Artículo 2 — Redes Bayesianas Gaussianas en México

AUTHORS
Jose Angel Govea Garcia
Diego Vértiz Padilla
Daniel Sánchez Fortiz
Augusto Ley Rodriguez

PUBLISHED
September 7, 2025

1 Link al repositorio de Github

https://github.com/Vertiz2405/Redes_Bayesianas_Gaussianas_SP2.git

2 Abstract

Este estudio tuvo como objetivo implementar un modelo probabilístico que permitiera comprender cómo los factores ambientales influyen en la salud de la población mexicana, con el fin de aportar recomendaciones para políticas públicas y estrategias de prevención. Para ello, se utilizaron datos de calidad del aire y biomarcadores clínicos provenientes de ENSANUT, se aplicó un proceso de limpieza y estandarización, y se construyeron redes bayesianas gaussianas evaluadas mediante criterios de información (BICg, AICg, BGe). Los resultados mostraron que la DAG 3 fue la estructura con mejor desempeño, indicando que los contaminantes atmosféricos (NO, CO y SO_2) influyen en biomarcadores clave como homocisteína, creatinina y hemoglobina glucosilada, los cuales a su vez medían efectos en lípidos y riesgo metabólico. Las consultas probabilísticas evidenciaron asociaciones comprobadas, destacando la relación del CO con alteraciones renales y metabólicas. En conclusión, el modelo confirma la utilidad de las redes bayesianas gaussianas para explorar dependencias complejas entre contaminación y salud, aunque futuras investigaciones deben ampliar el espectro de contaminantes, considerar variables categóricas y validar los hallazgos con estudios pertinentes.

3 Introducción

La calidad del aire en México constituye uno de los principales retos de salud pública, ya que la exposición crónica a contaminantes como dióxido de azufre (SO_2), óxidos de nitrógeno (NOx) y monóxido de carbono (CO) se ha vinculado con un aumento en enfermedades cardiovasculares, respiratorias y metabólicas (Secretaría de Salud, 2022; World Health Organization [WHO], 2021). Estos contaminantes pueden alterar biomarcadores clave, como la hemoglobina glucosilada, los lípidos en sangre o la función renal, lo que los convierte en variables fundamentales para estudiar la relación entre ambiente y salud.

El objetivo de este trabajo es implementar un modelo probabilístico que ayude a comprender cómo los factores ambientales influyen en la salud de la población mexicana, con el fin de ofrecer insumos para políticas públicas y estrategias de prevención. Dicho modelo busca integrar información sobre contaminantes atmosféricos y biomarcadores clínicos, permitiendo visualizar relaciones causales y dependencias condicionales que no siempre son evidentes con enfoques estadísticos tradicionales.

En este contexto, las redes bayesianas gaussianas (RBG) ofrecen una herramienta poderosa, pues permiten representar la estructura de dependencias entre variables continuas bajo supuestos de normalidad multivariada. Cada nodo de la red representa una variable, y las aristas describen relaciones de dependencia probabilística. Esto facilita la estimación de probabilidades conjuntas y condicionales, incluso en presencia de datos incompletos, así como la simulación de escenarios hipotéticos (Koller & Friedman, 2009). Su importancia radica en que integran el conocimiento experto con los datos empíricos, generando modelos interpretables que apoyan la toma de decisiones en salud pública y políticas ambientales (Pearl, 2009).

De esta manera, aplicar redes bayesianas gaussianas a la relación entre contaminación y salud en México no solo permite identificar qué contaminantes afectan más a ciertos biomarcadores, sino también diseñar intervenciones más focalizadas, orientadas a la prevención y al cuidado integral de la población.

4 Metodología

4.1 Limpieza de Datos

4.1.1 Limpieza 1

Durante la primera limpieza se integraron y depuraron las bases de datos originales de ENSANUT. Primero se cargaron los archivos de muestras biológicas y de datos sociodemográficos.

Luego se revisaron los nombres de todas las variables y se identificaron aquellas que se repetían en ambos archivos. Para nuestro caso, la variable “FOLIO_INT” fue la que seleccionamos para unir ambos datasets.

En estos casos se resolvieron las duplicaciones unificando la información: cuando una misma variable aparecía con dos sufijos diferentes (por ejemplo .x y .y), se consolidó en una sola columna, conservando los valores no nulos. Finalmente, se generó una nueva base integrada llamada datos.csv, que contiene la información de cada persona con sus indicadores de salud y características demográficas, sin redundancias ni columnas duplicadas.

FOLIO_INT	t_hora	t_min	t_sumai	t_sumaf	hora_ini_1	fecha_ini_1	hora_fin_1	fecha_fin_1	tiempo1
2022_01001004_01	13	35	813	815	13:33:14	10/11/2022	13:35:20	10/11/2022	2
2022_01001008_01	20	44	1244	1244	20:34:12	20/11/2022	20:34:55	20/11/2022	0
2022_01001008_03	20	43	1243	1243	20:41:51	20/11/2022	20:42:23	20/11/2022	1

FOLIO_INT	t_hora	t_min	t_sumai	t_sumaf	hora_ini_1	fecha_ini_1	hora_fin_1	fecha_fin_1	tiempo1
2022_01001009_02	9	26	561	566	09:21:48	10/11/2022	09:26:03	10/11/2022	5
2022_01001011_01	20	49	1249	1249	20:45:31	20/11/2022	20:46:06	20/11/2022	1
2022_01001011_03	20	50	1250	1250	20:47:47	20/11/2022	20:48:21	20/11/2022	1

4.1.2 Limpieza 2

El objetivo de la segunda limpieza es incluir los datos de calidad de aire, recolectados por SEMARNAT en el 2025, con la información previamente del ENSANUT acerca de estudios de sangre de personas y sus respectivos datos demográficos. Asumiremos que esta información son estimaciones adecuadas para los contaminantes del 2022.

Los datos con los que contamos en calidad de aire son los siguientes (las 5 observaciones son ejemplo):

Entidad_federativa	Municipio	Tipo_de_Fuente	SO_2	CO	NOx	COV	PM_010	PM_
Aguascalientes	Aguascalientes	Fuentes fijas	546.316	94.308	209.404	2219.396	246.873	182
Aguascalientes	Aguascalientes	Fuentes de área	15.877	2396.298	333.306	12438.827	1567.037	483
Aguascalientes	Aguascalientes	Fuentes móviles carreteros	308.001	48527.124	10486.580	5148.961	721.389	659
Aguascalientes	Aguascalientes	Fuentes móviles que no circulan por carretera	36.981	548.796	1042.305	96.899	66.318	63
Aguascalientes	Aguascalientes	Fuentes naturales	NA	NA	440.268	1663.833	NA	
Aguascalientes	Asientos	Fuentes fijas	0.195	0.683	1.303	0.032	451.809	333

Lo primero que hicimos fue determinar un tipo de fuente de datos, debido a la magnitud de estos, nos quedamos solo con los datos provenientes de fuentes fijas. También quitamos la columna llamada “Entidad”, ya que incluye la misma información que la columna llamada “Entidad_federativa”. La tabla actualizada la vemos de la siguiente manera:

	Entidad_federativa	Municipio	Tipo_de_Fuente	SO_2	CO	NOx	COV	PM_010	PM_
1	Aguascalientes	Aguascalientes	Fuentes fijas	546.316	94.308	209.404	2219.396	246.873	18
6	Aguascalientes	Asientos	Fuentes fijas	0.195	0.683	1.303	0.032	451.809	333

	Entidad_federativa	Municipio	Tipo_de_Fuente	SO_2	CO	NOx	COV	PM_010	PM
19	Aguascalientes	Jesús María	Fuentes fijas	23.235	27.727	39.878	83.952	136.274	9
24	Aguascalientes	Pabellón de Arteaga	Fuentes fijas	43.906	0.959	4.849	0.053	5.977	3
29	Aguascalientes	Rincón de Romos	Fuentes fijas	0.115	1.568	9.004	0.903	9.277	1
38	Aguascalientes	Tepezalá	Fuentes fijas	2611.680	1763.088	2516.466	5.578	175.984	10

El siguiente paso fue verificar el nombre de entidad y municipio de los 2 data frames, esto con el objetivo de ver cómo relacionar ambas tablas y poder unificar la información. En el caso de la tabla aire algunos ejemplos de entidades federativas y municipios son los siguientes:

- Aguascalientes
- Baja California
- Asientos
- Pabellón de Arteaga

Mientras que esta misma información, en la tabla referente a la informacion del ensanut la visualizamos de la siguiente manera:

- 01 AGUASCALIENTES
- 02 BAJA CALIFORNIA
- 002 ASIENTOS
- 006 PABELLÓN DE ARTEAGA

Con esta información decidimos que al primer grupo de datos (SEMARNAT), tanto a las entidades y municipios las pasaríamos a un formato en mayúsculas únicamente. Mientras que en el segundo grupo de datos (ENSANUT), eliminaríamos los números y el primer espacio. Con estos cambios ambas tablas tendrían el mismo formato para sus entidades y muncipios.

Una vez realizado estos cambios realizamos un antijoin entre ambas tablas, esto con la intención de conocer qué elementos seguían sin coincidir. En el caso de las entidades no coincidían 3, para los datos acerca del aire son los siguientes:

- COAHUILA
- MICHOACÁN
- VERACRUZ

Mientras que en el conjunto de datos del ENSANUT aparecían de la siguiente manera:

- COAHUILA DE ZARAGOZA
- MICHOACÁN DE OCAMPO
- VERACRUZ DE IGNACIO DE LA LLAVE

Como se puede leer, estos son exactamente los mismos estados pero pequeñas variaciones en los nombres. Debido a esto, modificamos el segundo conjunto de datos para que coincidiera con el primero

En el caso del los municipios, cuando realizamos un autojoin para conocer aquellas que no coincidía, nos percatamos que no era por variaciones en los nombres, si no porque información de un municipio en cierta tabla, no se encontraba presente en la otra, esto sucedía en ambas direcciones. En este caso no realizamos ninguna modificación.

Una vez realizados estos cambios ya se pudo realizar un inner join, para poder contar en una misma con todos los datos. Esta unión la podemos visualizar en la siguiente tabla:

FOLIO_INT	t_hora	t_min	t_sumai	t_sumaf	hora_ini_1	fecha_ini_1	hora_fin_1	fecha_fin_1	tiempo1
2022_01001004_01	13	35	813	815	13:33:14	10/11/2022	13:35:20	10/11/2022	2
2022_01001008_01	20	44	1244	1244	20:34:12	20/11/2022	20:34:55	20/11/2022	0
2022_01001008_03	20	43	1243	1243	20:41:51	20/11/2022	20:42:23	20/11/2022	1
2022_01001009_02	9	26	561	566	09:21:48	10/11/2022	09:26:03	10/11/2022	5
2022_01001011_01	20	49	1249	1249	20:45:31	20/11/2022	20:46:06	20/11/2022	1
2022_01001011_03	20	50	1250	1250	20:47:47	20/11/2022	20:48:21	20/11/2022	1

4.2 Expertos químicos

Al tener nuestras bases de datos listas para usarse, el siguiente paso fue el comunicar nuestro objetivo y buscar la ayuda de expertos químicos. Los expertos contactados fueron: el Dr. Alberto Sánchez Estrada, ingeniero agrónomo en sistemas de producción agrícola, maestría en fruticultura y doctorado en agricultura protegida. Cuenta con gran conocimiento de el efecto de estos contaminantes en el aire; la Mtra. Judith Fortiz Hernández, ingeniera agroindustrial y maestra en ciencias agrícolas, su enfoque fue parecido al del Dr. Alberto; y por último contamos con la opinión de un experto en salud, el Dr. Víctor Zepeda Fortiz, licenciado en medicina. Para poder empezar el proceso, se desecharon las variables que no irrelevantes para el estudio, solo dejando todos los contaminantes y todos los resultados de las pruebas de salud, en estos no desechamos ninguna hasta obtener las opiniones de los expertos, junto con las variables esenciales sociodemográficas. Se colocaron en un archivo listo para que los expertos pudieran elegir las que creyeran más convenientes y pasar a la proposición de DAGs.

El Dr. Alberto propuso enfocarse en evaluar cómo la exposición a contaminantes atmosféricos se relaciona con biomarcadores sanguíneos y diferencias demográficas. En particular, sugiere:

i. analizar si el amoníaco modula la proteína C reactiva (PCR) mediado por el perfil lipídico (colesterol y triglicéridos);

ii. estudiar la asociación de NOx y compuestos orgánicos volátiles con los niveles de ferritina considerando estrato social, edad, sexo y residencia;

iii. vincular sulfatos, amoníaco y óxidos nitrosos/volátiles con folatos, con especial atención a mujeres gestantes y resultados al nacer;

iv. relacionar CO_2 con albúmina y su posible interacción con colesterol y el estado nutricional; y

v. examinar homocisteína en sinergia con albúmina ante exposición a CO_2, nitritos, amonio y sulfatos, explorando además la sensibilidad por edad y diferencias por sexo.

El Dr. Victor propone centrar el análisis en la relación entre contaminación del aire y salud mediante biomarcadores clave y patrones de uso de servicios de salud. En particular, sugiere:

i. priorizar la proteína C reactiva (PCR) como biomarcador principal y compararla frente a distintos contaminantes —con énfasis en PM2.5—, además de estratificar por variables sociodemográficas (entidad, edad, sexo, antecedente de COVID-19, zona de trabajo y estrato);

ii. evaluar asociaciones específicas de contaminantes con enfermedades respiratorias, cardiovasculares y metabólicas;

iii. explorar si vitaminas (B12 y folatos) modulan la necesidad de consulta médica;

iv. formular preguntas guía sobre la relación entre niveles de contaminantes y elevación de PCR; y

v. analizar qué contaminantes se vinculan más con enfermedades crónicas.

La Mtra. Judith remitió un listado estructurado de variables para el análisis de contaminación y salud. En particular, especifica:

i. **estratificadores sociodemográficos**: entidad, municipio, sexo, edad, estrato, zona de trabajo y antecedentes (necesidad de atención médica en 3 meses y haber sido paciente COVID);

ii. **panel de biomarcadores** para inflamación, metabolismo, función renal y micronutrientes: proteína C reactiva (PCR), ferritina, receptor soluble de transferrina (sTfR), perfil lipídico (colesterol total, HDL, LDL, triglicéridos), HbA1c, glucosa promedio estimada (EAG), insulina, creatinina, albúmina, ácido úrico, homocisteína, folato, vitaminas B12 y D;

iii. **variables ambientales**: tipo de fuente emisora y niveles de SO_2, CO, NOx, compuestos orgánicos volátiles (COV), PM_{10}, $PM_{2.5}$ y NH_3.

En conjunto, su aporte funciona como glosario/plantilla de variables para vincular exposición a contaminantes con biomarcadores y demanda de atención, habilitando comparaciones por edad, sexo, entidad, estrato y ocupación; no formula hipótesis explícitas, sino que sienta la base para construir preguntas de investigación y modelos estadísticos.

De manera convergente, las tres recomendaciones coinciden en la necesidad de vincular la exposición a contaminantes atmosféricos con biomarcadores sanguíneos y analizarlos bajo un marco sociodemográfico, dando a entender que las variables que les fueron comparidas sí tenían relevancia. Tanto el Dr. Alberto como el Dr. Victor subrayan la importancia de la proteína C reactiva (PCR) y otros marcadores de inflamación en relación con contaminantes como PM2.5, NOx, COV y amoníaco, mientras que la Mtra. Judith aporta un glosario estructurado que refuerza esta misma línea al incluir biomarcadores de inflamación, metabolismo y micronutrientes junto con variables ambientales y sociales. En conjunto, los tres enfoques concluyen en priorizar la PCR como indicador central, considerar múltiples contaminantes relevantes y dar peso a los factores demográficos para sustentar modelos comparativos y preguntas de investigación sólidas.

4.3 DAGs

Para la creación de las DAGS utilizamos las variables en común de los 3 expertos, asimismo nos basamos completamente en sus opiniones. Aunque no utilizamos todas las variables de los expertos en sus DAGS ya que necesitamos hacer una comparación de las DAGS para saber cuál fue la mejor. Y así utilizarla para las queries que ellos mismos nos propusieron.

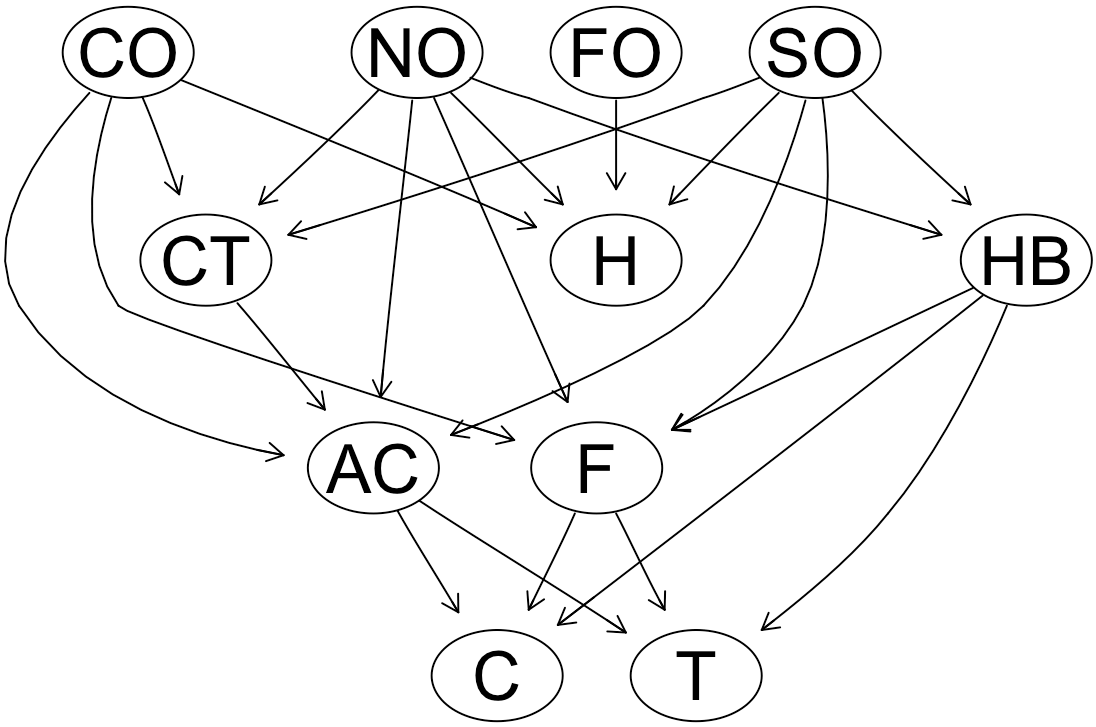
Pero antes de la creación de las DAGs hicimos una pequeña limpieza para poder quedarnos con un dataframe con solo las variables. Lo primero que hicimos fue seleccionar las variables de interés. Al tenerlas, eliminamos las observaciones que tuvieran algún valor faltante. Ya teniendo esto listo renombramos las columnas para tener mayor control que son las siguientes: E = Entidad M = Municipio ES = Estrato S = Sexo C19 = COVID 19 NO = Nivel NOx CO = Nivel de COv SO = Nivel de S0_2 MN = Micronutrientes C = Colesterol Total T = Triglicéridos F = valor de Ferritina FO = Valor de Fol. H = Valor de HCST CT = Valor de Creatinina AC = Valor de Ácido Úrico HB = Valor de Hemoglobina Glucosilada

Después lo que hicimos fue estandarizar la variable C19, si la variable es 1,2,3,4 = 1 (si ha tenido COVID) y si la variable es 98 ,99 = 0 (no ha tenido COVID). Esto para facilitar las DAGs. Como la mayoría de nuestras variables eran caracteres y tenían comas (,) en lugar de puntos (.). Las convertimos a numéricas y remplazamos las comas por puntos. Como por el momento vamos a hacer las DAGs con solo variables continuas y no categóricas o dicotómicas eliminamos las siguientes columnas (variables). E (Entidad), M (Municipio), S (Sexo), ES(Estrato) C19 (COVID) y el MN (Micronutrientes). Y así se ve nuestros dataframe final.

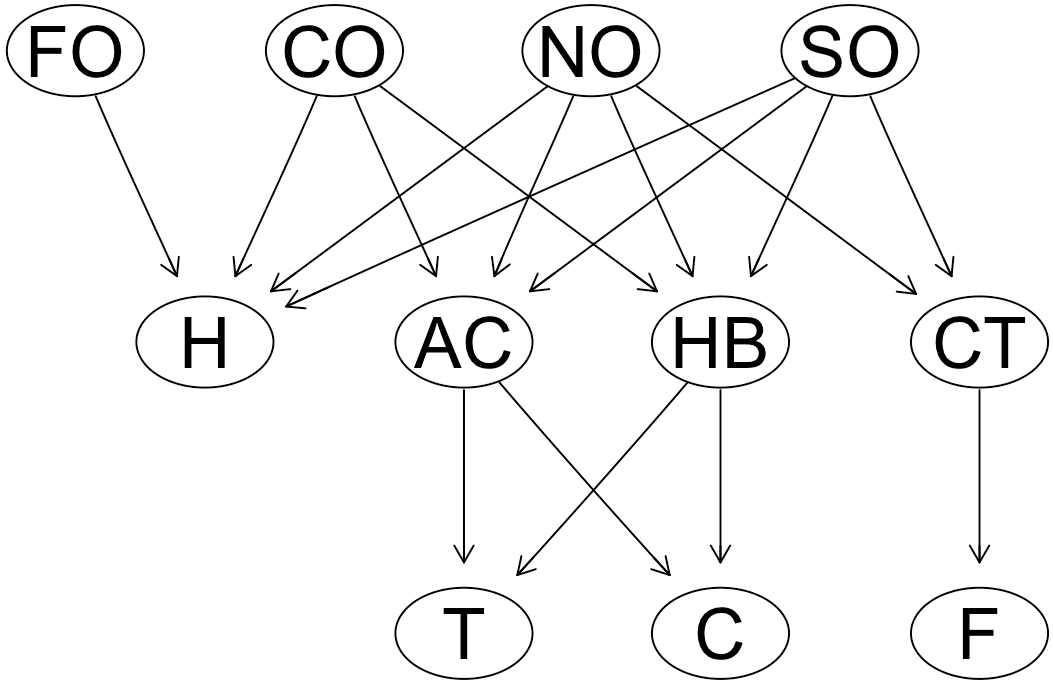
NO	CO	SO	C	T	F	FO	H	CT	AC	HB
209.404	2219.396	546.316	139	121	44	106	694	0.65	46	59
209.404	2219.396	546.316	168	117	573	141	976	0.69	81	56
209.404	2219.396	546.316	161	146	194	138	687	0.60	34	52
209.404	2219.396	546.316	169	50	12	195	739	0.54	39	44
209.404	2219.396	546.316	162	129	175	219	934	0.82	42	54
209.404	2219.396	546.316	149	94	704	174	926	1.01	72	53

Ya teniendo nuestro dataframe final creamos las 3 DAGS.

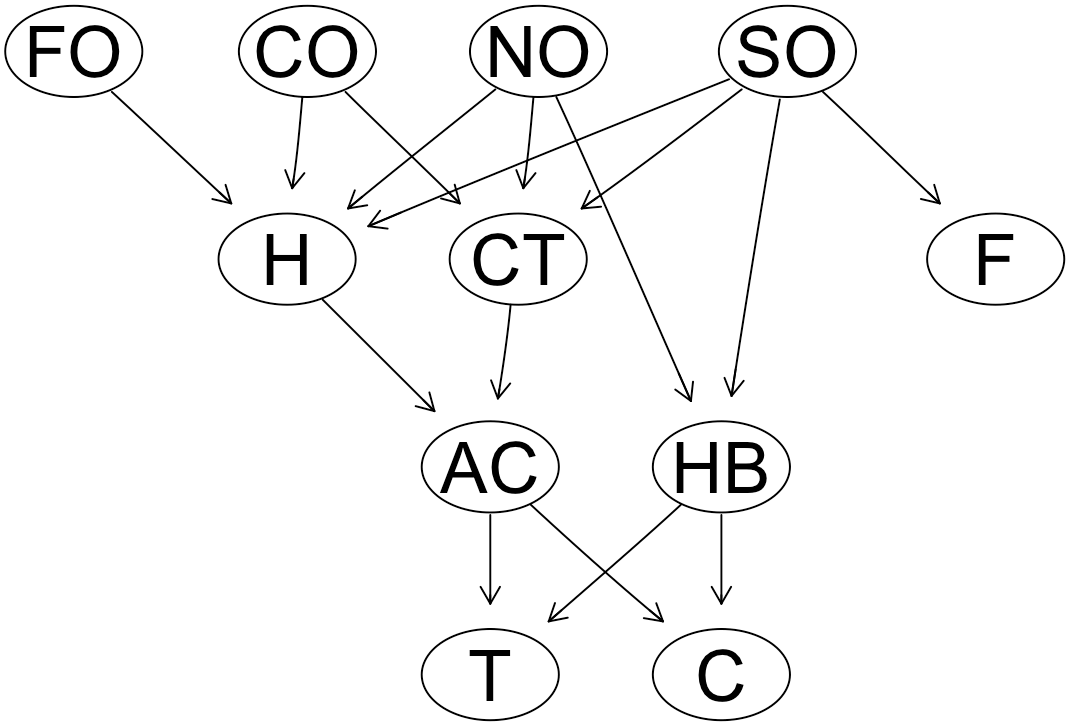
DAG 1:



DAG 2:



DAG 3:



4.4 Evaluación de DAGs

A partir de la selección de variables continuas y siguiendo la propuesta de los expertos, se definieron tres estructuras candidatas de red bayesiana (DAG1, DAG2 y DAG3). Posteriormente, se evaluaron sus desempeños mediante criterios de información para distribuciones gaussianas: **BICg**, **AICg** y **BGe**, todos implementados en `bnlearn`.

Los resultados obtenidos se resumen en la siguiente tabla:

DAG	BICg	AICg	BGe
dag1	-34816.00	-34721.31	-35252.25
dag2	-34803.71	-34721.64	-35165.96
dag3	-34798.21	-34718.25	-35143.19

De acuerdo con los resultados, **la DAG 3 presenta el mejor desempeño**, al obtener los valores más altos (menos negativos) tanto en BICg como en AICg, superando a las otras dos propuestas. Por ello, se selecciona la **DAG 3** como la estructura base para los análisis posteriores. El criterio **BGe** también se reporta como medida adicional de robustez, aportando una perspectiva bayesiana complementaria.

Es decir, el mejor modelo relacional, dados los datos y la opinión de los expertos, sugiere que los contaminantes del aire (NO, CO y SO_2) impactan de manera directa marcadores metabólicos y renales clave: elevan homocisteína (H) y se asocian con mayor creatinina (CT); además, SO_2 también se vincula con ferritina (F) (reserva de hierro e inflamación). El folato (FO) actúa sobre homocisteína, consistente con su papel bioquímico en su metabolismo (posible efecto “protector”). A partir de ahí, H y CT

determinan ácido úrico (AC), y en paralelo NO y SO_2 influyen directamente en el control glucémico crónico (HbA1c, HB). Finalmente, HB y AC son los padres de colesterol total (C) y triglicéridos (T), lo que indica que la disglucemia y el metabolismo de purinas (ácido úrico) median el camino entre la exposición a contaminantes y la dislipidemia/riesgo cardiometabólico. En conjunto, el DAG plantea una cadena plausible: contaminación → (homocisteína/función renal) → ácido úrico y HbA1c → lípidos, con folatos modulando el eje de homocisteína.

4.5 Variables categóricas

Un punto de interés muy fuerte en la creación de un modelo como el empleado es el sexo de la persona. Sin embargo, las redes bayeasianas gaussianas solo nos permiten trabajar con variables continuas que siguen una distribución normal multivariada. Por lo que no es posible incluir variables categóricas dentro de este marco.

4.5.1 Discretización

El primer método consiste enuna discretización de todas las variables continuas, de modo que, estas variables ahora estarán representadas en un espacio discreto. Este cambio hace que pasemos de un modelo bayesiano gaussiano a un modelo bayesiano discreto. Diversos métodos han sido propuestos para elegir los puntos de corte, entre ellos estrategias simples (intervalos de igual amplitud o frecuencia) y enfoques más avanzados que incorporan la discretización dentro del propio proceso de aprendizaje de la red, buscando un equilibrio entre la complejidad del modelo y su capacidad de representar los datos (Friedman & Goldszmidt, 1996).

Este procedimiento presenta ciertas limitaciones. En primer lugar, implica una pérdida de información, ya que se reducen valores continuos a categorías. Además, una discretización inadecuada puede llevar a modelos poco generalizables o que no reflejen adecuadamente las relaciones entre variables.

4.5.2 Modelos mixtos/hibridos

El segundo enfoque corresponde a las redes bayesianas gaussinas condicionales (CGBN), también llamadas redes híbridas. En este marco, las variables categóricas se incluyen como nodos que condicionan la distribución de las variables continuas. Esto significa que cada categoría de una variable discreta define una distribución gaussiana distinta sobre las variables continuas relacionadas. Este enfoque permite modelar de manera más fiel la interacción entre categorías y mediciones numéricas, evitando la necesidad de discretizar los datos originales (Lauritzen & Wermuth, 1989). No obstante, su implementación es más compleja

4.5.3 Implementación

En la implementación con el paquete bnlearn en R nos encontramos con una limitación importante. No es posible combinar variables categóricas y continuas en una misma red. Este nos exige que los datos sean homogéneos, es decir, que todas las variables sean discretas o que todas sean continuas. Esto impide la construcción de redes híbridas dentro de este entorno, incluso si se recodifican las variables categóricas. (Scutari, 2010) En el trabajo realizado se intentó incluir variables como el sexo y el estrato

social en una red bayesiana gaussiana, pero el modelo rechazó estas variables al no cumplir con los supuestos previamente mencionaods.

5 Aplicación

En esta sección usamos la DAG seleccionada para responder preguntas clínicas y químicas con probabilidades condicionadas. Para cada pregunta definimos un evento de interés (p.ej., "colesterol o triglicéridos altos") y comparamos su probabilidad en la población general (línea base) contra su probabilidad cuando un contaminante está alto (percentil 90). Reportamos: • Prob.: $P(\text{evento} \mid \text{evidencia})$. • $\Delta(p.p.)$: diferencia en puntos porcentuales vs. la línea base. • Lift: El lift se define como la razón $\frac{P(\text{evento}|\text{evidencia})}{P(\text{evento})}$.Un lift > 1 indica mayor riesgo asociado al contaminante elevado.

Los umbrales (percentiles) se definieron al inicio de esta sección y las probabilidades se estimaron con muestreo por importancia (cp_dist, método "ls") sobre el modelo ajustado. (Koller & Friedman, 2009)

Aunque bnlearn ofrece cpquery() para calcular $P(\text{evento} \mid \text{evidencia})$, en redes continuas/gaussianas su evaluación interna del evento y la evidencia puede exigir vectores lógicos estrictos y fallar según la versión o la forma de las expresiones. (Scutari, 2017) Por robustez usamos un estimador por muestreo condicional con cpdist() (logic sampling) y rechazo de muestras que no cumplen la evidencia. Implementado en el helper p_rs(), esto equivale a estimar la probabilidad condicional como el promedio de la condición del evento en las muestras que satisfacen la evidencia.(Neapolitan, 2004)

Ventajas: (i) funciona de forma estable con variables continuas y umbrales, (ii) permite escribir eventos/evidencias complejos con quote(...) sin pelear con coerciones internas, y (iii) mantiene la misma interpretación probabilística que cpquery. Para controlar la variabilidad Monte Carlo usamos un número grande de simulaciones (n, típicamente 10^{5-5}) y fijamos semilla (set.seed) para reproducibilidad.(Robert & Casella, 2004)

5.0.1 Query 1: Dislipidemia (colesterol o triglicéridos altos)

¿Cuando aumentan los niveles de contaminantes (NO, CO, SO), se incrementa la probabilidad de observar colesterol o triglicéridos elevados?

En esta consulta se estima la probabilidad de presentar colesterol total o triglicéridos elevados cuando los niveles de contaminantes (NO, CO, SO_2) están altos, comparándola con la probabilidad base sin evidencia.

Query	Contaminante	Prob	Base	Δ (p.p.)	Lift
Q1: Dislipidemia (C o T altos)	NO alto (p90)	55.0%	56.9%	-1.8	0.97
Q1: Dislipidemia (C o T altos)	CO alto (p90)	57.2%	56.9%	0.3	1.01
Q1: Dislipidemia (C o T altos)	SO_2 alto (p90)	57.3%	56.9%	0.4	1.01
Q1: Dislipidemia (C o T altos)	Base (sin evidencia)	56.9%	56.9%	0.0	1.00

5.0.2 Query 2: HbA1c alta (diabetes / resistencia a la insulina)

¿Cuando aumentan los niveles de contaminantes (NO, CO, SO), se incrementa la probabilidad de observar HbA1c elevada?

En esta consulta se evalúa la probabilidad de que la hemoglobina glucosilada (HbA1c) esté por encima del umbral (indicador de diabetes o resistencia a la insulina) cuando los contaminantes NO, CO o SO_2 son altos, comparado con la probabilidad base.

Query	Contaminante	Prob	Base	Δ (p.p.)	Lift
Q2: HbA1c alta (diabetes/resistencia)	NO alto (p90)	33.9%	41.3%	-7.4	0.82
Q2: HbA1c alta (diabetes/resistencia)	CO alto (p90)	41.7%	41.3%	0.4	1.01
Q2: HbA1c alta (diabetes/resistencia)	SO_2 alto (p90)	43.3%	41.3%	2.0	1.05
Q2: HbA1c alta (diabetes/resistencia)	Base (sin evidencia)	41.3%	41.3%	0.0	1.00

5.0.3 Query 3: Creatinina alta (función renal)

¿Cuando aumentan los niveles de contaminantes (NO, CO, SO), se incrementa la probabilidad de observar creatinina elevada?

Aquí se estima la probabilidad de que la creatinina se encuentre elevada (marcador de alteración en la función renal) cuando los contaminantes están altos, en comparación con la probabilidad base.

Query	Contaminante	Prob	Base	Δ (p.p.)	Lift
Q3: Creatinina alta (función renal)	NO alto (p90)	38.0%	41.6%	-3.6	0.91
Q3: Creatinina alta (función renal)	CO alto (p90)	44.4%	41.6%	2.9	1.07
Q3: Creatinina alta (función renal)	SO_2 alto (p90)	40.2%	41.6%	-1.4	0.97
Q3: Creatinina alta (función renal)	Base (sin evidencia)	41.6%	41.6%	0.0	1.00

5.0.4 Query 4: Riesgo metabólico compuesto

¿Niveles altos de contaminantes (NO, CO, SO) se asocian con ferritina o folatos anormales y mayor riesgo cardiovascular (C, T, HB, CT, H)?

En este caso se agrupan varios marcadores (colesterol, triglicéridos, HbA1c, homocisteína y ferritina) para evaluar un riesgo metabólico compuesto. Se compara la probabilidad de alteraciones cuando los contaminantes están altos contra la base.

Query	Contaminante	Prob	Base	Δ (p.p.)	Lift
Q4: Riesgo metabólico compuesto	NO alto (p90)	85.3%	87.6%	-2.2	0.97
Q4: Riesgo metabólico compuesto	CO alto (p90)	89.1%	87.6%	1.6	1.02
Q4: Riesgo metabólico compuesto	SO_2 alto (p90)	88.5%	87.6%	0.9	1.01
Q4: Riesgo metabólico compuesto	Base (sin evidencia)	87.6%	87.6%	0.0	1.00

5.0.5 Query 4 (opcional): Componente nutricional

También se explora la relación entre contaminantes y biomarcadores nutricionales: ferritina (almacenamiento de hierro) y folatos (asociados a metabolismo de homocisteína).

Indicador	Probabilidad
Ferritina alta | NO alto	43.3%
Ferritina alta | SO_2 alto	44.9%
Folatos bajos | NO alto	46.6%
Folatos bajos | SO_2 alto	46.4%

5.0.6 Resumen combinado de queries

Finalmente, se integran los resultados de todas las consultas en una tabla resumen para facilitar su interpretación.

Query	Contaminante	Prob	Base	Δ (p.p.)	Lift
Q1: Dislipidemia (C o T altos)	NO alto (p90)	55.0%	56.9%	-1.8	0.97
Q1: Dislipidemia (C o T altos)	CO alto (p90)	57.2%	56.9%	0.3	1.01
Q1: Dislipidemia (C o T altos)	SO_2 alto (p90)	57.3%	56.9%	0.4	1.01
Q1: Dislipidemia (C o T altos)	Base (sin evidencia)	56.9%	56.9%	0.0	1.00
Q2: HbA1c alta (diabetes/resistencia)	NO alto (p90)	33.9%	41.3%	-7.4	0.82
Q2: HbA1c alta (diabetes/resistencia)	CO alto (p90)	41.7%	41.3%	0.4	1.01
Q2: HbA1c alta (diabetes/resistencia)	SO_2 alto (p90)	43.3%	41.3%	2.0	1.05
Q2: HbA1c alta (diabetes/resistencia)	Base (sin evidencia)	41.3%	41.3%	0.0	1.00
Q3: Creatinina alta (función renal)	NO alto (p90)	38.0%	41.6%	-3.6	0.91
Q3: Creatinina alta (función renal)	CO alto (p90)	44.4%	41.6%	2.9	1.07
Q3: Creatinina alta (función renal)	SO_2 alto (p90)	40.2%	41.6%	-1.4	0.97
Q3: Creatinina alta (función renal)	Base (sin evidencia)	41.6%	41.6%	0.0	1.00
Q4: Riesgo metabólico compuesto	NO alto (p90)	85.3%	87.6%	-2.2	0.97
Q4: Riesgo metabólico compuesto	CO alto (p90)	89.1%	87.6%	1.6	1.02
Q4: Riesgo metabólico compuesto	SO_2 alto (p90)	88.5%	87.6%	0.9	1.01
Q4: Riesgo metabólico compuesto	Base (sin evidencia)	87.6%	87.6%	0.0	1.00

6 Conclusiones

El análisis desarrollado mostró que los contaminantes atmosféricos (NO, CO y SO_2) sí presentan vínculos con diversos biomarcadores clínicos relevantes para la salud de la población mexicana. En particular, los resultados de las queries indican que: i) el impacto sobre lípidos (colesterol y

triglicéridos) es marginal, con cambios mínimos en la probabilidad de dislipidemia; ii) la exposición a SO_2 se asocia a un incremento ligero en la probabilidad de HbA1c elevada, mientras que NO incluso se relaciona con una reducción de este riesgo en la muestra; iii) en el caso de la función renal, el CO mostró la mayor asociación con niveles altos de creatinina; y iv) en un riesgo metabólico compuesto, el CO nuevamente resaltó como el contaminante con mayor incremento de probabilidad, aunque en general los cambios fueron pequeños

Estos hallazgos son relevantes porque confirman, con un modelo probabilístico robusto, la utilidad de las redes bayesianas gaussianas para explorar dependencias complejas entre variables ambientales y biomarcadores de salud, ofreciendo un marco flexible para generar insumos que orienten políticas públicas y estrategias de prevención. No obstante, también se identifican limitaciones: el análisis se restringió a un conjunto reducido de contaminantes y biomarcadores, los datos ambientales fueron estimados con supuestos temporales y no se incorporaron variables categóricas sociodemográficas que podrían matizar las asociaciones.

Para investigaciones futuras, se sugiere ampliar la muestra a más regiones y contaminantes (incluyendo $PM_{2.5}$, ozono y compuestos orgánicos volátiles), probar modelos híbridos que integren variables continuas y categóricas, así como validar los resultados en estudios longitudinales que permitan confirmar direccionalidad causal. De esta manera, se podrán afinar los modelos y fortalecer la evidencia disponible para el diseño de estrategias de prevención en salud ambiental en México. # Referencias

Koller, D., & Friedman, N. (2009). Probabilistic Graphical Models: Principles and Techniques. MIT Press.

Pearl, J. (2009). Causality: Models, Reasoning and Inference (2nd ed.). Cambridge University Press.

Secretaría de Salud. (2022). Informe sobre la calidad del aire y salud en México. Gobierno de México.

World Health Organization. (2021). Air pollution. https://www.who.int/health-topics/air-pollution

Friedman, N., & Goldszmidt, M. (1996). Discretizing Continuous Attributes While Learning Bayesian Networks. Proceedings of the Thirteenth International Conference on Machine Learning. Recuperado de https://www.cs.huji.ac.il/w~nir/Papers/FrG2.pdf

Lauritzen, S. L., & Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. The Annals of Statistics.

Neapolitan, R. E. (2004). Learning Bayesian Networks. Pearson Prentice Hall.

Robert, C. P., & Casella, G. (2004). Monte Carlo Statistical Methods (2nd ed.). Springer.

Scutari, M. (2010). Learning Bayesian Networks with the bnlearn R Package. Journal of Statistical Software, 35(3), 1–22. https://doi.org/10.18637/jss.v035.i03

Scutari, M. (2017). Bayesian Network Models of Discrete and Continuous Data. British Journal of Mathematical and Statistical Psychology, 70(2), 272–291. https://doi.org/10.1111/bmsp.12073