Análisis de Patrones de Movilidad Urbana mediante Redes Bayesianas

José Emilio Martínez Hernández *A01403100*

Josué Tapia Hernández *A01621056*

Andrés Martínez Almazán *A01621042*

Diego Arechiga Bonilla *A01621045*

31/08/2025

Resumen

Este trabajo presenta un análisis de patrones de movilidad urbana utilizando redes bayesianas para modelar las relaciones causales entre características socioeconómicas, demográficas y decisiones de transporte. El objetivo principal es responder a consultas probabilísticas específicas sobre comportamientos de movilidad, incluyendo: la probabilidad de uso de transporte no motorizado según estrato socioeconómico, patrones de uso de autobús por mujeres en diferentes localidades, frecuencia de viajes en automóvil por estudiantes mujeres, y comparación entre transporte motorizado y no motorizado según duración del viaje. Utilizando datos de la Encuesta Origen-Destino 2017 de la Zona Metropolitana del Valle de México, se proponen modelos DAG (Directed Acyclic Graph) que capturan diferentes perspectivas de análisis: influencia socioeconómica, patrones temporales y un modelo integrado de decisión de viaje. La metodología emplea el paquete bnlearn de R para la construcción, evaluación y comparación de los modelos mediante criterios BIC y análisis de fortaleza de arcos. Los resultados muestran que el estrato socioeconómico es el factor determinante en las decisiones de movilidad, con diferencias

significativas en el uso de transporte no motorizado entre estratos (mayor uso en estrato bajo), baja utilización del transporte público en localidades medias (11 % en mujeres), y que los viajes cortos (¡40 min) se distribuyen casi equitativamente entre transporte motorizado (53 %) y no motorizado (47 %). Este enfoque proporciona una herramienta valiosa para el análisis y comprensión de la movilidad urbana, con aplicaciones potenciales en la planificación del transporte y políticas públicas.

Repositorio del proyecto: [https://github.com/Emilio-Mtz-bit/Multinomial_BN.git]

Palabras clave: Redes Bayesianas, Movilidad Urbana, DAG, Transporte, Análisis Causal

Índice

1. Introducción

La movilidad urbana constituye uno de los aspectos más complejos y desafiantes de las ciudades modernas, especialmente en megaurbes como la Zona Metropolitana del Valle de México (ZMVM). La comprensión de los patrones de desplazamiento y los factores que influyen en las decisiones de transporte de los ciudadanos es fundamental para el desarrollo de políticas públicas efectivas y la planificación urbana sustentable.

Durante este proyecto estamos abarcando varias hipótesis para verificar varios aspectos, como el tiempo de viaje, vehículos usados, y características socioeconómicas de las personas.

1.1. Preguntas a contestar

Query 1: ¿Cuál es la probabilidad de que una persona del estrato socioeconómico más bajo camine o use bicicleta como medio de transporte versus una persona del estrato más alto?

En este primer query nuestra hipótesis es que va a ser más probable que una persona del estrato socioeconómico más bajo camine o use bicicleta esto ya que como se ha visto en los últimos años la situación de pobreza en el valle de México ha ido empeorando por lo que muchas personas optan por moverse caminando o en bicicletas.

Query 2: ¿Cuál es la probabilidad de que una persona de sexo femenino de la entidad de Sonora que habita en una localidad de 15,000 a 99,999 habitantes utilice un autobús?

En esta query nuestra hipótesis es que la probabilidad va a ser baja por el tamaño de localidad ya que en este tipo de localidades el transporte público no es efectivo y escasea mucho.

Query 3: ¿Cuál es la probabilidad de que una estudiante mujer haga más de 2 viajes al día en un automóvil?

En esta query nuestra hipótesis es que la probabilidad va a ser baja por lo mismo del estrato social, ya que la mayoría de las personas en la encuesta son de estrato bajo lo que significa que muchas de ellas no tienen acceso a automóviles propios.

Query 4: ¿Cuál es la probabilidad de que un viaje con duración menor a 40 minutos

se realice en transporte motorizado versus un viaje con transporte no motorizado?

En esta última query nuestra hipótesis es que va a ser mayor la probabilidad de que se usen transportes no motorizados en viajes de menos de 40 minutos, esto porque hoy en día la gente suele usar más los medios de transporte no motorizados ya que el tráfico en el valle de México es muy pesado y a veces las personas suelen hacer más tiempo en coche o camión que yendo en bicicleta o caminando.

2. Metodología

2.1. Análisis Exploratorio y Comprensión de Datos

El proceso metodológico inició con un análisis exhaustivo del dataset principal de la Encuesta Origen-Destino 2017 de la Zona Metropolitana del Valle de México. Se consultaron los diccionarios de datos disponibles en `eod_2017_csv/` para comprender el significado de variables, identificadores y códigos utilizados.

El primer acercamiento se realizó mediante `main.qmd`, donde se procesó el dataset original `tviaje.csv` que contenía información detallada sobre viajes realizados. Se identificaron las variables más relevantes para el análisis de patrones de movilidad y se eliminaron variables redundantes o con alta proporción de datos faltantes.

2.2. Construcción de Datasets Especializados

2.2.1. Dataset de viaje Inicial (data_limpio.csv)

Inicialmente se editó el dataset principal `tviaje.csv` eliminando algunas columnas y agregando unas nuevas, algunas de las columnas eliminadas fueron: `p5_11a`, `p5_5_02`, `p5_12_7`, `p5_13`, etc. Todas estas columnas las decidimos eliminar por completo ya que eran irrelevantes para las querys que teníamos que resolver.

Por otro lado también renombramos algunas de las columnas para poder entender mejor a que se refiere cada una, `p5_7_7` la cambiamos por `entidad`, `p5_9_1` la cambiamos por `hora_ini`, etc. Decidimos renombrar todas las variables para que se nos facilitara más

utilizar los datos para responder las querys.

Después tuvimos que crear 3 columnas nuevas con la información específica que se necesitaba, creamos las columnas: `motorizado`, `no_motorizado` y `estudiante`. Todo esto para que finalmente nos quedará un dataset con únicamente la información necesaria para responder las querys.

El dataset final quedó con las siguientes columnas: `id_via`, `id_soc`, `entidad`, `hora_ini`, `min_ini`, `hora_fin`, `min_fin`, `used_auto`, `veces_auto`, `used_autobusM1`, `used_bicicleta`, `used_autobus`, `used_caminar`, `estrato`, `tam_localidad`, `sexo`, `edad`, `motorizado`, `no_motorizado` y `estudiante`.

Donde `id_via` nos muestra el id del viaje en específico, `id_soc` cuenta con el id específico de cada persona, `entidad` el estado de la persona, `hora_ini` la hora de inicio del viaje, `hora_fin` la hora de fin del viaje, `min_ini` el minuto de inicio del viaje, `min_fin` el minuto de fin del viaje, `veces_auto` las veces que utilizó un auto, `used_` muestra si la persona utilizó ese tipo de transporte en específico, `estrato` muestra el estrato socioeconómico al que pertenece, `tam_localidad` tamaño de la localidad a la que pertenece, `sexo` el sexo de la persona, `edad` edad de la persona, `motorizado` si la persona utilizó algún tipo de transporte motorizado, `no_motorizado` si la persona utilizó algún tipo de transporte no motorizado y finalmente `estudiante` si la persona es estudiante o no. Con esto pudimos contestar la query 4.

2.2.2. Dataset de Personas (data_persona.csv, tsdem.csv)

Posteriormente nos dimos cuenta que íbamos a necesitar otro dataset diferente para poder responder algunas de las querys, esto ya que en el dataset `data_limpio.csv` únicamente cuenta con información de los viajes en general. Pero necesitábamos uno que nos diera información sobre las personas en específico. Por lo que decidimos crear el dataset `data_persona.csv`.

Para poder conseguir esto igual que en el anterior tuvimos que eliminar completamente todas las columnas que eran irrelevantes para responder las querys, después con ayuda de ciclos for y funciones if combinamos todas las observaciones de una persona en específico

en una sola fila, mostrándonos de esta manera toda la información de la persona en una sola fila no en varias, facilitando bastante el manejo de la información para resolver estas querys.

Por ejemplo, tuvimos que sumar todas las veces que la persona utilizó un carro, contar la cantidad de días que se registraron sobre las personas, etc. Finalmente obtuvimos un dataset con las siguientes columnas: `id_soc`, `dias`, `sexo`, `edad`, `estrato`, `entidad`, `used_auto`, `used_autobusM1`, `used_bicicleta`, `used_autobus`, `used_caminar`, `veces_auto`.

Donde `id_soc` muestra el id de la persona en específico, `dias` muestra la cantidad de días registrados por cada persona, `sexo` el sexo de la persona, `edad` edad de la persona, `estrato` estrato socioeconómico de la persona, `entidad` estado en el que vive la persona, `used_` muestran si en cualquier observación se utilizó ese método de transporte en específico y `veces_auto` muestra la suma de las veces que la persona utilizó un automóvil. Con esto pudimos responder las querys 1, 2 y 3.

2.2.3. Dataset de Viajes Final (DF_DAG(Viajes).csv y PreProcessing Viajes.csv)

A través del primer dataset que fue el `data_limpio.csv` se hicieron algunos cambios para que este quedara mejor con las DAG que teníamos en mente. Por ejemplo se quitaron las variables de sexo y edad ya que en ninguna de nuestras querys las ocupábamos y las consideramos variables extra que no aportaban al contexto de nuestras querys.

Igualmente se quitaron las variables como hora de inicio, hora final, estudiantes y los números de identificación de los viajes ya que todas estas variables las discretizamos poniéndolas en rangos y así categorizándolas para que el análisis de las DAGs fuera más fácil.

2.3. Propuesta de Modelos DAG

2.3.1. DAGs para Dataset de Personas (Personas Dag.qmd)

Se propusieron tres enfoques diferentes para modelar los patrones de movilidad:

DAG 1: Enfoque Sociodemográfico: Las referencias hablan de cómo los factores sociales determinan la ocupación. Y a su vez la ocupación y los factores sociales determinan el medio de transporte que cada persona utiliza. Por lo tanto todos los factores sociales son independientes.

DAG 2: Enfoque de Ocupación independiente y suposición del auto Está basado en lo mismo que lo anterior pero aquí asumimos que ocupación también es independiente y en el hecho de que si usa auto es un transporte privado y por lo tanto es probable que no utiliza autobús ni bicicleta ni camina.

DAG 3: Hill Climbing: Se utilizó el algoritmo de hill climbing para construir la DAG.

2.3.2. DAGs para Dataset de Viajes (Optimal_DAG(Viajes).qmd)

Para el dataset específico de viajes se desarrollaron dos modelos:

DAG 1: Determinantes socioeconómicos y duración del viaje Este modelo se centra en cómo las características del contexto socioeconómico y territorial (estrato, entidad y tamaño de la localidad) influyen en el acceso a transporte motorizado o no motorizado, y cómo esta decisión determina los modos de viaje utilizados. Posteriormente, los diferentes modos de transporte, junto con el horario del viaje, impactan directamente en la duración total del traslado, la cual finalmente clasifica el viaje como corto o largo.

DAG 2: Impacto directo del estrato y el tamaño de localidad en los modos de transporte En esta propuesta se hace explícita la relación directa entre el nivel socioeconómico y el tamaño de la localidad con el uso de modos de transporte específicos (auto, autobús, bici, caminar). Es decir, no solo median a través de las categorías motorizado/no motorizado, sino que influyen de manera directa en los modos concretos. El modelo resalta cómo el contexto social y urbano determina el uso real de cada vehículo y cómo, a partir de ahí, junto con el horario, se explica la duración del traslado y su clasificación final como corto o largo.

2.4. Redes Bayesianas y Modelos DAG

2.4.1. Fundamentos Teóricos

Una red Bayesiana es un modelo probabilístico que se utiliza para representar y razonar con la incertidumbre, o en otras palabras es un gráfico dirigido acíclico o DAG. Una DAG está compuesta por 3 cosas principales: los nodos que representan variables aleatorias ya sean continuas o discretas, arcos que representan las relaciones de dependencia condicional entre los nodos y finalmente las tablas de probabilidad condicional, cada nodo cuenta con una de estas tablas que describe cómo el mismo nodo depende de sus nodos "padres".

2.4.2. Construcción de DAGs

Computacionalmente podemos construir DAGs mediante el uso de diferentes algoritmos como hill climbing, para posteriormente utilizar métricas como BIC o AIC, para comparar dos o más propuestas de DAG y poder concluir cuál de las DAGs es mejor. Aunque siempre se recomienda que estas sean creadas por una persona especializada en el área que se está trabajando.

2.5. Evaluación y Comparación de Modelos

La evaluación de los modelos DAG se realizó utilizando criterios estadísticos estándar, principalmente el Criterio de Información Bayesiano (BIC) y el análisis de fortaleza de arcos mediante información mutua. Estas métricas permiten comparar objetivamente diferentes estructuras de red y determinar cuál captura mejor las dependencias presentes en los datos. El BIC penaliza la complejidad del modelo, favoreciendo estructuras que logran un buen ajuste con menor número de parámetros, mientras que el análisis de fortaleza de arcos cuantifica qué tan fuerte es la dependencia entre variables conectadas.

2.6. Herramientas y Software

Se utilizó R con:

- `bnlearn`

- `Rgraphviz`
- Paquetes base de R

3. Implementación y Evaluación

3.1. Construcción de los Modelos

El proceso de implementación se llevó a cabo en R utilizando el paquete `bnlearn` para la construcción y evaluación de las redes bayesianas. Se implementaron tanto modelos teóricos basados en conocimiento del dominio como modelos generados algorítmicamente mediante hill climbing.

3.2. Métricas de Evaluación

- BIC
- Fortaleza de arcos

3.3. Resultados

3.3.1. Comparación de Modelos

Modelos para Dataset de Viajes:

Los tres modelos DAG propuestos para el dataset de viajes fueron evaluados utilizando el criterio BIC (Bayesian Information Criterion). Los resultados muestran que el modelo generado por Hill Climbing (DAG 3) obtiene consistentemente el mejor score BIC, seguido por el DAG 2 (Impacto directo del estrato) y finalmente el DAG 1 (Determinantes socioeconómicos).

La superioridad del modelo de Hill Climbing sugiere que la estructura óptima de dependencias causales es más compleja que las propuestas teóricas iniciales, capturando relaciones no evidentes a priori entre las variables de movilidad urbana.

Modelos para Dataset de Personas:

Para el dataset de personas se evaluaron tres enfoques diferentes:

- **DAG 1 (Enfoque Sociodemográfico):** Obtuvo un score BIC superior, validando la hipótesis de que los factores sociales determinan la ocupación, y tanto la ocupación como los factores sociales influyen en las decisiones de transporte.
- **DAG 2 (Ocupación independiente):** Presentó un score BIC inferior al DAG 1, sugiriendo que la asunción de independencia de la ocupación no captura adecuadamente las relaciones causales.
- **DAG 3 (Hill Climbing):** Generó la mejor estructura según el criterio utilizado (AIC), proporcionando el modelo óptimo para las inferencias probabilísticas.

La comparación confirma que el DAG 1 es superior al DAG 2, validando que la ocupación sí depende de los factores sociales, tal como sugiere la literatura especializada.

3.3.2. Análisis de Dependencias

El análisis de fortaleza de arcos mediante información mutua revela que las conexiones más robustas en los modelos se establecen entre:

- Variables socioeconómicas (estrato, entidad) y decisiones de transporte
- Características temporales (hora_cat) y duración del viaje
- Modos de transporte específicos y la categorización motorizado/no motorizado

3.3.3. Interpretación de Patrones

Query 4 - Duración de viajes y modalidad de transporte:

Los resultados del análisis probabilístico muestran que para viajes con duración menor a 40 minutos:

- Probabilidad de uso de transporte motorizado: 53 %
- Probabilidad de uso de transporte no motorizado: 47 %

Esta distribución casi equitativa contradice parcialmente la hipótesis inicial y revela un patrón interesante: en la ZMVM, los viajes cortos no están fuertemente determinados por el tipo de transporte. Esto sugiere que tanto el transporte motorizado como el

no motorizado son opciones viables para distancias cortas, reflejando la realidad urbana donde factores como el tráfico vehicular pueden hacer que caminar o usar bicicleta sea tan eficiente como el transporte motorizado para trayectos breves.

Este hallazgo es consistente con las tendencias de movilidad urbana observadas en megaciudades, donde la congestión vehicular ha promovido el uso de modos de transporte alternativos para distancias cortas.

Queries 1, 2 y 3 - Análisis del dataset de personas:

Query 1 - Transporte no motorizado por estrato socioeconómico: Los resultados confirman marcadamente la hipótesis inicial: existe una diferenciación socioeconómica significativa en el uso de transporte no motorizado. Las personas del estrato socioeconómico más bajo muestran una probabilidad considerablemente mayor de caminar o usar bicicleta comparado con personas del estrato más alto. Este patrón refleja tanto limitaciones económicas para acceder a transporte motorizado como adaptaciones necesarias de la población de menores recursos.

Query 2 - Uso de autobús por mujeres en localidades medias: La probabilidad de uso de autobús por mujeres en localidades de 15,000 a 99,999 habitantes en el Estado de México es aproximadamente 11 %, confirmando la hipótesis de baja utilización. Este resultado evidencia las deficiencias del transporte público en localidades de tamaño medio, donde la infraestructura de transporte masivo es limitada y la cobertura insuficiente para satisfacer las necesidades de movilidad.

Query 3 - Uso intensivo de automóvil por estudiantes mujeres: Los resultados revelan una probabilidad muy baja para estudiantes mujeres en general de realizar más de 2 viajes diarios en automóvil. Sin embargo, cuando se controla por estrato socioeconómico alto, la probabilidad aumenta significativamente. Este contraste subraya el papel determinante del estrato social en el acceso a transporte privado, donde incluso para poblaciones con alta movilidad potencial (estudiantes), el factor socioeconómico es limitante.

4. Conclusiones

4.1. Hallazgos Principales

Nuestros principales descubrimientos fueron que en realidad bastantes cosas dependen de los factores sociales, más que nada del estrato ya que en todas las DAGs propuestas el estrato siempre era independiente y determinaba a muchísimas variables de la DAG. Esto nos hace preguntarnos más que nada algunas querys que tengan que ver con el estrato y así observar todas las diferencias sociales que conlleva el ser de estrato bajo en relación al transporte y qué puede hacer el gobierno para solucionar estas problemáticas sociales.

4.2. Implicaciones

Las implicaciones que tiene este modelo pueden llegar a ser desde calcular por qué las personas de estrato social más bajo tienen dificultades para llegar a cualquier lugar y por qué y cómo enmendar la situación en cuanto al transporte público.

4.3. Limitaciones

Las limitaciones de este modelo es que solo se limita a variables discretas cuando en realidad hay un gran campo de estudio que se puede hacer con variables continuas para así tener un modelo más preciso en cuanto a variables numéricas ya sea duración del viaje, edad, ingresos fijos, etc.

4.4. Trabajo Futuro

Dentro del trabajo futuro se podrían incluir las variables numéricas así como un estudio más profundo acerca de las dependencias de las variables. Además de que se pueden formular querys más complejas para resolver problemáticas sociales que tal vez no son visibles sin estos modelos.

Las direcciones futuras de investigación incluyen la incorporación de variables continuas como ingresos exactos, tiempo de viaje preciso y distancias reales, que permitirían modelos más granulares.

Referencias

[1] Smith, D. L. (2023). Social and environmental determinants of occupation: An intersectional concept focused on occupational justice and participation. *Journal of Occupational Science*. https://doi.org/10.1080/14427591.2023.2212676

[2] Laliberte-Rudman, D., Beagan, B. L., Phelan, S., & Kiepek, N. (2019). Silences around occupations framed as unhealthy, illegal, and deviant. *Journal of Occupational Science, 18*(3), 254–276. https://doi.org/10.1080/14427591.2018.1499123

[3] Instituto Nacional de Estadística y Geografía (INEGI). (2017). *Encuesta Origen-Destino en Hogares de la Zona Metropolitana del Valle de México 2017*. INEGI. https://www.inegi.org.mx/programas/eod/2017/

[4] Berrones-Sanz, L. D. (2022). Walking in Mexico City: Sociodemographic characteristics of the pedestrian. *Transportation Research Part D: Transport and Environment, 108*, 103312. https://doi.org/10.1016/j.trd.2022.103312