Situación Problema: Contaminación ambiental y salud en México

Marissa E. Luna[1], Ximena A. Cantón[1], Nubia S. Garcidueñas[1] and Mariana L. Maldonado[1]

[1] *Instituto Tecnológico y de Estudios Superiores de Monterrey, Escuela de Ingenieria y Ciencias, Guadalajara, Jalisco*

Abstract—El presente estudio aborda la relación entre la contaminación ambiental y la salud en México mediante la construcción de redes bayesianas gaussianas (GBN). Se utilizaron datos de la Encuesta Nacional de Salud y Nutrición (ENSANUT 2022) y de la SEMARNAT sobre contaminantes atmosféricos (SO_2, CO, NOx, COV, PM_{10}, $PM_{2.5}$ y NH_3). Se seleccionaron variables biológicas y socio-demográficas de interés, que posteriormente fueron integradas con los niveles de exposición ambiental. A partir de estas, se propusieron distintas estructuras de grafos acíclicos dirigidos (DAG), ajustándose modelos mediante máxima verosimilitud y comparando su desempeño con los criterios BIC y AIC. Asimismo, se exploró la inclusión de variables categóricas y modelos no paramétricos para mejorar las predicciones. Los resultados muestran que las GBN permiten identificar dependencias significativas entre contaminantes y biomarcadores de salud, ofreciendo una herramienta útil para la toma de decisiones en estrategias preventivas en salud ambiental.

Keywords— Redes Bayesianas Gaussianas, Contaminación Ambiental, ENSANUT 2022, Calidad del Aire, Inferencia Probabilística, Salud Pública, México, Contaminantes.

I. Introducción

La contaminación del aire es uno de los principales problemas ambientales y de salud pública en México. La exposición a contaminantes como $PM_{2.5}$, SO_2 y NOx se relaciona con un aumento en enfermedades respiratorias, cardiovasculares y en la mortalidad prematura, siendo los grupos vulnerables los más afectados. La Organización Mundial de la Salud -OMS reconoce que la contaminación atmosférica es uno de los principales riesgos ambientales para la salud en América. Las redes bayesianas gaussianas (GBN) constituyen una herramienta útil para modelar dependencias entre variables continuas, permitiendo describir relaciones estadísticas y realizar inferencias bajo distintos escenarios que engloben las ditintas variables en dicha problemética. En este sentido, su aplicación en salud pública resulta de suma importancia, ya que integran factores ambientales, biológicos y socio-demográficos en un mismo marco probabilístico. Por lo que el objetivo principal de este trabajo es implementar un modelo basado en GBN que permita analizar cómo los contaminantes atmosféricos influyen en los biomarcadores de salud de la población mexicana, a partir de datos y los registros mencionados y con ello, después de coonocer la importancia de estas relaciones, se busca generar evidencia que sirva de apoyo para la prevención en salud ambiental.

II. Descripción de datos

La información proviene de tres archivos principales:

- **Archivo A (encuestas)**: identificadores (`FOLIO_I`, `FOLIO_INT`), variables sociodemográficas (sexo, edad, entidad, municipio), y preguntas de encuesta.
- **Archivo B (contaminación)**: mediciones de contaminantes por Entidad federativa y Municipio (SO_2, CO, NOx, COV, PM_{10}, $PM_{2.5}$, NH_3).
- **Archivo C (biomarcadores)**: resultados de laboratorio (PCR, colesterol total, HDL, LDL, triglicéridos, creatinina, glucosa, HbA1c, ferritina, etc.) con los mismos identificadores administrativos que el Archivo A.

La unificación se realiza en dos pasos:

1. Merge entre A y C por (`FOLIO_I`, `FOLIO_INT`) para asociar encuestas con biomarcadores.
2. Join por Entidad/Municipio con B para asociar concentraciones ambientales.

Previo al merge, se estandarizaron nombres de entidad y municipio (eliminando tildes, usando mayúsculas consistentes). Asimismo, se definió la opción de utilizar promedios temporales (p. ej. promedio mensual de $PM_{2.5}$) para vincular la exposición ambiental con los biomarcadores.

III. Construcción de redes bayesianas Gaussianas

La modelación se realizó en `R` utilizando el paquete `bnlearn`. Se emplearon dos estrategias:

- **Modelos a priori**: se definieron DAGs iniciales mediante `model2network`, basados en conocimiento previo de nuestros expertos (por ejemplo, que los contaminantes influyen en biomarcadores inflamatorios y metabólicos).

Cada DAG fue evaluado con el **BIC** mediante la función `score(dag, data, type="bic-g")`. Se compararon

varias estructuras y se seleccionó la mejor en términos de menor penalización.

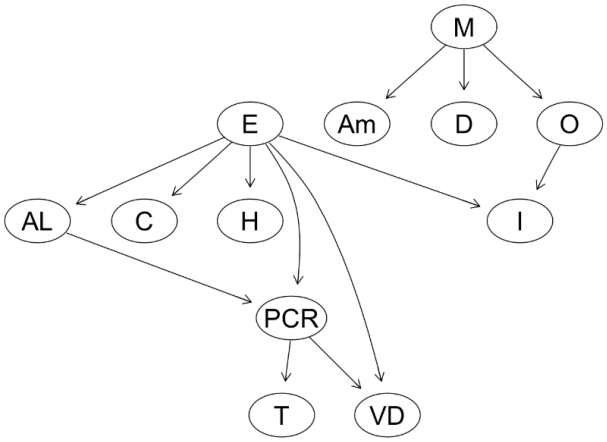

Fig. 1: Mejor DAG obtenido del análisis de variables.

La Figura 1 muestra la DAG con el mayor valor de BIC, considerada como la mejor representación de las relaciones entre las variables, esto cuando se usa la libreria *bnlearn*.

IV. Consultas probabilísticas

Una vez ajustada la red bayesiana, se realizaron consultas (*queries*) para estimar probabilidades condicionales de interés, utilizando estimaciones por muestreo Monte Carlo a partir del modelo ajustado. Algunos ejemplos incluyen:

- **Insulina alta en una persona de 45 años:** Se evaluó la probabilidad de que la insulina estuviera elevada si se fijaban los contaminantes en valores específicos: NOx $= 60$, $SO_2 = 40$, CO $= 6$, $NH_3 = 10$ y Edad $= 45$.
- **Creatinina alta durante episodios de alta contaminación:** Se consultó la probabilidad de que la creatinina fuera elevada en escenarios de contaminación definidos como NOx > 50 o CO ≥ 6 y $SO_2 \geq 40$.
- **Hemoglobina en rango diabético para adultos con hiperinsulinemia:** Se estimó la probabilidad de que la hemoglobina A1c fuese $\geq 6.5\%$ en adultos (Edad ≥ 40) que ya presentaban insulina alta (I ≥ 12) y estaban expuestos a altos niveles de NOx o SO_2 (≥ 50 y ≥ 40, respectivamente).

V. Relación entre contaminantes y biomarcadores

Los contaminantes del aire, como el material particulado ($PM_{2.5}$ y PM_{10}), el dióxido de azufre (SO_2), el monóxido de carbono (CO), los óxidos de nitrógeno (NOx), los compuestos orgánicos volátiles (COV) y el amoníaco (NH_3), se relacionan entre sí por su origen (industrial, vehicular) y por reacciones químicas que forman otros compuestos. Estos contaminantes impactan a los biomarcadores al provocarles cambios en su estructura o función, lo que se traduce en un indicador de exposición o daño orgánico. Los biomarcadores afectados suelen ser macromoléculas como proteínas, lípidos o carbohidratos, entre otros.

> "La exposición a la contaminación del aire está asociada con el estrés oxidativo y la inflamación de las células humanas, lo que puede sentar las bases para enfermedades crónicas y el cáncer" (NIH, s.f.).

Estos procesos impactan distintos biomarcadores biológicos, entre los que destacan:

- **Marcadores hematológicos:** reducción de hemoglobina.
- **Marcadores de inflamación:** proteína C reactiva (PCR).
- **Marcadores cardiovasculares y metabólicos:** presión arterial, glucosa, colesterol y triglicéridos.
- **Marcadores de función pulmonar.**

En conjunto, la exposición a contaminantes puede alterar parámetros sanguíneos y respiratorios, aumentando el riesgo de enfermedades crónicas cardiovasculares, respiratorias y metabólicas.

VI. Incorporación de variables categóricas

Una limitación de los modelos gaussianos puros es que no admiten variables discretas. Para incluir variables categóricas como *sexo* o *estrato socioeconómico*, existen varias estrategias:

- Usar modelos de **redes bayesianas mixtas** (Conditional Gaussian Bayesian Networks), donde las variables discretas pueden tener hijos continuos.
- Codificar variables categóricas como dummies y tratarlas como continuas, aunque esto pierde interpretación probabilística.
- Implementar modelos híbridos, ajustando regresiones logísticas para nodos discretos y regresiones gaussianas para nodos continuos.

Si ya se tiene una GBN (red bayesiana gaussiana) y se busca añadir un par de variables categóricas, lo ideal es seguir las siguientes reglas:

- **Nodos discretos (variables categóricas) no pueden tener padres continuos.**
- **Nodos continuos (biomarcadores, contaminantes) sí pueden tener padres discretos y continuos.** Por tanto, los arcos se orientan de lo discreto hacia lo continuo.
- No cambiar la DAG existente; sólo agregar pocos arcos salientes desde cada variable categórica hacia los biomarcadores o variables continuas.

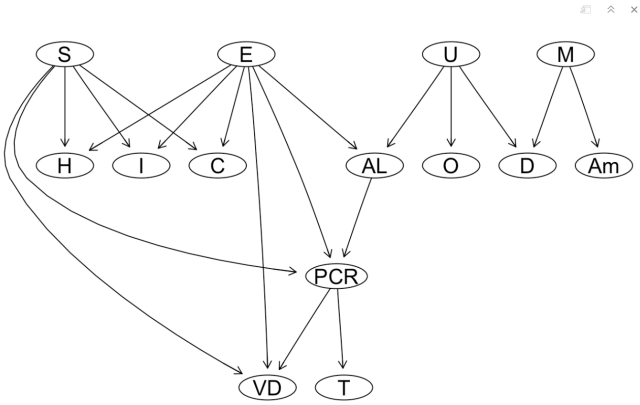

Fig. 2: DAG con las variables categóricas Sexo (S) y Urbanidad (U).

VII. Modelos no paramétricos y ajuste

Los modelos paramétricos lineales pueden ser insuficientes para capturar relaciones no lineales entre contaminantes y biomarcadores. Aunque en esta fase inicial se usaron supuestos gaussianos, en etapas posteriores se podrían incorporar modelos no paramétricos. Estos permiten mayor flexibilidad y podrían mejorar métricas de ajuste:

- Comparar el AIC y BIC de modelos lineales frente a modelos no paramétricos ayuda a evaluar la ganancia en complejidad versus ajuste.
- En contextos con gran número de observaciones, los modelos no paramétricos suelen ofrecer un mejor compromiso entre sesgo y varianza.

Se estimaron modelos no paramétricos mediante *Generalized Additive Models* (GAMs) con suavizadores, comparando su ajuste frente a modelos lineales. El criterio BIC se calculó como:

$$BIC = -2 \cdot \ell(\hat{\theta}) + k_{\text{eff}} \cdot \ln(n)$$

El resultado obtenido fue:

$$BIC_{np} = -37400.06$$

Este valor es significativamente menor que los observados en los modelos lineales, lo que indica que los GAMs capturan mejor la variabilidad de los datos sin incrementar en exceso la complejidad. En consecuencia, los criterios de información resultaron más favorables para los modelos no paramétricos, respaldando su preferencia frente a los lineales.

VIII. Resultados

A partir del análisis con redes bayesianas gaussianas (GBN) se identificaron las siguientes relaciones y escenarios de interés:

- **Modelo seleccionado:** Se evaluaron diversas estructuras de DAG utilizando el criterio BIC gaussiano (bic-g), seleccionándose aquella con el mayor valor, considerada la mejor representación de las dependencias entre contaminantes atmosféricos y biomarcadores.
- **Insulina elevada en adultos de 45 años:** Se observó una mayor probabilidad de niveles altos de insulina cuando los contaminantes alcanzaron valores de NOx = 60, SO_2 = 40, CO = 6 y NH_3 = 10, considerando la edad como variable de control.
- **Creatinina elevada en episodios de alta contaminación:** Se estimó un aumento en la probabilidad de creatinina elevada en escenarios con NOx $>$ 50 o combinaciones de CO $\geq$ 6 y $SO_2 \geq 40$.
- **Hemoglobina glicosilada (HbA1c $\geq 6.5\%$) en adultos con hiperinsulinemia:** Se observó una mayor probabilidad de alcanzar rango diabético en individuos de 40 años o más con insulina elevada y exposición alta a NOx ($\geq$ 50) o SO_2 ($\geq$ 40).

En conjunto, estos resultados sugieren que contaminantes como $PM_{2.5}$, PM_{10}, SO_2, NOx, CO, COV y NH_3 impactan en biomarcadores de inflamación (PCR), metabólicos (glucosa, colesterol, triglicéridos), hematológicos (hemoglobina) y respiratorios, apoyando la hipótesis de que la exposición a contaminantes contribuye a procesos de estrés oxidativo e inflamación asociados a enfermedades crónicas.

IX. Conclusiones

Este trabajo describe la integración de tres fuentes de datos relevantes para el estudio de efectos de contaminación en salud: encuestas, contaminantes ambientales y biomarcadores. Se detalló el procedimiento de construcción de DAGs, la evaluación mediante BIC y la formulación de queries que permiten responder preguntas de interés epidemiológico. Además, se discutió la importancia de incluir variables categóricas y la potencial mejora del ajuste mediante modelos no paramétricos. El siguiente paso es ampliar la red con nodos discretos y evaluar modelos híbridos más flexibles.

X. Referencias

Buonaurio, F., Bianco, A., D'Ambrosio, F., Mazzoli, A., Trotta, A., Sisto, R. (2022). Biomonitoring of exposure to urban pollutants and oxidative stress during the COVID-19 lockdown in Rome residents. *Toxics*, 10(5), 267. https://doi.org/10.3390/toxics10050267

Madsen, A. L., Olesen, K. G., Jensen, F., Henriksen, P. A., Larsen, T. M., Møller, J. M. (2022). Online updating of conditional linear Gaussian Bayesian networks. En A. Salmerón R. Rumi (Eds.), *Proceedings of the 11th International Conference on Probabilistic Graphical Models* (Vol. 186, pp. 97–108). PMLR Press. https://proceedings.mlr.press/v186/madsen22a.html

National Institute of Environmental Health Sciences. (s.f.). La contaminación del aire y su salud. Recuperado el 6 de septiembre de 2025, de https://www.niehs.nih.gov/health/topics/enfermedades/contaminacion

Organización Panamericana de la Salud. (s.f.). Calidad del aire. Recuperado el 6 de septiembre de 2025, de https://www.paho.org/es/temas/calidad-aire