https://doi.org/10.37815/rte.v34n3.952
Artículos originales

Extracción de conocimiento mediante ventanas de tiempo en variables atmosféricas

Knowledge extraction through time windows on atmospheric variables

Nicolás Alvarez[1] https://orcid.org/0000-0002-4649-6844, Jaime Panata[1] https://orcid.org/0000-0001-6233-4899, Marcos Orellana[1] https://orcid.org/0000-0002-3671-9362, Priscila Cedillo[2] https://orcid.org/0000-0002-6787-0655, Jorge Luis Zambrano-Martinez[1] https://orcid.org/0000-0002-5339-7860, Juan Fernando Lima[1] https://orcid.org/0000-0003-3500-3968

[1]*Universidad del Azuay*, Cuenca, Ecuador
nicolas.alvarez@es.uazuay.edu.ec, panatta3004@es.uazuay.edu.ec, marore@uazuay.edu.ec, jorge.zambrano@uazuay.edu.ec, flima@uazuay.edu.ec

[2]*Universidad de Cuenca*, Cuenca, Ecuador
priscila.cedillo@ucuenca.edu.ec

Enviado: 2022/07/01
Aceptado: 2022/09/16
Publicado: 2022/11/30

Resumen

La industrialización y el rápido crecimiento de zonas urbanas aumentan alarmantemente la presencia de contaminantes atmosféricos. Estos contaminantes afectan la calidad de vida de las personas y se crea una oportunidad de estudio para determinar su comportamiento atmosférico y la relación entre variables meteorológicas presentes en el ambiente. Previo a esto, se aplicaron ventanas rodantes de tiempo para eliminar datos anómalos. A continuación, se identificaron variables y se segmentaron los datos a través del algoritmo X-means. También, dos clústeres que representan las relaciones entre pares de variables y la temporalidad de las ventanas de tiempo. Como resultado, se encontró una correlación inversa de -0,78 entre las variables de ozono y punto de rocío dentro de las horas de la jornada laboral.

Palabras clave: clúster, correlación, X-means, contaminantes atmosféricos, variables meteorológicas, ventanas de tiempo.

Como citar: Alvarez, N., Panata, J., Orellana, M., Cedillo, P., Zambrano-Martinez, J. L. & Lima, J. F. (2022). Extracción de conocimiento mediante ventanas de tiempo en variables atmosféricas. *Revista Tecnológica - Espol, 34(3), 72-83.* http://www.rte.espol.edu.ec/index.php/tecnologica/article/view/952

Abstract

Industrialization and the rapid growth of urban areas are alarmingly increasing the presence of air pollutants. These pollutants affect the quality of life of people and present an opportunity for study is created to determine the atmospheric behavior and relationship between meteorological variables present in the environment. Prior to this, rolling windows of time were applied to remove anomalous data. Next, variables were identified, and the data was segmented through the X-means algorithm. Also, two clusters that represent the relationships between pairs of variables and the temporality of the time windows. As a result, an inverse correlation of -0.78 was found between the ozone and dew point variables within the hours of the working day.

Keywords: clustering, correlated, X-means, atmospheric pollutants, meteorological variables, time windows.

Introducción

La actividad humana ha contribuido significativamente a la contaminación del aire, mediante la industrialización y el crecimiento acelerado de zonas urbanas (Parker, 1983). Esta problemática la aborda la comunidad científica con el objetivo de controlar las fuentes de alta emisión de contaminantes. El incremento de la contaminación del aire genera exponencialmente consecuencias en el ambiente, y, por ende, el cambio climático a nivel mundial (Chong et al., 2019), (Clifford et al., 2016), (Franchini et al., 2016).

La contaminación del aire es entendida como la presencia de sustancias químicas o material particulado (PM) en el medio ambiente. Este último, por su cantidad o su composición química causa perjuicio a los seres humanos y otros organismos vivos impidiendo el funcionamiento de procesos naturales (Goel et al., 2012). Así mismo, Brook, et al. (2010) demuestran que el contenido de PM contribuye al aumento de las tasas de mortalidad y disminuye la esperanza de vida de la población. Por otro lado, expertos declaran que, el índice alto de contaminación del ambiente, está relacionado con implicaciones psicológicas. Así lo demuestran Zhang et al. (2017), donde afirman que la existencia de contaminantes en el aire reduce significativamente la felicidad hedónica y aumenta los síntomas de depresión.

Además, la contaminación del aire se ve afectada por diferentes elementos, entre ellos se destacan los contaminantes atmosféricos, como: el dióxido de carbono (CO_2), óxidos de azufre (SO_x), Compuestos Orgánicos Volátiles (COV), material particulado grueso (PM_{10}) grueso y material particulado fino ($PM_{2.5}$). Estos contaminantes conllevan a efectos nocivos y una inestabilidad de la calidad del aire y del medio ambiente (Brook, et al.,2010), (Clifford et al., 2016), (Zhang et al., 2017). Por lo tanto, es primordial comprender la influencia de los contaminantes atmosféricos en la calidad del aire, ya que proporcionan información relevante para quienes tratan de mitigar este problema a través de desarrollos de programas y políticas de salud pública (Simioni & United Nations, 2003).

La predicción de la cantidad de sustancias contaminantes en el aire, puede ser un soporte en tareas para mitigar el problema de la contaminación. Una forma de realizar esto es la definición de indicadores que determinen el grado de contaminación del aire, mediante una implementación técnicas de extracción de conocimiento. Mediante el uso de estadística clásica, correlación de variables y ventanas temporales Orellana, et al. (2021) explican la relación existente entre los contaminantes atmosféricos y las variables meteorológicas. La relación entre un par de variables (correlación entre dos variables) es representada mediante una ecuación (Sperman o Pearson) dependiendo de la distribución de los datos. Y, por otro lado, las ventanas de tiempo rodantes, que es una técnica que permite dividir una serie de tiempo en intervalos o

secciones para profundizar el análisis en dicho intervalo, ambas ya utilizadas en el estudio mencionado. Sin embargo, es posible incluir técnicas de aprendizaje de máquina para que la asociación de los datos dentro de las ventanas de tiempo sea más fuerte.

Existen diversas técnicas para detectar asociaciones entre variable dentro del aprendizaje automático. Por un lado, técnicas supervisadas predicen el comportamiento mediante la extracción de información de un gran conjunto de datos con técnicas de minería de datos. En este campo, los modelos de predicción necesitan una entrada correctamente etiquetada con la salida esperada; es decir, construir patrones que deriven en una salida esperada (Russell & Norvig, 2003). Otro grupo de estas técnicas plantean descubrir las relaciones existentes sobre un gran conjunto de datos, pero sin la información de salida como en el caso de los sistemas no supervisados (Russell & Norvig, 2003).

Sin embargo, cuando se dispone de datos sin etiquetar, se da paso a otro tipo de técnicas, las no supervisadas, siendo las técnicas de agrupamiento o clusterización las de mayor demanda (Paulose et al., 2018), (Fränti & Sieranoja, 2018). La clusterización es una técnica de aprendizaje automático no supervisado que encuentra patrones y conocimiento dentro de un grupo de datos sin una etiqueta. Mediante la relación de los datos se generan grupos a base de la distancia que tienen entre sí, y así forman conglomerados con similitud en los atributos seleccionados. El resultado obtenido en esta técnica es un modelo de comportamiento de la información capaz de predecir similares situaciones (Russell & Norvig, 2003). Lo cual, permite el descubrimiento y asignación de etiquetas a los datos que, de otra manera, no se podría identificar.

La presente investigación se aplica la técnica no supervisada de clusterización para generar conocimiento a partir de las relaciones existentes entre los contaminantes atmosféricos y las variables meteorológicas del ambiente de la ciudad de Cuenca, Ecuador. Este estudio está estructurado de la siguiente manera: la Sección II expone trabajos relacionados con métodos similares, la Sección III presenta la metodología utilizada para realizar esta investigación, la Sección IV describe los resultados obtenidos, la Sección V presenta las conclusiones de esta investigación y sus trabajos futuros.

Trabajos relacionados

Existen diversos trabajos que estudian tanto a los contaminantes atmosféricos, como a las variables meteorológicas, permitiendo extraer información para diferentes propósitos como: la creación y mejoramiento de sistemas predictivos o para solventar falencias a la hora de obtener los datos. A continuación, se presenta los trabajos que han vinculado diversas técnicas de aprendizaje de maquina en búsqueda de la mejora de resultados haciendo parte de nuevos sistemas predictivos.

Los autores Lan Yuxiao y Dai Yifan (2020), predicen la calidad del aire para una estación de monitoreo, considerando datos de calidad del aire, datos meteorológicos y datos de circulación vehicular para construir un modelo de predicción tradicional desde dimensiones espaciales y temporales. Además, los autores plantean un enfoque de predicción de calidad de aire mediante el uso de un modelo de optimización espacio-temporal o STOM por el abreviado de space-time optimization model. El mismo se basa en una red neuronal de memoria a corto plazo o LSTM (long short-term memory). El estudio optimiza el tamaño de la ventana de tiempo y considera la dispersión de los contaminantes en el aire, lo que mejora la precisión en las predicciones.

En el trabajo de Othman et al., (2017) se utilizaron datos de la ciudad de Putrajaya – Malaysia, entre los años de 2005 a 2012, para examinar las relaciones entre la temperatura y el Ozono (O_3). Los investigadores utilizan técnicas de agrupamiento y reglas de asociación para superar los resultados que hasta ese momento se obtenían con clásicas técnicas estadísticas. En este artículo, se afirma que los métodos de pronóstico estadísticos presentan defectos como la necesidad de analizar los datos con anterioridad y no utilizar datos sintéticos, dado que éstos generan resultados incorrectos si se los compara con la utilización de datos reales.

De manera similar Gu, et al., (2018) realizan un predictor recurrente de calidad de aire (RAQP). Según los autores, el RAQP es el primero en aplicar estrategias recurrentes para la predicción del aire. Este predictor fue construido a partir de la combinación del Support Vector Regression (SVR) y un framework presentado por el mismo autor en este artículo. También se utilizaron estrategias para introducir ruido en los datos y mejorar la generalización de los módulos de regresión.

En la investigación realizada por Yang et al., (2009) se demuestra un descubrimiento en las reglas de asociación basadas en técnicas evolutivas, con el fin de obtener relaciones entre series temporales correlacionadas. De esta manera, dicha contribución propone un algoritmo genético que determina los intervalos, luego el algoritmo forma reglas sin discretizar atributos y por último permite la superposición de las regiones cubiertas por las reglas. Este algoritmo ha sido probado en series temporales climáticas del mundo real, como la temperatura, el viento y el ozono.

Otro artículo presentado por Otham, et al. (2016) proponen la obtención de un pronóstico de lluvia preciso, con la ayuda de la extracción de datos para una predicción con mayor precisión de las precipitaciones de lluvia. Los investigadores utilizan la minería de datos desarrollando una distribución del modelo de pronóstico de lluvias basada en una representación de datos simbólicos usando Piecewise o regresión lineal por partes. Ésta es una forma de regresión que permite ajustar múltiples modelos lineales a los datos para diferentes rangos de la variable analizada. Así, cada dato almacenado de las precipitaciones, se presenta gráficamente; pues se limita a descubrir patrones comprensibles, debido a que son visualizados con base en los valores de la serie temporal.

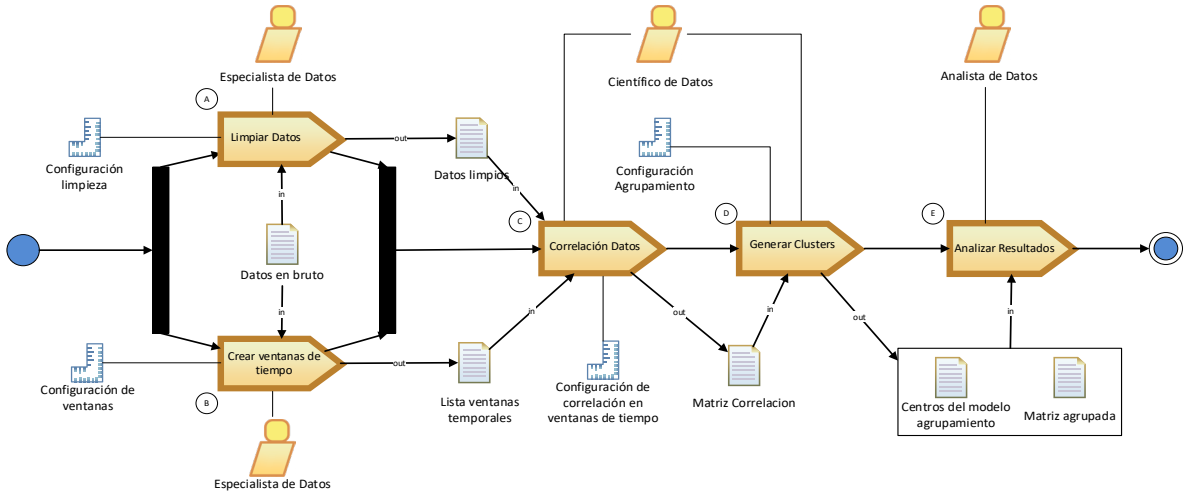
En los artículos antes mencionados, los autores utilizaron en conjunto técnicas de relevancia, sobresaliendo: redes neuronales, técnicas de asociación, y SVR; en el cual todas las variables son parte del algoritmo y generan los resultados prometedores. Sin embargo, es necesario profundizar en el análisis entre pares de variables para detectar asociaciones específicas. Por lo tanto, en este estudio se presenta el análisis específico de pares de variables siendo estos pares las entradas mínimas para los algoritmos.

Metodología

En este apartado se presenta una metodología para extraer conocimiento relevante sobre la interacción de los contaminantes atmosféricos y las variables meteorológicas en diferentes horas del día. La metodología propuesta en esta investigación consta de las siguientes actividades: A) Limpiar datos, B) Crear ventanas de tiempo, C) Correlacionar datos, D) Generar clústeres, E) Análisis de la generación de clústeres. La limpieza de datos y la creación de ventanas de tiempo se realizan para manipular los datos sin generar alteraciones en los resultados. La Figura 1 representa las etapas de la metodología propuesta en un diagrama Software & Systems Process Engineering Metamodel (SPEM) 2.0.

Figura 1

Diagrama de procesos en SPEM 2.0



Los datos utilizados para presentar la instancia de la metodología propuesta, representan datos de variables meteorológicas y contaminantes atmosféricos, fueron proporcionados por el Instituto de Estudios de Régimen Seccional del Ecuador (IERSE). Los datos correspondientes al año 2018 fueron recolectados en la ciudad de Cuenca, Ecuador con un intervalo de un minuto y corregidos mediante la presión barométrica local del aire. La Tabla 1 muestra un resumen general con los datos estadísticos descriptivos de las variables meteorológicas y los contaminantes atmosféricos, estos valores son equivalentes a cero, es decir, lecturas bastante pequeñas, las mismas que truncadas a dos dígitos generan este valor.

Tabla 1

Estadísticas descriptivas de las variables recopiladas en Cuenca, Ecuador

VARIABLE	UNIDAD	N	MEDIA	DS	MÍNIMO	25%	MEDIANA	50%	MÁXIMO
Contaminantes del Aire									
Ozono(O3)	ug/m3	42291,00	30,89	25,13	0,00	9,23	26,85	45,57	140,56
Monóxido de Carbono (CO)	ug/m3	42521,00	0,86	0,38	0,26	0,60	0,79	1,00	3,60
Dióxido de Sulfuro (SO2)	ug/m3	42374,00	7,92	8,66	2,85	2,85	4,44	9,10	88,24
Dióxido de Nitrógeno (NO2)	ug/m3	42394,00	17,35	14,99	6,16	6,16	14,23	24,45	94,81
Material Particulado (PM2.5)	ug/m3	43044,00	14,85	14,20	5,70	5,70	10,20	18,60	156,00
Variables Meteorológicas									
Temperatura del Aire (TEMP)	°C	3936,00	7,44	2,18	-0,60	6,40	7,80	9,00	12,30
Punto de Rocío (PR)	°C	3936,00	7,44	2,18	-0,60	6,40	7,80	9,00	12,30

VARIABLE	UNIDAD	N	MEDIA	DS	MÍNIMO	25%	MEDIANA	50%	MÁXIMO
Variables Meteorológicas									
Velocidad del Aire (VA)	m/s	3936,00	1,59	1,01	0,10	0,80	1,30	2,20	6,00
Precipitación	Mm	3936,00	0,01	0,17	0,00	0,00	0,00	0,00	8,20
Radiación Global	w/m^2	3936,00	189,30	293,64	0,00	0,00	1,00	305,20	1363,00
Radiación UVA	w/m^2	3936,00	12,50	18,55	0,00	0,00	0,25	21,20	74,30
Radiación UVE	w/m^2	3936,00	0,01	0,00	0,00	0,00	0,02	0,01	0,03
Humedad Relativa (HR)	%	3936,00	63,82	24,00	24,00	51,00	67,00	77,00	97,00
Presión Atmosférica (PATM)	hPa	3936,00	751,80	746,40	746,40	750,40	751,60	752,60	876,80

Limpiar datos

Se depuran lo datos para prepararlos para el proceso de minería de datos que se describen en las secciones siguientes. El proceso de limpieza de datos mejora la calidad de los datos al eliminar y los fragmentos que disminuyen su usabilidad. Los pasos involucran la eliminación de datos inexactos, incompletos e irrelevantes para extraer su máximo beneficio (Juneja & Das, 2019). Como resultado se obtiene una colección de datos depurados para la aplicación de técnicas de extracción de conocimiento.

Para este estudio, se seleccionaron las variables descritas en la Tabla I; a excepción de radiación global y precipitación. Los que presentaban largos tramos donde los datos estaban formados únicamente por valores de cero. De acuerdo a la fórmula de la correlación (Lind et al., 2012) no se puede tener una variable cuyo vector está formado únicamente por valores en cero, pues resultaría en una indeterminación causada por una división por cero.

Similar al estudio de Orellana et al. (2021) se encuentran patrones durante las horas del día, sin considerar horarios nocturnos. Los datos se filtraron con base en la selección de horas, entre 05:00 y 20:00. En este rango de horas, se encuentran mediciones completas de las variables seleccionadas y representan un comportamiento coherente de las mismas.

Crear ventanas de tiempo

Las ventanas rotativas de tiempo se caracterizan por suavizar el comportamiento de una variable y unificar días en una secuencia de tiempo única (Orellana et al., 2021). Para realizar una buena correlación y agrupamiento adecuado, fue necesario corregir problemas de datos anómalos, los mismos que se presume se produjeron durante la captura de datos por parte de los sensores. Esta corrección se realiza mediante el uso de ventanas de tiempo con un rango de 10 minutos. Los rangos de 10 minutos fueron promediados, entregando un conjunto de datos suavizado, pero sin perder el significado de los mismos. Es importante recalcar que, las ventanas de tiempo no solo se realizaron dividiendo los datos en intervalos de 10 minutos, sino que se dividieron también en intervalos de 60 minutos como límite superior en el ancho total de las ventanas, antes de iterar nuevamente en los intervalos.

Correlación de datos

El análisis de correlación es el grupo de técnicas para medir la asociación entre dos variables. El coeficiente de correlación fue creado por Karl Pearson, también llamado coeficiente de Pearson. El mismo que indica la fuerza de la relación entre dos conjuntos de variables, logrando ser de tipo intervalo o razón. Su rango oscila entre -1 y 1, lo que significa que una correlación de -1 o 1 es una relación perfecta. Es decir, si el coeficiente es 1, significa una relación lineal directa o positiva, y si es -1 se entiende como una relación indirecta o negativa. Sin embargo, el coeficiente 0 indica que no existe correlación lineal aparente, así como los valores muy cercanos a este indican una correlación débil (Murray & Larry, 2009).

Este estudio creó pares de variables para demostrar la correlación de contaminantes atmosféricos y variables meteorológicas, donde cada par de variables se encuentra en una respectiva ventana de tiempo, lo que permite entender cómo se relacionan los pares y cómo fueron evolucionando a lo largo del día.

El objetivo de este proceso fue la preparación de los datos para generar clústeres como se muestra en la Tabla 2. Cada columna de la tabla representa un par de correlación y los índices de las filas están dados por la ventana de tiempo que se utilizó en su generación.

Tabla 2

Matriz de correlación y ventanas de tiempo

VENTANA DE TIEMPO	(O3, TEMP)	(O3, HR)	…	(PM2.5, UVA)	(PM2.5, UVE)
5:00	0,3022	-0,1170	…	0,1054	0,4478
5:10	0,4188	-0,2065	…	0,0721	0,4365
5:20	0,3461	-0,2226	…	0,1441	0,4618
…	…	…	…	…	…
19:30	-0,2984	-0,1851	…	0,2872	0,4500
19:40	-0,4761	-0,0140	…	0,2072	0,4261
19:50	-0,3386	-0,1403	…	0,2862	0,4047

Generación de clústeres

Para generar patrones de relación entre pares de variables en diferentes horarios, se utilizaron las técnicas de clusterización k-means y x-means, mismas que se describen a continuación.

K-means

Es el algoritmo de agrupación más utilizado. Este permite trabajar con una gran cantidad de datos numéricos de alta dimensionalidad y es capaz de proporcionar un método eficaz para la clasificación de datos similares (Fränti & Sieranoja, 2018). K-means asocia todos los objetos con características semejantes, mediante la minimización de las distancias entre cada objeto y un centroide de grupo. Para la minimización, se hace uso de la ecuación de la distancia euclidiana (Fränti & Sieranoja, 2018), la cual está expresada de la siguiente manera.

$$d(P,Q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$$

El algoritmo selecciona los puntos de datos aleatorios como centroides iniciales, para luego mejorar reiteradamente la solución en dos pasos denominados asignación y actualización (Yu et al., 2018).

X-means

El algoritmo es una variación de k-means, donde se repite la aplicación de k-means hasta optimizarlo, esto se logra al alcanzar el número de clústeres eficientes y optimizar el criterio de información bayesiano (BIC). El objetivo del algoritmo es calcular el número de clústeres dinámicamente, para ello, utiliza el límite superior e inferior proporcionado por el usuario. Si, el límite superior proporcionado es $k \geq k_{max}$, se considera que el modelo tiene la mejor puntuación durante la búsqueda, caso contrario se regresa a la iteración inicial hasta que la condición llegue a cumplirse (Kumar & Krishan Wasan, 2010).

La generación de clústeres se realiza mediante la aplicación del algoritmo X-means, consiguiendo así la mejor aproximación posible en cada clúster. Para la aplicación del algoritmo se consideró primordial que la cantidad de clústeres esté entre un mínimo de dos y un máximo de cincuenta. En los clústeres generados, cada par de variables se encuentran representadas por un punto en el espacio vinculadas con todas las correlaciones de los pares de variables dentro de una ventana de tiempo específica. Por tal razón, se consideró importante incluir las ventanas de tiempo como una variable espacial más, sin embargo, por la naturaleza categórica de la variable, se la transformó en diversas variables indicadoras/ficticias.

Análisis de la generación de clústeres

Los resultados de aplicar las técnicas de clusterización, dieron como resultado una matriz con los grupos generados, sus centroides y el grupo al que pertenecen los datos. Tal como se indicó en la sección anterior, los centros de los clústeres formados por el algoritmo k-means y x-means, constituyen la media de todos los datos que conforman el clúster. Por este motivo, se estudiaron y compararon los centros de cada clúster, para determinar el contenido de cada agrupación y sus particularidades. Este método permitió establecer la relación de las ventanas temporales en el día, comprender el comportamiento de los pares y examinar los parentescos que se puedan presentar.

Resultados y Discusión

Debido a la estructura que presentan los centros resultantes descritos en el apartado anterior, fue posible dividirlos y estudiarlos en dos partes: A) Análisis ventanas temporales, y B) Análisis de correlación de pares.

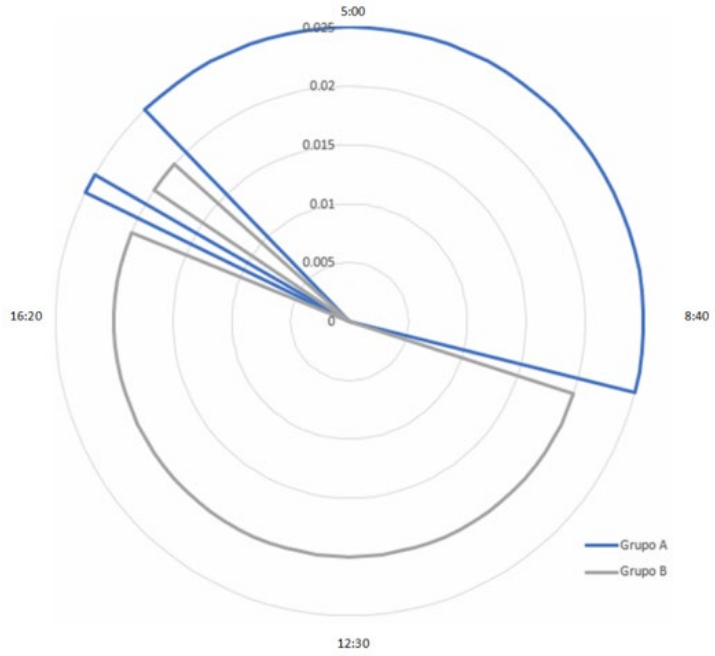
Análisis de ventanas temporales

Una parte de cada centro estaba compuesta por un grupo de variables ficticias, para evidenciarlas, se analizó la correspondencia de cada ventana temporal con uno de los grupos identificados, tal como se presenta en Figura 2.

En la Figura 2, se puede observar que, al clúster A, le pertenecen las ventanas desde las 18:10 hasta las 9:20 y de 17:20 hasta 17:30. Así mismo, al clúster B, le pertenecen las ventanas desde 9:30 hasta las 17:10 y desde las 17:40 hasta las 18:00. Por lo tanto, el clúster B engloba principalmente las horas laborales del día. Mientras que desde las 17:20 hasta las 18:00 la clasificación se invierte. Este espacio de tiempo presume que corresponde a la hora en que los trabajadores culminan su jornada laboral.

Figura 2

Ventanas en clústeres A y B

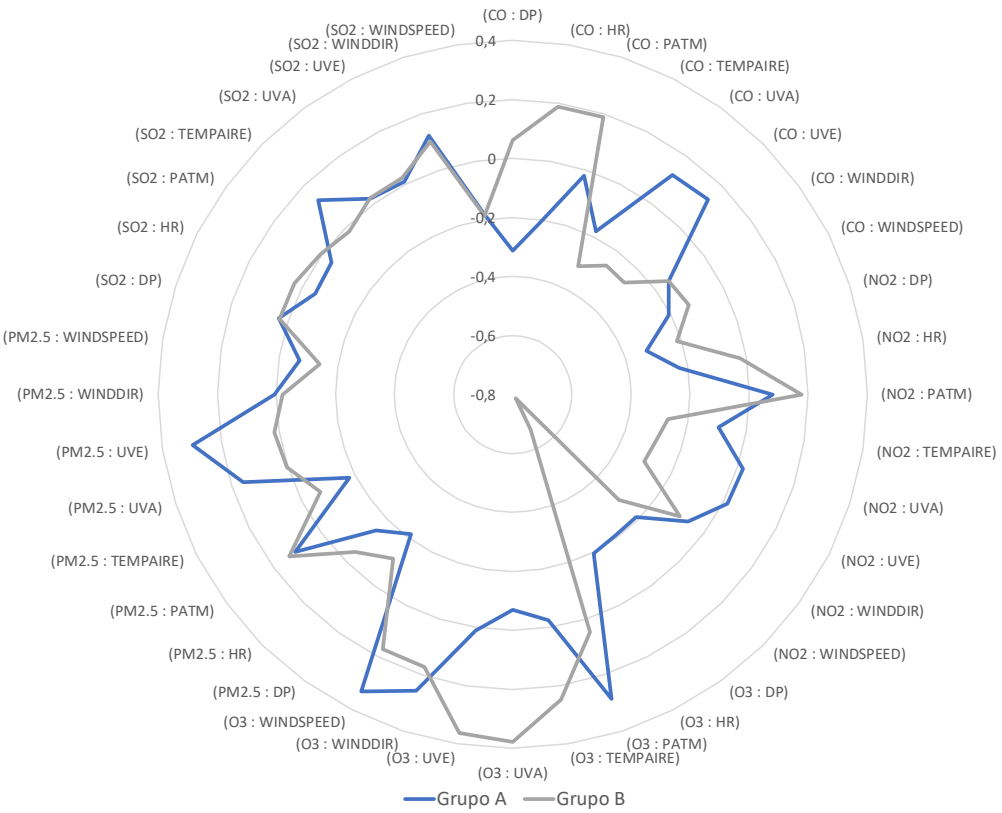


Análisis de correlación de pares

Una vez comprendido el tiempo que engloba para cada uno de los clústeres al que pertenece la media de correlación de pares en cada grupo, se realiza la diferenciación de correlación en los marcos de tiempo de cada par, como se puede observar en la Figura 3.

Figura 3

Correlación de pares en clústeres A y B



En ambos clústeres, se aprecia como algunos de los valores de correlación oscilan entre el rango de 2 a -2. Lo que significa que la mayoría de estos pares no presentan una correlación significativa en cualquiera de los dos clústeres. Sin embargo, en el clúster B se pueden rescatar tres puntos altamente significativos: la relación entre el ozono y punto de roció que llega hasta -0,78 (correlación fuerte); la de ozono y UVA con 0,38 (correlación débil); y la del ozono con el UVE 0,36 (correlación débil).

Es importante recalcar que la relación más alta entre un contaminante y una variable atmosférica se da entre el ozono y punto de rocío, esto se refleja en la alta correlación presentada dentro del clúster B (horas laborales del día), ya que su valor dentro del grupo A es de -0.2 (una diferencia significativa cuando se revisó la correlación).

Conclusiones

La actividad humana y la contaminación del aire en los últimos años ha puesto a la comunidad científica en la tarea de encontrar las fuentes de contaminación. Esta búsqueda ha usado técnicas de aprendizaje supervisado, que extrae información mediante minería de datos aplicadas a grandes volúmenes de datos, permitiendo predecir de esta manera el comportamiento y relaciones de los contaminantes presentes en el aire. Sin embargo, para encontrar estas relaciones mediante técnicas de aprendizaje supervisado, se necesita etiquetar las variables a analizar, dejando a la deriva a las variables que no pueden ser etiquetadas.

El trabajo expuesto, demostró que la aplicación de técnicas no supervisadas, como es el caso de la clusterización, permite identificar relaciones y patrones de comportamiento en las variables analizadas. Para efectos prácticos, en este trabajo se tomó como referencia los datos de variables meteorológicas y contaminantes atmosféricos de datos del IERSE, del año 2018 y seleccionados en un rango entre 05:00 y 20:00.

Con la finalidad de corregir problemas de datos anómalos y suavizar los datos a tratar, se utilizó ventanas de tiempo de 10 y 60 minutos, permitiendo, eliminar datos erróneos producidos al momento de la toma de medidas. La técnica de aprendizaje no supervisado para encontrar la relación entre las variables analizadas, se realizó mediante aplicación de la técnica de clusterización x-means, la cual, dio como resultado dos clústeres, clúster A y clúster B.

El comportamiento del clúster A es inversamente proporcional al clúster B, en donde el clúster B representa las horas laborales del día y permite observar una correlación significativa con las variables ozono y punto de rocío, que fluctúa dependiendo del tiempo. Por lo tanto, la limitación que se presenta actualmente son las pocas estaciones automáticas de monitoreo para el registro de contaminantes atmosféricos en Ecuador, lo que imposibilita tener una mayor granularidad en los resultados más precisos.

La relación de las variables encontradas mediante los clústeres A y B, presenta una oportunidad para próximos análisis. En trabajos futuros, se puede encontrar la relación de los clústeres con las jornadas u horarios de trabajo. Por otro lado, el análisis futuro no se ve limitado a un mismo escenario o tamaño en ventanas de tiempo, es decir, la actual metodología se puede aplicar en diferentes entornos o rangos de tiempo.

Las variables analizadas en este trabajo, podrían ser reemplazadas por la relación entre pares de contaminantes y pares de variables meteorológicas, dando lugar a una nueva propuesta. Así como otro trabajo futuro es estudiar con los datos atmosféricos actuales la aplicación distintos tipos de correlación entre sus variables continuas y representarlos adecuadamente.

Reconocimientos

Los autores desean agradecer al Vicerrectorado de Investigaciones de la Universidad del Azuay por el apoyo financiero y académico, así como a todo el personal de la escuela de Ingeniería de Ciencias de la Computación, y el Laboratorio de Investigación y Desarrollo en Informática - LIDI.

Referencias

Brook, R. D., Rajagopalan, S., Pope, C. A., Brook, J. R., Bhatnagar, A., Diez-Roux, A. V., Holguin, F., Hong, Y., Luepker, R. V., Mittleman, M. A., Peters, A., Siscovick, D., Smith, S. C., Whitsel, L., & Kaufman, J. D. (2010). Particulate matter air pollution and cardiovascular disease: An update to the scientific statement from the american heart association,. *Circulation*, *121*(21), 2331–2378. https://doi.org/10.1161/CIR.0b013e3181dbece1

Chong, K. C., Lau, W. J., Lai, S. O., Thiam, H. S., & Ismail, A. F. (2019). Preparation and Characterization of Chromium Metal Organic Framework for Carbon Dioxide Adsorption. *IOP Conference Series: Earth and Environmental Science*, *268*(1), 012010. https://doi.org/10.1088/1755-1315/268/1/012010

Clifford, A., Lang, L., Chen, R., Anstey, K. J., & Seaton, A. (2016). Exposure to air pollution and cognitive functioning across the life course—A systematic literature review. *Environmental Research*, *147*, 383–398. https://doi.org/10.1016/j.envres.2016.01.018

Franchini, M., Mengoli, C., Cruciani, M., Bonfanti, C., & Mannucci, P. M. (2016). Association between particulate air pollution and venous thromboembolism: A systematic literature review. *European Journal of Internal Medicine*, *27*, 10–13. https://doi.org/10.1016/j.ejim.2015.11.012

Fränti, P., & Sieranoja, S. (2018). K-means properties on six clustering benchmark datasets. *Applied Intelligence*, *48*(12), 4743–4759. https://doi.org/10.1007/s10489-018-1238-7

Goel, A., Ray, S., Agrawal, P., & Chandra, N. (2012). Air Pollution Detection Based on Head Selection Clustering and Average Method from Wireless Sensor Network. *2012 Second International Conference on Advanced Computing Communication Technologies*, 434–438. https://doi.org/10.1109/ACCT.2012.18

Gu, K., Qiao, J., & Lin, W. (2018). Recurrent Air Quality Predictor Based on Meteorology- and Pollution-Related Factors. *IEEE Transactions on Industrial Informatics*, *14*(9), 3946–3955. https://doi.org/10.1109/TII.2018.2793950

Juneja, A., & Das, N. N. (2019). Big Data Quality Framework: Pre-Processing Data in Weather Monitoring Application. *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 559–563. https://doi.org/10.1109/COMITCon.2019.8862267

Kumar, P., & Krishan Wasan, S. (2010). Analysis of X-means and global k-means USING TUMOR classification. *2010 The 2nd International Conference on Computer and Automation Engineering (ICCAE)*, *5*, 832–835. https://doi.org/10.1109/ICCAE.2010.5451883

Lan, Y., & Dai, Y. (2020). Urban Air Quality Prediction Based on Space-Time Optimization LSTM Model. *2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, 215–222. https://doi.org/10.1109/ICAIBD49809.2020.9137441

Lind, D. A., Marchal, W. G., & Wathen, S. A. (2012). *Estadística aplicada a los negocios y a la economía* (15th ed.). McGraw-Hill.

Murray, R. S., & Larry, J. S. (2009). *Estadística* (4th ed.).

Orellana, M., Lima, J.-F., & Cedillo, P. (2021). Discovering Patterns of Time Association Among Air Pollution and Meteorological Variables. In K. Arai (Ed.), *Advances in Information and Communication* (pp. 205–215). Springer International Publishing. https://doi.org/10.1007/978-3-030-73103-8_13

Ostadabbas, S., & Jafari, R. (2010). Spectral Spatio-Temporal template extraction from EEG signals. *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, 4678–4682. https://doi.org/10.1109/IEMBS.2010.5626411

Othman, Z. A., Ismail, N., & Latif, M. T. (2017). Association rules of temperature towards high and low ozone in putrajaya. *2017 6th International Conference on Electrical Engineering and Informatics (ICEEI)*, 1–5. https://doi.org/10.1109/ICEEI.2017.8312438

Othman, Z. A., Risdiyanto Ismail, N., Aziz Hamdan, A., & Mahmoud, S. (2016). *KLANG VALLY RAINFALL FORECASTING MODEL USING TIME SERIES DATA MINING TECHNIQUE*. *92*, 8.

Parker, A. (1983). *Contaminación del aire por la industria* (1st ed.). Editorial Reverté. https://www.reverte.com/libro/contaminacion-del-aire-por-la-industria_91542/

Paulose, B., Sabitha, S., Punhani, R., & Sahani, I. (2018). Identification of Regions and Probable Health Risks Due to Air Pollution Using K-Mean Clustering Techniques. *2018 4th International Conference on Computational Intelligence Communication Technology (CICT)*, 1–6. https://doi.org/10.1109/CIACT.2018.8480232

Russell, S. J., & Norvig, P. (2003). *Artificial intelligence: A modern approach* (2nd ed.). Prentice Hall/Pearson Education.

Simioni, D., & United Nations (Eds.). (2003). *Contaminación atmosférica y conciencia ciudadana*. Naciones Unidas, CEPAL.

Yang, X., Tang, K., & Yao, X. (2009). The Minimum Redundancy – Maximum Relevance Approach to Building Sparse Support Vector Machines. In E. Corchado & H. Yin (Eds.), *Intelligent Data Engineering and Automated Learning—IDEAL 2009* (pp. 184–190). Springer. https://doi.org/10.1007/978-3-642-04394-9_23

Yu, S. S., Chu, S. W., Wang, C. M., Chan, Y. K., & Chang, T. C. (2018). Two improved k-means algorithms. *Applied Soft Computing Journal*, *68*, 747–755. https://doi.org/10.1016/j.asoc.2017.08.032

Zhang, X., Zhang, X., & Chen, X. (2017). Happiness in the air: How does a dirty sky affect mental health and subjective well-being? *Journal of Environmental Economics and Management*, *85*, 81–94. https://doi.org/10.1016/j.jeem.2017.04.001