Redes Bayesianas Gaussianas: Un estudio de la influencia de los contaminantes del aire en la salud de la población mexicana

Romina Nájera Fuentes - A01424411
Humberto Mondragón García - A01711912
Juan Braulio Olivares Rodríguez - A01706880
Edgar Andrey Balvaneda - A01644770

Análisis de métodos de razonamiento e incertidumbre

7 de Septiembre del 2025

Abstract

Debido al tráfico vehicular, la actividad industrial y a la alta densidad poblacional, México tiene altos niveles de contaminantes del aire en grandes ciudades. Esta contaminación pone en riesgo la salud de la población mexicana, quienes interactúan continuamente con diferentes concentraciones de contaminantes de aire, dependiendo de la zona en la que se reside. El objetivo de este artículo es el de comprender el impacto que dicha contaminación tiene en la salud de la población mexicana. Se proponen diferentes modelos de redes bayesianas gaussianas, utilizando como variables a los contaminantes del aire en México y biomarcadores de la Encuesta Nacional de Salud y Nutrición (ENSANUT), cuyas relaciones se determinaron con apoyo de expertos en medicina y química. Entre estos, se escogió el modelo que presentó el mejor ajuste, con el cual se respondieron consultas de interés para observar el impacto de los contaminantes en los biomarcadores. Entre los resultados obtenidos, se asoció una mayor probabilidad de niveles de ferritina altos a una persona de mayor edad, en comparación con una persona joven, cuando se tienen niveles altos de óxido de nitrógeno; asimismo, se encontró que con niveles elevados de material particulado, que la proteína C reactiva sea elevada resulta en una probabilidad mayor a 0.4. Con ello, se observaron altas probabilidades de niveles de ferritina, proteína C reactiva y ácido úrico, biomarcadores asociados al transporte de oxígeno en la sangre, a la inflamación, y a los riñones respectivamente, indicando un efecto transversal en la salud de la población. Una de las mayores limitaciones del artículo se encuentra en la falta de biomarcadores relacionados al nivel de oxígeno.

Palabras clave: Contaminación del aire, contaminantes, biomarcadores, redes bayesianas gaussianas, población mexicana, salud, proteína C reactiva.

Introducción

La contaminación ambiental se ha convertido en una problemática crítica para la salud pública a nivel mundial. Según la Organización Mundial de la Salud, 9 de cada 10 personas respiran aire con altos niveles de contaminantes, y 4.2 millones de personas mueren al año a causa de la contaminación del aire del exterior. [9]

En México, en las zonas metropolitanas, como la Ciudad de México, Monterrey y Guadalajara, se reportan concentraciones elevadas de contaminantes provenientes del intenso tráfico vehicular, sus altas densidades poblacionales y la actividad industrial con pocas o nulas restricciones en la liberación de contaminantes en el ambiente. Si bien esta mezcla de contaminantes es compleja, ciertos de ellos son especialmente preocupantes por su impacto en la salud pública. Los datos de la SEMARNAT nos proveen de los siguientes contaminantes:

El material particulado (PM_{10} y $PM_{2.5}$) constituye uno de los contaminantes atmosféricos más nocivos para la salud. Proviene principalmente de procesos de combustión como los de vehículos o actividades industriales, es culpable de agravamiento de los síntomas del asma, deterioro de la función pulmonar y aumento del riesgo de cáncer de pulmón, garganta o laringe.[1]

El dióxido de azufre (SO_2) es un gas que se produce principalmente por la quema de combustibles fósiles como el carbón y el petróleo. Su inhalación provoca irritación de las vías respiratorias y desencadena reacciones inflamatorias locales y sistémicas. [7]

El monóxido de carbono (CO), por su parte, es un gas inodoro y altamente tóxico que se produce principalmente por la combustión incompleta de gasolina, madera, carbón o gas. Su mecanismo de toxicidad se debe a su alta afinidad con la hemoglobina, lo que desplaza al oxígeno y disminuye la capacidad de transporte de este en la sangre. [4]

Los óxidos de nitrógeno (NOx), que incluyen principalmente el dióxido de nitrógeno (NO_2) y el óxido nítrico (NO), se generan a partir de procesos de combustión como en vehículos motorizados, plantas eléctricas e industrias. Estos gases son potentes irritantes de las vías respiratorias, provocan inflamación bronquial y reducen la función pulmonar. [2]

Por último, los compuestos orgánicos volátiles (COV) y el amoníaco (NH_3), aunque menos estudiados en comparación con los contaminantes anteriores, también tienen efectos relevantes en la calidad del aire y la salud. Se han vinculado con irritación ocular y respiratoria, así como con procesos de estrés oxidativo. [3] Los COV provienen de solventes, combustibles, pinturas y procesos industriales, contribuyen junto con los NOx a la formación de ozono troposférico. [8]

Por la ya mencionada complejidad de los contaminantes, resulta necesario estudiar no solo su presencia ambiental, sino también sus repercusiones biológicas medibles en la población.

El presente trabajo tiene como objetivo principal implementar un modelo probabilístico que permita identificar en qué medida estos contaminantes influyen en la salud de la población mexicana, analizando sus biomarcadores. Comprender la magnitud de este impacto, el origen de los contaminantes, y los componentes biológicos afectados es fundamental para el diseño de políticas públicas efectivas y estrategias de prevención.

Para ello utilizamos los datos de muestras biológicas de la Encuesta Nacional de Salud y Nutrición (ENSANUT) 2022, y registros de contaminantes atmosféricos provenientes de la SEMARNAT.

Existen diversos enfoques estocásticos que podemos utilizar para modelar la influencia de los contaminantes en los biomarcadores. Entre ellos se encuentran métodos como la regresión, que permite evaluar asociaciones entre variables, los procesos de Poisson, que modelan conteos de eventos discretos, y las simulaciones Monte Carlo, que permiten explorar el comportamiento de modelos complejos. Sin embargo, estas aproximaciones frecuentemente se centran en relaciones simples entre variables o en estructuras lineales, lo que limita su capacidad de capturar la red de dependencias entre las variables de este estudio. Frente a ello, las redes bayesianas gaussianas ofrecen una alternativa más flexible representando las relaciones de dependencia en forma de un grafo acíclico dirigido y modelar la incertidumbre mediante distribuciones gaussianas.

Metodología

Una red bayesiana es definida como un grafo acíclico dirigido, conocido como DAG, $D = (V, E)$, en donde $V = \{X_1, X_2, ..., X_n\}$ representa a las variables del problema, y E representa las relaciones de dependencia entre las variables. Asimismo, se cuenta con P, el conjunto de distribuciones de probabilidad asociadas a cada nodo.

Se utilizará una red bayesiana de tipo gaussiana, la cual es utilizada para modelar variables aleatorias continuas que siguen una distribución normal. En este caso, la distribución conjunta de las variables del problema es normal multivariada; es decir, sigue una distribución $N(\mu, \Sigma)$, siendo μ el vector de medias y Σ la matriz de covarianzas. [6]

En la versión gaussiana de la red bayesiana, los nodos raíz, o nodos sin padres, siguen marginalmente una distribución univariada normal. Los demás nodos siguen de manera marginal una distribución condicionada normal, con una media que se representa a través de un modelo lineal respecto a las medias de los padres, y con una varianza propia [10].

Entre las variables presentadas en la ENSANUT, se encuentra información sociodemográfica de las personas encuestadas, como el sexo, la edad, la zona de residencia, el acceso a servicios de salud pública, entre otros. Sin embargo, de estos solamente se tomará en cuenta la edad, que es una variable que se puede considerar continua, y el lugar en el que se reside, para asignar los niveles de contaminantes a los que los encuestados se exponen de manera constante.

Además de la información sociodemográfica, la encuesta también contiene información relacionada con biomarcadores. Estos incluyen los niveles de ácido úrico, albúmina, colesterol HDL y LDL, colesterol total, creatinina, glucosa, insulina, proteína C reactiva, triglicéridos, hemoglobina glucosilada, ferritina, folato, homocisteina, receptor de transferrina, vitamina B12 y vitamina D de los encuestados. Si bien se cuenta con más información, estas variables son las que son continuas, condición necesaria en las redes bayesianas gaussianas, por lo que el impacto en la salud se determinará solamente con estas variables.

Con la información de la SEMARNAT, como se mencionó previamente en la introducción, se utilizarán los niveles de los contaminantes SO_2, CO, COV, NOx, PM_{10} y $PM_{2.5}$ medidos en los municipios de cada estado del país. Los contaminantes se dividen en diferentes tipos de fuentes causantes, que son las fuentes fijas, de área, móviles carreteros, móviles que no circulan por carretera y naturales. Por observación del Dr. Ernesto Reyes Villegas, experto en calidad del aire, estas fuentes son complementarias, por lo que la suma de los contaminantes en cada fuente es equivalente a la contaminación total en dicho municipio. Sin embargo, en vista de que no todos los municipios presentan información completa en las fuentes fijas y las fuentes naturales, se considerará que el nivel de contaminación a la que se somete una persona que reside en un municipio, es igual a la suma de la contaminación de fuentes de área, de móviles carreteros y de móviles que no circulan por carretera para cada contaminante por municipio.

Las variables seleccionadas son exclusivamente continuas, con el propósito de utilizar la red bayesiana gaussiana. Agregar una variable categórica, como lo podría ser el sexo de la persona encuestada, representaría utilizar un modelo distinto al aquí presentado. Una manera de realizarlo es a través de la discretización de las variables continuas. Sin embargo, para una mayor precisión, la discretización se tendría que llevar con múltiples categorías para cada variable, lo cual incrementa la complejidad del modelo y su tiempo de cómputo. [5]

Otra manera de realizar el modelo con la combinación de variables discretas y continuas es a través de una red bayesiana mixta. En esta, los nodos discretos siguen una distribución multinomial, y sus nodos padres pueden ser solamente otras variables discretas. En cambio, los nodos continuos simplemente siguen una distribución normal, cuando ninguno de sus nodos padre son discretos; y siguen un conjunto de distribuciones normales, cada una para cada posible combinación de las variables de las que dependen, en caso de tener alguna variable discreta como nodos padre. [10]

Utilizando estas variables, se consultará a especialistas en las áreas de medicina y de bioquímica, quienes apoyarán en la realización de la red bayesiana gaussiana, seleccionando las variables de interés, y proponiendo las relaciones que consideren relevantes.

Con ello, se generarán 3 DAGs, las cuales serán ajustadas a los datos para ser posteriormente comparadas entre sí a través del Criterio de Información Bayesiano (BIC), el cual considera la verosimilitud y la complejidad de cada modelo para asignar un puntaje que se busca maximizar [10], y del Criterio de Información Akaike (AIC), el cual considera los mismos parámetros que el BIC, pero penaliza en menor escala a la complejidad del modelo.

Estos criterios de información se calculan de la siguiente manera:

$$AIC = -2\ln(\hat{L}) + 2k \tag{1}$$

$$BIC = -2\ln(\hat{L}) + k\ln(n) \tag{2}$$

Después de seleccionar un modelo óptimo en base al puntaje BIC, se formarán 3 preguntas a resolver con dicha red bayesiana, las cuales se propondrán con el apoyo de los especialistas contactados para la propuesta de las redes.

Aplicación

La primera propuesta de la DAG fue dada por la estudiante de medicina Leslie Aleydis Ocampo Hernández, quien recalcó el efecto de los contaminantes en el aumento de la proteína C reactiva (*pcr*), debido a la inflamación que estos generan. Asimismo, estableció el efecto que tiene el material particulado (PM_{10} y $PM_{2.5}$) y el amoníaco (NH_3) en los niveles de ácido úrico y creatinina, debido al efecto en los riñones dada por la inflamación generada

por estos contaminantes. En la Figura 1 se puede observar el grafo generado con la guía de la Dra. Ocampo.

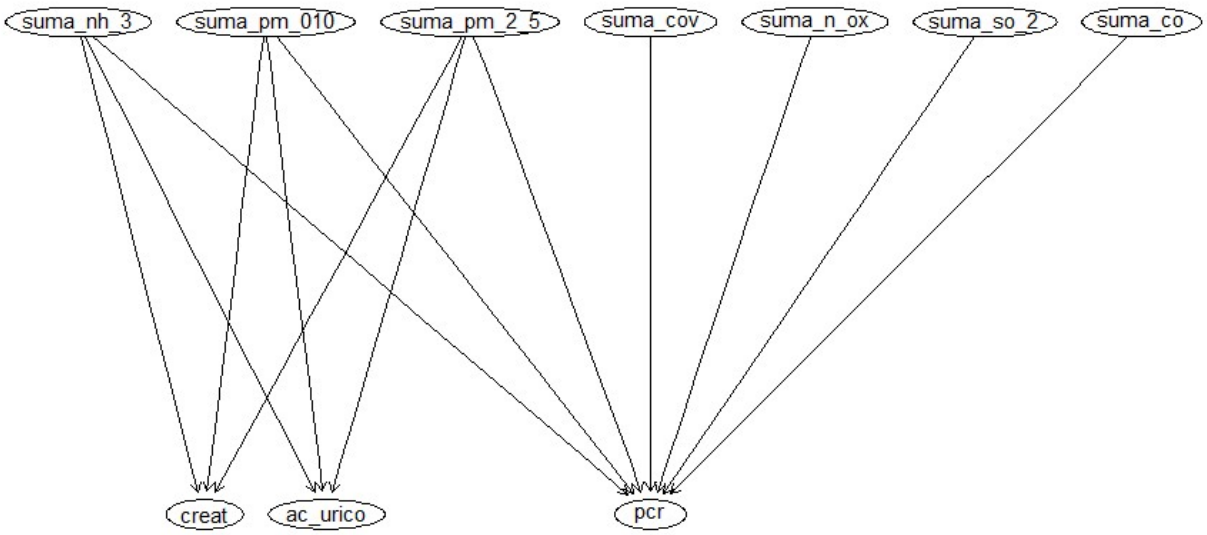

Figura 1: Grafo propuesto con el apoyo de la Dra. Ocampo.

La segunda propuesta para el grafo se obtuvo con ayuda de la Dra. Ma. Silvia Rodríguez López y del Dr. Ricardo Cano Pérez, quienes no consideraron que hubiera relación relevante entre el monóxido de carbono (CO) y los compuestos orgánicos volátiles (COV), respecto a alguno de los biomarcadores. Los demás contaminantes consideraron que todos influyen en la proteína C reactiva, en la ferritina y en el ácido úrico. Asimismo, decidieron añadir la edad como un factor que también influye a los niveles de los 3 biomarcadores mencionados. En la Figura 2 se observa el grafo generado con las observaciones realizadas por ambos doctores.

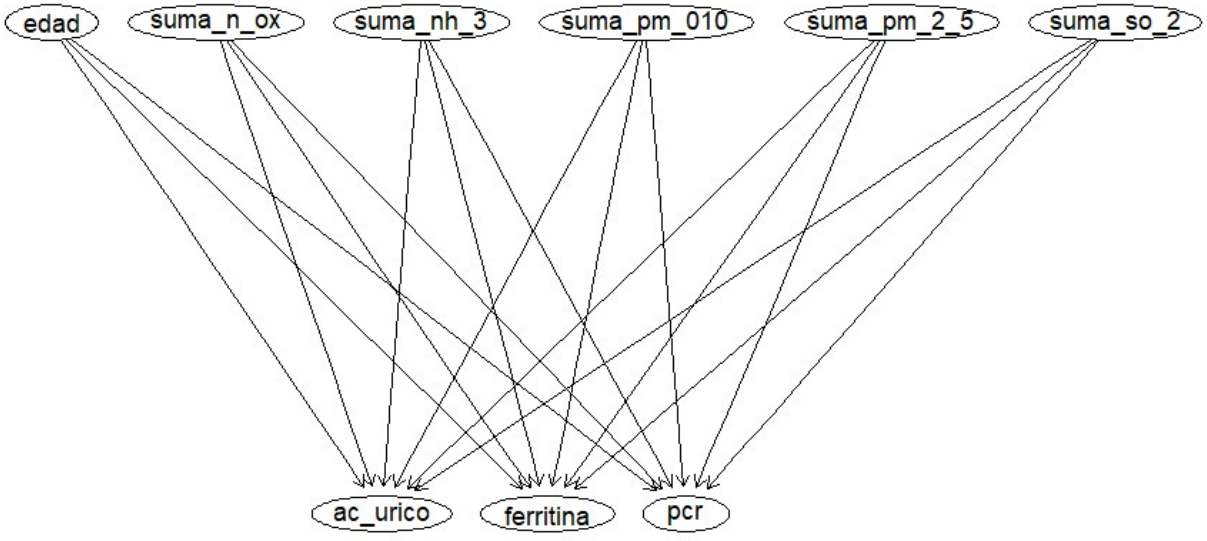

Figura 2: Grafo propuesto con el apoyo de la Dra. Rodríguez y el Dr. Cano.

La tercer propuesta de DAG se obtuvo con la ayuda del estudiante de medicina Jean Emmanuel Zabre Pando, quien estableció el efecto de los contaminantes como posible generador de inflamación, aumentando el nivel de proteína C reactiva, y el impacto del aumento de esta proteína, así como de los niveles de material particulado, en el metabolismo, particularmente en los triglicéridos. Posterior a ello, estableció las relaciones entre múltiples biomarcadores, información con la cual se generó la DAG de la Figura 3.

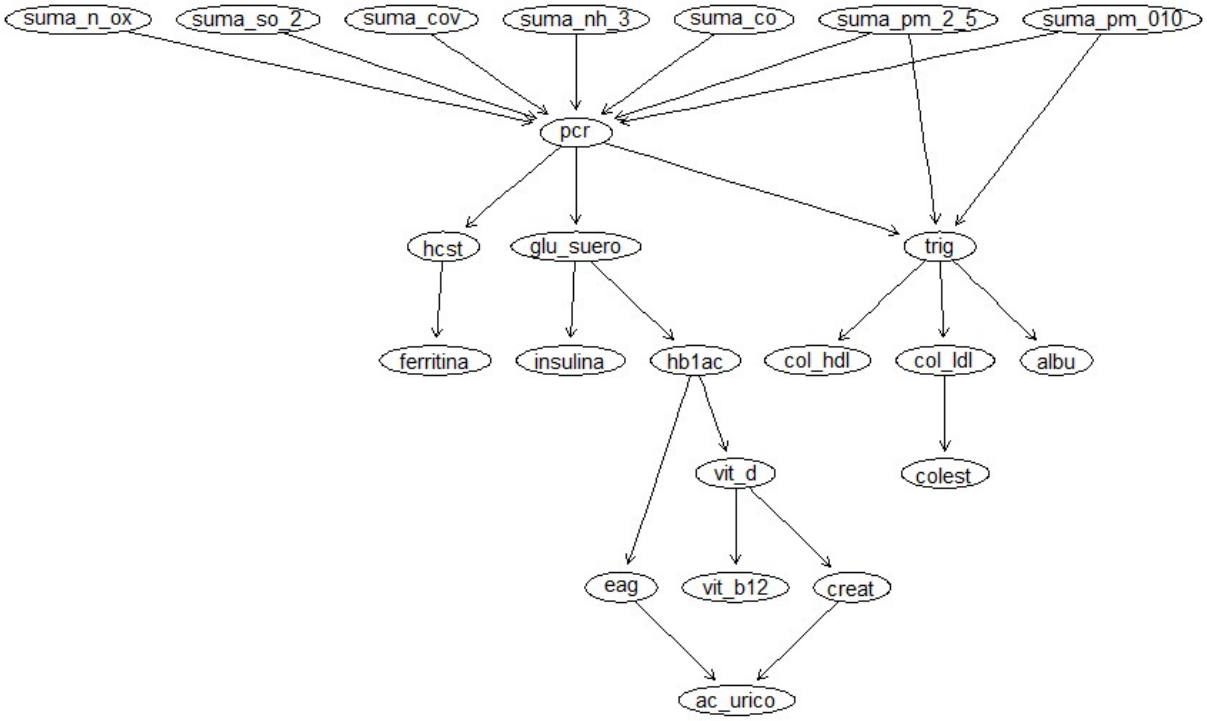

Figura 3: Grafo propuesto con el apoyo del Dr. Zabre.

Evaluando los modelos en los criterios de información seleccionados, se obtuvieron los resultados de la Tabla 1. Como ambos criterios se buscan maximizar para determinar un mejor modelo, los resultados muestran ser consistentes entre ambos criterios. Es decir, se puede concluir que el segundo modelo resultó en los valores más altos de los criterios, por lo que es la DAG con mejor ajuste entre las propuestas. Por otro lado, el tercer modelo tuvo los resultados más bajos, lo cual quiere decir que fue el que tuvo un peor ajuste a los datos. Esto se puede deber a la diferencia en complejidad de los modelos, puesto que el segundo modelo es el que tiene una menor cantidad de nodos, mientras que el tercer modelo es el que utiliza una mayor cantidad de variables. Sin embargo, los puntajes de los criterios también se ven influenciados por la verosimilitud del modelo, por lo que un menor puntaje también podría indicar un mal ajuste.

Modelo	BIC	AIC
1	-42502.98	-42418.14
2	-32122.3	-32027.29
3	-77035.66	-76779.56

Tabla 1: Criterios de información bayesiano y akaike en los modelos propuestos.

En una red bayesiana gaussiana lineal, como la aquí utilizada, se asume que la media de la distribución normal es una combinación lineal de sus nodos padre; en un modelo no paramétrico, no se impone la regresión lineal para la media, sino que se usan splines o procesos más flexibles que se adaptan a los datos. Como el modelo tiene más "libertad" en los modelos no paramétricos para captar las irregularidades de los datos, la verosimilitud suele aumentar.

En un modelo paramétrico k es bajo, (coeficientes + varianzas), pero en un modelo no paramétrico, k tiende a ser más alto. Por ejemplo, en splines, aumentan la cantidad de parámetros

Como lo mencionamos anteriormente, en un modelo no paramétrico la verosimilitud suele aumentar, y como la penalización por los parámetros, que podemos observar en la fórmula de AIC1 se comporta como $2k$, normalmente la mejora en la verosimilitud lo compensa, por lo que AIC suele ser más bajo. (mejor)

En BIC, la verosimilitud también mejora, pero la penalización cambia 2 , si n es pequeño (número de datos), $ln(n)$ no es tan grande y la penalización puede matar al modelo no paramétrico, en cambio, si n es grande, la verosimilitud aumentada puede compensar la penalización. Por lo que dependiendo de n puede mejorar o no.

Como podemos darnos cuenta, el modelo que dio un mejor resultado para las métricas planteadas fue el propuesto en colaboración con la Dra. Silvia, por lo tanto, a partir de las variables relacionadas en esta DAG podemos plantear ciertas preguntas a responder.

1. ¿Cuál es la probabilidad de que una persona de 20 años o menos tenga niveles de ferritina mayores a 100, respecto a las personas mayores de 50 años, dado que se tiene contacto con niveles de NOx mayores a 20,000?

En lugares con niveles de NOx superiores a 20,000, la probabilidad de que una persona de 20 años o menos presente niveles de ferritina mayores a 100 es de aproximadamente 0.13, mientras que en personas mayores de 50 años esta probabilidad aumenta a cerca de 0.19. Esto indica que, bajo las mismas condiciones de exposición a contaminantes, los adultos mayores tienen un riesgo relativo más alto de presentar ferritina elevada en comparación con los jóvenes.

2. ¿Cuál es la probabilidad de que una persona viva en un municipio con niveles de PM_{10} mayores a 2,000 dado que su nivel de ácido úrico es menor a 3, respecto a las personas con nivel de 6 o más?

En los municipios con niveles de PM_{10} superiores a 2,000, la probabilidad de que una persona con ácido úrico bajo (menor a 3) viva en estas condiciones es de aproximadamente 0.176, mientras que para las personas con ácido úrico alto (mayor o igual a 6), esta probabilidad aumenta a cerca de 0.299. Esto indica que parece existir una asociación entre la exposición a mayores concentraciones de PM_{10} y los niveles elevados de ácido úrico.

Conclusión

Referencias

[1] Airly. *What is PM2.5 and PM10? Info about particulate matter (particle pollution)*. Consultado: 06-09-2025. s.f. URL: https://airly.org/en/what-is-pm10-and-what-is-pm2-5/.

[2] ATSDR. *Resúmenes de Salud Pública – Amoníaco (Ammonia)*. Consultado: 06-09-2025. 2016. URL: https://www.atsdr.cdc.gov/es/phs/es_phs126.html.

[3] ATSDR. *ToxFAQs™ – Óxidos de nitrógeno (monóxido de nitrógeno, dióxido de nitrógeno, etc.) (Nitrogen Oxides)*. Consultado: 06-09-2025. 2026. URL: https://www.atsdr.cdc.gov/es/toxfaqs/es_tfacts175.html.

[4] Mayo Clinic. *Intoxicación con monóxido de carbono*. Consultado: 06-09-2025. 2025. URL: https://www.mayoclinic.org/es/diseases-conditions/carbon-monoxide/symptoms-causes/syc-20370642.

[5] R.G. Cowell et al. "Probabilistic Networks and Expert Systems". In: vol. 43. Jan. 2001, pp. 125–126.

[6] Rosario Susi García. "Análisis de sensibilidad en redes bayesianas gaussianas". PhD thesis. Madrid: Universidad Complutense de Madrid, 2007.

[7] IVHHN. *Dióxido de azufre (SO2)*. Consultado: 06-09-2025. 2003. URL: https://www.ivhhn.org/es/guidelines/guia-sobre-gases-volcanicos/dioxido-de-azufre.

[8] Agencia de Protección Ambiental de los Estados Unidos. *El impacto de los compuestos orgánicos volátiles en la calidad del aire interior*. Consultado: 06-09-2025. 2025. URL: https://espanol.epa.gov/cai/el-impacto-de-los-compuestos-organicos-volatiles-en-la-calidad-del-aire-interior.

[9] Organización Mundial de la Salud. *Nueve de cada diez personas de todo el mundo respiran aire contaminado*. Consultado: 06-09-2025. 2018. URL: https://www.who.int/es/news/item/02-05-2018-9-out-of-10-people-worldwide-breathe-polluted-air-but-more-countries-are-taking-action.

[10] Marco Scutari and Jean-Baptiste Denis. *Bayesian Networks: With Examples in R*. 2nd ed. Texts in Statistical Science. CRC Press, 2021. ISBN: 978-0-367-36651-3.