Redes bayesianas gaussianas

Oliver Arturo Casas Pontanillo | A01645764 Erik Abraham Jajan Díaz | A01644648
José Luis Santos Montaño | A01781721

2025-09-07

Table of contents

https://github.com/olivercasas17/MA2014_Redes_bayesianas_gaussianas

0.1 Abstract

Esta investigación examina las relaciones entre contaminantes atmosféricos y biomarcadores de salud mediante redes bayesianas gaussianas. Utilizando datos de ENSANUT 2022 y calidad del aire de SEMARNAT, se modelaron las dependencias entre siete contaminantes y diez biomarcadores sanguíneos. Se construyeron tres DAGs alternativos y se seleccionó el modelo óptimo mediante criterios AIC y BIC. Los resultados proporcionan evidencia cuantitativa del impacto de la contaminación atmosférica en biomarcadores metabólicos, lipídicos e inflamatorios, contribuyendo al diseño de políticas de salud ambiental.

0.2 Introducción

Diversos estudios han demostrado que la exposición crónica a contaminantes atmosféricos se asocia con un aumento en enfermedades respiratorias, cardiovasculares y metabólicas. Estos efectos usualmente se manifiestan en biomarcadores biológicos, los cuales permiten que se evalúe de manera indirecta el impacto de los contaminantes en la salud de las personas. Las redes bayesianas gaussianas nos permiten modelar las dependencias probabilísticas entre variables y generar representaciones gráficas. En este caso, el uso de los datos provenientes de la encuesta nacional ENSANUT 2022, que ofrecen información detallada sobre muestras de sangre e información sociodemográfica, además de información actualizada de la SEMARNAT sobre los contaminantes del aire, nos permitirán *plantear como hipótesis que los contaminantes atmosféricos tienen un efecto significativo sobre determinados biomarcadores de salud.* Com-

prender cómo interactúan factores ambientales con los biomarcadores de salud de la población genera un impacto importante, ya que puede contribuir al diseño de políticas públicas orientadas a la prevención y mitigación de los efectos de la contaminación atmosférica.

0.3 Metodología

El trabajo sigue un enfoque cuantitativo, exploratorio, explicativo e interdisciplinario, analizando las relaciones que tienen contaminantes del aire con biomarcadores de salud en la población, utilizando redes bayesianas gaussianas para representar estas relaciones de variables continuas. Inicialmente, se realizó una limpieza de la base de datos para extraer las variables con datos continuos relevantes para nuestro estudio, además de asegurarnos de que los datos estuvieran en un formato con el que pudiéramos trabajar. Basándonos en información proporcionada por especialistas, se propusieron tres gráficos acíclicos dirigidos, por sus siglas en inglés DAGs, para posteriormente comparar las 3 DAGs con métricas estadísticas, y seleccionar la mejor de las tres para continuar con el análisis. Se incluyó una variable categórica específica, en este caso sexo, y con ayuda de los especialistas en el tema se escogieron tres queries que nos llevarían a conclusiones prometedoras Las principales técnicas que se utilizaron en este trabajo fueron: Análisis de redes bayesianas gaussianas, construyendo grafos acíclicos dirigidos para modelar relaciones de dependencia entre variables continuas Evaluación de modelos, usando métodos estadísticos como el Criterio de Información de Akaike y el Criterio de Información Bayesiano, AIC y BIC por sus siglas en inglés. Inferencia probabilística, consultas condicionales para estimar probabilidades bajo evidencia Los datos utilizados fueron los siguientes: data.csv data_F.csv data_M.csv Los datos crudos fueron recolectados de: calidad_aire_2025.csv ensanut2022_muestras.csv ensanut2022_socdem.csv Las herramientas que utilizamos fueron las siguientes: Lenguaje de programación R Librerías: tidyverse, bnlearn, Rgraphviz Entorno de trabajo R Studio

0.4 Aplicacion

```r
library(tidyverse)
library(bnlearn)
```

```r
data = read_csv("../data/data.csv")
head(data)
```

```
# A tibble: 6 x 17
  valor_ALBU valor_COL_HDL valor_COL_LDL valor_CREAT valor_GLU_SUERO
       <dbl>         <dbl>         <dbl>       <dbl>           <dbl>
```

```
1        3.8            73           130        0.62            111
2        4.1            49           107        0.91            106
3        4.2            41            76        0.71            109
4        3.8            42            84        0.65             98
5        4.1            41           119        0.83            253
6        3.8            44           133        0.66            204
# i 12 more variables: valor_INSULINA <dbl>, valor_PCR <dbl>, valor_TRIG <dbl>,
#   valor_EAG <dbl>, valor_HB1AC <dbl>, SO_2 <dbl>, CO <dbl>, NOx <dbl>,
#   COV <dbl>, PM_010 <dbl>, PM_2_5 <dbl>, NH_3 <dbl>
```

```r
colnames(data) <- c("ALB", "HDL", "LDL", "CR", "GLU", "INS",
  ↪  "PCR", "TRI", "HB1AC", "EAG", "SO2", "CO", "NOx", "COV",
  ↪  "PM10", "PM2.5", "NH3")
```

```r
dag1 = empty.graph(nodes = c("ALB", "HDL", "LDL", "CR",
  ↪  "GLU", "INS", "PCR", "TRI", "HB1AC", "EAG", "SO2", "CO",
  ↪  "NOx", "COV", "PM10", "PM2.5", "NH3"))
```

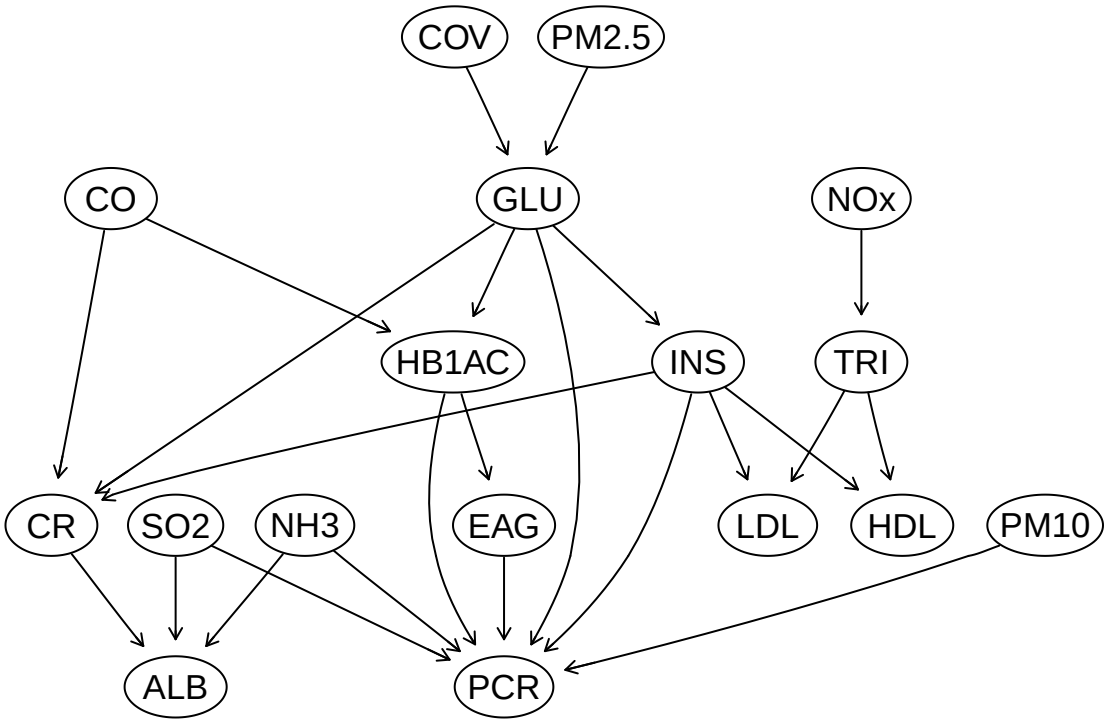
```r
arc_set1 = matrix(c("TRI", "LDL",
                    "TRI", "HDL",
                    "INS", "LDL",
                    "INS", "HDL",
                    "GLU", "INS",
                    "GLU", "HB1AC",
                    "GLU", "PCR",
                    "GLU", "CR",
                    "INS", "CR",
                    "INS", "PCR",
                    "HB1AC", "EAG",
                    "HB1AC", "PCR",
                    "EAG", "PCR",
                    "CR", "ALB",
                    "SO2", "ALB",
```

```
                        "SO2", "PCR",
                        "CO", "CR",
                        "CO", "HB1AC",
                        "NOx", "TRI",
                        "COV", "GLU",
                        "PM10", "PCR",
                        "PM2.5", "GLU",
                        "NH3", "PCR",
                        "NH3", "ALB"), byrow = TRUE, ncol = 2,
                      dimnames = list(NULL, c("from", "to")))
```

```
arcs(dag1) = arc_set1
```

```
graphviz.plot(dag1, shape = "ellipse")
```

```
Loading required namespace: Rgraphviz
```




```
gbn1 = bn.fit(dag1, data = data)
```

```
score_bic_dag1 = score(dag1, data = data, type = "bic-g")
score_bic_dag1
```

```
[1] -97442.1
```

```
score_aic_dag1 = score(dag1, data = data, type = "aic-g")
score_aic_dag1
```

```
[1] -97297.73
```

DAG 2:

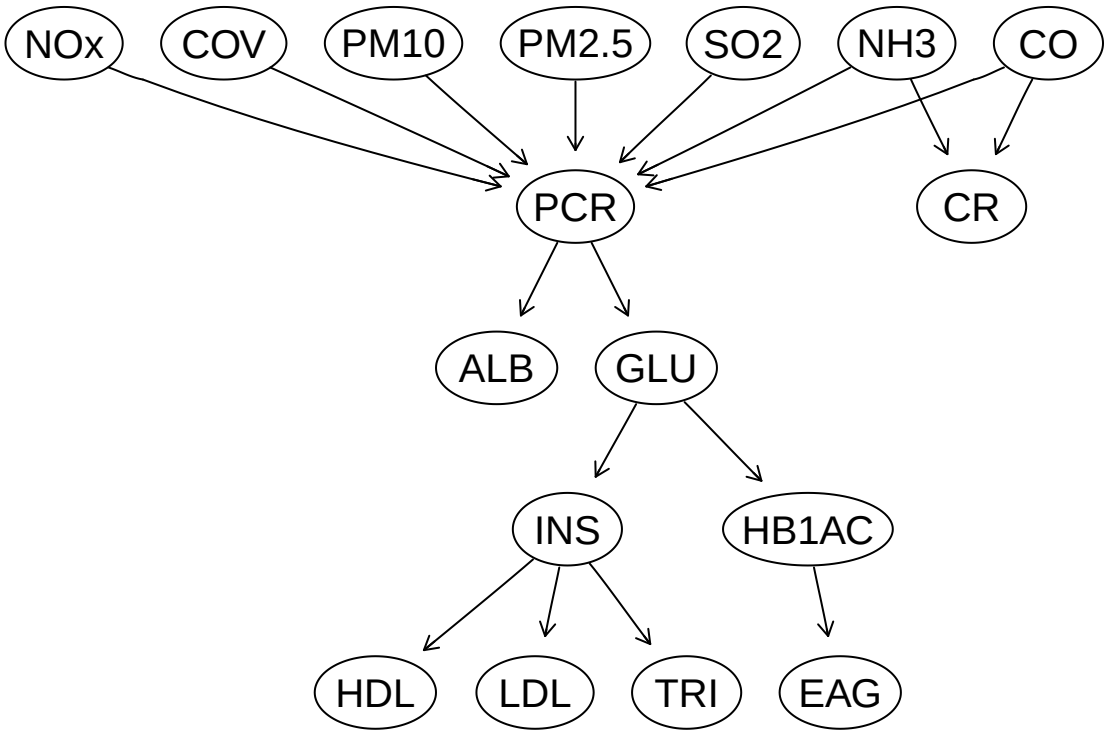
```
dag2 = empty.graph(nodes = c("ALB", "HDL", "LDL", "TRI",
  ↪  "GLU", "INS", "PCR",
                             "HB1AC", "EAG", "CR",
                             "SO2", "CO", "NOx", "COV",
                              ↪  "PM10", "PM2.5", "NH3"))
```

```
arc_set2 = matrix(c("SO2", "PCR",
                    "NOx", "PCR",
                    "COV", "PCR",
                    "PM10", "PCR",
                    "PM2.5", "PCR",
                    "CO", "PCR",
                    "NH3", "PCR",
                    "PCR", "ALB",
                    "NH3", "CR",
                    "CO", "CR",
                    "GLU", "INS",
                    "GLU", "HB1AC",
                    "HB1AC", "EAG",
                    "INS", "HDL",
                    "INS", "LDL",
                    "INS", "TRI",
```

```
                        "PCR", "GLU"), byrow = TRUE, ncol = 2,
                      dimnames = list(NULL, c("from", "to")))
```

```
arcs(dag2) = arc_set2
```

```
graphviz.plot(dag2, shape = "ellipse")
```




```
gbn2 = bn.fit(dag2, data = data)
```

```
score_bic_dag2 = score(dag2, data = data, type = "bic-g")
score_bic_dag2
```

```
[1] -97471.69
```

```
score_aic_dag2 = score(dag2, data = data, type = "aic-g")
score_aic_dag2
```

```
[1] -97344.75
```

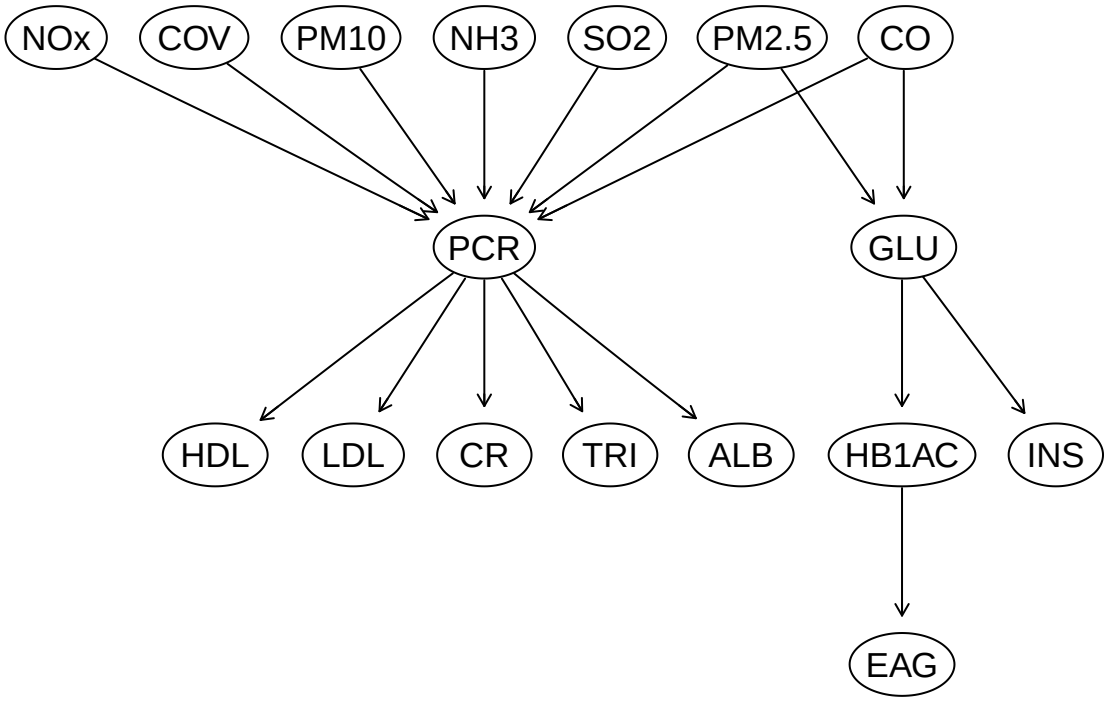
DAG 3:

```
dag3 = empty.graph(nodes = c("ALB", "HDL", "LDL", "CR", "GLU",
 ↪  "INS", "PCR", "TRI", "HB1AC",
                              "EAG", "SO2", "CO", "NOx", "COV",
                                ↪  "PM10", "PM2.5", "NH3"))
```

```
arc_set3 = matrix(c("SO2", "PCR",
                    "CO", "PCR",
                    "NOx", "PCR",
                    "COV", "PCR",
                    "PM10", "PCR",
                    "PM2.5", "PCR",
                    "NH3", "PCR",
                    "PCR", "ALB",
                    "PCR", "LDL",
                    "PCR", "HDL",
                    "PCR", "TRI",
                    "PCR", "CR",
                    "PM2.5", "GLU",
                    "CO", "GLU",
                    "GLU", "INS",
                    "GLU", "HB1AC",
                    "HB1AC", "EAG"), byrow = TRUE, ncol = 2,
                  dimnames = list(NULL, c("from", "to")))
```

```
arcs(dag3) = arc_set3
```

```
graphviz.plot(dag3, shape = "ellipse")
```




```r
gbn3 = bn.fit(dag3, data = data)
```

```r
score_bic_dag3 = score(dag3, data = data, type = "bic-g")
score_bic_dag3
```

```
[1] -97488.91
```

```r
score_aic_dag3 = score(dag3, data = data, type = "aic-g")
score_aic_dag3
```

```
[1] -97361.97
```

```r
#tabla con los scores de las dag para que se vea bonita la
 ↪  comparacion
tabla_scores <- data.frame(
  DAG = c("DAG1", "DAG2", "DAG3"),
  BIC = c(score_bic_dag1, score_bic_dag2, score_bic_dag3),
  AIC = c(score_aic_dag1, score_aic_dag2, score_aic_dag3)
)

```

```
tabla_scores
```

```
   DAG       BIC       AIC
1 DAG1 -97442.10 -97297.73
2 DAG2 -97471.69 -97344.75
3 DAG3 -97488.91 -97361.97
```

Podemos observar que la DAG1 es la que presenta un menor error por lo que va a ser la que usaremos en el resto del análisis. Una forma de agregar una variable categórica, en este caso sexo, a la red gaussiana sería recodificando la variable para volverla continua, o separar nuestra base de datos en una de hombres y una de mujeres, trabajarlas por separado, y ver cómo cambian las relaciones en ambas y comparar los resultados de las inferencias condicionales que realicemos, este método es el que realizaremos.

```
data_M = read_csv("../data/data_M.csv")
data_F = read_csv("../data/data_F.csv")
colnames(data_M) <- c("ALB", "HDL", "LDL", "CR", "GLU",
  ↪  "INS", "PCR", "TRI", "HB1AC", "EAG", "SO2", "CO", "NOx",
  ↪  "COV", "PM10", "PM2.5", "NH3")
colnames(data_F) <- c("ALB", "HDL", "LDL", "CR", "GLU",
  ↪  "INS", "PCR", "TRI", "HB1AC", "EAG", "SO2", "CO", "NOx",
  ↪  "COV", "PM10", "PM2.5", "NH3")
```

```
gbn_masc = bn.fit(dag1, data = data_M)
gbn_fem = bn.fit(dag1, data = data_F)
```

Queries: 1. ¿Cuál es la probabilidad de que una persona tenga los triglicéridos saludables dado que están expuestas a altos niveles de NOx?

```
cpquery(gbn1, event = (TRI < 200), evidence = (NOx >
  ↪  quantile(data$NOx, 0.8)), n=10^6)
```

```
[1] 0.6536755
```

¿Cuál sería la probabilidad sin condición?

```
cpquery(gbn1, event = (TRI < 200), evidence = TRUE, n=10^6)
```

```
[1] 0.64335
```

2. ¿Cuál es la probabilidad de que los niveles de glucosa sean saludables en mujeres expuestas a altos niveles de COV y PM2.5?

```
cpquery(gbn_fem, event = (GLU < 100), evidence = (COV >
  ↪  quantile(data$COV, 0.8) & PM2.5 > 500), n=10^6)
```

```
[1] 0.4508285
```

¿Cuál sería la probabilidad sin condición?

```
cpquery(gbn_fem, event = (GLU < 100), evidence = TRUE, n=10^6)
```

```
[1] 0.461291
```

3. Cual es la probabilidad de que un hombre tenga un valor saludable de hemoglobina glucosilada dado que estan expuestos a altos niveles de CO

```
cpquery(gbn1, event = (HB1AC < 6), evidence = (CO >
  ↪  quantile(data$CO, 0.8)), n=10^6)
```

```
[1] 0.01352375
```

¿Cuál sería la probabilidad sin condición?

```
cpquery(gbn1, event = (HB1AC < 6), evidence = TRUE, n=10^6)
```

```
[1] 0.013185
```

En general podemos observar que las probabilidades cambian muy poco, por lo que podemos concluir que los contaminantes del aire no tienen impacto en los biomarcadores de las personas.

Incluir modelos no paramétricos puede ayudar a mejorar el BIC y el AIC si las relaciones entre los datos son no lineales o no gaussianas, ya que los modelos no paramétricos permiten más flexibilidad en las distribuciones condicionales, lo que permite capturar relaciones no lineales.

Sin embargo, este acercamiento necesita una gran cantidad de datos para evitar un sobreajuste y suele ser más difícil de interpretar.

0.5 Conclusiones

Se construyeron tres modelos (DAGs) y, mediante los criterios AIC y BIC, se seleccionó la estructura más adecuada. Sin embargo, los resultados de las consultas condicionales mostraron que los cambios en la probabilidad de presentar biomarcadores saludables ante altos niveles de contaminantes fueron mínimos. Esto sugiere que, bajo los supuestos del modelo, los contaminantes no presentan un impacto directo fuerte sobre los biomarcadores analizados.

Una posible explicación es la limitación de las redes gaussianas, que asumen relaciones lineales y distribuciones normales, lo cual puede no reflejar la complejidad real de las interacciones entre contaminación y salud. Para futuros trabajos, se recomienda el uso de modelos no paramétricos o híbridos, así como bases de datos más amplias que incluyan variables contextuales como hábitos de vida y condiciones socioeconómicas. Aunque no se encontraron efectos significativos, la metodología implementada constituye una base sólida para estudios posteriores y puede convertirse en una herramienta útil para el diseño de políticas públicas orientadas a la prevención de enfermedades relacionadas con la contaminación.

0.6 Referencias

https://www.jstatsoft.org/article/view/v035i03