Situación Problema: Transporte 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— Este trabajo tiene como objetivo analizar y responder un conjunto de consultas asignadas mediante la construcción de redes bayesianas multinomiales. Para ello, se discretizaron variables continuas, se propusieron distintos grafos acíclicos dirigidos (DAGs) y se evaluaron sus relaciones de dependencia. Posteriormente, se aplicó el algoritmo de hill-climbing para identificar la estructura que mejor se ajusta a los datos. Los resultados muestran que la red obtenida permite responder de manera consistente las queries planteadas, confirmando la utilidad de este enfoque para modelar dependencias probabilísticas.

Keywords— Redes Bayesianas, Grafos Acíclicos Dirigidos, Inferencia Probabilística, Discretización de Variables, Hill-Climbing, Modelado Multinomial, Dependencias entre Variables

I. Introducción

En 2017, el Instituto Nacional de Estadística y Geografía (INEGI) realizó la Encuesta Origen Destino en Hogares de la Zona Metropolitana del Valle de México (EOD 2017). Esta encuesta recopiló información detallada sobre los viajes diarios de los habitantes de la zona metropolitana, incluyendo origen y destino, motivo, duración, medio de transporte y horarios de desplazamiento. Dichos datos constituyen una fuente estadística clave para analizar la movilidad urbana y sirven como base para el presente trabajo.

Las redes bayesianas se han consolidado como una herramienta fundamental para el modelado de incertidumbre y la representación de relaciones de dependencia entre variables aleatorias. Estas estructuras han sido muy útiles en áreas como la bioinformática, el diagnóstico médico,clasificaciónde documentos, la inteligencia artificial y la toma de decisiones bajo incertidumbre (DataFlair Team, s. f.). [1]

El presente reto se centra en la construcción de una red bayesiana multinomial con el propósito de responder un conjunto de consultas previamente asignadas. La motivación de este trabajo se basa en la necesidad de contar con modelos que permitan inferir de manera eficiente la probabilidad de ciertos aspectos con respecto al transporte en México dado ciertas condiciones específicas en función de múltiples variables.

El problema abordado consiste en determinar qué estructura de red representa mejor las relaciones presentes en los datos y cómo esa estructura influye en la calidad de las respuestas a las consultas. Para ello, se comparan diferentes propuestas de grafos acíclicos dirigidos y se aplica un algoritmo de búsqueda que optimiza el ajuste de la red. Este trabajo resulta relevante al demostrar la importancia de las redes bayesianas como herramienta para la inferencia probabilística y el análisis de datos.

II. Metodología

El enfoque metodológico en este caso, se basa en la construcción y análisis de redes bayesianas multinomiales para responder un conjunto de queries específicas asignadas. A continuación, se detallan los procedimientos, técnicas y herramientas empleadas:

a. Preparación del entorno de trabajo

El desarrollo se realizó en un entorno de Python administrado con Anaconda, en combinación con R mediante la librería reticulate. Esto permitió integrar el preprocesamiento de datos en Python con la construcción y validación de redes bayesianas en R utilizando el paquete bnlearn.

El preprocesamiento consistió en integrar tablas de viajes y transportes a través del identificador de viaje, calcular la duración como diferencia entre la hora de inicio y fin del trayecto, y discretizar la variable continua de tiempo en dos categorías.

b. Definición y construcción de variables

A partir de los conjuntos de datos de la Encuesta Origen Destino 2017 (EOD 2017), se seleccionaron y transformaron variables relevantes para representar los atributos de interés. Se definieron las siguientes variables discretas:

- Estrato (E): nivel de estrato social (4 niveles).
- Semana (S): Si el viaje se realizó entre semana o fin de semana).
- Origen (O): lugar de inicio del viaje.
- Propósito (P): motivo principal del viaje.
- Transporte (T): medio de transporte utilizado.
- Duración (D): tiempo total de viaje (0 = viaje 60 minutos; 1 = viaje > 60 minutos).

c. Propuesta de estructuras iniciales (DAGs)

Con las variables definidas se propusieron tres posibles grafos acíclicos dirigidos (DAGs) que representarán diferentes "hipótesis" sobre las relaciones de dependencia entre las variables. Estas estructuras se definieron manualmente con base en criterios de lógica y conocimiento previo sobre movilidad urbana.

d. Estimación de parámetros

Para cada una de las estructuras propuestas se estimaron los parámetros de las redes bayesianas mediante el método de Máxima Verosimilitud (MLE), obteniendo así las distribuciones de probabilidad condicional correspondientes a cada nodo, bajo la fórmula que describe ese método:

$$L(\theta \mid x) = \prod_{i=1}^{n} f(x_i, \theta)$$

e. Evaluación de las dependencias y ajuste del modelo

La significancia de las relaciones entre variables se evaluó mediante la medida de información mutua, siendo equivalente a la prueba de G^2, lo que permitió comparar la relevancia de los arcos en cada DAG, dada por:

$$G^2 = 2 \sum_{t \in T} \sum_{e \in E} \sum_{k \in O \times R} n_{tek} \log \left(\frac{n_{tek}\, n_{..k}}{n_{t \cdot k}\, n_{\cdot ek}} \right)$$

Posteriormente, se calculó el Bayesian Information Criterion (BIC) para cada estructura, seleccionando aquella con mayor puntaje como el modelo de mejor ajuste.

f. Optimización de la estructura mediante Hill-Climbing

Con el objetivo de mejorar la estructura, se aplicó el algoritmo de Hill-Climbing implementado en bnlearn, empleando el criterio BIC como función de puntuación. Este procedimiento permitió obtener de manera automática una DAG optimizada que refleja las relaciones más consistentes con los datos observados.

III. Aplicación

Se llevó a cabo la implementación práctica de la metodología con base en la EOD 2017: construcción de variables, propuestas de DAG, ajuste de redes bayesianas, validación por BIC y *hill-climbing*, así como la resolución y discusión de cuatro *queries* específicas.

a. Descripción de los datos y construcción de variables

Se parte de un set de viajes de la EOD 2017 (CDMX y zona metropolitana). Se integran tablas de viaje y transporte (vía `id_via`). Las variables utilizadas y su codificación (todas categóricas) son:

- **Estrato (E)**: 1=bajo, 2=medio bajo, 3=medio alto, 4=alto.
- **Semana (S)**: 1=entre semana, 2=fin de semana.
- **Origen (O)**: 1=hogar, 2=trabajo, 3=escuela, 4=otro.
- **Propósito (P)**: codificado según la EOD; en las consultas se usa, por ejemplo, P=2 para “ir al trabajo”.
- **Transporte (T)**: múltiple categoría (auto particular, modos de transporte público como colectivo/micro, metro, autobús, metrobús, trolebús, etc.; y otros como taxi, bicicleta, bicitaxi). Para “transporte público” se considera el conjunto $\{2, 5, 6, 8, 10, 11, 12, 13, 15\}$.
- **Duración (D)**: binaria por discretización del tiempo de viaje; D=1 si la duración es > 60 minutos y D=0 en otro caso.

b. DAGs propuestas

Se proponen tres estructuras causales como hipótesis iniciales:

DAG 1. El *modo* de transporte (T) y la *duración* (D) influyen en las condiciones del viaje junto con O, P y E. Hipótesis central: D depende directamente de T.

DAG 2. T depende de E (elección modal condicionada por estrato) y D depende de T; S depende de P; E depende de O y P.[1]

DAG 3. T depende de E y D depende de T; E depende de O y P; S depende de P. Esta estructura se utiliza explícitamente en los ejercicios de ajuste.

c. Modelación (redes bayesianas) y estimación

Para cada DAG se ajusta una red bayesiana con parámetros estimados por Máxima Verosimilitud (MLE), empleando las frecuencias observadas en el conjunto de datos discretizados.

d. Relaciones de dependencia

Para cada arco se realizaron pruebas de independencia condicional y se estimaron medidas de dependencia (información mutua) y fuerza de arco mediante bootstrap.

Resultados generales En las tres DAGs todas las relaciones evaluadas muestran p-valores extremadamente pequeños (prácticamente 0), lo cual indica rechazo consistente de la hipótesis nula de independencia condicional para cada par de nodos conectados por un arco. Esto se debe a:

1. *Especificación atenta de las DAGs:* las estructuras propuestas se diseñaron con base en conocimiento de dominio y en la codificación de las variables, por lo que las dependencias esperadas aparecen claramente en los datos.

Esto mismo fue aplicado para la `bestdag` obteniendo mismos resultados.

[1] Ver sección “Relaciones de dependencia” y modelos impresos en el `modelstring`.

e. Validación por BIC y selección del mejor ajuste

Se compara el ajuste de las tres DAG mediante el Criterio de Información Bayesiano (BIC). Los resultados son:

Modelo	BIC
DAG 1	−6,258,404
DAG 2	−6,238,546
DAG 3	−6,267,011

El mayor (menos negativo) BIC corresponde a la **DAG 2**, por lo que es la que mejor se ajusta entre las propuestas.

*f. Búsqueda de estructura (***hill-climbing***)*

Además de las DAG propuestas, se aprende una estructura automática con *hill-climbing* (score BIC). El *model string* obtenido es:

```
[E][T|E][D|E:T][O|T:D][P|O:D][S|O:P:D]
```

Esto implica, entre otros, que T depende de E; D depende de E y T; O depende de T y D; P depende de O y D; y S depende de O, P y D. Se verifican las *arc strengths* por información mutua como chequeo de pertinencia.

g. Queries, resultados y discusión

A continuación se responden las cuatro consultas planteadas, usando `cpquery` sobre la red ajustada.

Query 1. Probabilidad de que un viaje del **hogar al trabajo** dure $>$ **60 min** si la persona usa **transporte público**.
Evento: $O{=}1 \wedge P{=}2 \wedge D{=}1$. Evidencia: $T \in \{2,5,6,8,10,11,12,13,15\}$.
Resultado: $\hat{p} = 0.1665$.
Interpretación: Aproximadamente 1 de cada 6 viajes hogar–trabajo en transporte público supera la hora. Esto concuerda con la mayor congestión y transferencias de los modos colectivos frente al auto particular en zonas densas.

Query 2. Probabilidad de que el **estrato sea bajo** dado que el **modo es auto**.
Evento: $E{=}1$. Evidencia: $T{=}1$ (auto).
Resultado: $\hat{p} = 0.3801$.
Interpretación: Existe una probabilidad sustancial (38%) de observar estrato bajo entre quienes viajan en auto. Esto sugiere diversidad socioeconómica en la tenencia/uso del auto y posibles viajes en auto de corta distancia y costo compartido.

Query 3. Probabilidad de que un **estrato bajo** utilice **transporte público** (micro/metro/autobús).
Evento: $T \in \{2,5,8,6\}$ (representativo de modos públicos). Evidencia: $E{=}1$.
Resultado: $\hat{p} = 0.3798$.
Interpretación: La probabilidad de elección de transporte público entre estrato bajo ronda 38%, consistente con restricciones de costo y disponibilidad modal esperables.

Query 4. Probabilidad de viajar **entre semana** y con **estrato medio bajo**, dado que el **modo es bicitaxi**.
Evento: $S{=}1 \wedge E{=}2$. Evidencia: $T{=}16$ (bicitaxi).
Resultado: $\hat{p} = 0.3911$.
Interpretación: Casi 4 de cada 10 viajes en bicitaxi ocurren entre semana y en estrato medio bajo, lo que puede asociarse a trayectos de primera/última milla en zonas específicas.

IV. Conclusiones

- La codificación y discretización de variables demostraron ser un recurso eficaz para ajustar redes bayesianas multinomiales. Esto permitió representar y evaluar hipótesis causales plausibles sobre la movilidad urbana en la Zona Metropolitana del Valle de México, generando modelos consistentes y estadísticamente válidos.
- Entre las estructuras propuestas, el **DAG 2** mostró el mejor balance entre ajuste y simplicidad, según el criterio *BIC*. La estructura aprendida automáticamente mediante *hill-climbing* reforzó esta elección al identificar dependencias adicionales entre duración, propósito y origen de viaje, lo que confirma la robustez de la metodología.
- Los resultados muestran que el contexto del viaje (origen, propósito y duración) influye de manera decisiva en la elección modal y en el calendario de desplazamientos. Esta evidencia es valiosa para diseñar políticas públicas de movilidad, ya que permite identificar patrones específicos que podrían guiar mejoras en la infraestructura de transporte público, estrategias de reducción de tiempos de traslado y la promoción de modos alternativos y sostenibles.
- En conjunto, este estudio confirma que las redes bayesianas son una herramienta poderosa para el análisis probabilístico y causal de la movilidad urbana. Su aplicación a los datos de la EOD 2017 no solo valida hipótesis sobre los factores que afectan los desplazamientos, sino que también ofrece un marco para orientar decisiones en la planificación y gestión del transporte en México.

References

[1] D. Team, "Top 10 real-world bayesian network applications – know the importance!" https://data-flair.training/blogs/bayesian-network-applications/, n.d., recuperado el 30 de agosto de 2025.

[2] D. J. E. G. Guillén, "Redes bayesianas multinomiales," file:///Users/ximenacanton/Downloads/Redes%20bayesianas%20multinomiales%20(5).html, 2025, archivo HTML. Recuperado el 30 de agosto de 2025.