Artículo 2 - Redes Bayesianas gaussianas

Barush Caliel C. Luna Diego B. Castillo
Pedro Emilio S. Rodríguez Vidal Alejandro G. López

2025-09-07

Abstract

This study employs Gaussian Bayesian networks (GBNs) to analyze the complex relationships between air pollution, socio-demographic factors, and health biomarkers in Mexico. Using data from the National Health and Nutrition Survey (ENSANUT) 2022 and air quality data from SEMARNAT, we integrated domain expertise from medical and environmental specialists to propose multiple network structures. Our methodology involved data fusion, variable selection, and model optimization using Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC). The best-performing GBN incorporated sex as a categorical variable and revealed significant dependencies between pollutants like PM . and PM and health outcomes such as inflammatory markers. Our findings highlight the utility of GBNs in environmental health research and provide insights for targeted public health interventions.

Table of contents

1 Introducción

La contaminación ambiental y el cambio climático constituyen una crisis global con impactos profundos en la salud humana. En México, la exposición a contaminantes atmosféricos como material particulado (PM . y PM), ozono (O), dióxido de nitrógeno (NO) y otros ha sido vinculada a alteraciones en biomarcadores biológicos asociados con inflamación, función respiratoria y riesgo cardiovascular. Estas interacciones son inherentemente complejas, ya que los contaminantes no actúan de forma aislada, sino que interactúan entre sí y con factores socio demográficos y fisiológicos, creando una red de dependencias que desafía los enfoques analíticos tradicionales.

Las redes bayesianas gaussianas (GBN) emergen como una herramienta prometedora para modelar estas interacciones, permitiendo no solo la descripción de dependencias probabilísticas, sino también la inferencia predictiva y la simulación de escenarios bajo diferentes condiciones de exposición. Este estudio se centra en implementar un modelo probabilístico que capture cómo los factores ambientales influyen en la salud de la población mexicana, con el objetivo de generar evidencia para políticas públicas y estrategias de prevención.

Utilizamos datos de la Encuesta Nacional de Salud y Nutrición (ENSANUT) 2022, que incluye información sobre biomarcadores y variables socio demográficas, junto con datos de calidad del aire de la SEMARNAT (2025), asumiendo que estos últimos son representativos de las condiciones en 2022. Mediante la colaboración con especialistas en medicina y química, propusimos varias estructuras de redes bayesianas y evaluamos su desempeño mediante criterios de información. Además, exploramos la inclusión de variables categóricas como el sexo y aplicamos modelos no paramétricos para mejorar el ajuste del modelo. # Metodología

Las herramientas utilizadas en este reporte constaron de Rstudio como el entorno de programación junto con el lenguaje R, donde se efectuó la mayoría de ejecución del proyecto; Excel como la base de datos; y GitHub para el control de versiones y el trabajo colaborativo.

1.1 Entrevistas y estructuración de DAGs

La conexión con profesores y especialistas en los temas de salud fue fundamental, pues permitió conocer una faceta diferente del enfoque del objetivo del proyecto y saber específicamente cómo se comportan las variables del medioambiente con los biomarcadores.

En concreto, se entrevistaron a tres especialistas por medio de una encuesta de 19 preguntas, la cual estaba diseñada para responder con *sí*, *no*, o dar detalles si era muy compleja la pregunta. Esta encuesta se pensó de manera que se descartara o confirmara relaciones de dependencia entre las variables de biomarcadores, demográficas y del medioambiente; mismas preguntas fueron separadas en secciones. (La encuesta se encuentra en el repositorio, así como las respuestas).

Concretamente, las personas entrevistadas fueron: la Dra. Mariana Sofía Flores Jiménez, profesora e investigadora del Instituto Tecnológico de Estudios Superiores de Monterrey (ITESM) ligada al área de salud y biomédica; al profesor Rafael Alberto Pérez San Lázaro así como estudiante de doctorado ligado al área de salud del ITESM; y la Dra. Adriana Barragán Castillo, médica urgencióloga del ISSTE.

1.2 Construcción de base de datos

Una vez que se tuvo la información sobre las variables, así como propuestas de DAGs, se enlistaron las variables para el proyecto. Bajo el supuesto de que la información de la encuestas (SEMARNAT 2025 y ENSANUT 2022) se puede relacionar y hay representación de los contaminantes sobre los biomarcadores, se construyó el base de datos. El análisis se buscó realizarse por entidad federativa, por tanto, la variable que representaba el Estado en ambos base de datoss fue el medio de unión.

Como primer paso, se renombraron las variables que coincidian en la base de datos, así como cambiar el tipo de dato de *chr* a numérico reemplazando las comas por puntos decimales. Se conservó únicamente el conjunto de variables consideradas relevantes tras las entrevistas y recomendaciones de especialistas, incluyendo los biomarcadores como PCR, colesterol, triglicéridos y hemoglobina glicosilada, además de información demográfica como edad, sexo y la entidad federativa. Asimismo, para las variables de entidad federativa, se reemplazaron los nombres tipo *str* por valores *int*, de modo que su unión fuera más dinámica.

Para cada observación en la base de datos de ENSANUT se promedió una observación de SEMARNAT y se le asignó por cada coincidencia que había por entidad federativa. Esto se hizo con el objetivo de vincular los datos de salud con las exposiciones ambientales de manera representativa, evitando la duplicación de filas. Y por último, se eliminaron las variables con valores faltantes.

1.3 Construcción de GBNs

Se construyeron tres redes bayesianas gausianas (GBN) basadas en las DAGs propuestas por los especialistas. La calidad del ajuste de cada DAG se evaluó mediante los criterios BIC y AIC, donde valores menos negativos indican un mejor ajuste.

1.4 Queries

Posterior a la entrevista de las 19 preguntas, se pidió a los especialistas que dieran una propuesta de query basada en la información presentada. Esta última fue guardada y reacomodada de modo que pudiera ser interpretada matemáticamente, y por ello, compatible con el lenguaje de programación R.

Sumado a ello, al ser las queries técnicas, se tuvo que investigar términos como qué representa que un biomarcador sea *alto*, *bajo* o *estable*. Para dicha tarea se utilizaron citios web de confianza como la National Center for Biotechnology Information (NCBI) para obtener los rangos específicos e interpretar correctamente lo que se requería en las queries.

1.5 Modelo no paramétrico

Para complementar el análisis lineal y la GBN, se ajustaron modelos aditivos generalizados (GAMs) multivariables (`gam`) para los biomarcadores de interés, incluyendo PCR, colesterol, triglicéridos y glucosa. Cada modelo incorporó efectos suaves por medio del método `s()` para las variables continuas, de modo que se capturaran las relaciones no lineales complejas entre los contaminantes, los datos demográficos y los biomarcadores.

Asimismo, se evaluó el desempeño de las GAMs mediante el cálculo del BIC total, sumando los modelos de los biomarcadores y variables demográficas, comparando con la de la DAG seleccionada. El fin de este procedimiento fue la modelación de la estructura de la GBN.

2 Resultados (aplicación)

2.1 Imágenes de DAGs

2.1.1 DAG propuesta 1

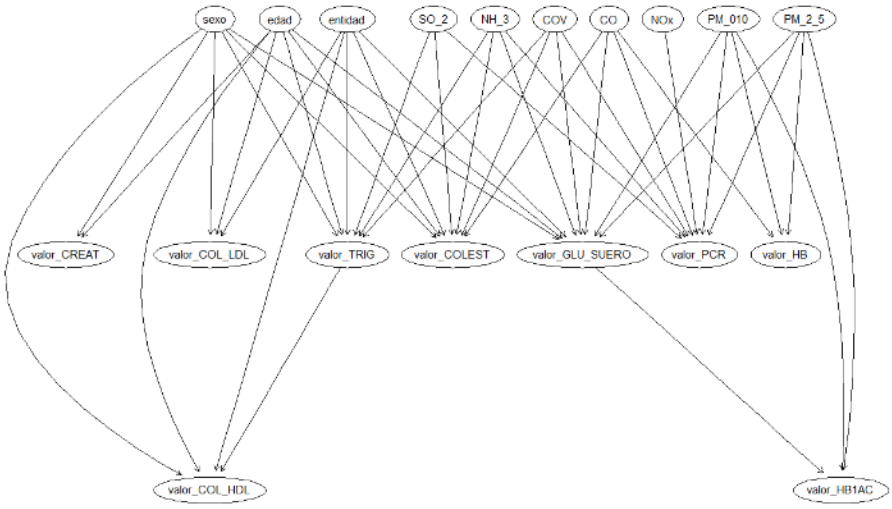

Figure 1: DAG 1

2.1.2 DAG propuesta 2

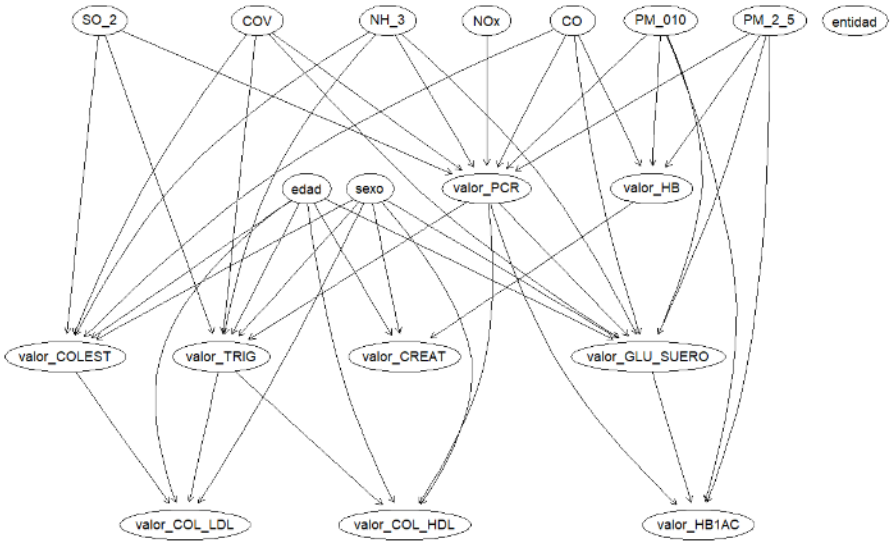

Figure 2: DAG 2

2.1.3 DAG propuesta 3

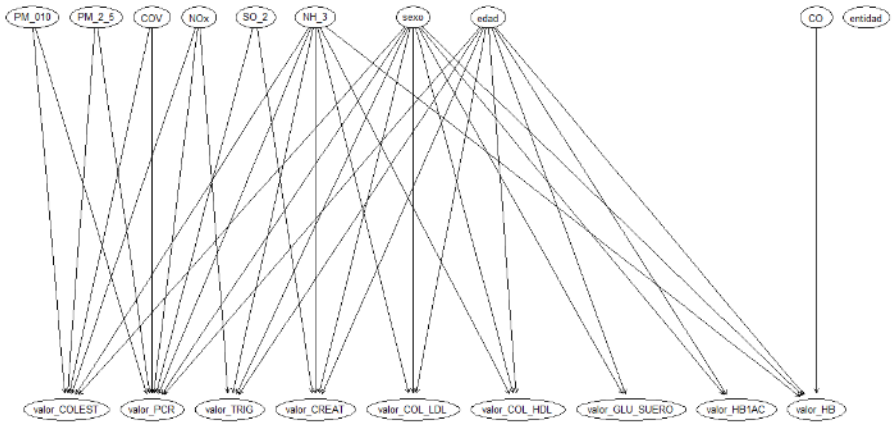

Figure 3: DAG 3

2.2 Resultados de BIC & AIC de las GBNs

Modelo GBN	BIC	AIC
DAG 1	-187,889.7	-187,661.8
DAG 2	-186,319.3	-186,085.8
DAG 3	-188,021.4	-187,813.3

2.3 Modelos con variables categóricas

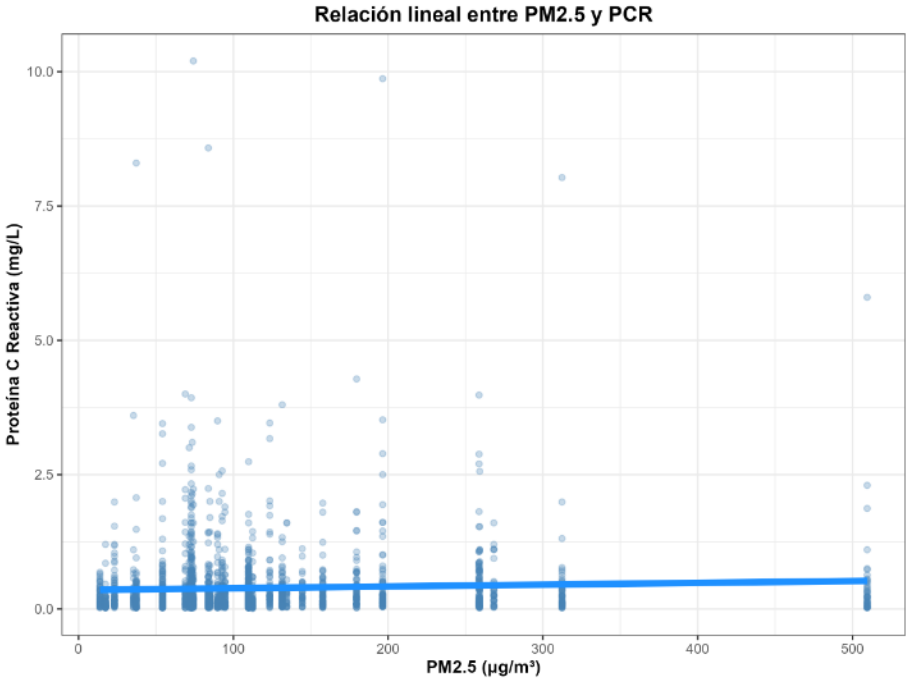

Figure 4: Modelo de regresión lineal (LM / MLE)

2.4 Modelo no paramétrico

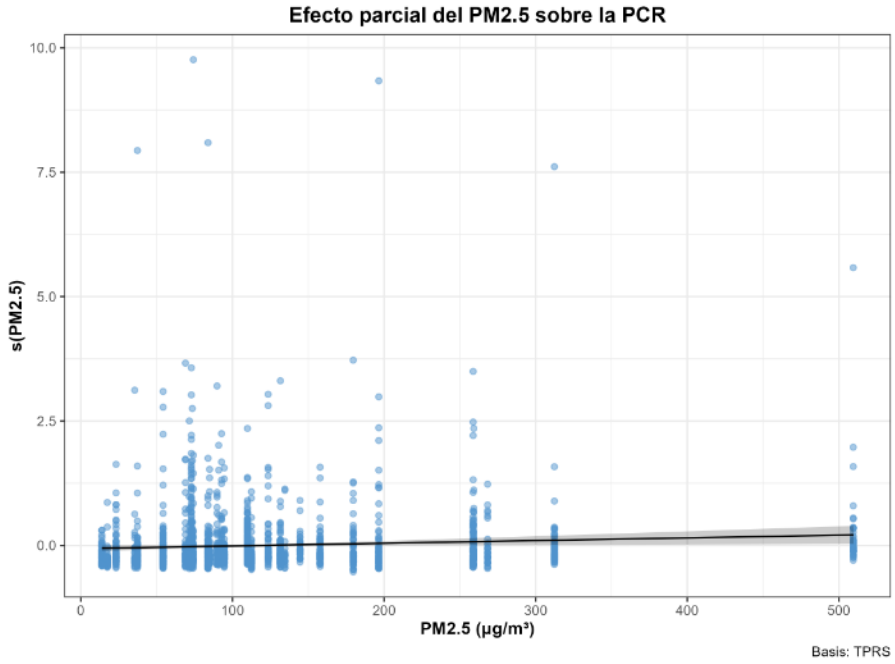

Figure 5: Modelo no paramétrico (Splines)

3 Análisis de resultados

3.1 Análisis de DAGs, BIC y AIC

Para evaluar la calidad y ajuste de los diferentes modelos de Redes Gaussinas-Bayesianas que se construyeron, se aplicaron dos criterios estadísticos que son el BIC-G y AIC-G, una versión adaptada a datos continuos, estos criterios proporcionan una medida comparativa para la selección de las variables.

Los valores de BIC-G para los 3 modelos analizados (dag_1, dag_2, dag_3) en el conjunto de datos usado fueron de $-187889.7, -186319.3, -188021.4$ respectivamente. De acuerdo a esta métrica, el modelo dag_2 presentó el valor menos negativo, indicando un mejor valance en comparación con los otros modelos.

De forma similar, el AIC-G arrojó valores de $-187661.8, -186058.8, -197813.3$ para los mismos modelos, reiterando que la dag_2 es la mejor opción para esta métrica.

Estos resultados sugieren que la dag_2 es el modelo que logra una mejor representación del conjunto de datos.

3.2 Análisis de modelos con variables categóricas

Dado que algunas variables originales del data set eran categóricas, (por ejemplo, sexo y entidad), estas fueron transformadas mediante funciones de codificación, como lo es *mutate* dentro de R, para convertirlas en variables numéricas.

Esta transformación facilita el cálculo de probabilidades condicionales en las DAGs y evita sesgos que puedan surgir del manejo indiscriminado de variables categóricas en el modelo.

El proceso permitió integrar de forma homogénea todas las variables en los modelos bayesianos, contribuyendo a la interpretabilidad de los resultados.

3.3 Métricas del modelo no paramétrico

El modelo no paramétrico basado en GAMs demostró un desempeño superior con un BIC total de $-54,816.66$, superando considerablemente a los modelos bayesianos gaussianos tradicionales. Esta mejora sustancial en la métrica confirma que la inclusión de componentes no paramétricos permite capturar relaciones no lineales de mayor complejidad entre contaminantes ambientales y biomarcadores de salud que los enfoques lineales convencionales no logran detectar, lo cual valida la efectividad del modelado flexible para este tipo de análisis epidemiológico.

3.4 Análisis de probabilidades

Se realizaron consultas clave para interpretar y validar las relaciones relevantes entre contaminación ambiental y biomarcadores de salud.

3.4.1 Query 1

Se evaluó la probabilidad de que un individuo presente valores elevados de Proteína C Reactiva, (mayores o iguales a 50 mg/dL, que indica una severa inflamación sistémica), condicionado a que sus niveles de exposivión a partículas $PM_{2.5}$ estén en 50 $\mu g/m^3$ y óxidos de nitrógeno (NOx) en 3160 $\mu g/m^3$, umbrales ya predefinidos.

La porbabilidad resulto ser 0%, lo que sugiere que en el conjunto de datos esta condición es muy poco frecuente o que existe poca evidencia empírica para estos valores extremos, aunque no se descarta su presencia en otros contextos.

3.4.2 Query 2

Se evaluó la probabilidad de que un individuo presente niveles elevados de creatinina sérica (>1.2 mg/dL, indicativo de posible daño renal), condicionado a un valor específico de hemoglobina glicosilada (HbA1c = 6.5%) y una concentración específica de material particulado fino (PM . = 12 g/m³).

La probabilidad condicional estimada fue de 8.11%, lo que sugiere una asociación moderada entre estas condiciones específicas y la creatinina elevada. Este análisis ayuda a explorar la relación entre la contaminación ambiental, el control glucémico y la función renal.

3.4.3 Query 3

Se evaluó si los contaminantes aéreos que afectan vías respiratorias también se pueden asociar en cambios a los biomarcadores metabólicos e inflamatorios.

Dichos biomarcadores son: Proteína C Reactiva $\geq 50\ mg/dL$, hemoglobina glicosilada $(HbA1c) \geq 6.5\%$ y triglicéridos $\geq 150\ mg/dL$. Condicionado a la presencia de contaminantes atmosféricos elevados pero más moderados que en la query 1: $COV < 0.5ppm$, $NH_3 > 10ppb$, $NOx > 40ppb$ y $PM_{2.5} > 12\mu g/m^3$

Nuevamente, la probabilidad condicional estimada fue 0%, sugiriendo que la combinación simultánea de estos valores elevados en biomarcadores e indicadores contaminantes no se observa en el conjunto de datos o es extremadamente rara.

4 Conclusiones

Este estudio mostró que las redes bayesianas gaussianas sirven como una herramienta útil para modelar las interacciones entre factores de contaminación ambiental, factores sociodemográficos y biomarcadores de salud en la población mexicana. Los resultados dados confirman que la colaboración con especialistas médicos y estudios ambientales es crucial para construir modelos probabilísticos que, además de optimizar métricas estadísticas, mantengan coherencia científica e interpretabilidad clínica.

Este estudio mostró que las redes bayesianas gaussianas son una herramienta útil para modelar las interacciones entre contaminación ambiental, factores sociodemográficos y biomarcadores de salud en la población mexicana. Los resultados confirman que la colaboración con especialistas médicos y ambientales es clave para construir modelos probabilísticos que, además de optimizar métricas estadísticas, mantengan coherencia científica e interpretabilidad clínica.

El análisis comparativo de las tres grafos acíclicos propuestos reveló que el **DAG 2** obtuvo el mejor desempeño, con valores de $BIC = -186{,}319.3$ y $AIC = -186{,}058.8$, superando a las alternativas. La incorporación de variables

categóricas como *sexo* mejoró la capacidad predictiva del modelo, haciendo relevante incluir factores demográficos dentro del análisis de exposición ambiental y salud pública. Asimismo, la aplicación de modelos no paramétricos (GAMs) permitió observar relaciones no lineales que complementaron el enfoque flexible bayesiano.

No obstante, el proyecto tuvo dificultades en la integración de bases de datos. La unión entre ENSANUT 2022 y SEMARNAT 2025 solo pudo hacerse a través de la variable *entidad federativa*, lo que redujo el nivel de detalle del análisis y pudo generar sesgos al promediar contaminantes por estado, perdiendo así variabilidad importante dentro de cada entidad.

Otro reto fue la definición de relaciones de dependencia entre variables. Con la participación de expertos, fue posible establecer conexiones con sentido entre contaminantes específicos y biomarcadores. Sin embargo, formular preguntas lo suficientemente específicas para los especialistas sin inducir sesgos requirió tiempo de planeación de la encuesta aplicada, que constó de 19 preguntas.

Los resultados de las consultas probabilísticas mostraron probabilidades muy bajas (0%) para escenarios de exposición severa. Esto puede interpretarse de dos maneras: por un lado, que estas condiciones son poco frecuentes en los datos analizados; por otro, que el modelo aún requiere refinamiento para capturar eventos extremos de exposición. A esta dificultad se suma el hecho de que una gran parte de los registros presentaba valores faltantes en distintos biomarcadores, además de que muchos estaban en formato de caracteres, por lo que fue necesario aplicar técnicas de conversión de *strings* a valores numéricos para que pudieran emplearse en el análisis de DAGs. Asimismo, se encontraron numerosas celdas vacías o con espacios, lo que obligó a generar dos bases de datos distintas.

La primera base conservó únicamente los registros completos en todos los biomarcadores, lo que redujo el tamaño a cerca de 2,000 filas. La segunda base aplicó imputación con KNN ($k = 3$) para rellenar los datos faltantes, alcanzando aproximadamente 13,000 registros y con un sesgo cercano al 5%, lo cual se consideró bajo y consistente con lo esperado en una base diversa. No obstante, se decidió trabajar principalmente con la primera base, ya que aquella cuenta los datos completos y sin imputación. De esta manera, se evita el riesgo de introducir dependencias derivadas de información que no provenía directamente de la encuesta de salud original.

En este contexto, las consultas produjeron resultados limitados. Era difícil esperar altos niveles de dependencia entre bases de datos que solo pudieron vincularse a través de una variable común (*entidad federativa*). Además, los contaminantes en el aire son factores complejos de reflejar en los biomarcadores considerados, pues intervienen múltiples elementos "invisibles" ó que al menos no se encontraban dentro de los datos que varían de manera importante entre individuos. A ello se añade la restricción del número de registros disponibles (solo alrededor de 2,000 en la base limpia), lo cual redujo considerablemente la

robustez de las consultas probabilísticas que pudieron realizarse.

En conjunto, este trabajo construye un primer acercamiento metodológico para el análisis probabilístico de factores ambientales y de salud pública en México. Aunque presenta limitaciones, este sienta las bases para investigaciones más amplias que, con mejoras en la calidad y cobertura de los datos, podrían llegar a ofrecer recursos para el diseño de políticas de prevención y estrategias de salud ambiental mejor fundamentadas.

5 Referencias

Air Quality, Energy and Health (AQE). (2021, 22 septiembre). WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. https://www.who.int/publications/i/item/9789240034228

Creatinine test - Mayo Clinic. (s.f.). https://www.mayoclinic.org/tests-procedures/creatinine-test/about/pac-20384646

Instituto Nacional de Salud Pública. (2022). *Encuesta Nacional de Salud y Nutrición (ENSANUT) 2022: Base de datos de muestras de sangre* [Conjunto de datos]. https://ensanut.insp.mx/

Instituto Nacional de Salud Pública. (2022). *Encuesta Nacional de Salud y Nutrición (ENSANUT) 2022: Base de datos sociodemográficos* [Conjunto de datos]. https://ensanut.insp.mx/

Jarvis, D. J., Adamkiewicz, G., Heroux, M., Rapp, R., & Kelly, F. J. (2010). Nitrogen dioxide. WHO Guidelines For Indoor Air Quality: Selected Pollutants - NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK138707/

NGSP: Clinical use. (s.f.). https://ngsp.org/ADA.asp

Secretaría de Medio Ambiente y Recursos Naturales (SEMARNAT). (2025). *Base de datos de calidad del aire: Contaminantes atmosféricos por entidad federativa* [Conjunto de datos]. https://historico.datos.gob.mx/busca/dataset/indicadores-basicos-del-desempeno-ambiental–atmosfera–calidad-del-aire

Singh, B., Goyal, A., & Patel, B. C. (2025, 3 mayo). C-Reactive Protein: Clinical Relevance and Interpretation. StatPearls - NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK441843/