Caracterización de regiones espacialmente homogéneas de monóxido de carbono en Lima Metropolitana mediante el algoritmo de clustering k-means

José Abel Espinoza Guillen Agramonte , Marleni Beatriz Alderete Malpartida

Universidad Nacional Agraria La Molina, Lima, Perú

Recibido: 16/06/2020 Revisado: 26/07/2020 Aceptado: 20/09/2020 Publicado: 28/01/2021

Resumen

Los análisis estadísticos de series de tiempo o datos espaciales se han utilizado extensamente para examinar el comportamiento de los contaminantes atmosféricos. Debido a que los datos de contaminación del aire comúnmente se recopilan en una vasta área de interés durante un período de tiempo relativamente largo, dichos análisis deben tener en cuenta tanto las características espaciales como las temporales. El objetivo de este estudio es caracterizar regiones espacialmente homogéneas basadas en patrones temporales de monóxido de carbono en el Área Metropolitana de Lima y Callao (AMLC) utilizando el algoritmo de clustering k-means. Este estudio utilizó concentraciones horarias promedios de CO medidas durante un periodo de 5 años (2015 – 2019) en las diez estaciones de monitoreo que conforman la Red de Monitoreo Automático de la Calidad del Aire (REMCA) del AMLC. Se utilizó el algoritmo de clustering (agrupamiento) de k-means empleando la distancia euclidiana para investigar la semejanza en los patrones entre los perfiles temporales observados en las estaciones de monitoreo. El análisis de agrupamiento de k-means identificó tres grupos de áreas con patrones temporales distintos que pudieron identificar y caracterizar zonas espacialmente homogéneas en el AMLC.

Palabras claves

Análisis de clúster, algoritmo k-means, monóxido de carbono, Lima Metropolitana, distribución espacial

Abstract

Statistical analyzes of time series or spatial data have been extensively used to examine the performance of air contaminants. Because air contamination data is commonly collected over an extensive area of interest over a relatively long period of time, such analyzes must take into account

both spatial and temporal features. The aim of this study is to characterize spatially homogeneous regions based on temporal patterns of carbon monoxide in the Metropolitan Area of Lima and Callao (MALC) using the k-means clustering algorithm. This study used average hourly CO concentrations measured over a 5-year period (2015 - 2019) in the ten monitoring stations that make up the MALC Automatic Air Quality Monitoring Network (AAQMN). The k-means clustering algorithm using Euclidean distance was used to research the likeness in the patterns among the temporal profiles observed at the monitoring stations. The k-means clustering analysis identified three groups of areas with different temporal patterns that were able to recognize and characterize spatially homogeneous zones in the MALC.

Keywords

Cluster analysis, k-means algorithm, carbon monoxide, Metropolitan Lima, spatial distribution

Introducción

El monóxido de carbono (CO) es un contaminante gaseoso que por su naturaleza es incoloro, inodoro e insípido, lo que lo convierte en una amenaza invisible y debido a sus implicancias en la salud de las personas, los animales y los vegetales es considerado como un contaminante criterio (Reumuth *et al*., 2019; USEPA, 2020). Su afinidad por las moléculas de hemoglobina es aproximadamente 200 veces mayor que la del oxígeno (O_2), por lo que en ciertas concentraciones forma con la hemoglobina, la carboxihemoglobina (COHb) que interfiere en el transporte del oxígeno a la sangre, causando una hipoxia tisular que afecta principalmente a áreas de alto flujo sanguíneo y demandantes de oxígeno (Cope, 2020; USEPA, 2020).

El CO surge del proceso de combustión incompleta de los combustibles fósiles, sus derivados y la biomasa. Éste es emitido de forma directa durante el arranque de vehículos a través de sus tubos de escape, causado por el suministro limitado de aire o la inapropiada afinación de los vehículos (USEPA, 2020).

Debido a sus múltiples efectos perjudiciales, muchas investigaciones buscan determinar la distribución de este contaminante en el ambiente para la prevención y control de la contaminación ambiental por CO. Es así que muchos estudios se han concentrado en la identificación de patrones en la distribución espacial y temporal de este gas contaminante (Stolz, 2020).

Las técnicas estadísticas como la agrupación jerárquica, el algoritmo de clustering k-means (Govender y Sivakumar, 2020) y el análisis de componentes principales (Kim *et al*., 2008) se pueden aplicar a los datos de calidad del aire para analizar los patrones de concentración de monóxido de carbono. Entre estas técnicas, el algoritmo de clustering k-means es una excelente herramienta estadística, ya que permite agrupar elementos (estaciones) con patrones similares de una variable dada y puede usarse para evaluar la influencia de cada estación individual y cómo se relacionan con el monóxido de carbono u otros patrones de concentración de contaminantes. Es importante mencionar que, a pesar de las posibilidades que puede ofrecer la técnica, los estudios que aplican el algoritmo k-means sobre datos de contaminación atmosférica para el monóxido de carbono son escasos en la literatura (Govender y Sivakumar, 2020).

Sin embargo, diversos estudios aplicaron esta técnica en la evaluación de otros contaminantes atmosféricos con resultados interesantes. Por ejemplo, Munir *et al*. (2015) utilizaron el algoritmo k-means en la caracterización de las variaciones temporales del ozono a nivel del suelo (O_3) en La Meca. El estudio muestra las variaciones diurnas del O_3 y menciona que el algoritmo k-means presenta de manera clara las diferencias de estas en el tiempo y espacio, diferencias que no son obvias cuando se usan solo presentaciones gráficas como diagramas de variación de tiempo que solo promedian las concentraciones durante un período de tiempo. Por otro lado, un estudio reciente ha demostrado cómo la agrupación de k-means se puede emplear para categorizar diferentes ubicaciones en una ciudad grande y poblada que representa la variabilidad de la contaminación de acuerdo con las variables empleadas para el estudio (Govender y Sivakumar, 2020).

Aunque varios estudios investigaron patrones de contaminantes atmosféricos en el AMLC (Silva *et al*., 2017; Silva *et al*., 2018), no existen investigaciones que consideren mediciones de monóxido de carbono en toda la ciudad, siendo esto importante dado que, en esta existen zonas con diferentes tamaños, población, fuentes e influencias. Además, ninguno aplicó la técnica de clustering k-means que como ya se mencionó puede proporcionar innumerables beneficios y ser una herramienta que proporcione información importante para el desarrollo de políticas públicas enfocadas al control de la contaminación atmosférica. Por lo tanto, el objetivo principal de este estudio es caracterizar regiones espacialmente homogéneas basadas en patrones temporales de monóxido de carbono durante 5 años en el AMLC utilizando el algoritmo de clustering k-means.

Materiales y métodos

Datos de contaminación del aire

Se analizaron los datos de calidad de aire de CO para el periodo 2015-2019. Los datos fueron recopilados en las diez estaciones de monitoreo de la Red de Monitoreo Automático de la Calidad del Aire (REMCA) del Área Metropolitana de Lima – Callao (AMLC) administrada por el Servicio Nacional de Meteorología e Hidrología del Perú (SENAMHI). Las estaciones de monitoreo están situadas en las diferentes zonas de Lima Metropolitana.

En este artículo se analizan las variaciones espaciales del CO utilizando el algoritmo de clustering de k-means (Govender y Sivakumar, 2020) y después las variaciones temporales de los clústers formados utilizando gráficos de variación de tiempo (Carslaw y Ropkins, 2012).

El análisis de los datos estadísticos se realizó con el lenguaje de programación del software estadístico R versión 4.0.3 (R Development Core Team, 2020) y una serie de paquetes complementarios, incluidos "*factoextra*" (Kassambara y Mundt, 2020), "*NbClust*" (Charrad *et al*., 2015) y "*openair*" (Carslaw y Ropkins, 2012). Éste último diseñado específicamente para el tratamiento de datos de contaminación del aire.

Algoritmo de clustering k-means

El método de agrupamiento de k-means es una técnica no jerárquica que se utiliza para agrupar observaciones en k grupos. Cada elemento se asigna a un grupo con el centro más cercano. El algoritmo actualiza iterativamente los grupos para minimizar la variación de sus elementos. El algoritmo básico de k-means, que se utilizó en este artículo, se refiere a la métrica euclidiana para definir la distancia entre los elementos y los centros de los conglomerados (Stolz *et al*., 2020).

Además, el resultado del método de k-means depende en gran medida del número de agrupaciones que se defina de antemano. En este estudio se utilizó el paquete *NbClust* la cual es la herramienta más utilizada para definir el número final de conglomerados (Charrad *et al*., 2015).

En general, el método de agrupamiento iterativo de k-means se implementa de la siguiente manera: Paso 1: Se elige un valor de k. Se usa como el conjunto inicial de k centroides. Paso 2: Se asigna cada uno de los objetos al grupo con el centroide más cercano. Paso 3: Se determina los nuevos centroides de los k grupos, calculando la media de los miembros del grupo. Paso 4: Se repiten los pasos 3 y 4 hasta que no haya cambios en la función de criterio después de una iteración (Govender y Sivakumar, 2020).

Las principales ventajas del algoritmo k-means son la baja complejidad, es computacionalmente alto, capacidad para manejar grandes conjuntos de datos y la flexibilidad para el ajuste del número de clúster (Govender y Sivakumar, 2020).

Variación diurna, semanal y mensual

La función *timeVariation* se utilizó para evaluar las diferencias en los patrones de concentraciones medias de monóxido de carbono horaria semanal, horaria, mensual y semanal (Carslaw y Ropkins, 2012).

La función *timeVariation* facilita la visualización de la variación de las concentraciones (y muchos otros tipos de variables) según la hora del día y el día de la semana. Los gráficos también muestran los intervalos de confianza del 95% en la media. Los intervalos de confianza del 95% en la media se calculan mediante simulaciones de *bootstrap*, que proporcionan estimaciones más sólidas de los intervalos de confianza (particularmente cuando hay relativamente pocos datos) (Carslaw, 2020).

Resultados y discusiones

Las estadísticas descriptivas que caracterizan a las diez estaciones para el contaminante monóxido de carbono (CO) se presentan en la **Tabla 1**. En ésta se puede observar que las estaciones ATE y CDM presentan los mayores y menores niveles promedio de CO respectivamente. Esto se debe a la ubicación de las estaciones, la primera se encuentra en la zona este del AMLC caracterizada por las industrias y el parque automotor pesado los cuales transitan por la carretera central, mientras que la segunda se encuentra en la zona centro del AMLC, caracterizada por abundantes áreas verdes y buena ventilación lo cual permite que el CO se disperse (Pacsi, 2016).

Tabla 1. Estadísticas descriptivas para el CO en las diez estaciones de la REMCA del AMLC

Zona de Lima	Estación	Abreviatura	Mínimo ($\mu g/m^3$)	Media ($\mu g/m^3$)	Máximo ($\mu g/m^3$)
Norte	Puente Piedra	PPD	0.5	1017.1±503.4	4534.4
	Carabayllo	CRB	5.6	771.7±342.9	4186.0
	San Martín de Porres	SMP	0.7	581.8±319.0	3122.7
Este	Ate	ATE	34.5	1490.9±723.3	5917.3
	Santa Anita	STA	0.3	881.6±451.6	4202.1
	San Juan de Lurigancho	SJL	1.3	902.9±741.2	5225.6

Sur	Huachipa	HCH	0.1	886.9±616.4	4547.1
	Villa María del Triunfo	VMT	0.2	788.9±328.4	3251.0
Centro	Campo de Marte	CDM	0.1	579.3±644.2	4743.8
	San Borja	SBJ	0.7	807.8±565.5	3671.9

El algoritmo de clúster k-means permitió agrupar a las diez estaciones de monitoreo de la REMCA en 3 grupos (Charrad *et al*., 2015). El clúster 1 conformado por las estaciones CRB y PPD; el clúster 2 conformado por las estaciones HCH, SJL, STA, ATE y VMT; y el clúster 3 conformado por las estaciones CDM, SBJ y SMP (**Figura 1**). Las estaciones que pertenecen al clúster 1 están ubicadas en la zona norte del AMLC, las del clúster 2, en la zona este y sur del AMLC; y las de clúster 3, en la zona centro y norte del AMLC (Pacsi, 2016).

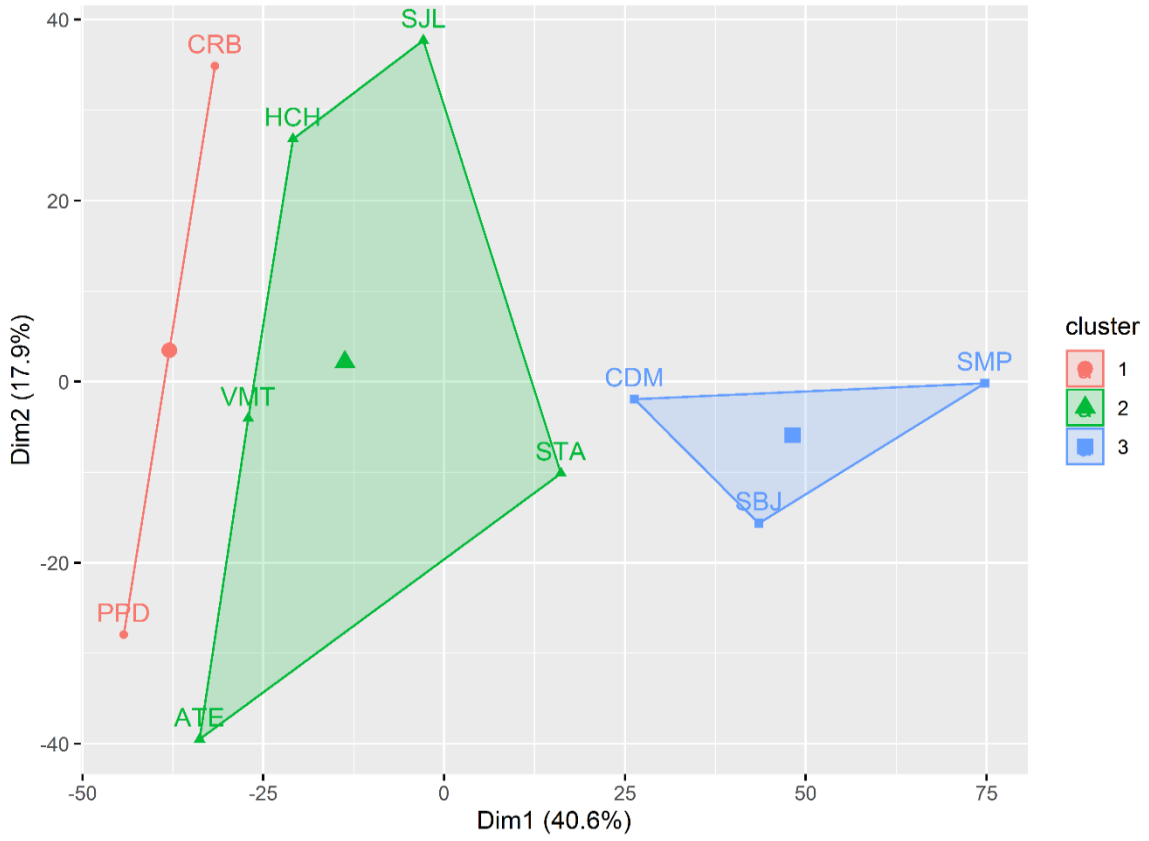

Figura 1. Distribución espacial de las estaciones de monitoreo de CO para el AMLC

De acuerdo con la **Tabla 2**, los clústers agruparon a las estaciones de monitoreo conforme al nivel de contaminación por CO que presentaban, es así que el CL2 agrupó a las estaciones con mayor nivel

de CO, el CL1 agrupó a las estaciones con nivel intermedio de CO y el CL3 a las estaciones con niveles bajos de concentraciones del CO.

Tabla 2. Estadísticos descriptivos según clúster

	CL1	CL2	CL3
Máximo	4534.4	5917.3	4743.8
promedio	883.6±441.0	992.5±648.1	655.4±529.2
Mínimo	0.5	0.1	0.1

Se ploteó la variación horaria semanal, horaria, mensual y semanal para las concentraciones de monóxido de carbono de las estaciones de monitoreo agrupadas en los 3 clústers. Se producen cuatro gráficas independientes por cada clúster.

En la **Figura 2** se presenta la variación temporal del CO del clúster 1. Se observa que la variación horaria y horaria semanal presentan dos picos, una en la mañana y otra en noche, los cuales guardan relación con la intensidad de las actividades antropogénicas, sobre todo de las emisiones del parque automotor (Pacsi, 2016). Semanalmente, la concentración de CO se mantiene constante de lunes a jueves, para luego aumentar los viernes y finalmente disminuir los sábados y domingos. En cuanto la variación mensual se muestra que las concentraciones más altas de CO se presentan en el mes de setiembre (1262 µg/m^3) para la estación PPD y junio (1002.2 µg/m^3) para la estación CRB y las concentraciones mínimas de CO se dan en los meses de marzo (685.5 µg/m^3) y noviembre (591.8 µg/m^3) respectivamente. Se evidencia que los mayores valores se dan en el periodo de invierno posiblemente debido al fortalecimiento del anticiclón del pacífico sur (APS) que genera condiciones de estabilidad atmosférica que caracteriza el AMLC (Sánchez y Ordoñez, 2016; Silva *et al*., 2018).

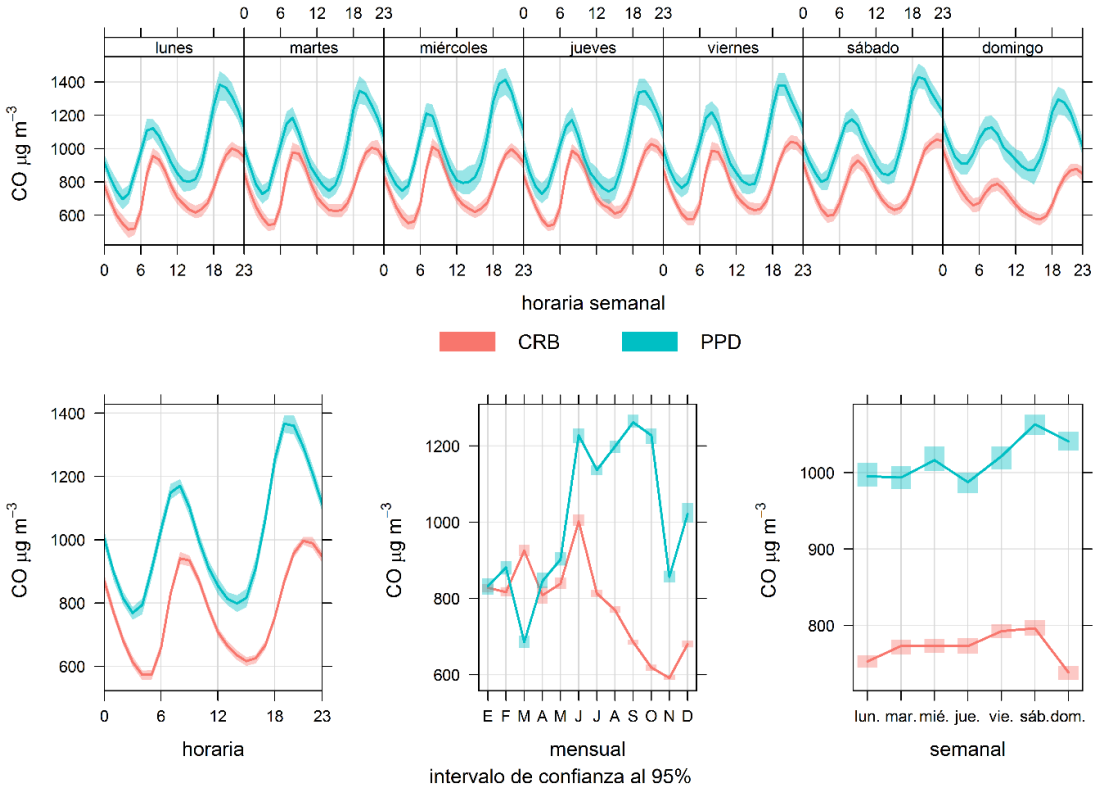

Figura 2. Variación temporal del CO en el clúster 1

En la **Figura 3** se presenta la variación temporal del CO del clúster 2. La variación horaria y horaria semanal, se caracterizan por dos picos a lo largo del día (Pacsi, 2016). Así también la variación semanal de los niveles de CO muestra un comportamiento constante de lunes a sábado y decae los domingos (Sánchez y Ordoñez, 2016). Respecto a la variación mensual se muestra que las concentraciones más altas de CO se presentan en los meses de marzo (1170.3 µg/m^3), abril (1059.1 µg/m^3 y 1365.1 µg/m^3), junio (2026.6 µg/m^3) y agosto (931.2 µg/m^3) para las estaciones HCH, STA, SJL, ATE y VMT respectivamente.

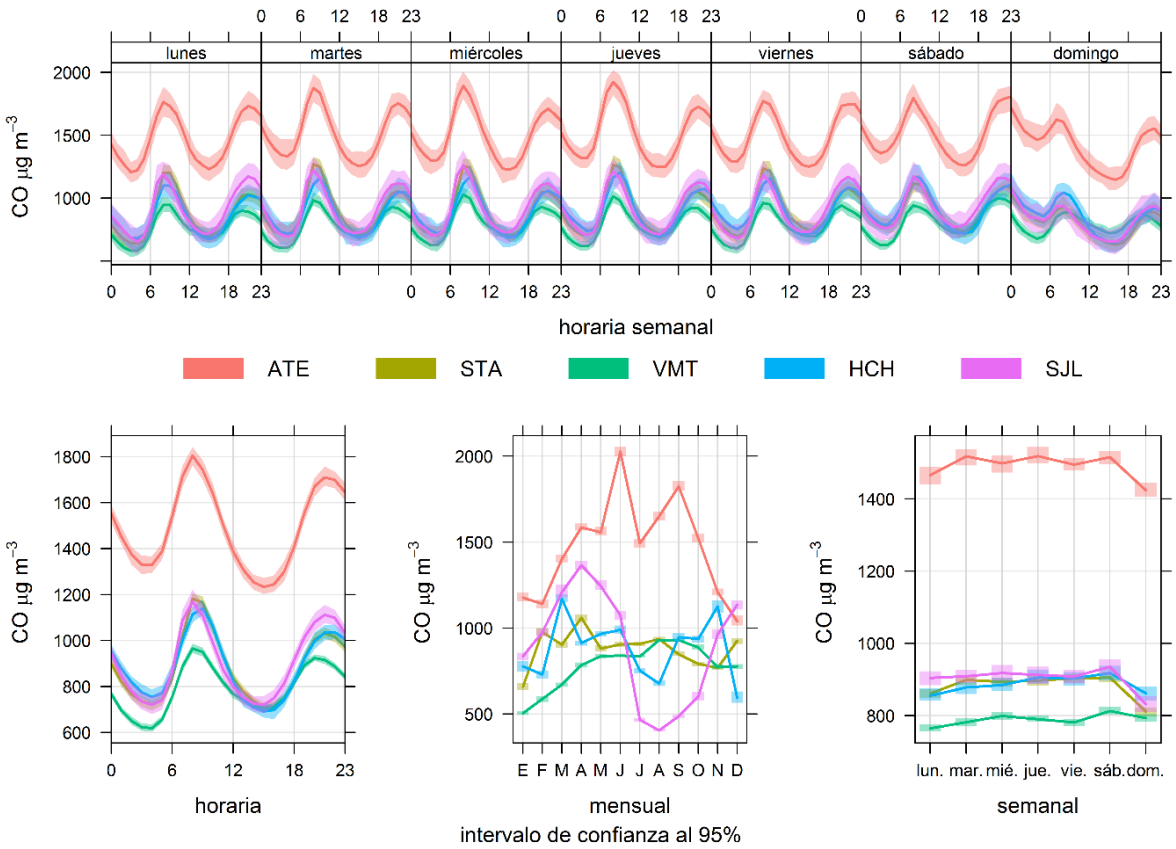

Figura 3. Variación temporal del CO en el clúster 2

La **Figura 4** presenta la variación temporal del CO del clúster 3. Se observa que la variación semanal de los niveles de CO muestra un comportamiento constante de lunes a viernes y decae los fines de semana. Respecto a la variación mensual se muestra que las concentraciones más altas de CO se presentan en los meses de diciembre (883.3 µg/m^3, 1026.9 µg/m^3) y enero (864.6 µg/m^3), para las estaciones CDM, SBJ y SMP respectivamente. En cuanto a las concentraciones mínimas de CO, éstas se dan en los meses de julio (189.1 µg/m^3 y 399.9 µg/m^3) para las estaciones CMD y SBJ y octubre (362.1 µg/m^3) para la estación SMP. Resultados similares se muestran en (Sánchez y Ordoñez, 2016).

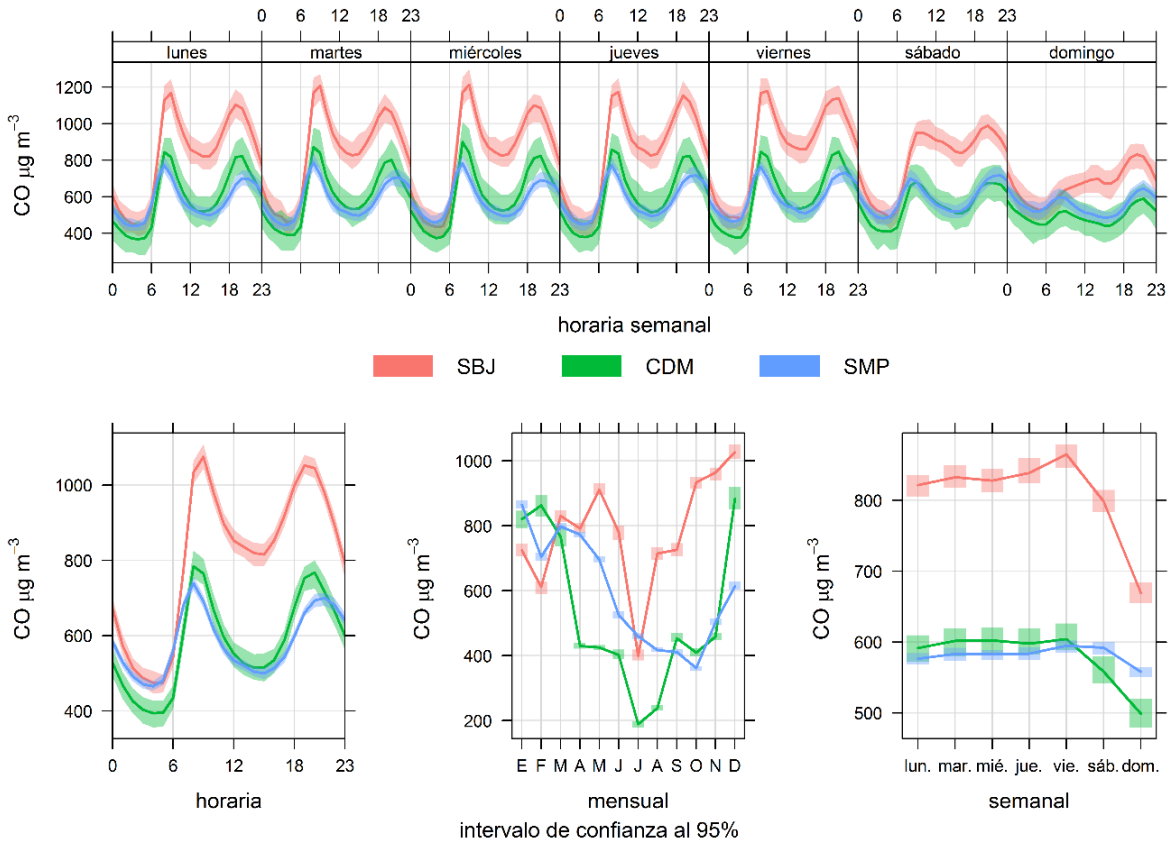

Figura 4. Variación temporal del CO en el clúster 3

Conclusiones y recomendaciones

En este estudio se utilizaron las concentraciones medias de CO para el periodo 2015-2019 de las diez estaciones de monitoreo de la REMCA en el AMLC. El algoritmo k-means permitió caracterizar regiones espacialmente homogéneas de contaminación por CO. Se identificó tres áreas con alta, intermedia y baja contaminación por CO en el AMLC. El clúster 2 se caracteriza por presentar mayores niveles de CO, mientras que el clúster 3, por los menores valores. Así también la estación ATE es aquella que presenta valores superiores para este contaminante. El análisis de agrupamiento de k-means agrupó las estaciones de monitoreo con estrecha cercanía espacial. Se recomienda complementar el estudio con un análisis de componentes principales (ACP) para caracterizar los patrones espaciales de CO y analizar las diferencias entre el algoritmo de clustering k-means y el ACP.

Bibliografía

Carslaw, D. C., & Ropkins, K. (2012). Openair—an R package for air quality data analysis. *Environmental Modelling & Software*, *27*, 52-61. https://doi.org/10.1016/j.envsoft.2011.09.008

Carslaw, D.C. (2020). Package "Openair". Tools for the Analysis of Air Pollution Data. https://cloud.r-project.org/web/packages/openair/openair.pdf

Charrad, M., Ghazzali N., Boiteau V., Niknafs A. (2015). Package "NbClust". Determining the Best Number of Clusters in a Data Set. Available from: https://cran.r-project.org/web/packages/NbClust/NbClust.pdf

Cope, R. (2020). Carbon monoxide: can't see, can't smell, body looks red but they are dead. In *Handbook of Toxicology of Chemical Warfare Agents*. https://doi.org/10.1016/b978-0-12-819090-6.00024-6

Govender, P., & Sivakumar, V. (2020). Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). *Atmospheric Pollution Research*, *11*(1), 40-56. https://doi.org/10.1016/j.apr.2019.09.009

Kassambara, A., Mundt, F. (2020). Package "factoextra". Extract and Visualize the Results of Multivariate Data Analyses. https://cran.r-project.org/web/packages/factoextra/index.html

Kim, S. B., Temiyasathit, C., Chen, V. C., Park, S. K., Sattler, M., & Russell, A. G. (2008). Characterization of spatially homogeneous regions based on temporal patterns of fine particulate matter in the continental United States. *Journal of the Air & Waste Management Association*, *58*(7), 965-975. https://doi.org/10.3155/1047-3289.58.7.965

Munir, S., Habeebullah, T. M., Mohammed, A. M., Morsy, E. A., Awad, A. H. A., Seroji, A. R., & Hassan, I. A. (2015). An Analysis into the Temporal Variations of Ground Level Ozone in the Arid Climate of Makkah applying k-means Algorithms. *EnvironmentAsia*, *8*(1), 53-60.

Pacsi, S. (2016). Análisis temporal y espacial de la calidad del aire determinado por material particulado PM_{10} y $PM_{2,5}$ en Lima Metropolitana. In *Anales Científicos* (Vol. 77, No. 2, pp. 273-283). Universidad Nacional Agraria La Molina. http://dx.doi.org/10.21704/ac.v77i2.699

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Reumuth, G; Alharbi, Z; Houschyar, K; Kim, B; Siemers, F; Fuchs, P y Grieb, G. (2019). Carbon monoxide intoxication: What we know. *Burns*, 45(3), 526–530. https://doi.org/10.1016/j.burns.2018.07.006

Sánchez, O., & Ordóñez, C. (2016). Evaluación de la calidad del aire en Lima Metropolitana 2015. Lima: Servicio Nacional de Meteorología e Hidrología del Perú. SENAMHI.

Silva, J., Rojas, J., Norabuena, M., Molina, C., Toro, R. A., & Leiva-Guzmán, M. A. (2017). Particulate matter levels in a South American megacity: the metropolitan area of Lima-Callao, Peru. *Environmental monitoring and assessment*, *189*(12), 635. https://doi.org/10.1007/s10661-017-6327-2

Silva, J. S., Rojas, J. P., Norabuena, M., & Seguel, R. J. (2018). Ozone and volatile organic compounds in the metropolitan area of Lima-Callao, Peru. *Air Quality, Atmosphere & Health*, *11*(8), 993-1008. https://doi.org/10.1007/s11869-018-0604-2

Stolz, T., Huertas, M. E., & Mendoza, A. (2020). Assessment of air quality monitoring networks using an ensemble clustering method in the three major metropolitan areas of Mexico. *Atmospheric Pollution Research*. https://doi.org/10.1016/j.apr.2020.05.005

United States Environmental Protection Agency USEPA. (2019). Carbon Monoxide (CO) Pollution in Outdoor Air. Disponible en: https://www.epa.gov/co-pollution.

Zhao, S., Yu, Y., Yin, D., He, J., Liu, N., Qu, J., & Xiao, J. (2016). Annual and diurnal variations of gaseous and particulate pollutants in 31 provincial capital cities based on in situ air quality monitoring data from China National Environmental Monitoring Center. *Environment international*, *86*, 92-106.

Zhao, S., Yu, Y., Qin, D., Yin, D., Dong, L., & He, J. (2019). Analyses of regional pollution and transportation of PM2.5 and ozone in the city clusters of Sichuan Basin, China. *Atmospheric Pollution Research*, *10*(2), 374-385. https://doi.org/10.1016/j.apr.2018.08.014