento de los datos, como por ejemplo la existencia de patrones, predicciones de datos futuros, o comportamientos que pueden realizar a partir de unas acciones (Bellinger et al., 2017). La minería de datos es una disciplina híbrida en donde se integran múltiples disciplinas como son matemáticas, computación, estadística etc., para el tratamiento de grandes volúmenes de datos que pueden estar almacenados en base de datos, depósitos de información o almacenes de datos. Su aplicación permite incorporar una ventaja competitiva a una organización mediante la explotación del conocimiento (Li & Shue, 2004).

1.1.2 Redes Neuronales Artificiales

Las redes neuronales artificiales son una potente herramienta de minería de datos, que inspiradas en una red neuronal biológica, permiten utilizar grupo de neuronas que se comunican entre sí por medio de señales, recibiendo información, aprendiendo y generando una salida (Fuentes et al., 2018). Debido a que imitan al sistema nervioso humano, replican el comportamiento también, permitiendo aprender y reconocer en base de entradas y salidas proporcionadas. De esta manera una vez entrenada la red neuronal artificial es posible tomar decisiones solamente mediante entradas, debido a que existió aprendizaje previo (Fernanda & Logroño, 2018).

1.1.3 Contaminantes del aire

La contaminación del aire es provocada por el cambio en su composición natural, esto debido a efectos naturales como emisiones de gases, cenizas volcánicas, polen, polvo, bacterias, entre otros, o efectos derivados por actividades del ser humano, los cuales representan el mayor problema para la salud (Rojas, 2017). La contaminación tiene impacto en todo el mundo, tanto a países desarrollados como en desarrollo, esta contaminación afecta a personas, animales e incluso las plantas (Alvarez et al., 2017). Según la OMS (Organización Mundial de la Salud) el 94% de muertes relacionadas con la contaminación en el 2012 se deben a enfermedades no transmisibles como son las enfermedades cardiovasculares, accidentes cerebrovasculares, la neumopatía obstructiva y el cáncer de pulmón (OMS, 2016)

Los contaminantes una vez inhalados viajan a distintas partes del sistema respiratorio. El NO2 y O3 pueden alcanzar en algunos casos hasta los alvéolos de los pulmones, el SO2 por lo general no suele pasar la región traqueobronquial y el más perjudicial es el material particulado, el cual mientras más pequeño viajará más lejos y ocasionará mayores inconvenientes para la salud (Boldo, 2016).

Dióxido de nitrógeno (NO2): el dióxido de nitrógeno se forma principalmente por la oxidación del óxido nítrico (NO) (Brunekreef & Holgate, 2002). El NO2 se presenta como un gas de color café rojizo, es un contaminante dañino en grandes concentraciones que afecta principalmente al área de los pulmones produciendo una disminución en la resistencia a infecciones (Espinoza Molina, 2011).

Material Particulado (PM2_5): el material particulado PM2_5 son partículas muy pequeñas, clasificadas por tener un diámetro menor a 2.5 micrómetros (μm). Estas partículas se generan principalmente por motores de vehículos que funcionan a diésel, generación eléctrica en centrales térmicas y la combustión residencial como industrial (Espinoza Molina, 2011). Este contaminante afecta a los pulmones, penetrando hasta las zonas de intercambio de gases, siendo peligroso debido a que es un contaminante cancerígeno (Brunekreef & Holgate, 2002).

Dióxido de Azufre (SO2): es un gas incoloro, no inflamable ni explosivo que en grandes concentraciones causa alteraciones en vías respiratorias así como problemas oculares (Espinoza Molina, 2011). El SO2 causa problemas principalmente a niños,

personas de edad avanzada y personas que sufren problemas cardiovasculares, o respiratorios como son bronquitis o asma (EPA, 2009). El dióxido de azufre se genera por la oxidación del azufre que se encuentran en los combustibles fósiles incluyendo gasolina, diésel, carbón (Espinoza Molina, 2011).

Monóxido de carbono (CO): sin lugar a duda es un contaminante altamente peligroso, siendo un contaminante emitido principalmente por los tubos de escape de vehículos a gasolina. Existe una alta concentración en zonas urbanas debido al tráfico vehicular, también aparece en procesos de combustión en industrias (Espinoza Molina, 2011). El Monóxido de Carbono en gran cantidad y en un tiempo de exposición relevante, puede causar la privación de oxígeno produciendo alteraciones del flujo sanguíneo, perturbación visual, dolor de cabeza, vómito, desmayo, convulsiones y hasta la muerte por asfixia (EPA, 2009).

En Cuenca existe una menor disponibilidad de oxígeno, debido a que se encuentra en una altura considerable sobre el nivel del mar (2550 msnm), permitiendo una mayor emisión de CO debido a los procesos de combustión menos eficientes a comparación de otras ciudades (EMOV, 2017).

OZONO (O3): es un componente natural de la atmósfera, absorbe la mayor cantidad de radiación ultravioleta proveniente del sol, actuando como una capa protectora (Rojas, 2017). Sin embargo, las concentraciones en la troposfera afecta a la salud de las personas causando irritando los ojos, nariz, garganta, dolores de cabeza y dolor pectoral (Espinoza Molina, 2011). Además, afecta el desarrollo normal de las plantas y produce desgaste de materiales como caucho, pinturas y colorantes textiles debido a que es un potente agente oxidante (Boldo, 2016).

1.1.4 Variable Meteorológica

Temperatura: la temperatura es un parámetro físico asociada al movimiento de partículas que caracteriza el calor o transferencia de energía térmica. Muchas propiedades de los materiales dependen del estado de la temperatura, interviniendo en su densidad, presión de vapor, conductividad eléctrica. La temperatura sin lugar a duda es uno de los parámetros más sensibles del clima. Cuando dos sistemas en contacto se encuentran con la misma temperatura se dice que están en el equilibrio térmico, pero cuando existe

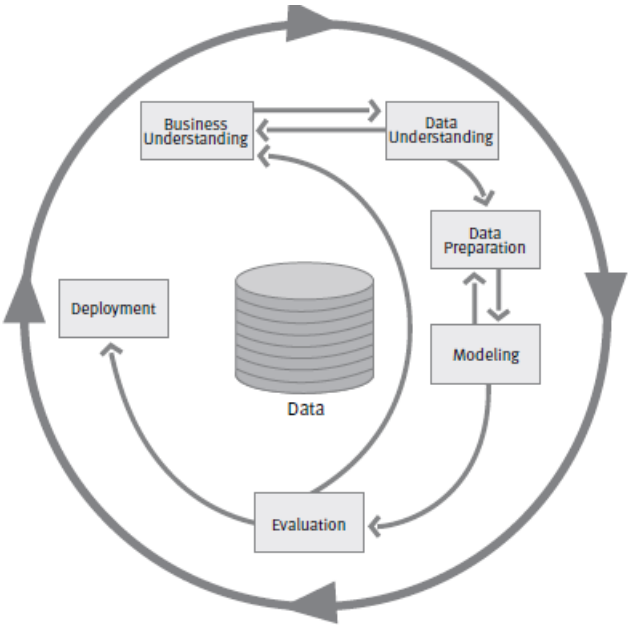
diferencia de temperatura existe el traspaso de energía del sistema con mayor temperatura al menor, hasta llegar al punto de equilibrio térmico (Sierra, 2006).

La temperatura tiene relación con algunos contaminantes atmosféricos, por ejemplo existe un comportamiento similar con el ozono con una correlación de 0.7, de esta manera cuando la temperatura sube, el ozono sube y cuando disminuye tiende a bajar también el contaminante. De igual manera existe relación con los contaminantes dióxido de nitrógeno, dióxido de azufre y el material particulado pm2_5, existiendo un comportamiento inversamente proporcional con estos contaminantes (Andrade, 2018).

1.1.5 Modelo CRISP-DM

Es un método probado para orientar la realización de trabajos sobre minería de datos. Esta metodología tiene un ciclo de vida muy descriptivo, en donde muestra las fases normales de un proyecto, las tareas necesarias para cada fase y explicación entre tareas (IBM, 2012).

Figura 1
Modelo CRISP - DM



Fuente: Pete, C. (2000). Fases del modelo de referencia CRISP-DM. [Figura].

El ciclo de vida de la metodología está compuesto por seis etapas, cada una señalada con flechas como se muestra en la Figura 1, las cuales muestran la trayectoria de las fases y la importancia entre las mismas. La metodología no es secuencial estrictamente, los

proyectos pueden avanzar o retroceder las fases si es necesario, permitiendo ser flexible y personalizable según se necesite, además CRISP-DM permite crear modelos de minería de datos que se adapten a las necesidades concretas (IBM, 2012).

Las etapas o fases con las que cuenta el modelo CRISP-DM son las siguientes: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación. Desde la evaluación se puede analizar si se puede implementar directamente o seguir en el ciclo del modelo para mejorar ciertos aspectos.

Comprensión del negocio.

La comprensión del negocio es indispensable y una de las fases más importantes del modelo, con la comprensión del negocio se sabe hacia dónde va dirigido el análisis, teniendo claro el problema que se necesita solucionar y los objetivos a cumplir, de esta manera en las siguientes fases se permiten obtener los datos correctos e interpretar de manera adecuada los resultados (IBM, 2012; Pete et al., 2000).

Comprensión de datos.

La parte de comprensión implica el primer contacto con los datos, siendo necesario estudiar más detenidamente o de una manera más profunda la información que se tienen para realizar las acciones correspondientes en futuras fases (IBM, 2012). De esta manera se familiariza con los datos sabiendo a que hace referencia cada uno de ellos e identificando la calidad de la información (Pete et al., 2000).

- Recolección de datos: describir las técnicas utilizadas para la recolección de los datos.
- Descripción de los datos: la descripción permite saber o comprender acerca de cada campo que se tiene en el set de datos.
- Verificación de la calidad de datos: se determina la consistencia de los datos obtenidos, la cantidad de registros, los valores nulos o vacíos, valores fuera de rango. Teniendo como objetivo asegurar la corrección de datos.

Preparación de datos

Una vez que se define los objetivos del problema y se obtienen los datos iniciales se debe proceder a tratar la información para que se adecue al estudio que se realizará (Pete et al., 2000). Un trabajo adecuado se logra a partir de una buena selección de datos

y el correcto tratamiento de información. Se estima que el 50-70% de tiempo y esfuerzo se emplea en la preparación de datos (IBM, 2012). Esta fase está relacionada con la fase de modelado interactuando de manera continua las dos fases, esto se debe a que los datos se tienen que tratar de diferentes formas según el modelo seleccionado (Pete et al., 2000).

- Selección de datos: se debe realizar un análisis para obtener solamente los campos del set que se necesitan para el modelado.
- Limpieza de los datos: la limpieza es una de las partes que más tiempo toma, en esta se debe optimizar la calidad de datos que se tienen. Algunas de las técnicas que se realizan en esta etapa son: normalización de datos, estandarizar, tratamiento de valores ausentes, etc.
- Estructurar los datos: la estructuración permite incorporar nuevos registros y convertir valores de atributos ya existentes dentro del conjunto de datos.
- Integrar los datos: crea nuevos campos a partir de otros existentes, fusión de otras tablas o incorporación de datos que sea requeridos para el modelado.
- Formateo de los datos: realizar transformaciones que no alteren al conjunto de datos, se puede realizar acciones como: eliminar comas, ordenar como se necesite, tabular, etc.

Modelado.

El modelado permite utilizar la información obtenida en todos los pasos anteriores, definiendo cual es el modelo adecuado y sus modificaciones para que se incorpore al proyecto que se quiere llevar a cabo, esto se suele ejecutar en múltiples iteraciones normalmente ajustando parámetros en varios modelos hasta llegar a determinar el mejor (IBM, 2012). En la etapa de modelado se debe considerar los objetivos y la relación que tiene con las herramientas, por ejemplo si el problema es de clasificación se puede utilizar herramientas como: árboles de decisión o razonamiento basado en casos; si el problema es de predicción: técnicas de visualización o redes neuronales. De esta manera se selecciona la técnica adecuada para el modelado (Pete et al., 2000).

Evaluación.

Una vez que se llegue a la etapa de evaluación, el proyecto estará terminado casi en su totalidad, en esta etapa simplemente se evaluará si los resultados que se obtienen del

proyecto cumplen con los criterios establecidos en los objetivos comerciales (IBM, 2012). En esta etapa se debe analizar si los resultados que se obtuvieron fueron los deseados y permitir pasar a la siguiente fase, caso contrario se podría decidir en regresar las fases para saber dónde está el error o incluso comenzar desde cero (Pete et al., 2000).

Implementación.

La fase de implementación permite transformar el conocimiento obtenido durante todas las anteriores fases y tomar acciones dentro de las organizaciones. Se debe redactar los puntos más importantes en el estudio explicando los resultados obtenidos en el proyecto, describiendo lo que se hizo bien, se hizo mal, lo que se logró y lo que se puede mejorar (Pete et al., 2000).

1.1.6 Trabajos relacionados

Varios trabajos se han realizado en el ámbito ambiental en diferentes partes del mundo, por ejemplo se elaboró un pronóstico para los siguientes tres días de contaminación del aire en el área metropolitana de Estambul con la aplicación de redes neuronales. Los contaminantes estudiados fueron CO, PM10Y SO2. Para este estudio se trabajó con *set* de datos completos de un año y se pudo comparar los resultados con las predicciones que comparte el sitio web *AirPolTool* debido a que también son predicciones para los siguientes tres días (Kurt et al., 2008).

Otra aplicación con redes neuronales se realizó en el centro de Chile, en donde utilizando una red neuronal *backpropagation* se predijo la temperatura mínima del día siguiente, teniendo como entradas datos referentes al clima como: temperatura, humedad relativa, radiación, precipitación, velocidad y dirección del viento. De esta manera se detecta las heladas en la zona y tiene como objetivo principal ayudar al área de la agricultura para evitar pérdidas en los cultivos. Los datos con los cuales se trabajaron fueron *sets* de datos con series temporales de aproximadamente ocho años, recopilados en la estación de la Red Nacional de Agrometeorología (Fuentes et al., 2018).

El análisis de componentes principales (PCA) y las redes neuronales permitieron de manera integrada tener una buena capacidad predictiva en Malasia. El estudio considera ocho parámetros, monóxido de carbono, ozono, PM10, metano, dióxido de azufre, dióxido de nitrógeno, hidrocarburos no metálicos e hidrocarburos totales,

recolectados de 10 estaciones monitoreadas cada hora durante siete años con un total de 202.080 puntos. Para el análisis se centraron en el reconocimiento de patrones para identificar los parámetros más significativos y para obtener predicciones para los contaminantes del aire con el menor índice de error (Azid et al., 2014).

De igual manera se realizaron pronósticos de los niveles de los contaminantes en la ciudad de Perugia a corto, mediano y largo plazo. Los pronósticos se realizaron con la ayuda de una red neuronal artificial en donde las variables fueron: dióxido de azufre, óxidos de nitrógeno, partículas, benceno, velocidad del viento, entre otras (Viotti et al., 2002). Otra ciudad en la cual se realizó predicciones con redes neuronales fue Hong Kong, con la diferencia que en este caso se adaptó el Algoritmo de Optimización de Enjambre de Partículas (PSO) con el fin de mejorar las predicciones. Como resultado se obtuvo las predicciones de los contaminantes NO2 y NOX en una época de 3 días (Lo et al., 2003).

El presente trabajo propone la utilización de la herramienta de redes neuronales, específicamente la red NARX. La red NARX permite tener una retro-alimentación con los valores verdaderos para su entrenamiento y utilizar las predicciones para imputar datos perdidos, a diferencia de otros trabajos que utilizan redes neuronales para predicciones futuras bien sea de contaminantes o temperaturas mínima. En algunos estudios se evidencia grandes volúmenes de datos de hasta siete años y hasta diez estaciones de monitoreo, lo que les permite contar con mejores condiciones de entrenamiento para la red neuronal. En este trabajo, debido a circunstancias particulares de disponibilidad de la información, incluyo datos de once meses y de una sola estación. Además, las entradas de la red propuesta en este trabajo incluyen datos de tiempo, niveles de contaminantes relacionados, variable meteorológica de temperatura y un valor angular experimental que se crea a partir del nivel inicial del contaminante y el nivel final del mismo en un intervalo de tiempo de 10 minutos.

El modelo utilizado para aplicar minería de datos al problema de valores perdidos en datos relacionados a la contaminación atmosférica fue CRISP- DM, esto debido a que es un modelo muy descriptivo y permite la adecuación en todas sus fases de los requerimientos para el estudio. El modelo es uno de los más utilizados en cuanto a trabajos de minería de datos.

2. MÉTODO

2.1 Zona de estudio

Este trabajo de titulación se centró específicamente en datos de la ciudad de Cuenca – Ecuador (Figura 2). Cuenca se encuentra ubicada en la región sierra, perteneciente a la provincia del Azuay. Esta ciudad está localizada a 2.560 metros sobre el nivel de mar, en las coordenadas 2° 53′ 51″S y 79° 00′ 16″ O. Cuenca cuenta con un clima subtropical de montaña, que se caracteriza por abundantes precipitaciones y temperaturas regularmente frías.

Figura 2
Ubicación del Cantón Cuenca

Fuente: EMOV. (2017). Ubicación Del Cantón Cuenca.

2.2 Materiales

Los datos que fueron utilizados para realizar la investigación se obtuvieron desde la estación automática de monitoreo del aire con coordenadas 2° 89′ 74″S y 79° 00′ 31″S (Figura 3), estación que pertenece a la empresa EMOV EP (Red de Monitoreo de Calidad del Aire de Cuenca de la EMOV EP), en la cual se miden y almacenan los datos referentes a la calidad del aire dentro de la ciudad de Cuenca (EMOV, 2017). Los datos con los que se trabajaron en la investigación pertenecieron al periodo enero del 2018 hasta noviembre del 2018. Los contaminantes utilizados fueron los siguientes:

Contaminantes Atmosféricos

- Ozono (O3).
- Monóxido de Carbono (CO).
- Dióxido de nitrógeno (NO2).
- Dióxido de Azufre (SO2).
- Material Particulado (PM2.5).

Variables Meteorológicas

- Temperatura del Aire.

Figura 3
Estación Automática de Calidad del Aire y Meteorología

Fuente: EMOV. (2017). Estación automática de calidad del aire y meteorología localizada en la estación MUN. [Figura]

2.3 Metodología

2.3.1 CRISP-DM

Para el desarrollo del trabajo se utilizó la metodología CRISP-DM (*Cross-Industry Standard Process for Data Mining*), analizando que se adecua de mejor manera al estudio en comparación con otras metodologías como son KDD y SEMMA. Con esta metodología lo que se realizó es dividir por partes el proyecto para tener una forma escalonada de avanzar, partiendo del objetivo del proyecto, lo cual consiste en tener claro

qué es lo que se debe lograr, luego obtener los datos necesarios tanto de contaminantes como de la variable meteorológica, posteriormente tratar los datos, para adecuar a la manera en la cual se necesita esta información, luego realizar el entrenamiento de la red neuronal, parte fundamental del proyecto, siguiendo con el relleno de datos faltantes y por último la evaluación cumpliendo con el objetivo del trabajo.

2.3.2 Comprensión del negocio

En la comprensión del negocio es importante destacar que no se encontró trabajo existente relacionado con valores angulares para imputación de datos, en donde se permitió rellenar o imputar datos faltantes con una nueva técnica experimental. Esto se determinó mediante la búsqueda en repositorios como son SCIELO, GOOGLE SCHOLAR, en donde no se encontró trabajos de imputación de datos que utilicen redes neuronales conjuntamente con una técnica angular, niveles de contaminantes y nivel de la variable meteorológica de temperatura.

2.3.3 Comprensión de los datos

Recolección de datos: los datos fueron obtenidos mediante la estación de monitoreo del aire de la empresa EMOV. Los datos fueron proporcionados por el Instituto de Estudios de Régimen Seccional del Ecuador (IERSE) en donde se manifestó que hubo un primer tratamiento de los contaminantes. Este tratamiento consistió en obtener la media de los valores para tener los datos cada minuto en lugar de cada segundo. Posteriormente se realizó un segundo tratamiento a los valores de los contaminantes obteniendo nuevamente la media, esta vez para obtener valores cada 10 minutos, debido a que los valores de las variables meteorológicas tienen dicha granularidad.

Descripción de los datos: a continuación se describen los datos incluyendo una explicación del significado de todos los campos que incluye en el conjunto de datos.

La Tabla 1 muestra los campos referentes al tiempo en el cual fueron obtenidos los datos.

Tabla 1
Campos de Tipo Fecha

EC_TIME_STAMP	Mes	Día	Hora	Min	Día_sem
1/1/2018 0:10	1	1	1	10	1
1/1/2018 0:20	1	1	1	20	1

1/1/2018 0:30	1	1	1	30	1

Fuente: Elaboración Propia

Todos los contaminantes atmosféricos se representan en la Tabla 2, O3 hace referencia a los valores del ozono, el CO significa al nivel de monóxido de carbono, NO2 los valores del dióxido de nitrógeno, SO2 dióxido de azufre y por último PM2_5 que representa los valores de material particulado. Todos los contaminantes tienen como unidad de medida "partes por billón" a diferencia del material particulado el cual es medido en "Microgramos por metro cúbico".

Tabla 2
Campos de Contaminantes

O3	CO	NO2	SO2	PM2_5
9,4654	1,0113	54,869	20,668	43,2
10,733	1,0155	54,548	24,039	43,2
12,663	0,9606	53,897	35,271	43,12

Fuente: Elaboración Propia

Las variables meteorológicas que se muestran en la Tabla 3 tienen el siguiente significado; TEMPAIRE_AV temperatura, HR_AV humedad relativa, DP_AV punto de rocío, PATM_AV presión atmosférica, RADGLOBAL_AV radiación global, PRECIP_SUM precipitación, WINDSPEED_AV velocidad del viento, WINDDIR_AV dirección del viento, UVA_AV radiación ultravioleta A y UVE_AV radiación ultravioleta E.

Tabla 3
Campos de Variables Meteorológicas

TEMPAIRE_AV	HR_AV	DP_AV	PATM_AV	RADGLOBAL_AV	PRECIP_SUM	WINDSPEED_AV	WINDDIR_AV	UVA_AV	UVE_AV
12,2	87	10	752	0	0	1,1	51	0	0,00278
12,4	85	10	752	0	0	0,9	42	0	0,00278
12,5	84	9,9	752	0	0	0,6	66	0	0,00281

Fuente: Elaboración Propia

Verificación de la calidad de datos: los datos recolectados corresponden a registros tomados cada 10 minutos desde el uno de enero del 2018 hasta el 31 de

noviembre del 2018. El total de registros obtenidos fue de 48.095 en donde existen varios registros vacíos que fueron tratados en la fase de preparación de datos.

2.3.4 Preparación de datos

La información que se obtuvo en el conjunto de datos fue de todos los contaminantes atmosféricos y variables meteorológicas, así como información perteneciente al tiempo en la cual se tomó los datos.

Selección de datos: Con la finalidad de decidir cuáles son las variables de contaminantes relevantes para la construcción de la red neuronal, se analizaron estudios anteriores realizados por (Andrade, 2018; Ortega, 2018). Estos trabajos también fueron realizados en el ámbito de la calidad del aire en la ciudad de Cuenca.

Existen patrones de comportamiento entre ciertos contaminantes, estas relaciones están presentes en cada uno de ellos de manera distinta, no todos intervienen de igual manera, de esta forma depende del contaminante que se quiere analizar para definir la relación con el resto.

Ozono (O3)

El ozono tiene una correlación con el dióxido de nitrógeno (NO2), en algunos casos hasta llega a una correlación mayor a 0.45. De igual manera el contaminante dióxido de azufre (SO2) presenta una buena correlación, en algunos casos llegando a 0.44. Y por último teniendo una relación indirecta el ozono con el contaminante monóxido de carbono (CO) (Ortega, 2018).

Dióxido de Nitrógeno (NO2)

El dióxido de nitrógeno tiene una relación directa con el dióxido de azufre (SO2), de igual manera existe una correlación directa que supera 0.5 con el contaminante monóxido de carbono (CO) y una fuerte correspondencia con el material particulado PM2_5 (Ortega, 2018).

Dióxido de Azufre (SO2)

El dióxido de azufre presenta una relación más armónica con el monóxido de carbono (CO) y teniendo una correlación positiva con el material particulado PM2_5 (Ortega, 2018).

Monóxido de Carbono (CO)

El monóxido de carbono tiene relación con tres contaminantes, como son: material particulado PM2_5 con el cual se tiene una correlación positiva y similitud en los gráficos, dióxido de azufre (SO2) con una correlación cercana a 0.44 y con el contaminante ozono (O3) con el cual tiene una correlación positiva de 0.45 (Ortega, 2018).

Material Particulado PM2_5

Por último el contaminante material particulado PM2_5 está relacionado con el monóxido de carbono (CO) con el cual tiene una relación en gran parte y el dióxido de azufre (SO2) con una correlación positiva entre 0.44 (Ortega, 2018).

La variable meteorológica de la temperatura se involucra en varios contaminantes, analizando se obtuvo que tiene relación con el ozono (O3), dióxido de nitrógeno (NO2), dióxido de azufre (SO2), y material particulado PM2_5 (Andrade, 2018).

Tabla 4
Intervención de Contaminantes Atmosféricos y Temperatura

Contaminante	O3	NO2	SO2	CO	PM2_5	TEMP
O3		X	X	X		X
NO2			X	X	X	X
SO2				X	X	X
CO	X		X		X	
PM2_5			X	X		X

Relación existente tanto entre los contaminantes atmosféricos como con la variable temperatura. Fuente: Elaboración Propia.

La selección de datos permitió tener un conjunto solamente con datos relevantes y necesarios para el este trabajo. El conjunto original de datos tenía información de tiempo como es: fecha, mes, día, hora, minuto, día de la semana. Información de contaminantes atmosféricos como el: Ozono, Monóxido de Carbono, Dióxido de Azufre, Dióxido de Nitrógeno, Material Particulado PM2_5. Y de variables meteorológicas como son: Temperatura, Humedad Relativa, Punto de Rocío, Presión Atmosférica, Radiación Global, Precipitación, Velocidad del Viento, Dirección del Viento, Radiación Ultravioleta A y Radiación Ultravioleta E. De todos estos campos solamente se necesitó

información de los contaminantes, la fecha y la variable meteorológica de temperatura, desechando o eliminando el resto de información.

De esta manera los campos con los cuales se formó el nuevo conjunto de datos son los que se muestran en la Tabla 5. Para cada contaminante se creó un conjunto de datos diferente, esto debido a que las entradas que intervienen para cada contaminante son diferentes, como se ilustra en la Tabla 4.

Tabla 5
Conjunto de Datos Modificado

EC_TIME_STAMP	Día_sem	O3	CO	NO2	SO2	PM2_5	TEMPAIRE_AV
1/1/2018 0:10	1	9,4654088	1,0113283	54,8690290	21	43	12,2
1/1/2018 0:20	1	10,733239	1,0155263	54,5482279	24	43	12,4
1/1/2018 0:30	1	12,662551	0,9605711	53,897223	35	43	12,5

Fuente: Elaboración Propia

Limpieza de los datos: fue necesario realizar la limpieza de registros con la eliminación de valores ausentes y tratamiento de valores atípicos.

La eliminación de registros que contienen valores nulos o ausentes permite tener datos consistentes y de esta manera evitar alterar los resultados (Matute Rivera, 2018). Se realizó la eliminación de vacíos en las entradas predictores para cada contaminante, analizando que la red neuronal para tener un buen entrenamiento necesita que todas sus entradas estén completas, mientras que la variable de salida consta tanto de valores como de celdas vacías que posteriormente se imputaran.

El conjunto de datos completo estaba compuesto de 48.095 registros. Luego del proceso de limpieza de datos para tratar para cada contaminante eliminando los registros que contenían celdas vacías en las variables de entrada, los conjuntos se vieron modificados como se representa en la Tabla 6.

Tabla 6
Número de Registros Actuales

Contaminante	Registros Actuales	Porcentaje %
O3	35.235	73,26
CO	42.542	88,85

NO2	38.015	79,04
SO2	38.898	80,87
PM2_5	38.711	80,48

Los valores atípicos son valores de la serie que se encuentran fuera del rango normal y pueden causar problemas al momento de tratar los datos, para la eliminación de estos valores se utilizó el método de puntuación Z (Z - Score). El método de puntuación Z es un método popular muy utilizado en el tratamiento de valores atípicos, se basa en la distribución normal en donde utiliza el promedio y la desviación estándar en su fórmula, la regla indica que los valores absolutos calculados de Z mayores a 3 desviaciones estándar son valores atípicos (Garcia, 2012).

$Z\ Score = \frac{D - P}{DS}$	(Ecuación 1)

En la Ecuación 1 (Garcia, 2012), D representa el dato original; P el promedio; y DS la desviación estándar. Para verificar si es o no un valor atípico se debe calcular el valor absoluto del resultado y si es mayor a 3 indicaría que está fuera del rango normal de los datos, por lo tanto es considerado valor atípico.

Tabla 7
Cálculos para Valores Atípicos

Contaminante	**Promedio**	**Desviación Estándar**	**Coeficiente de variación**	**Valores atípicos encontrados**	**Porcentaje %**
O3	24.02	17.30	72.04	270	0.76
CO	0.8	0.4	49	672	1.57
NO2	17	11	63	250	0.65
SO2	10	9.1	90	691	1.77
PM2_5	10	7.7	76	530	1.36

Valores calculados en el conjunto de cada contaminante y representación de la cantidad de valores atípicos en porcentaje, referente al conjunto completo de datos. Fuente: Elaboración Propia.

Estructurar los datos: la conversión de atributos o valores ya existentes en el conjunto de datos se realiza en la fase de estructurar los datos, en esta fase se estandarizó las variables de entrada, de esta manera quedando sin unidad de medida, debido a que se eliminan al tener en la fórmula tanto en la parte del numerador como en el denominador la misma unidad. La fórmula utilizada es la siguiente:

$VE = \frac{D - P}{DS}$	(Ecuación 2)

De igual manera que en la Ecuación 1, D representa el dato original; P el promedio; DS la desviación estándar y con la diferencia de VE, que representa el valor estandarizado en la Ecuación 2 (Avila, De Hernández, Rodríguez Pérez, & Caraballo, 2012).

En la Tabla 8 se muestra un ejemplo de los valores sin estandarizar y en la Tabla 9 los valores estandarizados.

Tabla 8
Campos Sin Estandarizar

EC_TIME_STAMP	Día_sem	CO	NO2	SO2	TEMPAIRE_AV	O3
1/1/2018 0:20	1	1,016	54,55	24	12,4	11
1/1/2018 0:30	1	0,961	53,9	35	12,5	13
1/1/2018 0:40	1	1,187	59,71	33	12,5	13

Fuente: Elaboración Propia.

Tabla 9
Campos Estandarizados

Fecha Final	Día_sem	CO	NO2	SO2	TEMPAIRE_AV	O3
1/1/2018 0:20	1	0,672	3,443	1,6	-0,825688748	11
1/1/2018 0:30	1	0,528	3,383	2,8	-0,793094749	13
1/1/2018 0:40	1	1,121	3,92	2,6	-0,793094749	13

Fuente: Elaboración Propia.

Integrar los datos: la incorporación de nuevos atributos como el valor angular, y demás modificaciones en el conjunto de datos permitieron manejar de una mejor manera la información. De esta manera se obtuvo un nuevo *set* de datos en donde consta la fecha

inicial, fecha final y los niveles de los contaminantes que se relacionan con el contaminante a imputar.

Tabla 10
Incorporación de Nuevos Campos

Fecha Inicial	Fecha Final	Día_ sem	O3 Ini	O3 Fin	Ángulo	SO2 Ini	SO2 Fin	Ángulo	PM2_5 Ini	PM2_5 Fin	Ángulo	CO
1/1/2018 0:10	1/1/2018 0:20	1	-0,8079	-0,7339	0,424	1,1794	1,5542	2,1465	4,338678	4,338676	0	1,0113
1/1/2018 0:20	1/1/2018 0:30	1	-0,7339	-0,6213	0,6451	1,5542	2,8031	7,1186	4,338678	4,328214	-0,06	1,0155
1/1/2018 0:30	1/1/2018 0:40	1	-0,6213	-0,6025	0,1081	2,8031	2,5874	-1,235	4,328214	4,300747	-0,1574	0,9606

Fuente: Elaboración Propia.

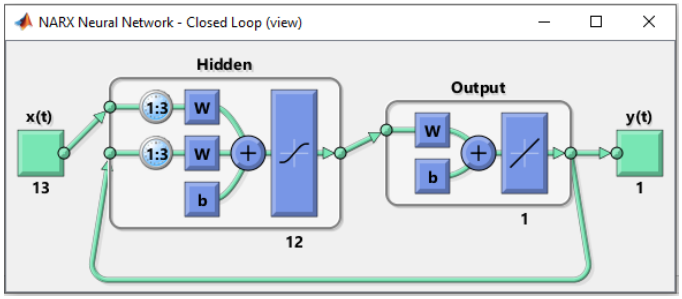
Para obtener el valor del ángulo se aplicó la fórmula trigonométrica del arco tangente. Para la aplicación de esta fórmula se necesitó el valor del lado opuesto que se obtiene de la diferencia entre el valor final del contaminante y el valor inicial y el adyacente el cual tiene un valor constante de 10, debido a que es el valor del tiempo en el cual se toma los datos.

2.3.5 Modelado

Para el modelado se utilizó la herramienta de redes neuronales, específicamente la red NARX, la cual es una red estándar de dos capas con líneas de retardo. Además es una red de retro-propagación la cual tiene como alimentación las salidas verdaderas, por lo tanto es una red supervisada. La red NARX consta de dos maneras de ejecutarse red de bucle cerrado la cual se retroalimenta con los valores que se predicen, la segunda es la red de bucle abierto la cual se retroalimenta con los valores verdaderos (MATLAB & Simulink, 2019).

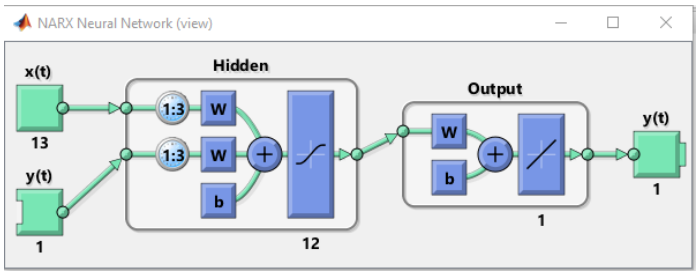
Para realizar el entrenamiento se usó un conjunto completo de datos, tanto con las variables de entrada como con los targets, es decir sin ningún tipo de celda vacía. Se seleccionó una red de bucle abierto, debido a que se tenía todos los valores verdaderos para retroalimentar a la red.

Figura 4
Red de Bucle Cerrado



Fuente: Elaboración Propia.

Figura 5
Red de Bucle Abierto



Fuente: Elaboración Propia.

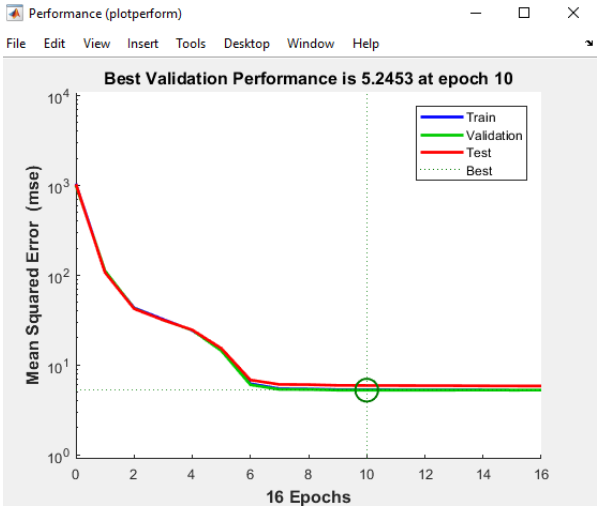
Para seleccionar el algoritmo de entrenamiento se realizaron pruebas, para determinar cuál es el que mejor se adecua al estudio, dando como resultado el algoritmo de "*levenberg marquardt*".

Para definir la división del conjunto de datos se tomó el siguiente criterio; 70% para el entrenamiento de la red neuronal, el 15% para validación y el 15% restante para las pruebas. Esta división del *set* de datos se tomó como referencia a otros trabajos de rellenos o predicciones los cuales dividían de esta manera los datos (Fernanda & Logroño, 2018) y también por defecto en la generación del código de la red neuronal el cual definía de esa misma manera.

El entrenamiento más adecuado se obtuvo mediante prueba y error. Varias gráficas de Matlab permitieron analizar la manera en la cual estaba entrenada la red y quedarse con los mejores parámetros. Es importante mencionar que cada vez que se entrena la red, aun con los mismos parámetros los resultados no serán exactamente iguales para cada entrenamiento.

Una gráfica importante es la del "*Performance*", la cual permite visualizar el desempeño de la red, mientras más cercanas estén las líneas de validación y pruebas quiere decir que el entrenamiento de la red está realizada de manera correcta, sin presentar sobreajustes (MATLAB, 2018). En este caso se muestra la Figura 6 que pertenece al entrenamiento de la red NO2.

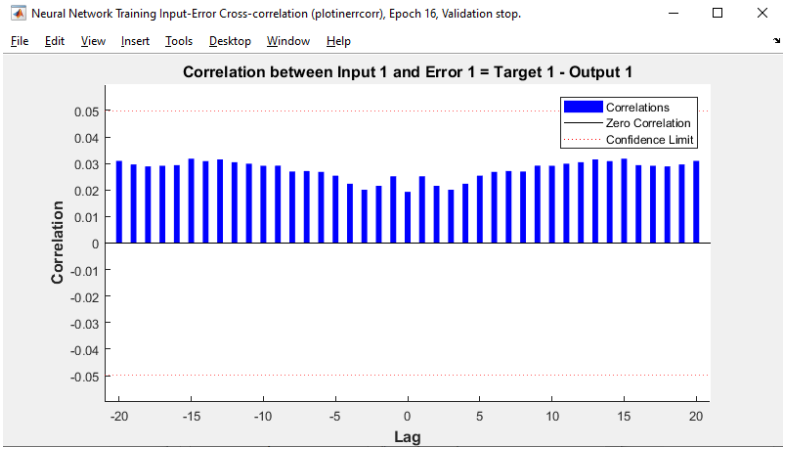
Figura 6
Rendimiento



Fuente: Elaboración Propia.

De igual manera otra grafica importante es la de "*Input-Error Cross-Correlation*", la cual indica como los errores tienen una correlación con las entradas. Para saber si se tiene un modelo predictivo adecuado los valores se deben encontrar dentro del rango que se indica con las líneas entrecortadas, caso contrario si los valores no se acerca al rango se pueden modificar los parámetros principalmente de retardos para adecuar de mejor manera el modelo (MATLAB, 2018).

Figura 7
Error de Correlación NO2



Fuente: Elaboración Propia.

Todos estos pasos se tuvieron que realizar para cada contaminante, debido a que cada uno tiene una manera distinta de entrenamiento, con diferentes entradas y distintos parámetros.

Para ver el código con el que se creó la red se puede ver en el Anexo A.

2.3.6 Evaluación

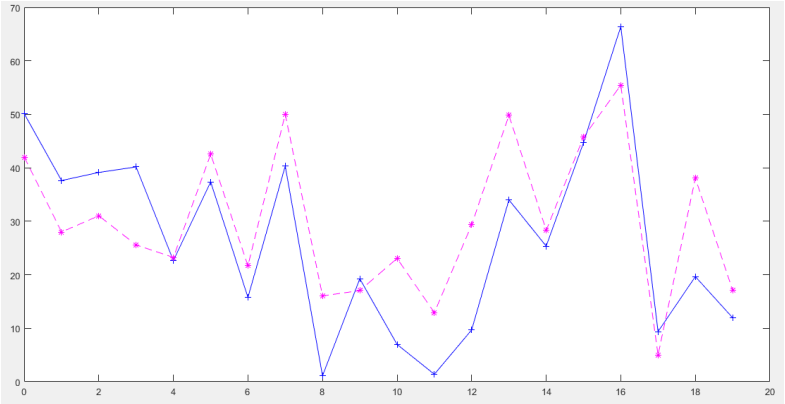
En esta fase de la metodología CRISP-DM se analiza sí los resultados obtenidos del proyecto cumplen con los criterios establecidos en los objetivos comerciales.

La evaluación se determinó mediante el valor de R – Ajustado para cada conjunto de datos de prueba que utilizo la red neuronal. El valor de resultado R – Ajustado, mientras más cercano a uno se encuentre, mejor relación existe entre los valores verdaderos y los valores que predice la red. Otro parámetro utilizado es la raíz del error cuadrático medio (RMSE), el cual calcula el error entre los valores verdaderos y las predicciones. Estos dos parámetros han sido utilizados para evaluar las predicciones que realiza la red. En la sección de resultados se puede evidenciar estas pruebas realizadas de una manera más detallada.

Para ampliar la evaluación se realizó la toma de 20 datos de manera aleatoria para comprobar la red de cada uno de los contaminantes, graficando los puntos verdaderos y los que se predicen con la utilización de la red neuronal NARX. En la gráfica se representa

con línea azul los valores verdaderos, mientras que la línea entrecortada a los valores que predice la red.

Figura 8
Gráfica de Evaluación Ozono



Fuente: Elaboración Propia.

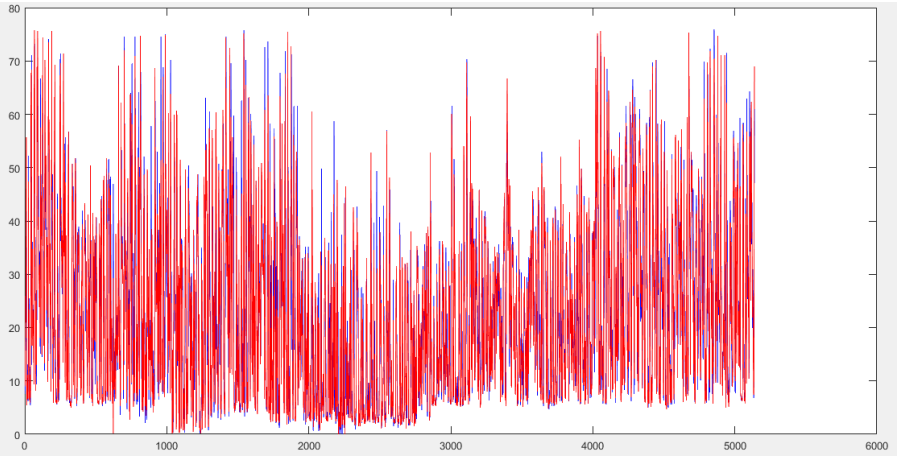
La manera de evaluar con solamente 20 datos se estableció debido a que la cantidad de valores totales es muy grande y no permite visualizar con claridad las gráficas correspondientes. Analizando las gráficas se puede observar que se cumple con una adecuada imputación de datos, debido a que estos valores fueron borrados del conjunto de datos, apareciendo como vacíos y sin embargo se asemejan al comportamiento de los valores verdaderos.

Para poder visualizar todas las evaluaciones, correspondientes para cada contaminante se puede acceder al Anexo B.

3. RESULTADOS Y DISCUSIÓN

Una vez realizada la fase de evaluación, se analizaron los resultados obtenidos para cada contaminante. Para analizar cuál fue el contaminante con mejores predicciones para la imputación de datos se tomó en cuenta el conjunto de prueba, esto debido a que la red neuronal internamente elimina estos valores y no los toma en cuenta para el entrenamiento, de esta forma se pudo comparar las predicciones con los valores verdaderos en una buena cantidad con el 15% del conjunto total de datos completo. En la Figura 9 se evidencia el comportamiento de los dos conjuntos. Los datos verdaderos están representados con el color azul, mientras que los datos que se predicen se muestran en color rojo.

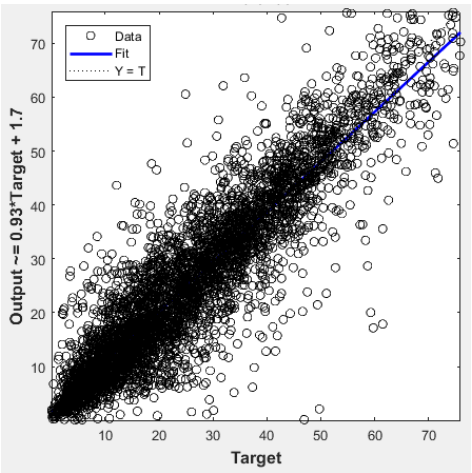
Figura 9
Visualización de Datos del Conjunto de Pruebas - Ozono



Fuente: Elaboración Propia.

La manera de evaluar las predicciones fue mediante el R cuadrado ajustado, en el caso del ozono se obtuvo un R de 0.8537 y un RMSE de 6.29, lo cual indica que los valores que se predicen para la imputación son cercanos a los verdaderos, pudiéndose observar en la Figura 10 el comportamiento central de gran cantidad de datos.

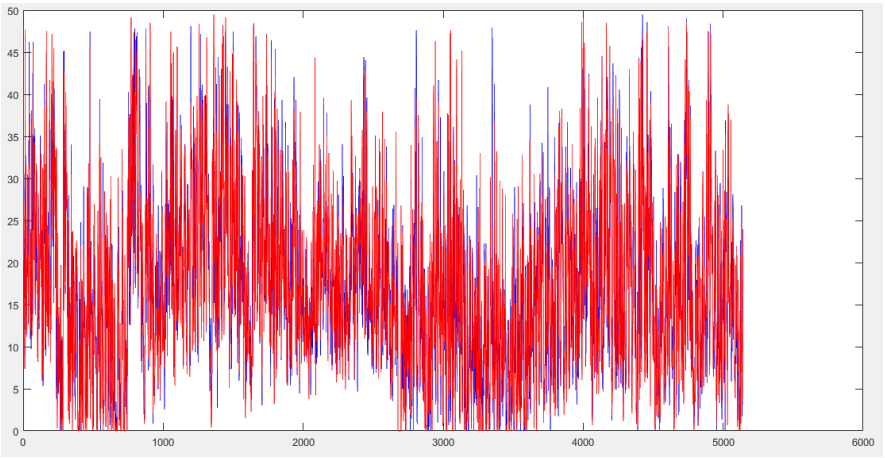
Figura 10
Correlación del Ozono en el Conjunto de Prueba



Fuente: Elaboración Propia.

Otro de los contaminantes con buenos resultados fue el dióxido de nitrógeno (NO2). De igual manera se puede observar en la Figura 11 el comportamiento parecido que tienen los valores en el conjunto de pruebas, también representando el 15% del conjunto completo total.

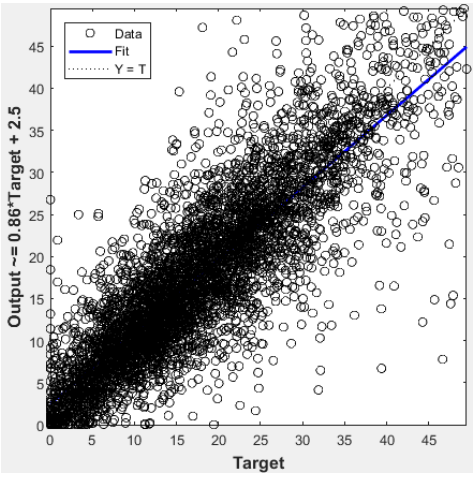
Figura 11
Visualización de Datos del Conjunto de Pruebas - Dióxido de Nitrógeno



Fuente: Elaboración Propia.

Los datos del conjunto se pueden observar en la Figura 12 que se encuentran más dispersos en comparación con el ozono, sim embargo existe una adecuada centralización de datos, obteniendo un valor R ajustado de 0.7370 y un RMSE de 5.32.

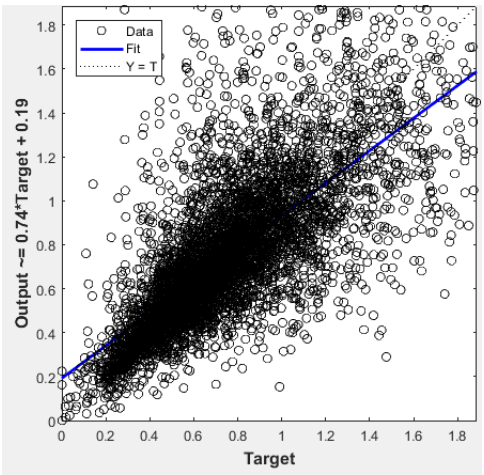
Figura 12
Correlación del NO2 en el Conjunto de Prueba



Fuente: Elaboración Propia.

Mientras que para el resto de contaminantes como son: monóxido de carbono, dióxido de azufre y material particulado pm2_5 se obtuvieron resultados más bajos que el O3 y NO2. Los datos del conjunto de pruebas se encontraron más dispersos, dando valores de R ajustado de 0.53 para el monóxido de carbono, 0.49 para el dióxido de azufre y 0.55 para el material particulado pm2_5.

Figura 13
Correlación del CO en el Conjunto de Prueba



Fuente: Elaboración Propia.

Para visualizar los resultados de evaluación de todos los contaminantes se puede acceder al Anexo C.

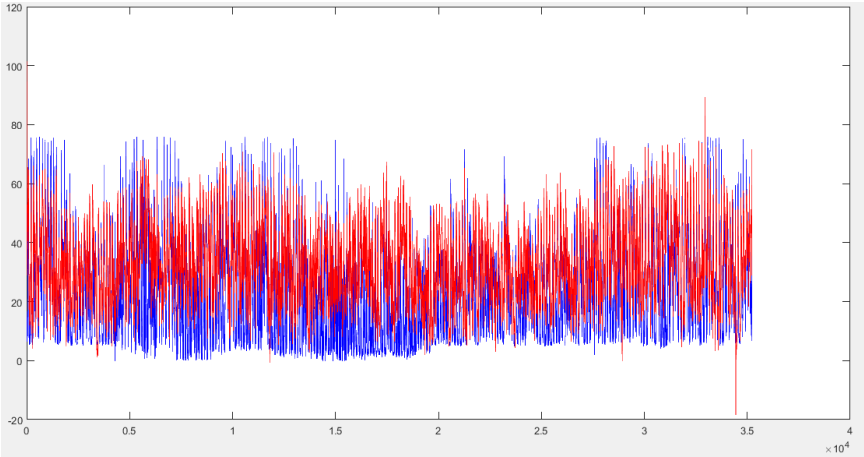
Una manera adecuada de visualizar los resultados es con gráficos, que permite entender de una manera más fácil, mostrando los valores verdaderos, las predicciones y el relleno de los datos en cada contaminante.

Para la representación de todo el conjunto de datos se graficó con colores azul para los datos verdaderos del conjunto y rojo para las predicciones. En las figuras de relleno se representa los datos verdaderos con verde y negro para la imputación de datos, de esta manera tratando de resaltar la imputación y visualizar de mejor manera.

Las mejores imputaciones con predicciones se realizaron en los contaminantes de ozono (O3) y dióxido de nitrógeno (SO2), sin embargo para el resto de contaminantes se puede observar una adecuada manera de imputar.

En la Figura 14 se puede observar el comportamiento parecido que existe entre las dos líneas, esto quiere decir que los datos que se predicen se apegan a los datos verdaderos en gran medida.

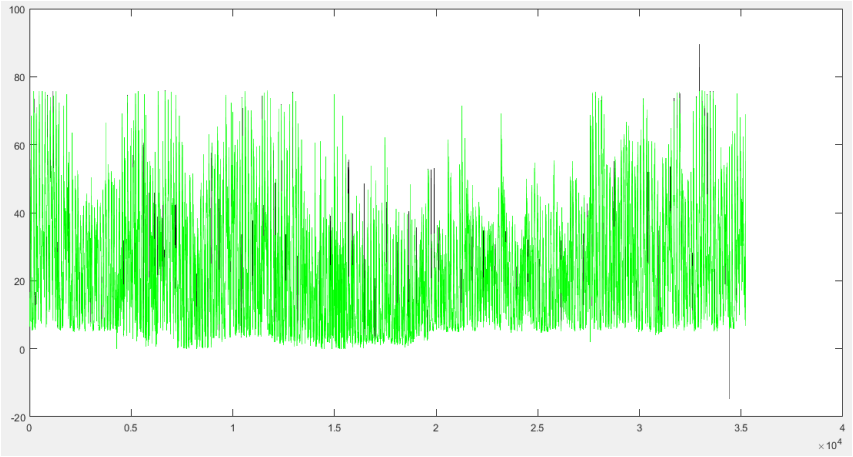
Figura 14
Conjunto de Datos Completo - O3



Fuente: Elaboración Propia.

La imputación de datos se puede observar en la Figura 15, en donde los puntos negros son los datos vacíos que están siendo rellenados, analizando de una manera visual se puede observar que estos datos siguen una secuencia lógica que se adapta a los datos verdaderos que están siendo representados con color verde.

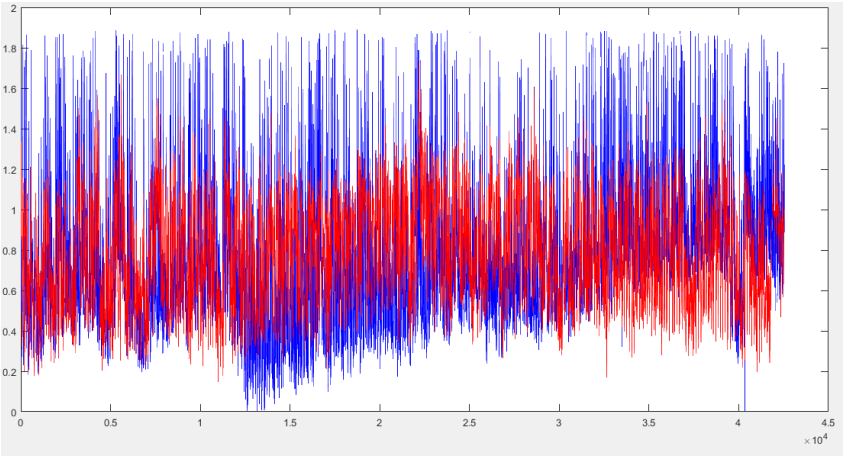
Figura 15
Imputación de Datos del Conjunto Completo - O3



Fuente: Elaboración Propia.

En la Figura 16 que representa los valores verdaderos y las predicciones del monóxido de carbono (CO), se puede observar que no se apega tanto a los valores verdaderos sobre todo en los picos más altos, en la figura se observa que los datos que se predicen se encuentran más centralizados en la gráfica.

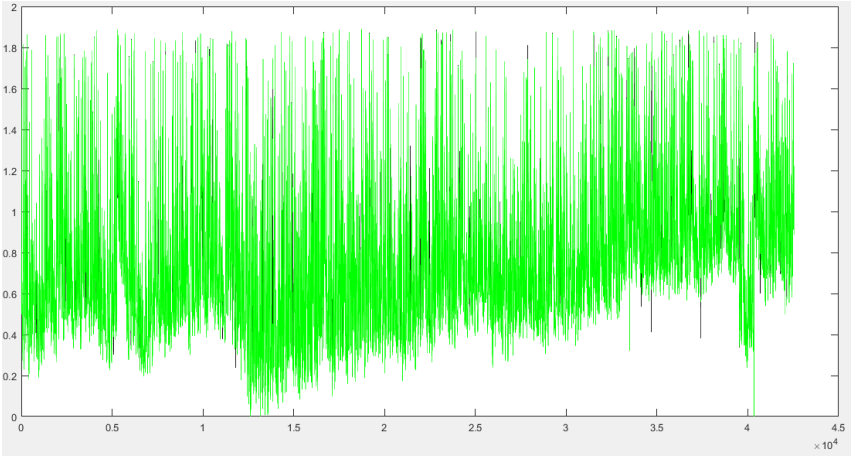
Figura 16
Conjunto de Datos Completo - CO



Fuente: Elaboración Propia.

Sin embargo para la imputación de los valores del monóxido de carbono se puede observar que no se distorsionan drásticamente con el comportamiento normal de los datos verdaderos (Figura 17).

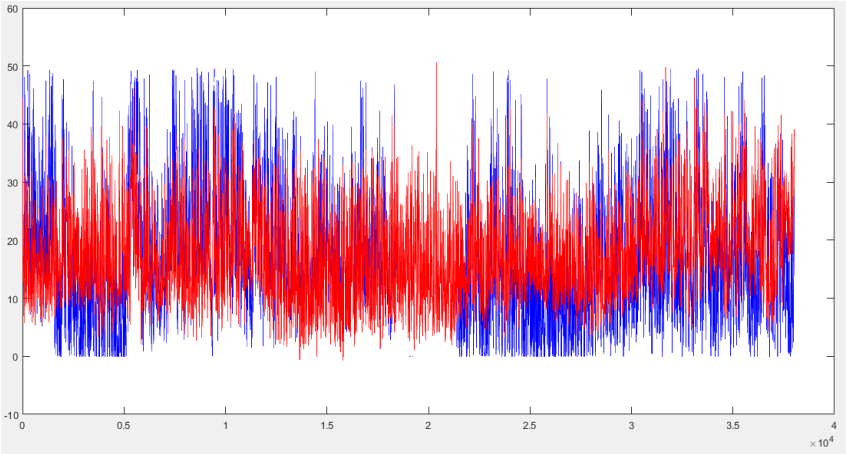
Figura 17
Imputación de Datos del Conjunto Completo - CO



Fuente: Elaboración Propia.

Uno de los contaminantes con más datos perdidos que se pudo evidenciar fue el dióxido de nitrógeno (NO2). Los puntos que se predijeron tienen un comportamiento muy cercano a los valores verdaderos, la raíz del error cuadrático medio del conjunto completo de datos fue de 9.03, lo cual representa un buen valor en comparación al número de datos del conjunto total.

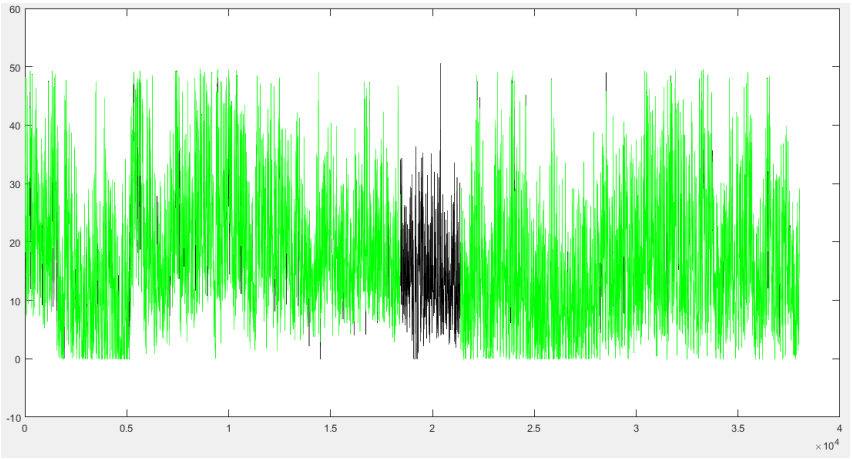
Figura 18
Conjunto de Datos Completo - NO2



Fuente: Elaboración Propia.

En la Figura 19 se puede evidenciar que un gran tramo de datos se encuentra vacío, pero la imputación ha sido adecuada para el caso, siguiendo el comportamiento de los datos verdaderos y permitiendo rellenar toda esta parte perdida.

Figura 19
Imputación de Datos del Conjunto Completo - NO2

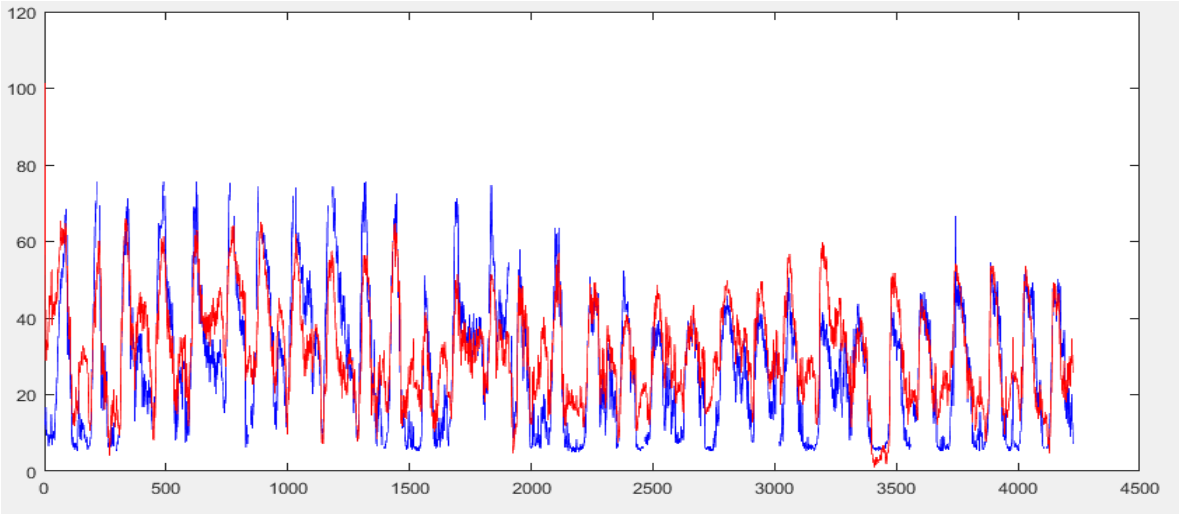


Fuente: Elaboración Propia.

Todos estos casos descritos tienen un gran volumen de datos, lo cual no permite visualizar de una manera adecuada los resultados, por lo tanto se realizó la división por meses para representar una menor cantidad de datos y mejorar la comprensión de los resultados, de esta manera visualizando tanto el comportamiento de los datos verdaderos, las predicciones y sobre todo la imputación de datos en meses que se han visto afectados por gran cantidad de datos vacíos. En el Anexo D, se puede visualizar las gráficas con el conjunto de datos completo para cada uno de los contaminantes.

La Figura 20 representa los valores del contaminante ozono en el mes de enero, trazando los valores verdaderos y valores que predice la red, en donde se puede evidenciar el seguimiento de la curva, aunque no sea idéntica se apega mucho a los valores verdaderos y al comportamiento de la misma, de igual manera la línea azul muestran los valores verdaderos y la línea roja los valores que se predicen.

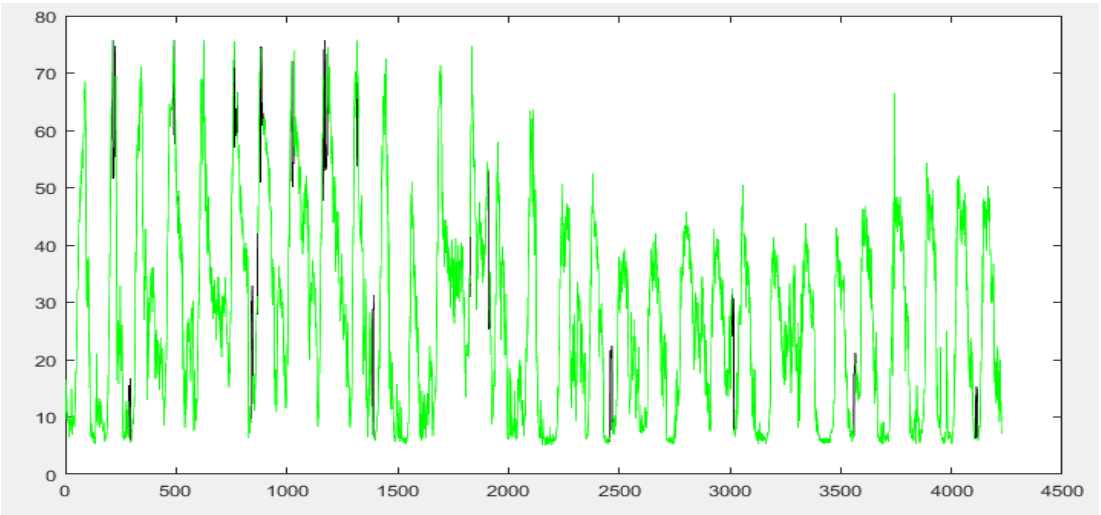
Figura 20
Datos del Mes de Enero - O3



Fuente: Elaboración Propia.

Una vez que se ha visto el comportamiento de la red en el mes de enero, se indica la imputación de los datos vacíos, los cuales están representados por el color negro y se observa una similitud con el resto de la gráfica (Figura 21).

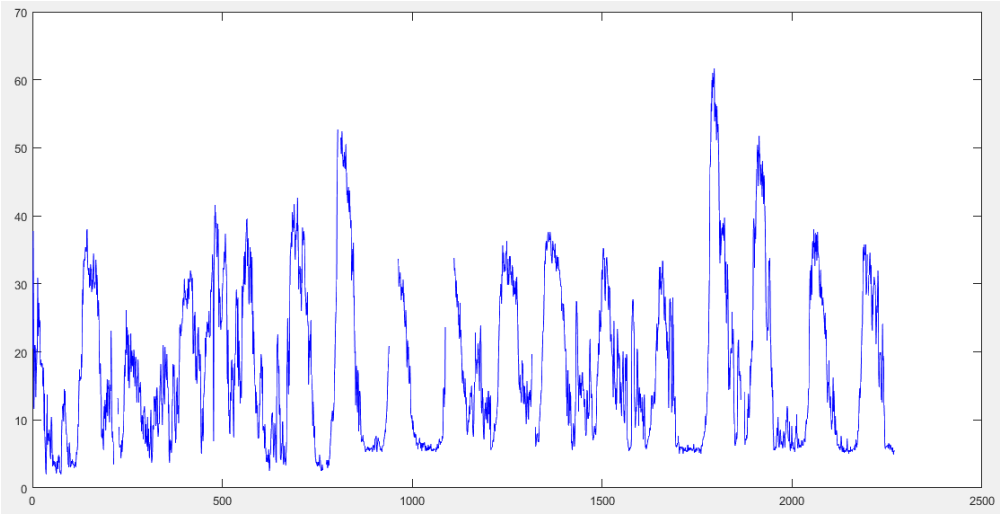
Figura 21
Imputación del Mes de Enero - O3



Fuente: Elaboración Propia.

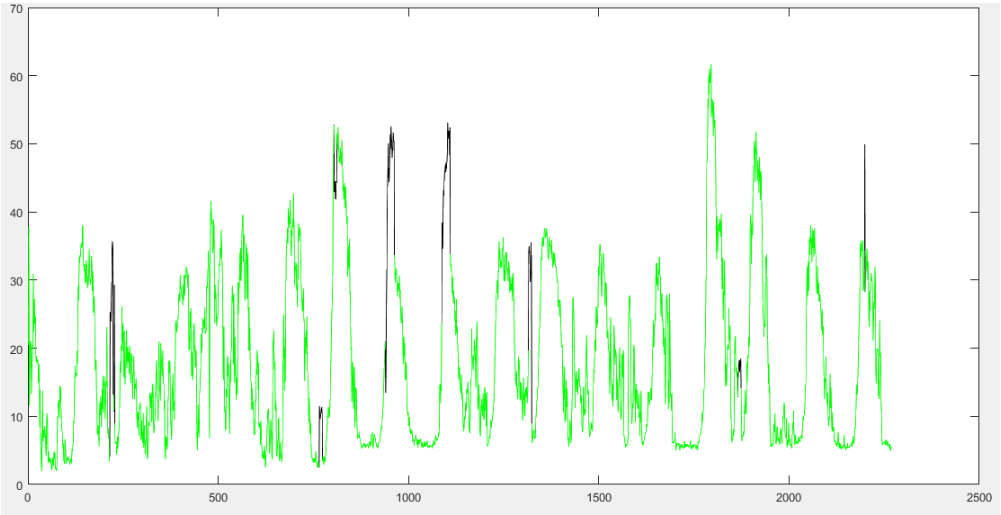
Uno de los meses con más valores perdidos del contaminante del ozono fue el mes de junio, el cual se puede observar el la Figura 22, quedando vacíos que se imputaron con las predicciones realizadas por la red (Figura 23).

Figura 22
Valores Verdaderos Mes de Junio - O3



Fuente: Elaboración Propia.

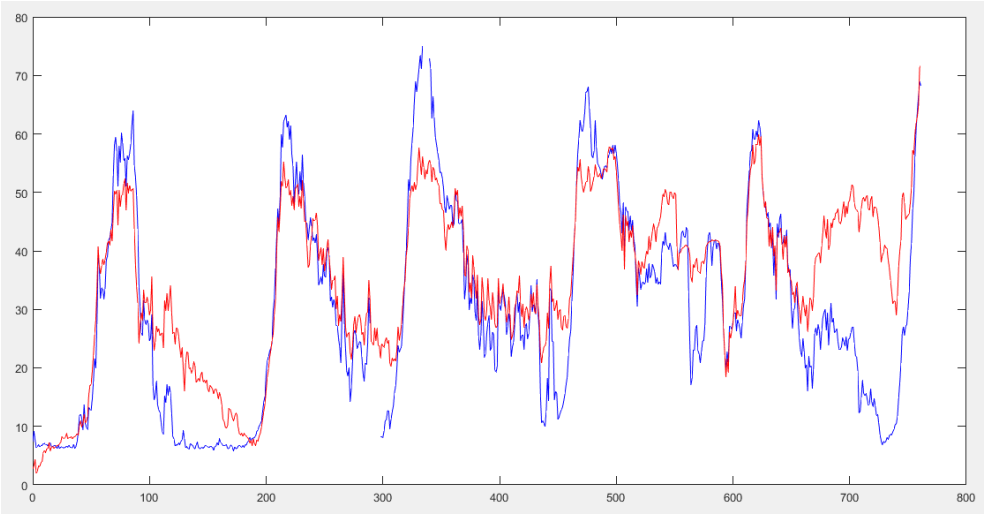
Figura 23
Imputación de Datos Mes de Junio - O3



Fuente: Elaboración Propia.

Uno de los meses con menor número de datos es el mes de noviembre, pero que fue de utilidad para visualizar de una manera más amplia los valores, representado esta menor cantidad de datos en una manera dispersa y poder interpretar el comportamiento cercano entre las dos líneas (Figura 24).

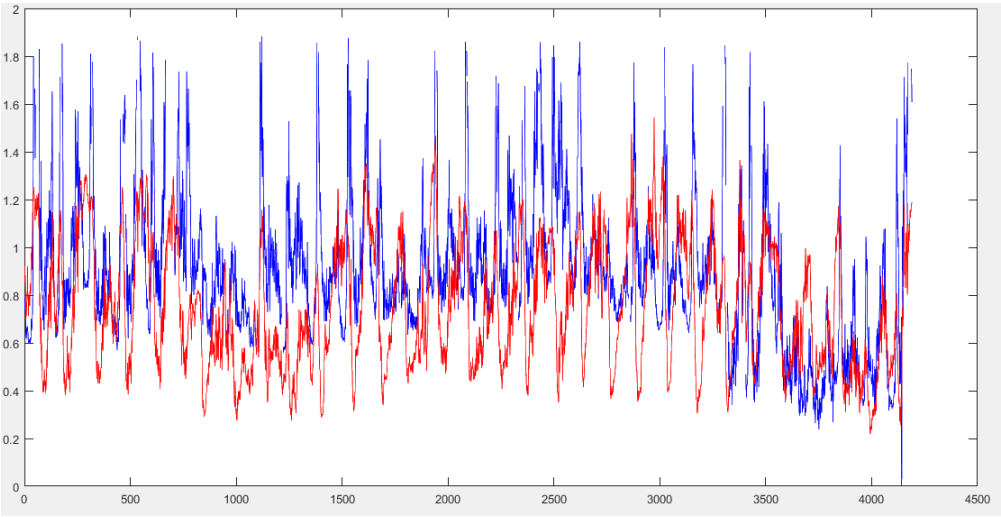
Figura 24
Datos del Mes de Noviembre - O3



Fuente: Elaboración Propia.

Con el contaminante monóxido de carbono (CO) no se tuvo resultados como fueron para los contaminantes del ozono y del dióxido de azufre que posteriormente se explicará a detalle. En la Figura 25 se puede evidenciar que aunque sigue el comportamiento de los valores verdaderos, no se acerca demasiado a estos valores, quedando por debajo de su comportamiento normal.

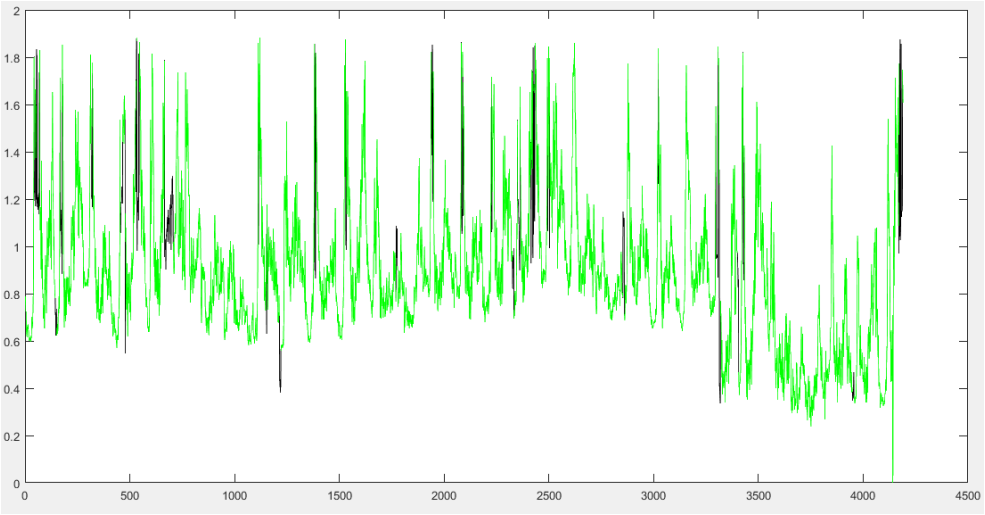
Figura 25
Datos del Mes de Octubre - CO



Fuente: Elaboración Propia.

De igual manera la imputación de datos se realizó en el mes de octubre, el cual fue uno de los meses con más datos perdidos (Figura 26).

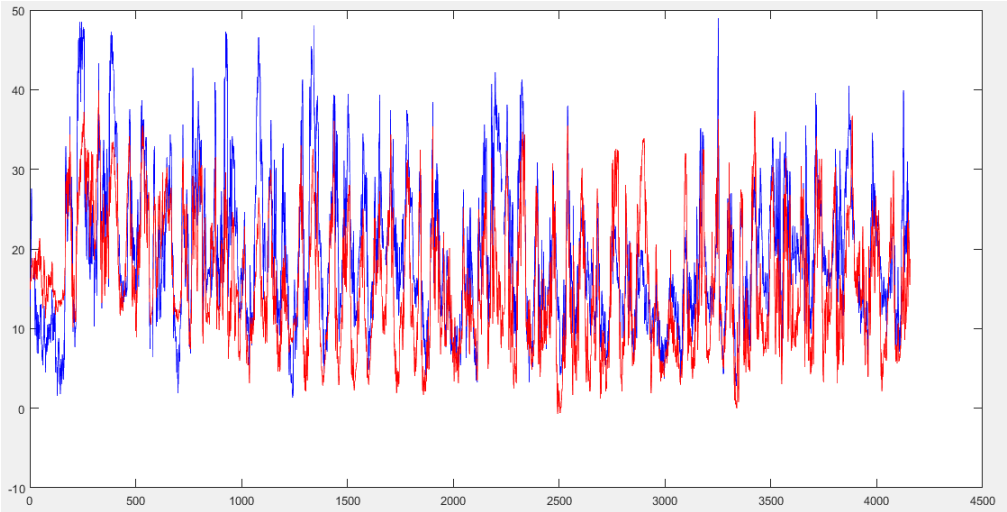
Figura 26
Imputación Mes de Octubre - CO



Fuente: Elaboración Propia.

Otro contaminante con buenos resultados fue el dióxido de azufre, el cual tuvo un acercamiento muy alto en comparación con los valores verdaderos, esto se pudo evidenciar en varios meses, pero para ejemplificar se tomó el mes de abril (Figura 27), en donde se puede evidenciar el comportamiento de las predicciones.

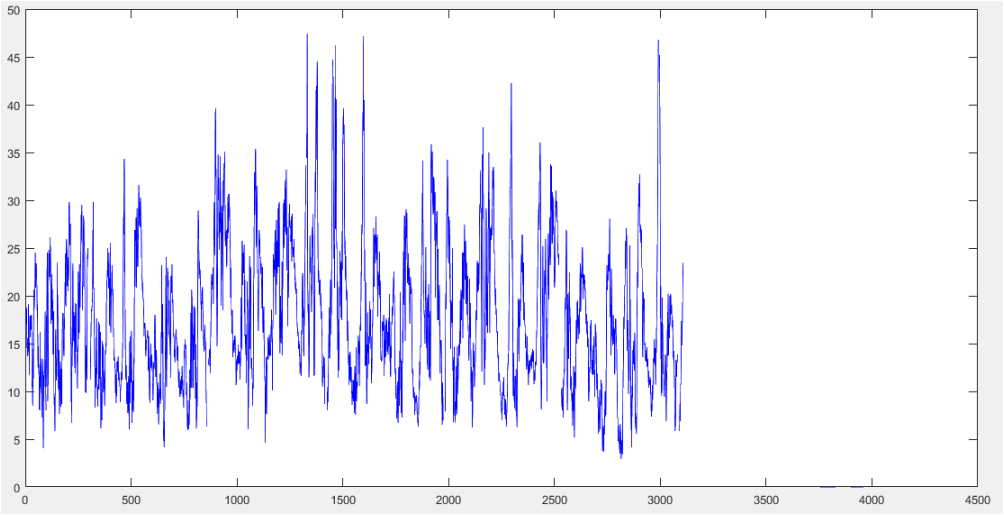
Figura 27
Datos del Mes de Abril - NO2



Fuente: Elaboración Propia.

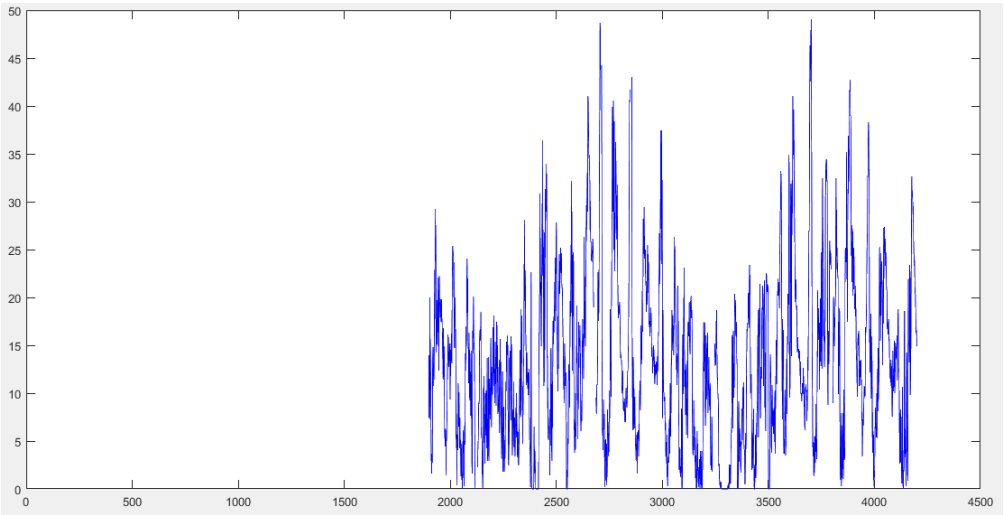
El caso particular del dióxido de nitrógeno con el resto de contaminantes, fue que existió gran pérdida de datos al final del mes de mayo (Figura 28) y al inicio del mes de junio (Figura 29)

Figura 28
Datos Verdaderos Mes de Mayo - NO2



Fuente: Elaboración Propia.

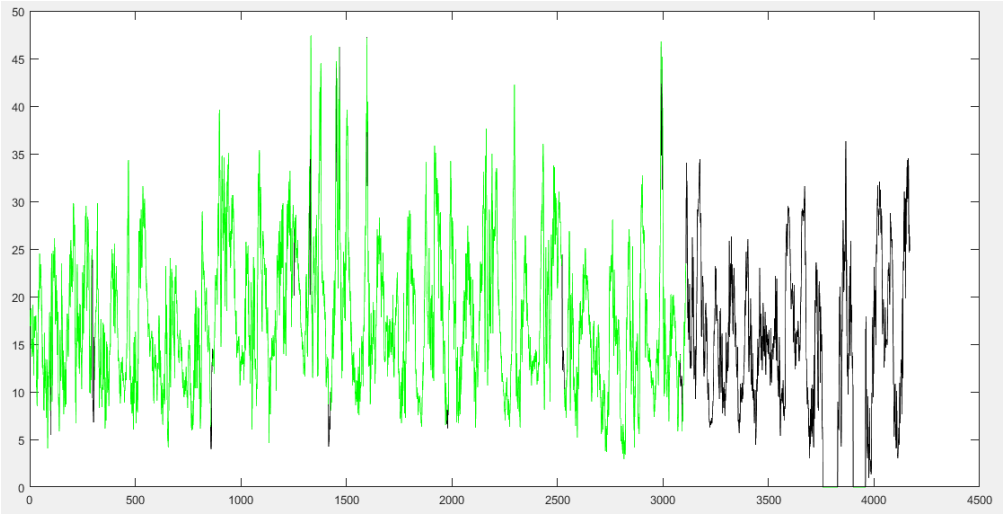
Figura *29*
Datos Verdaderos Mes de Junio - NO2



Fuente: Elaboración Propia.

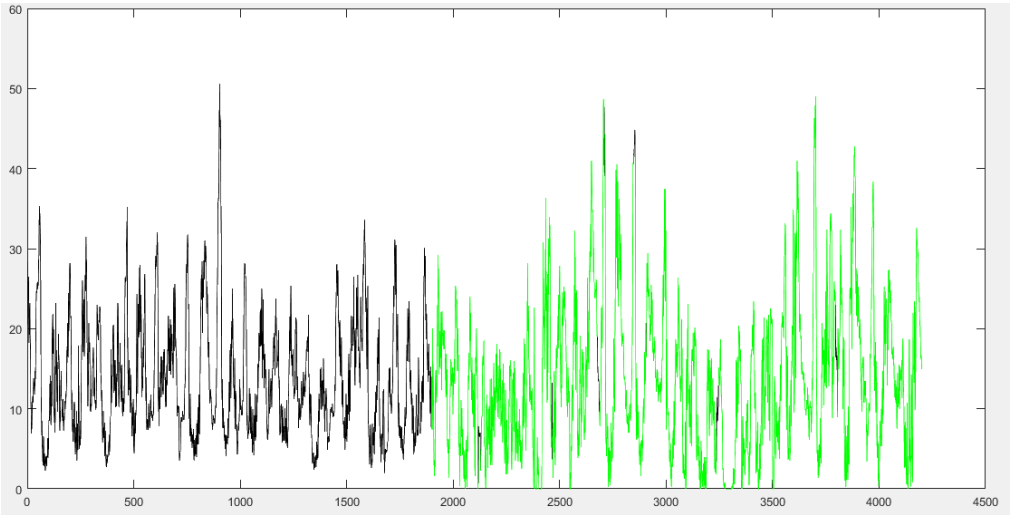
De igual forma se realizó la imputación de datos con las predicciones tanto para el mes de mayo (Figura 30), como para el mes de junio (Figura 31).

Figura 30
Imputación de Datos Mes de Mayo - NO2



Fuente: Elaboración Propia.

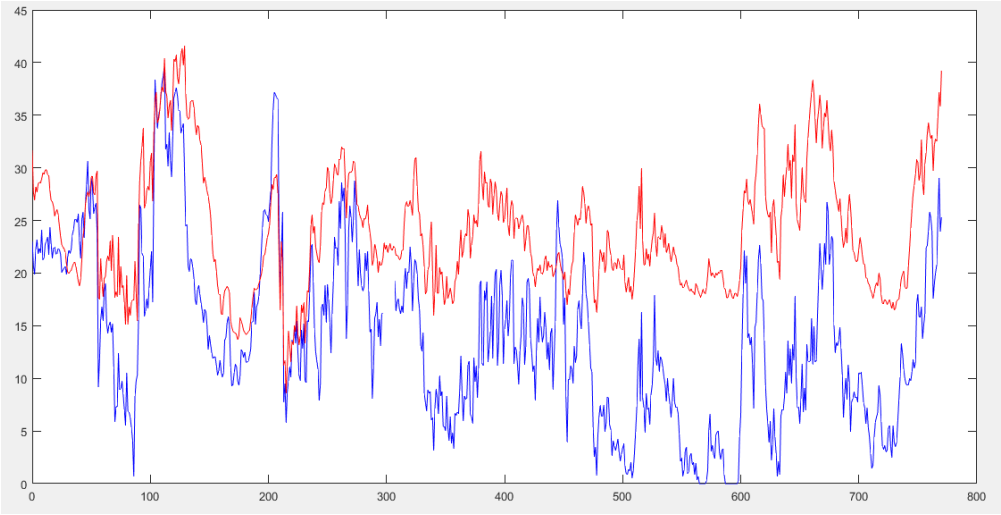
Figura 31
Imputación de Datos Mes de Junio - NO2



Fuente: Elaboración Propia.

De igual manera en el mes de noviembre existe menos cantidad de datos, lo cual ayuda a visualizar el comportamiento de las líneas de valores en la Figura 32, en donde se comprueba que existe una buena relación entre los valores verdaderos y los que se predicen en el contaminante dióxido de nitrógeno.

Figura 32
Datos del Mes de Noviembre - NO2



Fuente: Elaboración Propia.

Para poder visualizar las gráficas de todos los contaminantes, tanto de los valores verdaderos con las predicciones y la imputación en todos los meses, se puede observar en el Anexo E.

4. CONCLUSIONES

El estudio realizado proporcionó resultados muy buenos en la imputación con predicciones en los contaminantes del ozono (O3) y dióxido de nitrógeno (NO2), siendo estos los que mejores resultados obtuvieron en las pruebas correspondientes con la comparación entre los valores verdaderos y los valores que predecían la red neuronal.

Dentro del estudio se contó con un conjunto de datos inicial de 48.095 registros, el cual se procedió a pre-procesar para que de esta manera se adecuen al modelo que se utilizó. Se elaboró 5 conjuntos de datos, uno para cada contaminante solamente con los datos que se requerían para la predicción de la variable.

El valor angular se obtuvo conjuntamente con los niveles de las variables de entrada en un intervalo de tiempo de 10 minutos, siendo este valor calculado una entrada más en la red neuronal. Mediante la utilización de la red neuronal NARX fue posible realizar la imputación de datos usando como variables de entrada el tiempo en el cual se toman los valores, los valores angulares, los niveles de los contaminantes que interactúan entre sí y la variable meteorológica de la temperatura. Esto dio como resultado valores más cercanos a los valores verdaderos en ciertos contaminantes que en otros.

Se analizó los valores de las predicciones comparando precisamente el conjunto de datos de prueba que tomaba la red neuronal. Los dos casos con mejores resultados fueron el ozono (O3) con un R de 0.85 y RMSE de 6.29, el dióxido de nitrógeno (NO2) con un R de 0.73 y RMSE de 5.32, seguido del material particulado PM2_5 con un R de 0.55 y RMSE de 4.32, y los dos casos con menor resultado fueron: el monóxido de carbono (CO) con un R de 0.53 y RMSE de 0.22 y por último el dióxido de azufre (SO2) con un R de 0.49 y RMSE de 4.76.

5. TRABAJOS FUTUROS

Para el proyecto se utilizó un conjunto de datos que comprendía información desde enero de 2018 hasta noviembre de 2018 en donde se tuvo que realizar la eliminación de varios registros para procesar los datos de mejor manera, permitiendo de esta forma cumplir con el objetivo del proyecto, sin embargo para posteriores trabajos se recomienda la utilización de mayor cantidad de datos, debido a que esto permitirá entrenar de mejor manera la red y así obtener mejores resultados.

Este estudio se centró en la imputación de datos perdidos en contaminantes atmosféricos, sin embargo se pueden adecuar el enfoque para otras áreas. Por último, para un futuro proyecto se puede comparar el modelo del estudio con otros modelos con distintas entradas, debido a que en este caso en particular se experimentó con el valor angular de los contaminantes como una entrada más. De esta manera se puede analizar qué modelo entrega los mejores resultados.

6. BIBLIOGRAFÍA

Alvarez, I., Méndez Martínez, J., Bello Rodríguez, B. M., Benítez Fuentes, B., Escobar Blanco, L. M., & Zamora Monzón, R. (2017). Influencia de los contaminantes atmosféricos sobre la salud. *Revista Médica Electrónica*, *39*(5), 1160–1170. Retrieved from http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1684-18242017000500017

Andrade, P. S. (2018). *Aplicación de minería de datos en el análisis de contaminantes atmosféricos y variables meteorológicas*.

Avila, R. M. ´, De Hernández, G., Rodríguez Pérez, V., & Caraballo, E. A. H. (2012). PREDICCIÓN DEL RENDIMIENTO DE UN CULTIVO DE PLÁTANO MEDIANTE REDES NEURONALES ARTIFICIALES DE REGRESIÓN GENERALIZADA. *Publicaciones En Ciencias y Tecnología*, *6*(1), 200702–202730.

Azid, A., Juahir, H., Toriman, M. E., Kamarudin, M. K. A., Saudi, A. S. M., Hasnam, C. N. C., … Yamin, M. (2014). Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: A case study in Malaysia. *Water, Air, and Soil Pollution*, *225*(8). https://doi.org/10.1007/s11270-014-2063-1

Bellinger, C., Mohomed Jabbar, M. S., Zaïane, O., & Osornio-Vargas, A. (2017). A systematic review of data mining and machine learning for air pollution epidemiology. *BMC Public Health*, *17*(1), 1–19. https://doi.org/10.1186/s12889-017-4914-3

Boldo, E. (2016). *La Contaminación del Aire*. Retrieved from https://repisalud.isciii.es/bitstream/20.500.12105/7274/1/LaContaminaciónDelAire_2016.pdf

Brunekreef, B., & Holgate, S. T. (2002). Air pollution and health. *Lancet*, *360*(9341), 1233–1242. https://doi.org/10.1016/S0140-6736(02)11274-8

Cevallos, V. M., Díaz, V., & Sirois, C. M. (2017). Particulate matter air pollution from the city of Quito, Ecuador, activates inflammatory signaling pathways in vitro.

Innate Immunity, *23*(4), 392–400. https://doi.org/10.1177/1753425917699864

EMOV. (2017). *Informe de calidad del aire Cuenca*. 1–123.

EPA. (2009). *Air Quality A Guide to Air Quality and Your Health*. (August), 1–7.

Espinoza Molina, A. (2011). *Diseño de un sistema de información geográfica para la Red de Monitoreo Ambiental de la ciudad de Cuenca*.

Estrella, B., Sempértegui, F., Franco, O. H., Cepeda, M., & Naumova, E. N. (2019). Air pollution control and the occurrence of acute respiratory illness in school children of Quito, Ecuador. *Journal of Public Health Policy*, *40*(1), 17–34. https://doi.org/10.1057/s41271-018-0148-6

Fernanda, P., & Logroño, B. (2018). *ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO METEOROLÓGICA CHIMBORAZO*.

Fuentes, M., Campos, C., & García-Loyola, S. (2018). Application of artificial neural networks to frost detection in central Chile using the next day minimum air temperature forecast. *Chilean Journal of Agricultural Research*, *78*(3), 327–338. https://doi.org/10.4067/s0718-58392018000300327

Garcia, F. (2012). Tests to identify outliers in data series. *Pontifical Catholic University of Rio de Janeiro, …*, 1–16. Retrieved from http://habcam.whoi.edu/HabCamData/HAB/processed/Outlier Methods_external.pdf

Geetha, A., & Nasira, G. M. (2015). Data mining for meteorological applications: Decision trees for modeling rainfall prediction. *2014 IEEE International Conference on Computational Intelligence and Computing Research, IEEE ICCIC 2014*, 0–3. https://doi.org/10.1109/ICCIC.2014.7238481

Gore, R. W., & Deshpande, D. S. (2017). An approach for classification of health risks based on air quality levels. *Proceedings - 1st International Conference on Intelligent Systems and Information Management, ICISIM 2017*, *2017-Janua*, 58–61. https://doi.org/10.1109/ICISIM.2017.8122148

IBM, I. B. M. (2012). Manual CRISP-DM de IBM SPSS Modeler. *IBM Corporation*, 56. Retrieved from

ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/es/CRISP-DM.pdf

Kingsy, G. R., Manimegalai, R., Geetha, D. M. S., Rajathi, S., Usha, K., & Raabiathul, B. N. (2017). Air pollution analysis using enhanced K-Means clustering algorithm for real time sensor data. *IEEE Region 10 Annual International Conference, Proceedings/TENCON*, (August 2006), 1945–1949. https://doi.org/10.1109/TENCON.2016.7848362

Kukkonen, J., Partanen, L., Karppinen, A., Ruuskanen, J., Junninen, H., Kolehmainen, M., … Cawley, G. (2003). Extensive evaluation of neural network models for the prediction of NO 2 and PM10 concentrations, compared with a deterministic modelling system and measurements in central Helsinki. *Atmospheric Environment*, *37*(32), 4539–4550. https://doi.org/10.1016/S1352-2310(03)00583-1

Kurt, A., Gulbagci, B., Karaca, F., & Alagha, O. (2008). An online air pollution forecasting system using neural networks. *Environment International*, *34*(5), 592–598. https://doi.org/10.1016/j.envint.2007.12.020

Li, S. T., & Shue, L. Y. (2004). Data mining to aid policy making in air pollution management. *Expert Systems with Applications*, *27*(3), 331–340. https://doi.org/10.1016/j.eswa.2004.05.015

Lo, S. M., Lu, W. Z., & Fan, H. Y. (2003). Application of evolutionary neural network method in predicting pollutant levels in downtown area of Hong Kong. *Neurocomputing*, *51*, 387–400.

MATLAB. (2018). Shallow Neural Network Time-Series Prediction and Modeling. Retrieved from https://www.mathworks.com/help/deeplearning/gs/neural-network-time-series-prediction-and-modeling.html

MATLAB & Simulink. (2019). Design Time Series NARX Feedback Neural Networks - MATLAB & Simulink - MathWorks Deutschland. Retrieved from MATLAB Documentation website: https://de.mathworks.com/help/deeplearning/ug/design-time-series-narx-feedback-neural-networks.html;jsessionid=44f1e9acc4fc3d65278ab59a923f

Matute Rivera, M. A. (2018). *Evaluación de las herramientas de minería de datos en*

variables de contaminación atmosférica. 100. Retrieved from http://dspace.uazuay.edu.ec/handle/datos/8203

Ortega, J. J. (2018). Impacto de la aplicación de algoritmos de minería de datos en variables de contaminación del aire. *Director*, *15*(2), 2017–2019. https://doi.org/10.22201/fq.18708404e.2004.3.66178

Pete, C., Julian, C., Randy, K., Thomas, K., Thomas, R., Colin, S., & Wirth, R. (2000). Crisp-Dm 1.0. *CRISP-DM Consortium*, 76.

Rojas, F. J. (2017). *MODELACIÓN NUMÉRICA DEL TRANSPORTE DE CONTAMINANTES ATMOSFÉRICOS Y SU RELACIÓN CON LAS CONDICIONES METEOROLÓGICAS EN LIMA METROPOLITANA*.

Sierra, M. M. (2006). Asociación existente entre las variables meteorológicas Temperatura, Velocidad del viento y Precipitación y las concentraciones de PM10 registradas en la Red de Calidad del Aire de Bogota. *Mycological Research*, *106*(11), 171. Retrieved from http://repository.lasalle.edu.co/bitstream/handle/10185/14805/00798291.pdf?sequence=1&isAllowed=y

Viotti, P., Liuti, G., & Di Genova, P. (2002). Atmospheric urban pollution: Applications of an artificial neural network (ANN) to the city of Perugia. *Ecological Modelling*, *148*(1), 27–46. https://doi.org/10.1016/S0304-3800(01)00434-3

ANEXOS

Anexo A. Código Red Neuronal NARX

```
>> % Carga los datos
filenamel = 'Path\1 Ozono completo.xlsx'; %Ejemplo de carga, datos externos en excel
entradas = xlsread(filenamel,'A:O'); %Columnas de excel, desde A hasta O son entradas
ozono = xlsread(filenamel,'P:P'); %Targets
%   entradas - input time series.
%   ozono - feedback time series.

inputSeries = tonndata(entradas,false,false);
targetSeries = tonndata(ozono,false,false);

% Función de entrenamiento
% 'trainlm' suele ser el más rápido.
% 'trainbr' lleva más tiempo, pero puede ser mejor para problemas difíciles. regulacion bayesiana
% 'traincg' usa menos memoria. Adecuado en situaciones de poca memoria. gradiente conjugado
trainFcn = 'trainlm';  % Levenberg-Marquardt backpropagation seleccionada
%trainFcn = 'trainbr';
%trainFcn = 'trainscg';

% Crear una red autorregresiva no lineal con entrada externa
% Se puede modificar los parametros segun se requiera
inputDelays = 1:3;
feedbackDelays = 1:3;
hiddenLayerSize = 12;
net = narxnet(inputDelays,feedbackDelays,hiddenLayerSize,'open',trainFcn);

% Preparar los datos para entrenamiento y simulación
[inputs,inputStates,layerStates,targets] = preparets(net,inputSeries,{},targetSeries);

% Configuración de la división de datos para capacitación, validación, prueba
% Para obtener una lista de todas las funciones de división de datos, escriba: help nndivide
net.divideFcn = 'dividerand';  % Divide datos al azar
net.divideMode = 'time';  % Divide cada muestra
net.divideParam.trainRatio = 70/100;
net.divideParam.valRatio = 15/100;
net.divideParam.testRatio = 15/100;

% Elija una función de rendimiento
% Para obtener una lista de todas las funciones de rendimiento, escriba: help nnperformance
net.performFcn = 'mse';  % Mean Squared Error

% Elegir funciones de trazado
% Para obtener una lista de todas las funciones de trazado, escriba: help nnplot
net.plotFcns = {'plotperform','plottrainstate', 'ploterrhist', ...
    'plotregression', 'plotresponse', 'ploterrcorr', 'plotinerrcorr'};

% Entrenar a la red
[net,tr] = train(net,inputs,targets,inputStates,layerStates);

% Prueba la red
outputs = net(inputs,inputStates,layerStates);
errors = gsubtract(targets,outputs);
performance = perform(net,targets,outputs)

% Recalcular entrenamiento, validación y rendimiento de prueba
trainTargets = gmultiply(targets,tr.trainMask);
valTargets = gmultiply(targets,tr.valMask); ...
```

```
% Parcelas que permiten visualizar el entrenamiento que se realizó
figure, plotperform(tr)
%figure, plottrainstate(tr)
%figure, ploterrhist(errors)
%figure, plotregression(targets,outputs)
%figure, plotresponse(targets,outputs)
%figure, ploterrcorr(errors)
%figure, plotinerrcorr(inputs,errors)

% Red de bucle cerrado
% Use esta red para hacer predicciones de varios pasos.
% La función CLOSELOOP reemplaza la entrada de retroalimentación con un
% de conexión desde la capa externa.
netc = closeloop(net);
netc.name = [net.name ' - Closed Loop'];
%view(netc)
[xc,xic,aic,tc] = preparets(netc,inputSeries,{},targetSeries);
yc = netc(xc,xic,aic);
closedLoopPerformance = perform(net,tc,yc);

% Red de predicción progresiva
nets = removedelay(net);
nets.name = [net.name ' - Predict One Step Ahead'];
%view(nets)
[xs,xis,ais,ts] = preparets(nets,inputSeries,{},targetSeries);
ys = nets(xs,xis,ais);
stepAheadPerformance = perform(nets,ts,ys);
```

Código para realizar las predicciones.

```
>> filename3 = 'Path ejemplo\1 predecir Ozono.xlsx'; %Carga del archivo con entradas a la red

predict_input = xlsread(filename3,'A:O');%Se define las columnas que tiene las entradas
% Formato de datos
toPredict = tonndata(predict_input,false,false);

[y1,xf,af] = net(inputs,inputStates,layerStates);

% Cerrar el ciclo para realizar multiples predicciones
[netc,xi,ai] = closeloop(net,xf,af);%Predicciones en bucle cerrado
y2= netc(toPredict,xi,ai);%Valor que se predice

B = transpose(y2);
xlswrite('Prueba.xlsx',B,'Hoja1','A1:A35235');%Se puede exportar los resultados a excel para comparar
```

Una vez realizada el entrenamiento se puede predecir los valores de otro conjunto de datos que se desee.

Anexo B. Evaluación de los Contaminantes

Ilustración 1.
Monóxido de Carbono (CO)

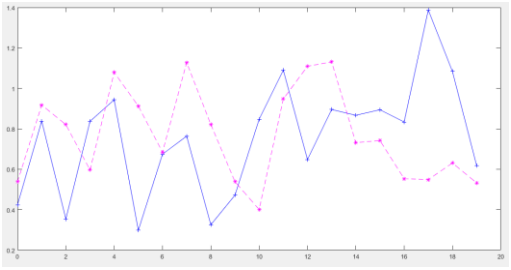


Ilustración 2
Dióxido de Nitrógeno (NO2)

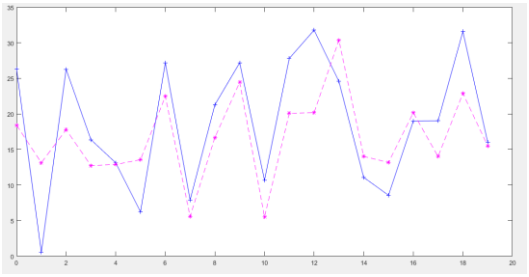


Ilustración 3
Dióxido de Azufre (SO2)

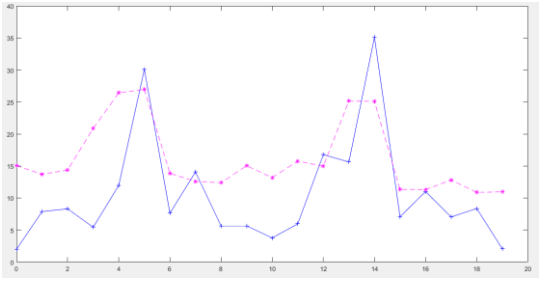
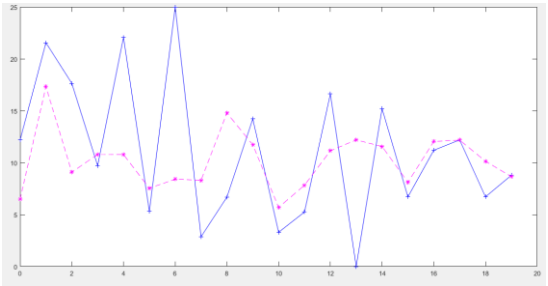


Ilustración 4
Material Particulado PM2_5



Anexo C. Correlación en el Conjunto de Pruebas

Ilustración 5
Ozono – R Ajustado 0.85

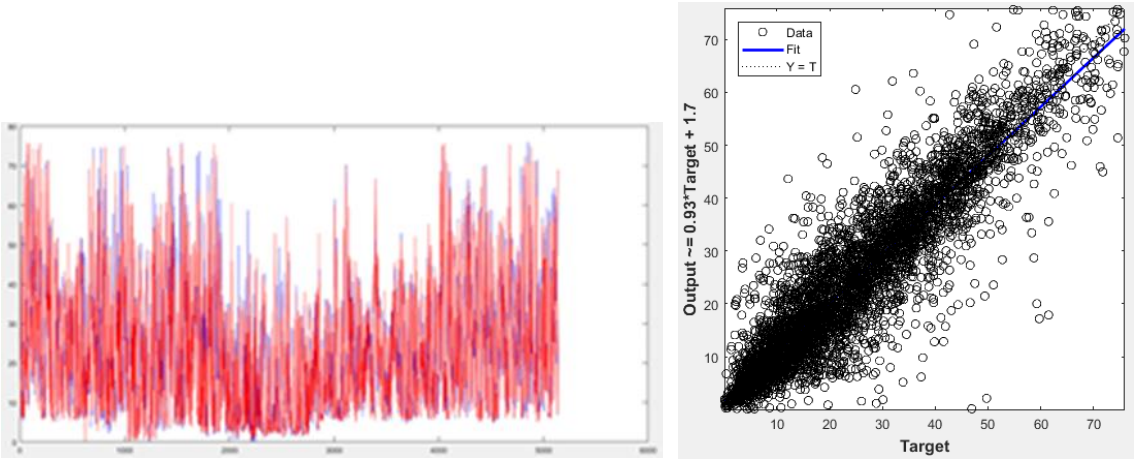


Ilustración 6
Monóxido de Carbono - R Ajustado 0.53

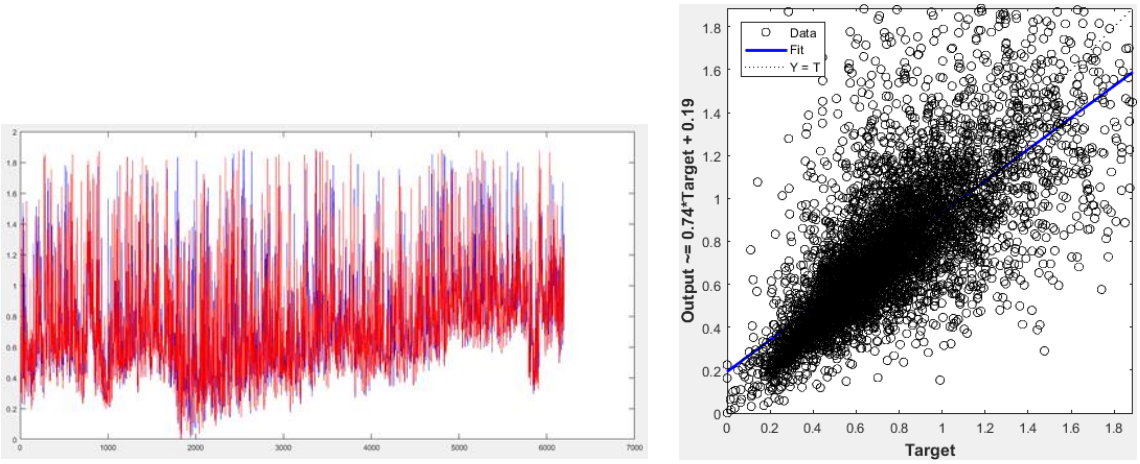
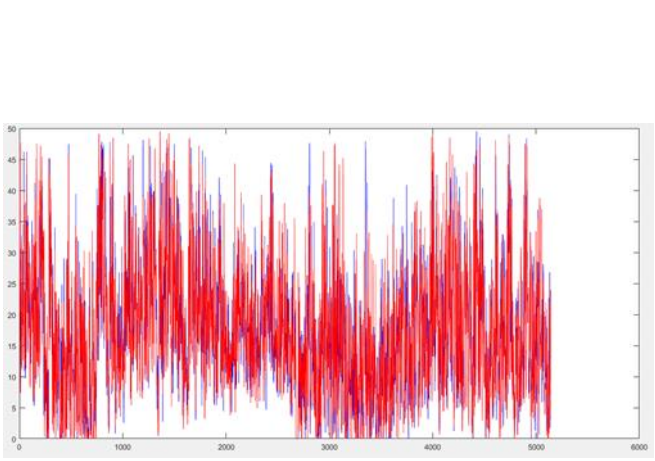


Ilustración 7
Dióxido de Nitrógeno - R Ajustado 0.73



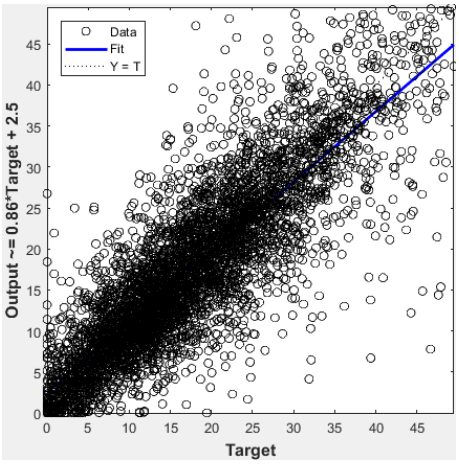
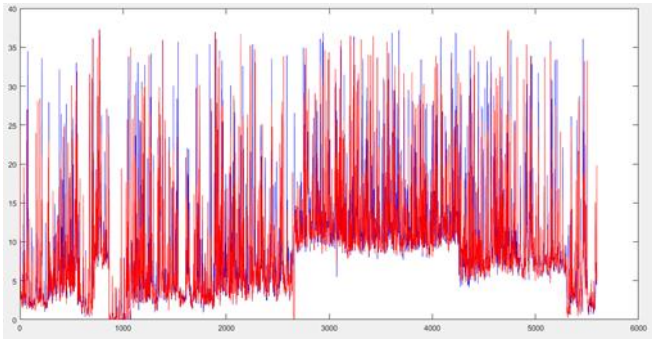


Ilustración 8
Dióxido de Azufre - R Ajustado 0.49



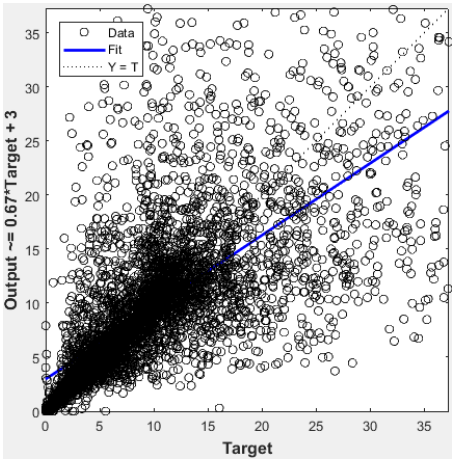
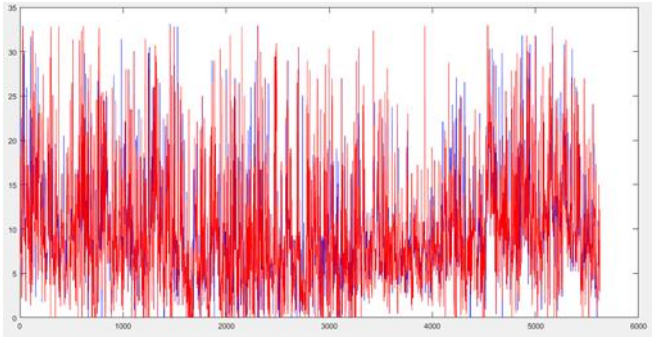
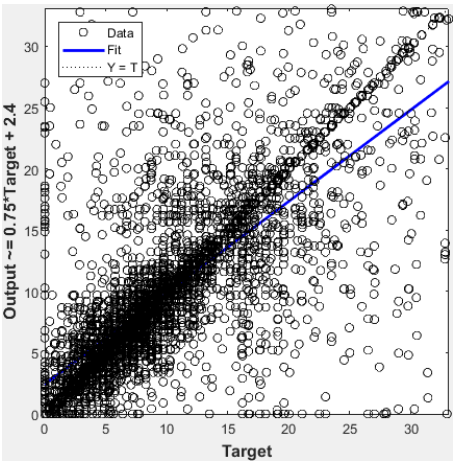


Ilustración 9
PM2_5 - R Ajustado 0.55





Anexo D. Resultados del Conjunto Completo de Datos

Ilustración 10
Verdaderos y Predicciones - O3

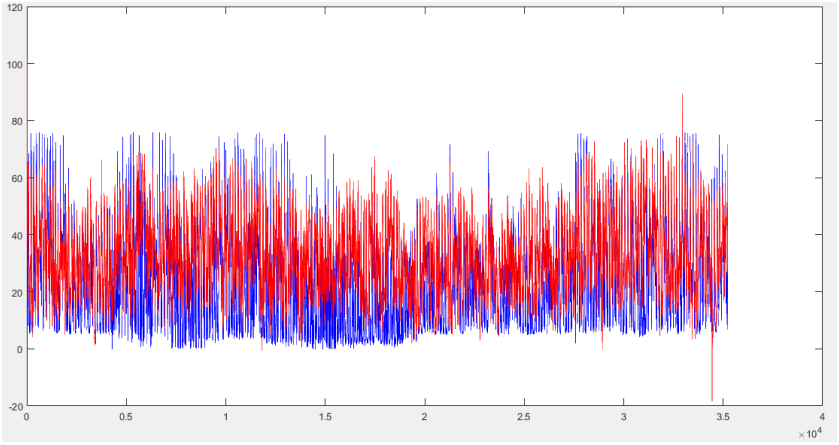


Ilustración 11
Imputación Completa - O3

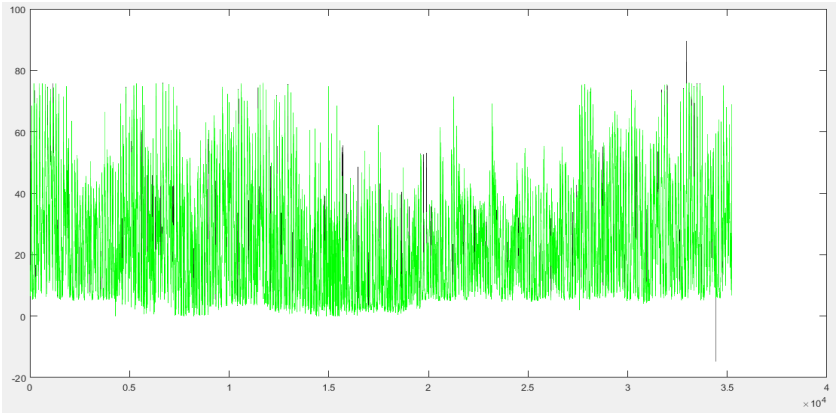


Ilustración 12
Verdaderos y Predicciones - CO

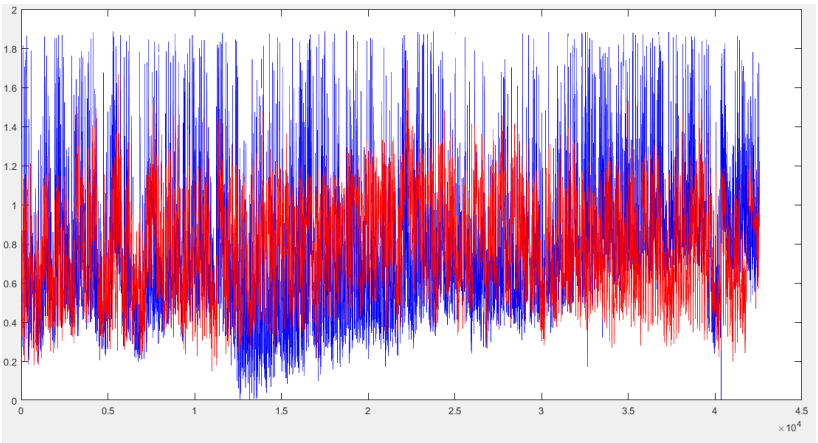


Ilustración 13
Imputación Completa - CO

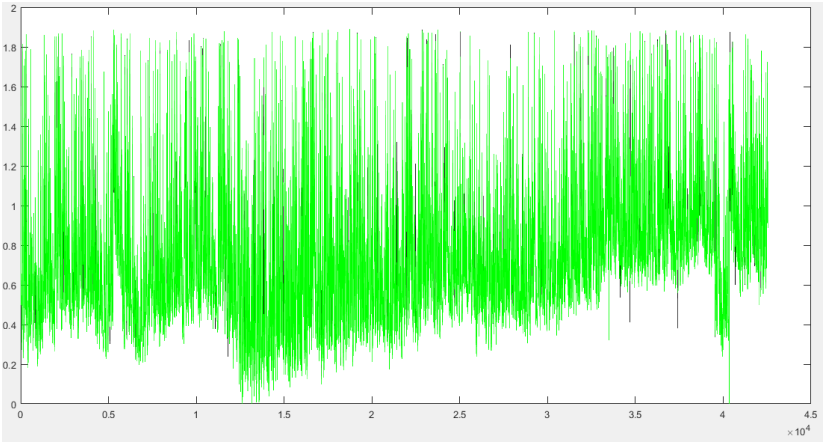


Ilustración 14
Verdaderos y Predicciones - NO2

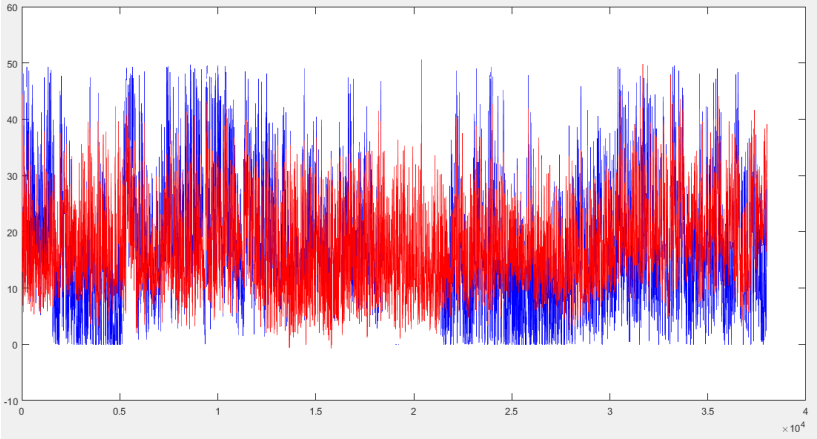


Ilustración 15
Imputación Completa - NO2

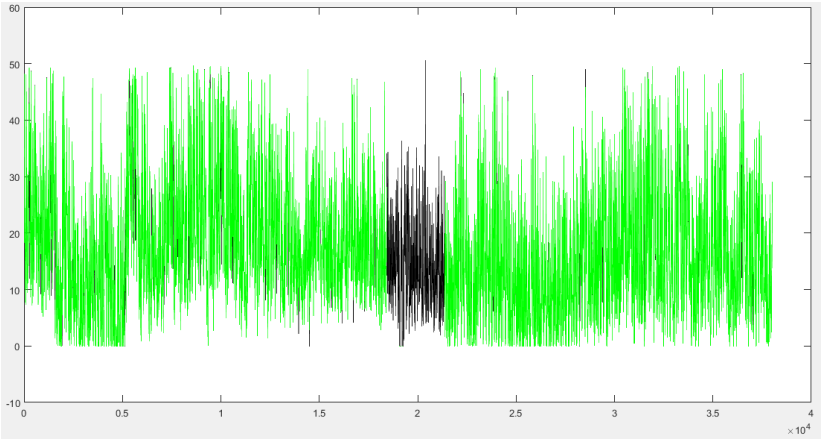


Ilustración 16
Verdaderos y Predicciones - SO2

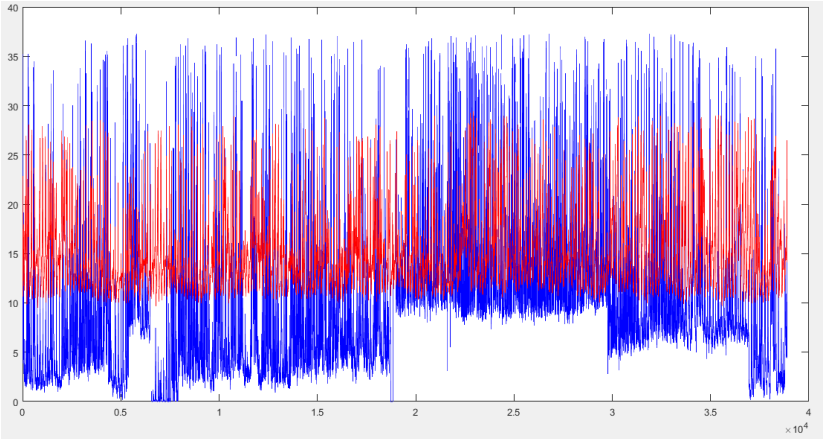


Ilustración 17
Imputación Completa - SO2

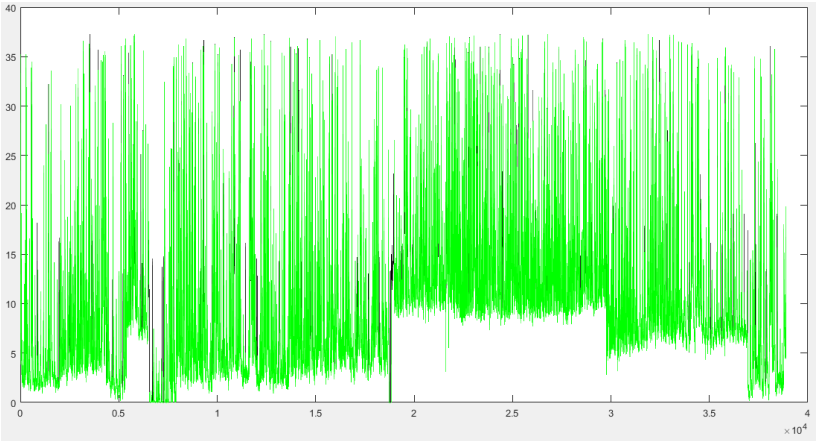


Ilustración 18
Verdaderos y Predicciones - PM2_5

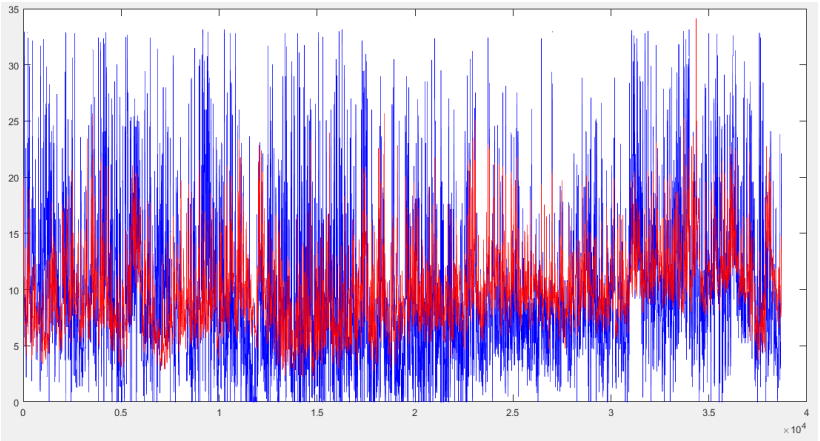
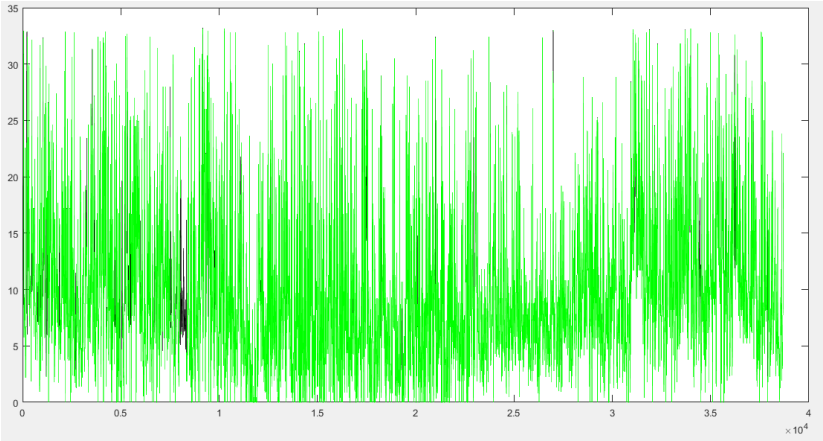


Ilustración 19
Imputación Completa - PM2_5



Anexo E. Resultados Por Meses de Todos los Contaminantes

Ilustración 20. *O3 Mes de Enero*

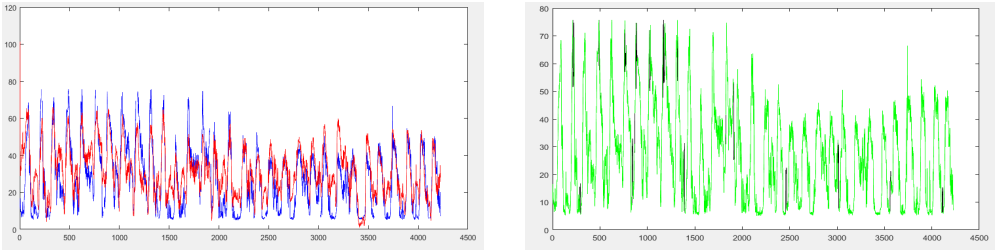


Ilustración 21. *CO Mes de Enero*

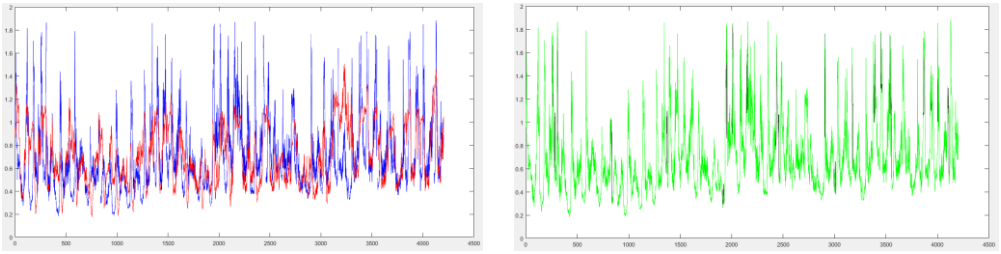


Ilustración 22. *NO2 Mes de Enero*

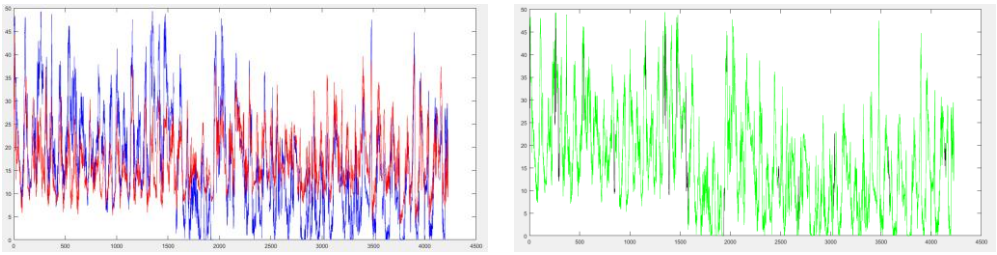


Ilustración *23. SO2 Mes de Enero*

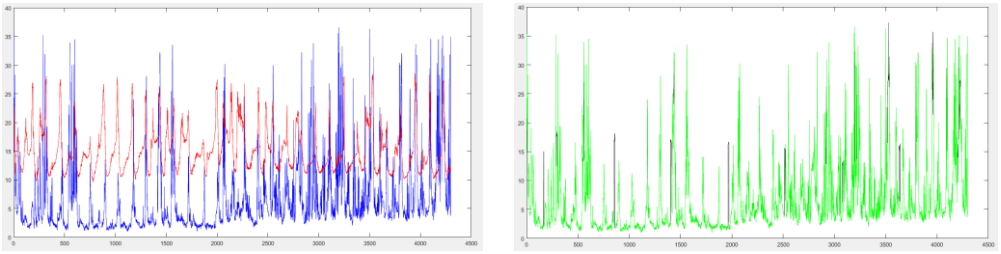


Ilustración 24. *PM2_5 Mes de Enero*

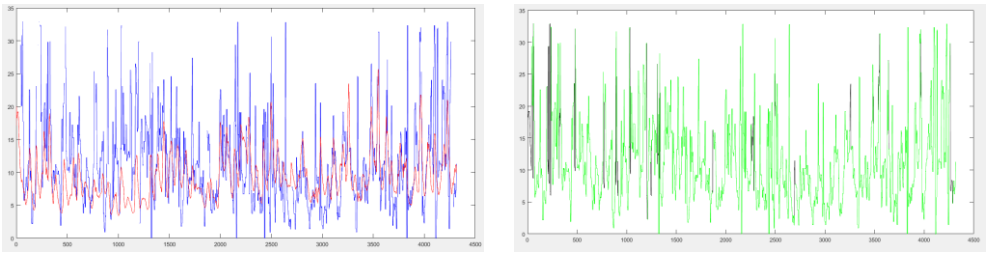


Ilustración 25. *O3 Mes de Febrero*

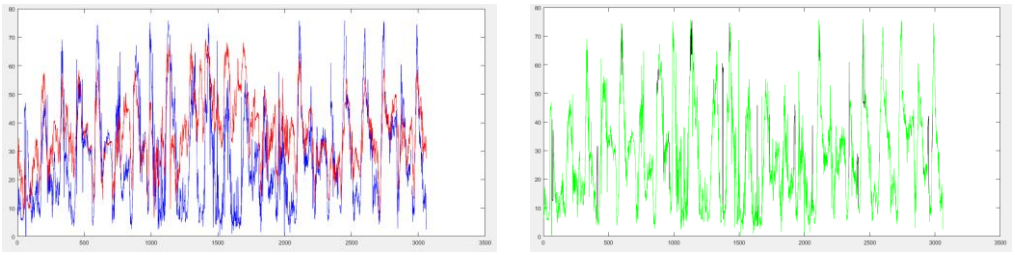


Ilustración 26. *CO Mes de Febrero*

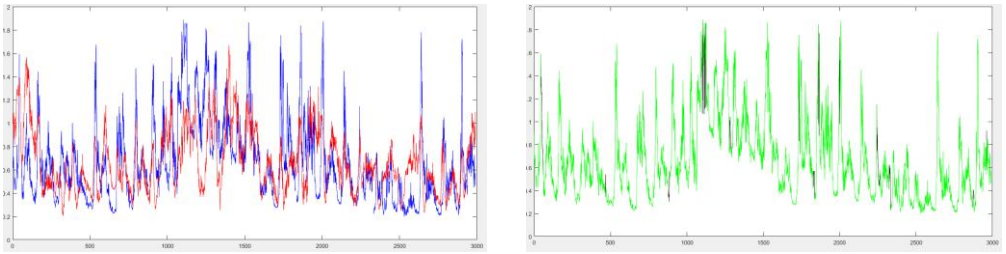


Ilustración 27. *NO2 Mes de Febrero*

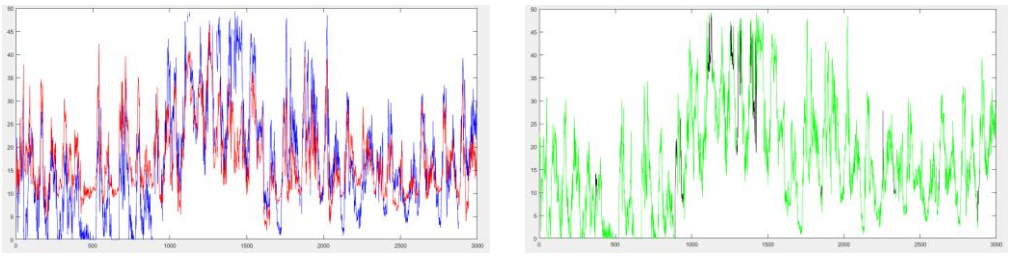


Ilustración 28. *SO2 Mes de Febrero*

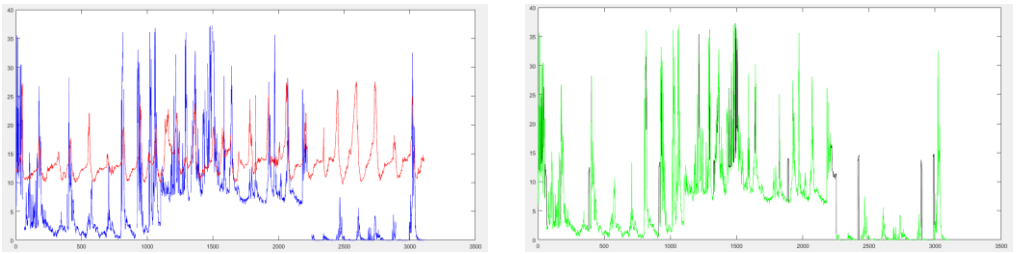


Ilustración 29. *PM2_5 Mes de Febrero*

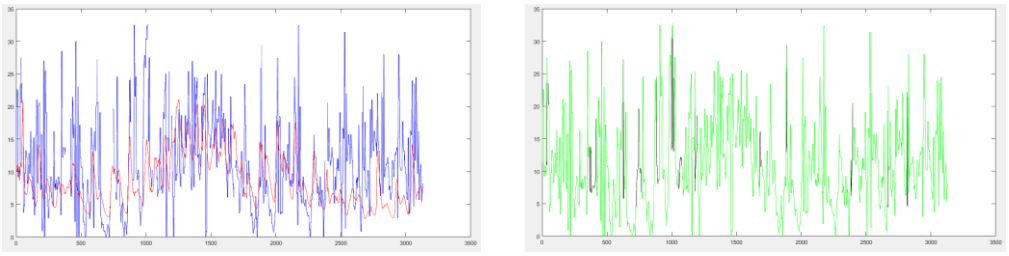


Ilustración 30. *O3 Mes de Marzo*

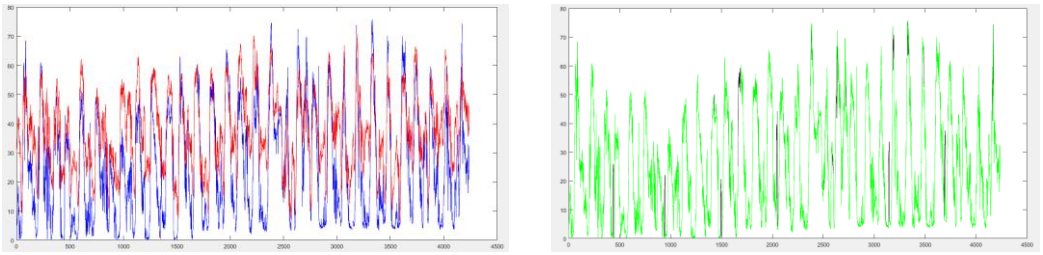


Ilustración 31. *CO Mes de Marzo*

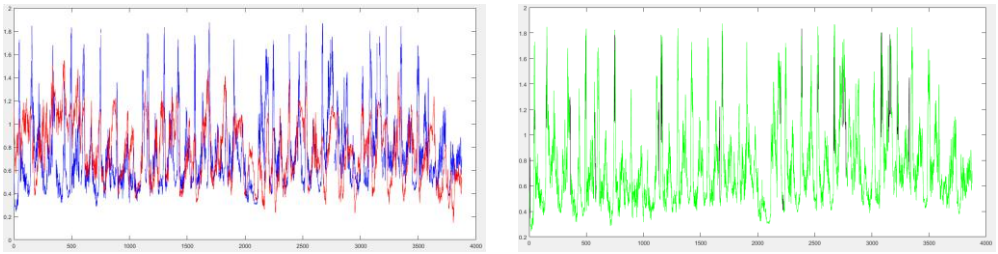


Ilustración 32. *NO2 Mes de Marzo*

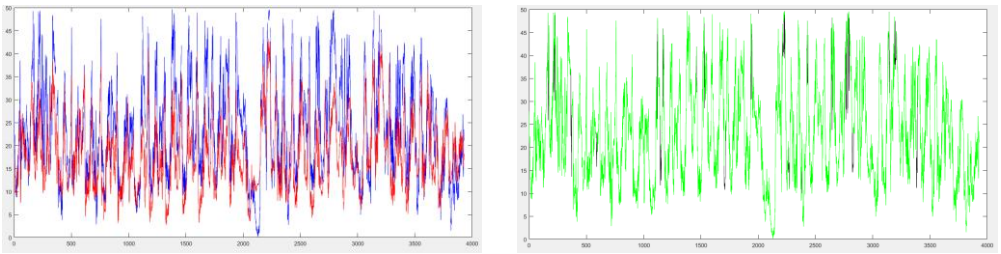


Ilustración 33. *SO2 Mes de Marzo*

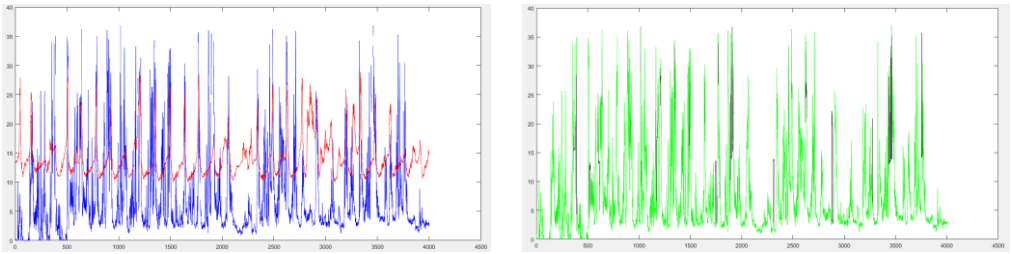


Ilustración 34. *PM2_5 Mes de Marzo*

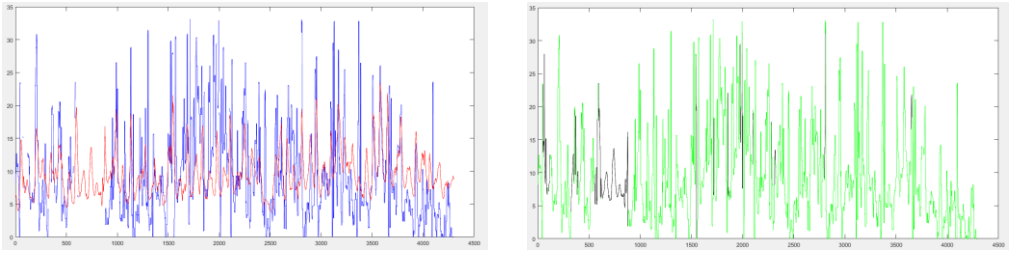


Ilustración 35. *O3 Mes de Abril*

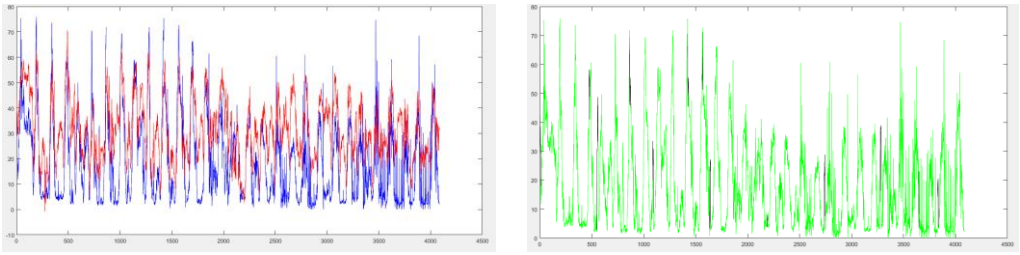


Ilustración 36. *CO Mes de Abril*

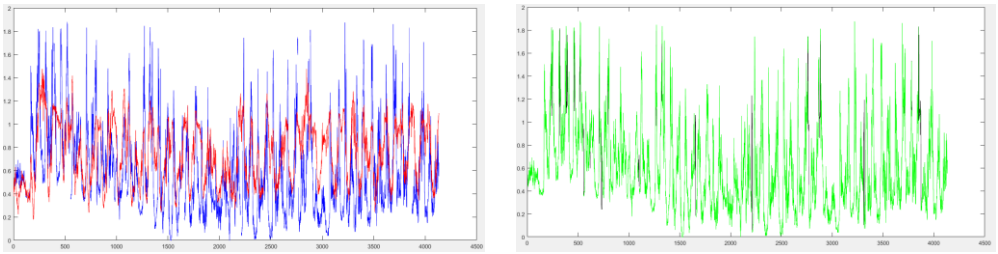


Ilustración 37. *NO2 Mes de Abril*

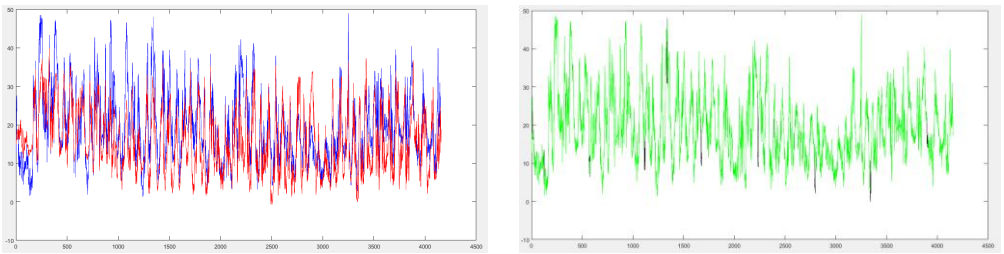


Ilustración 38. *SO2 Mes de Abril*

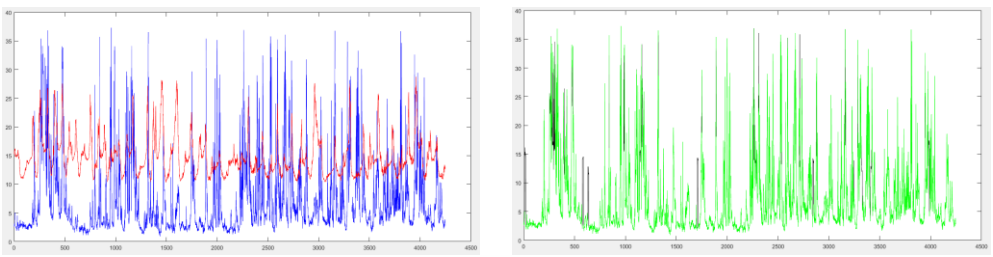


Ilustración 39. *PM2_5 Mes de Abril*

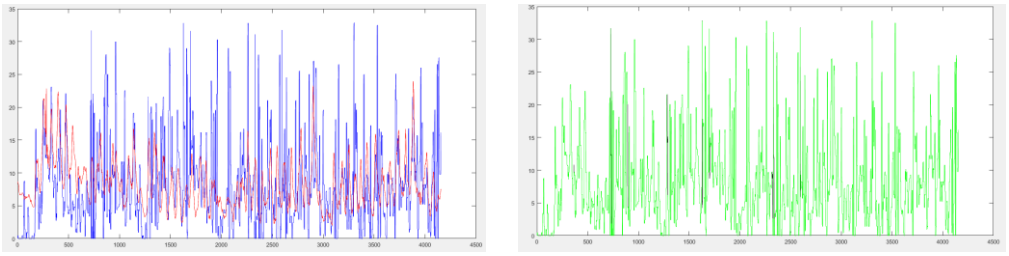


Ilustración 40. *O3 Mes de Mayo*

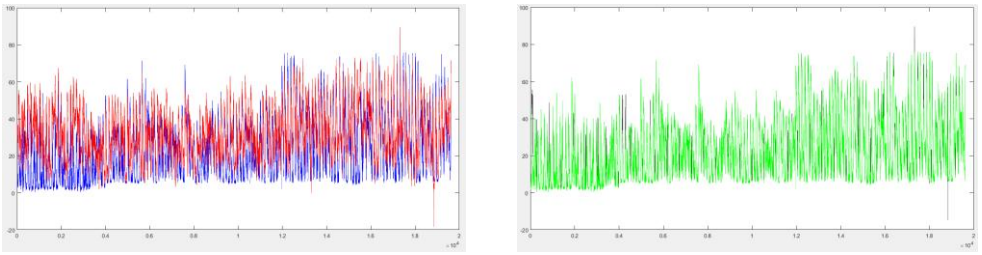


Ilustración 41. *CO Mes de Mayo*

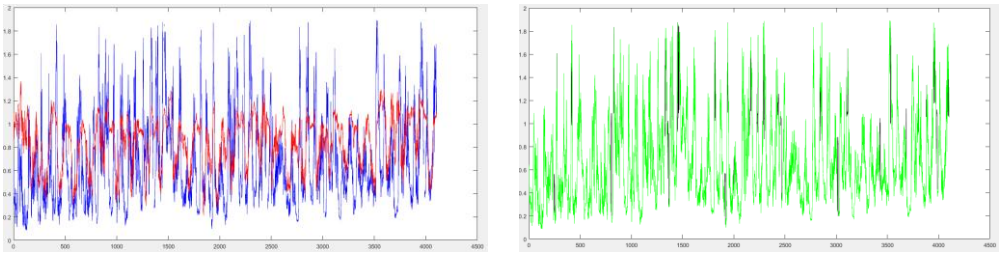


Ilustración 42. *NO2 Mes de Mayo*

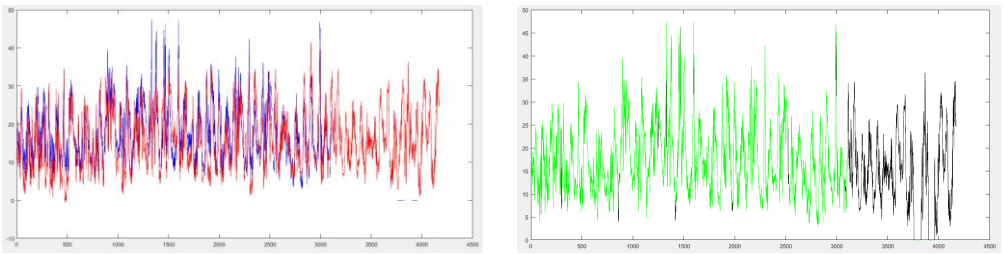


Ilustración 43. *SO2 Mes de Mayo*

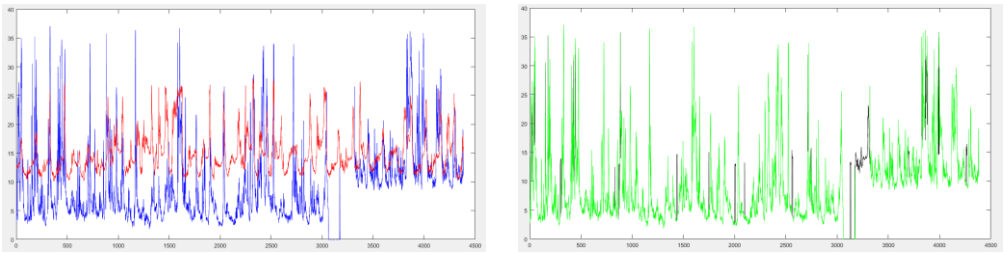


Ilustración 44. *PM2_5 Mes de Mayo*

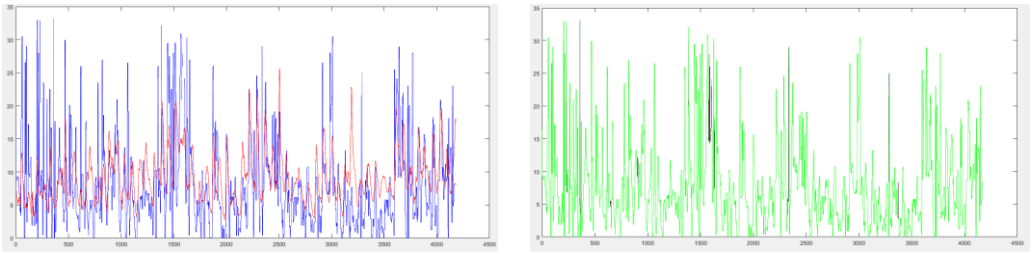


Ilustración 45. *O3 Mes de Junio*

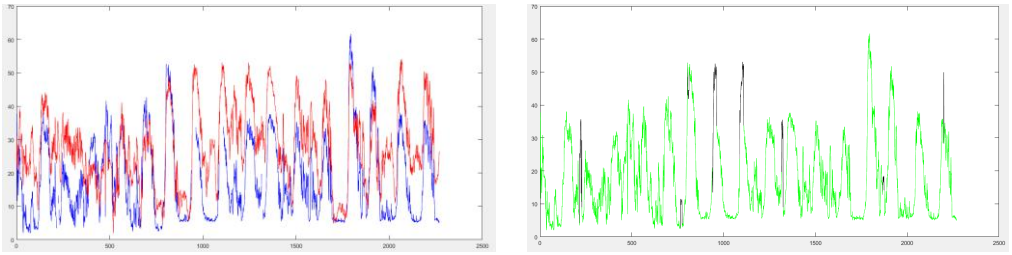


Ilustración 46. *CO Mes de Junio*

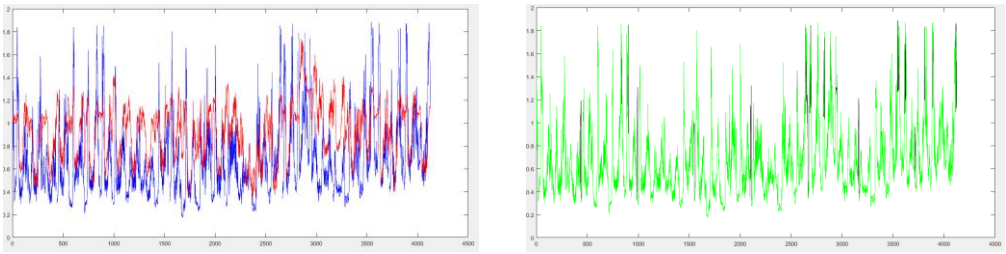


Ilustración 47. *NO2 Mes de Junio*

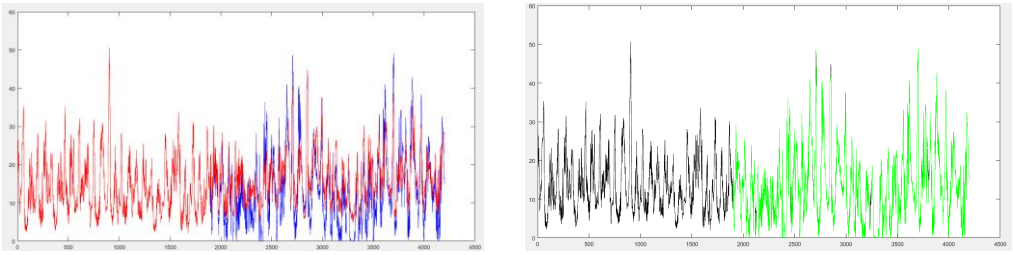


Ilustración 48. *SO2 Mes de Junio*

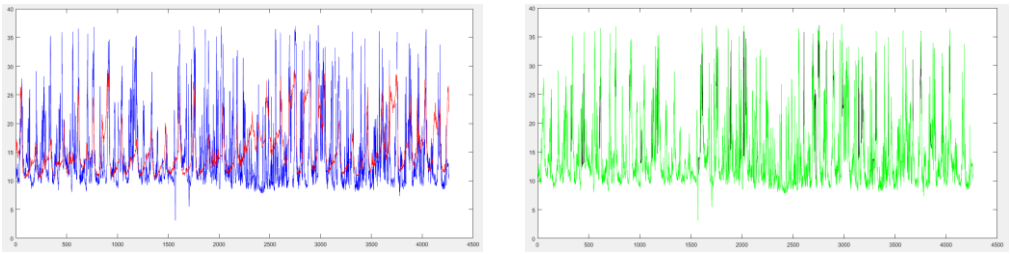


Ilustración 49. *PM2_5 Mes de Junio*

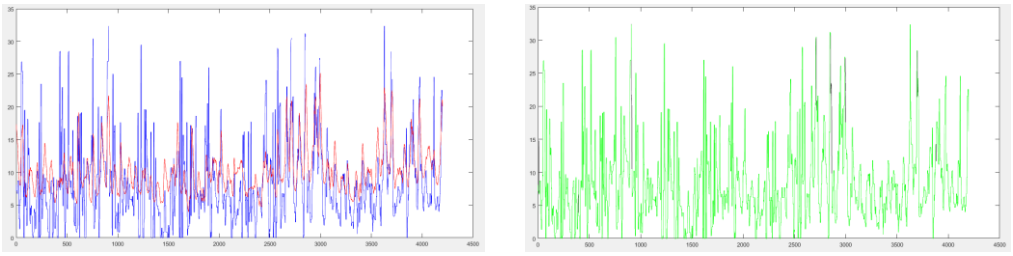


Ilustración 50. *O3 Mes de Julio*

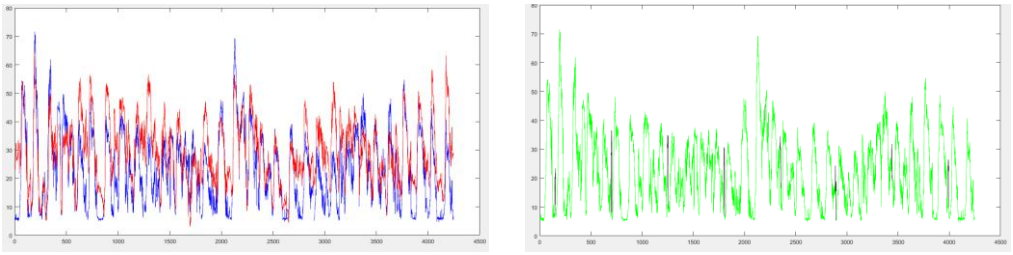


Ilustración 51. *CO Mes de Julio*

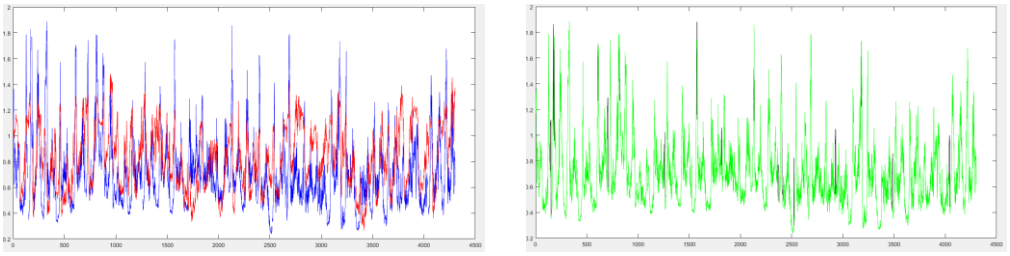


Ilustración 52. *NO2 Mes de Julio*

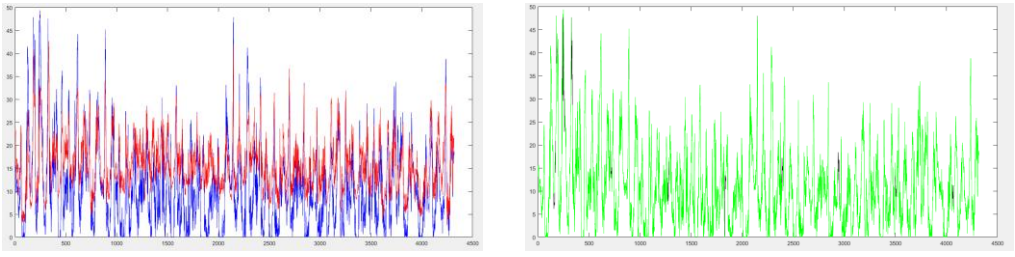


Ilustración 53. *SO2 Mes de Julio*

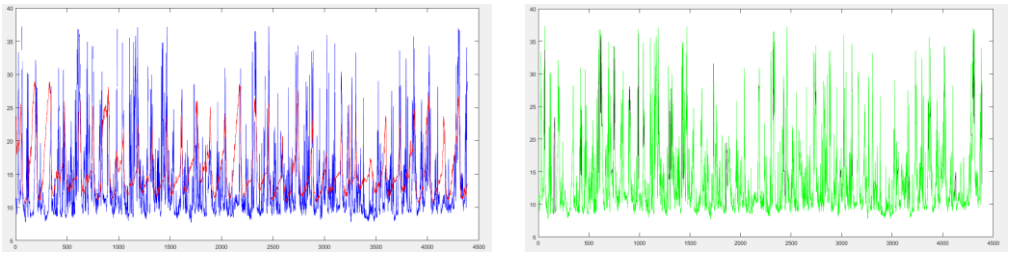


Ilustración 54. *PM2_5 Mes de Julio*

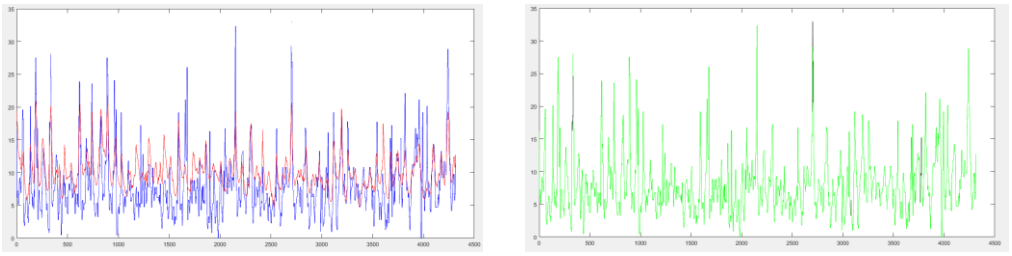


Ilustración 55. *O3 Mes de Agosto*

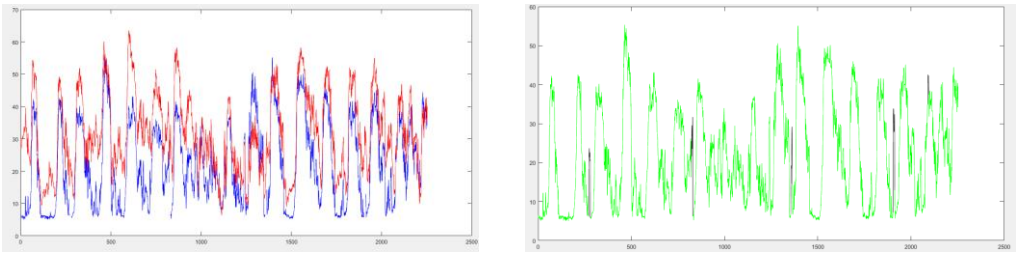


Ilustración 56. *CO Mes de Agosto*

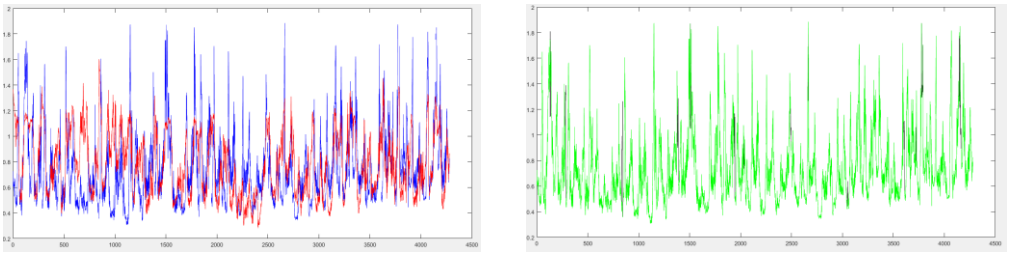


Ilustración 57. *NO2 Mes de Agosto*

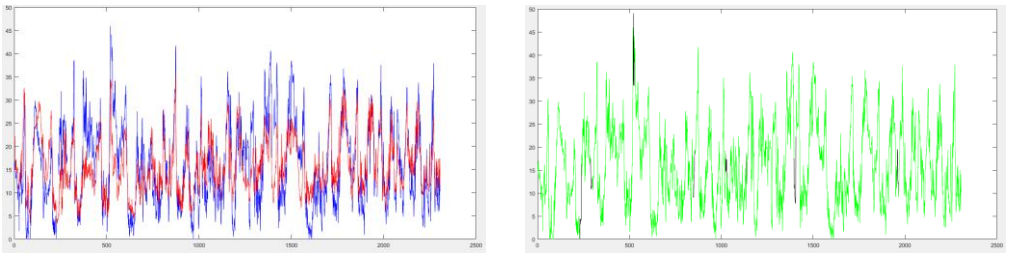


Ilustración 58. *SO2 Mes de Agosto*

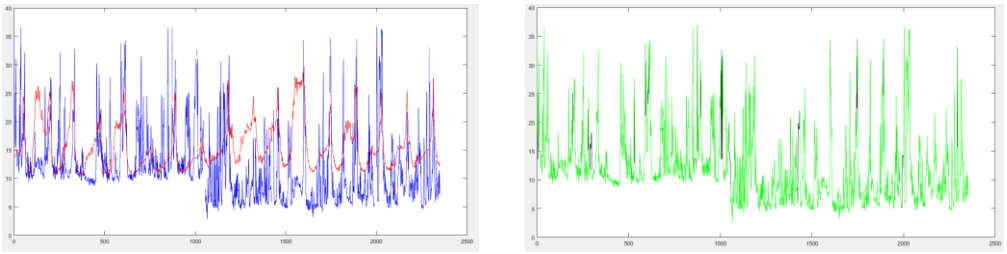


Ilustración 59. *PM2_5 Mes de Agosto*

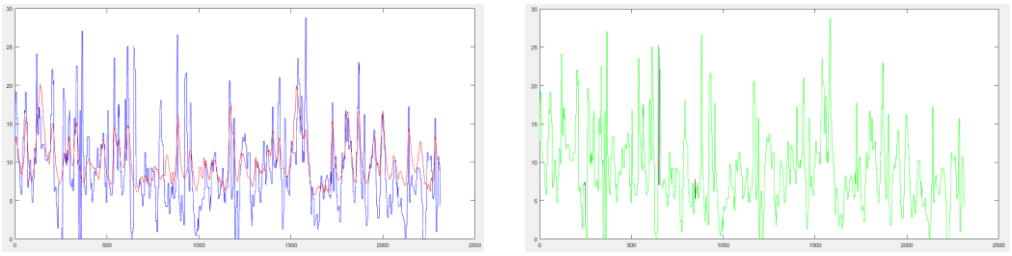


Ilustración 60. *O3 Mes de Septiembre*

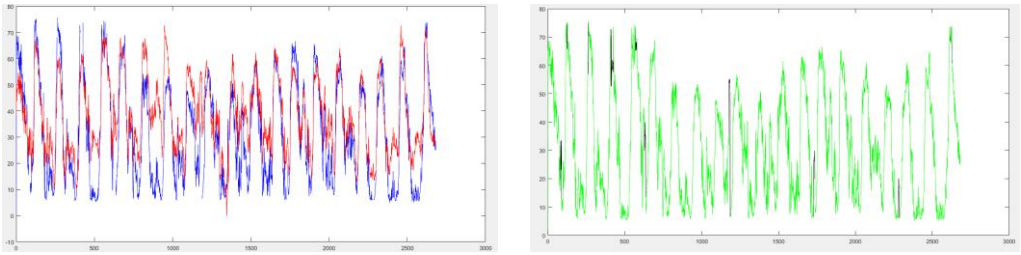


Ilustración 61. *CO Mes de Septiembre*

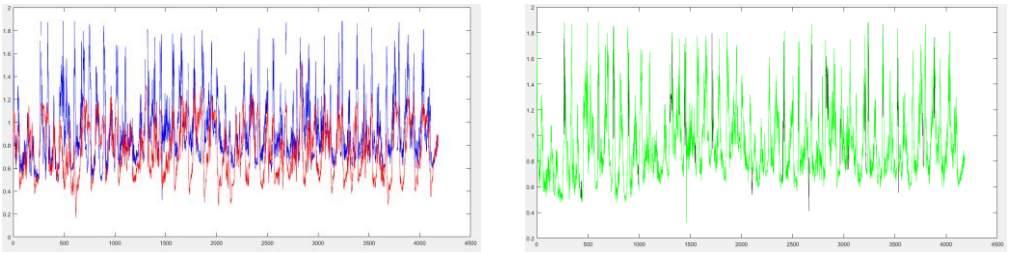


Ilustración 62. *NO2 Mes de Septiembre*

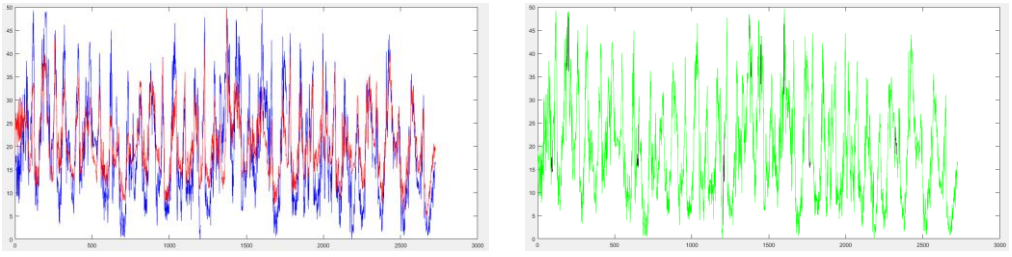


Ilustración 63. *SO2 Mes de Septiembre*

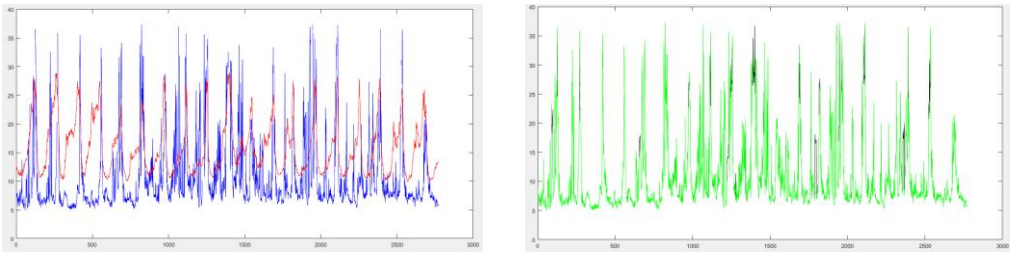


Ilustración 64. *PM2_5 Mes de Septiembre*

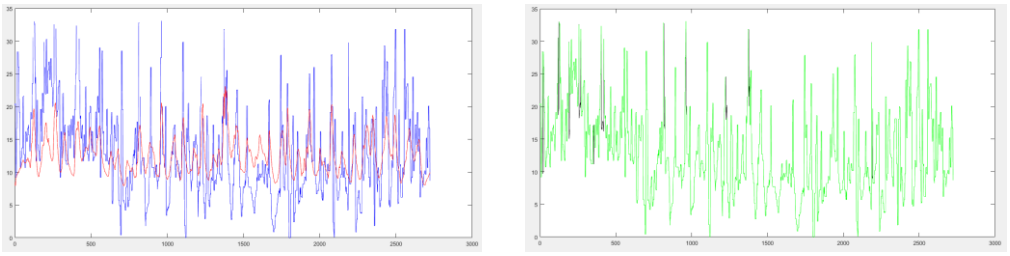


Ilustración 65. *O3 Mes de Octubre*

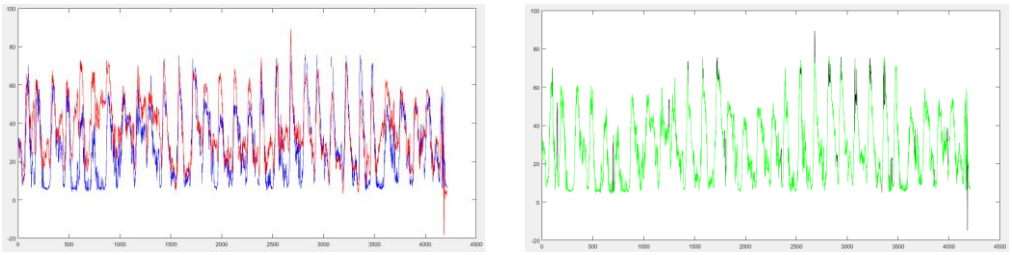


Ilustración 66. *CO Mes de Octubre*

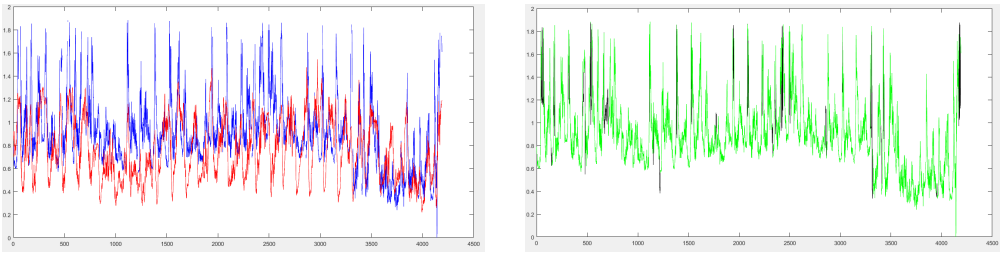


Ilustración 67. *NO2 Mes de Octubre*

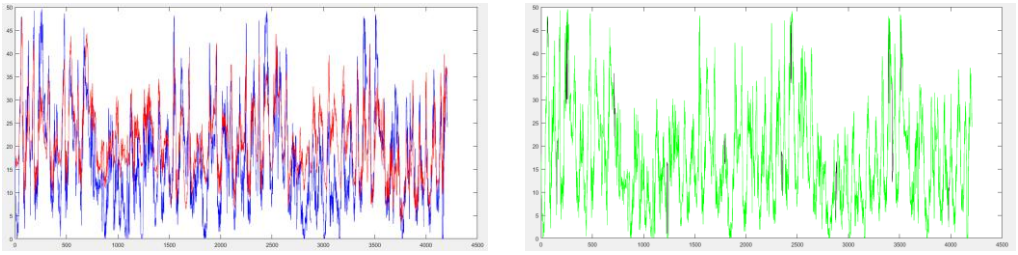


Ilustración 68. *SO2 Mes de Octubre*

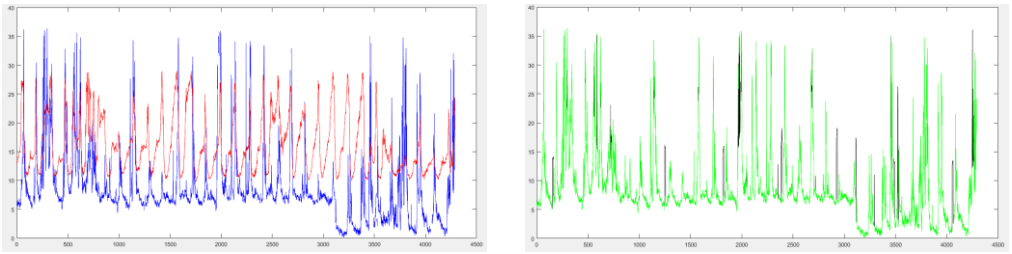


Ilustración 69. *PM2_5 Mes de Octubre*

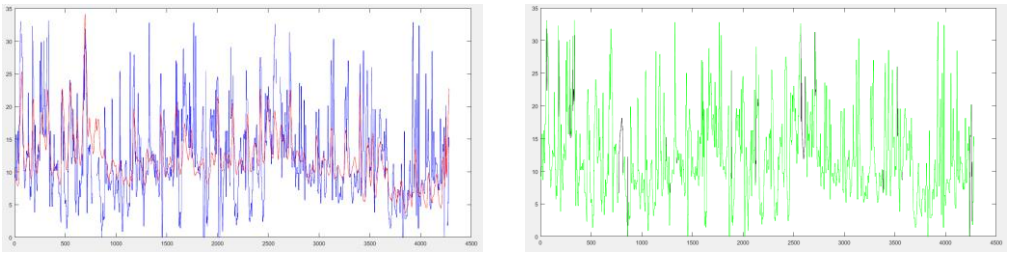


Ilustración 70. *O3 Mes de Noviembre*

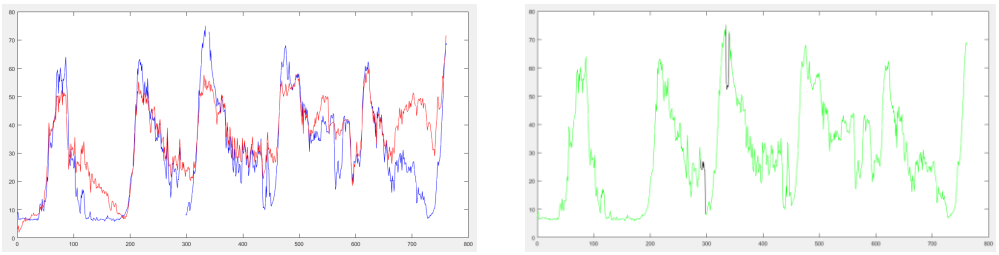


Ilustración 71. *CO Mes de Noviembre*

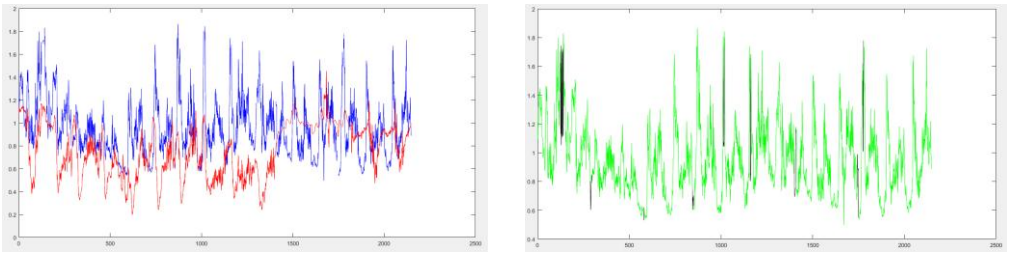


Ilustración 72. *NO2 Mes de Noviembre*

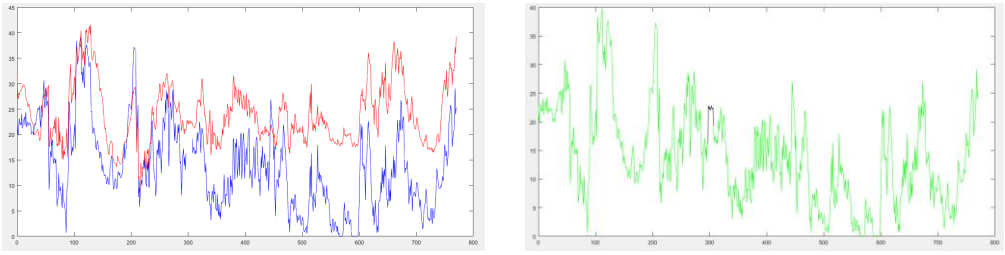


Ilustración 73. *SO2 Mes de Noviembre*

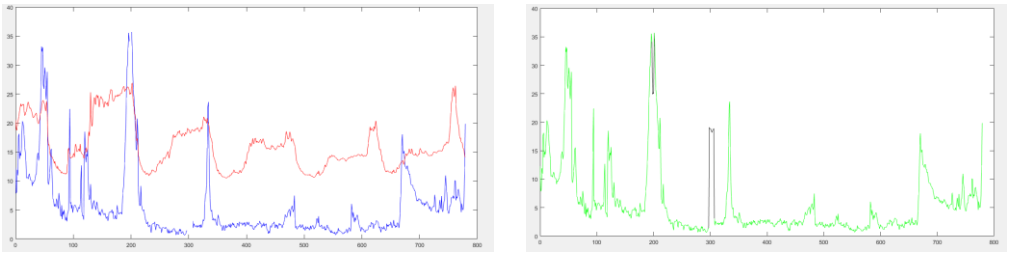
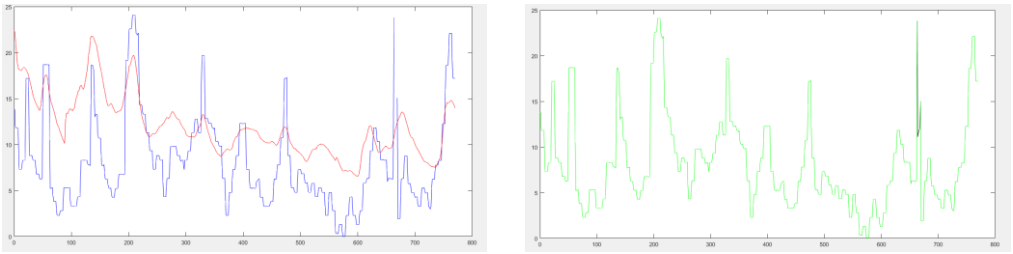


Ilustración 74. *PM2_5 Mes de Noviembre*



Doctora María Elena Ramírez Aguilar, Secretaria de la Facultad de Ciencias de la Administración de la Universidad del Azuay

CERTIFICA:

Que, el Consejo de Facultad de Ciencias de la Administración, en sesión del 31 de julio de 2019, conoció y aprobó la solicitud para la realización del trabajo de titulación y el respectivo protocolo presentado por:

Estudiante: Miguel Angel Calle Beltrán con código 73644
Tema: **Imputación de datos a partir de la búsqueda de patrones en un conjunto de datos de contaminantes atmosféricos**
Previo a la obtención del título de **Ingeniero de Sistemas y Telemática**
Director: Ing. Patricia Ortega Chasi
Tribunal: Ing. María Inés Acosta Uriguen e Ing. Chester Sellers Walden

Plazo de presentación del trabajo de titulación: El Consejo de Facultad resolvió establecer el plazo de seis meses para la presentación del trabajo de titulación concluido y calificado por el Director; este plazo se contará desde la fecha de aprobación del protocolo, esto es hasta el 31 de enero de 2020.

Cuenca, 1 de agosto de 2019

Dra. María Elena Ramírez Aguilar
Secretaria de la Facultad de Ciencias de la Administración

UNIVERSIDAD DEL AZUAY
Facultad de Ciencias de la Administración
SECRETARIA

UNIVERSIDAD DEL AZUAY

CONVOCATORIA

Por disposición de la Junta Académica de la escuela de Ingeniería de Sistemas y Telemática se convoca a los Miembros del Tribunal Examinador, a la sustentación del Protocolo del Trabajo de Titulación: **Imputación de datos a partir de la búsqueda de patrones en un conjunto de datos de contaminantes atmosféricos,** presentado por el estudiante Miguel Angel Calle Beltrán con código 73644, previa a la obtención del título de Ingeniero de Sistemas y Telemática, para el día **Jueves, 27 de junio de 2019 a las 09:40**

Tomar en cuenta que posterior a la sustentación del Diseño del Trabajo de Titulación, por ningún concepto se puede realizar modificaciones ni cambios en los documentos; únicamente, en caso de diseño aprobado con modificación, el Director adjuntará al esquema un oficio indicando que se procede con los cambios sugeridos.

Cuenca, 24 de junio de 2019

Dra. María Elena Ramírez Aguilar
Secretaria de la Facultad

Ing. Patricia Ortega Chasi .....

Ing. María Inés Acosta Uriguen .....

Ing. Chester Sellers Walden .....

Oficio Nro. 051-2019-DIST-UDA

Cuenca, 18 de junio de 2019

Ingeniero,
Oswaldo Merchán Manzano
DECANO DE LA FACULTAD DE CIENCIAS DE LA ADMINISTRACIÓN
UNIVERSIDAD DEL AZUAY

De nuestras consideraciones,

La Junta Académica de la Escuela de Ingeniería de Sistemas y Telemática, reunida el día 18 de junio del 2019, revisó la documentación del trabajo de titulación denominado**"IMPUTACIÓN DE DATOS A PARTIR DE LA BÚSQUEDA DE PATRONES EN UN CONJUNTO DE DATOS DE CONTAMINANTES ATMOSFÉRICOS "**, por la/el estudiante **MIGUEL ANGEL CALLE BELTRÁN**, con código/s estudiantil **73644**, estudiante/s de la Escuela de Ingeniería de Sistemas y Telemática, y revisado por **PATRICIA ORTEGA**, previo a la obtención del título de Ingeniero de Sistemas y Telemática.

La Junta Académica considera que la documentación cumple con las normas legales y reglamentarias de la Universidad y de la Facultad de Ciencias de la Administración y designa como miembros del tribunal a María Inés Acosta y Chester Sellers, así por su digno intermedio, el conocimiento y aprobación por parte del Consejo de Facultad.

Atentamente,

Marcos Orellana Cordero
Coordinador de la Escuela de Ingeniería de Sistemas y Telemática
Universidad del Azuay

1. FECHA DE RECEPCIÓN DE PROTOCOLO: 17-06-2019 FIRMA

2. REVISIÓN DE ESTADO ACADÉMICO DEL ALUMNO:

NOMBRE: Miguel Angel Valle Beltrán

CÓDIGO: 73644.

CARRERA: Ingeniería de Sistemas y Telemática

FECHA DE INICIO DE ESTUDIOS: 17 Sep/2013

FECHA CULMINACIÓN DE ESTUDIOS: No termina.

HOMOLOGACIONES: Ninguna CARRERA PROCEDENTE X Tramitado homologación

CONVALIDACIONES: NO UNIVERSIDAD PROCEDENTE:

FECHA DE ESTA REVISIÓN: FIRMA:

DE: DRA. MARÍA ELENA RAMÍREZ, SECRETARIA

ASUNTO: ENVÍO DE PROTOCOLO DE TRABAJO DE TITULACIÓN

PARA: JUNTA ACADÉMICA DE LA CARRERA DE Ingeniería de Sistemas

TÍTULO A OTORGARSE: Ingeniero de Sistemas y Telemática

Observación:
Pendiente homologar Humanismo Cristiano

Fecha de revisión: 10/julio/2019

FIRMA:

TÍTULO DEL TRABAJO: Imputación de datos a partir de la búsqueda de patrones en un conjunto de datos de contaminantes atmosféricos

REALIZADO EN EL CURSO DE METODOLOGÍA: X SI ____ NO

FECHA DE APROBACIÓN DEL CONSEJO DE FACULTAD: 31 de julio de 2019

DIRECTOR: Ing. Patricia Ortega Chasi

TRIBUNAL: Ing. Ma. Inés Acosta Urigüen e Ing. Chester Sellers.

Firma:

UNIVERSIDAD DEL AZUAY

ACTA

SUSTENTACIÓN DE PROTOCOLO/DENUNCIA DEL TRABAJO DE TITULACIÓN

1. Nombre del estudiante: Miguel Angel Calle Beltrán
2. Código: 73644
3. Director sugerido: Ing. Patricia Ortega Chasi
4. Codirector (opcional): _____
5. Tribunal: Ing. María Inés Acosta Uriguen e Ing. Chester Sellers Walden
6. Título propuesto: **Imputación de datos a partir de la búsqueda de patrones en un conjunto de datos de contaminantes atmosféricos**
7. Aceptado sin modificaciones: ✓
8. Aceptado con las siguientes modificaciones:

_____

_____

_____

9. No aceptado
10. Justificación:

_____

_____

_____

Tribunal

Ing. Patricia Ortega Chasi

Ing. María Inés Acosta Uriguen

Ing. Chester Sellers Walden

Sr. Miguel Angel Calle Beltrán

Dra. María Elena Ramírez Aguilar
Secretaria de la Facultad

UNIVERSIDAD DEL AZUAY

RÚBRICA PARA LA EVALUACIÓN DEL PROTOCOLO DE TRABAJO DE TITULACIÓN
(Tribunal)

1. Nombre del estudiante: Miguel Angel Calle Beltrán
2. Código: 73644
3. Director sugerido: Ing. Patricia Ortega Chasi
4. Codirector (opcional):
5. Título propuesto: **Imputación de datos a partir de la búsqueda de patrones en un conjunto de datos de contaminantes atmosféricos**
6. Revisores tribunal: Ing. María Inés Acosta Uriguen e Ing. Chester Sellers Walden

	Cumple	No cumple
Problemática y/o pregunta de investigación		
1. ¿Presenta una descripción precisa y clara?	✓	
2. ¿Tiene relevancia profesional y social?	✓	
Objetivo general		
3. ¿Concuerda con el problema formulado?	✓	
4. ¿Se encuentra redactado en tiempo verbal infinitivo?	✓	
Objetivos específicos		
5. ¿Permiten cumplir con el objetivo general?	✓	
6. ¿Son comprobables cualitativa o cuantitativamente?	✓	
Metodología		
7. ¿Se encuentran disponibles los datos y materiales mencionados?	✓	
8. ¿Las actividades se presentan siguiendo una secuencia lógica?	✓	
9. ¿Las actividades permitirán la consecución de los objetivos específicos planteados?	✓	
10. ¿Las técnicas planteadas están de acuerdo con el tipo de investigación?	✓	
Resultados esperados		
11. ¿Son relevantes para resolver o contribuir con el problema formulado?	✓	
12. ¿Concuerdan con los objetivos específicos?	✓	
13. ¿Se detalla la forma de presentación de los resultados?	✓	
14. ¿Los resultados esperados son consecuencia, en todos los casos, de las actividades mencionadas?	✓	

Nota sobre 10 puntos: : 10/

Ing. Patricia Ortega Chasi

Ing. María Inés Acosta Uriguen

Ing. Chester Sellers Walden

UNIVERSIDAD DEL AZUAY

Facultad de Ciencias de la Administración
Escuela de Ingeniería de Sistemas y Telemática

Cuenca, 17 de Junio de 2019

Ingeniero,
Oswaldo Merchán Manzano
DECANO DE LA FACULTAD DE CIENCIAS DE LA ADMINISTRACIÓN
UNIVERSIDAD DEL AZUAY

De mi consideración,

Estimado Señor Decano, yo **Miguel Angel Calle Beltrán** con C.I. **0106546179**, código estudiantil **73644**; estudiante de la Carrera de Ingeniería de Sistemas y Telemática, solicito encarecidamente a usted, la aprobación del protocolo de trabajo de titulación con el tema **"Imputación de datos a partir de la búsqueda de patrones en un conjunto de datos de contaminantes atmosféricos"** previo a la obtención del título de Ingeniero de Sistemas y Telemática para lo cual adjunto la documentación respectiva.

Por la favorable acogida que sepa dar a la presente, anticipo mi más sincero agradecimiento.

Atentamente,

Miguel Angel Calle Beltrán
Estudiante de la Escuela de Ingeniería de Sistemas y Telemática
Universidad del Azuay

0898959

UNIVERSIDAD
DEL AZUAY

LA SECRETARIA DE LA FACULTAD DE CIENCIAS DE LA ADMINISTRACION DE LA UNIVERSIDAD DEL AZUAY

CERTIFICA:

Que, el señor MIGUEL ANGEL CALLE BELTRAN, con número de cédula de identidad 0106546179, código de estudiante Nro. 73644, alumno de la carrera de INGENIERÍA DE SISTEMAS Y TELEMÁTICA, tiene aprobado el 89,96% de créditos de su malla curricular.

Cuenca, 28 de mayo de 2019

Dra. María Elena Ramírez Aguilar

SECRETARIA DE LA FACULTAD DE CIENCIAS DE LA ADMINISTRACION

UNIVERSIDAD DEL AZUAY
Facultad de Ciencias de la Administración
SECRETARÍA

Derecho Nro. N

rgp.-

Edición Autorizada
10 000 ejemplares
895 501 - 905 500
Nº 0904371

UNIVERSIDAD DEL AZUAY

UNIVERSIDAD DEL AZUAY

Facultad de Ciencias de la Administración
Escuela de Ingeniería de Sistemas y Telemática

Cuenca, 17 de Junio de 2019

Ingeniero,
Oswaldo Merchán Manzano
DECANO DE LA FACULTAD DE CIENCIAS DE LA ADMINISTRACIÓN
UNIVERSIDAD DEL AZUAY

De mi consideración,

Yo, **Patricia Ortega** informo que he revisado el protocolo de trabajo de titulación, elaborado previo a la obtención del título de Ingeniero de Sistemas y Telemática, **"Imputación de datos a partir de la búsqueda de patrones en un conjunto de datos de contaminantes atmosféricos"**, realizado por el estudiante **Miguel Angel Calle Beltrán,** con código estudiantil **73644,** protocolo que a mi criterio, cumple con los lineamientos y requerimientos establecidos por la carrera.

Por lo expuesto, me permito sugerir que sea considerado para la revisión y sustentación del mismo,

Sin otro particular, me suscribo.

Atentamente,

Patricia Ortega, PhD
Universidad del Azuay

Edición Autorizada
10.000 ejemplares
895.501 - 905.500
N° 0898961

UNIVERSIDAD DEL AZUAY

DENUNCIA/PROTOCOLO DE TRABAJO DE TITULACIÓN

1. DATOS GENERALES

1.1 Nombre del estudiante: Miguel Angel Calle Beltrán

1.1.1 Código: 73644

1.1.2 Contacto:

Telf.: 07-2262-732

Cel.: 0984649681

Correo: miguel_angel2995@hotmail.com

1.2 Director sugerido: Patricia Margarita Ortega Chasi, PhD

1.2.1 Contacto:

Telf.: 0992997533

Correo: portega@uazuay.edu.ec

1.3 Co-director sugerido: Ing. Marcos Patricio Orellana Cordero

1.3.1 Contacto:

Telf.: 0999955611

Correo: marore@uazuay.edu.ec

1.4 Asesor metodológico: Daniela Elisabet Ballari, PhD.

1.5 Tribunal designado:

1.6 Aprobación:

1.7 Línea de investigación de la carrera:

1.7.1 Código UNESCO:

1203 Ciencia de Los Ordenadores

120304 Inteligencia Artificial

120317 Informática

1.7.2 Tipo de Trabajo: Proyectos Técnicos – Investigación

1.8 Área de estudio:

Minería de Datos

1.9 Título propuesto:

"Imputación de datos a partir de la búsqueda de patrones en un conjunto de datos de contaminantes atmosféricos"

1.10 Estado del proyecto: Nuevo

2. CONTENIDO

2.1 Motivación de la investigación:

El crecimiento exponencial del suelo urbano y la industrialización en los últimos años ha aumentado significativamente la contaminación atmosférica afectando principalmente el ámbito de la salud (Alvarez et al., 2017). El problema de la contaminación ha afectado la salud humana con las exposiciones tanto de material particulado (PM), como de monóxido de carbono, dióxido de azufre y otros gases (Cevallos, Díaz, & Sirois, 2017). Los gases tóxicos existentes en el aire son realmente preocupantes, debido a que causan problemas a la salud de todos los habitantes, sin embargo los más vulnerables son los niños, esto debido al tamaño pequeño de sus órganos respiratorios y bajas defensas en sus mecanismos (Estrella, Sempértegui, Franco, Cepeda, & Naumova, 2019). Esta contaminación es algo que está causando problemas, tanto en la salud como al medio ambiente, siendo la causante de lluvias ácidas, destrucción de la capa de ozono, y además del calentamiento global (Kingsy et al., 2017).

2.2 Problemática:

Desde un punto de vista tecnológico, los datos que se tienen acerca de los contaminantes, exposición, comportamiento, factores ambientales y salud pública es cada vez más relevante. La abundante cantidad de datos permite tener un gran potencial para procesar información y observar sus comportamientos, sin embargo los métodos tradicionales existentes son poco adecuados para los grandes volúmenes de datos que se obtienen. Esto ha dado como resultado la búsqueda de nuevas formas de análisis para avanzar con la comprensión de datos (Bellinger, Mohomed Jabbar, Zaïane, & Osornio-Vargas, 2017). Por ello, se puede describir que la minería de datos es una disciplina hibrida, en donde se permite la integración de tecnologías como bases de datos, estadísticas, aprendizaje automático, computación de alto rendimiento, etc. (Li & Shue, 2004). Los algoritmos de minería de datos, y entre ellos las redes neuronales, son cada vez más utilizados en la investigación de temas relacionados a la contaminación del aire (Bellinger et al., 2017).

Dentro de la ciudad de Cuenca, y específicamente en la Universidad del Azuay, se han realizado investigaciones conjuntamente con el Instituto de Estudios de Régimen Seccional del Ecuador (IERSE), departamento que cuenta con datos de los contaminantes atmosféricos de la ciudad. Estos datos han sido combinados con algoritmos de minería de datos en proyectos de investigación anteriores. En una primera fase se realizó el trabajo de titulación "Búsqueda de patrones a partir del comportamiento de los contaminantes atmosféricos" (Ortega, 2018), y posteriormente el trabajo de titulación "Relación de los contaminantes con datos de tipo meteorológico" (Andrade, 2018), sin embargo estos estudios tuvieron varios datos

perdidos al realizar su análisis, debido a que el sensor en muchas ocasiones tiene inconvenientes y no obtiene toda la información. Es por ello, que es necesario direccionar la imputación de datos perdidos. En este trabajo se parte de la premisa que las redes neuronales, combinadas con una técnica angular desarrollada en el marco de este trabajo, la cual consiste en obtener el valor del ángulo del contaminante y el nivel en el que se encuentra, permitiría obtener resultados satisfactorios en cuanto a la imputación de datos.

2.3 Pregunta de investigación:

¿Qué tan buena resulta la predicción mediante la implementación de una técnica angular para la imputación de datos?

2.4 Resumen

En la ciudad de Cuenca se ha reconocido la importancia del monitoreo de contaminantes atmosféricos. De allí la existencia de muchos proyectos en el ámbito de la contaminación dentro de la ciudad, sin embargo todos ellos se basan en los datos medidos con sensores en estaciones de monitoreo que frecuentemente pierden datos. Por ello, este proyecto tiene como objetivo solucionar el inconveniente de pérdida de información, a partir de la imputación de datos de los contaminantes atmosféricos. Para cumplir con el objetivo establecido en este trabajo primero se llevará a cabo un análisis de comportamiento de patrones y posteriormente con la aplicación de redes neuronales se establecerá los valores de variables de contaminantes para imputar con la mayor exactitud posible, logrando esto a través de predicciones.

2.5 Estado del arte y marco teórico

Data mining es una disciplina que ha permitido extraer conocimiento de gran cantidad de datos relacionados con la contaminación del medio ambiente, de esta manera estudiar los patrones y su comportamiento (Gore & Deshpande, 2017). Una de las herramientas para minería de datos son las redes neuronales, usadas en muchos proyectos para la predicción del comportamiento de contaminantes en distintas partes del mundo (Azid et al., 2014; Kurt, Gulbagci, Karaca, & Alagha, 2008; Lo, Lu, & Fan, 2003). Las redes neuronales han sido utilizadas en su gran mayoría debido a la gran certeza en comparación con otras herramientas que realizan la misma función (Viotti, Liuti, & Di Genova, 2002).

Las redes neuronales han sido utilizadas en diferentes ámbitos en minería de datos donde han sido aplicadas por ejemplo en la predicción del comportamiento meteorológico para la toma de decisiones en las actividades agrícolas (Fuentes, Campos, & García-Loyola, 2018). En ambientes urbanos contaminados como por ejemplo la ciudad de Perugia se realizó un análisis de predicción de contaminantes utilizando variables monitorizadas como dióxido de azufre, óxidos de nitrógeno, PM10, benceno, monóxido de carbono, ozono, velocidad del viento horizontal, humedad, presión, temperatura, radiación solar total (Viotti et al., 2002), sin embargo las predicciones del comportamiento meteorológico son difíciles de realizar, debido a la alta cantidad de variación de parámetros y comportamiento que se deben considerar (Geetha & Nasira, 2015). Desde este punto de vista, las redes neuronales son una herramienta muy utilizada, ya que una vez entrenadas incorporan un análisis más rápido para poder predecir series de tiempo con alta exactitud en sus predicciones (Kukkonen et al., 2003).

Varios estudios se han centrado en el análisis de los contaminantes atmosféricos, dichos estudios se han realizado utilizando diferentes herramientas como son: redes neuronales, arboles de decisiones, lógica difusa, algoritmos genéticos, entre otros (Geetha & Nasira, 2015), sin embargo algo que no se ha podido evidenciar en trabajos anteriores es una técnica que permita buscar patrones dentro del comportamiento de los contaminantes utilizando el ángulo que se forma entre los niveles en un determinado intervalo de tiempo y el nivel en el cual se encuentra dicho contaminante. Permitiendo a partir del ángulo buscar patrones, para posteriormente realizar predicciones de una mejor manera.

2.6 Objetivo general

Imputar datos perdidos a través de predicciones de los contaminantes atmosféricos, utilizando redes neuronales y una técnica angular.

2.7 Objetivos específicos

- Obtener el valor del ángulo de los contaminantes atmosféricos y su nivel.
- Entrenamiento de la red neuronal para realizar la predicción del comportamiento del contaminante en el tramo perdido.
- Analizar los resultados de las predicciones obtenidas.

2.8 Metodología

CRISP-DM

Para el desarrollo del trabajo de titulación se tiene previsto la utilización la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining), esto debido a un análisis previo entre otras metodologías como son KDD y SEMMA. CRISP-DM es un método probado para orientar a la realización de trabajos sobre minería de datos. Esta metodología tiene un ciclo de vida muy descriptivo, en donde muestra las fases normales de un proyecto, las tareas necesarias para cada fase y explicación entre tareas (IBM, 2012).

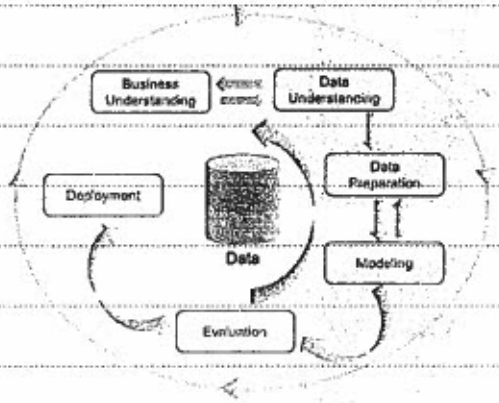


Figura 1. Modelo CRISP-DM. Fases en las que consiste el modelo CRISP-DM (IBM, 2012).

El ciclo de vida de la metodología está compuesto por 6 etapas, cada una señalada con flechas, las cuales muestran la trayectoria de las fases y la importancia entre las mismas. La

metodología no es secuencial estrictamente, los proyectos pueden avanzar o retroceder las fases si es necesario, permitiendo ser flexible y personalizable según se necesite, además CRISP-DM permite crear modelos de minería de datos que se adapten a las necesidades concretas.

Las etapas o fases con las que cuenta el modelo CRISP-DM son las siguientes: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación. Desde la evaluación se puede analizar si se puede implementar directamente o seguir en el ciclo del modelo para mejorar ciertos aspectos.

2.9.1 Área de estudio

Este trabajo de titulación se centrará específicamente en datos de la ciudad de Cuenca – Ecuador (Figura 2). Cuenca se encuentra ubicada en la región sierra, perteneciente a la provincia del Azuay. Esta ciudad está localizada a 2560 metros sobre el nivel de mar, en las coordenadas 2° 53' 51"S y 79° 00' 16" O. Cuenca cuenta con un clima subtropical de montaña, que se caracteriza por abundantes precipitaciones y temperaturas regularmente frías.

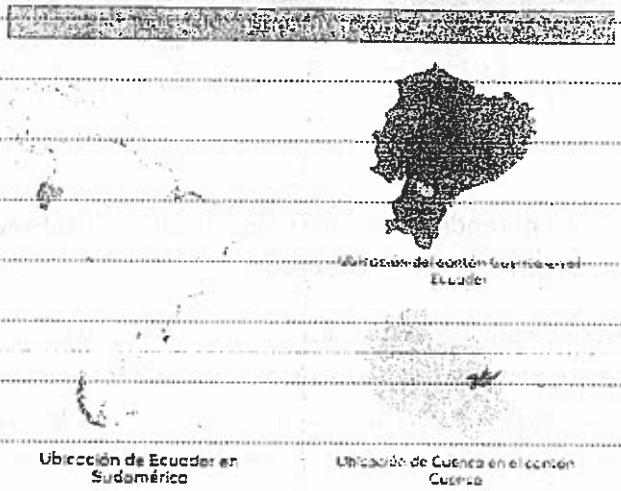


Figura 2. *Ubicación del cantón Cuenca* Fuente: (EMOV, 2017).

2.9.2 Materiales

Los datos que serán utilizados para realizar la investigación fueron obtenidos desde la estación automática de monitoreo del aire con coordenadas 2° 89' 74"S y 79° 00' 31"S (Figura 3), estación que pertenece a la empresa EMOV EP (Red de Monitoreo de Calidad del Aire de Cuenca de la EMOV EP), en la cual se miden y almacenan los datos referentes a la calidad del aire dentro de la ciudad de Cuenca (EMOV, 2017). Los datos con los que se trabajarán en la investigación pertenecen al periodo enero del 2018 hasta noviembre del 2018. Los contaminantes utilizados serán los siguientes:

Contaminantes Atmosféricos

- Ozono (O3).
- Monóxido de Carbono (CO).
- Dióxido de nitrógeno (NO2).
- Dióxido de Azufre (SO2).
- Material Particulado (PM2.5).

Variables Meteorológicas

- Temperatura del Aire.

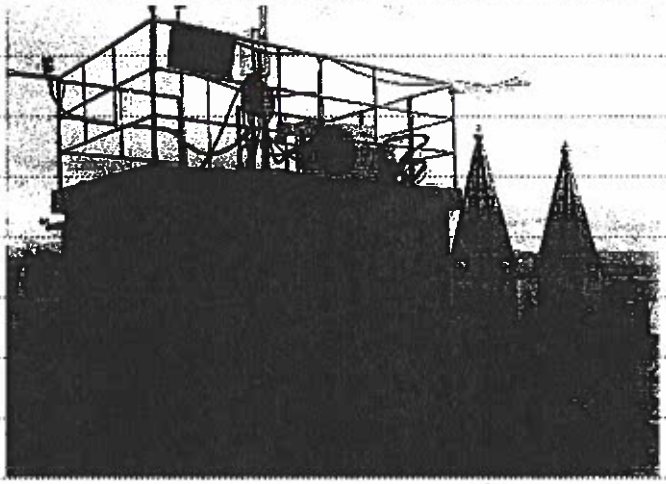


Figura 3. *Estación automática de calidad del aire y meteorología ubicada en la estación MUN (Municipio)*: Fuente (EMOV, 2017)

2.9.3 Métodos

El método a utilizar esta comprendido por varias fases (Figura 4): Datos de los contaminantes, tratamiento de datos, entrenamiento de red neuronal, predicción y análisis de resultados.

La primera fase se describe como "Datos de los contaminantes" y consiste en la obtención de datos descritos en la parte de materiales. Estos datos son obtenidos mediante la estación automática de la empresa EMOV. En esta fase se necesita clasificar los datos, obteniendo solo los contaminantes que se van a analizar en este proyecto y familiarizarse con las medidas en las que se encuentran.

En una segunda fase, se describe la manera en la cual son tratados todos los datos obtenidos, en donde se deberá obtener el ángulo de los contaminantes en un tramo de una hora, logrando esto mediante un punto de inicio y un punto final, además de tener en cuenta el nivel en el que se encuentra el contaminante. Para obtener el ángulo se realizarán varias pruebas para seleccionar la mejor opción. En un inicio se tiene previsto tener el valor del ángulo por medio de la formula trigonométrica del arco tangente, en la cual se necesitan valores del lado opuesto y lado adyacente para obtener el valor del ángulo, datos que se tienen.

En esta fase también se necesita obtener la información de los contaminantes, datos que son relevante al estudio realizado en un tramo dado, datos como son:

- Nombre del contaminante
- Nivel inicial del contaminante
- Nivel final del contaminante
- Valor del ángulo
- Hora de inicio
- Hora final
- Fecha

En una tercera fase se realizará una de las partes más importantes en la investigación como es el Entrenamiento de la red neuronal. Este paso consiste en realizar el entrenamiento de la red neuronal con un set de datos. Para determinar el set de datos para entrenamiento, pruebas y validación, esta etapa se considerarán técnicas de validación como: validación simple, leave-one-out validation, leave-group validation o k-fold Cross validation.

La cuarta fase del proyecto consiste en la parte de predicción para el relleno de datos, que es el objetivo de la investigación, logrando esto a través con la red neuronal ya entrenada, conjuntamente con el 20% del set de validación.

Para finalizar el proyecto se tiene previsto realizar un análisis de los resultados dados con las predicciones realizadas, en donde se compararán los datos que se predicen con los verdaderos que fueron separados en el 20% de validación. De esta manera se analizará el nivel de error tiene la predicción y se determinará cuán confiables son los datos que se predicen. Esta validación permite visualizar que tan fiable es la imputación de datos con la técnica utilizada. Para lograr la validación es necesario utilizar una herramienta que permite sacar el error de todos los datos procesados, la herramienta que permite realizar los cálculos de una manera adecuada es el software R, la cual trabaja de una excelente manera en entornos de computación estadística y gráficos.

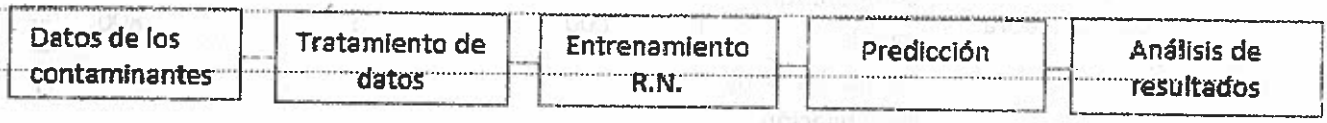


Figura 4. Diagrama que muestra las fases en las que va a consistir le elaboración de la investigación; Fuente (Propio)

2.9 Alcances y resultados esperados

Obtener de manera correcta el valor angular del nivel del contaminante atmosférico, algo primordial y necesario para el proyecto. Posteriormente se tiene previsto entrenar la red neuronal de buena manera con el número de capas adecuado para obtener mejores resultados y por último que las predicciones sean las más cercanas posibles a los datos perdidos.

2.10 Supuestos y riesgos

Riesgo	Probabilidad	Alternativa de solución
No obtención del ángulo del contaminante con una escala dada en la medición de niveles	Media	Buscar otra alternativa como pueden ser con identidades trigonométricas o creación de tablas que permitan obtener el valor angular, según el nivel del contaminante

Predicción no satisfactoria de datos faltantes	Alta	Relacionar con otro tipo de patrones, que permitan solucionar el inconveniente
No cumplir con el tiempo establecido	Alta	Adecuarse en siguientes etapas para igualarse en tiempo, o realizar la petición de prórroga para cumplir con el proyecto de investigación

2.11 Presupuesto

Rubro	Justificación	Costo	Cantidad	Costo Total
Mano de obra	Tiempo para elaborar la investigación.	800	4 (meses)	3200
Computadora	Uso de computadora para la realización de la investigación	800	1	800
Impresiones	Gastos de impresiones para presentar la documentación del tema de titulación	50	1	50
				4050 $

2.12 Financiamiento: Propio

2.13 Esquema tentativo

Capítulo 1: Introducción

Capítulo 2: Estado del arte

Capítulo 3: Desarrollo

- **3.1** Cálculo del valor angular
- **3.2** Red neuronal
- **3.3** Imputación de datos

Capítulo 4: Resultados

- **4.1** Validación de resultados
- **4.2** Discusión

a. Cronograma

Objetivo específico	Actividad	Mes 1				Mes 2				Mes 3				Mes 4			
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Tratamiento de información	Revisión de documentos, tutoriales y técnicas realizadas para imputación de datos																
	Obtener información de los contaminantes, calculando el ángulo																
Imputación de datos	Revisar metodologías para implementar redes neuronales																
	Entrenamiento de la red neuronal																
	Predicción y relleno de datos de contaminantes atmosféricos																
Análisis de resultados	Analizar que tan buenos resultados brinda la predicción con la redes neuronales																
Escritura de la investigación	Redacción de la investigación realizada																

b. Referencias

Alvarez, I., Méndez Martínez, J., Bello Rodríguez, B. M., Benítez Fuentes, B., Escobar Blanco, L. M., & Zamora Monzón, R. (2017). Influencia de los contaminantes atmosféricos sobre la salud. *Revista Médica Electrónica*, *39*(5), 1160–1170. Retrieved from http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1684-18242017000500017

Andrade, P. S. (2018). *Aplicación de minería de datos en el análisis de contaminantes atmosféricos y variables meteorológicas*.

Azid, A., Juahir, H., Toriman, M. E., Kamarudin, M. K. A., Saudi, A. S. M., Hasnam, C. N. C., … Yamin, M. (2014). Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: A case study in Malaysia. *Water, Air, and Soil Pollution*, *225*(8). https://doi.org/10.1007/s11270-014-2063-1

Bellinger, C., Mohomed Jabbar, M. S., Zaïane, O., & Osornio-Vargas, A. (2017). A systematic review of data mining and machine learning for air pollution epidemiology. *BMC Public Health*, *17*(1), 1–19. https://doi.org/10.1186/s12889-017-4914-3

Cevallos, V. M., Díaz, V., & Sirois, C. M. (2017). Particulate matter air pollution from the city of Quito, Ecuador, activates inflammatory signaling pathways in vitro. *Innate Immunity*, *23*(4), 392–400. https://doi.org/10.1177/1753425917699864

EMOV. (2017). *Informe de calidad del aire Cuenca*. 1–123.

Estrella, B., Sempértegui, F., Franco, O. H., Cepeda, M., & Naumova, E. N. (2019). Air pollution control and the occurrence of acute respiratory illness in school children of Quito, Ecuador. *Journal of Public Health Policy*, *40*(1), 17–34. https://doi.org/10.1057/s41271-

018-0148-6

Fuentes, M., Campos, C., & García-Loyola, S. (2018). Application of artificial neural networks to frost detection in central Chile using the next day minimum air temperature forecast. *Chilean Journal of Agricultural Research*, *78*(3), 327–338. https://doi.org/10.4067/s0718-58392018000300327

Geetha, A., & Nasira, G. M. (2015). Data mining for meteorological applications: Decision trees for modeling rainfall prediction. *2014 IEEE International Conference on Computational Intelligence and Computing Research, IEEE ICCIC 2014*, 0–3. https://doi.org/10.1109/ICCIC.2014.7238481

Gore, R. W., & Deshpande, D. S. (2017). An approach for classification of health risks based on air quality levels. *Proceedings - 1st International Conference on Intelligent Systems and Information Management, ICISIM 2017*, *2017-Janua*, 58–61. https://doi.org/10.1109/ICISIM.2017.8122148

IBM, I. B. M. (2012). Manual CRISP-DM de IBM SPSS Modeler. *IBM Corporation*, 56. Retrieved from ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/es/CRISP-DM.pdf

Kingsy, G. R., Manimegalai, R., Geetha, D. M. S., Rajathi, S., Usha, K., & Raabiathul, B. N. (2017). Air pollution analysis using enhanced K-Means clustering algorithm for real time sensor data. *IEEE Region 10 Annual International Conference, Proceedings/TENCON*, (August 2006), 1945–1949. https://doi.org/10.1109/TENCON.2016.7848362

Kukkonen, J., Partanen, L., Karppinen, A., Ruuskanen, J., Junninen, H., Kolehmainen, M., … Cawley, G. (2003). Extensive evaluation of neural network models for the prediction of NO 2 and PM10 concentrations, compared with a deterministic modelling system and measurements in central Helsinki. *Atmospheric Environment*, *37*(32), 4539–4550. https://doi.org/10.1016/S1352-2310(03)00583-1

Kurt, A., Gulbagci, B., Karaca, F., & Alagha, O. (2008). An online air pollution forecasting system using neural networks. *Environment International*, *34*(5), 592–598. https://doi.org/10.1016/j.envint.2007.12.020

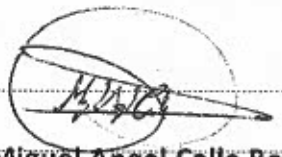
Li, S. T., & Shue, L. Y. (2004). Data mining to aid policy making in air pollution management. *Expert Systems with Applications*, *27*(3), 331–340. https://doi.org/10.1016/j.eswa.2004.05.015

Lo, S. M., Lu, W. Z., & Fan, H. Y. (2003). Application of evolutionary neural network method in predicting pollutant levels in downtown area of Hong Kong. *Neurocomputing*, *51*, 387–400.

Ortega, J. J. (2018). Impacto de la aplicación de algoritmos de minería de datos en variables de contaminación del aire. *Director*, *15*(2), 2017–2019. https://doi.org/10.22201/fq.18708404e.2004.3.66178

Viotti, P., Liuti, G., & Di Genova, P. (2002). Atmospheric urban pollution: Applications of an artificial neural network (ANN) to the city of Perugia. *Ecological Modelling*, *148*(1), 27–46. https://doi.org/10.1016/S0304-3800(01)00434-3

c. Firma de responsabilidad (estudiante)



Miguel Angel Calle Beltrán

d. Firma de responsabilidad (director sugerido)

Patricia Ortega, PhD

e. Fecha de entrega: 17-06-2019