



Universidad del Azuay

Facultad de Ciencias de la Administración

Escuela de Ingeniería de Sistemas y Telemática

APLICACIÓN DE MINERÍA DE DATOS EN EL
ANÁLISIS DE CONTAMINANTES
ATMOSFÉRICOS Y VARIABLES
METEOROLÓGICAS

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

Autor:

Pedro Sebastián Andrade Durazno

Director:

Ing. Marcos Patricio Orellana Cordero

**Cuenca – Ecuador
2018**

DEDICATORIA

El presente trabajo está dedicado a toda mi familia, en especial a mi madre Anita Durazno M. y mi padre Rigoberto Andrade, que con su apoyo incondicional supieron guiarme e incentivarme a lo largo de estos años motivando así mi desarrollo personal y profesional.

AGRADECIMIENTO

Quiero agradecer en primer lugar a mis padres Anita y Rigoberto, ya que con el apoyo que me han sabido brindar a lo largo de todos estos años he podido culminar una etapa más en mi vida, a mis hermanos Pablo y Daniel y en especial a mi hermana Isabel, ya que ha sido un pilar fundamental a lo largo de mi vida, quiero también agradecer a todas las personas que han creído en mí y han puesto su confianza en mi persona y a todas las personas que han colaborado para que este objetivo se cumpla.

Quiero agradecer también a mi tutor de tesis el Ing. Marcos Orellana Cordero, ya que con el apoyo y tiempo que me ha brindado se han alcanzado los objetivos planteados para el presente trabajo.

RESUMEN:

La contaminación del aire afecta directamente a la salud de las personas y causa grandes problemas al medio ambiente, se han implementado en varios países sensores que recolectan información del aire en tiempo real. Éste trabajo se enfocó en la aplicación de minería de datos en el análisis de los datos de contaminantes atmosféricos y variables meteorológicas aplicando algoritmos de clustering no supervisados. El mejor algoritmo de clustering tanto en precisión como agrupamiento de datos fue el algoritmo DBSCAN. La información se presentó a través; de matrices de correlación y series de tiempo entre variables con ventanas de duración. Entre los resultados más relevantes, por ejemplo, se obtuvo que el Ozono tiene una fuerte correlación con la temperatura.

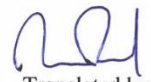
Palabras clave: Minería de datos, K-means, DBSCAN, Contaminantes atmosféricos, Variables meteorológicas.

ABSTRACT

Air pollution directly affected the health of people and caused great problems to the environment. Sensors that collected air information in real time have been implemented in several countries. This work focused on the application of data mining in the analysis of air pollutant data and meteorological variables applying unsupervised clustering algorithms. The best clustering algorithm both in precision and data grouping was the DBSCAN algorithm. The information was presented through correlation matrices and time series between variables with windows of duration. It was obtained among the most relevant results that Ozone had a strong correlation with temperature.

Keywords: Data mining, K-means, DBSCAN, Atmospheric pollutants, Weather variables.




Translated by
Ing. Paul Arpi

ÍNDICE

DEDICATORIA.....	I
AGRADECIMIENTO.....	II
RESUMEN:.....	III
ABSTRACT:.....	IV
1. INTRODUCCIÓN	1
2. MÉTODO.....	3
2.1 Zona de Estudio.....	3
2.2 Datos utilizados	3
Contaminantes Atmosféricos	4
Variables Meteorológicas.....	4
2.3 Software Utilizado.....	5
2.3.1 Rapidminer	5
2.3.2 Software Estadístico R	5
2.4 Metodología	5
2.4.1 CRISP-DM.....	5
2.4.2 Comprensión del negocio o problema.....	6
2.4.3 Comprensión de los datos	7
2.4.4 Preparación de los datos	7
2.4.5 Modelado de los datos	8
2.4.6 Evaluación.....	8
2.4.7 Implementación.....	8
2.5 Desarrollo	9
2.5.1 Comprensión de los datos	9
2.5.1.1 Variables meteorológicas	9
2.5.1.2 Contaminantes atmosféricos.....	11
2.5.2 Preparación de los datos	13
2.5.3 Modelado de datos	14
2.5.4 Evaluación.....	16
2.5.5 Implementación.....	19
3. RESULTADOS.....	20
4. DISCUSIÓN	27
5. CONCLUSIONES	29
BIBLIOGRAFÍA	30
ANEXOS	32

FIGURAS

Figura 1 Mapa de la ciudad Cuenca – Ecuador.....	3
Figura 2 Modelo de proceso CRISP-DM.....	6
Figura 3 Etapas de preparación de los datos	13
Figura 4 Set de datos final.....	14
Figura 5 Modelo de Clustering	15
Figura 6 Prueba de normalidad de datos	16
Figura 7 Clasificación de cada dato dentro del cluster más grande de cada algoritmo	18
Figura 8 Suma de acuerdo al número de coincidencias por cada algoritmo	18
Figura 9 Matriz de correlación en base al set de datos obtenido.....	19
Figura 10 Comportamiento entre el Ozono(O3) y la Velocidad del viento	20
Figura 11 Comportamiento entre el Ozono(O3) y la Presión Atmosférica.....	21
Figura 12 Comportamiento entre el Monóxido de Carbono y la Radiación UVA.....	22
Figura 13 Comportamiento entre el Monóxido de Carbono y el Punto de Rocío.....	22
Figura 14 Comportamiento entre Dióxido de Nitrógeno y la Humedad Relativa.....	23
Figura 15 Comportamiento entre Dióxido de Nitrógeno (NO2) y la radiación UVE	24
Figura 16 Comportamiento entre Dióxido de Azufre y material particulado PM2.5.....	24
Figura 17 Comportamiento entre material particulado PM2.5 y Temperatura del aire	25
Figura 18 Comportamiento entre Material particulado PM2.5 y Presión atmosférica.....	26
Figura 52 Comportamiento entre Material particulado PM2.5 y Radiación UVA	48
Figura 53 Comportamiento entre Material particulado PM2.5 y Radiación UVE.....	49
Figura 54 Comportamiento entre Material particulado PM2.5 y Velocidad del viento.....	49

TABLAS

Tabla 1 Variables Meteorológicas utilizadas	9
Tabla 2 Contaminantes Atmosféricos	11
Tabla 3 Proceso de unión de set de datos.....	14
Tabla 4 Prueba de algoritmos de minería de datos.....	17

ANEXOS

Figura 19 Script para realizar unión entre set de datos	32
Figura 20 Comportamiento entre Ozono (O3) y Punto del rocío.....	32
Figura 21 Comportamiento entre Ozono (O3) y Humedad relativa.....	33
Figura 22 Comportamiento entre Ozono (O3) y Precipitación	33
Figura 23 Comportamiento entre Ozono (O3) y Radiación Global	34
Figura 24 Comportamiento entre Ozono (O3) y Temperatura del aire	34
Figura 25 Comportamiento entre Ozono (O3) y Radiación UVA	35
Figura 26 Comportamiento entre Ozono (O3) y Radiación UVE.....	35
Figura 27 Comportamiento entre Monóxido de carbono (CO) y Humedad relativa	36
Figura 28 Comportamiento entre Monóxido de carbono (CO) y Presión atmosférica	36
Figura 29 Comportamiento entre Monóxido de carbono (CO) y Precipitación	37
Figura 30 Comportamiento entre Monóxido de carbono (CO) y Radiación global.....	37
Figura 31 Comportamiento entre Monóxido de carbono (CO) y Temperatura del aire.....	38
Figura 32 Comportamiento entre Monóxido de carbono (CO) y Radiación UVE.....	38
Figura 33 Comportamiento entre Monóxido de carbono (CO) y Velocidad del viento.....	39
Figura 34 Comportamiento entre Dióxido de nitrógeno (NO2) y Punto del rocío	39
Figura 35 Comportamiento entre Dióxido de nitrógeno (NO2) y Presión atmosférica	40
Figura 36 Comportamiento entre Dióxido de nitrógeno (NO2) y Precipitación.....	40
Figura 37 Comportamiento entre Dióxido de nitrógeno (NO2) y Radiación global.....	41

Figura 38 Comportamiento entre Dióxido de nitrógeno (NO ₂) y Temperatura del aire.....	41
Figura 39 Comportamiento entre Dióxido de nitrógeno (NO ₂) y Radiación UVA	42
Figura 40 Comportamiento entre Dióxido de nitrógeno (NO ₂) y Velocidad del viento.....	42
Figura 41 Comportamiento entre Dióxido de azufre (SO ₂) y Punto del rocío	43
Figura 42 Comportamiento entre Dióxido de azufre (SO ₂) y Humedad Relativa	43
Figura 43 Comportamiento entre Dióxido de azufre (SO ₂) y Radiación Global.....	44
Figura 44 Comportamiento entre Dióxido de azufre (SO ₂) y Temperatura del aire.....	44
Figura 45 Comportamiento entre Dióxido de azufre (SO ₂) y Radiación UVA	45
Figura 46 Comportamiento entre Dióxido de azufre (SO ₂) y Radiación UVE	45
Figura 47 Comportamiento entre Dióxido de azufre (SO ₂) y la Velocidad del viento.....	46
Figura 48 Comportamiento entre Material particulado PM _{2.5} y Punto del rocío.....	46
Figura 49 Comportamiento entre Material particulado PM _{2.5} y Humedad relativa.....	47
Figura 50 Comportamiento entre Material particulado PM _{2.5} y Precipitación	47
Figura 51 Comportamiento entre Material particulado PM _{2.5} y Radiación global	48
Figura 52 Comportamiento entre Material particulado PM _{2.5} y Radiación UVA	48
Figura 53 Comportamiento entre Material particulado PM _{2.5} y Radiación UVE.....	49
Figura 54 Comportamiento entre Material particulado PM _{2.5} y Velocidad del viento.....	49

1. INTRODUCCIÓN

En la actualidad uno de los mayores problemas a nivel mundial es la contaminación del aire, la cual tiene efectos sobre el medio ambiente, y a su vez, tiene una gran incidencia sobre los problemas de la salud de las personas, llegando a causar irritación en los ojos, piel y problemas respiratorios. En casos agudos, puede causar cáncer de pulmón y neumonía. La contaminación del aire es también la mayor causante de muertes prematuras según la Organización Mundial de la Salud (OMS). Además, para ésta última, 2.4 millones de personas mueren cada año a causa de la contaminación del aire (Lanjewar & Shah, 2012).

El fenómeno se debe principalmente al incremento de la congestión vehicular y al incremento desconsiderado de la industrialización (Christina & Komathy, 2013). Esto está estrechamente relacionado con actividades de desarrollo dentro de las zonas urbanas, las cuales son responsables de cambios a largo plazo en la atmósfera, ocasionado por industrias, automóviles, aumento de la carga de tráfico y uso desmedido de combustibles (Christian, Jariwala, & Gohil, 2010).

La contaminación del aire no solamente causa efectos sobre las personas, sino también afecta al medio ambiente, ya que causa principalmente smog o lluvias ácidas y es causante principal del llamado calentamiento global (Ri et al., 2016). Es por esto que, desde hace muchos años la predicción del comportamiento de la meteorología ha sido una tarea muy difícil, debido a la gran cantidad y variabilidad de parámetros y comportamientos que deben considerarse (Geetha & Nasira, 2015).

Debido a esto, en la actualidad se utiliza la minería de datos como herramienta para extraer conocimiento de grandes cantidades de datos relacionados con el medio ambiente y, de esta manera, obtener predicciones meteorológicas (Gore & Deshpande, 2017). Hasta el momento se han realizado investigaciones en el área de la contaminación del aire en todo el planeta. Uno de estos trabajos es el realizado en Teherán, que es una de las ciudades más contaminadas del mundo. En ella se analizó tres principales contaminantes como son: material particulado (PM10), óxido de nitrógeno (NO), y dióxido de azufre (SO2), presentando niveles muy por encima de lo normal, indicando de esta manera, la mala calidad del aire dentro de esta ciudad (Aghdam & Hosseini, 2015).

Se ha demostrado que los factores meteorológicos desfavorables como la presión atmosférica, pueden agravar en gran medida los impactos causados por ciertos contaminantes del aire, como es el caso del monóxido de carbono (CO) y el ozono (O₃). Es decir, en presencia de altos índices de presión atmosférica aumentan su impacto contra la calidad del aire. Por lo tanto, en las últimas décadas, la calidad del aire en las áreas urbanas ha sido analizada en combinación de varios factores (Esfandani & Nematzadeh, 2016). Así, en un estudio realizado por (Banerjee, Singh, & Srivastava, 2011), respecto a la calidad del aire se analizaron contaminantes como: dióxido de azufre (SO₂), dióxido de nitrógeno (NO₂) y material particulado (PM) en la ciudad de Pantnagar – India. Conjuntamente se analizaron datos meteorológicos como temperatura del aire y precipitación, dando como resultado que el material particulado PM₅ es el mayor contaminante presente en concentraciones anuales, alterando la calidad de vida de las personas ya que provoca afecciones a los órganos respiratorios y en casos más graves cáncer de pulmón. De igual manera, en la investigación realizada por (Elminir, 2005), los resultados mostraron que los contaminantes relacionados con el tráfico se encontraban en la concentración más alta cuando el nivel de velocidad de viento era bajo mientras que, a velocidades altas el viento arrastraba el material particulado contribuyendo a la calidad positiva del aire.

Dentro de la ciudad de Cuenca, se ha realizado una investigación a cargo del Departamento de Investigaciones (IERSE), en colaboración del Departamento de Investigaciones (LIDI), en el cual como primera fase se analizó el comportamiento de los contaminantes atmosféricos, encontrando como resultado que existen patrones entre los mismos (Ortega Guamán, 2018). Sin embargo, no se ha realizado un estudio que analice la relación de los contaminantes con datos de tipo meteorológico, mediante la aplicación de minería de datos. Por ello, el objetivo de éste trabajo es analizar la relación entre contaminantes atmosféricos y variables meteorológicas con la finalidad de identificar patrones de comportamiento entre estos dos tipos de variables y, de esta manera, generar conocimiento en beneficio de la sociedad y mejorar la calidad de vida de las personas.

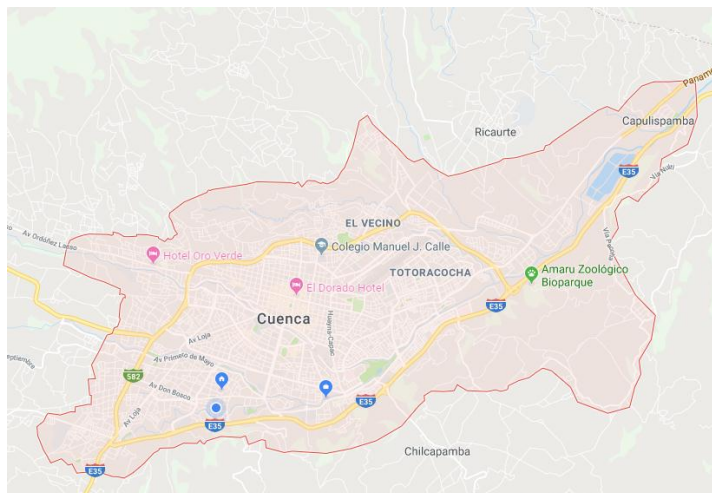
2. MÉTODO

2.1 Zona de Estudio

El trabajo se realizó en la ciudad de Cuenca – Ecuador (Figura 1), la cual se encuentra ubicada en las coordenadas $2^{\circ} 53' 51''\text{S}$ y $79^{\circ} 00' 16''\text{O}$, cuenta con un clima templado que se caracteriza por abundantes precipitaciones y climas regularmente fríos. Cuenca se encuentra a 2,560 sobre el nivel del mar perteneciendo a la sierra ecuatoriana.

Figura 1

Mapa de la ciudad Cuenca – Ecuador



2.2 Datos utilizados

Los datos utilizados para realizar éste trabajo fueron recolectados por una estación automática de monitoreo del aire a cargo de la empresa EMOV EP (Red de Monitoreo de Calidad del Aire de Cuenca de la EMOV EP, 2015), que mide y almacena los datos referentes a la calidad del aire dentro de la ciudad de Cuenca – Ecuador. Los datos utilizados en el trabajo corresponden al periodo de tiempo comprendido desde el día 16 de septiembre del año 2017 hasta el día 14 de octubre del año 2017, y corresponden a las mediciones realizadas cada 10 minutos. Las variables que se utilizaron para el desarrollo del trabajo son:

Contaminantes Atmosféricos

- Ozono (O₃).
- Monóxido de Carbono (CO).
- Dióxido de nitrógeno (NO₂).
- Dióxido de Azufre (SO₂).
- Material Particulado (PM_{2.5}).

Variables Meteorológicas

- Temperatura del Aire.
- Humedad Relativa.
- Punto del rocío.
- Presión Atmosférica.
- Radiación Global.
- Precipitación.
- Velocidad del viento.
- Radiación ultravioleta A.
- Radiación ultravioleta E.

El set de datos (*dataset*) original incluye los valores máximo, promedio y mínimo por cada variable meteorológica. Dentro del trabajo realizado se utilizó los valores promedios ya que los valores máximos y mínimos no presentan una diferencia considerable con el valor promedio, además se consideró descartar estos valores ya que pueden representar situaciones no cotidianas dentro del set de datos. Se estableció también descartar del set de datos el valor perteneciente a la dirección del viento, ya que éste valor se utiliza para discretizar los datos relacionados con el aire (Cagliero et al., 2016), tarea que no forma parte del objetivo de este trabajo.

2.3 Software Utilizado

2.3.1 Rapidminer

Rapidminer es una herramienta informática que se utiliza para el análisis y minería de datos mediante un entorno gráfico. Es utilizado principalmente en ámbitos de investigación, educación y en aplicaciones empresariales. La herramienta cuenta con más de 450 operadores para el tratamiento de datos, procesamiento de los mismos y visualización (Rapidminer, 2018). Rapidminer puede ser ejecutada en cualquier plataforma y es una herramienta gratuita, además de ser de fácil uso y amigable con el usuario.

Dentro del trabajo realizado Rapidminer se utilizó como herramienta base para el análisis de contaminantes y variables meteorológicas, además se construyó el modelo de clustering y sobre éste modelo se probaron los algoritmos de minería de datos, posteriormente dentro de la herramienta se generó una matriz de correlación que indicó que variables se correlacionan entre sí y su comportamiento.

2.3.2 Software Estadístico R

R es una herramienta enfocada al análisis estadístico de datos. Es también un lenguaje de programación y una herramienta muy poderosa que cuenta con una gran variedad de técnicas estadísticas, muy útil al momento de realizar gráficas, operaciones matemáticas y cálculos en general (The R Project for Statistical Computing, 2018). La herramienta R se utilizó dentro del trabajo para realizar las funciones y gráficas que ayudaron a analizar el comportamiento de los contaminantes en relación a las variables meteorológicas y su correlación en un determinado tiempo.

2.4 Metodología

2.4.1 CRISP-DM

Para el desarrollo del trabajo se utilizó como base de referencia la metodología CRISP – DM (Figura 2), la cual es una de las más utilizadas a nivel general para la realización de proyectos de minería de datos. Esta cuenta con seis

etapas las cuales son flexibles y ajustables a cada proyecto. La metodología sirve de guía a través del proceso de realización de un proyecto de minería de datos (Gallardo, 2017). En la realización del trabajo la metodología indica el proceso y cada etapa que se debe seguir para cumplir de forma correcta un proyecto de minería de datos.

Figura 2

Modelo de proceso CRISP-DM

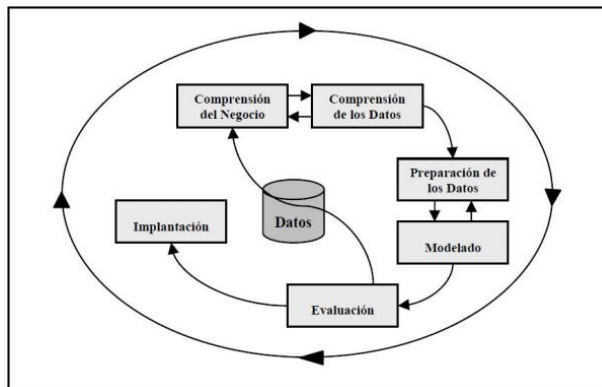


Imagen que describe las etapas de la metodología CRISP-DM para proyectos de minería de datos.

La metodología CRISP-DM cuenta con las siguientes etapas: comprensión del negocio o problema, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación.

2.4.2 Comprensión del negocio o problema

La fase de comprensión del problema es una de las más importantes a lo largo de la realización del proyecto ya que en base a ésta se convertirán los objetivos del problema en objetivos técnicos de la investigación (Gallardo, 2017). La comprensión del problema o del caso de estudio sirve para elegir correctamente las herramientas necesarias y obtener resultados de calidad.

Para realizar éste trabajo es importante conocer que dentro del medio local no existe un estudio que analice el comportamiento entre contaminantes del aire y los factores meteorológicos.

2.4.3 Comprensión de los datos

Esta etapa inicia con la recolección de los datos con la finalidad de un mejor entendimiento de los mismos, además familiarizarse con los datos y establecer posibles relaciones entre ellos (Gallardo, 2017).

Para Gallardo (2017) esta etapa cuenta con las siguientes fases:

- Recolección de datos.
- Descripción de los datos.
- Exploración de los datos.
- Verificación de calidad de los datos.

Como se mencionó anteriormente los datos son recolectados por una estación de monitoreo del aire, la cual los recolecta cada minuto. Para la realización del trabajo se consideró variables meteorológicas y los datos utilizados en un trabajo anterior sobre los contaminantes del aire.

Los datos meteorológicos utilizados se muestran en la tabla 1, y los contaminantes utilizados se presentan en la tabla 2. La información originalmente es almacenada cada segundo y posteriormente se promedia en frecuencias por minuto dentro de la base de datos perteneciente a la empresa EMOV-EP (Red de Monitoreo de Calidad del Aire de Cuenca de la EMOV EP, 2015).

2.4.4 Preparación de los datos

La preparación de los datos es la etapa más importante del proyecto, ya que somete a un proceso de selección de los datos más relevantes para el estudio. En esta fase se depuran los datos para que sean óptimos en un estudio de minería de datos. Por lo general, la etapa de preparación de los datos conlleva el 80% o 90% del trabajo realizado (Perdomo, 2007).

Perdomo (2007) afirma que la etapa de preparación de los datos consta de las siguientes fases:

- Selección de datos.
- Limpieza de datos.

- Estructura de datos.
- Integración de los datos.

2.4.5 Modelado de los datos

En ésta fase de la metodología CRISP-DM se seleccionan las técnicas adecuadas para una correcta aplicación de minería de datos. Ésta etapa se realiza luego de la etapa de preparación de los datos, ya que luego de una correcta depuración de los mismos es importante definir y construir el modelo que se utilizara para el análisis (Gallardo, 2017).

En éste trabajo, se utilizó agrupamiento de datos (*Clustering*) ya que se necesita asociar los valores de los contaminantes con los factores meteorológicos; por lo que la aplicación de algoritmos de agrupamiento resulta efectiva en el desarrollo de éste trabajo.

2.4.6 Evaluación

En ésta etapa de la metodología CRISP-DM son evaluados los algoritmos establecidos en base al modelo generado (Gallardo, 2017).

En ésta sección se aplican los algoritmos de agrupamiento previamente definidos sobre el modelo definido y construido para obtener resultados de dichas pruebas.

2.4.7 Implementación

En esta fase, cuando el modelo haya sido definido, construido y validado, la información tratada mediante este último, se convierte en conocimiento. Por lo general, la etapa de implementación no es el fin del proyecto ya que el conocimiento debe ser presentado de forma entendible al cliente para lograr una comprensión total del conocimiento generado (Gallardo, 2017).

2.5 Desarrollo

2.5.1 Comprensión de los datos

2.5.1.1 Variables meteorológicas

El set de datos proporcionado por el Instituto de Estudios de Régimen Seccional del Ecuador (IERSE) cuenta con un total de 3.936 datos recolectados cada 10 minutos, registros recolectados durante 27 días, el set de datos contiene los valores de máximo, promedio y mínimo por cada variable (Tabla1), de los cuales se utilizarán los valores promedios.

Tabla 1

Variables Meteorológicas utilizadas para la realización del trabajo

Variable meteorológica	Abreviatura	Unidad
Temperatura del aire	TEMPAIRE_AV	°C (Grados Centígrados)
Humedad relativa	HR_AV	% (Valor porcentual de agua)
Punto del rocío	DP_AV	°C (Grados Centígrados)
Presión Atmosférica	PATM_AV	hPa (Hectopascal)
Radiación Global	RADGLOBAL_AV	w/m2 (Vatios por metro cuadrado)
Precipitación	PRECIP_SUM	mm (Milímetros de agua)
Velocidad del viento	WINDSPEED_AV	m/s (metros por segundo)
Radiación UVA	UVA_AV	w/m2 (Vatios por metro cuadrado)
Radiación UVE	UVE_AV	w/m2 (Vatios por metro cuadrado)

Temperatura del aire

La temperatura es una magnitud relacionada con la rapidez de movimiento de las partículas, que constituyen la materia. Mientras mayor es la agitación mayor es la temperatura. Es una de las magnitudes más utilizadas para indicar el estado de la atmósfera, la temperatura varía entre el día o la noche y de estación a otra (Rodríguez Jimenez, Capa, & Portela Lozano, 2004).

Humedad relativa

La humedad relativa es la cantidad de vapor de agua presente en el aire, dicha cantidad no es constante en el medio ambiente y depende si recientemente ha llovido o si está cerca del mar. La humedad del aire se mide en tanto por ciento (%) e indica el grado de saturación del aire (Rodríguez Jimenez et al., 2004). La variación de la humedad con respecto a la altura se aprecia de mejor forma en las zonas ecuatoriales, tomando en cuenta que el aire que se encuentra a mayor altura

presentara poca humedad, lo que se considera un aire seco, sin embargo, el aire a menor altura presenta una mayor saturación (Basantes & García, 2018).

Punto del rocío

El punto de rocío es la medida de la cantidad de vapor de agua presente en un gas; es la más alta temperatura a la cual se condensa el vapor de agua produciendo neblina o cualquier tipo de nube y en temperaturas más bajas escarcha (Vaisala, 2013).

Presión Atmosférica

La presión atmosférica es la presión que ejerce el aire sobre la superficie, depende de muchas variables, pero sobre todo la altitud, mientras una superficie se encuentre a mayor altitud en la atmósfera la cantidad de aire será menor, caso contrario mientras una superficie se encuentre por debajo de la atmósfera existirá una mayor presencia de aire (Rodríguez Jimenez et al., 2004).

Radiación global

La radiación global es la energía que se transfiere del sol a la tierra, ésta se define como la radiación recibida en un ángulo de 2π , la radiación global incluye la que se recibe directamente del sol y la radiación celeste que ocurre cuando la misma atraviesa la atmósfera (Rodríguez Jimenez et al., 2004).

Precipitación

La precipitación es la cantidad de agua presente en la superficie terrestre proveniente de la humedad atmosférica, se presenta en estado líquido como la lluvia, o en estado sólido como la escarcha o granizo (Piura, 2006).

Velocidad del viento

La velocidad del viento produce energía porque siempre está en movimiento, cerca de la superficie terrestre, se encuentra una menor velocidad del viento, la misma que aumenta mientras la altura es mayor, mientras la superficie sea más

accidentada como valles, etc., menor será la velocidad del viento (Rodríguez Jimenez et al., 2004).

Radiación UVA - UVE

El sol emite energía constantemente en una alta gama de ondas hacia la tierra, la longitud de onda de la radiación es más corta que la luz visible normal o la violeta, ésta produce quemaduras y más efectos adversos a la salud de las personas. La capa de ozono es la encargada de filtrar la mayor cantidad de radiación ultravioleta, es caso especial que la radiación UVA o UVE no son absorbidas por la capa de ozono (Aire, 2001).

2.5.1.2 Contaminantes atmosféricos

Los datos de contaminantes atmosféricos proporcionados son los mismos que se utilizaron en una etapa anterior del trabajo, en la cual solo se analizaban patrones entre los diferentes contaminantes (Tabla 2). Los datos proporcionados pertenecen al mismo rango de tiempo y recolectados por el mismo sensor, los datos están almacenados por cada minuto, por lo que se necesita realizar una unión con cada dato perteneciente a los factores meteorológicos en la misma hora y fecha.

Tabla 2

Contaminantes Atmosféricos

Contaminante atmosférico	Abreviatura	Unidad
Ozono	O3	Partes por Billón
Monóxido de Carbono	CO	Partes por Billón
Dióxido de Nitrógeno	NO2	Partes por Billón
Dióxido de Azufre	SO2	Partes por Billón
Material Particulado	PM2.5	Microgramos por metro cubico

Contaminantes atmosféricos utilizados para la realización del trabajo.

Ozono (O3)

El ozono es un gas compuesto de tres átomos de oxígeno, el aire que respiramos consta únicamente de dos átomos, cuando el gas se encuentra presente en la atmósfera sirve de filtro contra los rayos ultra violetas, sin embargo cuando se

encuentra presente en la superficie terrestre causa afecciones a las personas como: enfermedades respiratorias crónicas y la incapacidad del sistema inmunológico para defender al sistema respiratorio (PNUMA, 2013).

Monóxido de Carbono (CO)

El monóxido de carbono (CO) es un gas incoloro, se produce cuando el combustible de carbón no se quema correctamente y es producido por personas o acciones de la naturaleza, el principal causante de monóxido de carbono (CO) es el tubo de escapes de los carros (Atsdr & Sciences, n.d.).

Dióxido de nitrógeno (NO₂)

El dióxido de nitrógeno (NO₂), es emitido principalmente por la combustión de motores de automóviles, forma parte de un grupo de contaminantes gaseosos que contribuyen a la formación y modificación de otros contaminantes como el ozono (O₃) y puede colaborar en la formación de lluvia ácida (Nitrógeno, 2004).

Dióxido de Azufre (SO₂)

Es un contaminante que proviene principalmente de la combustión de combustibles fósiles como: madera, derivados del petróleo, etc., puede ocasionar daño respiratorio y cardiovascular agudo, además de causar afecciones graves a personas que padecen asma (Del, En, & Salud, 2010).

Material Particulado (PM_{2.5})

Es el material respirable presente en la atmósfera del planeta, debido a su tamaño se lo puede dividir en dos grupos pm₁₀ y pm_{2.5} que corresponden a la milésima parte de un milímetro, a nivel de la superficie terrestre puede ocasionar: daños al sistema respiratorio e incrementar el desarrollo de alergias (Linares & Díaz, 2008).

2.5.2 Preparación de los datos

Debido a que la etapa de preparación de los datos es la más importante del trabajo se la dividió en cuatro fases (Figura 3). Las etapas se llevaron a cabo con la finalidad de obtener un set de datos uniforme para el análisis de los contaminantes y variables meteorológicas. Dentro de la primera etapa de preparación de los datos está el análisis de los sets de datos, en la cual, identificamos los valores pertenecientes a cada variable dentro del *dataset*, en ésta etapa se identificó como estaban estructurados los datos. El análisis presentó que el set de datos de contaminantes atmosféricos almacenaba valores cada diez minutos para cada registro, sin embargo, el set de datos de variables meteorológicas presentó un almacenamiento cada minuto para cada registro, lo que presentó una dificultad para realizar la unión de estos dos set de datos; en una segunda etapa se realizó el tratamiento de la información, en la cual, se procedió a calcular el valor de la media entre un rango de 10 minutos del set de datos de contaminantes para que coincidiera con el set de datos de variables meteorológicas (Tabla 3), dando como resultado que por cada registro meteorológico en determinada hora, exista un valor de la media calculada en la misma hora y fecha, en este caso se puede observar en la gráfica que el campo FECHA consta de diez registros, los cuales dan como resultado un valor promedio que coincide con el campo FECHOR del set de datos de variables meteorológicas. Luego de éste proceso se realizó el *match* entre los set de datos; se tomó en cuenta los campos de fecha como llave principal de los mismos, de ésta manera, en base a hora y fecha de registro se realizó la unión de los *dataset* mediante la utilización del software PLSQL (Figura 19).

Figura 3

Etapas de preparación de los datos

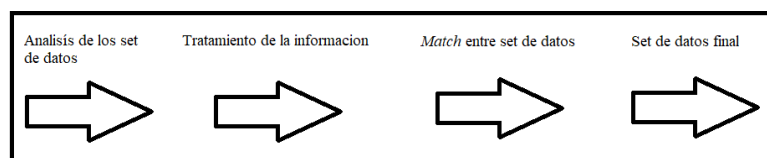


Tabla 3

Proceso de unión de set de datos, contaminantes atmosféricos (Izquierda) y variables meteorológicas (Derecha).

FECHA	O3	CO	NO2	SO2	PM2_5	fechor	tempaire_av	tempaire_max	tempaire_min	
16/9/2017 0:01	20.4153056	0.789981	23.4809132	3.01132514	7.2	Promedio 0:01:00 - 0:10:00	16/9/2017 0:10	14.3	14.4	14.3
16/9/2017 0:02	23.0873838	0.789981	23.445185	3.00346951	9.7					
16/9/2017 0:03	24.7549804	0.789981	22.800198	2.99299533	11.7					
16/9/2017 0:04	23.6269004	0.80143	22.800198	2.74161515	11.7					
16/9/2017 0:05	22.1201778	0.80143	22.5181337	2.89872777	11.7					
16/9/2017 0:06	24.53525	0.767083	22.1176024	3.12654106	11.7					
16/9/2017 0:07	27.8331674	0.732736	21.0363559	3.26270533	11.7					
16/9/2017 0:08	25.6025118	0.68694	20.4496621	3.42243649	11.7					
16/9/2017 0:09	22.6341899	0.698389	20.8595956	3.60311599	11.7					
16/9/2017 0:10	21.4825674	0.709838	21.177388	3.49313716	11.7					

De esta manera, se obtuvo como resultado un set de datos uniforme (figura 4) que coincide tanto en horas como en fechas, lo cual se puede verificar en la similitud de los campos HORADEC y FECHOR.

Figura 4

Set de datos final que contiene contaminantes atmosféricos y variables meteorológicas de acuerdo a hora y fecha.

HORADEC	O3_AVG	CO_AVG	NO2_AVG	SO2_AVG	PM2_5_AVG	FECHOR	TEMPAIRE_...	HR_AV	DP_AV	PATM_AV
16/9/2017 0:10	41.768	0.554	13.083	5.858	11.660	16/9/2017 0:10	14.300	65	7.900	752.990
16/9/2017 0:20	48.411	0.476	3.172	10.395	11.700	16/9/2017 0:20	14.300	65	7.900	752.840
16/9/2017 0:30	44.794	0.507	3.409	4.877	11.700	16/9/2017 0:30	14.300	65	7.900	752.656
16/9/2017 0:40	40.198	0.575	5.841	3.600	11.690	16/9/2017 0:40	14.400	65	7.900	752.563
16/9/2017 0:50	19.225	0.742	21.397	3.542	11.700	16/9/2017 0:50	14.400	65	7.900	752.541
16/9/2017 1:00	25.266	0.690	20.682	3.107	16.220	16/9/2017 1:00	14.100	67	8.100	752.473
16/9/2017 1:10	31.834	0.592	12.219	2.572	17.590	16/9/2017 1:10	14	68	8.200	752.433

Luego de reestructurar el set de datos se procede a eliminar del mismo los valores que sean nulos, ya que algunos algoritmos de minería de datos no permiten su ejecución si el *dataset* contiene este tipo de datos en alguno de los atributos, el número total de elementos descartados por ser nulos es 212, lo cual representa una pérdida del 5.81% de datos del set de datos original.

2.5.3 Modelado de datos

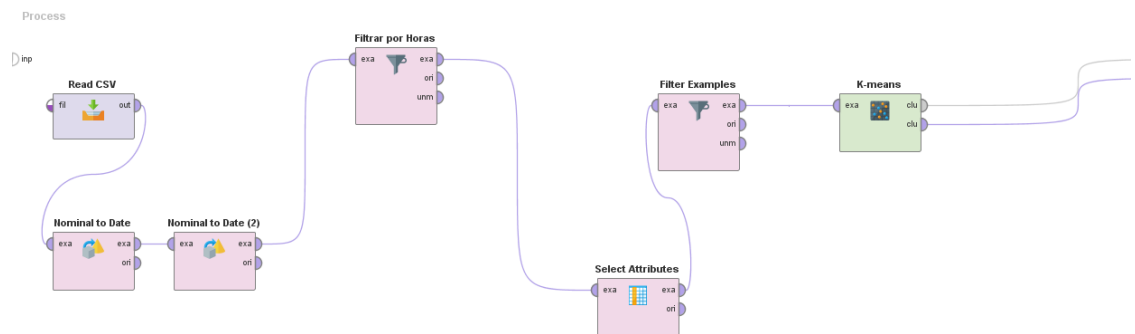
Para el modelado de datos se utilizó la herramienta Rapidminer, donde se definió el tipo de modelado y se realizó la construcción del mismo. Para éste trabajo se realizó el análisis utilizado clustering; que es una técnica de minería de datos que tiene por objetivo encontrar agrupamientos de tal forma que los objetos de un grupo sean similares entre sí y diferentes de otro grupo (Berzal, 2012).

Mediante el software Rapidminer se construyó un modelo de clustering utilizando los siguientes componentes (Figura 5):

- Read CSV
 - Lee una fuente externa al software en formato CSV.
- Nominal to Date
 - Convierte un valor nominal a un valor de fecha.
- Filter examples
 - Establece un filtro a la información.
- Componente de algoritmo
 - Componente utilizado para tratar la información mediante un algoritmo de minería de datos.

Figura 5

Modelo de Clustering el cual se utiliza para analizar la información del dataset.



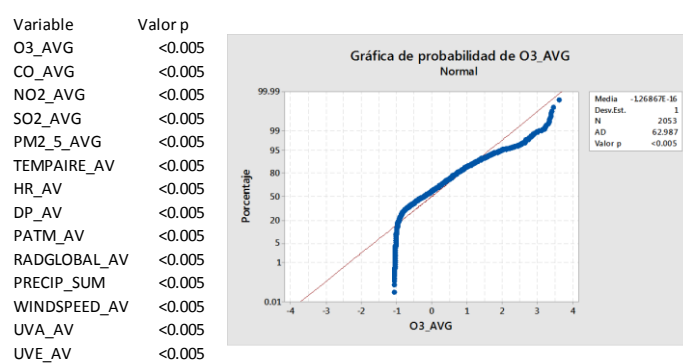
Se implementaron cinco algoritmos de minería de datos: *K-means*, *Xmeans*, *DBSCAN*, *CobWeb* y *Expectation Maximization Clustering (EMC)*. Los algoritmos son no supervisados lo que implica que no necesitan de una variable de clasificación y pueden ser utilizados en grandes cantidades de datos (Moreno García & López Batista, n.d.).

2.5.4 Evaluación

Para realizar una correcta evaluación del modelo se procedió a realizar una prueba de normalidad de datos. Según (Rojas Dávila, 2003), la prueba de normalidad se realiza con la finalidad de determinar si las muestras de datos siguen o no una distribución normal, la prueba genera una gráfica y una prueba de hipótesis, por lo que, en el set de datos se analizó la distribución de los elementos del mismo. Según las pruebas de Anderson – Darling ninguno de los elementos del set de datos tiene distribución normal (figura 6), por lo que en base a éste resultado se utiliza la técnica de correlación de Spearman.

Figura 6

Prueba de normalidad de datos realizada a cada elemento del set de datos.



Luego se realizó la evaluación de los diferentes algoritmos de minería de datos antes mencionados, dando como resultado (Tabla 4), que el algoritmo DBSCAN es el que mayor agrupación de datos presenta, lo que significa que el algoritmo agrupa un 67.59% de los datos, en su cluster más grande, señalando que este es el mejor algoritmo para la realización del trabajo.

Tabla 4

Prueba de algoritmos de minería de datos

Algoritmo	Número de datos en cluster mas alto	Porcentaje	Tiempo de ejecución
K-means	2319	63.59%	321ms
X-means	2265	62.11%	406ms
DBSCAN	2465	67.59%	8181ms
EMC	2415	66.21%	20993ms
COBWEB	2090	57.31%	1222ms

Prueba realizada al set de datos con cada algoritmo.

Para validar la elección del algoritmo adecuado se realizó una comprobación en cuanto al set de datos y se utilizó el cluster más grande o con mayor cantidad de datos de cada algoritmo, ya que dentro del cluster más grande se encuentra la mayor cantidad de datos con relación entre sí; además de identificar en cuántos de estos se posiciona cada dato después de realizar las pruebas de *clustering* (Figura 7). Los datos que se encuentren dentro del cluster más grande se identificaron con un valor de uno, caso contrario será cero. Después se realizó una operación en la cual se suman los valores de cada dato dando como resultado un valor de cero a cuatro; si el valor es tres o cuatro significa que el dato es relevante en el análisis (figura 8), al realizar esta operación se verifica mediante la suma que el dato se encuentre posicionado en tres de los cuatro, o cuatro de los cuatro clusters de cada algoritmo, garantizando que el dato sea relevante para el análisis. Éste proceso demostró que el mejor algoritmo para evaluación del set de datos es DBSCAN el cual presenta mayor cantidad de datos para el análisis y mayor precisión; durante el proceso se descartó el algoritmo Expectation Maximization Clustering (EMC), ya que por su funcionamiento no permite ubicar cada dato dentro de un cluster, sino que éste indica una probabilidad de que un dato se posicione dentro de un cluster del algoritmo lo cual no es significativo a la investigación.

Figura 7

Clasificación de cada dato dentro del cluster más grande de cada algoritmo, si el dato se encuentra en el cluster más grande se identifica con un valor de 1 caso contrario 0.

id Dato	Cluster Kmeans	Cluster X means	Cluster cobweb	Cluster dbscan	K-means	X-means	COBWEB	DBSCAN
1	cluster_0	cluster_1	cluster1	cluster_2	1	1	1	0
2	cluster_0	cluster_1	cluster1	cluster_3	1	1	1	0
3	cluster_0	cluster_1	cluster1	cluster_0	1	1	1	1
4	cluster_0	cluster_1	cluster1	cluster_1	1	1	1	0
5	cluster_0	cluster_1	cluster1	cluster_2	1	1	1	0
6	cluster_0	cluster_1	cluster1	cluster_0	1	1	1	1
7	cluster_0	cluster_1	cluster1	cluster_1	1	1	1	0
8	cluster_0	cluster_1	cluster1	cluster_0	1	1	1	1

Figura 8

Suma de acuerdo al número de coincidencias de cada dato en los clusters más grandes de cada algoritmo, siendo relevantes para la investigación aquellos que tengan un valor de 3 o 4.

id Dato	Cluster Kmeans	Cluster X means	Cluster cobweb	Cluster dbscan	K-means	X-means	COBWEB	DBSCAN	SUMA
1	cluster_0	cluster_1	cluster1	cluster_2	1	1	1	0	3
2	cluster_0	cluster_1	cluster1	cluster_3	1	1	1	0	3
3	cluster_0	cluster_1	cluster1	cluster_0	1	1	1	1	4
4	cluster_0	cluster_1	cluster1	cluster_1	1	1	1	0	3
5	cluster_0	cluster_1	cluster1	cluster_2	1	1	1	0	3
6	cluster_0	cluster_1	cluster1	cluster_0	1	1	1	1	4
7	cluster_0	cluster_1	cluster1	cluster_1	1	1	1	0	3
8	cluster_0	cluster_1	cluster1	cluster_0	1	1	1	1	4

Además de este proceso, dentro del set de datos se seleccionaron aquellos datos que presentaron una suma de tres o cuatro, ya que éste número indica la relevancia de los datos para el análisis, dando como resultado un nuevo set de datos con 2,265 registros que se utilizarán para generar la matriz de correlación en la cual se identificará el comportamiento entre contaminantes y variables meteorológicas.

Mediante la incorporación de un nuevo elemento dentro del modelo definido en la herramienta Rapidminer y en base al nuevo set de datos, se obtiene una matriz de correlación (figura 9), en la cual se analiza los valores entre los dos tipos de variables.

Figura 9

Matriz de correlación en base al set de datos obtenido

Attribut...	TEMPAIRE_AV	HR_AV	DP_AV	PATM_AV	RADGLOBAL_AV	PRECIP_SUM	WINDSPEED_AV	UVA_AV	UVE_AV
O3_AVG	0.734	-0.822	-0.664	-0.041	0.592	-0.039	0.564	0.592	0.491
CO_AVG	-0.403	0.265	0.007	0.068	-0.244	-0.043	-0.289	-0.279	0.014
NO2_AVG	-0.496	0.334	0.025	0.131	-0.268	0.010	-0.422	-0.303	-0.251
SO2_AVG	-0.420	0.283	0.037	0.150	-0.080	-0.060	-0.456	-0.100	-0.086
PM2_5_...	-0.185	-0.097	-0.367	0.175	0.241	-0.064	-0.183	0.215	0.169

Matriz de correlación obtenida en base al set de datos depurado.

Luego de generar la matriz de correlación se evaluó el comportamiento de los contaminantes en relación a las variables meteorológicas en base a la matriz, por lo que se utilizó un Script desarrollado en R por el departamento de investigaciones LIDI, en el cual se graficó el comportamiento de los dos grupos de variables dentro de un rango de tiempo, además se estableció que éste rango de tiempo sea desde las 6:00 AM hasta las 6:00 PM en formato de 12 horas con una ventana de tiempo dinámica de 15 minutos, ya que dentro de éste rango de tiempo se presenta una mayor cantidad de datos para el análisis.

Por motivos de representación se procedió previamente a normalizar el set de datos, tomando en cuenta una validación previa en la cual se normalizó todo el *dataset* y luego se normalizaron variables en pares, dando como resultado que el proceso no afecta a los valores de correlación; por lo que, se normalizó los datos para una mejor representación del comportamiento de los dos grupos de variables.

2.5.5 Implementación

Los resultados que se encuentran dentro de esta sección serán especificados en el área de resultados del documento, debido a que la metodología utilizada propone un área de resultados.

3. RESULTADOS

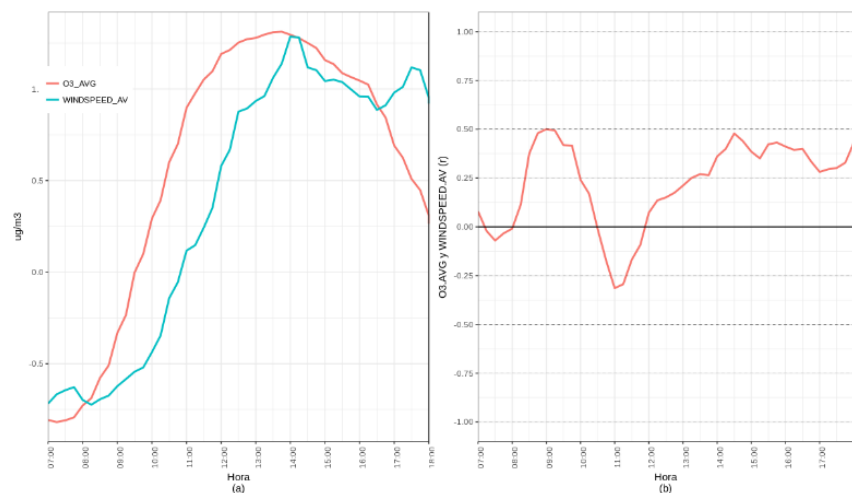
Luego de realizar la etapa de evaluación, se analiza los resultados de los diferentes valores de correlación, presentándolos en forma de gráficas que mejoran el entendimiento de los mismos. Éstas gráficas presentan singularidades entre los contaminantes y las variables meteorológicas como es el caso del ozono (O3) y la temperatura del aire, que presentan una correlación de 0.7 lo cual indica que este tiene un comportamiento proporcional a la temperatura.

En este caso, el análisis se enfoca en casos excepcionales en los cuales la correlación presenta un valor muy bajo, pero de acuerdo a las gráficas obtenidas a lo largo del día se muestra un comportamiento inusual, como en los siguientes casos:

En la figura 10, se aprecia que entre el Ozono (O3) y la velocidad del viento existe una correlación superior a 0.5, lo cual indica que éste es proporcional a la velocidad del viento; sin embargo, después de las 10:00 AM se puede observar que existe una correlación negativa entre las dos variables, sin embargo, a partir de las 12:00 PM la correlación se vuelve positiva, lo que indica que a lo largo del día existe una variación en la correlación no respaldada por el valor inicial de 0.5.

Figura 10

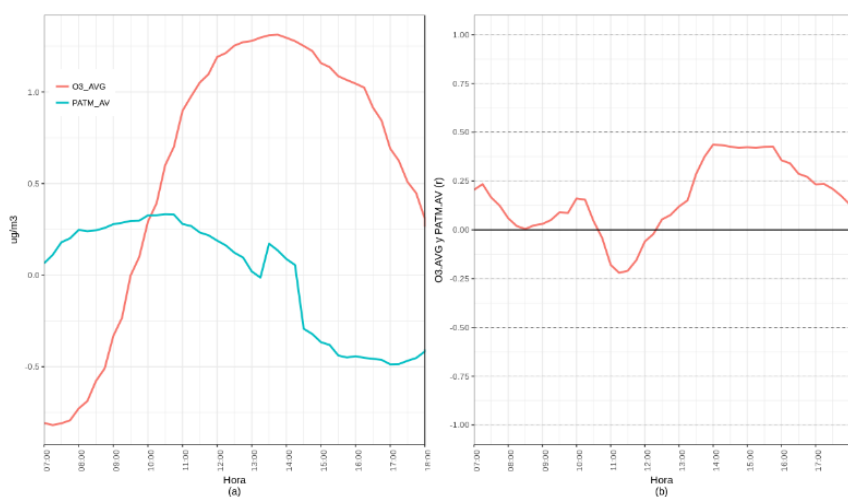
Comportamiento entre el Ozono(O3) y la Velocidad del viento



Se puede observar en la figura 11, que existe un valor de -0.041 de correlación entre el Ozono (O3) y la presión atmosférica de acuerdo a la matriz de correlación; sin embargo, graficando el comportamiento entre las dos variables y la correlación a lo largo del día se observa que la correlación pasa de ser inversamente proporcional en un principio a ser positiva a partir de las 12:00 PM y llega a un valor significativo a las 14:00 PM, acercándose a un valor de 0.5 lo cual indica que el valor inicial de la correlación no representa a la correlación a lo largo del día.

Figura 11

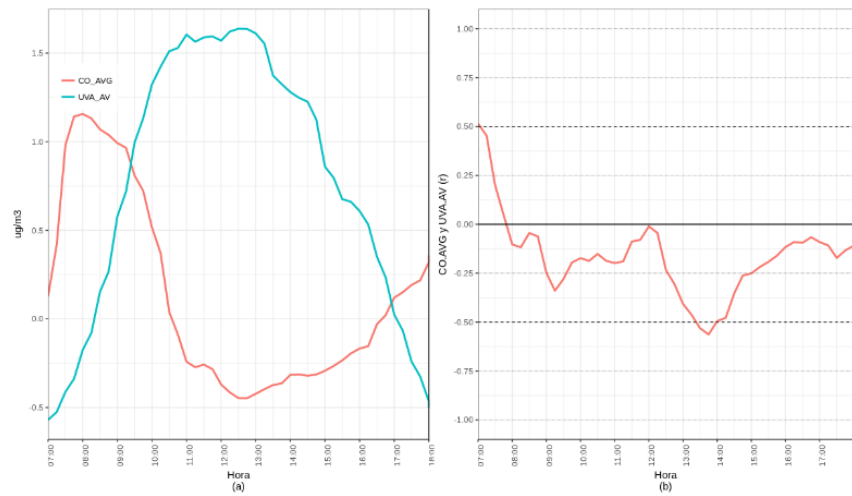
Comportamiento entre el Ozono(O3) y la Presión Atmosférica



En la figura 12, se observa que existe una relación inversamente proporcional entre el monóxido de carbono (CO) y la radiación UVA, con una correlación de -0.2 en general; sin embargo, en la gráfica se observa que a las 7:00 AM parte de una correlación positiva, descendiendo a un valor negativo hasta alcanzar un valor de correlación superior a -0.5, lo que indica que el valor de correlación obtenido en la matriz, no representa el comportamiento a lo largo del día de la correlación.

Figura 12

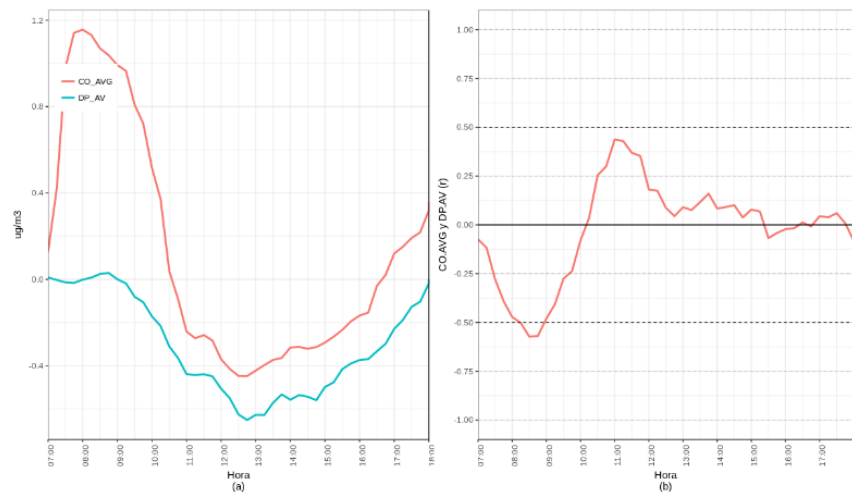
Comportamiento entre el Monóxido de Carbono y la Radiación UVA



Como se aprecia en figura 13, el valor de la correlación entre el monóxido de carbono y el punto de rocío es de 0.007 según la matriz obtenida; sin embargo, se observa que a lo largo del día éste valor cambia llegando a valores significativos como a las 8:30 AM donde alcanza un valor superior a -0.5 y a las 11:00 AM alcanza un valor positivo cercano a 0.5 demostrando de igual manera que la correlación obtenida en la matriz en algunos casos no corresponde al comportamiento de la misma a lo largo del día ya que en ciertas horas del día el valor de la correlación cambia de forma significativa.

Figura 13

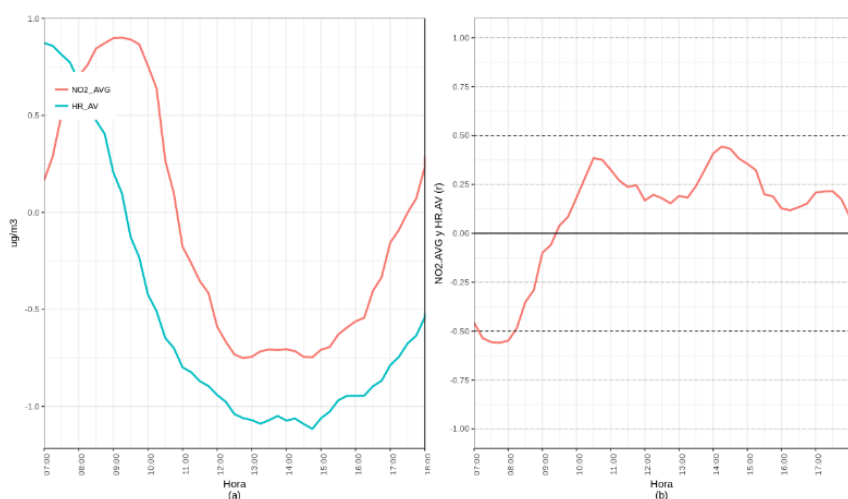
Comportamiento entre el Monóxido de Carbono y el Punto de Rocío



En la figura 14, se observa que existe una correlación positiva entre el Dióxido de Nitrógeno (NO₂) y la Humedad Relativa, de acuerdo al valor de la matriz de correlación es 0.3; sin embargo se observa que a lo largo del día éste valor cambia considerablemente partiendo a las 7:00 AM de una fuerte correlación inversa superior a -0.5 y a partir del medio día la correlación aumenta a valores muy fuertes sobrepasando el valor de 0.5; es decir, a lo largo del día mientras la humedad relativa aumenta el Dióxido de Nitrógeno (NO₂) sigue un comportamiento similar.

Figura 14

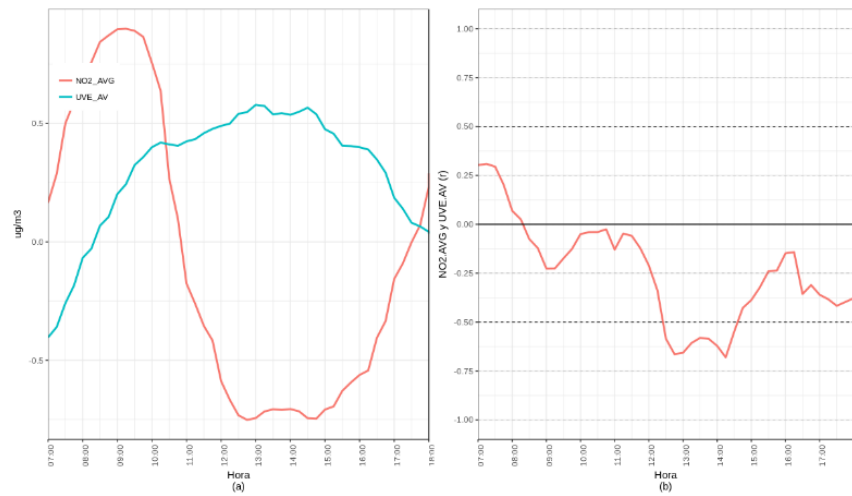
Comportamiento entre Dióxido de Nitrógeno y la Humedad Relativa



Como se observa en la figura 15, existe una correlación negativa entre el dióxido de nitrógeno (NO₂) y la radiación UVE, sin embargo, en la matriz de correlación obtenida se aprecia un valor de correlación de -0.2 para las dos variables, pero en la gráfica de comportamiento se puede apreciar que en horas de la mañana la correlación parte de ser positiva llegando a tener valores negativos muy altos en horas del medio día; llegando a valores próximos a -0.75 lo que indica que las dos variables tiene una relación inversa fuerte en ciertas horas del día.

Figura 15

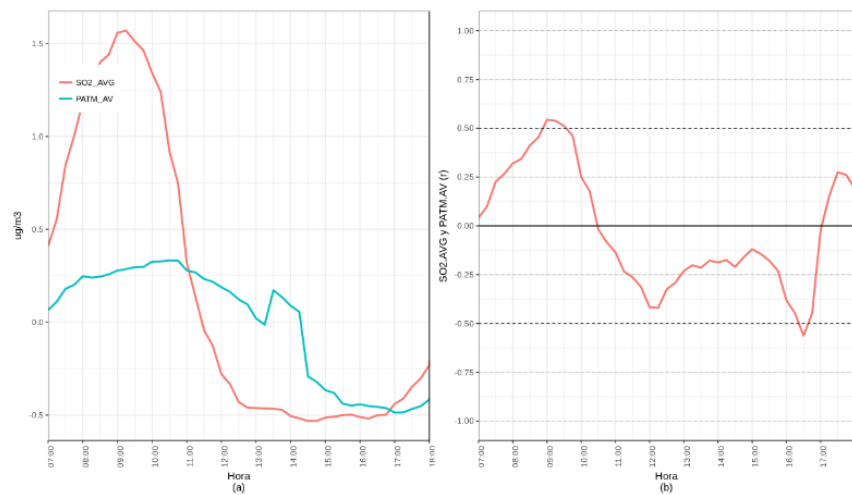
Comportamiento entre Dióxido de Nitrógeno (NO₂) y la radiación UVE



La matriz de correlación obtenida presenta que la figura 16, tiene un valor de correlación de 0.1 entre el dióxido de azufre (SO₂) y el material particulado PM_{2.5}; sin embargo, en el gráfico de comportamiento se puede identificar que en horas de la mañana parte de una fuerte correlación positiva sobrepasando valores de 0.5, y en horas del mediodía la correlación se vuelve inversamente proporcional llegando a valores negativos muy altos por la noche sobrepasando valores de -0.5, lo que demuestra una vez más que muchas de las veces el valor de correlación no representa el comportamiento de las variables a lo largo del día.

Figura 16

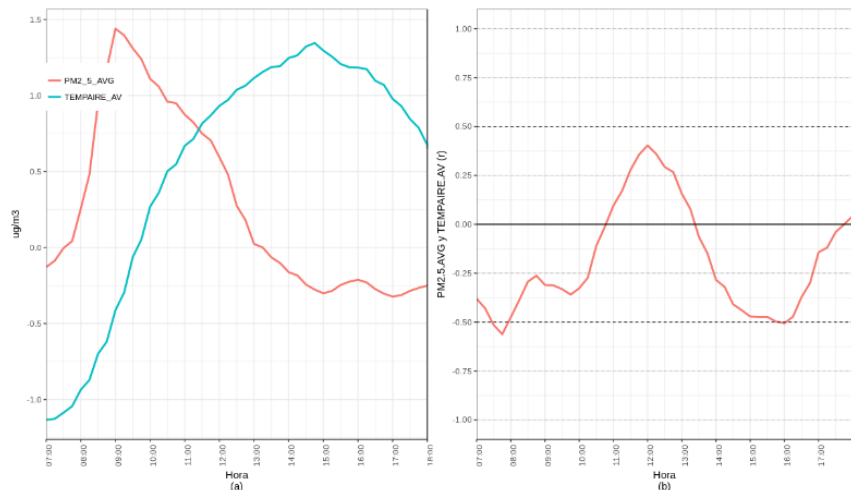
Comportamiento entre Dióxido de Azufre y material particulado PM_{2.5}



Como se aprecia en la figura 17, existe un comportamiento inversamente proporcional entre el material particulado PM2.5 y la temperatura del aire, la matriz de correlación indica un valor de -0.1; sin embargo, en el gráfico de comportamiento se aprecia que la correlación inicia en horas de la mañana con un valor superior a -0.5, y al medio día la correlación se vuelve positiva llegando a valores próximos a 0.5; en horas de la noche la correlación se vuelve inversamente proporcional con valores fuertes, lo que significa que al medio día cuando la temperatura aumenta, también aumenta la presencia de material particulado.

Figura 17

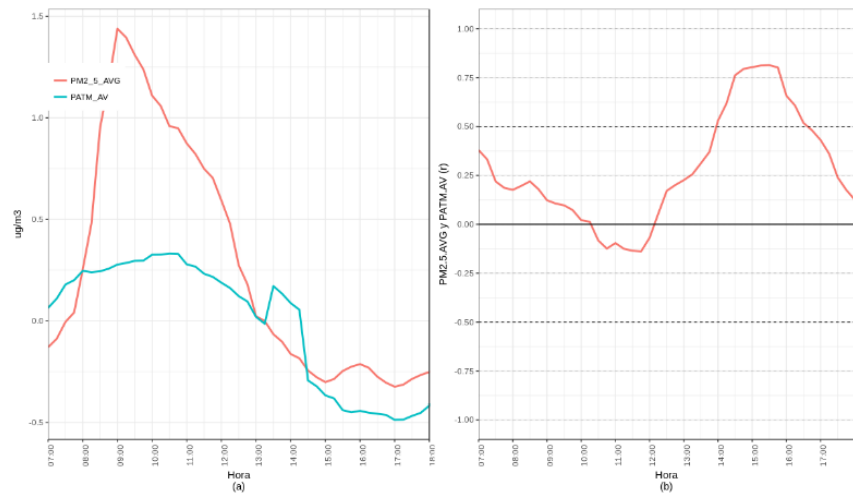
Comportamiento entre material particulado PM2.5 y Temperatura del aire



En la figura 18, se aprecia que existe una correlación positiva entre el material particulado PM2.5 y la presión atmosférica; dentro de la matriz obtenida la correlación entre estas dos variables indica un valor de 0.1; sin embargo, el gráfico de comportamiento indica que en horas de la tarde, mientras el valor de la presión atmosférica es incrementa, el valor del material particulado aumenta proporcionalmente llegando a valores superiores a 0.75 de correlación, lo que indica un fuerte comportamiento entre estas dos variables y demostrando que el valor de la matriz de correlación no respalda a los valores obtenidos a lo largo del día.

Figura 18

Comportamiento entre Material particulado PM2.5 y Presión atmosférica



4. DISCUSIÓN

El objetivo principal de éste trabajo fue la aplicación de minería de datos en el análisis de contaminantes atmosféricos y variables meteorológicas. Por ésta razón, se utilizó un set de datos proporcionado por el Instituto de Estudios de Régimen Seccional del Ecuador (IERSE). El análisis dio como resultado que el mejor algoritmo de minería de datos tanto en precisión, como en agrupamiento de datos es el algoritmo DBSCAN, el cual se eligió en base a las pruebas realizadas. Dichas pruebas dieron como resultado que el algoritmo trabaja mejor para esta investigación que otros algoritmos como *K-means* o *X-means*. Ya que por su funcionamiento utiliza la agrupación de datos por densidad, la cual asocia datos basándose en un punto en común, además de generar poco desperdicio de datos (Navarrete, Francisco Javier, Huertas, n.d.) .

De acuerdo a trabajos similares se indica que el Ozono (O₃), en presencia de la luz solar específicamente de la Radiación Ultravioleta aumenta o reacciona positivamente (Athanasiadis et al., 2006), lo que concuerda con el trabajo realizado ya que dentro de la ciudad de Cuenca – Ecuador el comportamiento del Ozono (O₃) en relación a la Radiación Ultravioleta presentó una correlación superior a 0.5 indicando que hay un comportamiento proporcional entre las dos variables.

De igual manera, en un estudio realizado por (Othman, Ismail, & Latif, 2017), se ha comprobado que la temperatura reacciona como uno de los principales factores para que los contaminantes atmosféricos incrementen su comportamiento, se ha comprobado que a mayores temperaturas mayor es el índice de Ozono (O₃); lo cual concuerda con el trabajo realizado ya que la relación entre el Ozono (O₃) y la temperatura dentro de la ciudad presentó una correlación superior a 0.7 indicando que entre las dos variables existe una fuerte relación.

Sin embargo, en un estudio realizado por (Ramos-Herrera, Bautista-Margulis, & Valdez-Manzanilla, 2010), se indica que existe una correlación negativa entre la humedad relativa y el dióxido de nitrógeno (NO₂); lo cual no concuerda con el trabajo realizado ya que dentro de la ciudad se presentó un comportamiento positivo entre estas dos variables. Dentro de la misma investigación mencionada se indica que existe una correlación negativa entre el material particulado (PM_{2.5}) y la

humedad relativa, que concuerda con el trabajo realizado ya que a lo largo del análisis se demostró que en horas del día la correlación entre estas dos variables es superior a -0.5 indicando un comportamiento similar a la investigación mencionada.

En el trabajo realizado se utilizó un set de datos que almacena registros de un mes aproximadamente, del año 2017 de la ciudad de Cuenca – Ecuador, permitiendo cumplir con los objetivos planteados para este trabajo. Sin embargo, para una mejor estimación de resultados se podría considerar utilizar un set de datos más grande en trabajos próximos, ya que el actual set de datos presento un total de 2.053 registros después de su etapa de preparación. Por lo que, en un trabajo futuro al incluir una mayor cantidad de datos se podría analizar el comportamiento de los contaminantes y las variables meteorológicas de una forma más exacta.

Dentro de la ciudad de Cuenca – Ecuador el estudio realizado proporciona información muy importante, ya que analizar el comportamiento entre contaminantes atmosféricos y variables meteorológicos, brinda información que puede ser utilizada para diversos fines como predicción y conocimiento. El trabajo dio como resultado patrones de comportamiento entre estos dos grupos de variables.

5. CONCLUSIONES

Dentro del estudio se evaluaron cinco algoritmos de minería de datos: *K-means*, *X-means*, *CobWeb*, *Maximization Expectation Clustering* (EMC) y DBSCAN, con una data set final de 2.053 registros, dando como resultado después de la aplicación de los mismos que el mejor algoritmo para el análisis de este trabajo fue el algoritmo DBSCAN, ya que presenta una mejor agrupación de datos y una mayor concentración de los mismos dentro de su cluster más grande.

Mediante el uso de la matriz de correlación se obtuvieron los valores pertenecientes a la correlación entre los dos grupos de variables indicando que existen patrones de comportamiento entre los contaminantes atmosféricos y las variables meteorológicas, mediante el uso de gráficas de series de tiempo que representan el comportamiento de los dos grupos de variables, así como de su correlación, se logró identificar que el valor obtenido en la matriz de correlación muchas de las veces no concuerda con el comportamiento entre dos variables a lo largo del día.

En la mayoría de los casos se muestran fuertes correlaciones entre los contaminantes y las variables meteorológicas como es el caso del Ozono (O₃) que presenta una fuerte relación con la radiación global especialmente en horas de la mañana que alcanza valores cercanos a -0.75, y casos como: el Ozono (O₃) y la humedad relativa que indican una fuerte correlación inversamente proporcional de -0.8; sin embargo, se presentaron casos como el comportamiento entre el material particulado PM_{2.5} y la presión atmosférica, en que la correlación obtenida fue de 0.1, pero el análisis realizado mediante gráficas con series de tiempo dio como resultado que la correlación a lo largo del día alcanzaba un valor superior a 0.7.

Mediante la aplicación de técnicas de minería de datos en el análisis de contaminantes atmosféricos, se identificaron patrones de comportamiento entre estos dos grupos de variables cumpliendo con los objetivos planteados dentro del trabajo.

BIBLIOGRAFÍA

- Aghdam, A. E. F., & Hosseini, V. (2015). Decision Support System in Air Pollution Using Data Mining, 9(4), 22802.
- Aire, E. U. (2001). 1EPA El sol , la radiación Guía SunWise para la exposición al sol.
- Athanasiadis, I. N., Athanasiadis, I. N., Karatzas, K. D., Karatzas, K. D., Mitkas, P., & Mitkas, P. (2006). Classification techniques for air quality forecasting. *Fifth ECAI Workshop on Binding Environmental Sciences and Artificial Intelligence, 17th European Conference on Artificial Intelligence*, (August), 1–7.
- Atsdr, C. D. C., & Sciences, H. H. (n.d.). ToxFAQs™ sobre el monóxido de carbono., 1–2.
- Banerjee, T., Singh, S. B., & Srivastava, R. K. (2011). Development and performance evaluation of statistical models correlating air pollutants and meteorological variables at Pantnagar, India. *Atmospheric Research*, 99(3–4), 505–517. <https://doi.org/10.1016/j.atmosres.2010.12.003>
- Basantes, A. C. N., & García, E. H. (2018). Altitud , variables climáticas y tiempo de permanencia de las personas en plazas de Ecuador. *Urbe. Revista Brasileira de Gestão Urbana*, 10(2), 414–425. <https://doi.org/10.1590/2175-3369.010.002.ao11>.
- Berzal, F. (2012). Introducción Medidas de similitud. *Decsai*, 1(5), 35–50. Retrieved from elvex.ugr.es/decsai/intelligent/slides/dm/D3_Clustering.pdf
- Cagliero, L., Cerquitelli, T., Chiusano, S., Garza, P., Ricupero, G., & Xiao, X. (2016). Modeling Correlations among Air Pollution-Related Data through Generalized Association Rules. *2016 IEEE International Conference on Smart Computing, SMARTCOMP 2016*. <https://doi.org/10.1109/SMARTCOMP.2016.7501707>
- Christian, R. A., Jariwala, N. D., & Gohil, D. B. (2010). Summer episode temperature inversion and its impact on CO concentration in urban environment. *ICBEE 2010 - 2010 2nd International Conference on Chemical, Biological and Environmental Engineering, Proceedings*, (Icbee), 327–330. <https://doi.org/10.1109/ICBEE.2010.5649079>
- Christina, J., & Komathy, K. (2013). Analysis of hard clustering algorithms applicable to regionalization. *2013 IEEE Conference on Information and Communication Technologies, ICT 2013*, (Ict), 606–610. <https://doi.org/10.1109/CICT.2013.6558166>
- Del, E., En, S. O., & Salud, L. A. (2010). 3. Efectos Del So2 En La Salud De Las Personas 3.1., 167–219.
- Elminir, H. K. (2005). Dependence of urban air pollutants on meteorology. *Science of the Total Environment*, 350(1–3), 225–237. <https://doi.org/10.1016/j.scitotenv.2005.01.043>
- Esfandani, M. A., & Nematzadeh, H. (2016). Predicting air pollution in Tehran: Genetic algorithm and back propagation neural network. *Journal of Artificial Intelligence and Data Mining*, 4(1), 49–54. <https://doi.org/10.5829/idosi.JAIDM.2016.04.01.06>
- Gallardo, J. (2017). Modelos de proceso para proyectos de Data Mining (DM) CRISP-DM (Cross Industry Standard Process for Data Mining). Retrieved from http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM.2385037.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. *2017 1st International Conference on Intelligent Systems and Information Management (ICISIM)*, 58–61. <https://doi.org/10.1109/ICISIM.2017.8122148>
- Lanjewar, U. M., & Shah, J. J. (2012). Air Pollution Monitoring & Tracking System Using Mobile Sensors and Analysis of Data Using Data Mining. *International Journal of Advanced Computer Research*, 2(4), 4–8.
- Linares, C., & Díaz, J. (2008). Las PM 2,5 y su afección a la salud. *El Ecologista*, 58, 46–49.
- Moreno García, M. N., & López Batista, V. F. (n.d.). Uso de Técnicas no Supervisadas en la Construcción de Modelos de Clasificación en Ingeniería del Software. *Departamento de Informática y Automática. Universidad de Salamanca*, 11. Retrieved from <http://www.lsi.us.es/redmidas/Capitulos/LMD14.pdf>
- Navarrete, Francisco Javier, Huertas, D. (n.d.). Dilver Huertas Guerrero Francisco Javier Navarrete Cristian Andres Melo. *Dbscan*, 1–14. Retrieved from http://www.disi.unal.edu.co/profesores/eleonguz/cursos/md/presentaciones/C_DBscan.pdf
- Nitrógeno, D. De. (2004). Consenso Científico sobre la Contaminación del Aire, 3–7.
- Othman, Z. A., Ismail, N., & Latif, M. T. (2017). Association rules of temperature towards high and low ozone in putrajaya. *Electrical Engineering and Informatics (ICEEI), 2017 6th International Conference On*, 1–5.
- Perdomo, J. G. (2007). Minería de Datos II, 284.
- Piura, U. de. (2006). Precipitación. *Insuficiencia Cardíaca Crónica*, 14–20. Retrieved from http://www.biblioteca.udep.edu.pe/bibvirudep/tesis/pdf/1_136_147_89_1257.pdf
- PNUMA. (2013). La Capa de Ozono y las SAO. *Campus*, 1–7. Retrieved from <http://www.pnuma.org/ozono/curso/pdf/m1.pdf>
- Ramos-Herrera, S., Bautista-Margulis, R., & Valdez-Manzanilla, a. (2010). Statistical study of the correlation between atmospheric pollutants and meteorological variables in northern Chiapas, Mexico. *Universidad y Ciencia*, 26(1), 65–80.
- Rapidminer. (2018). Lightning Fast Data Science Platform | RapidMiner. Retrieved August 20, 2018, from <https://rapidminer.com/>
- Red de Monitoreo de Calidad del Aire de Cuenca de la EMOV EP. (2015). Informe de la Calidad del Aire de Cuenca, 124. <https://doi.org/10.1017/CBO9781107415324.004>
- Ri, K. G., Manimegalai, R., Si, G. D. M., Si, R., Ki, U., & Ni, R. B. (2016). Air Pollution Analysis Using Enhanced K-Means Clustering Algorithm for Real Time Sensor Data, (August 2006), 1945–1949.
- Rodríguez Jimenez, R. M., Capa, Á. B., & Portela Lozano, A. (2004). *Meteorología y Climatología*. <https://doi.org/M-XXXXX-2004>
- Rojas Dávila, M. A. (2003). Técnicas Estadísticas Paramétricas y No Paramétricas Equivalentes: Resultados Comparativos por Simulación, 312.
- The R Project for Statistical Computing. (2018). R: El Proyecto R para Computación Estadística. Retrieved August 20, 2018, from <https://www.r-project.org/>
- Vaisala. (2013). Punto de rocío en el aire comprimido, 1–4. Retrieved from <https://goo.gl/ats0aA>

ANEXOS

Figura 19

Script para realizar unión entre set de datos

```
with cort as (select c.ec_time_stamp,
to_date(to_char(c.ec_time_stamp, 'DD-MM-YYYY HH24:') ||
substr(to_char(c.ec_time_stamp, 'MI'), 0, 1) || '0:00',
'DD-MM-YYYY HH24:MI:SS') HORADEC,
c.o3,
c.co,
c.no2,
c.so2,
c.pm2_5
from tc_con_ori piv c)
select *
from tc_envmet m left join cort o
on m.fechor=o.HORADEC

create table tc_ca$met_avgs as(
select tc_med_ca.*,tc_envmet.*
from tc_envmet,
(select to_date(to_char(c.ec_time_stamp, 'DD-MM-YYYY HH24:') ||
substr(to_char(c.ec_time_stamp, 'MI'), 0, 1) ||
'0:00',
'DD-MM-YYYY HH24:MI:SS') HORADEC,
avg(c.o3) o3_avg,
avg(c.co) co_avg,
avg(c.no2) no2_avg,
avg(c.so2) so2_avg,
avg(c.pm2_5) pm2_5_avg
from tc_con_ori piv c
group by to_date(to_char(c.ec_time_stamp, 'DD-MM-YYYY HH24:') ||
substr(to_char(c.ec_time_stamp, 'MI'), 0, 1) ||
'0:00',
'DD-MM-YYYY HH24:MI:SS')) tc_med_ca
where tc_med_ca.HORADEC = tc_envmet.fechor)
```

Script desarrollado en PLSQL para el proceso de *match* entre set de datos.

Figura 20

Comportamiento entre Ozono (O3) y Punto del rocío

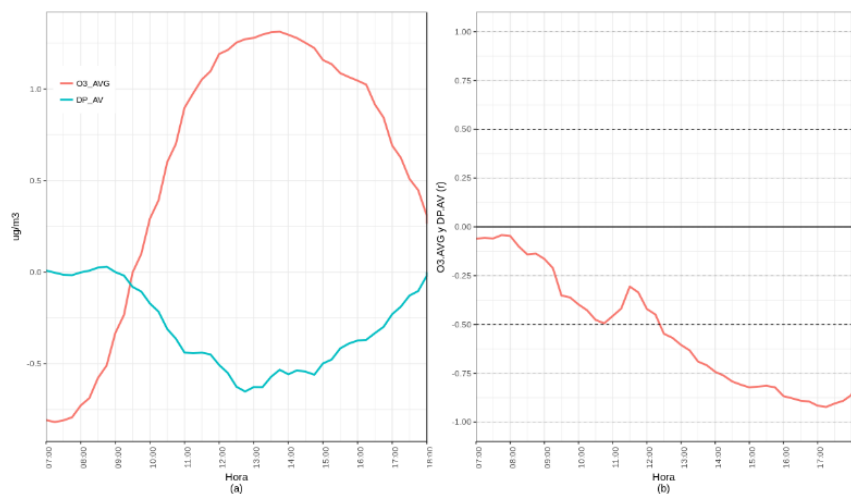


Figura 21

Comportamiento entre Ozono (O3) y Humedad relativa

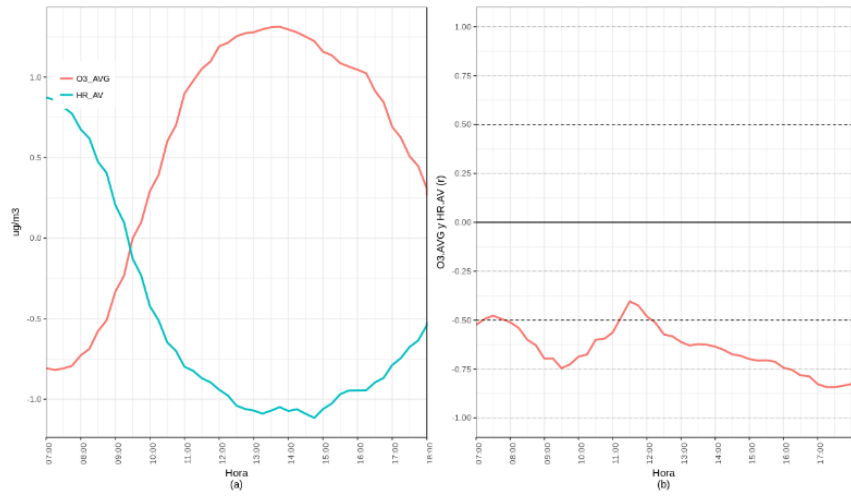


Figura 22

Comportamiento entre Ozono (O3) y Precipitación

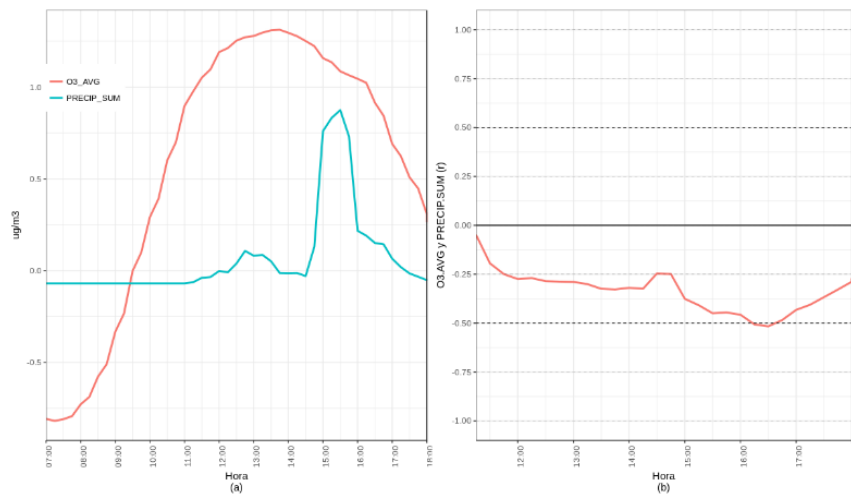


Figura 23

Comportamiento entre Ozono (O3) y Radiación Global

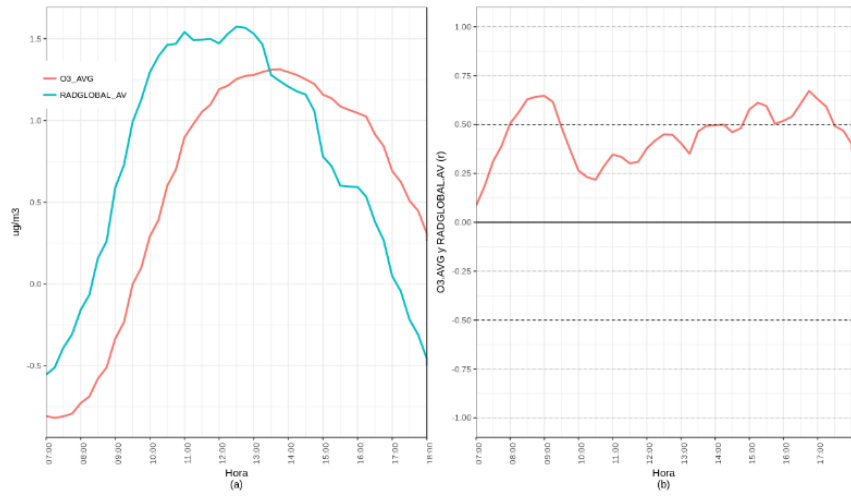


Figura 24

Comportamiento entre Ozono (O3) y Temperatura del aire

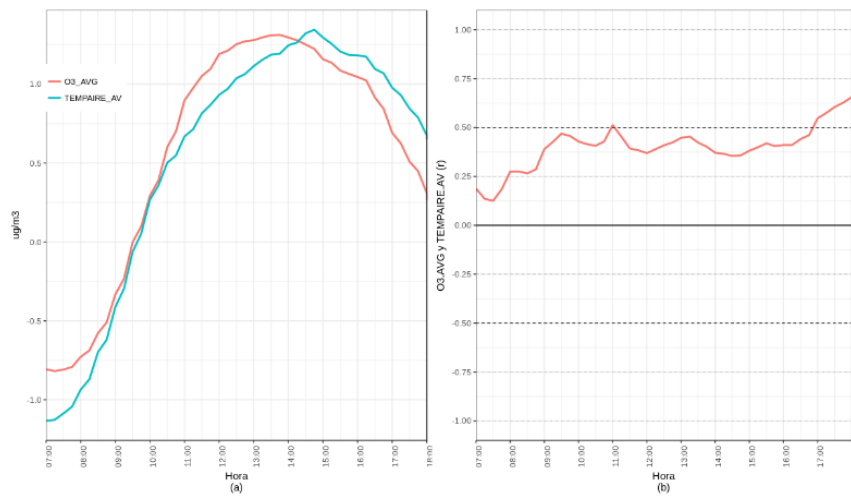


Figura 25

Comportamiento entre Ozono (O3) y Radiación UVA

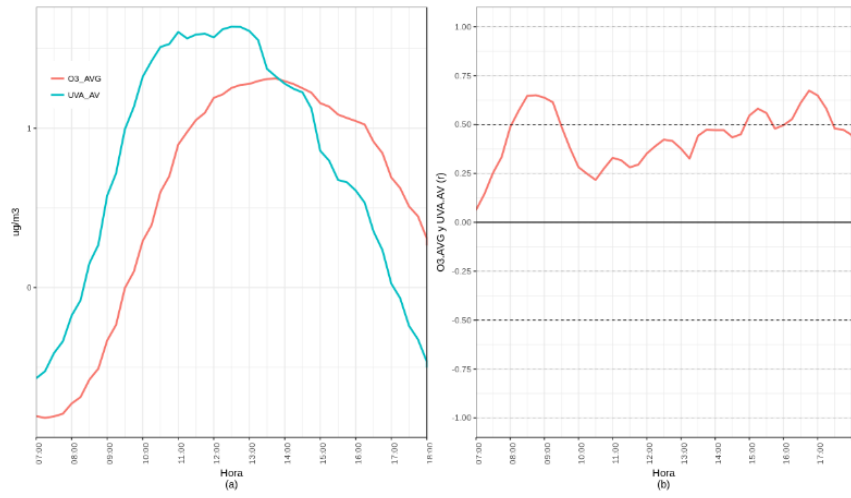


Figura 26

Comportamiento entre Ozono (O3) y Radiación UVE

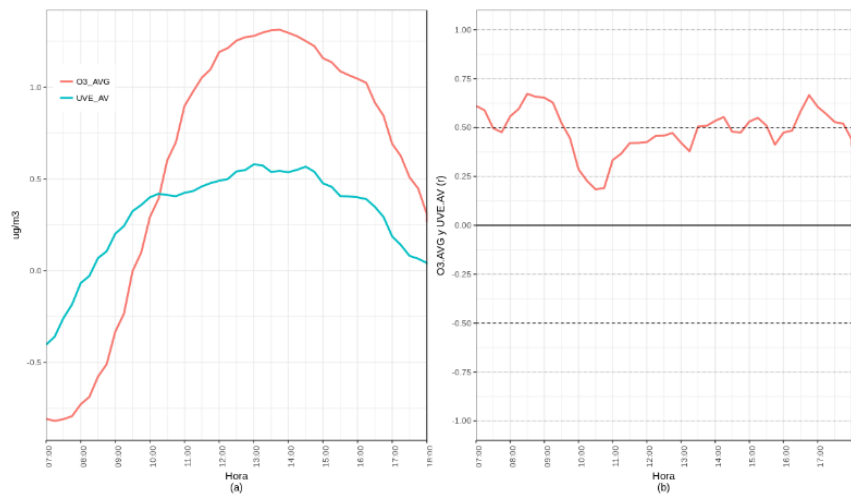


Figura 27

Comportamiento entre Monóxido de carbono (CO) y Humedad relativa

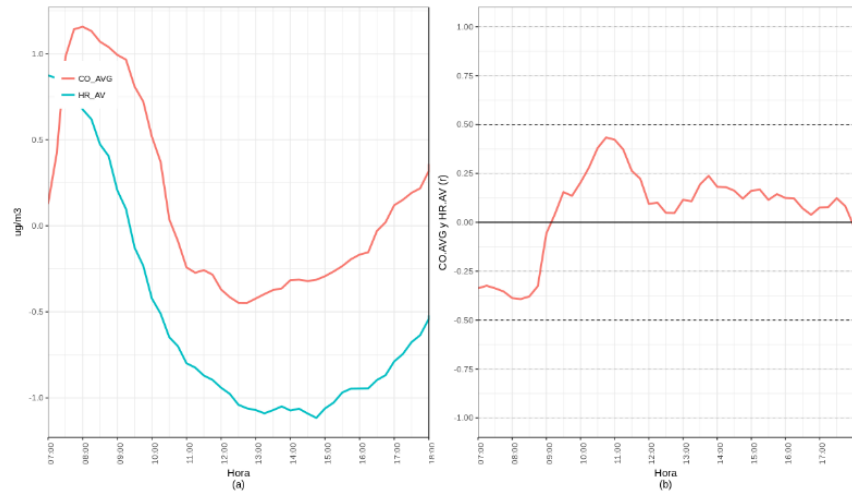


Figura 28

Comportamiento entre Monóxido de carbono (CO) y Presión atmosférica

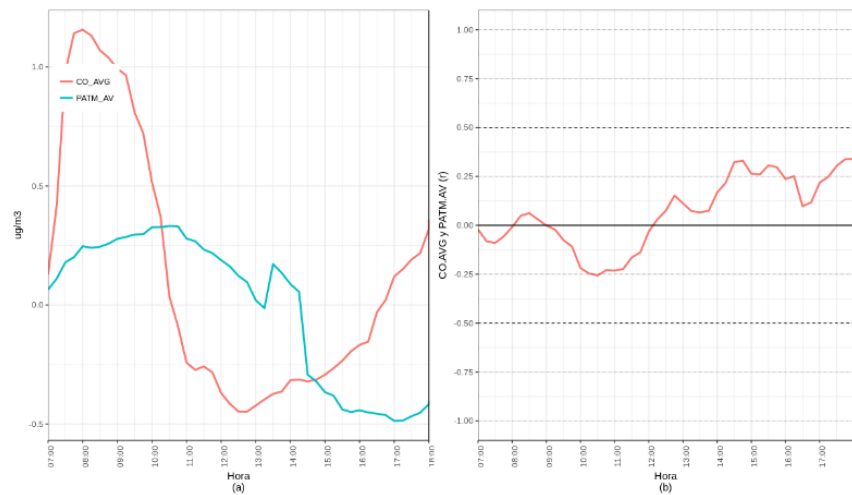


Figura 29

Comportamiento entre Monóxido de carbono (CO) y Precipitación

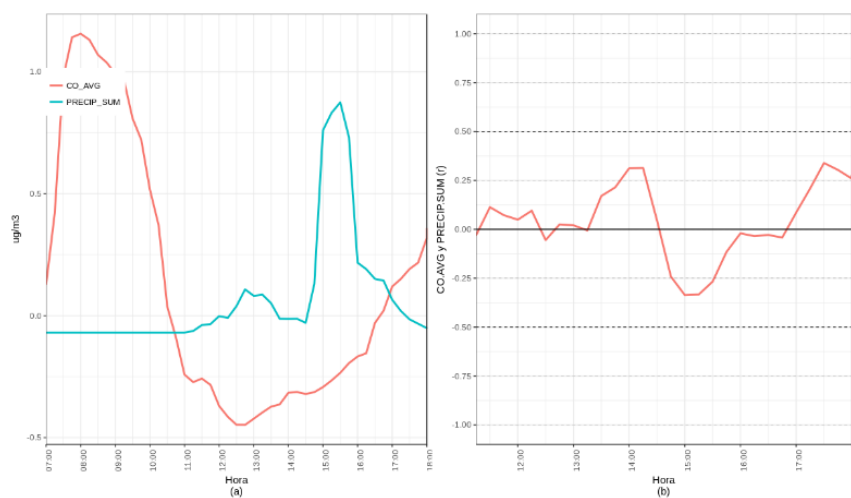


Figura 30

Comportamiento entre Monóxido de carbono (CO) y Radiación global

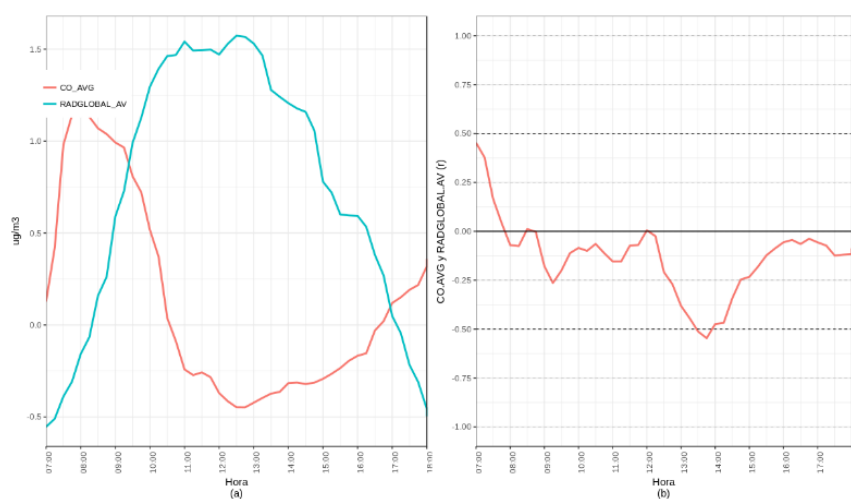


Figura 31

Comportamiento entre Monóxido de carbono (CO) y Temperatura del aire

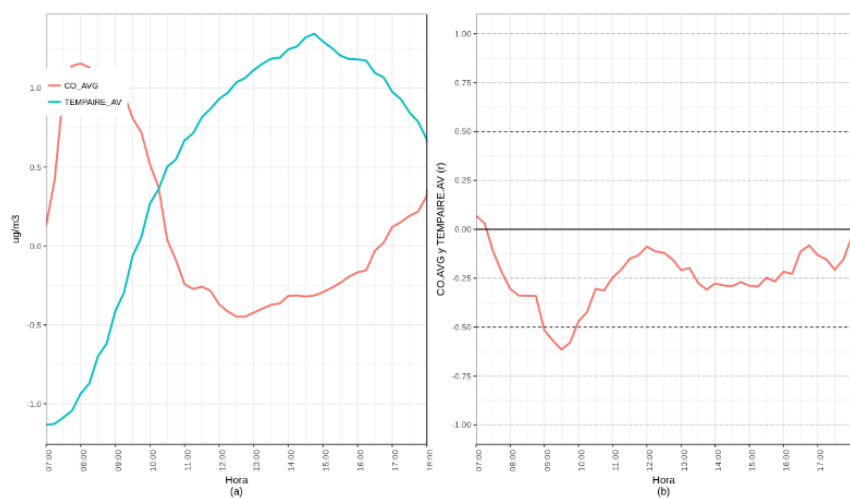


Figura 32

Comportamiento entre Monóxido de carbono (CO) y Radiación UVE

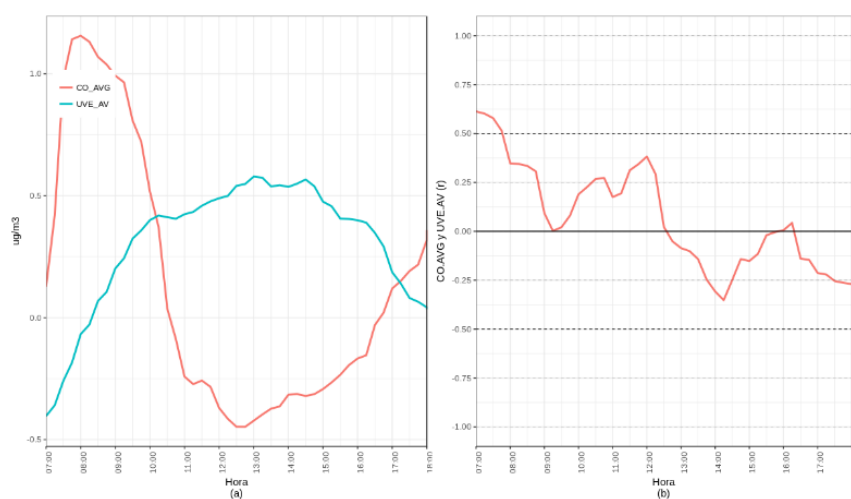


Figura 33

Comportamiento entre Monóxido de carbono (CO) y Velocidad del viento

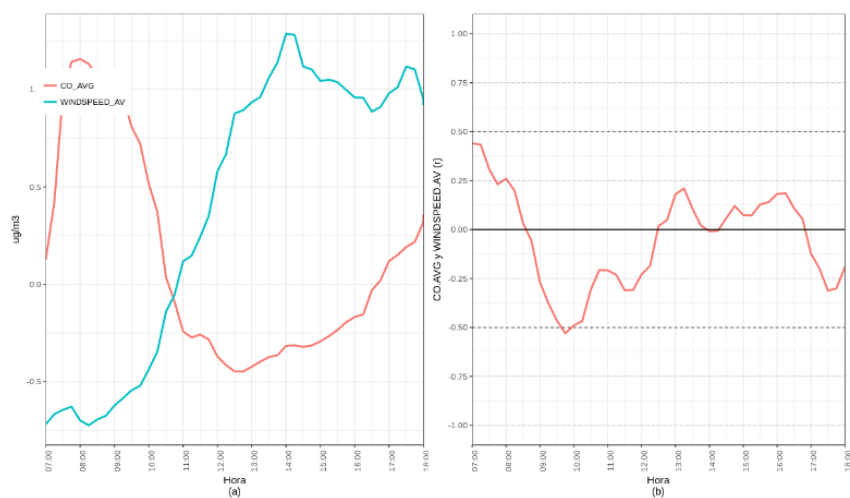


Figura 34

Comportamiento entre Dióxido de nitrógeno (NO2) y Punto del rocío

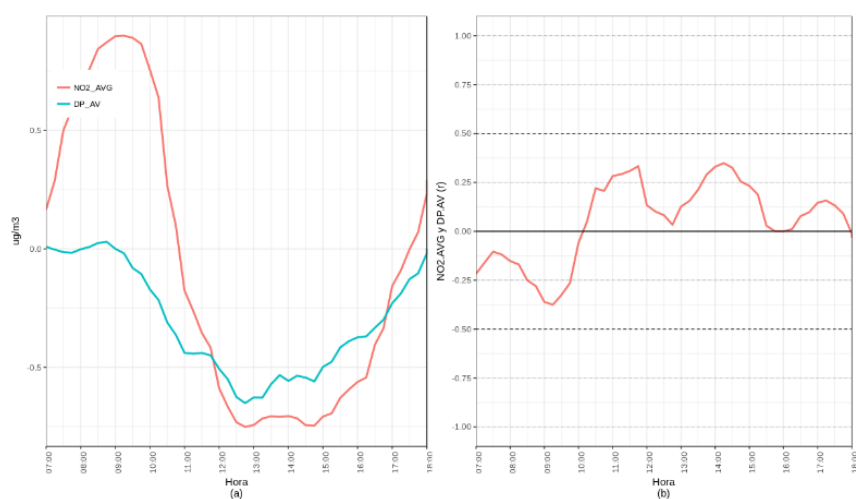


Figura 35

Comportamiento entre Dióxido de nitrógeno (NO₂) y Presión atmosférica

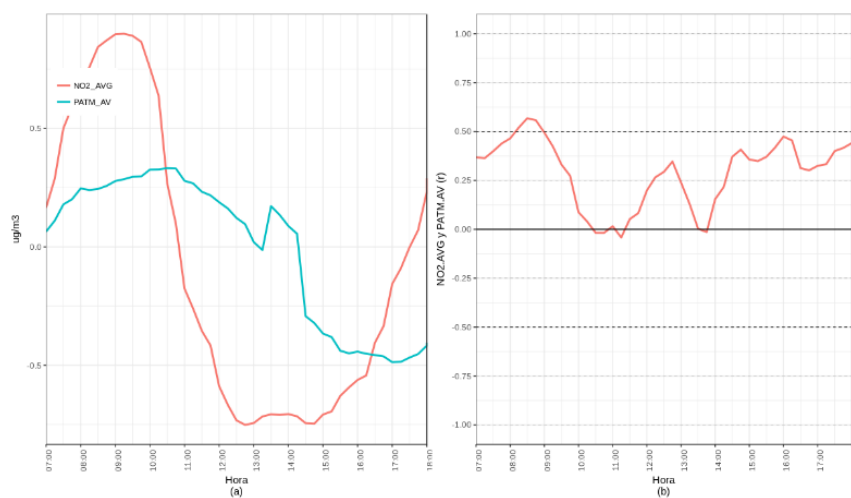


Figura 36

Comportamiento entre Dióxido de nitrógeno (NO₂) y Precipitación

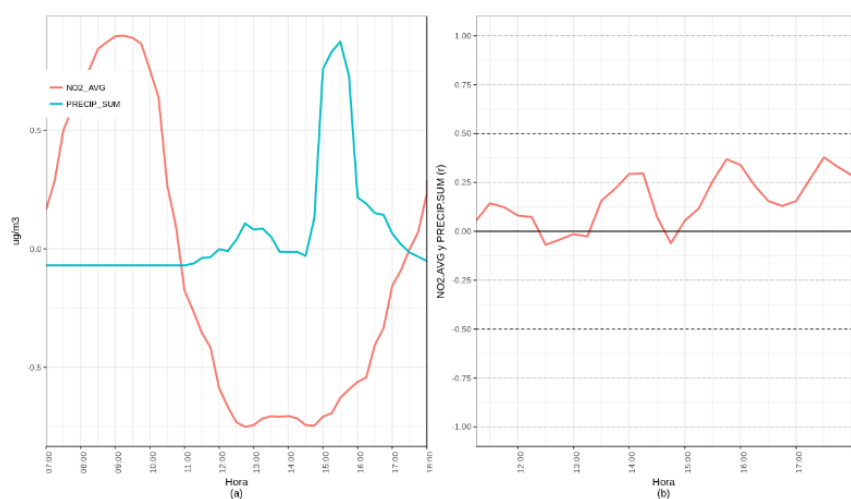


Figura 37

Comportamiento entre Dióxido de nitrógeno (NO₂) y Radiación global

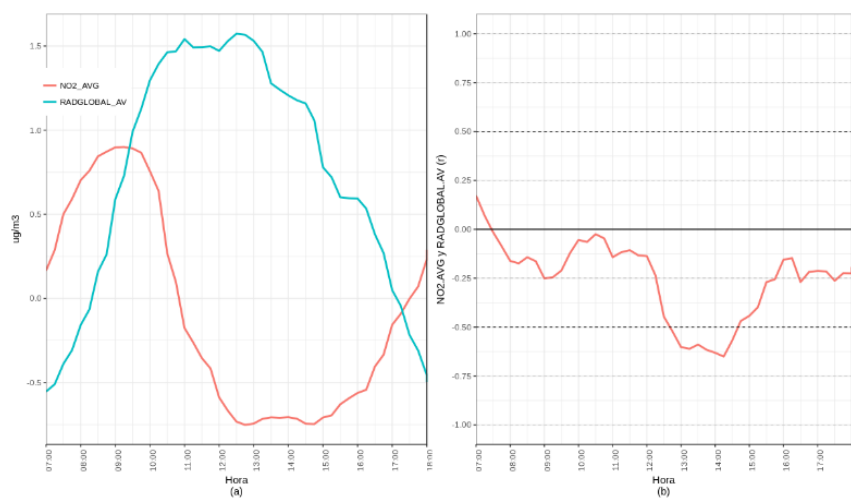


Figura 38

Comportamiento entre Dióxido de nitrógeno (NO₂) y Temperatura del aire

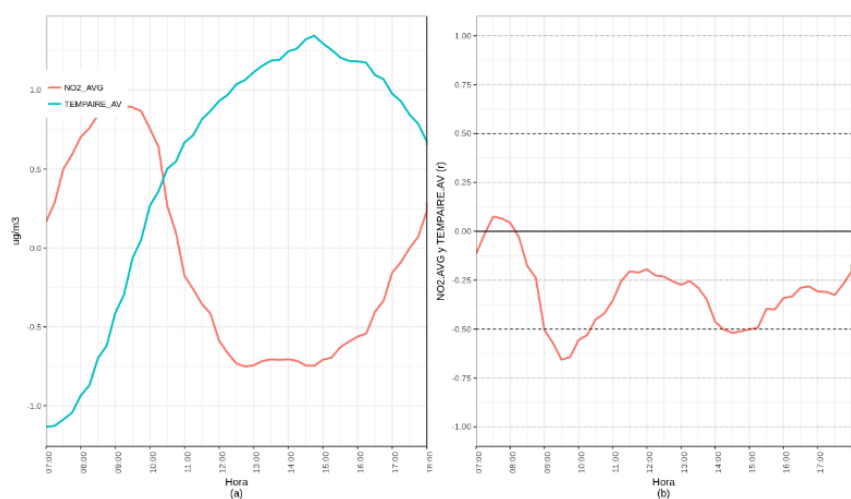


Figura 39

Comportamiento entre Dióxido de nitrógeno (NO₂) y Radiación UVA

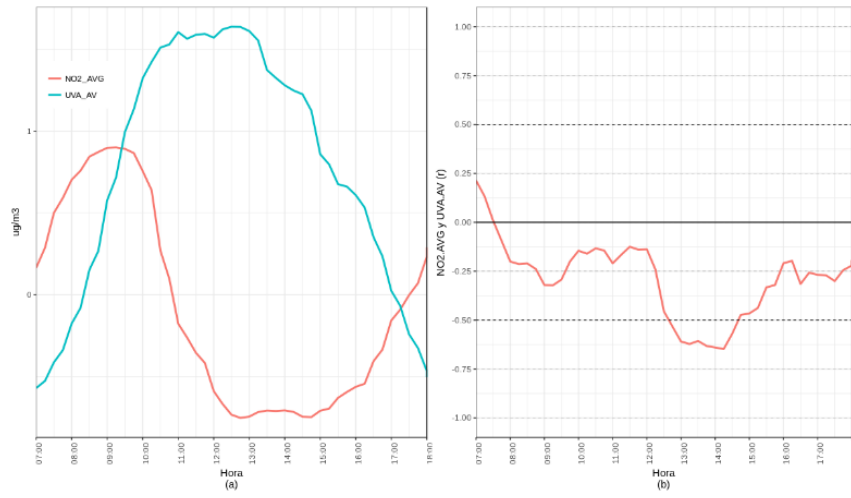


Figura 40

Comportamiento entre Dióxido de nitrógeno (NO₂) y Velocidad del viento

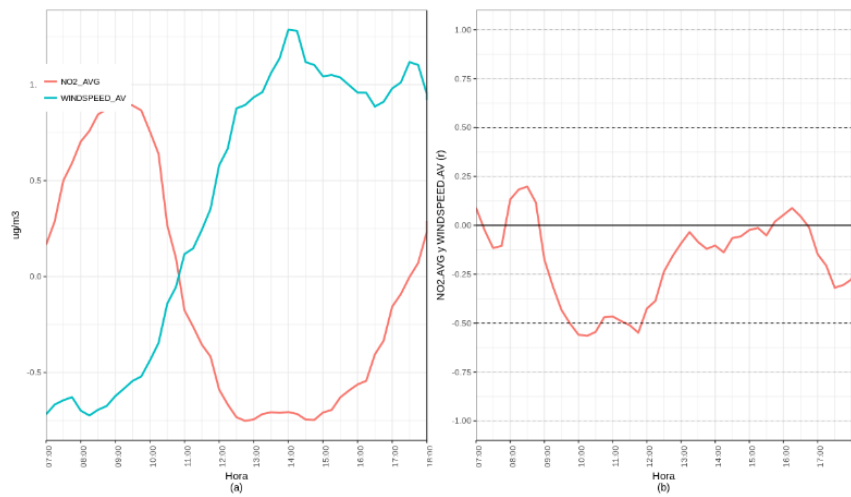


Figura 41

Comportamiento entre Dióxido de azufre (SO₂) y Punto del rocío

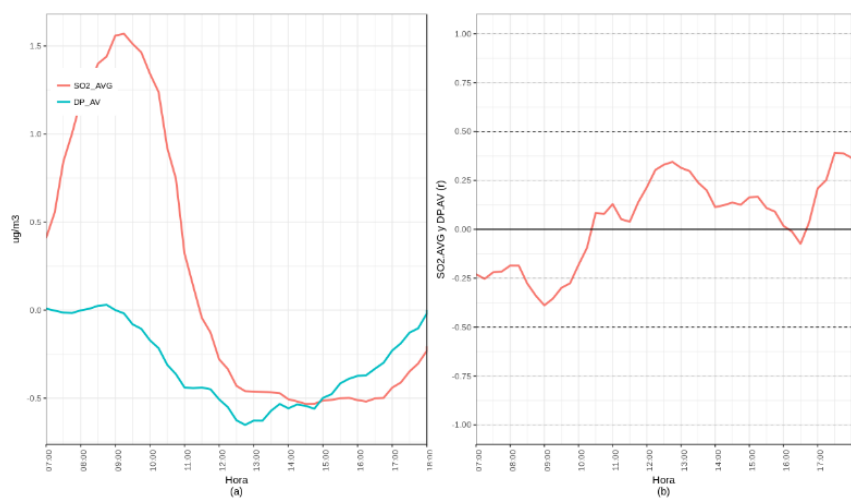


Figura 42

Comportamiento entre Dióxido de azufre (SO₂) y Humedad Relativa

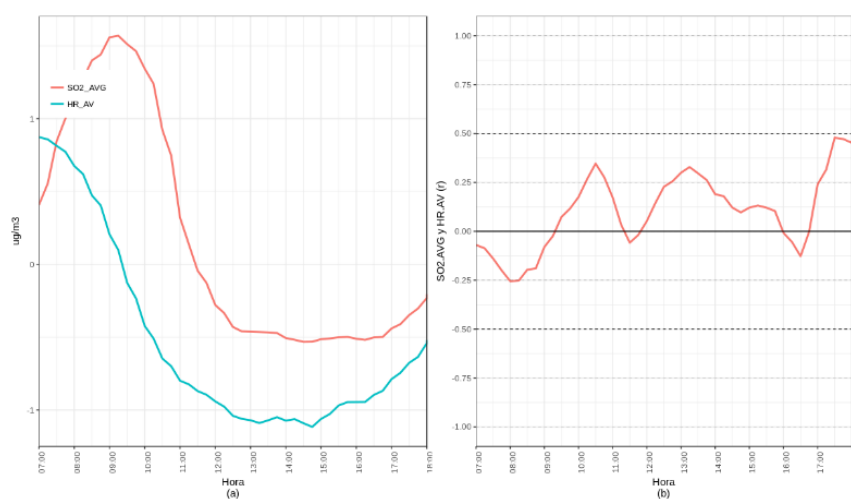


Figura 43

Comportamiento entre Dióxido de azufre (SO₂) y Radiación Global

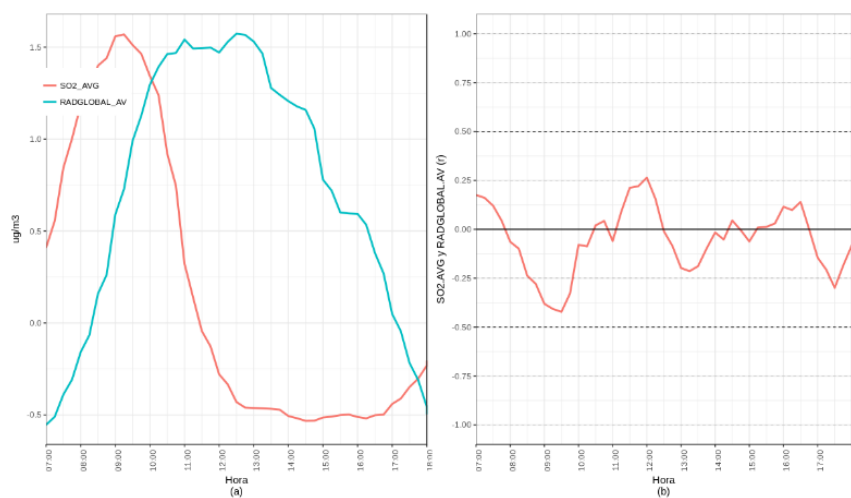


Figura 44

Comportamiento entre Dióxido de azufre (SO₂) y Temperatura del aire

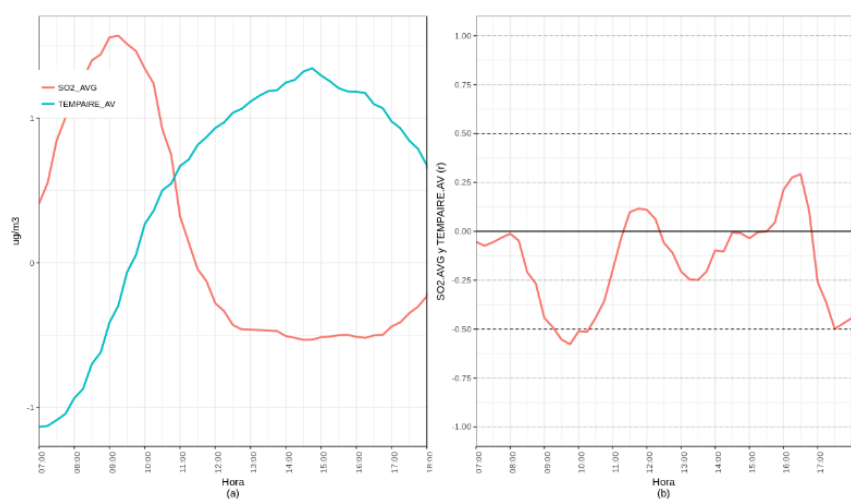


Figura 45

Comportamiento entre Dióxido de azufre (SO₂) y Radiación UVA

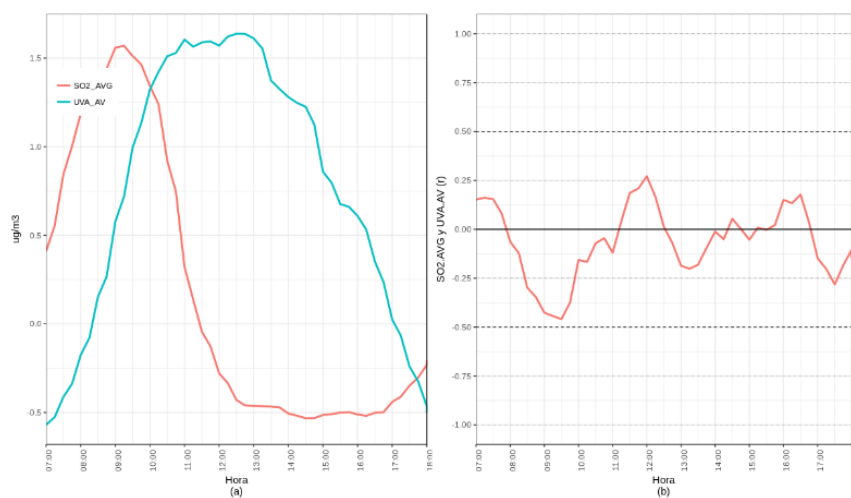


Figura 46

Comportamiento entre Dióxido de azufre (SO₂) y Radiación UVE

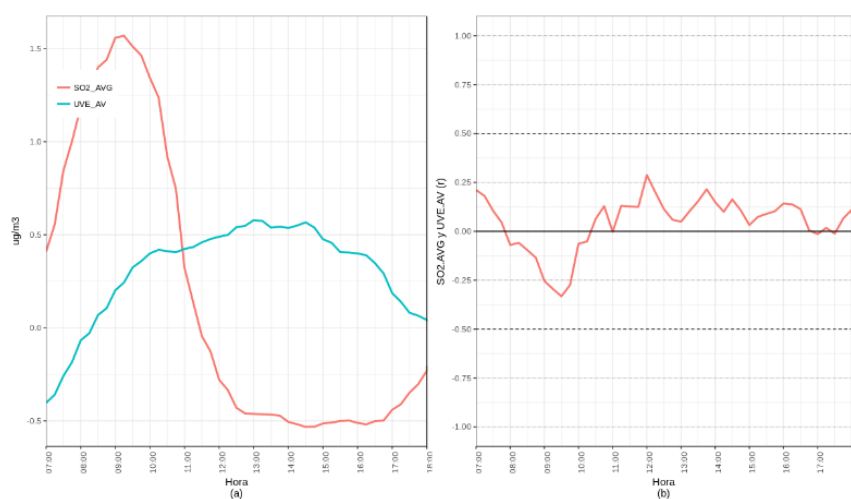


Figura 47

Comportamiento entre Dióxido de azufre (SO₂) y la Velocidad del viento

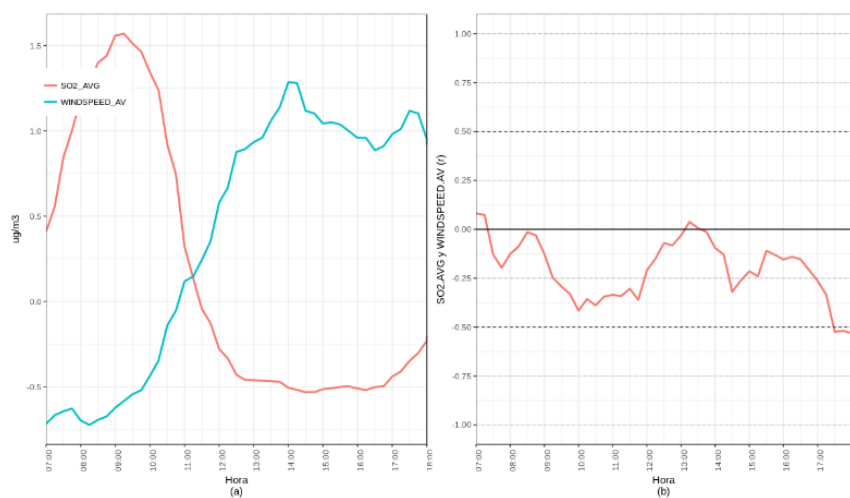


Figura 48

Comportamiento entre Material particulado PM_{2.5} y Punto del rocío

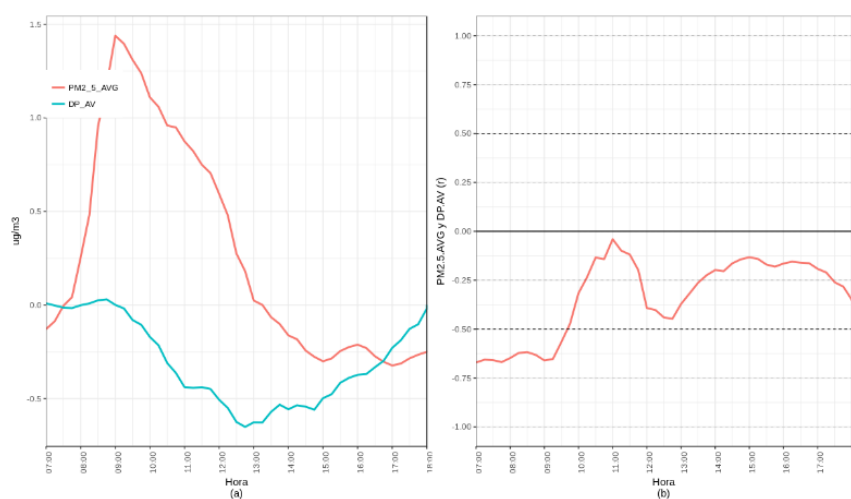


Figura 49

Comportamiento entre Material particulado PM2.5 y Humedad relativa

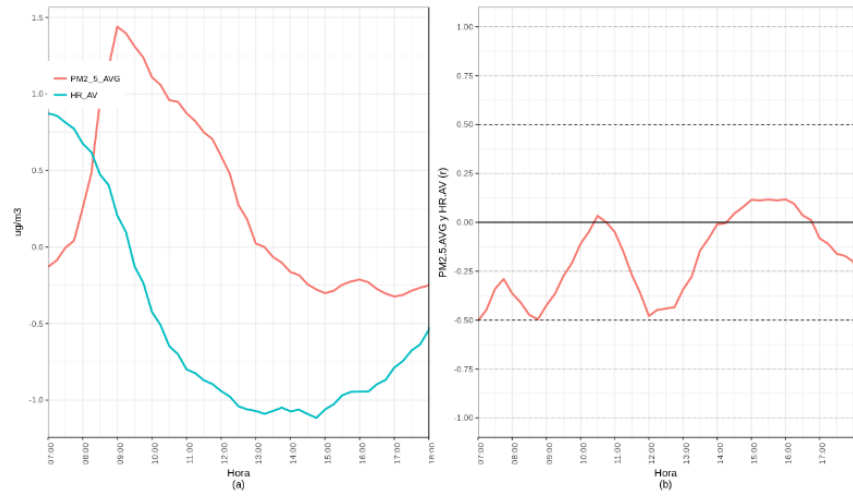


Figura 50

Comportamiento entre Material particulado PM2.5 y Precipitación

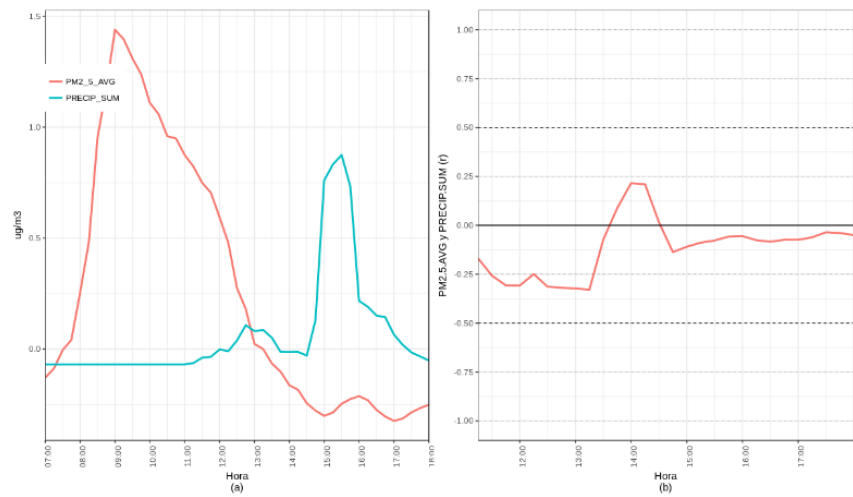


Figura 51

Comportamiento entre Material particulado PM2.5 y Radiación global

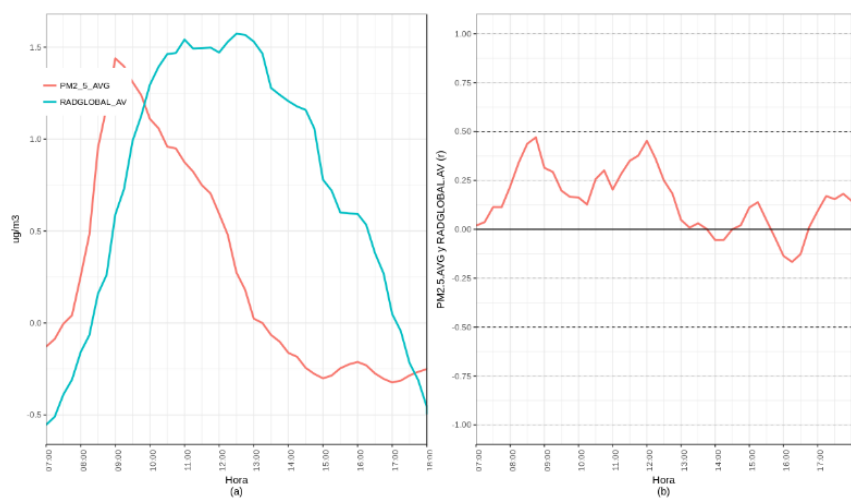


Figura 52

Comportamiento entre Material particulado PM2.5 y Radiación UVA

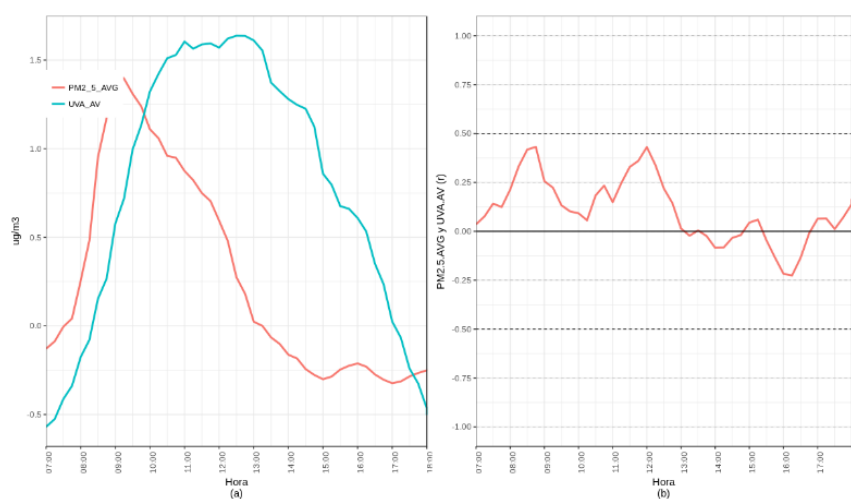


Figura 53

Comportamiento entre Material particulado PM2.5 y Radiación UVE

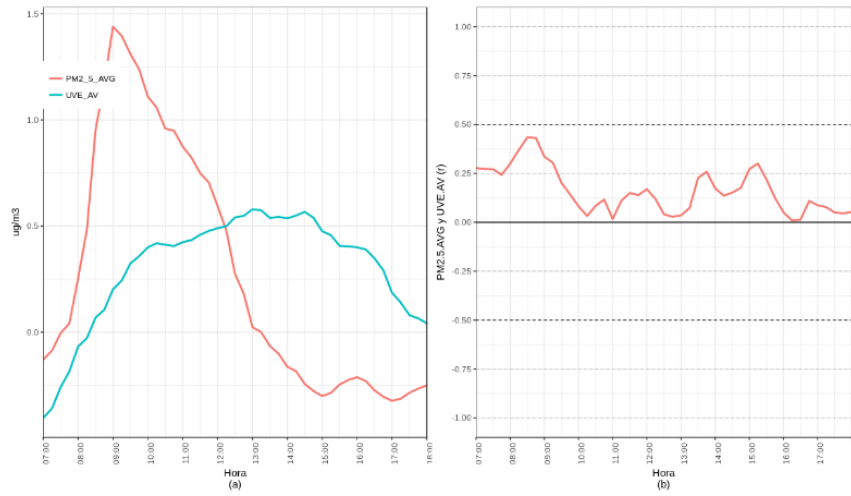
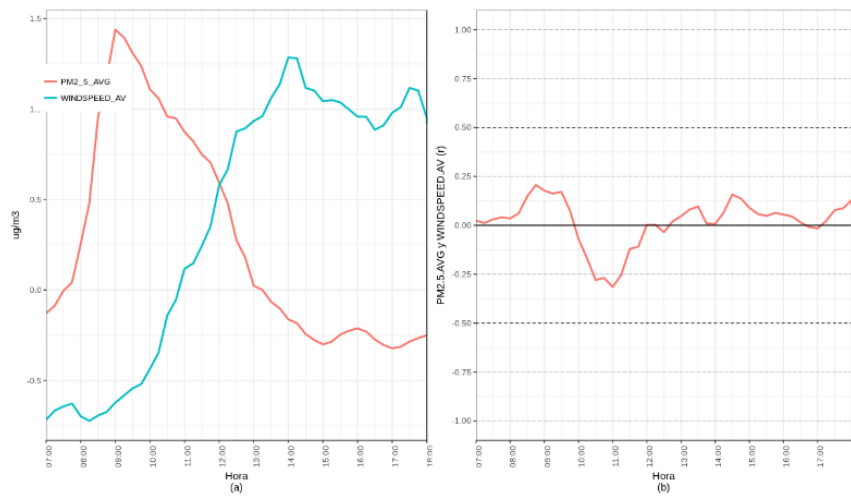


Figura 54

Comportamiento entre Material particulado PM2.5 y Velocidad del viento



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 en sesión del 11 de mayo de 2018, conoció y aprobó la solicitud para realización del trabajo de titulación, presentada por :

Estudiante: Pedro Sebastián Andrade Durazno (cód. 63744)

Tema: “Aplicación de minería de datos en el análisis de contaminantes atmosféricos y variables meteorológicas”

Previo a la obtención del título de Ingeniero de Sistemas y Telemática

Director: Ing. Marcos Orellana

Tribunal: Agrim. Daniela Ballari

Ing. Patricia Ortega

Plazo de presentación del trabajo de titulación, con la calificación del Director: seis meses a partir de la fecha de aprobación, esto es hasta el 11 de noviembre de 2018.

E INFORMA:

Que en aplicación de la Disposición General Cuarta del Reglamento de Régimen Académico vigente, en caso de que el estudiante no culmine y apruebe el trabajo de titulación luego de dos períodos académicos contados a partir de la fecha de culminación de estudios, deberá realizar la actualización de conocimientos previa a su titulación.

Cuenca, 14 de mayo de 2018



UNIVERSIDAD DEL AZUAY | Ciencias de la Administración
Facultad
Dra. María Elena Ramírez Aguilar
Secretaría - Abogada

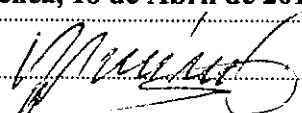


DOCTORA MARÍA ELENA RAMÍREZ AGUILAR SECRETARIA DE LA
FACULTAD DE CIENCIAS DE LA ADMINSTRACION DE LA UNIVERSIDAD
DEL AZUAY

CERTIFICA:

Que, el Señor Pedro Sebastián Andrade Durazno, registrado con código 63744,
perteneciente a la escuela de Ingeniería de Sistemas y Telemática tiene aprobado más
del 80% de su pensum de estudios.

Cuenca, 18 de Abril de 2018


Dra. María Elena Ramírez Aguilar
SECRETARIA DE LA FACULTAD DE
CIENCIAS DE LA ADMINISTRACIÓN



Derecho.63744
vcf



Cuenca 4 de Mayo del 2018

Ingeniero,

Oswaldo Merchán Manzano

DECANO DE LA FACULTAD DE CIENCIAS DE LA ADMINISTRACION

UNIVERSIDAD DEL AZUAY

De mi consideración,

Estimado Señor Decano, yo **Pedro Sebastián Andrade Durazno** con C.I. **0104636972**, código estudiantil **63744**; estudiante de la carrera de Sistemas y Telemática, solicito muy comedidamente a usted y por su intermedio al Consejo de Facultad, la aprobación del protocolo de trabajo de titulación con el tema **"Aplicación de minería de datos en el análisis de contaminantes atmosféricos y variables meteorológicas"** previo a la obtención del título de Ingeniero en Sistemas y Telemática para lo cual adjunto la documentación respectiva.

Por la favorable acogida que brinde a la presente, anticipo mi agradecimiento.

Atentamente

Pedro Sebastián Andrade Durazno

Estudiante de la Carrera de Sistemas y Telemática

Cuenca, 4 de Mayo de 2018

Señor Ingeniero

Oswaldo Merchán Manzano

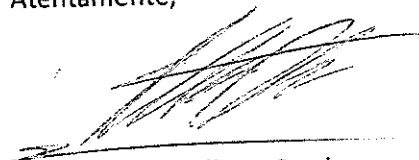
DECANO DE LA FACULTAD DE CIENCIAS DE LA ADMINISTRACIÓN

Presente.-

De mis consideraciones:

Luego de revisar el diseño del trabajo de graduación denominado "Aplicación de minería de datos en el análisis de contaminantes atmosféricos y variables meteorológicas", presentado por Pedro Sebastián Andrade Durazno estudiante de la Escuela de Ingeniería de Sistemas y Telemática. Considero que el documento cumple con las normas legales y reglamentarias de la Universidad y de la Facultad de Ciencias de la Administración, por lo que recomiendo su aprobación por parte del Consejo de Facultad.

Atentamente,



Ing. Marcos Orellana Cordero
Director

Cuenca, 9 de mayo de 2018

Señor Ingeniero
Oswaldo Merchán Manzano
DECANO DE LA FACULTAD DE CIENCIAS DE LA ADMINISTRACIÓN
Presente.-

De mis consideraciones:

El tribunal designado para el trabajo de titulación denominado "Aplicación de minería de datos en el análisis de contaminantes atmosféricos y variables meteorológicas", presentado por Pedro Sebastián Andrade, estudiante de la Escuela de Ingeniería de Sistemas y Telemática, sugirió cambios en el diseño del trabajo de titulación. Debo ratificar que los cambios fueron realizados con los ajustes sugeridos, por lo que solicito se proceda con la aprobación por parte del Consejo de Facultad.

Atentamente,

MARCOS	Firmado digitalmente
PATRICIO	por MARCOS
ORELLANA	PATRICIO ORELLANA
CORDERO	CORDERO
	Fecha: 2018.05.09
	15:58:02 -05'00'

Ing. Marcos Orellana Cordero
Director
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: "**Aplicación de minería de datos en el análisis de contaminantes atmosféricos y variables meteorológicas**", presentado por el estudiante Pedro Sebastián Andrade Durazno con códigos 63774 , previa a la obtención del título de Ingeniero de Sistemas y Telemática, para el día **Miércoles, 09 de mayo de 2018 a las 12:40.**

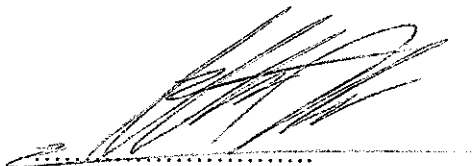
Tomar en cuenta que posterior a la sustentación del Diseño del Trabajo de Titulación, por ningún concepto se puede realizar modificaciones ni cambios en los documentos; únicamente, en caso de diseño aprobado con modificación, el Director adjuntará al esquema un oficio indicando que se procede con los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

Cuenca, 07 de mayo de 2018

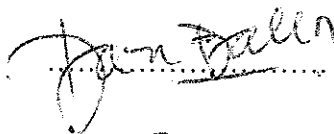


Dra. María Elena Ramírez Aguilar
Secretaria de la Facultad

Ing. Marcos Orellana



Ing. Daniela Ballari



Ing. Patricia Ortega



ESCUELA DE INGENIERÍA DE SISTEMAS Y TELEMÁTICA

FECHA: 07 de MAYO DE 2018.

Estudiante: Pedro Sebastián Andrade Durazno

Oficio Nro. 014-2018-DIST-UDA

Cuenca, 2 de mayo de 2018

Señor Ingeniero
Oswaldo Merchán Manzano
DECANO DE LA FACULTAD DE CIENCIAS DE LA ADMINISTRACIÓN
Presente.-

De nuestras consideraciones:

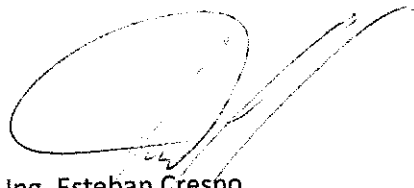
La Junta Académica de la Escuela de Ingeniería de Sistemas y Telemática, reunida el día 2 de mayo del 2018, recibió el proyecto de tesis titulado "Aplicación de minería de datos en el análisis de contaminantes atmosféricos y variables meteorológicas", presentado por Pedro Sebastián Andrade Durazno estudiante de la Escuela de Ingeniería de Sistemas y Telemática, y revisado por el Ing. Marcos Orellana, previo a la obtención del título de Ingeniero de Sistemas y Telemática.

Por lo expuesto, y de conformidad con el Reglamento de Graduación de la Facultad, recomendamos como director y responsable de aplicar cualquier modificación al diseño del trabajo de graduación posterior al Ing. Marcos Orellana y como miembros del tribunal a la Ing. Daniela Ballari Ph.D. e Ing. Patricia Ortega.

Atentamente,



Ing. Catalina Astudillo
Junta Académica de la Escuela de
Ingeniería de Sistemas y Telemática
Universidad del Azuay



Ing. Esteban Crespo
Junta Académica de la Escuela de
Ingeniería de Sistemas y Telemática
Universidad del Azuay



ACTA
SUSTENTACIÓN DE PROTOCOLO/DENUNCIA DEL TRABAJO DE TITULACIÓN

Fecha de sustentación: **Miércoles, 09 de mayo de 2018 a las 12:40**

1.1. Nombre del estudiante: Pedro Sebastián Andrade Durazno

1.2. Código: 63774

Director sugerido: Ing. Marcos Orellana

1.3. Codirector (opcional): _____

1.3.1. Tribunal: Agrim. Daniela Ballari e Ing. Patricia Ortega

1.3.2. Título propuesto: **"Aplicación de minería de datos en el análisis de contaminantes atmosféricos y variables meteorológicas"**

1.3.3. Aceptado sin modificaciones :

1.3.4. Aceptado con las siguientes modificaciones:

Detallar la metodología. (Temporalidad y datos)
Definir los métodos y las técnicas de minería.

1.3.5. No aceptado

1.3.6. Justificación:

Tribunal

Ing. Marcos Orellana

Agrim. Daniela Ballari

Ing. Patricia Ortega

Sr. Pedro Sebastián Andrade Durazno

Dra. María Elena Ramírez Aguilar
Secretaría de la Facultad

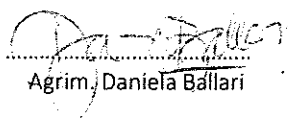


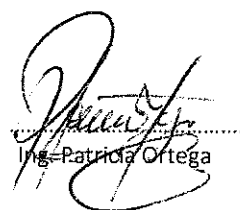
RÚBRICA PARA LA EVALUACIÓN DEL PROTOCOLO DE TRABAJO DE TITULACIÓN
(Tribunal)

- 1.1. Nombre del estudiante: Pedro Sebastián Andrade Durazno
1.2. Código : 63774
1.3. Director sugerido:
Codirector (opcional): Ing. Marcos Orellana
1.4. Título propuesto: **“Aplicación de minería de datos en el análisis de contaminantes atmosféricos y variables meteorológicas”**
1.5. Revisores tribunal: Agrim. Daniela Ballari e Ing. Patricia Ortega
1.6. Recomendaciones generales de la revisión:

	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?	/	


Ing. Marcos Orellana


Agrim. Daniela Ballari


Ing. Patricia Ortega



UNIVERSIDAD DEL AZUAY
INGENIERIA EN SISTEMAS Y TELEMATICA
DISEÑO DE TESIS

1. DATOS GENERALES

1.1 Nombre del estudiante: Pedro Sebastián Andrade Durazno

1.1.1 Código: 63744

1.1.2 Contacto:

Teléfono convencional: 072817788

Celular: 0995155552

Correo electrónico: thundersad3878@outlook.com

1.3 Director sugerido: Orellana Cordero, Marcos Ing.

1.2.1 Contacto:

Teléfono convencional: 4128640

Celular: 0999955611

Correo electrónico: marore@uazuay.edu.ec

1.3 Co-director sugerido:

1.4 Asesor metodológico:

1.5 Tribunal designado:

1.6 Aprobación: Junta Académica:

Consejo de Facultad:

1.7 Línea de Investigación de la carrera:

1.7.1 Código UNESCO:

1203. Informática de computadores.

1203.17 Informática

1.7.2 Tipo de trabajo: Investigación.

1.8 Área de estudio: Minería de Datos.

1.9 Título propuesto:

Aplicación de minería de datos en el análisis de contaminantes atmosféricos y variables meteorológicas.

1.10 Subtítulo:

1.11 Estado del proyecto:

Actualmente existe un proyecto en la ciudad de Cuenca –Ecuador, encargado de monitorear los niveles de contaminación del aire en base a factores atmosféricos, además, se encarga de la recopilación de datos de variables meteorológicas. En una fase anterior del proyecto se analizó mediante minería de datos la relación que existe entre factores de contaminación atmosférica. En esta etapa se analizará la relación entre variables atmosféricas y variables meteorológicas aplicando técnicas de minería de datos.

2. CONTENIDO

2.1 Motivación de la investigación:

De acuerdo a la problemática actual que presenta la contaminación ambiental es importante realizar estudios que ayuden a analizar el comportamiento de los factores que intervienen en dicha problemática. Un análisis mediante la minería de datos es el principal objetivo de esta investigación. Se pretende definir patrones de relación entre contaminantes atmosféricos y variables meteorológicas, para de esta manera contribuir al proyecto del departamento de investigación IERSE y al grupo de investigación LIDI.

2.2 Problemática:

La contaminación ambiental es un problema que afecta directamente a la salud de las personas, por lo que en base a esta realidad, el centro de investigaciones IERSE inició un proyecto para monitorear la calidad del aire en la ciudad de Cuenca – Ecuador, en una fase anterior de este proyecto se realizó el análisis de factores de correlación entre variables de contaminación atmosférica, sin embargo, no se ha realizado un estudio que identifique patrones de relación entre variables de tipo atmosférica y meteorológica, lo cual se realizará en una nueva fase del proyecto.

2.3 Pregunta de investigación:

¿Cuáles son los patrones de comportamiento entre variables de contaminación atmosférica y meteorológica?

2.4 Resumen:

Este proyecto es la continuación de una investigación ya iniciada por el departamento de investigaciones IERSE en colaboración con el grupo de investigación LIDI. Se realizó un estudio de variables de contaminación atmosférica presentes en la ciudad de Cuenca – Ecuador y dicha investigación analizó la información mediante minería de datos, dando como resultado patrones de correlación entre estas variables. En este trabajo se analizará la información relacionada con datos de variables atmosféricas y meteorológicas como son: Temperatura del aire, Humedad Relativa, Velocidad del Viento, Radiación Ultravioleta (UVA), y se aplicará técnicas de minería de datos a la información recolectada de acuerdo a estos grupos de variables. El estudio de esta información generará conocimiento, que puede aplicarse a la predicción o construir indicadores de correlación entre estos dos grupos de variables.

2.5. Indagación Exploratoria:

Moscoso Zea & Luján Mora (2016) mencionan que la minería de datos es una ciencia que ha tenido gran repercusión a lo largo de los años en el área computacional; sin embargo, con el paso del tiempo y con el aumento exponencial que han tenido diversos campos como: la educación, construcción, salud, etc., se ha visto necesario fusionar el área tecnológica con otros campos, estableciendo relaciones que generen conocimiento para la toma de decisiones y la predicción.

El proceso de minería de datos cuenta con varias etapas las cuales son fundamentales para la obtención de resultados adecuados, por lo que es importante contar con un modelado de datos, es decir, convertir los datos en información para realizar un análisis y generar conocimiento. Todo proceso de minería de datos inicia con la recolección de datos; luego, se procede a la limpieza, que consiste en clasificar en base a caracteres o condiciones, datos (Snee, 2015). Hernández Orallo (2004) propone varias etapas en el proceso de minería de datos como son: recolección, preparación, almacenamiento, limpieza y transformación de datos, y por último define el grado de impacto que tiene la minería de datos para la predicción de eventos relacionados con la información tratada.

Para extraer y determinar patrones de relación entre los datos se hace uso de una serie de métodos a través de algoritmos como: regresión lineal, regresión logística, de asociación, lógica difusa, árboles de decisión, redes neuronales, entre otros; esto permite obtener conocimiento histórico y predictivo que apoyan la toma de decisiones (Riquelme Santos & Otros, 2006). Por ejemplo, Sahafizadeh & Ahmadi (2009) utilizan la minería de datos para predecir los días polvorientos basándose en datos de 53 años recolectados en la ciudad de Bushehr – Irán. Para el estudio se analizaron muestras del índice de polvo durante tres meses por cada año, tomando en cuenta también los índices de presión del aire y la humedad del mismo. Además, dentro de su trabajo analizan la predicción de los datos para conocer la variación mensual de polución del aire.

Dentro del estudio se utilizó el algoritmo K-means y un árbol de decisión para el análisis de los datos.

En la investigación realizada por J. & K. (2013) se aplican técnicas de minería de datos para analizar la calidad del aire en la ciudad de Chennai – India, se ha analizado que el aumento en la congestión vehicular repercute directamente en la salud de las personas, influyendo directamente en enfermedades como el asma o la bronquitis. La investigación indica que se encuentran patrones entre variación de gases producidos por el tráfico y por factores meteorológicos. Se analizaron variables atmosféricas como: Monóxido de Carbono (CO), Dióxido de Carbono (CO₂) y Dióxido de Nitrógeno (NO₂) y variables meteorológicas como la presión del aire. Se analiza también patrones de relación entre dos grupos de variables y el impacto que tienen en la salud de las personas; en este estudio se utiliza el algoritmo K-means y se realiza una comparación con otros dos algoritmos como son DBSCAN y Ward. Se concluye que para un correcto análisis de datos es importante que previamente se analice el número correcto de clusters para el tratamiento de la información.

La calidad del aire repercute directamente en la salud de las personas. Souza & Rabelo (2015) analizaron la relación que existe entre las enfermedades respiratorias y el índice de crecimiento vehicular en el país, el estudio abarco tres ciudades de Brasil como son: Curitiba, Araucaria y Colombo. El estudio indicó que durante los últimos veinte años el incremento de vehículos en la ciudad se ha magnificado llegando al punto de que nueve de cada diez personas poseen un vehículo. Mediante la minería de datos se analizó el índice de contaminación del aire y de la misma forma se analizaron los datos recolectados en busca de patrones que indiquen este comportamiento, En el estudio se utilizaron las siguientes variables atmosféricas: Monóxido de Carbono (CO), Ozono (O₃), Dióxido de Nitrógeno (NO₂). Por ello, se debe tomar en cuenta la

creciente gravedad de la contaminación ambiental en países de todo el mundo y que cada vez demanda más atención por parte de entidades regulatorias de todo el planeta (Ming & Otros, 2015). Otro estudio realizado en la ciudad de Chiapas – México indica los diferentes tipos de correlaciones entre variables atmosféricas y las variables meteorológicas. Se presenta particularidades propias del entorno, en las cuales se describe principalmente una relación entre la radiación solar, temperatura y el ozono. También se indica una relación directa entre la humedad relativa y el material particulado, dando la posibilidad de predecir temporadas de lluvia o sequía.

De acuerdo a naturaleza cambiante de las variables tanto atmosféricas como meteorológicas se buscan patrones que muestren dichos comportamientos referentes a los dos grupos de variables presentes (Herrera, & Otros, 2010).

En nuestro medio la calidad del aire se ha visto afectada por la gran cantidad de empresas alrededor de la ciudad y por el incremento vehicular, esto causa que la calidad del aire se deteriore.

Este efecto produce reacciones en el cuerpo humano, en casos particulares como el del dióxido de nitrógeno (NO_2) que en elevadas concentraciones produce irritación a los alveolos e incrementa el riesgo de infecciones pulmonares. La ciudad de Cuenca-Ecuador tiene implementado un sensor que mide datos atmosféricos por minuto. Los contaminantes atmosféricos que se miden son:

material particulado, dióxido de azufre (SO_2), ozono (O_3), dióxido de nitrógeno (NO_2), partículas sedimentales. También datos meteorológicos como: precipitación, radiación solar global, velocidad y dirección del viento, temperatura y humedad relativa. Los datos se presentan con mayor relevancia en el centro de la ciudad y de sectores alejados como: el sector industrial; la evaluación de la calidad del aire estuvo a cargo de la institución pública Cuenca Aire, y debido al cese de funciones dicho estudio paso a cargo de la empresa EMOV-EP. Esta institución tiene como objetivo analizar constantemente la calidad del aire dentro de la ciudad de acuerdo a guías de la

Organización Mundial de la Salud (OMS) así como de predecir el comportamiento de la contaminación a largo plazo (Red de Monitoreo de Calidad del Aire de Cuenca - EMOV_EP, 2016).

Las técnicas de minería de datos en el tratamiento de la información son capaces de encontrar patrones de relación entre variables. A lo largo de los años uno de los principales usos de la minería de datos ha sido el estudio de información relacionada con el medio ambiente, con el uso de estas técnicas se ha logrado predecir cambios en la calidad del aire en varios países. Sin embargo, dentro de la ciudad de Cuenca solo se han realizado estudios para encontrar patrones en relación de variables atmosféricas, por lo tanto, en contribución al estudio realizado, se procederá a identificar patrones que indiquen una relación entre variables atmosféricas y variables meteorológicas. El proyecto generará patrones de relación en base a la información tratada y generará conocimiento a partir de dicha información.

2.6 Objetivo general:

- Aplicar técnicas de minería de datos en el análisis de contaminantes atmosféricos y variables meteorológicas, para establecer patrones de relación.

2.7 Objetivos específicos:

1. Indagar sobre los proyectos que se han emprendido en la captura, proceso y análisis de datos que indiquen la relación o patrones referentes a variables atmosféricas y meteorológicas.
2. Elaborar un marco teórico referente a la técnica estadística de correlación y a la técnica de minería de datos de clusterización aplicable al problema.
3. Identificar los patrones relevantes que determinen la relación entre las variables de contaminación atmosférica y variables meteorológicas.

4. Analizar los resultados obtenidos de la aplicación de las técnicas seleccionadas y generar conocimiento.

2.8 Metodología:

Se estudiará los fundamentos teóricos de la minería de datos sustentado en una búsqueda de fuentes de carácter científico como son: revistas científicas, conferencias, libros y repositorios de universidades que brinden información certera relacionada al tema, se seleccionaran las fuentes bibliográficas más relevantes y se creará un documento que sirva como base y construir el estado del arte.

Será necesario aplicar técnicas de minería de datos en la información recolectada por la estación de monitoreo de la ciudad de Cuenca – Ecuador y experimentar sobre estos grupos de datos, identificando de patrones de relación. Se aplicará la técnica de minería de datos denominada clusterización, en la cual se procederá a analizar y agrupar datos de acuerdo a un criterio, se evaluará el uso del algoritmo K-MEANS, en el cual se agrupan los datos de acuerdo al valor medio más cercano en relación de los mismos. Luego de analizar los datos por grupos, se identificarán patrones mediante la técnica estadística de la correlación. En este trabajo se analizará la información relacionada con datos de variables atmosféricas y meteorológicas como son: Temperatura del aire, Humedad Relativa, Velocidad del Viento, Radiación Ultravioleta (UVA).

2.9 Alcances y resultados esperados:

El alcance de este proyecto determinará patrones de relación entre variables atmosféricas y variables meteorológicas, con la finalidad de generar conocimiento aplicable a la predicción. Además, se buscará justificar el uso de la minería de datos para el estudio de información relacionada con el medio ambiente y su uso aplicable a nuestro entorno local.

2.10 Supuestos y riesgos:

Riesgos	Probabilidad	Alternativas de solución
Datos irrelevantes que distorsionan los resultados de la investigación.	Alta	Realizar una depuración de datos previo al análisis.
Capacidad de procesamiento baja que distorsionaría los resultados.	Media	Utilización de más equipos para mejorar el procesamiento
Insuficiente consistencia y variabilidad de datos recolectados que respalden los resultados.	Media	Mejorar las fuentes de datos
Tiempo insuficiente para terminar la investigación	Media	Solicitar prórroga para la entrega final del trabajo

2.11 Esquema tentativo:

Capítulo 1: Introducción

Capítulo 2: Estado del Arte

Capítulo 3: Aplicación de minería de datos

Capítulo 4: Análisis de resultados

Conclusiones

Bibliografía

2.12 Cronograma:

[illegible]

2.13 Referencias:

Hernández Orallo, J. (2004). *Introducción a la minería de datos*. Madrid : Pearson Educación.

Herrera, S., Margulis, R., & Manzanilla, A. (2010). Statistical study of the correlation between atmospheric pollutants and meteorological variables in northern Chiapas, Mexico. *Universidad y ciencia*, 65-80.

J., C., & K., K. (2013). Analysis of hard clustering algorithms applicable to regionalization. *Information & Communication Technologies (ICT)*.

Ming , H., Tao, Z., & Jingyang , W. (2015). A new air quality forecasting model using data mining and artificial neural network. *Software Engineering and Service Science (ICSESS)*.

Moscoso Zea, O., & Luján Mora, S. (2016). Educational data mining: An holistic view. *Information Systems and Technologies (CISTI)*.

Red de Monitoreo de Calidad del Aire de Cuenca - EMOV_EP. (2016). *Informe de la Calidad del Aire de Cuenca*. Cuenca.

Riquelme Santos , J. C., Ruiz , R., & Gilbert, K. (2006). Minería de Datos: Conceptos y Tendencias. *Inteligencia Artificial: Revista Iberoamericana de Inteligencia Artificial*, 11-18.

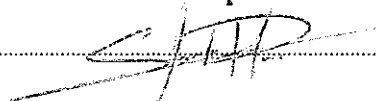
Sahafizadeh, E., & Ahmadi, E. (2009). Prediction of Air Pollution of Boushehr City Using Data Mining. *Environmental and Computer Science*.

Snee, R. D. (2015). A Practical Approach to Data Mining: I Have All This Data: What Should I Do? *Quality Engineering*.

T. Souza, F., & S. Rabelo, W. (2015). A Data Mining Approach to Study the Air Pollution. *Natural Computation (ICNC)*.

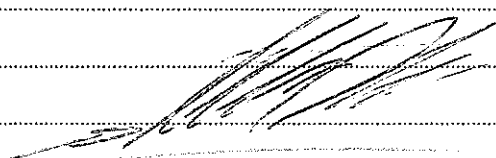
2.14 Anexos: para casos en los que se requiera respaldar el proyecto.

2.15 Firma de responsabilidad (estudiante).



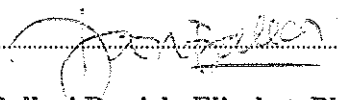
Pedro Sebastián Andrade Durazno

2.16 Firma de responsabilidad (director sugerido)



Ing. Marcos Orellana Cordero

2.17 Firma de responsabilidad (Asesor Metodológico)



Ballari Daniela Elisabet, PhD

2.18 Fecha de entrega: