



Universidad Autónoma del Estado de Hidalgo

Instituto de Ciencias Básicas e Ingeniería
Área Académica de Matemáticas y Física

Impacto de la formalidad laboral en los
ingresos de egresados universitarios en
México: una evaluación mediante
emparejamiento por puntajes de
propensión

T E S I S

que para obtener el título de:

Licenciado en Matemáticas Aplicadas

presenta

Arturo Moctezuma Ramos

Bajo la dirección de la

Dra. Alma Sofía Santillán Hernández

Mineral de la Reforma, Hidalgo.

Junio 2026



Universidad Autónoma del Estado de Hidalgo

Instituto de Ciencias Básicas e Ingeniería

School of Engineering and Basic Sciences

Dirección

Office of the Director

Mineral de la Reforma, Hgo., a 24 de junio de 2026

Número de control: ICBI-D/1202/2026

Asunto: Autorización de impresión.

MTRA. OJUKY DEL ROCÍO ISLAS MALDONADO DIRECTORA DE ADMINISTRACIÓN ESCOLAR DE LA UAEH

Con Título Quinto, Capítulo II, Capítulo V, Artículo 51 Fracción IX del Estatuto General de nuestra Institución, por este medio, le comunico que el Jurado asignado al egresado de la Licenciatura en Matemáticas Aplicadas **Arturo Moctezuma Ramos**, quien presenta el trabajo de titulación "**Impacto de la formalidad laboral en los ingresos de egresados universitarios en México: una evaluación mediante emparejamiento por puntajes de propensión**", ha decidido, después de revisar fundamento en lo dispuesto en el Título Tercero, Capítulo I, Artículo 18 Fracción IV; dicho trabajo en la reunión de sinodales, **autorizar la impresión del mismo**, una vez realizadas las correcciones acordadas.

A continuación, firman de conformidad los integrantes del Jurado:

Presidente: Dr. Roberto Ávila Pozos

Secretario: Mtra. Mariana Sánchez Licon

Vocal: Dra. Alma Sofía Santillán Hernández

Suplente: Dr. Ronald Richard Jiménez Munguía

Sin otro particular por el momento, reciba un cordial saludo.

Atentamente
"Amor, Orden y Progreso"

Dr. Gabriel Vergara Rodríguez
Director del ICBI



GVR/MMM

"Amor, Orden y Progreso"

Ciudad del Conocimiento, Carretera Pachuca-Tulancingo
Km. 4.5 Colonia Carboneras, Mineral de la Reforma, Hidalgo,
México. C.P 42184
Teléfono: 52 (771) 71 720 00 Ext. 40001
direccion_icbi@uaeh.edu.mx, vergara@uaeh.edu.mx



uaeh.edu.mx

A mis padres, por su amor y apoyo incondicional.

Agradecimientos

Expreso mi agradecimiento a la Dra. Alma Sofía Santillán Hernández, directora de esta tesis, por su orientación, paciencia y acompañamiento durante el desarrollo de este trabajo. Asimismo, agradezco a los profesores que contribuyeron a mi formación académica, por compartir sus conocimientos y enseñanzas a lo largo de mi trayectoria universitaria.

Resumen

Esta investigación analiza el impacto de la formalización laboral sobre el ingreso mensual de los egresados universitarios en México durante el periodo 2005–2025. El objetivo principal es determinar si transitar de la informalidad a la formalidad se asocia con un incremento en el ingreso, en comparación con egresados que permanecen en la informalidad.

Para ello, se utiliza información de la Encuesta Nacional de Ocupación y Empleo (ENOE), aprovechando su diseño de panel rotatorio, que permite observar a los individuos durante cinco trimestres consecutivos. El tratamiento se define como el tránsito de un empleo informal en el primer trimestre a un empleo formal en el quinto trimestre. El grupo de control está conformado por egresados que permanecen en la informalidad durante el mismo periodo.

Dado que la transición hacia la formalidad no ocurre de manera aleatoria, se emplea la metodología de Emparejamiento por Puntajes de Propensión (PSM). Esta técnica permite comparar egresados con características observables similares, reduciendo el sesgo de selección. El parámetro de interés es el efecto promedio del tratamiento sobre los tratados (ATT), que mide el impacto de la formalización en quienes efectivamente realizaron dicha transición.

Los resultados muestran que la formalización laboral se asocia con incrementos positivos y estadísticamente significativos en el ingreso mensual de los egresados universitarios. En particular, el método principal de emparejamiento Kernel muestra efectos positivos en todos los años analizados. Estos hallazgos sugieren que la formalidad representa un segmento laboral con mayores retornos económicos para los egresados universitarios.

En conjunto, la evidencia es consistente con una interpretación de segmentación laboral, en la cual el acceso al empleo formal no depende únicamente del nivel educativo, sino también de condiciones del mercado laboral. No obstante, los resultados deben interpretarse bajo los supuestos del PSM, especialmente la selección sobre observables, el soporte común y la calidad del balance alcanzado.

Índice general

Introducción	1
1. Antecedentes	3
1.1. Revisión de literatura	3
1.2. Movilidad Laboral en México: transiciones entre estados laborales	6
1.2.1. Procesos de Markov, persistencia y matriz de transición	7
1.2.2. Del análisis de flujos al análisis de impacto	8
1.3. Fundamentos de Inferencia Causal	9
1.3.1. El Marco de Resultados Potenciales (Rubin Causal Model)	10
1.3.2. SUTVA y el problema fundamental de la inferencia causal	11
1.4. Comparación causal y sesgo de selección	12
1.4.1. Contrafactuales y estudios observacionales	12
1.4.2. El Sesgo de Selección y la Descomposición del Impacto	13
1.5. Medidas Fundamentales de Impacto	14
1.5.1. Efecto promedio del tratamiento: ATE y ATT	14
1.5.2. Estimación empírica del ATT	15
1.6. Metodología de Emparejamiento por Puntajes de Propensión	16
1.6.1. Puntaje de propensión y reducción de dimensionalidad	16
1.6.2. Supuestos y fundamentos teóricos del Emparejamiento por Puntajes de Propensión	16
1.6.3. Especificación del Modelo Logit	20
1.6.4. Evaluación del balance: sesgo estandarizado y prueba t	21
2. Datos y estadística descriptiva	23
2.1. Fuente de Datos: Encuesta Nacional de Ocupación y Empleo (ENOE)	23
2.1.1. Descripción general de la ENOE	23
2.1.2. Estructura de bases de datos	24
2.2. Construcción del Panel Longitudinal	25
2.3. Construcción de Variables	27
2.4. Definición de la Muestra de Análisis	30
2.5. Estadística Descriptiva	34

2.5.1.	Evolución general del panel (2005–2025)	34
2.5.2.	Perfil sociodemográfico de la muestra de análisis	38
2.5.3.	Transición a la muestra de análisis	45
2.5.4.	Caracterización de la muestra de análisis	46
3.	Resultados	52
3.1.	Estimación del puntaje de propensión	52
3.1.1.	Especificación empírica del modelo Logit	52
3.1.2.	Efectos marginales promedio	53
3.1.3.	Distribución y evolución temporal del puntaje de propensión	55
3.2.	Calidad del emparejamiento	56
3.2.1.	Diagnóstico longitudinal de balance bajo el emparejamiento por Kernel	57
3.2.2.	Análisis de robustez: comparación con los métodos Vecino Más Cercano y Caliper	60
3.2.3.	Alcances del balance y criterios para interpretar el ATT	62
3.3.	Resultados de las estimaciones del ATT	64
3.3.1.	Efecto promedio del tratamiento sobre los tratados bajo Kernel	64
3.3.2.	Comparación de resultados con Vecino Más Cercano y Caliper	66
3.3.3.	Síntesis de resultados e implicaciones para el mercado laboral	68
3.3.4.	Alcances y limitaciones del estudio	69
4.	Conclusiones	70

Introducción

La informalidad laboral constituye uno de los problemas más persistentes del mercado de trabajo mexicano (OCDE, 2025; Perry et al. 2007). En los últimos años, la tasa de informalidad se mantuvo por encima del 54 % de la población ocupada sin registrar una reducción significativa (OCDE, 2025). Si bien este fenómeno suele asociarse con trabajadores de baja calificación, la evidencia empírica y teórica revisada en este trabajo muestra que su presencia también alcanza a personas con educación superior, lo que puede limitar el aprovechamiento del capital humano acumulado y afectar sus trayectorias de ingreso a lo largo del tiempo.

Surge entonces una pregunta relevante: ¿La formación universitaria garantiza el acceso a empleos formales y mejor remunerados? Esta investigación analiza el impacto de la formalidad laboral sobre los ingresos de egresados universitarios en México. En particular, se estudia si transitar de un empleo informal a uno formal se asocia con un incremento en el ingreso mensual, en comparación con egresados que permanecen en la informalidad. La pregunta central del trabajo es si la formalización representa una mejora salarial atribuible al cambio de estado laboral o si la brecha observada responde únicamente a diferencias previas entre quienes logran formalizarse y quienes no.

Para responder esta pregunta se utiliza información de la Encuesta Nacional de Ocupación y Empleo (ENOE) durante el periodo 2005–2025. A partir del diseño rotatorio de la encuesta se construye un panel de cinco trimestres, donde el primer trimestre, denotado por $t1$, se utiliza como línea base y el quinto trimestre, denotado por $t5$, como periodo de resultado. El tratamiento se define como la transición de la informalidad en $t1$ a la formalidad en $t5$, mientras que el grupo de control está compuesto por egresados que permanecen en la informalidad durante el mismo intervalo de tiempo. La variable de resultado es el ingreso mensual observado en el quinto trimestre.

Dado que la transición hacia la formalidad no ocurre de manera aleatoria, una comparación simple entre egresados formales e informales podría estar afectada por sesgo de selección. Por ello, la investigación emplea la metodología de Emparejamiento por Puntajes de Propensión (PSM, por sus siglas en inglés), la cual permite comparar individuos con características observables similares, pero que difieren en su transición a la formalidad. El efecto estimado corresponde al efecto promedio del tratamiento sobre los tratados (ATT, por sus siglas en inglés), es decir, el impacto de la formalización sobre los egresados que efectivamente transitaron hacia el empleo formal. Los resultados sugieren que dicha transición se asocia con incrementos estadísticamente

significativos en el ingreso mensual: dependiendo del algoritmo de emparejamiento empleado, el ATT estimado oscila entre 0.25 y 0.56, equivalentes a incrementos de entre el 28 % y 75 % en el ingreso mensual, lo que sugiere que la transición hacia la formalidad se asocia con una mejora salarial, bajo los supuestos del emparejamiento y la comparación observable alcanzada.

El primer capítulo del presente trabajo expone los antecedentes teóricos y metodológicos, incluyendo la revisión de la literatura sobre informalidad, segmentación laboral, movilidad entre estados laborales e inferencia causal. El segundo capítulo describe la fuente de datos, la construcción del panel longitudinal, la definición de variables y la estadística descriptiva de la muestra. El tercer capítulo presenta los diagnósticos de balance del emparejamiento y las estimaciones del ATT mediante tres algoritmos de emparejamiento. Finalmente, se presentan las conclusiones, alcances y limitaciones del estudio.

Capítulo 1

Antecedentes

Este capítulo presenta los antecedentes teóricos y metodológicos que sustentan la investigación. En primer lugar, se revisa la literatura sobre informalidad laboral en México, contrastando la visión de segmentación con la de elección voluntaria. Enseguida, se modelan las trayectorias laborales de los egresados universitarios mediante procesos de Markov de primer orden, describiendo los flujos de movilidad entre estados laborales. Posteriormente, se desarrolla el marco de resultados potenciales y se formaliza el sesgo de selección que motiva el uso de métodos de inferencia causal. El capítulo cierra con la metodología de Emparejamiento por Puntajes de Propensión (PSM), incluyendo sus supuestos teóricos, la especificación del modelo Logit y los criterios de evaluación del balance de covariables.

1.1. Revisión de literatura

La informalidad laboral es uno de los fenómenos más persistentes en los mercados de trabajo de América Latina y, en particular, en México. Su relevancia no radica únicamente en su magnitud. En el caso mexicano, más del 54 % de la población ocupada trabajó en condiciones informales entre 2020 y 2024, sin registrar una reducción significativa (OCDE, 2025). Además, la informalidad tiene implicaciones sobre productividad, protección social, ingresos laborales y movilidad económica. Ante este panorama, surgen dos preguntas: ¿Por qué persiste la informalidad? y ¿Cuáles son sus consecuencias para el bienestar de los trabajadores?

Desde el plano teórico, la primera interpretación es la visión de exclusión o segmentación laboral. Por segmentación se entiende la existencia de barreras que impiden que trabajadores con características similares accedan en igualdad de condiciones al empleo formal. Bajo esta perspectiva, el sector informal funciona como un espacio laboral secundario al que ingresan trabajadores que no logran acceder a empleos formales. La informalidad no representa una alternativa equivalente al empleo formal, sino una condición asociada a menores ingresos, menor estabilidad y ausencia de protección social (Harris y Todaro, 1970; Rauch, 1991). Estas lecturas se fundamentan en los modelos de mercados laborales duales, en los cuales existen barreras de entrada al sector formal

y las transiciones entre sectores no son completamente libres.

La segunda interpretación teórica es la visión de elección voluntaria o de mercados laborales integrados. Esta perspectiva supone que los trabajadores pueden desplazarse con libertad entre formalidad e informalidad y que, por tanto, la permanencia en el sector informal puede responder a una decisión racional basada en la comparación de costos y beneficios. Bajo esta lógica, la informalidad puede ofrecer flexibilidad, menores obligaciones fiscales o regulatorias y, en ciertos casos, mayor autonomía laboral. Maloney (2004) desarrolla esta argumentación para el caso latinoamericano, señalando que el sector informal no debe entenderse exclusivamente como un espacio residual, sino como un conjunto heterogéneo de actividades que puede incluir decisiones voluntarias de autoempleo y trayectorias laborales más fluidas entre sectores. Perry et al. (2007), por su parte, presenta ambas visiones, exclusión y elección voluntaria, como marcos complementarios más que excluyentes, reconociendo que distintos trabajadores y contextos pueden responder a situaciones distintas.

En el caso mexicano, la hipótesis de Levy (2008) resulta especialmente relevante porque traslada el debate del comportamiento individual al diseño institucional de la política social, ofreciendo una explicación específicamente mexicana de la persistencia de la informalidad. Su argumento central es que el esquema contributivo, en el que los trabajadores formales acceden a la seguridad social mediante descuentos de nómina, mientras que los informales pueden acceder a programas de protección social financiados con recursos fiscales generales¹, genera incentivos que reducen el atractivo del empleo formal (Levy, 2008). Bajo esta lógica, la expansión de estos programas podría inducir mayor informalidad si los mercados laborales fueran suficientemente integrados y los trabajadores pudieran desplazarse libremente entre sectores. Desde la perspectiva de esta investigación, esta hipótesis resulta parcialmente aplicable al grupo de egresados universitarios. Para que la permanencia en la informalidad pueda interpretarse principalmente como una decisión racional, la compensación total entre sectores, monetaria y no monetaria, debería ser equivalente. Sin embargo, para trabajadores con alta calificación, los diferenciales salariales a favor del sector formal documentados en la literatura (Esquivel Hernández, 2008; Gong y van Soest, 2002) sugieren que esa equivalencia no siempre se sostiene. En ese sentido, la permanencia en la informalidad podría no responder únicamente a una elección racional libre, sino también a la existencia de barreras de acceso al empleo formal. Esto no descarta que, en perfiles específicos, como profesionistas independientes o sectores con alta evasión fiscal o de prestaciones laborales, también operen factores asociados a la elección individual. Sin embargo, para el grupo analizado en esta investigación, la hipótesis de segmentación ofrece un marco adecuado para examinar si la transición hacia la formalidad se asocia con mejores resultados salariales.

En el plano empírico, diversos estudios han buscado contrastar estas interpretaciones en el caso mexicano.

El estudio de Esquivel Hernández (2008) es una referencia central para esta investigación. Con datos de la Encuesta Nacional de Empleo Urbano (ENEU), publicada por el INEGI para zonas

¹Los programas sociales no contributivos son transferencias o servicios de protección social que no requieren cotización laboral para su acceso. En México, ejemplos representativos son el Seguro Popular (2004-2019) y, actualmente, el IMSS-Bienestar.

urbanas de México, el autor examina una condición necesaria de los modelos de mercados integrados: si los sectores formal e informal son competitivos, los trabajadores informales deberían recibir una compensación salarial que equilibre la ausencia de prestaciones y seguridad social. En cambio, si los trabajadores formales obtienen mayores ingresos que informales con características similares, ello constituiría evidencia compatible con segmentación laboral, pues sugeriría que el acceso al empleo formal y a sus beneficios asociados no ocurre en igualdad de condiciones. Para evaluar esta hipótesis, el autor aplica la metodología de Emparejamiento por Puntajes de Propensión (PSM).

Los resultados de Esquivel Hernández (2008) muestran la existencia de un premio salarial asociado al empleo formal: trabajadores con características similares obtienen, en promedio, mayores salarios cuando se ubican en el sector formal que cuando trabajan en el sector informal. Este hallazgo no es consistente con la hipótesis de mercados laborales plenamente integrados y favorece una interpretación de segmentación en el mercado laboral mexicano. En consecuencia, el autor concluye que la expansión de programas sociales no puede considerarse, por sí sola, la causa principal del aumento o persistencia de la informalidad, dado que la movilidad entre sectores parece estar limitada por barreras de acceso al empleo formal.

Otros estudios han llegado a conclusiones más matizadas. Pratap y Quintin (2006), por ejemplo, analizan el sector informal en países en desarrollo y encuentran que, aunque existen diferencias importantes entre empresas formales e informales en términos de tamaño, activos y productividad, no siempre hay evidencia contundente de segmentación laboral. En una línea similar, Maloney (2004) argumenta que parte de la informalidad puede reflejar decisiones voluntarias, particularmente en el caso del autoempleo. Estos trabajos sugieren que la informalidad no debe tratarse como un fenómeno homogéneo, pues puede responder a mecanismos distintos según el tipo de ocupación, el nivel de calificación, la edad, el sector de actividad y la etapa de la trayectoria laboral.

La dimensión dinámica de la informalidad ha sido abordada por Bosch y Maloney (2010), quienes analizan transiciones laborales mediante procesos de Markov en Argentina, Brasil y México. Su trabajo muestra que los estados laborales presentan diferentes grados de persistencia y que la informalidad no constituye un bloque uniforme. En particular, distinguen entre formas de informalidad más asociadas al autoempleo, que pueden reflejar decisiones voluntarias, y formas de informalidad asalariada, que se aproximan más a una lógica de exclusión o cola de entrada al empleo formal.

Esta distinción es relevante para estudios como el presente, porque permite interpretar las transiciones laborales no solo como cambios entre categorías, sino como evidencia de diferentes mecanismos de inserción y movilidad en el mercado de trabajo. En este sentido, observar una condición laboral en un solo momento no permite distinguir entre permanencia y movilidad. Para esta investigación, la dimensión temporal es relevante porque permite diferenciar a quienes permanecen en la informalidad de quienes logran transitar hacia la formalidad, lo que ofrece una base más adecuada para analizar si la formalización se asocia con cambios en los ingresos laborales.

Alcaraz et al. (2015) encuentran evidencia de barreras de entrada al sector formal para una fracción de trabajadores informales, aunque también reconocen que una parte de la informalidad

puede responder a decisiones de autoselección². Este hallazgo permite superar una lectura dicotómica del fenómeno: la informalidad mexicana no parece ser completamente voluntaria, pero tampoco puede explicarse únicamente como exclusión. La evidencia apunta a un mercado laboral heterogéneo donde coexisten mecanismos de segmentación, restricciones de acceso y decisiones individuales.

La presente investigación se inserta en esta discusión al analizar un grupo específico: egresados universitarios en México. A diferencia de los estudios que consideran al conjunto de trabajadores urbanos, este trabajo se concentra en individuos con educación superior, para quienes podría esperarse una mayor probabilidad de acceso al empleo formal. Sin embargo, la presencia de egresados universitarios en la informalidad sugiere que la certificación educativa no garantiza por sí misma la inserción formal, lo que convierte a este grupo en un caso especialmente relevante: si incluso trabajadores con educación superior enfrentan dificultades para acceder o permanecer en el empleo formal, ello sugiere que la informalidad no puede explicarse únicamente por diferencias en capital humano, sino también por restricciones de acceso dentro del mercado laboral. La pertinencia de este estudio radica, además, en que la literatura sobre informalidad ha prestado escasa atención específica a los egresados universitarios como grupo de análisis, concentrándose principalmente en trabajadores de baja y mediana calificación.

La investigación retoma la motivación de Esquivel Hernández (2008), quien documenta empíricamente el premio salarial formal en México mediante PSM con datos transversales de la ENEU. Sin embargo, este estudio amplía ese enfoque en dos sentidos. Primero, se concentra específicamente en egresados universitarios. Segundo, adopta una perspectiva dinámica: en lugar de comparar trabajadores formales e informales en un momento dado, analiza egresados que transitan de la informalidad a la formalidad frente a egresados observacionalmente similares que permanecen informales, utilizando el diseño panel de la ENOE durante el periodo 2005–2025.

Este enfoque permite aproximar el efecto de la formalización sobre los ingresos bajo el supuesto de selección sobre observables, al comparar egresados que presentan características preexistentes similares antes de la transición laboral. Así, el trabajo contribuye a la literatura al aportar evidencia sobre el efecto de la formalización en los ingresos de egresados universitarios mexicanos, bajo el supuesto de selección sobre observables propio del emparejamiento por puntajes de propensión.

1.2. Movilidad Laboral en México: transiciones entre estados laborales

Antes de analizar el impacto de la formalidad sobre los ingresos de egresados universitarios, es necesario comprender cómo los egresados universitarios transitan entre los distintos estados del mercado laboral a lo largo del tiempo. Como señala Bosch y Maloney (2010), el análisis

²La autoselección se refiere al proceso mediante el cual los individuos no se distribuyen aleatoriamente entre estados laborales, sino que su ubicación puede depender de sus características, preferencias, expectativas, restricciones u oportunidades.

dinámico permite ir más allá de una descripción estática del mercado laboral, al observar no solo la distribución de trabajadores entre estados, sino también sus movimientos de entrada, salida y permanencia.

Para ello, esta investigación emplea matrices de transición, las cuales permiten describir los movimientos entre estados laborales e identificar patrones de permanencia y cambio a partir de sus probabilidades de transición. Este análisis resulta fundamental para determinar si la formalidad constituye un estado de estabilidad o una etapa transitoria dentro de la trayectoria laboral de los egresados universitarios.

La trayectoria laboral de un egresado universitario en México puede modelarse mediante un proceso estocástico. Sea $\{S_t\}_{t \in \mathbb{N}}$ una secuencia de variables aleatorias que representan el estado laboral en el tiempo t . Se define el espacio de estados como

$$\mathcal{S} = \{F, I, D, N\},$$

donde F representa empleo formal, I empleo informal, D desempleo y N inactividad.

A fin de reducir la dependencia temporal del proceso y permitir su representación mediante matrices de transición, se asume que el sistema satisface la propiedad de Markov de primer orden³. En términos formales, esto implica que la distribución de probabilidad del estado futuro depende únicamente del estado inmediatamente anterior, y no de toda la trayectoria pasada:

$$P(S_t = j \mid S_{t-1} = i, S_{t-2} = k, \dots, S_0 = m) = P(S_t = j \mid S_{t-1} = i)$$

para todo $t \in \mathbb{N}$ y cualquier secuencia de estados en $\mathcal{S}$. De esta definición se derivan las probabilidades de transición,

$$p_{ij} = P(S_t = j \mid S_{t-1} = i),$$

las cuales resumen la probabilidad condicional de pasar del estado i al estado j entre dos periodos consecutivos.

1.2.1. Procesos de Markov, persistencia y matriz de transición

Los flujos laborales se resumen en la matriz de transición estocástica $\mathbf{P}$. En este marco, cada elemento p_{ij} representa la probabilidad condicional de tránsito entre los estados del espacio $\mathcal{S} = \{F, I, D, N\}$:

$$\mathbf{P} = \begin{pmatrix} p_{FF} & p_{FI} & p_{FD} & p_{FN} \\ p_{IF} & p_{II} & p_{ID} & p_{IN} \\ p_{DF} & p_{DI} & p_{DD} & p_{DN} \\ p_{NF} & p_{NI} & p_{ND} & p_{NN} \end{pmatrix}$$

³La denominación primer orden distingue este supuesto de procesos de Markov de orden superior, en los que el estado futuro depende de los últimos $k > 1$ estados observados, y no exclusivamente del estado anterior. Véase Karlin y Taylor (1975, cap. 2).

En particular, sus entradas son no negativas y la suma de las probabilidades de cada fila debe ser igual a uno, de modo que la probabilidad total de transición desde un estado dado queda completamente distribuida entre los posibles estados de destino (Karlin y Taylor, 1975, cap. 2).
Propiedades:

- **Condición de Estocasticidad:** Dado que cada fila representa una distribución de probabilidad condicionada al estado inicial, sus elementos deben ser no negativos y satisfacer

$$\sum_{j \in \mathcal{S}} p_{ij} = 1.$$

Esto garantiza que la probabilidad total de transición desde el estado i se distribuye entre todos los estados posibles de destino.

- **Persistencia (Diagonal Principal):** Los coeficientes ubicados en la diagonal principal, p_{ii} , miden la probabilidad de que el proceso permanezca en el mismo estado entre dos periodos consecutivos, es decir, $P(S_t = i \mid S_{t-1} = i)$. Valores cercanos a uno indican alta persistencia o estados cuasi-absorbentes.
- **Movilidad entre estados:** Los elementos fuera de la diagonal principal, p_{ij} con $i \neq j$, miden la probabilidad de transición desde el estado i hacia un estado distinto j . En este estudio, dichos elementos permiten cuantificar las probabilidades de transición entre estados laborales.

Desde la perspectiva de los procesos estocásticos, el interés central de esta investigación se encuentra en la relación entre la probabilidad de permanencia en la informalidad, p_{II} , y la probabilidad de transitar hacia la formalidad, p_{IF} . Esta relación permite evaluar, en términos descriptivos, si la informalidad opera principalmente como un estado transitorio o si presenta patrones de alta persistencia dentro de la trayectoria laboral. En particular, valores elevados de p_{II} acompañados de valores relativamente bajos de p_{IF} no permiten afirmar por sí mismos la existencia de una trampa de informalidad, pero sí serían compatibles con la posibilidad de que ciertos individuos enfrenten dificultades para salir de la informalidad y transitar hacia la formalidad. Esta interpretación se vincula con la literatura que entiende la informalidad no solo como una elección voluntaria, sino también como resultado de exclusión o barreras de acceso al empleo formal (Levy, 2008; Perry et al. 2007)

1.2.2. Del análisis de flujos al análisis de impacto

El análisis mediante procesos de Markov permite construir un mapa de trayectorias laborales: identifica hacia dónde se mueven los egresados y con qué frecuencia lo hacen. Sin embargo, la matriz de transición, por su naturaleza descriptiva, revela el movimiento entre estados, pero no explica el beneficio económico asociado a dichos movimientos. En consecuencia, queda abierta una interrogante fundamental para el mercado laboral:

¿Refleja la formalidad un efecto causal sobre los ingresos de egresados universitarios, derivado de sus características propias, o constituye únicamente un resultado de procesos de selección asociados a diferencias preexistentes entre individuos?

En particular, si la matriz de transición muestra que la probabilidad de transitar hacia la formalidad, p_{IF} , varía entre grupos con distintas características observables (años de educación, edad, región, estado civil o tipo de localidad), surge la necesidad de adoptar un enfoque de inferencia causal. Dicho enfoque permite distinguir si el beneficio económico asociado a la formalidad representa un impacto atribuible al estado laboral o si está contaminado por ventajas previas de los individuos que lograron formalizarse.

1.3. Fundamentos de Inferencia Causal

La caracterización de la movilidad laboral mediante procesos estocásticos permite establecer que el tránsito hacia la formalidad no puede asumirse, a priori, como un fenómeno de asignación aleatoria. La estructura de la matriz de transición $\mathbf{P}$ proporciona evidencia sobre posibles regímenes de persistencia y barreras de entrada que condicionan la trayectoria de los egresados en función de sus características individuales y del contexto del mercado laboral.

En este contexto, un análisis basado exclusivamente en correlaciones resulta insuficiente, ya que no permite separar las diferencias salariales atribuibles a características preexistentes de aquellas derivadas estrictamente del estado laboral. Por lo tanto, con el fin de cuantificar si la transición hacia la formalidad se asocia con un incremento en el ingreso atribuible a dicho estado, se requiere adoptar un enfoque de inferencia causal. Este enfoque permite trascender la descripción de flujos y situar el problema dentro de un marco de identificación de efectos de tratamiento, diseñado para mitigar el sesgo de selección en entornos observacionales.

En el estudio de mercados laborales segmentados, la inserción de los individuos en el empleo formal o informal no puede asumirse como aleatoria. Por el contrario, puede responder a procesos de selección asociados tanto a características observables (educación, edad, región, estado civil o tipo de localidad), como a características no observables (habilidades, redes de contacto o motivación laboral) (Esquivel Hernández, 2008; Gong y van Soest, 2002; Perry et al. 2007). En este contexto, una simple comparación de medias entre trabajadores formales e informales podría atribuir erróneamente a la formalidad diferencias salariales que, al menos en parte, responden a características previas de los egresados.

La necesidad de controlar por estas diferencias resulta especialmente relevante en el caso de los trabajadores con educación superior. Huesca Reynoso y Padilla Arriola (2012) muestran que, aunque la escolaridad reduce la probabilidad de inserción en la informalidad como asalariado, no la elimina: incluso los trabajadores calificados pueden desempeñarse en el sector informal, de manera particular en regiones con bajo dinamismo económico, y en el caso del trabajo por cuenta propia la educación no favorece su formalización. Ello sugiere que la segmentación laboral no responde únicamente a diferencias de capital humano, sino también a factores estructurales del mercado, como las distorsiones laborales y la débil demanda local de trabajo calificado.

1.3.1. El Marco de Resultados Potenciales (Rubin Causal Model)

Para abordar el impacto de la formalidad de manera rigurosa, esta investigación se fundamenta en el enfoque de resultados potenciales, introducido por Neyman (1990) y formalizado posteriormente por Rubin (1974). Este marco conceptual redefine la causalidad como una comparación entre escenarios contrafactuales, lo cual resulta especialmente útil en estudios observacionales donde no es posible realizar experimentos aleatorios.

El enfoque de resultados potenciales se compone de dos elementos centrales.

1. **Comparaciones entre resultados potenciales:** se contrasta el resultado observado con el resultado que habría ocurrido si la unidad hubiera recibido un tratamiento distinto.
2. **Mecanismo de asignación:** Se refiere al proceso probabilístico que determina la asignación del tratamiento a las unidades, formalizado como la probabilidad condicional de recibir el tratamiento dado un conjunto de covariables observables.

Sea $i \in \{1, \dots, N\}$ un individuo perteneciente al conjunto de egresados universitarios, para cada unidad de análisis, se definen las siguientes variables:

1. **Variable de Tratamiento (T_i):** Denota la condición laboral del individuo i e identifica su pertenencia a los grupos de estudio:

$$T_i = \begin{cases} 1 & \text{si el individuo transita al empleo formal (Grupo de tratamiento)} \\ 0 & \text{si el individuo se mantiene en empleo informal (Grupo de control)} \end{cases}$$

Se asume que ambos grupos se encuentran activos en el mercado laboral, es decir, $S_i \in \{F, I\}$, excluyendo los estados de desempleo (D) e inactividad (N), pues estos no constituyen un grupo de comparación válido para evaluar el impacto salarial de la formalidad.

2. **Resultados Potenciales (Y_{1i}, Y_{0i}):** Para cada unidad i , definimos el ingreso mensual en ambos estados posibles:

- Y_{1i} : Ingreso del individuo i bajo condiciones de empleo formal ($T_i = 1$).
- Y_{0i} : Ingreso del individuo i bajo condiciones de empleo informal ($T_i = 0$).

El efecto causal individual se define como la diferencia entre ambos resultados potenciales:

$$\delta_i = Y_{1i} - Y_{0i}.$$

Sin embargo, esta diferencia no puede observarse directamente, ya que cada individuo solo puede encontrarse en uno de los dos estados laborales en un momento determinado.

1.3.2. SUTVA y el problema fundamental de la inferencia causal

Para transitar del marco teórico de resultados potenciales a una estimación estadística válida, resulta necesario asumir la estabilidad del sistema de tratamiento. Esto se establece mediante el Supuesto de Estabilidad del Valor del Tratamiento por Unidad, conocido como SUTVA por sus siglas en inglés.

1. **Ausencia de interferencia (No interacción):** Postula que el resultado potencial de una unidad i es independiente de la asignación de tratamiento de cualquier otra unidad j . Formalmente, si definimos el vector de tratamientos de la población como $Z = (T_1, T_2, \dots, T_N)$, el supuesto implica que:

$$Y_i(Z) = Y_i(T_i)$$

En el contexto de este trabajo, esto significa que el ingreso mensual de un egresado no se ve afectado por la decisión de formalización de sus pares. Sin esta restricción, el efecto causal estaría contaminado por efectos de red o saturación del mercado.

2. **Consistencia del tratamiento (Unicidad de la versión):** Postula que existe una única versión del tratamiento, de modo que el resultado potencial $Y_i(\cdot)$ está bien definido para cada valor $z \in \{0, 1\}$. Formalmente:

$$T_i = z \implies Y_i = Y_i(z)$$

Es decir, el resultado observado del individuo i bajo el tratamiento T coincide exactamente con su resultado potencial bajo ese mismo tratamiento. Este supuesto descarta la posibilidad de que existan múltiples versiones del tratamiento con efectos distintos, que se mezclarían de manera no identificable en la estimación.

Bajo el cumplimiento del SUTVA, el ingreso real percibido por el egresado se vincula con los resultados potenciales mediante una combinación lineal que depende del estado de la variable T_i :

$$Y_i = T_i Y_{1i} + (1 - T_i) Y_{0i}$$

Esta expresión confirma que el ingreso observado es una función de selección que “activa” uno de los dos escenarios teóricos, dejando al otro en un estado contrafactual inaccesible de forma directa:

- Si $T_i = 1$: $Y_i = Y_{1i}$ (ingreso observado bajo formalidad).
- Si $T_i = 0$: $Y_i = Y_{0i}$ (ingreso observado bajo informalidad).

De aquí surge el problema fundamental de la inferencia causal, formulado por Holland (1986): para cada individuo, nunca es posible observar simultáneamente Y_{1i} y Y_{0i} . Si un egresado accede

a la formalidad, su ingreso bajo la informalidad queda como un escenario hipotético; si permanece en la informalidad, su ingreso bajo formalidad no puede observarse.

Dada esta limitación, el problema se traslada del nivel individual al nivel estadístico. La inferencia causal se formula entonces como la estimación de efectos promedio en la población, donde el objetivo es reconstruir el contrafactual faltante mediante grupos de comparación adecuados.

1.4. Comparación causal y sesgo de selección

Habiendo establecido que el efecto causal individual es inobservable, el reto metodológico recae en la construcción de un contrafactual válido. En esta sección se contrasta la solución ideal, basada en experimentos aleatorios, con la realidad de los datos observacionales utilizados en esta investigación.

1.4.1. Contrafactuales y estudios observacionales

La inferencia causal en estudios observacionales se basa en la construcción de escenarios hipotéticos. Para un egresado que transitó al sector formal ($T_i = 1$), con ingreso observado Y_{1i} , su contrafactual es Y_{0i} : el ingreso que habría percibido si hubiera permanecido en la informalidad. Dado que Y_{0i} es inobservable para los tratados, la estrategia estadística consiste en sustituir el contrafactual individual por un promedio grupal. No obstante, la validez de esta sustitución depende críticamente del mecanismo de asignación al tratamiento.

Los diseños experimentales y los estudios observacionales son dos enfoques fundamentales en la investigación científica. Ambos enfoques permiten basar las decisiones en pruebas rigurosas y científicas, lo que, a su vez, contribuye de manera significativa al desarrollo de estrategias eficaces y eficientes para abordar problemas de interés.

Diseños experimentales.

En un experimento aleatorio controlado, la asignación del tratamiento T es independiente de los resultados potenciales. Matemáticamente, se satisface la propiedad de independencia fuerte:

$$(Y_1, Y_0) \perp T$$

Esta independencia garantiza que

$$E[Y_0 | T = 1] = E[Y_0 | T = 0].$$

En este escenario, el grupo de control es un contrafactual perfecto, y el efecto del tratamiento se reduce a una simple diferencia de esperanzas observadas:

$$E[Y | T = 1] - E[Y | T = 0] = E[Y_1 - Y_0] = ATE$$

Sin embargo, esta investigación se sitúa en el ámbito de los estudios observacionales. En el mercado laboral, el acceso a la formalidad no ocurre mediante asignación aleatoria, sino a través de mecanismos de selección asociados con características individuales y estructurales. Variables

como educación, edad, región, tipo de localidad, redes laborales, experiencia previa o habilidades no observadas pueden influir simultáneamente en la probabilidad de acceder al empleo formal y en el nivel de ingresos.

Para abordar esta limitación, los estudios observacionales recurren a métodos estadísticos que permiten controlar, en la medida de lo posible, el efecto de estas variables confusoras y aproximarse a inferencias causales más sólidas. Sin embargo, como señala Rubin (2008), estos estudios suelen estar sujetos a una serie de desafíos metodológicos que complican la tarea de establecer afirmaciones causales objetivas.

La asignación aleatoria en los experimentos garantiza que los grupos de tratamiento y control sean comparables en todas las variables, tanto observables como no observables. Esto permite estimar con mayor confianza el efecto causal del tratamiento. En cambio, en los estudios observacionales, la ausencia de asignación aleatoria introduce un sesgo de selección, ya que los individuos que deciden recibir un tratamiento pueden diferir sistemáticamente de aquellos que no lo reciben. Para mitigar este sesgo, técnicas como el emparejamiento por puntajes de propensión buscan crear grupos comparables en términos de sus características observables, aproximando así las condiciones de un experimento aleatorio.

1.4.2. El Sesgo de Selección y la Descomposición del Impacto

Cuando la independencia entre tratamiento y resultados potenciales se viola, surge el sesgo de selección. Para visualizar su magnitud, se descompone la diferencia de ingresos promedio observada entre egresados que transitaron a la formalidad y egresados que permanecieron en la informalidad:

$$\underbrace{E[Y | T = 1] - E[Y | T = 0]}_{\text{Diferencia Observada}} = \underbrace{E[Y_1 - Y_0 | T = 1]}_{\text{ATT}} + \underbrace{\{E[Y_0 | T = 1] - E[Y_0 | T = 0]\}}_{\text{Sesgo de Selección } (\alpha)}$$

Donde:

- **ATT (Average Treatment Effect on the Treated):** Es el impacto real de la formalidad que deseamos aislar.
- **Sesgo de Selección (α):** Representa la diferencia en los ingresos potenciales entre los grupos.

En el mercado laboral mexicano, es altamente probable que $\alpha > 0$. Esto ocurre porque los individuos con mayores habilidades no observadas tienen mayor probabilidad de transitar a la formalidad y, simultáneamente, tendrían mayores ingresos incluso si estuvieran en la informalidad, es decir,

$$E[Y_0 | T = 1] > E[Y_0 | T = 0]$$

En ese caso, la brecha salarial observada entre formales e informales sobrestimaría el efecto causal de la formalidad. Como señalan Gong y van Soest (2002), las diferencias salariales entre

sectores no pueden atribuirse enteramente al tipo de empleo, sino también a la composición de los individuos que los ocupan.

1.5. Medidas Fundamentales de Impacto

Una vez establecido el problema de identificación causal, es necesario definir los parámetros poblacionales que permiten cuantificar el beneficio monetario de la formalidad. En esta sección se presentan el efecto promedio del tratamiento y el efecto promedio del tratamiento sobre los tratados, enfatizando el papel central de este último en la investigación.

1.5.1. Efecto promedio del tratamiento: ATE y ATT

El Average Treatment Effect (ATE) representa el impacto esperado de la formalidad si esta se asignara aleatoriamente a un individuo tomado al azar de la población total de egresados activos en el mercado laboral. Matemáticamente, se define como:

$$ATE = E[\delta] = E[Y_1 - Y_0]$$

Este parámetro responde a una pregunta general:

¿Cuál sería el cambio promedio en el ingreso si la totalidad de la población de egresados activos fuera formalizada?

Sin embargo, en mercados con alta heterogeneidad y autoselección, el ATE puede diluir el efecto real que experimentan quienes efectivamente logran insertarse en el sector formal.

Por esta razón, el parámetro de mayor interés para esta investigación es el efecto promedio del tratamiento sobre los tratados (ATT). El ATT mide el impacto promedio de la formalidad exclusivamente para el grupo que recibió el tratamiento:

$$ATT = E[Y_1 - Y_0 \mid T = 1] = E[Y_1 \mid T = 1] - E[Y_0 \mid T = 1]$$

- El término $E[Y_1 \mid T = 1]$ es directamente identificable a partir de los datos, pues corresponde al ingreso promedio observado de los egresados que transitaron a la formalidad.
- El desafío reside en el segundo término, $E[Y_0 \mid T = 1]$, el cual representa el ingreso que esos mismos egresados habrían obtenido si hubieran permanecido en la informalidad. Este es el contrafactual que la metodología debe aproximar.

La elección del ATT como estimando central se justifica porque el objetivo de esta investigación no es evaluar el efecto hipotético de formalizar a toda la población, sino medir el impacto sobre aquellos egresados que efectivamente realizaron la transición hacia el sector formal. Este enfoque es especialmente relevante para evaluar políticas de formalización dirigidas a profesionistas universitarios.

1.5.2. Estimación empírica del ATT

Para llevar el ATT del plano teórico al plano empírico, se requiere un estimador que aproxime el contrafactual faltante mediante el uso de un grupo de comparación. El estimador del ATT, denotado por $\hat{\tau}_{ATT}$, se expresa como:

$$\hat{\tau}_{ATT} = \frac{1}{N_1} \sum_{i: T_i=1} (Y_i - \hat{Y}_{0i})$$

donde N_1 es el número de egresados tratados y $\hat{Y}_{0i}$ representa el contrafactual estimado para cada individuo tratado i .

Para construir este contrafactual, se utiliza una función de ponderación $W(i, j)$ que asigna importancia relativa a cada unidad de control j :

$$\hat{\tau}_{ATT} = \frac{1}{N_1} \sum_{i: T_i=1} \left(Y_i - \sum_{j: T_j=0} W(i, j) Y_j \right)$$

La función de ponderación es el núcleo del emparejamiento. Su rol es asignar una importancia relativa a cada unidad informal j para servir como espejo del formal i . La naturaleza de $W(i, j)$ depende del método elegido:

- **Vecino más cercano:** $W(i, j) = 1$ para el informal más parecido en términos de puntaje de propensión y 0 para los demás. Es el método más simple, pero puede generar emparejamientos de baja calidad si el vecino más cercano está muy distante en el espacio del puntaje.
- **Caliper:** Extensión del vecino más cercano que impone una restricción de distancia máxima tolerable ϵ entre los puntajes del tratado i y el control j . Formalmente, $W(i, j) = 1$ solo si $|e(X_i) - e(X_j)| \leq \epsilon$, y 0 en caso contrario. Esto garantiza que ningún emparejamiento se realice entre unidades con perfiles demasiado distantes, a costa de descartar tratados para los cuales no existe un control suficientemente cercano.
- **Kernel:** $W(i, j)$ asigna pesos decrecientes a los informales conforme su distancia al individuo i aumenta, permitiendo que múltiples unidades contribuyan a la construcción de un solo contrafactual, reduciendo así la varianza del estimador a cambio de introducir un ligero sesgo por incluir controles menos similares.

Esta estructura garantiza que la comparación no se realice entre cualquier egresado formal y cualquier egresado informal, sino entre individuos comparables en términos de características observables.

1.6. Metodología de Emparejamiento por Puntajes de Propensión

Para resolver el problema del sesgo de selección identificado anteriormente, esta investigación implementa la metodología de emparejamiento por puntajes de propensión (PSM). Esta técnica permite aproximar un entorno cuasi-experimental a partir de datos observacionales, equilibrando las características observables de los egresados que transitaron a la formalidad frente a aquellos que permanecieron en la informalidad.

1.6.1. Puntaje de propensión y reducción de dimensionalidad

El emparejamiento directo por covariables presenta un obstáculo: a medida que aumenta el número de características observables (edad, años de educación, región, sexo, estado civil o tipo de localidad), el espacio multidimensional se vuelve más disperso y la probabilidad de encontrar unidades comparables disminuye. Este fenómeno se conoce como maldición de la dimensionalidad. Rosenbaum y Rubin (1983) demostraron que es posible reducir el problema de emparejamiento desde un vector de covariables $X \in \mathbb{R}^k$ hacia una sola dimensión escalar mediante el puntaje de propensión, sin perder las propiedades de equilibrio necesarias para la inferencia causal.

Definición 1 (Puntaje de propensión). *El **puntaje de propensión**, denotado por $e(X_i)$, es la probabilidad condicional de recibir el tratamiento dado el vector de covariables observables:*

$$e(X_i) = \mathbb{P}(T_i = 1 \mid X_i) = \mathbb{E}[T_i \mid X_i].$$

Esta probabilidad recibe su denominación al reflejar la propensión de un individuo a participar en el tratamiento condicionada a sus covariables observables. Matemáticamente, se define como una aplicación $b : \mathbb{R}^k \rightarrow \mathbb{R}$, lo que implica una reducción del vector $X \in \mathbb{R}^k$ a un escalar.

Este puntaje resume toda la información relevante de las covariables en un único escalar, permitiendo el emparejamiento en una sola dimensión.

1.6.2. Supuestos y fundamentos teóricos del Emparejamiento por Puntajes de Propensión

Para que el puntaje de propensión funcione como mecanismo válido de comparación, deben cumplirse dos condiciones centrales: ignorabilidad condicional y soporte común.

La ignorabilidad condicional establece que, una vez controladas las covariables observables, los resultados potenciales son independientes de la asignación al tratamiento:

$$\{Y_0, Y_1\} \perp T \mid X.$$

Este supuesto implica que, condicional en X , la asignación al tratamiento se comporta como si fuera aleatoria: entre individuos con las mismas características observables. En consecuencia, las

diferencias de ingreso entre tratados y controles con covariables similares pueden atribuirse al tratamiento y no a diferencias preexistentes capturadas por X .

El segundo supuesto es el de soporte común u overlap, el cual exige que para todo perfil observable exista una probabilidad positiva de pertenecer tanto al grupo tratado como al grupo de control:

$$0 < P(T_i = 1 \mid X_i) < 1.$$

En aplicaciones de PSM, el soporte común se operacionaliza restringiendo el análisis a la intersección de los soportes empíricos de tratados y controles:

$$S = [\max\{\min(e \mid T = 1), \min(e \mid T = 0)\}, \min\{\max(e \mid T = 1), \max(e \mid T = 0)\}].$$

Las observaciones fuera de esta región son descartadas, de modo que los resultados se interpretan como válidos para la subpoblación donde existe comparabilidad empírica.

El fundamento teórico del PSM descansa en la propiedad de equilibrio del puntaje de propensión. Rosenbaum y Rubin (1983) muestran que el puntaje de propensión es un puntaje de equilibrio (balancing score).

Definición 2 (Puntaje de equilibrio). Sea $b : \mathbb{R}^k \rightarrow \mathbb{R}^p$ con $p \leq k$, una función de las covariables observables. Se dice que $b(X)$ es un **puntaje de equilibrio** (balancing score) si:

$$X \perp T \mid b(X).$$

La relevancia de esta definición reside en que un puntaje de equilibrio resume la información de X relevante para la asignación al tratamiento. Si dos individuos comparten valores similares de $b(X)$, su distribución de covariables es estadísticamente equivalente, independientemente de si recibieron o no el tratamiento. Esto permite realizar comparaciones válidas sin necesidad de igualar todos los valores del vector X .

Teoremas Fundamentales

El sustento teórico del PSM descansa en los resultados formales de Rosenbaum y Rubin (1983), quienes definen el puntaje de propensión como la probabilidad condicional de asignación al tratamiento dado un vector de covariables observadas, $e(X) = \mathbb{P}(T = 1 \mid X)$, y demuestran que dicho puntaje constituye un puntaje de equilibrio. En conjunto, estos resultados justifican la reducción del problema de emparejamiento multidimensional a una sola dimensión escalar, siempre que se cumplan los supuestos de ignorabilidad y soporte común Rosenbaum y Rubin (1983).

Teorema 1 (Propiedad de equilibrio del puntaje de propensión). El puntaje de propensión $e(X)$ es un puntaje de equilibrio. Es decir:

$$T \perp X \mid e(X)$$

Demostración. Es equivalente a demostrar que:

$$\mathbb{P}(T = 1 \mid X, e(X)) = \mathbb{P}(T = 1 \mid e(X)).$$

Por la definición del puntaje de propensión, se tiene que $\mathbb{P}(T = 1 \mid X) = e(X)$. Dado que $e(X)$ es función de X , condicionar simultáneamente en X y $e(X)$ es equivalente a condicionar solo en X :

$$\mathbb{P}(T = 1 \mid X, e(X)) = \mathbb{P}(T = 1 \mid X) = e(X).$$

Por la Ley de Esperanzas:

$$\mathbb{P}(T = 1 \mid e(X)) = \mathbb{E}[T \mid e(X)] = \mathbb{E}[\mathbb{E}[T \mid X] \mid e(X)] = \mathbb{E}[e(X) \mid e(X)] = e(X).$$

Por tanto:

$$\mathbb{P}(T = 1 \mid X, e(X)) = \mathbb{P}(T = 1 \mid e(X)) = e(X),$$

lo cual, por definición de independencia condicional, implica que $T \perp X \mid e(X)$. □

Este teorema garantiza que al condicionar en el puntaje de propensión, las covariables quedan balanceadas entre grupos de tratamiento y control. En otras palabras, individuos con valores similares de $e(X)$ son comparables en términos de la distribución de sus características observables, lo que justifica realizar el emparejamiento en una sola dimensión en lugar de en el espacio completo $X \in \mathbb{R}^k$.

Definición 3 (Asignación fuertemente ignorable). *Se dice que la asignación al tratamiento es fuertemente ignorable dado un vector de covariables v si se satisfacen simultáneamente las siguientes dos condiciones:*

1. Independencia condicional (Ignorabilidad):

$$\{Y_0, Y_1\} \perp T \mid v$$

Los resultados potenciales son independientes de la asignación al tratamiento, dado v .

2. Soporte común (Overlap):

$$0 < \mathbb{P}(T = 1 \mid v) < 1$$

Para todo valor de v , existe una probabilidad positiva de recibir cualquiera de los dos estados del tratamiento.

Cuando ambas condiciones se satisfacen con $v = X$, se dice simplemente que la asignación al tratamiento es fuertemente ignorable. Si se satisfacen con $v = e(X)$, se dice que es fuertemente ignorable dado el puntaje de propensión.

Esta definición es el supuesto central que habilita la inferencia causal en estudios observacionales. La condición Independencia Condicional (CIA) establece que una vez controladas las covariables relevantes, la asignación al tratamiento puede tratarse como si fuera aleatoria. Por su parte, la condición de Soporte Común garantiza que la comparación entre unidades tratadas y de control sea empíricamente posible.

Ambas condiciones deben satisfacerse conjuntamente para que el efecto promedio del tratamiento sea identificable. El siguiente teorema establece que, si la asignación es fuertemente ignorable dado X , entonces también lo es dado el escalar $e(X)$, lo cual es el resultado que habilita el PSM.

Teorema 2 (Ignorabilidad fuerte dado $e(X)$). *Si la asignación al tratamiento es fuertemente ignorable dado X , entonces también lo es dado $e(X)$. Es decir:*

$$\{Y_0, Y_1\} \perp T \mid X \quad y \quad 0 < \mathbb{P}(T = 1 \mid X) < 1$$

implica que:

$$\{Y_0, Y_1\} \perp T \mid e(X) \quad y \quad 0 < \mathbb{P}(T = 1 \mid e(X)) < 1.$$

Demostración. La segunda condición se sigue directamente del Teorema anterior, pues:

$$\mathbb{P}(T = 1 \mid e(X)) = e(X) = \mathbb{P}(T = 1 \mid X),$$

por lo que el soporte común se preserva al pasar de X a $e(X)$.

Para la independencia, sea $k \in \{0, 1\}$ y considérese el resultado potencial Y_k . Por la Ley de Esperanzas Iteradas:

$$\mathbb{P}(T = 1 \mid Y_k, e(X)) = \mathbb{E}[\mathbb{E}[T \mid Y_k, X] \mid Y_k, e(X)].$$

Bajo el supuesto de ignorabilidad fuerte $\{Y_0, Y_1\} \perp T \mid X$, la distribución de T condicional en (Y_k, X) colapsa a la distribución condicional en solo X :

$$\mathbb{E}[T \mid Y_k, X] = \mathbb{E}[T \mid X] = e(X).$$

Sustituyendo:

$$\mathbb{P}(T = 1 \mid Y_k, e(X)) = \mathbb{E}[e(X) \mid Y_k, e(X)] = e(X) = \mathbb{P}(T = 1 \mid e(X)).$$

La igualdad implica $T \perp Y_k \mid e(X)$. Aplicando el resultado para $k = 0$ y $k = 1$ conjuntamente:

$$\{Y_0, Y_1\} \perp T \mid e(X).$$

□

Este teorema es el resultado central del marco teórico del PSM. Establece que, si se controla por las covariables observables relevantes para la selección al tratamiento, entonces basta condicionar en el escalar $e(X)$ para que la asignación pueda tratarse como aleatoria. Esto permite recuperar

comparabilidad entre egresados que transitan hacia la formalidad y egresados que permanecen en la informalidad sin necesidad de un experimento controlado, siempre que el soporte común esté garantizado.

1.6.3. Especificación del Modelo Logit

El puntaje de propensión se estima mediante un modelo Logit⁴, dado que la variable de tratamiento es binaria y el objetivo es obtener probabilidades estimadas acotadas en el intervalo $[0, 1]$. Esta especificación resulta más adecuada que una regresión lineal de probabilidad estimada por MCO, ya que esta puede generar valores predichos fuera del rango admisible para una probabilidad. Si bien modelos no lineales como Logit y Probit suelen producir resultados similares en aplicaciones con variables dicotómicas, la elección del modelo Logit en esta investigación responde a tres criterios. Primero, facilita la implementación empírica del puntaje de propensión en los algoritmos de emparejamiento utilizados. Segundo, permite interpretar la asignación al tratamiento en términos de probabilidades estimadas de transición hacia la formalidad. Tercero, mantiene consistencia con la estrategia de emparejamiento por puntajes de propensión, en la cual el puntaje de propensión se define precisamente como la probabilidad condicional de recibir el tratamiento dado un conjunto de covariables observables.

La especificación general del modelo es

$$\begin{aligned} e(X_i) &= \mathbb{P}(T_i = 1 \mid X_i = x) \\ &= \frac{\exp(X_i' \beta)}{1 + \exp(X_i' \beta)} \\ &= \frac{\exp\{\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}\}}{1 + \exp\{\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}\}} \\ &= \frac{1}{1 + \exp\{-(\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik})\}}. \end{aligned}$$

Donde $\beta = (\beta_0, \beta_1, \dots, \beta_k)$ representa el vector de parámetros del modelo, estimado mediante máxima verosimilitud.

La relación anterior puede expresarse en términos de log-odds o función logit:

$$\ln\left(\frac{e(X_i)}{1 - e(X_i)}\right) = X_i' \beta$$

El vector X_i incluye covariables pretratamiento observables que pueden influir en la probabilidad de transitar hacia la formalidad. Estas variables se agrupan de la siguiente manera:

- **Variables binarias:** sexo, jefatura del hogar, estado civil y tipo de localidad.
- **Variables categóricas:** zona geográfica y rama de actividad económica.

⁴El modelo Logit es un modelo de respuesta binaria que permite estimar la probabilidad de ocurrencia de un evento a partir de un conjunto de variables explicativas. En este caso, el evento de interés es transitar de la informalidad a la formalidad laboral.

- **Variables continuas:** edad y años de educación.

La descripción detallada de cada variable y su construcción se presenta en el Capítulo 2.

A partir de la formalización general del ATT presentada anteriormente, el estimador utilizado en esta investigación incorpora explícitamente el puntaje de propensión estimado como criterio de comparación entre tratados y controles. Una vez obtenido $e(X_i)$ mediante el modelo Logit, el efecto promedio de la formalización sobre los egresados tratados se estima como:

$$\hat{\tau}_{ATT} = \frac{1}{N_1} \sum_{i:T_i=1} \left(Y_i - \sum_{j:T_j=0} W(e_i, e_j) Y_j \right)$$

Donde los pesos W dependen de la proximidad entre los puntajes estimados de los individuos que transitaron a la formalidad (i) y de los que permanecieron en la informalidad (j).

1.6.4. Evaluación del balance: sesgo estandarizado y prueba t

Una etapa fundamental en la implementación del Emparejamiento por Puntajes de Propensión consiste en evaluar si el procedimiento logró reducir las diferencias observables entre el grupo tratado y el grupo de control. Esta revisión se conoce como evaluación del balance de covariables y permite verificar si, después del emparejamiento, los individuos tratados y los controles emparejados presentan características observables suficientemente similares.

En esta investigación, el balance se evalúa principalmente mediante el sesgo estandarizado de las covariables. Este indicador mide la diferencia de medias entre tratados y controles, expresada en relación con la variabilidad conjunta de ambos grupos. De acuerdo con Caliendo y Kopeinig (2008), el sesgo estandarizado permite comparar la distancia entre las distribuciones marginales de las covariables antes y después del emparejamiento.

Para una covariable X , el sesgo estandarizado antes del emparejamiento puede expresarse como:

$$SB_{pre} = 100 \cdot \frac{\bar{X}_1 - \bar{X}_0}{\sqrt{0,5 \cdot (V_1(X) + V_0(X))}}$$

Mientras que después del emparejamiento se calcula como:

$$SB_{post} = 100 \cdot \frac{X_{1M} - X_{0M}}{\sqrt{0,5 \cdot (V_{1M}(X) + V_{0M}(X))}}$$

Donde $\bar{X}_1$ y $\bar{X}_0$ representan las medias de la covariable en el grupo tratado y de control antes del emparejamiento, respectivamente; $V_1(X)$ y $V_0(X)$ corresponden a sus varianzas. Los subíndices M indican que las medias y varianzas se calculan en la muestra emparejada. Por tanto, una reducción del sesgo estandarizado después del emparejamiento indica una mejora en la comparabilidad observable entre tratados y controles.

De manera complementaria, Caliendo y Kopeinig (2008) señala que puede utilizarse una prueba t de diferencia de medias para revisar si las covariables presentan diferencias estadísticamente significativas entre ambos grupos. Antes del emparejamiento, estas diferencias pueden ser esperables; después del emparejamiento, deberían reducirse o dejar de ser estadísticamente significativas. Sin

embargo, la prueba t se interpreta como un complemento del sesgo estandarizado, no como un sustituto.

Además del sesgo por covariable, se considera el sesgo estandarizado promedio, el cual resume el desequilibrio global del conjunto de covariables incluidas en el modelo. En términos generales, valores menores del sesgo promedio posterior al emparejamiento indican una mejor calidad del balance. Aunque no existe un umbral universal, Caliendo y Kopeinig (2008) señalan que en diversos estudios aplicados se consideran adecuados valores reducidos, comúnmente por debajo de 3 % o 5 %.

En consecuencia, los diagnósticos de balance permiten definir el grado de cautela con el que deben interpretarse los efectos estimados. Los años con menor sesgo posterior ofrecen una base más sólida para analizar el ATT, mientras que aquellos con mayor sesgo residual o diferencias significativas persistentes requieren una interpretación más prudente.

Capítulo 2

Datos y estadística descriptiva

Este capítulo presenta el proceso completo de construcción de la muestra analítica empleada en la investigación, desde la fuente original hasta la muestra final utilizada en el emparejamiento por puntajes de propensión. En primer lugar, se describe la Encuesta Nacional de Ocupación y Empleo (ENOE), su diseño de levantamiento y la estructura de sus bases de datos, argumentando la elección del módulo sociodemográfico como fuente principal. A continuación, se documenta la construcción del panel longitudinal, los criterios de inclusión de cada variable empleada en el modelo, los filtros poblacionales aplicados y la definición operativa del tratamiento y el grupo de control. El capítulo cierra con un análisis de estadística descriptiva que caracteriza la muestra de estudio y establece la comparación inicial entre grupos antes del emparejamiento, motivando así la necesidad del método de inferencia causal desarrollado en el Capítulo 3.

2.1. Fuente de Datos: Encuesta Nacional de Ocupación y Empleo (ENOE)

2.1.1. Descripción general de la ENOE

La Encuesta Nacional de Ocupación y Empleo es el instrumento estadístico de mayor alcance y continuidad para el análisis del mercado laboral en México. Es diseñada, coordinada y publicada por el Instituto Nacional de Estadística y Geografía (INEGI) con periodicidad trimestral, y su levantamiento es continuo a lo largo del año. Desde su implementación en 2005 cuando sustituyó a la Encuesta Nacional de Empleo (ENE) y a la Encuesta Nacional de Empleo Urbano (ENEU), la ENOE se ha consolidado como la fuente de referencia obligada para la medición de los indicadores laborales del país.

Su objetivo central es proporcionar información estadísticamente representativa sobre las características ocupacionales de la población de 15 años y más, abarcando dimensiones como el empleo, el desempleo, la subocupación, la informalidad laboral, las condiciones de trabajo y el nivel de ingresos, entre otras. La encuesta tiene cobertura nacional, urbana y rural, lo que permite dividir

el análisis por tamaño de localidad, entidad federativa y región geográfica.

Diseño de levantamiento rotatorio

Una de las características más valiosas de la encuesta para los propósitos de la investigación es su diseño de panel rotatorio. Bajo este esquema, cada vivienda seleccionada es visitada durante cinco trimestres consecutivos, de modo que la muestra se renueva de forma gradual: en cada trimestre, una fracción de las viviendas sale del panel y se reemplaza por viviendas nuevas, mientras que el resto continúa siendo entrevistado.

Este diseño es el fundamento metodológico del análisis de este trabajo. Al observar a cada egresado en su primer trimestre (línea base, t_1) y en su quinto trimestre (resultado, t_5), es posible identificar con precisión quiénes transitaron de la informalidad a la formalidad y quienes permanecieron en el sector informal durante el periodo de seguimiento.

2.1.2. Estructura de bases de datos

La información recabada por la ENOE se organiza en cinco bases de datos complementarias, cada una asociada a un cuestionario específico:

- **VIV** (Cuestionario de Vivienda): contiene la identificación de las viviendas seleccionadas, el número de residentes y la cantidad de hogares que las componen.
- **HOG** (Cuestionario de Hogar): registra la identificación de cada hogar dentro de la vivienda, las fechas de levantamiento y características.
- **SDEM** (Cuestionario Sociodemográfico): almacena las características sociodemográficas de todos los residentes del hogar como edad, sexo, estado civil, escolaridad, parentesco con el jefe del hogar, condición de residencia así como la clasificación laboral de cada individuo: su pertenencia a la Población Económicamente Activa (PEA) o a la Población No Económicamente Activa (PNEA).
- **COE1** (Cuestionario de Ocupación y Empleo, parte 1): contiene información detallada sobre las características del empleo principal: tipo de contrato, prestaciones, jornada laboral, ingresos y posición en el trabajo.
- **COE2** (Cuestionario de Ocupación y Empleo, parte 2): profundiza en aspectos como el empleo secundario, la búsqueda de trabajo y otras condiciones laborales específicas.

De estas cinco bases, el análisis emplea SDEM, ya que concentra en un mismo instrumento las dos dimensiones analíticas requeridas. Por un lado, las características sociodemográficas observables del individuo que sirven como covariables en el modelo Logit: edad, sexo, escolaridad, estado civil, jefatura del hogar, tipo de localidad, entidad de residencia y rama de actividad económica. Por otro lado, SDEM permite construir la clasificación del estado laboral a partir de las variables `clase1`, `clase2` y `emp_ppal`, con las cuales se identifica si un individuo se encuentra en empleo

formal, empleo informal, desempleo o inactividad en cada trimestre de observación. Adicionalmente, SDEM incluye el ingreso mensual (`ingocup`), que constituye la variable de resultado de interés.

Representatividad y factor de expansión

La ENOE es una encuesta en la cual cada observación tiene asociado un factor de expansión que indica cuántos individuos de la población nacional representa esa unidad muestral. Sin embargo, el uso del factor de expansión en el contexto del emparejamiento por puntajes de propensión requiere una distinción metodológica importante.

En este estudio, el factor de expansión se utiliza en dos etapas específicas:

- Primero, en la estadística descriptiva del panel, donde se emplean proporciones expandidas para caracterizar la evolución de los estados laborales a nivel poblacional a lo largo del periodo de análisis.
- Segundo, en la estimación del modelo Logit para el cálculo del puntaje de propensión, donde se incorpora como ponderador (`[pw=fac_tri]`). Esto garantiza que las probabilidades estimadas de formalización reflejen la estructura poblacional real y no solo la composición particular de la muestra observada, produciendo un puntaje representativo del mercado laboral mexicano para egresados universitarios.

Posteriormente, el emparejamiento se realiza utilizando el puntaje de propensión. Esta estrategia reconoce que el emparejamiento es un procedimiento de reducción de sesgo basado en la comparabilidad de covariables.

En consecuencia, el efecto se estima sin incorporar ponderadores en la etapa de cálculo del ATT, por lo que el resultado se interpreta como el efecto promedio del tratamiento sobre los tratados dentro de la muestra analizada. Esta distinción es relevante, ya que la inclusión o exclusión de los pesos de la encuesta modifica el alcance inferencial de los resultados: sin ponderadores, las estimaciones se refieren principalmente a la muestra observada; con ponderadores, la interpretación se extiende hacia la población representada por la encuesta.

2.2. Construcción del Panel Longitudinal

Estrategia de identificación de individuos

Dado el diseño rotatorio de la ENOE, es posible identificar la posición de cada observación dentro del ciclo de seguimiento mediante la variable `n_ent`, que indica el número de entrevista correspondiente.

A partir de esta información, se seleccionan los registros que permiten seguir a los mismos individuos desde su primera hasta su quinta entrevista. De manera que cada persona aparece una vez en cada periodo de observación, de t_1 a t_5 , lo que permite construir un panel de cinco observaciones por individuo.

Para que este panel sea trazable a lo largo del tiempo, es indispensable contar con un identificador único y estable por unidad observada. Este identificador se construye mediante la concatenación de siete variables administrativas que, en conjunto, identifican de manera única a cada observación registrada según su vivienda y su hogar:

$$\text{folio_ind} = \text{cd_a} + \text{ent} + \text{con} + \text{v_sel} + \text{n_hog} + \text{h_mud} + \text{n_ren}$$

donde **cd_a** es el código del área geográfica representada, **ent** la entidad federativa, **v_sel** la vivienda seleccionada, **n_hog** el número de hogar dentro de la vivienda, **h_mud** un indicador de cambio de residencia en la vivienda, y **n_ren** el número de renglón que identifica a cada persona dentro del hogar. La concatenación de estas variables permite el seguimiento longitudinal durante los cinco trimestres sin ambigüedad.

A partir del año 2021, INEGI rediseñó parcialmente el marco muestral de la ENOE como parte de la recuperación post-pandemia, introduciendo una variable adicional denominada **tipo**. Cuando esta variable está disponible, se incorpora como prefijo al folio.

Dado que **tipo** no aparece en todos los trimestres del periodo 2021–2025, su inclusión se maneja de forma condicional: cuando está disponible se agrega al folio, y cuando no lo está, la unicidad del identificador se preserva con las siete variables base.

Criterio de completitud del panel

Una vez construido el identificador único, se procede a verificar que cada individuo cuente con observaciones en los cinco trimestres que abarca el periodo de levantamiento. Solo se conservan en el panel aquellos individuos para quienes se dispone de información en todos los trimestres del ciclo, es decir, aquellos que satisfacen $n = 1, 2, 3, 4, 5$ consecutivamente. Este criterio de completitud responde a una exigencia metodológica del diseño del estudio: el análisis de impacto requiere conocer el estado laboral del individuo tanto en el primer trimestre que actúa como línea base pre-tratamiento como en el quinto trimestre que registra el resultado post-tratamiento. La ausencia de cualquiera de estas observaciones imposibilita la asignación del tratamiento y la medición del ingreso de cierre, razón por la cual los individuos con panel incompleto son excluidos del análisis.

Los individuos que no cumplen este criterio corresponden típicamente a viviendas que no pudieron ser entrevistadas en algún trimestre por razones ajenas al diseño, rechazo, ausencia prolongada del hogar o cambio de residencia fuera del área de seguimiento. En consecuencia, estas observaciones se excluyen con el fin de mantener la consistencia del seguimiento dentro del panel, aunque no puede descartarse completamente la posible existencia de algún sesgo asociado a la pérdida de seguimiento.

Panel completo: periodo 2005–2025

Aplicando la estrategia de identificación y el criterio de completitud a cada uno de los años del periodo, y apilando los resultados, se obtiene un panel longitudinal que abarca los años 2005–2025. En términos generales, cada individuo del panel cuenta con observaciones consecutivas a

lo largo del ciclo de entrevista, lo que permite registrar su estado laboral, sus características sociodemográficas y su ingreso entre el trimestre inicial y el trimestre final de seguimiento.¹

2.3. Construcción de Variables

A partir del panel longitudinal descrito en el apartado anterior, se procede a la construcción de las variables empleadas. Este proceso comprende tres tareas principales:

1. La aplicación de filtros poblacionales para delimitar la muestra de interés.
2. La construcción del estado laboral como variable central del diseño.
3. La construcción de las covariables demográficas, geográficas, educativas y de resultado.

La tabla 2.1 resume todas las variables construidas, su fuente dentro de la ENOE y la lógica de construcción empleada.

Filtros Poblacionales

La muestra se restringe a la población de interés mediante dos criterios aplicados sobre el panel completo.

- El primero es un **filtro de edad**: se conservan únicamente los individuos cuya edad se encuentra entre 15 y 65 años, conforme a la definición estándar de población en edad de trabajar adoptada tanto por la ENOE como por la Organización Internacional del Trabajo (OIT).
- El segundo es un **filtro de escolaridad**: se restringe la muestra a egresados universitarios, definidos como aquellos individuos con 16 o más años de educación acumulada. Este umbral corresponde a la conclusión de estudios de nivel licenciatura bajo el sistema educativo mexicano (12 años de educación básica y media superior más al menos 4 años de educación superior) y es consistente con la variable `años_edu` construida a partir de los campos de escolaridad de SDEM. Los individuos que no alcanzan este umbral son excluidos del análisis.

Adicionalmente, se eliminan las observaciones con inconsistencias estructurales: individuos que reportan edades incompatibles con su nivel educativo acumulado, aquellos cuya edad es menor a la que teóricamente requerirían para haber completado sus años de escolaridad.

¹Dos años del periodo requieren un tratamiento diferenciado debido a la disponibilidad de los datos. En 2020, el segundo trimestre no fue levantado por INEGI a causa de las restricciones de movilidad impuestas durante la emergencia sanitaria por COVID-19; por ello, el panel de ese año se construye con cuatro trimestres ($t1$, $t3$, $t4$ y $t5$), aplicando el criterio $n = 1, 3, 4, 5$. El año se mantiene en el análisis porque la comparación entre el estado laboral en $t1$ y en el último trimestre observado sigue siendo válida para definir el tratamiento. En 2025, el quinto trimestre requeriría el levantamiento del primer trimestre de 2026, aún no publicado al momento del corte; por ello, el panel se construye con cuatro trimestres ($t1$ a $t4$) y el cuarto trimestre se reclasifica como $t5$. De este modo, todos los bloques de construcción de variables y de definición del tratamiento operan sobre una estructura uniforme en la que el último trimestre observado siempre queda etiquetado como $t5$.

Estado Laboral

La variable central del diseño es `est_lab`, que clasifica el sector laboral del individuo en cada trimestre en cuatro estados mutuamente excluyentes. Esta variable se construye a partir de la combinación de tres variables de la base SDEM de la ENOE:

- `clase1`: distingue a la Población Económicamente Activa (PEA) de la Población No Económicamente Activa (PNEA).
- `clase2`: dentro de la PEA, separa a la población ocupada de la desocupada.
- `emp_ppal`: dentro de la población ocupada, clasifica el empleo principal como informal o formal. De acuerdo con la estructura de la ENOE, esta variable toma el valor de 1 cuando el empleo principal es informal y el valor de 2 cuando es formal. Esta clasificación no se limita a identificar si el trabajador labora en el sector informal, sino que responde al criterio de informalidad laboral definido por INEGI, el cual considera la naturaleza de la unidad económica, la posición en la ocupación y el acceso a seguridad social asociado al empleo principal.

A partir de estas tres variables, los cuatro estados se definen como:

- **Empleo formal** (`est_lab = 1`): individuos que pertenecen a la PEA, están ocupados y cuyo empleo principal es clasificado como formal por la ENOE mediante la variable `emp_ppal`.
- **Empleo informal** (`est_lab = 2`): individuos que pertenecen a la PEA, están ocupados y cuyo empleo principal es clasificado como informal por la ENOE mediante la variable `emp_ppal`. Esta categoría incluye no solo a quienes laboran en unidades económicas del sector informal, sino también a ocupados que, por las condiciones de su vínculo laboral o posición en la ocupación, se consideran laboralmente vulnerables bajo el criterio de informalidad laboral de INEGI.
- **Desempleo abierto** (`est_lab = 3`): individuos que pertenecen a la PEA, no están ocupados y se encuentran buscando activamente empleo.
- **Inactividad** (`est_lab = 4`): individuos que pertenecen a la PNEA, es decir, no están ocupados ni buscan activamente empleo.

Variables Demográficas y Geográficas

A partir del estado laboral y los filtros poblacionales, se construyen las covariables demográficas y geográficas que entran al modelo Logit como predictores de la probabilidad de formalización. En el plano demográfico, se construyen las siguientes variables binarias: `sexo` (1 = hombre, 0 = mujer), `jef_hog` (1 = jefe o jefa del hogar, 0 = otro parentesco) y `soltero` (1 = sin pareja estable; soltero, separado, divorciado o viudo, 0 = casado o en unión libre). Para la variable `edad` se considera la máxima registrada y se replica a todos los trimestres de la observación.

En el plano geográfico, se construye la variable `loc_urb` (1 = localidad urbana, 0 = localidad rural) y la variable de macrozona geográfica (`zona`) a partir de la entidad federativa de residencia, clasificando las observaciones en tres macroregiones: Norte, Centro y Sur.

En cuanto al sector de actividad económica, la variable `rama` clasifica el empleo principal según el sector productivo en ocho categorías: no aplica para inactivos y desocupados (`rama` = 0), construcción (`rama` = 1), industria manufacturera (`rama` = 2), comercio (`rama` = 3), servicios (`rama` = 4), otros sectores (`rama` = 5), agropecuario (`rama` = 6) y no especificado (`rama` = 7). Esta clasificación se utiliza como una aproximación general al entorno productivo del empleo principal, no como una desagregación detallada de todas las actividades económicas u ocupaciones. Su inclusión responde a que la literatura sobre informalidad laboral ha documentado que la propensión a la informalidad puede variar entre sectores productivos y tipos de empleo, por lo que resulta pertinente controlar por diferencias amplias en la estructura sectorial del empleo (Gasparini y Tornarolli, 2009).

Variables de Educación

La variable `años_edu` mide los años de educación acumulada del individuo y se construye a partir del grado máximo aprobado reportado en la SDEM (`cs_p13_1`, `cs_p13_2`), los años cursados dentro de ese grado y los requisitos de ingreso al nivel educativo correspondiente (`cs_p15`), cotejados con los criterios del sistema educativo mexicano.

Con fines exclusivamente descriptivos, se construyen además dos categorizaciones etarias que permiten caracterizar la composición de la muestra y presentar un panorama general de los egresados universitarios según su etapa de ciclo laboral. Estas variables no se incorporan como covariables en el modelo Logit ni en la estimación del PSM.

La primera categorización, `grupo_edad3`, agrupa a los individuos en tres grandes grupos de edad: jóvenes de 22 a 30 años, adultos de 31 a 45 años y senior de 46 a 65 años. La segunda, `grupo_edad5`, ofrece una desagregación más detallada en cinco cohortes laborales: inserción de 22 a 29 años, consolidación de 30 a 39 años, madurez plena de 40 a 49 años, pre-retiro de 50 a 59 años y transición al retiro de 60 a 65 años.

Variable de Resultado

La variable de resultado del análisis es el logaritmo natural del ingreso mensual por remuneraciones al trabajo registrado en el quinto trimestre, denotado `ln_ingm_t5`. El ingreso mensual base (`ing_men`) se construye a partir de `ingocup`, que registra el ingreso de la ocupación principal reportado en la ENOE, con los valores de cero convertidos a valores faltantes para excluirlos del cálculo del logaritmo.

La transformación logarítmica se justifica por dos razones complementarias. Desde el punto de vista estadístico, la distribución del ingreso laboral suele presentar una distribución asimétrica hacia la derecha, por lo que el uso del logaritmo permite aproximarla a una forma más simétrica, reduciendo la influencia de valores extremos sobre el estimador. Esta práctica es común en econometría aplicada cuando se trabaja con variables monetarias como salarios o ingresos (Gujarati y

Porter, 2009; Wooldridge, 2010). Desde el punto de vista interpretativo, bajo esta transformación el ATT estimado por el PSM se interpreta directamente como el efecto porcentual aproximado de la formalización sobre el logaritmo del ingreso mensual.

Es importante distinguir que la variable de resultado del PSM es el ingreso registrado en t_5 (el ingreso posterior al tratamiento), extraído como logaritmo del ingreso mensual en ese trimestre y propagado al registro de t_1 de cada observación. De esta manera, cuando en el paso siguiente se restringe la base a una sola fila por observación, ese registro ya lleva consigo el ingreso de cierre de t_5 . Esta operación garantiza que la base final del PSM contenga, en una única línea por `folio_ind`, todas las covariables pre-tratamiento junto con la variable de resultado post-tratamiento.

2.4. Definición de la Muestra de Análisis

Una vez construido el panel limpio con todas las variables descritas en la sección anterior, se procede a delimitar la muestra de análisis sobre la cual se estimará el efecto causal de la formalización. Este proceso sigue una lógica de cuatro pasos secuenciales:

- La identificación de los individuos elegibles.
- La asignación de la variable de tratamiento.
- La extracción de la variable de resultado.
- La depuración final de la muestra.

Al término de este proceso, la base de datos queda reducida a una sola observación por registro, correspondiente a su línea base en t_1 , con la información de resultado de t_5 ya incorporada. Esta es la base que alimenta directamente el modelo Logit y el algoritmo de emparejamiento del Capítulo 3.

Criterios de elegibilidad: informales registrados en t_1

El primer paso consiste en delimitar la población elegible para el análisis. Dado que el objetivo es estimar el efecto de los egresados que transitaron de la informalidad a la formalidad, es indispensable que todos los individuos de la muestra compartan el mismo punto de partida: estar ocupados en el sector informal al inicio del periodo de seguimiento. En consecuencia, se conservan únicamente los individuos que en el primer trimestre (t_1) presentan un estado laboral informal (`est_lab` = 2).

Se excluyen los individuos formales en t_1 (`est_lab` = 1) porque ya poseen el tratamiento de interés en la línea base y no tienen un contrafactual informal comparable dentro del diseño del estudio. Asimismo, se excluye a quienes están desempleados (`est_lab` = 3) o inactivos (`est_lab` = 4) en t_1 , pues su punto de partida es distinto al de la población de interés: la transición hacia el empleo formal desde el desempleo o la inactividad constituye un fenómeno diferente al de la

formalización de un trabajador ya ocupado, y requeriría un diseño de evaluación separado. Esta restricción garantiza que tanto el grupo tratado como el grupo de control compartan la misma condición laboral de partida, requisito fundamental para la validez del PSM.

Definición del tratamiento

Una vez restringida la muestra al sector informal en $t1$, se define la variable de tratamiento a partir del estado laboral observado en el quinto trimestre ($t5$), que representa el resultado del individuo i al término del periodo de seguimiento. Formalmente:

$$\text{tratamiento} = \begin{cases} 1 & \text{si } \text{est_lab} = 1 \text{ en } t5 \quad (\text{transitó a empleo formal}) \\ 0 & \text{si } \text{est_lab} = 2 \text{ en } t5 \quad (\text{permaneció en empleo informal}) \end{cases}$$

Esta definición del tratamiento aprovecha una propiedad central del diseño longitudinal de la ENOE: dado que las covariables del modelo logit se miden en $t1$ antes de que ocurra el tratamiento, el supuesto de independencia condicional (CIA) del PSM descansa sobre características genuinamente pre-tratamiento. Esto mejora la plausibilidad de la interpretación causal del estimador y reduce el riesgo de sesgo por causalidad inversa.

Variable de resultado: ingreso de cierre

La variable de resultado del PSM es el logaritmo del ingreso mensual registrado en el quinto trimestre ($\ln_ingm_t5$), cuya construcción se detalló en la Sección 2.3. La lógica de esta elección es directa: el ingreso de $t1$ refleja lo que el individuo ganaba siendo informal antes de cualquier transición, mientras que el ingreso de $t5$ refleja el efecto de lo ocurrido durante el periodo de seguimiento (haberse formalizado para los tratados, o haberse mantenido informal para los controles).

Depuración final

Con el tratamiento asignado y el ingreso de cierre disponible, se restringe la base de datos al trimestre 1 de cada registro, obteniendo una estructura de una sola fila por individuo. Cada registro contiene las covariables pre-tratamiento medidas en $t1$ (edad, sexo, escolaridad, zona, rama, entre otras), la variable de tratamiento (**tratamiento**) determinada por el estado laboral observado en $t5$, y el ingreso de cierre ($\ln_ingm_t5$) traído desde ese último trimestre.

Antes de cerrar la depuración, se aplica un filtro adicional: se eliminan los individuos que en $t1$ reportan cero horas trabajadas ($\text{hrs_ocup} = 0$). Estos corresponden a trabajadores clasificados como informales en $t1$ que simultáneamente declaran no haber trabajado ninguna hora en la semana de referencia, lo cual constituye una inconsistencia entre su clasificación de ocupación y su reporte de actividad laboral. Incluir estos casos en el Logit y en el emparejamiento introduciría ruido en la estimación del puntaje de propensión, pues su situación laboral en la línea base es ambigua.

Tabla 2.1: Construcción de variables para el modelo

Variable	Fuente (ENOE, SDEM)	Descripción y construcción
Identificadores y estructura del panel		
anio	Año calendario del panel	Variable numérica que identifica el año de la observación (2005–2025).
trimestre	n_ent (número de entrevista)	Variable categórica que indica el trimestre de observación dentro del periodo de levantamiento. Toma valores de 1 a 5; para 2025, el cuarto trimestre es reclasificado como trimestre 5.
folio_ind	cd_a , ent , con , v_sel , n_hog , h_mud , n_ren	Clave alfanumérica de seguimiento longitudinal, resultado de la concatenación de las variables de localización. Permite vincular al mismo individuo a través de los cinco trimestres del año.
fac	fac_tri (factor de expansión trimestral)	Factor de expansión de la ENOE tomado del trimestre 1 de cada individuo y propagado a todos sus trimestres.
Covariables del modelo logit		
sexo	sex (sexo del entrevistado)	Variable dicotómica: 1 si la persona es hombre, 0 si es mujer.
edad	eda (edad en años cumplidos)	Años de edad del individuo. Se utiliza la edad máxima observada entre los trimestres del panel anual para evitar exclusiones por fluctuaciones menores en el reporte.
anios_edu	cs_p13_1 (nivel de instrucción alcanzado), cs_p13_2 (años aprobados en ese nivel), cs_p15 (antecedente escolar requerido)	Variable continua de años de escolaridad acumulada, construida sumando al nivel educativo base los años aprobados en ese nivel. Se utiliza el valor máximo entre trimestres del individuo.
soltero	e_con (estado conyugal)	Variable dicotómica: 1 si la persona no tiene pareja estable (soltero, separado, divorciado, viudo o no especificado), 0 si está casado o en unión libre.
jef_hog	par_c (parentesco con el jefe del hogar)	Variable dicotómica: 1 si la persona es jefe o jefa del hogar, 0 en caso contrario.
loc_urb	ur (tipo de localidad)	Variable dicotómica: 1 si reside en zona urbana, 0 si reside en zona rural.

Continúa en la siguiente página

Tabla 2.1 – *Continuación*

Variable	Fuente (ENOE, SDEM)	Descripción y construcción
zona	ent (entidad federativa)	Variable categórica que clasifica la entidad de residencia en tres macroregiones: Norte (1), Centro (2) y Sur (3).
rama	rama_est2 (sector de actividad)	Variable categórica del sector de actividad económica del empleo principal: 0 no aplica (inactivos y desocupados), 1 construcción, 2 industria manufacturera, 3 comercio, 4 servicios, 5 otros sectores, 6 agropecuario, 7 no especificado.
Estado laboral y trayectoria		
est_lab	clase1 (PEA/PNEA), clase2 (ocupado/desocupado), emp_ppal (empleo formal/informal)	Variable categórica que clasifica la condición laboral: 1 empleo formal, 2 empleo informal, 3 desempleo, 4 inactividad. Es la variable central del diseño y el insumo de la definición del tratamiento.
Variable de tratamiento		
tratamiento	est_lab en t1 y en t5	Variable dicotómica construida sobre la submuestra de informales en t1: toma el valor de 1 si el individuo transitó de la informalidad a la formalidad entre t1 y t5, y 0 si permaneció informal en t5.
Variables de resultado		
ing_men	ingocup	Ingreso mensual por remuneraciones al trabajo.
ln_ingm	ingocup	Logaritmo natural del ingreso mensual (ing_men). Disponible para todos los trimestres del panel.
ln_ingm_t5	ingocup en t5	Logaritmo natural del ingreso mensual registrado en el quinto trimestre, extraído y propagado al registro del primer trimestre de cada individuo. Constituye la variable de resultado del PSM: refleja el ingreso post-tratamiento y permite estimar el efecto promedio de la formalización sobre los tratados (ATT).
Variables de uso descriptivo exclusivo		

Continúa en la siguiente página

Tabla 2.1 – Continuación		
Variable	Fuente (ENOE, SDEM)	Descripción y construcción
grupo_edad3	edad	Grupo etario de tres categorías: jóvenes (22–30 años), adultos (31–45 años) y senior (46–65 años). Uso exclusivamente descriptivo.
grupo_edad5	edad	Cohorte laboral de cinco categorías: inserción (22–29), consolidación (30–39), madurez plena (40–49), pre-retiro (50–59) y transición al retiro (60–65). Uso exclusivamente descriptivo.

Fuente: elaboración propia con datos de la ENOE (INEGI, 2005–2025).

2.5. Estadística Descriptiva

Como complemento a las figuras y tablas estadísticas de este capítulo, los resultados descriptivos de este trabajo se publicaron en una aplicación interactiva desarrollada en Streamlit, que permite explorar las cifras desagregadas por año y por covariables, así como reproducir de forma dinámica las matrices de transición, las distribuciones del panel y los diagnósticos del emparejamiento. La aplicación está organizada en páginas temáticas que se citan a lo largo de este capítulo y del capítulo 3.²

2.5.1. Evolución general del panel (2005–2025)

El panel de seguimiento laboral está constituido por egresados universitarios observados durante cinco trimestres consecutivos a lo largo del periodo 2005–2025. La unidad de análisis es el individuo en su cohorte anual, definida por el primer trimestre de observación (t_1) y seguida hasta el quinto trimestre (t_5), lo que permite capturar trayectorias laborales de corto plazo dentro de cada año. Todas las cifras descriptivas presentadas en esta sección se encuentran ponderadas por el factor de expansión, lo que garantiza un panorama de la población de egresados universitarios en México y no simplemente de la muestra encuestada. La evolución anual del panel y sus distribuciones pueden explorarse de forma interactiva en la página *Muestra Panel* de la aplicación.

El tamaño del universo representado experimentó un crecimiento sostenido a lo largo del periodo. En 2005, el panel representaba aproximadamente 920 mil egresados universitarios, cifra que se expandió progresivamente hasta superar los 2 millones en 2024 y 2025, reflejo del proceso de expansión de la educación superior en México. Este crecimiento no fue lineal: se registraron contracciones puntuales asociadas a choques socioeconómicos. La más pronunciada ocurrió en

²Aplicación interactiva del proyecto (Streamlit). <https://tesis-psm-formalidad-uaeh.streamlit.app/>.

2020, año en que la pandemia por COVID-19 interrumpió el levantamiento regular de la ENOE durante el segundo trimestre, lo que redujo la cobertura efectiva del panel y produjo una caída visible.

Más allá de su tamaño, la composición interna del panel revela una jerarquía de estados laborales notablemente estable. La ENOE permite clasificar a cada individuo durante el primer trimestre en uno de los cuatro estados mutuamente excluyentes: empleo formal, empleo informal, desempleo e inactividad. Esta taxonomía, estándar en la literatura sobre mercado de trabajo en países en desarrollo (Maloney, 2004; Perry et al. 2007), captura no solamente la condición de ocupación sino también la calidad del vínculo laboral, distinguiendo entre trabajadores con acceso a seguridad social (formales) y aquellos que operan al margen de estos esquemas de protección (informales). A lo largo del periodo, la distribución de estos estados se mantiene notablemente estable en términos cualitativos, con el empleo formal ocupando consistentemente la posición dominante. Considerando el primer trimestre de cada año, entre 2005 y 2025 aproximadamente el 61-65 % de los egresados universitarios se encontraban en empleo formal, en torno al 16-21 % en empleo informal, el 13-20 % en inactividad, y el 3-6 % en desempleo. Esta distribución, representada en la Figura 2.1a, es consistente a través de los cinco trimestres de seguimiento, lo que indica que la estructura laboral del panel conserva un patrón similar durante todo el ciclo observado.

La predominancia del empleo formal entre egresados universitarios no resulta sorprendente desde un punto de vista teórico. Diversos estudios han documentado que el nivel educativo constituye uno de los factores asociados a la formalidad laboral en México (Bosch y Maloney, 2010; Levy, 2008). La educación superior puede incrementar la productividad individual y funcionar como señal de capacidades para el empleador, lo que eleva la probabilidad de acceder a empleos formales, donde la cobertura de seguridad social constituye un rasgo definitorio de la relación laboral. No obstante, que alrededor de uno de cada cinco egresados universitarios se encuentre en la informalidad en t_1 revela que la certificación universitaria no es condición suficiente para garantizar la inserción formal. Este resultado puede interpretarse, por un lado, como evidencia de posibles restricciones de acceso al empleo formal, asociadas con la informalidad por exclusión o informalidad involuntaria (Fields, 1975; Levy y Székely, 2016). Por otro lado, también es consistente con la heterogeneidad del sector informal, en tanto una parte de la informalidad entre trabajadores calificados puede responder a decisiones voluntarias vinculadas con la independencia, el autoempleo o la valoración de atributos laborales distintos a los ofrecidos por el empleo formal (Maloney, 2004).

Dinámica de transición: matrices de movilidad laboral

El análisis de las distribuciones por condición laboral, si bien muestra un panorama estructural, oculta la dinámica de movilidad. Por ello, se construyeron matrices de transición que registran la probabilidad de que un individuo en el estado s en t_1 se encuentre en el estado s' en t_5 . Este enfoque es consistente con estudios que analizan la movilidad laboral mediante flujos brutos y procesos de transición entre empleo formal, empleo informal, desempleo e inactividad (Bosch y Maloney, 2008; Gallardo Del Ángel, 2019). Las matrices de transición completas, año por año a lo largo del periodo, pueden consultarse de forma interactiva en la página *Movilidad laboral* de

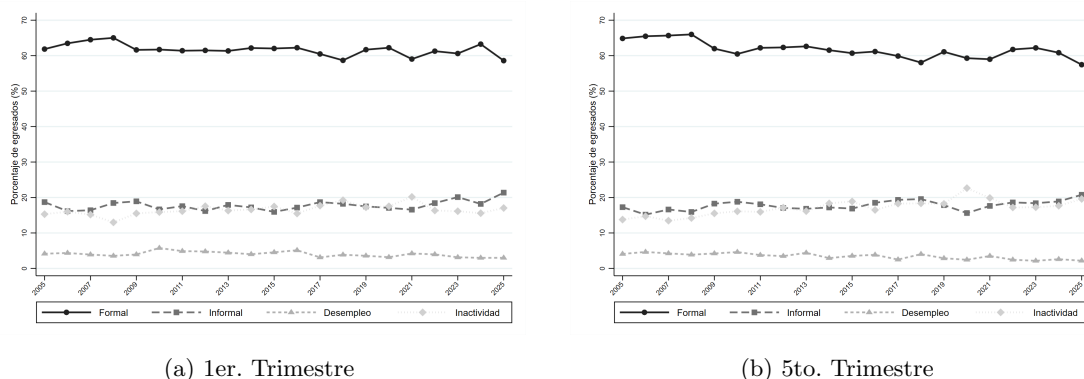


Figura 2.1: Evolución de la distribución laboral. Egresados universitarios, México 2005–2025.

Fuente: Elaboración propia con datos de la ENOE (INEGI, 2005–2025).

la aplicación.

Los resultados de estas matrices, promediados sobre el periodo completo, muestran que los estados laborales presentan distintos grados de persistencia. El empleo formal es el estado más estable: la probabilidad de permanecer formal en t_5 dado que se era formal en t_1 (p_{FF}), oscila entre 0.81 y 0.89 a lo largo del periodo. Este resultado es consistente con la evidencia de Bosch y Maloney (2010), quienes documentan mayores duraciones promedio en el empleo formal frente a otras formas de inserción laboral. Desde una perspectiva teórica, esta persistencia también puede interpretarse a la luz de los modelos de búsqueda y emparejamiento, en los cuales las transiciones laborales no son instantáneas, sino que dependen de fricciones de búsqueda, tasas de separación, disponibilidad de vacantes y decisiones de permanencia o destrucción del empleo (Mortensen y Pissarides, 1994).

El empleo informal, en contraste, muestra una persistencia menor que la observada en el empleo formal. La probabilidad de permanecer en la informalidad en t_5 dado que se era informal en t_1 (p_{II}), se sitúa, en promedio, en un rango de 0.43 a 0.54 durante el periodo analizado. Este resultado sugiere que, aproximadamente la mitad de quienes inician en un trabajo informal han cambiado de estado cinco trimestres después. Una fracción importante de esos movimientos corresponde a transiciones hacia la formalidad (p_{IF}), cuya probabilidad varía entre 0.31 y 0.40; mientras que la transición hacia la inactividad se ubica entre 0.08 y 0.20, y el paso hacia el desempleo presenta una probabilidad baja, entre 0.01 y 0.09. Esta movilidad es consistente con estudios para México que muestran que la informalidad laboral no constituye un estado homogéneo, sino que puede adoptar distintas formas y cumplir funciones diferenciadas dentro de las trayectorias laborales (Conover et al. 2022; Gallardo Del Ángel, 2019).

Si bien existe movilidad de la informalidad hacia la formalidad, en promedio esta resulta menos probable que la permanencia en la informalidad a lo largo de cinco trimestres. Esta diferencia puede interpretarse como evidencia compatible con la existencia de barreras a la formalización que van más allá de las características individuales observables, y que podrían estar relacionadas

con atributos del empleador, del sector de actividad o con la estructura de incentivos del sistema de seguridad social mexicano (Antón et al. 2013; Levy, 2008). Identificar si, y en qué magnitud, el acceso al empleo formal tiene un efecto causal sobre los ingresos requiere comparar individuos que transitaron a la formalidad con individuos observacionalmente similares que permanecieron informales durante el mismo periodo, lo que motiva el uso del emparejamiento por puntajes de propensión.

Permanencia y salida: dinámica por estado laboral

Un análisis adicional al de las matrices de transición lo ofrecen las estadísticas de permanencia y salida laboral en $t_1 \rightarrow t_5$ por año (Figuras 2.2a, 2.2b, 2.2c y 2.2d). Estas gráficas reportan, para cada año y estado inicial, qué proporción del panel total permaneció en ese estado al cabo de cinco trimestres y qué proporción lo abandonó.

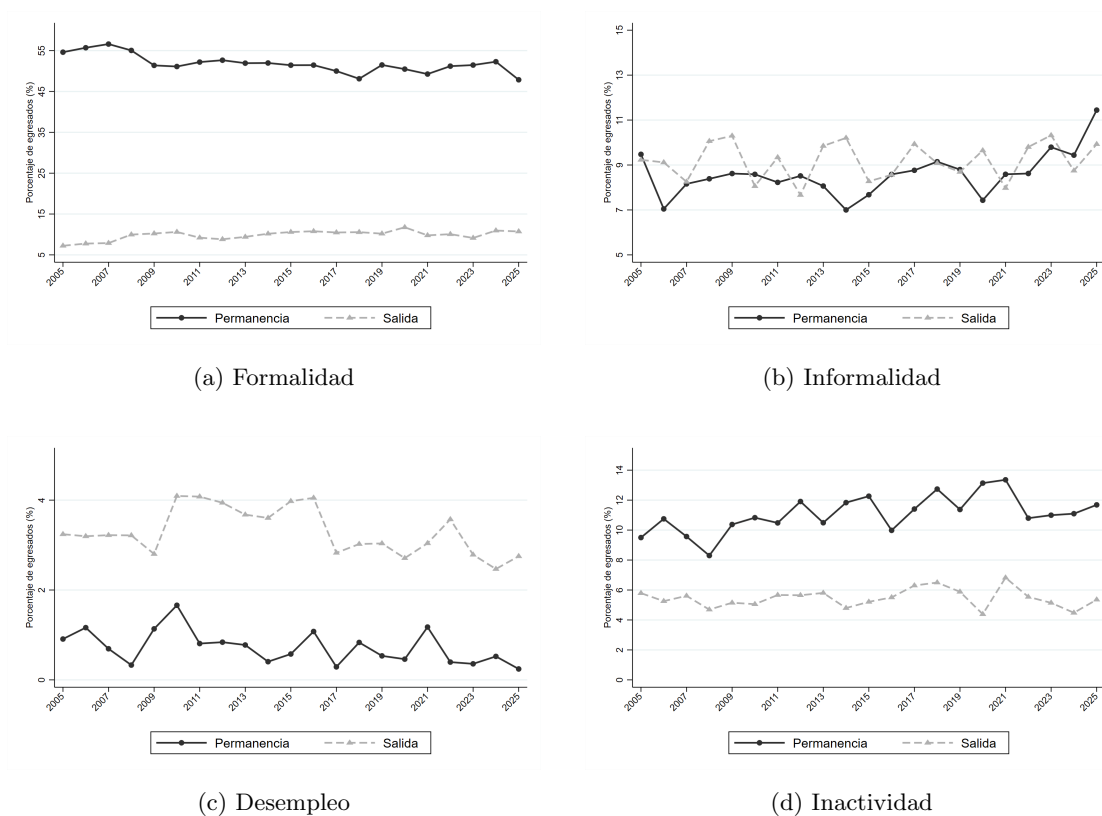


Figura 2.2: Flujos de permanencia y salida por estado laboral: trayectoria $t_1 \rightarrow t_5$. Egresados universitarios, México 2005–2025.

Fuente: Elaboración propia con datos de la ENOE (INEGI, 2005–2025).

Para el empleo formal (Figura 2.2a), la línea de estabilidad se mantiene consistentemente entre

el 47 % y el 56 % del total del panel, con una tendencia suavemente a la baja a lo largo del periodo: de valores cercanos al 55 % en 2005–2007 hacia valores de 47–52 % en 2017–2025. La inestabilidad formal, por su parte, se mantiene relativamente constante en torno al 7–11 %, con ligeras elevaciones alrededor de 2009 y 2020.

Para el empleo informal (Figura 2.2b), el contraste es llamativo: las líneas de estabilidad e inestabilidad se encuentran muy próximas entre sí a lo largo de todo el periodo, oscilando ambas en un rango de 7–12 % del total del panel. Esta proximidad refleja la alta rotación que caracteriza al sector informal: la probabilidad de salir de la informalidad es casi tan alta como la de permanecer en ella. Desde 2022, la línea de estabilidad informal muestra una tendencia creciente, alcanzando valores superiores al 11 % en 2025, lo que podría indicar una mayor consolidación de trabajadores calificados en el sector informal, posiblemente asociada a la expansión del trabajo por cuenta propia de alta calificación.

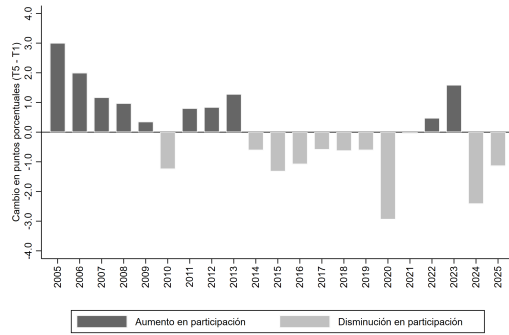
El desempleo, como era de esperarse por su carácter transitorio, exhibe una estructura muy diferente (Figura 2.2c): la estabilidad en desempleo es marginal, prácticamente inferior al 2 % del total de la población en cada año del periodo, mientras que la inestabilidad es la norma, con valores de 2.5–4 % del panel. Esto confirma que el desempleo entre egresados universitarios en México es fundamentalmente un estado de tránsito, no de larga duración, resultado coherente con la estructura del sistema de protección social mexicano, que no contempla seguro de desempleo universal y por tanto genera fuertes incentivos para salir del desempleo rápidamente, incluso hacia la informalidad o la inactividad.

La inactividad (Figura 2.2d) muestra la segunda mayor estabilidad después del empleo formal, con valores que oscilan entre el 8 % y el 13 % del panel y una tendencia levemente ascendente a partir de 2012. Este incremento en la inactividad estable podría reflejar el aumento en la participación femenina diferida (mujeres con educación superior que posponen su ingreso al mercado por responsabilidades de cuidado) o bien el crecimiento de la población en edad de pre-retiro dentro del panel.

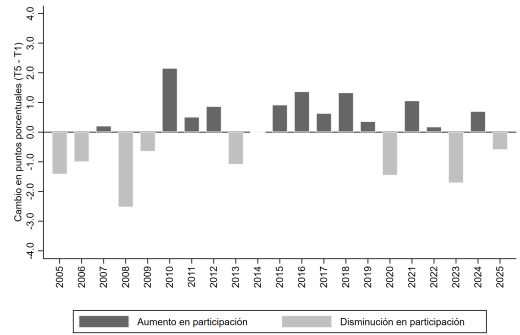
Las gráficas de cambio neto $t1 \rightarrow t5$ por estado (Figura 2.3) complementan este análisis con una perspectiva diferencial: muestran el signo y magnitud del cambio en la participación de cada estado entre $t1$ y $t5$. Para el empleo formal, la variación fue positiva durante la mayor parte de 2005–2013, lo que indica que al cabo de cinco trimestres el panel ganaba trabajadores formales. A partir de 2014 la variación se vuelve negativa con mayor frecuencia, sugiriendo que el empleo formal pierde participación relativa dentro de cada cohorte durante el año de seguimiento. Para la inactividad, la variación encontrada en 2020 es la más alta de toda la serie (+5 puntos porcentuales), testimonio directo del impacto que la pandemia tuvo sobre la participación laboral.

2.5.2. Perfil sociodemográfico de la muestra de análisis

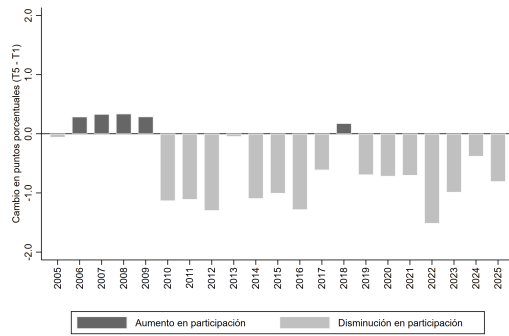
Esta sección caracteriza al universo de egresados universitarios del panel a lo largo de dimensiones sociodemográficas que la literatura ha identificado como factores asociados con la inserción en el empleo formal: sexo, edad, escolaridad, región de residencia, jefatura de hogar y estado civil (Bosch y Maloney, 2010; Esquivel Hernández, 2008; Levy, 2008; Maloney, 2004). Las dis-



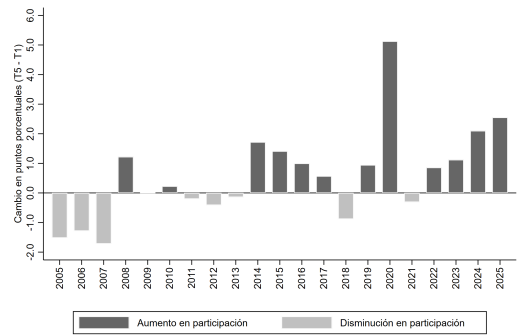
(a) Formalidad



(b) Informalidad



(c) Desempleo



(d) Inactividad

Figura 2.3: Variación neta en puntos porcentuales por estado laboral ($t1 \rightarrow t5$). Egresados universitarios, México 2005–2025.

Fuente: Elaboración propia con datos de la ENOE (INEGI, 2005–2025).

tribuciones anuales de cada una de estas covariables, desagregadas por estado laboral, pueden consultarse de forma interactiva en la página *Muestra Panel* de la aplicación.

Sexo

La composición por sexo del panel refleja una de las transformaciones más significativas del sistema educativo y del mercado laboral mexicano en las últimas dos décadas: la feminización de la educación superior. Como se aprecia en la Figura 2.4, en 2005 los hombres representaban aproximadamente el 56 % del panel, mientras que las mujeres constituían el 44 %. A partir de 2013-2014 esta brecha se cierra y, desde 2016 en adelante, las mujeres son mayoría, alcanzando valores superiores al 53-55 % en los años más recientes.

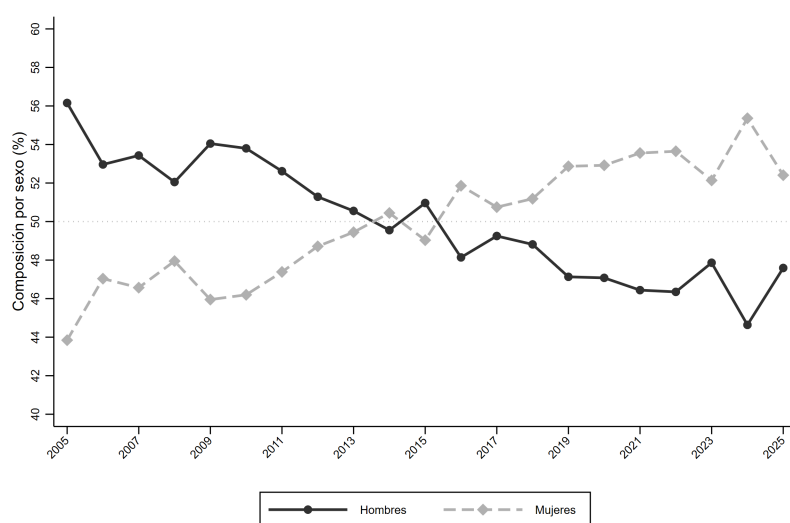


Figura 2.4: Evolución de la Composición por sexo, Trimestre 1 - Egresados universitarios, México 2005–2025.

Fuente: Elaboración propia con datos de la ENOE (INEGI, 2005–2025).

Edad y cohorte laboral

La distribución etaria del panel revela una muestra notablemente diversa en términos de ciclo de vida laboral. La Figura 2.5a muestra que, a lo largo de todo el periodo, el grupo adulto (31-45 años) es el más representado, oscilando entre el 35 % y el 39 % del panel. Le sigue el grupo senior (46-65 años), con una participación que creció de aproximadamente el 26 % en 2005 hasta 28-29 % en los años recientes. El grupo joven (22-30 años) es consistentemente el menos representado, con valores de alrededor del 27-28 % al inicio del periodo y una tendencia a la baja hacia el 25-26 % al final.

La Figura 2.5b introduce una distinción relevante: la edad promedio varía sistemáticamente entre estados laborales. Los trabajadores formales presentan una edad promedio cercana a los 40 años, relativamente estable a lo largo del periodo. Los informales son algo más jóvenes, con promedios de entre 36 y 38 años. El grupo más joven en promedio es el de desempleados, que se sitúa consistentemente por debajo de los 34 años, lo que refleja la concentración del desempleo friccional en las etapas de inserción laboral. Los inactivos, en contraste, muestran una edad promedio creciente a lo largo del periodo, sugiriendo que la inactividad en edades avanzadas (posiblemente pre-retiro) gana peso relativo dentro de este estado.

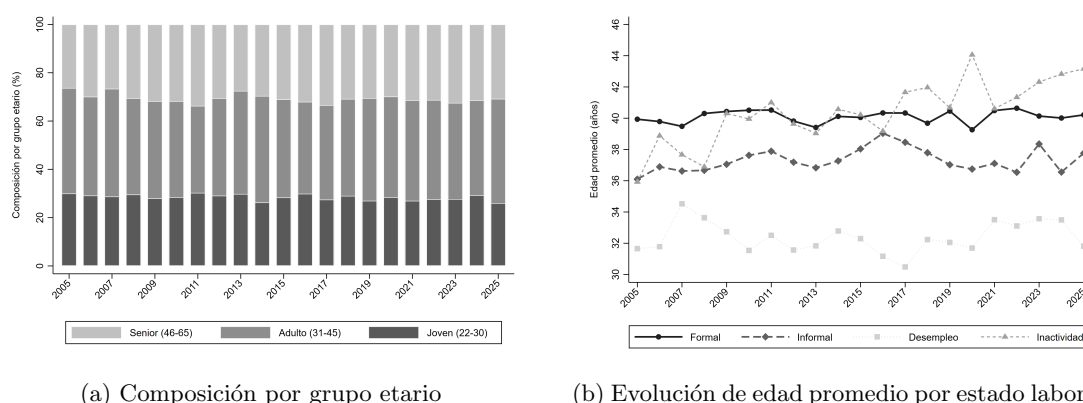


Figura 2.5: Composición y edad promedio por grupo y estado laboral, Trimestre 1. Egresados universitarios, México 2005–2025.

Fuente: Elaboración propia con datos de la ENOE (INEGI, 2005–2025).

La Figura 2.6 desagrega esta heterogeneidad etaria utilizando una clasificación de cinco cohortes laborales: inserción (22-29), consolidación (30-39), madurez (40-49), pre-retiro (50-59) y transición al retiro (60-65). Los resultados muestran que la distribución entre estados laborales es marcadamente heterogénea según cohorte. El desempleo se concentra de manera pronunciada en la cohorte de inserción, que representa alrededor del 55 % de los desempleados, lo que sugiere un componente friccional asociado a las primeras etapas de la trayectoria laboral. La informalidad, por su parte, también presenta una mayor concentración en las cohortes jóvenes, especialmente en inserción y consolidación, que en conjunto agrupan la mayor parte de los egresados informales. Sin embargo, su presencia no desaparece en las etapas posteriores: la cohorte de madurez conserva una participación relevante, aunque menor, mientras que pre-retiro y transición al retiro muestran proporciones más reducidas. Esto indica que la informalidad entre egresados universitarios no debe interpretarse únicamente como un fenómeno de entrada al mercado laboral, sino también como una condición que puede persistir durante etapas intermedias de la carrera laboral.

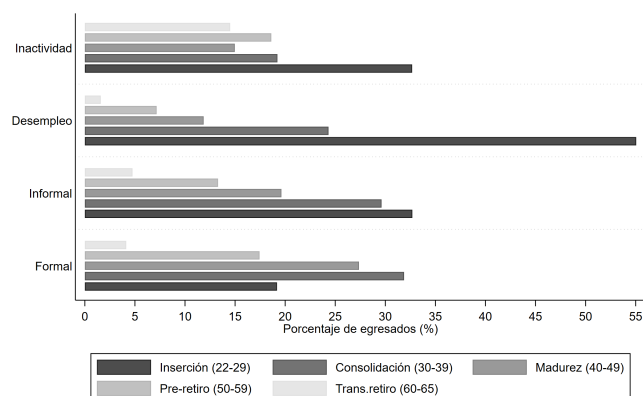


Figura 2.6: Distribución por etapa de vida laboral por estado, Trimestre 1. Egresados universitarios, México 2005–2025.

Fuente: Elaboración propia con datos de la ENOE (INEGI, 2005–2025).

Escolaridad

Dado que la muestra está restringida a egresados universitarios, la variación en escolaridad es acotada, aunque no trivial. La Figura 2.7 muestra que la escolaridad promedio oscila entre 16.5 y 17 años para todos los estados laborales, lo que corresponde aproximadamente a estudios de licenciatura completa y, en algunos casos, a formación adicional. Los trabajadores formales presentan de manera consistente la mayor escolaridad promedio, en torno a 16.6 – 17 años, mientras que los demás estados se agrupan en valores ligeramente inferiores, alrededor de 16.5 – 16.7 años.

Aunque la brecha es modesta en términos absolutos, de apenas 0.2 – 0.4 años, su persistencia a lo largo del periodo sugiere que incluso dentro de una muestra compuesta por egresados universitarios existen diferencias educativas asociadas con la inserción formal. Este patrón es consistente con la teoría de capital humano,³ según la cual la educación puede incrementar la productividad esperada y mejorar las oportunidades laborales. Asimismo, es coherente con la evidencia para México que vincula la educación con la movilidad social y el bienestar económico (De Hoyos et al. 2010), así como con estudios que identifican la escolaridad como un factor asociado positivamente con la probabilidad de trabajar en el sector formal (Esquivel Hernández, 2008).

Distribución Regional

La distribución regional del panel (Figura 2.8) muestra que la región Centro concentra consistentemente entre el 55 % y el 59 % de los egresados universitarios observados, seguida por la región

³La teoría del capital humano plantea que la educación y la formación incrementan las capacidades productivas de los individuos, lo que puede traducirse en mejores oportunidades laborales e ingresos más altos (Becker, 1964; Mincer, 1974).

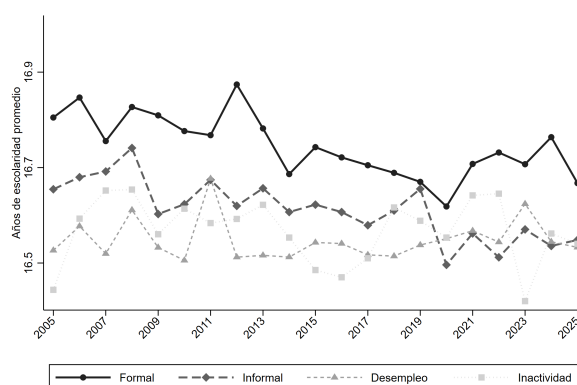


Figura 2.7: Evolución de la escolaridad promedio por estado laboral, Trimestre 1. Egresados universitarios, México 2005–2025.

Fuente: Elaboración propia con datos de la ENOE (INEGI, 2005–2025).

Norte, con una participación aproximada de 25–30 %, y la región Sur, con valores cercanos a 15–20 %. Esta distribución se mantiene relativamente estable a lo largo del periodo y refleja la mayor concentración de la muestra en entidades del centro del país.

La región de residencia es relevante para el análisis porque las condiciones de inserción laboral pueden diferir territorialmente. La literatura para México ha documentado diferencias regionales en la estructura del mercado laboral, particularmente en regiones vinculadas con actividades manufactureras y de exportación, como la frontera norte (Mendoza, 2010). Asimismo, estudios basados en la ENOE muestran diferencias regionales en el mercado laboral formal e informal, tanto en salarios como en composición sectorial (Rodríguez Pérez et al. 2019). Por ello, la región se incorpora como covariable de control en la estimación del puntaje de propensión con el objetivo de capturar diferencias regionales asociadas con la estructura productiva y las condiciones locales del mercado laboral.

Jefatura del hogar y estado civil

La Figura 2.9 muestra dos características del contexto familiar de los egresados universitarios del panel: la jefatura de hogar y la situación conyugal. La proporción de jefes de hogar presenta una tendencia decreciente a lo largo del periodo. En 2005, aproximadamente el 42 % de los egresados observados eran jefes o jefas de su hogar, mientras que en los años recientes la proporción disminuye a valores cercanos a 33 – 36 % entre 2023 y 2025. Este comportamiento puede estar relacionado con cambios en la composición del panel, particularmente con una mayor participación femenina, dado que las mujeres suelen registrar menores tasas de jefatura de hogar en la ENOE.

Por su parte, la proporción de individuos con pareja se mantiene relativamente estable hasta 2015, en torno al 60 %, pero muestra un descenso posterior hacia niveles cercanos a 50 – 55 % en los

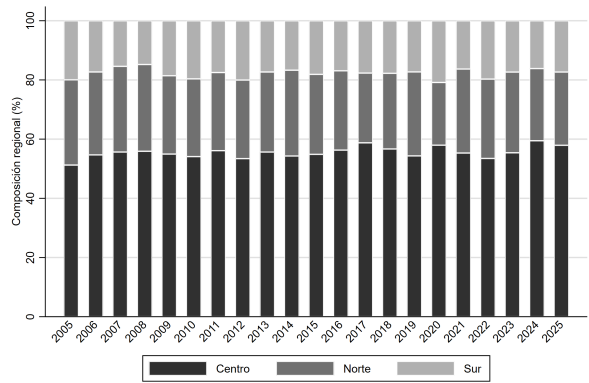


Figura 2.8: Evolución de la composición regional, Trimestre 1. Egresados universitarios, México 2005–2025.

Fuente: Elaboración propia con datos de la ENOE (INEGI, 2005–2025).

años recientes. En conjunto, ambas variables permiten contextualizar las condiciones familiares bajo las cuales los egresados participan en el mercado laboral. La jefatura de hogar identifica a quienes encabezan su hogar y puede estar asociada con mayores responsabilidades dentro de éste, mientras que la situación conyugal permite distinguir arreglos familiares distintos. Por ello, ambas variables se incorporan como controles sociodemográficos en el análisis.

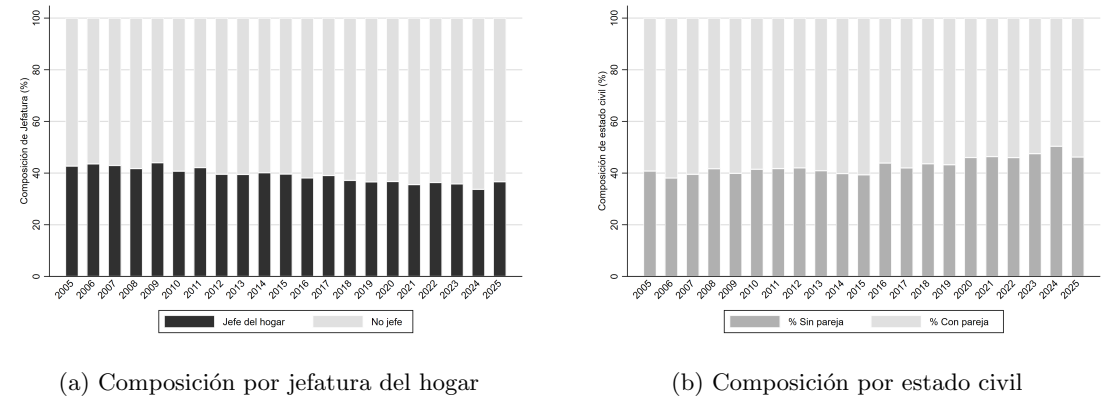


Figura 2.9: Evolución de la composición por jefatura del hogar y estado civil, Trimestre 1. Egresados universitarios, México 2005–2025.

Fuente: Elaboración propia con datos de la ENOE (INEGI, 2005–2025).

2.5.3. Transición a la muestra de análisis

El análisis descriptivo desarrollado en las secciones precedentes caracteriza el universo amplio de egresados universitarios observados en el panel de la ENOE a lo largo del periodo 2005–2025. A partir de este punto, la atención se desplaza hacia la subpoblación sobre la que se estimará el efecto causal de la formalización: los individuos que al inicio del periodo de seguimiento se encontraban ocupados en el sector informal (`est_lab` = 2) y que, al cabo de cinco trimestres, o bien transitaron al empleo formal (grupo de tratamiento) o bien permanecieron en la informalidad (grupo de control). Esta restricción define la muestra de análisis del emparejamiento por puntajes de propensión.

La construcción de la muestra de análisis sigue una lógica de tres filtros aplicados en cascada sobre el panel de egresados universitarios.

El primero es el filtro de estado laboral en t_1 : se conservan únicamente los individuos que en el primer trimestre reportan empleo informal (`est_lab` = 2), excluyendo a formales, desempleados e inactivos por las razones detalladas en la Sección 2.4. Aplicando este filtro, el universo de informales en t_1 asciende a 16,637 observaciones.

El segundo es la asignación del tratamiento a partir del estado en t_5 : quienes transitaron a la formalidad reciben `tratamiento` = 1 y quienes permanecieron en la informalidad reciben `tratamiento` = 0. Los individuos que en t_5 se encontraban desempleados o inactivos son excluidos, pues su trayectoria responde a una dinámica distinta a la de la permanencia o transición dentro del mercado laboral ocupado.

El tercero es el filtro de ingreso: se conservan aquellas observaciones que registran ingreso mensual por remuneraciones al trabajo en el quinto trimestre (`ing_men` $\neq$ `.`), que constituye la variable de resultado del PSM. Esta restricción es la que define operativamente la muestra efectiva del PSM: los individuos para quienes es posible estimar el efecto de la formalización sobre el ingreso. Aplicando los tres filtros, la muestra efectiva conserva 10,668 observaciones, de las cuales 4,338 corresponden al grupo tratado (40.7 %) y 6,330 al grupo de control (59.3 %).

La diferencia entre el universo de informales en t_1 (16,637) y la muestra efectiva del PSM (10,668) representa una pérdida neta de 5,969 observaciones, equivalente al 35 % del universo inicial. Este desgaste no es uniforme a lo largo del periodo: oscila entre 24.1 % en 2005 y 45.7 % en 2025, con una tendencia creciente sostenida a partir de 2010 que refleja, en parte, el incremento progresivo en la prevalencia de ocupaciones sin remuneración monetaria registrada (trabajo familiar no remunerado, autoempleo de subsistencia) entre los egresados universitarios informales en los años más recientes.

En esta investigación, el PSM no busca proyectar el efecto estimado sobre el universo completo de egresados informales en t_1 , sino identificar un efecto causal dentro de la muestra efectiva: individuos para quienes se dispone de covariables pretratamiento, condición de tratamiento claramente definida e ingreso observado en t_5 . En este sentido, el ATT estimado debe interpretarse como el efecto promedio de la formalización sobre los egresados inicialmente informales que transitaron a la formalidad, cuentan con ingreso observado al cierre del panel y se encuentran dentro de la región de soporte común. Esta delimitación es estándar en la literatura de evaluación de impac-

to, donde la validez del estimador depende de la disponibilidad de información pretratamiento, la comparabilidad entre tratados y controles y el cumplimiento del supuesto de soporte común (Caliendo y Kopeinig, 2008; Gertler et al. 2016; Heinrich et al. 2011; Imbens y Wooldridge, 2009).

2.5.4. Caracterización de la muestra de análisis

La muestra efectiva del PSM (10,668 observaciones) es el insumo de dos etapas metodológicamente distintas que requieren representaciones estadísticas también distintas, y por ello se caracteriza bajo dos visiones complementarias.

En la primera etapa, la estimación del modelo Logit que calcula el puntaje de propensión se realiza incorporando el factor de expansión de la ENOE como ponderador. Esta decisión, justificada en la Sección 2.1, garantiza que las probabilidades predichas de formalización reflejen la estructura del mercado laboral de egresados universitarios en México y no solo la composición particular de la muestra observada. En la segunda etapa, el emparejamiento propiamente dicho, se realiza sin ponderar, sobre las 10,668 observaciones brutas. En consecuencia, la muestra se presenta primero bajo una visión poblacional (ponderada) y luego bajo la visión observada (sin ponderar) que percibe el algoritmo de emparejamiento.

Visión poblacional: muestra expandida

La tabla 2.2 presenta las estadísticas descriptivas de la muestra efectiva del PSM ponderadas por el factor de expansión, desagregadas entre el grupo de control y el grupo tratado. El universo representado asciende a 2,513,887 egresados universitarios que se encontraban en el sector informal en *t1* y registraron ingreso en *t5* a lo largo del periodo 2005–2025: 1,513,167 permanecieron en la informalidad (controles, 60.2 %) y 1,000,720 transitaron al empleo formal (tratados, 39.8 %).

Composición demográfica y geográfica

La distribución por sexo muestra que los hombres constituyen el 55.3 % de la muestra total, con una diferencia de 3.2 puntos porcentuales entre controles (54.0 %) y tratados (57.2 %). Aproximadamente la mitad de la muestra no tiene pareja estable (48.3 %), sin diferencia apreciable entre grupos. La proporción de jefes o jefas de hogar alcanza el 39.6 % en total, ligeramente mayor entre tratados (40.6 %) que entre controles (39.0 %). La localidad de residencia exhibe una de las diferencias más notorias entre grupos antes del emparejamiento: el 58.9 % de los tratados reside en localidad urbana, frente al 49.9 % de los controles, una brecha de 9.0 puntos porcentuales que anticipa la mayor disponibilidad de empleos formales en mercados laborales urbanos. En términos regionales, la región Centro concentra a poco más de la mitad de la muestra (51.2 %), seguida por la región Sur (25.7 %) y la región Norte (23.1 %); los tratados están relativamente más concentrados en las regiones Norte (25.4 % frente a 21.7 %) y Centro (53.0 % frente a 49.9 %), mientras que los controles muestran mayor presencia en la región Sur (28.4 % frente a 21.6 %).

Sector de actividad económica

El sector servicios domina la composición sectorial de ambos grupos, concentrando al 65.6 % de la muestra total (64.6 % controles; 67.1 % tratados). El comercio es el segundo sector en importancia, aunque con mayor presencia entre controles (19.6 %) que entre tratados (15.8 %). La industria manufacturera agrupa al 7.1 % de la muestra, con mayor participación relativa entre tratados (8.0 % frente a 6.4 %), coherente con el mayor grado de formalización contractual en ese sector. El sector agropecuario y la construcción representan conjuntamente menos del 9 % de la muestra, con distribuciones similares entre grupos.

Distribución etaria y de escolaridad

La distribución por grupo etario muestra que el grupo adulto (31–45 años) es el más representado (41.5 %), seguido por el grupo joven (35.5 %) y el senior (23.0 %), con diferencias modestas entre controles y tratados. La edad promedio es prácticamente idéntica: 37.24 años para controles y 36.82 para tratados. Desagregada en cinco cohortes laborales, la cohorte de inserción (22–29 años) es la más numerosa (32.1 %), seguida por la de consolidación (30–39 años, 31.1 %). La cohorte de pre-retiro (50–59 años) representa el 12.7 % de la muestra, mientras que la de transición al retiro (60–65 años) apenas alcanza el 3.5 %, lo que sugiere que la informalidad persiste a lo largo de todo el ciclo de vida laboral entre egresados universitarios, incluidas las etapas avanzadas de la carrera. La escolaridad promedio es de 16.60 años para el total, con una diferencia mínima entre controles (16.59) y tratados (16.62), reflejo de que ambos grupos comparten por construcción el mismo umbral de elegibilidad de 16 años de educación acumulada.

Ingreso mensual: punto de partida y resultado

El logaritmo del ingreso mensual en t_1 , el ingreso informal previo al tratamiento, es de 8.56 para controles y 8.78 para tratados, una diferencia de 0.22 unidades logarítmicas que equivale aproximadamente a un 25 % de mayor ingreso informal inicial para el grupo tratado. Este diferencial sugiere que los egresados que eventualmente se formalizan no parten de una situación de precariedad absoluta en la informalidad, sino que ya muestran condiciones laborales relativamente superiores antes de la transición. El ingreso en t_5 , la variable de resultado del PSM, amplía esta brecha: el logaritmo del ingreso mensual post-tratamiento es de 8.63 para controles y 9.03 para tratados, una diferencia de 0.40 unidades logarítmicas equivalente a aproximadamente un 49 % de mayor ingreso para quienes se formalizaron. Esta diferencia cruda, sin ajuste por covariables ni emparejamiento, no puede interpretarse causalmente dado el desbalance inicial entre grupos; es precisamente la motivación del PSM desarrollado en el Capítulo 3. Las estadísticas descriptivas anuales desagregadas por covariable pueden consultarse de forma interactiva en la página *Muestra PSM* de la aplicación.

Balance pre-emparejamiento: muestra observada

La caracterización anterior refleja la estructura poblacional de la muestra efectiva una vez aplicado el factor de expansión. Esta visión presenta la misma muestra sin ponderar: las 10,668 observaciones brutas sobre las que operará el algoritmo de emparejamiento del PSM. El propósito no es retratar el fenómeno a escala nacional, sino documentar el balance inicial entre grupos tal como lo percibe el algoritmo: las diferencias observables entre los 4,338 tratados y los 6,330 controles que determinan la necesidad y la magnitud del ajuste por puntaje de propensión.

La tabla 2.3 presenta las características de balance pre-emparejamiento, incluyendo las medias no ponderadas de cada covariable para ambos grupos, la diferencia en puntos porcentuales o en unidades naturales, el sesgo estandarizado y el p -valor de la prueba t sobre el indicador de tratamiento, metodología idéntica a la empleada por el comando `pstest` de Leuven y Sianesi (2003).

Balance de variables binarias.

Entre las variables binarias, la localidad de residencia exhibe el mayor desbalance: el 77.8 % de los tratados reside en zona urbana frente al 71.2 % de los controles, generando un sesgo estandarizado de +15,3 %. Esta brecha refleja la concentración geográfica del empleo formal en mercados laborales urbanos y constituye una de las fuentes de sesgo de selección más relevantes a corregir mediante el emparejamiento. La composición por sexo muestra una diferencia modesta pero estadísticamente significativa: los hombres representan el 55.9 % del grupo tratado frente al 52.6 % del grupo de control (sesgo estandarizado +6,6 %, $p = 0,001$). La jefatura del hogar y el estado civil no presentan diferencias estadísticamente significativas entre grupos (sesgos de +3,2 % y -0,6 % respectivamente, $p > 0,10$). En términos regionales, la región Norte está ligeramente sobrerrepresentada entre los tratados (33.5 % frente a 30.4 %, sesgo +6,7 %, $p = 0,001$), mientras que la región Sur está sobre representada (29.1 % frente a 33.3 %, sesgo -9,0 %, $p < 0,001$). Por sector de actividad, el comercio presenta el mayor desbalance sectorial (sesgo -9,4 %, $p < 0,001$): los controles tienen mayor presencia en comercio (19.1 %) que los tratados (15.5 %), lo que refleja la naturaleza predominantemente informal de las relaciones laborales en el pequeño comercio, mientras que el sector servicios muestra el comportamiento opuesto, con mayor presencia entre tratados (68.9 % frente a 66.0 %, sesgo +6,3 %, $p = 0,001$).

Balance de variables continuas

Las variables continuas presentan diferencias estadísticamente significativas en los tres casos. La edad promedio es ligeramente menor entre tratados (36.31 años) que entre controles (36.87 años), con un sesgo estandarizado de -5,3 % ($p = 0,007$). Los años de escolaridad presentan un desbalance de +9,4 % ($p < 0,001$): los tratados acumulan en promedio 0.065 años adicionales de educación, diferencia que, aunque pequeña en términos absolutos, es sistemática y estadísticamente robusta. La variable de resultado, el logaritmo del ingreso mensual en t_5 , muestra el mayor sesgo estandarizado de toda la tabla: +56,1 % ($p < 0,001$), con una diferencia de 0.391 unidades logarítmicas entre tratados (9.01) y controles (8.61). Este desbalance es esperado y, de hecho,

Tabla 2.2: Estadísticas descriptivas de la muestra efectiva del PSM ponderadas (2005–2025).

Variable	Total	Control	Tratados
	N = 2,513,887	N = 1,513,167	N = 1,000,720
<i>Panel A: Variables binarias - % (desviación estándar)</i>			
Hombre	55.3 (49.7)	54.0 (49.8)	57.2 (49.5)
Sin pareja estable	48.3 (50.0)	48.3 (50.0)	48.4 (50.0)
Jefe(a) de hogar	39.6 (48.9)	39.0 (48.8)	40.6 (49.1)
Localidad urbana	53.5 (49.9)	49.9 (50.0)	58.9 (49.2)
<i>Zona geográfica</i>			
Norte	23.1 (42.2)	21.7 (41.2)	25.4 (43.5)
Centro	51.2 (50.0)	49.9 (50.0)	53.0 (49.9)
Sur	25.7 (43.7)	28.4 (45.1)	21.6 (41.2)
<i>Sector de actividad económica</i>			
Construcción	4.8 (21.5)	4.8 (21.5)	4.9 (21.5)
Industria manufacturera	7.1 (25.6)	6.4 (24.5)	8.0 (27.2)
Comercio	18.1 (38.5)	19.6 (39.7)	15.8 (36.5)
Servicios	65.6 (47.5)	64.6 (47.8)	67.1 (47.0)
Otros sectores	0.4 (6.2)	0.2 (4.2)	0.7 (8.4)
Agropecuario	3.5 (18.4)	3.8 (19.2)	3.0 (17.0)
No especificado	0.5 (7.3)	0.5 (7.3)	0.5 (7.3)
<i>Grupo etario laboral (3 categorías)</i>			
Joven (22–30 años)	35.5 (47.8)	34.7 (47.6)	36.7 (48.2)
Adulto (31–45 años)	41.5 (49.3)	42.6 (49.5)	39.9 (49.0)
Senior (46–65 años)	23.0 (42.1)	22.7 (41.9)	23.5 (42.4)
<i>Cohorte laboral (5 grupos)</i>			
Inserción (22–29)	32.1 (46.7)	31.3 (46.4)	33.2 (47.1)
Consolidación (30–39)	31.1 (46.3)	31.6 (46.5)	30.4 (46.0)
Madurez plena (40–49)	20.6 (40.5)	20.5 (40.4)	20.8 (40.6)
Pre-retiro (50–59)	12.7 (33.3)	12.8 (33.4)	12.6 (33.2)
Transición al retiro (60–65)	3.5 (18.4)	3.8 (19.1)	3.0 (17.1)
<i>Panel B: Variables continuas — media (desviación estándar)</i>			
Edad (años)	37.07 (10.81)	37.24 (10.86)	36.82 (10.71)
Años de educación	16.60 (0.69)	16.59 (0.68)	16.62 (0.70)
ln(Ingreso mensual) en t1	8.64 (0.80)	8.56 (0.80)	8.78 (0.77)
ln(Ingreso mensual) en t5 [†]	8.79 (0.74)	8.63 (0.75)	9.03 (0.65)

Notas: Estadísticas ponderadas por el factor de expansión trimestral de la ENOE.

Fuente: elaboración propia con datos de la ENOE (INEGI, 2005–2025).

constituye el objeto de interés del análisis: no corresponde a una covariable que deba balancearse en el modelo Logit, ya que es la variable de resultado, medida en t_5 , posterior al tratamiento, sino que su magnitud ilustra con claridad la diferencia cruda en ingreso entre quienes se formalizaron y quienes permanecieron informales antes de cualquier ajuste por características observables.

Síntesis: motivación del emparejamiento

El panorama que emerge de la tabla de balance pre-emparejamiento confirma la necesidad del PSM. Si bien la mayoría de los sesgos estandarizados se encuentran por debajo o cercanos a un umbral de $\pm 5\%$, la localidad urbana lo supera, mientras que otras covariables como región Sur, comercio, escolaridad, se aproximan a él, con diferencias estadísticamente significativas. En ausencia de corrección, una comparación directa de medias de ingreso entre tratados y controles mezclaría el efecto causal de la formalización con el efecto de estas diferencias de composición, produciendo un estimador sesgado. El emparejamiento por puntajes de propensión, estimando la probabilidad de tratamiento como función de estas covariables pre-tratamiento y emparejando a cada individuo tratado con su contraparte más similar entre los controles, permite aislar el efecto atribuible a la formalización del efecto de las características observables que diferencian a ambos grupos. Este es el procedimiento que se desarrolla en detalle en el Capítulo 3. Las tablas de balance anuales desagregadas por covariable demográfica y sectorial pueden consultarse de forma interactiva en la página *PSM Balance* de la aplicación; los resultados del emparejamiento, la estimación del modelo Logit y el ATT se presentan en las páginas *Muestra PSM Emparejada*, *Modelo Logit* y *PSM Resultados*, citadas en la siguiente sección.

Tabla 2.3: Tabla de balance pre-matching: comparación de covariables entre grupos (2005–2025). Muestra efectiva del PSM sin ponderar ($N = 10,668$; Control = 6,330; Tratados = 4,338).

Variable	Control N = 6,330	Tratados N = 4,338	Dif.	Sesgo std. (%)	p -valor
<i>Panel A: Variables binarias — media no ponderada (%)</i>					
Hombre	52.6 %	55.9 %	+3.3 pp	+6.6	0.001***
Sin pareja estable	49.2 %	48.9 %	-0.3 pp	-0.6	0.760
Jefe(a) de hogar	38.4 %	39.9 %	+1.5 pp	+3.2	0.110
Localidad urbana	71.2 %	77.8 %	+6.7 pp	+15.3	0.000***
<i>Zona geográfica</i>					
Norte	30.4 %	33.5 %	+3.1 pp	+6.7	0.001***
Centro	36.3 %	37.4 %	+1.0 pp	+2.1	0.277
Sur	33.3 %	29.1 %	-4.2 pp	-9.0	0.000***
<i>Sector de actividad económica</i>					
Construcción	4.5 %	5.5 %	+0.9 pp	+4.3	0.028**
Industria manufacturera	6.8 %	6.8 %	-0.0 pp	-0.0	0.998
Comercio	19.1 %	15.5 %	-3.6 pp	-9.4	0.000***
Servicios	66.0 %	68.9 %	+2.9 pp	+6.3	0.001***
Otros sectores	0.2 %	0.6 %	+0.4 pp	+6.6	0.000***
Agropecuario	2.9 %	2.3 %	-0.6 pp	-3.6	0.067*
No especificado	0.5 %	0.4 %	-0.1 pp	-2.1	0.300
<i>Panel B: Variables continuas</i>					
Edad (años)	36.87	36.31	-0.566	-5.3	0.007***
Años de educación	16.66	16.73	+0.065	+9.4	0.000***
ln(Ingreso mensual) en t5	8.61	9.01	+0.391	+56.1	0.000***

Notas: El p -valor corresponde a la prueba t . *** $p < 0,01$; ** $p < 0,05$; * $p < 0,10$.

Fuente: elaboración propia con datos de la ENOE (INEGI, 2005–2025).

Capítulo 3

Resultados

Este capítulo presenta los resultados empíricos de la investigación. En primer lugar, se expone la estimación del puntaje de propensión mediante un modelo Logit, con el objetivo de modelar la probabilidad de transitar de la informalidad a la formalidad laboral a partir de características observables de los egresados universitarios. Posteriormente, se analiza la calidad del emparejamiento, revisando el balance alcanzado entre tratados y controles. Finalmente se presentan las estimaciones del efecto promedio del tratamiento sobre los tratados, con énfasis en la comparación temporal de los resultados y en sus implicaciones para comprender la relación entre formalidad laboral e ingresos en el mercado laboral de los egresados universitarios de México.

3.1. Estimación del puntaje de propensión

3.1.1. Especificación empírica del modelo Logit

Como etapa previa a la estimación del efecto de la formalidad laboral sobre el ingreso, se estimó el puntaje de propensión de cada observación. En el contexto de esta investigación, la variable de tratamiento se define como una variable binaria que toma el valor de uno para los egresados universitarios que logran transitar a la formalidad y cero para aquellos que se mantienen informales.

La estimación del puntaje de propensión se realizó mediante un modelo Logit. Formalmente, para cada individuo i en el año t , se estimó la siguiente probabilidad condicional:

$$P(T_{it} = 1 \mid X_{it}) = e(X'_{it}\beta_t),$$

donde T_{it} indica la condición de tratamiento, X_{it} representa el vector de covariables observables y $e(\cdot)$ corresponde a la función logística acumulada, definida como:

$$e(X'_{it}\beta_t) = \frac{\exp(X'_{it}\beta_t)}{1 + \exp(X'_{it}\beta_t)}.$$

En el vector de covariables se incluyeron características sociodemográficas, geográficas y laborales observables: sexo, edad, años de educación, estado civil, condición de jefatura del hogar, tipo de localidad, macrozona geográfica y rama de actividad económica. La inclusión de estas variables responde a que dichos atributos pueden incidir en la probabilidad de inserción en el empleo formal, por lo que su incorporación permite controlar diferencias observables previas entre egresados universitarios formales e informales.

De manera específica, la especificación empírica puede expresarse como:

$$P(T_{it} = 1 | X_{it}) = e(\beta_{0t} + \beta_{1t} \text{sexo}_{it} + \beta_{2t} \text{edad}_{it} + \beta_{3t} \text{años_edu}_{it} + \beta_{4t} \text{soltero}_{it} + \beta_{5t} \text{jefhog}_{it} + \beta_{6t} \text{urbano}_{it} + \gamma_{zt} \text{zona}_{it} + \delta_{rt} \text{rama}_{it}).$$

El modelo se estimó de manera independiente para cada año del periodo 2005–2025. Esta decisión permite que la relación entre las características observables y la probabilidad de formalidad laboral varíe a lo largo del tiempo, en lugar de imponer una estructura constante para todo el periodo de análisis.

El puntaje de propensión estimado a partir de este modelo constituye el insumo central para el procedimiento de emparejamiento. A partir de él se comparan egresados universitarios formales e informales con probabilidades similares de transitar al sector formal, buscando mejorar la comparabilidad entre ambos grupos en términos de sus características observables.

3.1.2. Efectos marginales promedio

Los efectos marginales permiten expresar la asociación entre cada covariable y la probabilidad estimada de transitar a la formalidad en términos de cambios promedio de probabilidad. En el contexto de esta investigación, dichos efectos indican cómo varía la probabilidad de que un egresado universitario transite de la informalidad a la formalidad ante cambios en sus características observables. Los efectos marginales a detalle se muestran en la aplicación, específicamente en la página *Modelo logit*.

En términos generales, los resultados muestran que no existe una variable que mantenga significancia estadística durante el periodo 2005–2025. Por el contrario, las variables relevantes cambian entre años. Esto sugiere que la estructura observable asociada con la transición hacia la formalidad laboral es dinámica y que la probabilidad de tratamiento responde a distintos factores según el periodo analizado.

Una de las variables que aparece con mayor frecuencia es la edad, aunque su signo no es estable durante todo el periodo. En varios años, como 2005, 2009, 2014, 2018, 2019, 2024 y 2025, la edad se asocia negativamente con la probabilidad estimada de formalizar. Por ejemplo, en 2024 un año adicional de edad se asocia con una reducción de aproximadamente de 0.79 puntos porcentuales en la probabilidad de formalizarse, mientras que en 2025 la reducción es de aproximadamente 0.63 puntos porcentuales. No obstante, en 2021 y 2023, la edad presenta una asociación positiva. Esta variación indica que la relación entre edad y formalidad no debe interpretarse como un patrón estructural fijo, sino como una asociación que cambia según el año.

La variable años de educación resulta significativa únicamente en los años 2006 y 2013. En ambos casos, la asociación es positiva. En 2006, un año adicional de escolaridad se asocia con un incremento aproximado de 13.2 puntos porcentuales en la probabilidad de transitar hacia la formalidad; en 2013, el aumento es de aproximadamente de 11.5 puntos porcentuales. Sin embargo, en la mayoría de los años la escolaridad no resulta significativa. Aunque la educación puede estar asociada con la formalidad de egresados universitarios en ciertos periodos, no actúa como predictor estable durante el periodo de análisis.

La variable localidad urbana adquiere especial relevancia en los años recientes. En 2021, 2023, 2024 y 2025, residir en una localidad urbana se asocia positivamente con la probabilidad estimada de formalidad. El efecto más alto se observa en 2024, donde residir en una localidad urbana se asocia con una probabilidad 21 puntos porcentuales mayor de pertenecer al grupo formal. En 2021 el efecto es de 18.4 puntos porcentuales, en 2023 de 13.5 puntos porcentuales y en 2025 de 13.7 puntos porcentuales. Este patrón sugiere que, hacia el final del periodo 2005–2025, la dimensión urbano-rural adquiere un papel importante en la diferenciación entre individuos formales e informales.

En cuanto a la zona geográfica, la zona Sur aparece en varios años con una asociación negativa respecto a la zona de referencia. Este resultado se observa en 2005, 2007, 2011, 2020, 2021, 2022 y 2023. En 2005 y 2011, pertenecer a la zona Sur se asocia con una probabilidad 24.3 puntos porcentuales menor de formalizar; en 2007, con una reducción de 19.5 puntos porcentuales; en 2020, con una reducción de 25.7 puntos porcentuales; en 2021 y 2022, con una reducción de alrededor de 14 puntos porcentuales; y en 2023, con una reducción de 16.0 puntos porcentuales. Estos resultados sugieren que en ciertos años, la dimensión territorial contribuye a explicar diferencias observables en la probabilidad de transitar a la formalidad.

Las variables asociadas a la rama de actividad económica también muestran relevancia en periodos específicos. El sector agropecuario presenta asociaciones negativas importantes en 2011 y 2016, con efectos marginales de -52.7 y -55.5 puntos porcentuales, respectivamente. Esto indica que, en esos años, pertenecer a dicha rama se asocia con una menor probabilidad de pasar a la formalidad respecto al sector de construcción.

Asimismo, Comercio y Servicios presentan efectos negativos en algunos años, como 2007 y 2016. Sin embargo, Servicios aparece con signo positivo en 2010 y 2023, lo que confirma que la relación entre rama de actividad y formalidad cambia con el tiempo y depende de la categoría de referencia.

Otras variables presentan efectos puntuales. La variable Sexo resulta significativa en 2007, 2012 y 2021, con asociación positiva para el género masculino. La variable jefatura del hogar presenta asociaciones positivas en 2009 y 2013, pero una asociación negativa marginal en 2023. Finalmente, la variable soltero aparece como significativa en 2025, con una asociación negativa. Estos resultados deben entenderse como asociaciones estadísticas específicas de cada año, no como efectos causales ni como relaciones permanentes.

En conjunto, los efectos marginales promedio muestran que la probabilidad estimada de formalidad depende de una combinación cambiante de factores individuales, territoriales, urbanos y sectoriales. En los primeros años de la serie aparecen asociaciones relacionadas con escolaridad,

zona geográfica y rama de actividad; en los intermedios destacan los factores sectoriales; y en los años recientes la residencia urbana adquiere mayor importancia. Esta evolución refuerza la decisión metodológica de estimar el puntaje de propensión de manera separada para cada año.

3.1.3. Distribución y evolución temporal del puntaje de propensión

El puntaje de propensión estimado resume, para cada egresado universitario, la probabilidad predicha de ingresar a la formalidad dadas sus características observables. Por ello, analizar su distribución permite evaluar qué tan heterogénea es la probabilidad de tratamiento dentro de cada año y anticipar posibles problemas de soporte común en la etapa de emparejamiento.

En los primeros años con información reportada del puntaje, la media se mantiene cercana a 0.45. En 2006, la media fue 0.4510; en 2007, 0.4438; en 2008, 0.4732; y en 2009, 0.4552. Esto indica que, durante este subperiodo, la probabilidad promedio estimada de pertenecer al grupo formal se ubicó alrededor de 44 % a 47 %.

A partir de 2010 se observa una mayor variabilidad. En 2010, la media del puntaje disminuyó a 0.3678; en 2011 aumentó nuevamente a 0.4507; en 2012 se ubicó en 0.3869; y en 2013 volvió a aumentar a 0.4444. En 2014, la media alcanzó 0.4877, uno de los valores más altos del periodo. Posteriormente, en 2015 y 2016, la media se ubicó en 0.3853 y 0.3776, respectivamente.

En los años recientes también se observan cambios importantes. En 2021, la media fue 0.4086; en 2022, 0.4216; y en 2023, 0.4339. Sin embargo, en 2024 disminuyó a 0.3666 y en 2025 cayó a 0.3263, el valor promedio más bajo del periodo analizado. Esta caída indica que, hacia 2025, la probabilidad promedio estimada de pertenecer a la formalidad fue menor en comparación con años previos.

La dispersión del puntaje de propensión también varía de manera considerable entre años. Los valores más altos de desviación estándar se observan en 2021, con 0.1808; 2020, con 0.1728; 2011, con 0.1705; 2013, con 0.1674; y 2009, con 0.1604. En estos años, el modelo identifica perfiles con probabilidades más distintas de pertenecer al grupo formal, lo que indica una mayor heterogeneidad en la probabilidad estimada del tratamiento.

Esta mayor dispersión puede ser útil para distinguir entre individuos con baja y alta probabilidad de pertenecer al grupo formal. Sin embargo, también puede generar problemas si existen observaciones con probabilidades muy cercanas a cero o uno, ya que esas unidades pueden tener pocos individuos comparables en el grupo contrario. Por lo que, en estos años resulta especialmente importante revisar el soporte común antes de realizar el emparejamiento.

Por el contrario, los años con menor dispersión son 2015, con una desviación estándar de 0.0806; 2008, con 0.0889; 2018, con 0.0905; y 2010, con 0.0929. En estos años, las probabilidades estimadas se concentran más cerca de la media. Esta concentración puede favorecer la superposición entre tratados y controles, aunque también puede indicar que el modelo tiene menor capacidad para separar claramente a ambos grupos.

En cuanto a los valores extremos, algunos años presentan máximos cercanos a uno. Destacan 2017, con un máximo de 0.9811; 2007, con 0.9429; 2020, con 0.9394; 2013, con 0.9188; 2021, con 0.9145; y 2009, con 0.9116. Estos valores indican que existen individuos con probabilidad

estimada muy alta de pertenecer a la formalidad. En estos casos, es necesario revisar si existen unidades comparables dentro de grupo de control, ya que los valores extremos pueden dificultar el emparejamiento.

En contraste, años como 2008, 2010, 2018, 2023 y 2024 presentan máximos más acotados. Por ejemplo, en 2008 el máximo fue 0.6490; en 2010, 0.5409; en 2018, 0.6562; en 2023, 0.7688; y en 2024, 0.6761. Estos valores sugieren una menor presencia de probabilidades extremadamente altas, lo que puede favorecer el soporte común, aunque la calidad del emparejamiento debe evaluarse empíricamente mediante el diagnóstico de balance.

En conjunto, la evolución temporal del puntaje de propensión muestra tres patrones relevantes. Primero, la probabilidad promedio estimada de formalidad no es constante en el tiempo y representa una reducción marcada hacia 2025. Segundo, la dispersión del puntaje varía considerablemente entre años, lo que indica que la heterogeneidad observable de los individuos cambia a lo largo del periodo. Tercero, algunos años presentan valores extremos que obligan a revisar cuidadosamente el soporte común antes de implementar el emparejamiento.

Estos resultados son importantes porque el puntaje de propensión no solo resume la probabilidad de tratamiento, sino que también condiciona la calidad potencial del emparejamiento. Una buena estimación del puntaje debe permitir comparar egresados formales e informales con probabilidades similares de tratamiento. Por ello, la interpretación de la distribución del puntaje debe completarse con los diagnósticos de balance posteriores al emparejamiento.

3.2. Calidad del emparejamiento

La validez interna del estimador ATT depende fundamentalmente de que el procedimiento de emparejamiento logre reducir las diferencias observables entre el grupo tratado y el grupo de control. En esta investigación, ello implica que los egresados que transitaron de la informalidad a la formalidad sean comparados con egresados que permanecieron en la informalidad, pero que presentan características observables similares antes del tratamiento.

Dado que este supuesto de selección sobre observables no es directamente verificable a partir de los datos, resulta necesario documentar la calidad del balance alcanzado después del emparejamiento. Un balance adecuado en las covariables observables constituye una condición necesaria, aunque no suficiente, para que las estimaciones del ATT sean interpretadas como comparaciones más creíbles entre los grupos. Por lo tanto, la evaluación del balance cumple una función diagnóstica, pues permitiendo revisar si el procedimiento de emparejamiento redujo el sesgo de selección asociado a diferencias observables entre tratados y controles.

Esta investigación evalúa la calidad del emparejamiento mediante tres indicadores principales: el sesgo estandarizado por covariable, el sesgo estandarizado promedio y la prueba t de diferencias de medias. El sesgo estandarizado por covariable permite medir la magnitud de las diferencias entre tratados y controles en cada característica observable, antes y después del emparejamiento. A partir de estos valores, el sesgo estandarizado promedio resume el desequilibrio general del conjunto de covariables incluidas en el modelo de puntaje de propensión.

La prueba t de diferencias de medias se utiliza de manera complementaria para identificar si,

después del emparejamiento, persisten diferencias estadísticamente significativas entre tratados y controles en alguna covariable. Antes del emparejamiento, es esperable que algunas covariables presenten diferencias relevantes entre ambos grupos; después del emparejamiento, dichas diferencias deberían reducirse o dejar de ser estadísticamente significativas. No obstante, esta prueba se interpreta como complemento del sesgo estandarizado.

En esta investigación, un menor sesgo estandarizado posterior al emparejamiento se interpreta como evidencia de mejor balance. De manera particular, valores reducidos del sesgo residual, cercanos o inferiores al 5 %, se consideran indicativos de un balance aceptable (Caliendo y Kopeinig, 2008). No obstante, dado que no existe un umbral único, también se considera como referencia complementaria el criterio de 10 % (Austin, 2009).

La presente sección muestra los resultados diagnósticos de balance correspondientes a la aplicación anual del emparejamiento por puntajes de propensión al panel 2005–2025, estimado mediante la implementación `psmatch2` de Leuven y Sianesi (2003) en Stata y diagnosticando mediante el comando complementario `pstest`. La estrategia metodológica de la investigación contempló tres algoritmos de emparejamiento: Kernel, Vecino más cercano con un solo vecino ($n=1$) y Caliper con radio fijado en 0.2 veces la desviación estándar del puntaje de propensión. Kernel constituye la especificación principal del estudio, mientras que Vecino más cercano y Caliper se reportan como ejercicios de robustez metodológica, con el propósito de evaluar la sensibilidad de los resultados ante distintas formas de construir el grupo de comparación.

3.2.1. Diagnóstico longitudinal de balance bajo el emparejamiento por Kernel

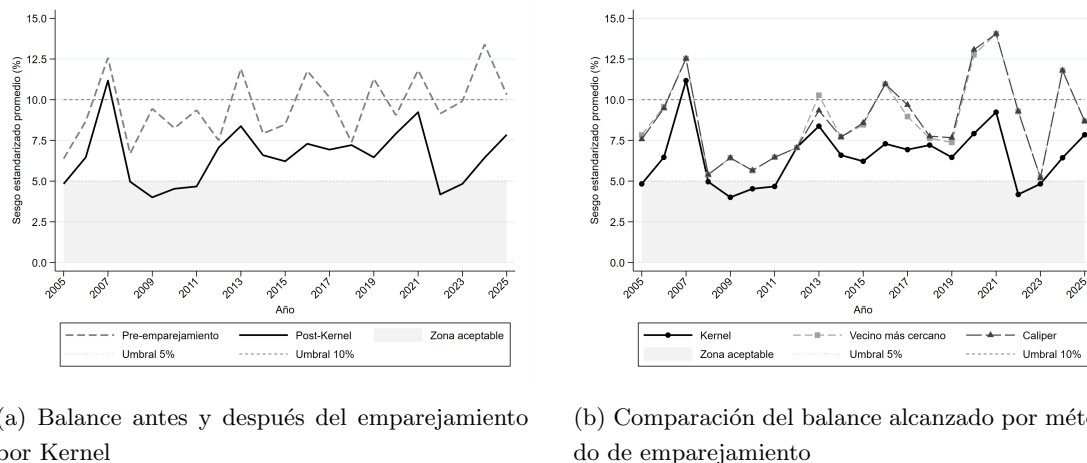
El diagnóstico longitudinal del balance bajo Kernel revela un patrón de cumplimiento heterogéneo a lo largo de los veintiún años analizados. En términos generales, este método reduce el desequilibrio observado entre tratados y controles en la mayoría de los años, aunque la magnitud de dicha reducción no es uniforme.

La Figura 3.1 resume la evolución del sesgo estandarizado promedio antes y después del emparejamiento, mientras que la tabla 3.1 presenta los valores resumidos de sesgo promedio pre y post por algoritmo de emparejamiento. Bajo Kernel, el sesgo promedio posterior disminuye en la mayor parte del periodo, lo que indica una mejora general en la comparabilidad observable entre los egresados que transitaron de la informalidad a la formalidad y aquellos que permanecieron en la informalidad.

Los años con mejor desempeño bajo Kernel corresponden a 2005, 2008, 2009, 2010, 2011, 2022 y 2023, en los cuales el sesgo promedio se ubica en valores reducidos, inferiores al umbral de referencia de 5 %. Estos resultados sugieren que, en dichos años, el emparejamiento Kernel logró construir grupos de comparación relativamente similares en términos de las covariables observables incluidas en el modelo.

Entre los años con mejor balance destaca 2009, al presentar una de las menores magnitudes de sesgo promedio posterior al emparejamiento y ausencia de diferencias estadísticamente significativas en las covariables después del emparejamiento. De manera similar, 2010, 2011 y 2023

muestran un desempeño favorable, tanto por la reducción del sesgo promedio como por la menor presencia de diferencias significativas en las pruebas t . En estos casos, la evidencia de balance posterior al emparejamiento ofrece una base más sólida para interpretar el ATT estimado.



(a) Balance antes y después del emparejamiento por Kernel

(b) Comparación del balance alcanzado por método de emparejamiento

Figura 3.1: Evolución del sesgo estandarizado promedio por método de emparejamiento. Egresados universitarios, México 2005–2025.

Fuente: Elaboración propia con datos de la ENOE (INEGI, 2005–2025).

No obstante, el balance no es uniforme a lo largo de la serie. Algunos años presentan reducciones parciales del sesgo pero conservan desequilibrios residuales en covariables específicas. Este comportamiento se observa particularmente en 2006, 2007, 2012, 2013, 2014, 2016, 2017, 2018, 2019, 2021 y 2025, años en los que algunas covariables continúan mostrando diferencias estadísticamente significativas después del emparejamiento. En estos casos, Kernel mejora la comparabilidad respecto a la muestra original, el balance posterior no alcanza el mismo nivel observado en los años de mejor desempeño.

El caso más desfavorable es 2007, debido a que covariables como sexo, edad, años de educación, soltero y jefe de hogar permanecen significativamente diferentes entre tratados y controles después del emparejamiento. Además, este año registra uno de los mayores niveles de sesgo promedio posterior al emparejamiento, superior al 10 %, lo que sugiere una capacidad limitada del Kernel para corregir el desequilibrio inicial. En 2021 también se observa un balance débil, asociado con la persistencia de diferencias significativas en variables como soltero e Industria Manufacturera. En contraste, algunos años muestran una mejora del balance promedio, aunque sin eliminar por completo las diferencias en todas las covariables. Estos casos pueden considerarse como situaciones de balance intermedio: el emparejamiento reduce el desequilibrio general, pero la presencia de diferencias significativas en covariables impide clasificarlos como años de balance fuerte. Para estos años, la interpretación del ATT debe reconocer que persisten diferencias observables que podrían afectar parcialmente la comparación entre tratados y controles.

A nivel de covariables individuales, los resultados muestran que el emparejamiento Kernel reduce las diferencias entre tratados y controles en la mayoría de los años. Sin embargo, la prueba t

posterior al emparejamiento permite identificar casos en los que algunas covariables continúan siendo estadísticamente diferentes entre ambos grupos. Por ello, la ausencia de diferencias significativas después del emparejamiento se interpreta como evidencia favorable de balance, mientras que la persistencia de significancia estadística en covariables se considera una señal de balance incompleto. Los resultados detallados de las pruebas t por covariable se presentan en la aplicación interactiva, específicamente en la página *Propensity Score Matching - Balance* de la aplicación. En conjunto, el diagnóstico longitudinal muestra que Kernel reduce el sesgo estandarizado promedio en buena parte del periodo y mejora la comparabilidad entre tratados y controles. Sin embargo, los resultados también muestran que la calidad del balance varía entre años, por lo que esta heterogeneidad se retoma en la subsección siguiente para establecer criterios diferenciados de interpretación del ATT.

Tabla 3.1: Sesgo estandarizado promedio pre y post emparejamiento por algoritmo, 2005–2025.

Año	Método	Sesgo promedio pre (%)	Sesgo promedio post (%)	p-valor post
2005	Kernel	6.4	4.8	0.987
	NN	6.4	7.8	0.163
	Caliper	6.4	7.6	0.207
2006	Kernel	8.7	6.5	0.721
	NN	8.7	9.6	0.420
	Caliper	8.7	9.5	0.287
2007	Kernel	12.6	11.2	0.103
	NN	12.6	12.5	0.005
	Caliper	12.6	12.5	0.005
2008	Kernel	6.7	5.0	0.972
	NN	6.7	5.4	0.774
	Caliper	6.7	5.4	0.774
2009	Kernel	9.5	4.0	0.985
	NN	9.5	6.4	0.529
	Caliper	9.5	6.4	0.529
2010	Kernel	8.3	4.5	0.987
	NN	8.3	5.7	0.555
	Caliper	8.3	5.7	0.555
2011	Kernel	9.4	4.7	0.907
	NN	9.4	6.5	0.769
	Caliper	9.4	6.5	0.769
2012	Kernel	7.5	7.1	0.905
	NN	7.5	7.1	0.694
	Caliper	7.5	7.1	0.694
2013	Kernel	11.9	8.4	0.406
	NN	11.9	10.3	0.163
	Caliper	11.9	9.3	0.184
2014	Kernel	7.9	6.6	0.469
	NN	7.9	7.7	0.489
	Caliper	7.9	7.7	0.489
2015	Kernel	8.5	6.2	0.836
	NN	8.5	8.5	0.423
Continúa en la siguiente página...				

Tabla 3.1 – Continuación

Año	Método	Sesgo promedio pre (%)	Sesgo promedio post (%)	<i>p</i> -valor post
	Caliper	8.5	8.6	0.299
2016	Kernel	11.8	7.3	0.849
	NN	11.8	11.0	0.065
	Caliper	11.8	11.0	0.065
2017	Kernel	10.1	6.9	0.767
	NN	10.1	9.0	0.322
	Caliper	10.1	9.7	0.238
2018	Kernel	7.4	7.2	0.376
	NN	7.4	7.6	0.485
	Caliper	7.4	7.8	0.439
2019	Kernel	11.3	6.5	0.832
	NN	11.3	7.4	0.605
	Caliper	11.3	7.7	0.536
2020	Kernel	9.0	7.9	0.950
	NN	9.0	12.8	0.343
	Caliper	9.0	13.1	0.248
2021	Kernel	11.8	9.2	0.485
	NN	11.8	14.0	0.025
	Caliper	11.8	14.0	0.025
2022	Kernel	9.1	4.2	0.909
	NN	9.1	9.3	0.038
	Caliper	9.1	9.3	0.038
2023	Kernel	9.9	4.8	0.925
	NN	9.9	5.2	0.696
	Caliper	9.9	5.2	0.696
2024	Kernel	13.4	6.4	0.828
	NN	13.4	11.8	0.010
	Caliper	13.4	11.8	0.010
2025	Kernel	10.3	7.8	0.514
	NN	10.3	8.7	0.328
	Caliper	10.3	8.7	0.328

El *p*-valor post corresponde al diagnóstico agregado posterior al emparejamiento; valores $\geq 0,10$ se resaltan en **negrita**.
Fuente: elaboración propia con datos de la ENOE (INEGI, 2005–2025).

3.2.2. Análisis de robustez: comparación con los métodos Vecino Más Cercano y Caliper

Como ejercicio de robustez metodológica, se replicó la estimación utilizando dos algoritmos alternativos de emparejamiento: el método de Vecino Más Cercano con un vecino y soporte común, denotado como NN(1), y el método Caliper. En este último, el radio de emparejamiento se fijó en 0.2 desviaciones estándar del puntaje de propensión. Esta elección se adopta como un criterio operativo de cercanía entre tratados y controles, inspirado en la literatura sobre emparejamiento con Caliper. No obstante, debe señalarse que la recomendación específica de Austin (2011) se refiere al uso de 0.2 desviaciones estándar del Logit del puntaje de propensión.

La comparación entre los tres algoritmos permite evaluar la sensibilidad de los diagnósticos de balance al método específico de emparejamiento y proporciona evidencia complementaria sobre la robustez de la estrategia empírica.

Los resultados resumidos en la tabla 3.1 muestran que el método Kernel presenta el desempeño más estable a lo largo del periodo 2005–2025. En comparación con NN(1) y Caliper, Kernel reduce de manera más consistente el sesgo estandarizado promedio después del emparejamiento. Esta diferencia es relevante porque el objetivo del emparejamiento no es únicamente encontrar controles cercanos para cada tratado, sino reducir el desequilibrio observado entre ambos grupos en el conjunto de covariables utilizadas para estimar el puntaje de propensión.

Un primer patrón se observa en los años donde el sesgo promedio posterior al emparejamiento aumenta bajo NN(1) o Caliper. En el caso de NN(1), el sesgo promedio aumenta en 2005, de 6.4 % a 7.8 %; en 2006, de 8.7 % a 9.6 %; en 2018, de 7.4 % a 7.6 %; en 2020, de 9.0 % a 12.8 %; en 2021, de 11.8 % a 14.0 %; y en 2022, de 9.1 % a 9.3 %. Además, en 2015 no se observa una mejora efectiva, pues el sesgo permanece en 8.5 %. Para Caliper, el deterioro es similar: el sesgo promedio aumenta en 2005, de 6.4 % a 7.6 %; en 2006, de 8.7 % a 9.5 %; en 2015, de 8.5 % a 8.6 %; en 2018, de 7.4 % a 7.8 %; en 2020, de 9.0 % a 13.1 %; en 2021, de 11.8 % a 14.0 %; y en 2022, de 9.1 % a 9.3 %. Estos resultados indican que, en determinados años, los métodos alternativos no lograron reducir el desbalance promedio inicial e incluso lo incrementan.

Los años más problemáticos para NN(1) y Caliper son 2007, 2016, 2020, 2021 y 2024, debido a que presentan niveles elevados de sesgo promedio posterior al emparejamiento. En particular, en 2007 ambos métodos presentan un sesgo promedio posterior al emparejamiento de 12.5 %; en 2016, de 11.0 %; en 2020, de 12.8 % bajo NN(1) y 13.1 % bajo Caliper; en 2021, de 14.0 % en ambos métodos; y en 2024 de 11.8 %. Estos valores, superiores al umbral de 10 %, sugieren un balance débil.

En contraste, Kernel muestra un comportamiento más favorable en la mayoría de esos mismos años. Por ejemplo, en 2016 reduce el sesgo promedio de 11.8 % a 7.3 %, mientras que NN(1) y Caliper permanecen en 11.0 %. En 2020, Kernel reduce el sesgo de 9.0 % a 7.9 %, mientras que NN(1) y Caliper lo incrementan por encima de 12 %. En 2021, aunque Kernel no alcanza un balance fuerte, registra un sesgo posterior de 9.2 %, menor que el 14 % observado bajo los métodos alternativos. Finalmente, en 2024 Kernel reduce el sesgo de 13.4 % a 6.4 %, frente al 11.8 % obtenido con NN(1) y Caliper. Esta comparación confirma que Kernel tiende a ser menos sensible a deterioros del balance promedio.

No obstante, la superioridad de Kernel no implica ausencia total de problemas. Algunos años, como 2007, 2013 y 2021, conservan niveles de sesgo posteriores al emparejamiento moderados o elevados incluso bajo este método. En 2007, el sesgo promedio bajo Kernel alcanzó 11.2 %, el valor más alto registrado para este algoritmo. En 2013 se ubica en 8.4 % y en 2021 en 9.2 %. Estos casos deben interpretarse con cautela, pues el emparejamiento mejora parcialmente la comparabilidad, pero no elimina por completo el desbalance residual.

La comparación entre métodos también permite identificar los años con mejor desempeño relativo. Bajo Kernel, los años más favorables son 2005, 2008, 2009, 2010, 2011 y 2023, debido a que presentan sesgos promedio inferiores al umbral de 5 %. En particular, 2009 registra un sesgo de

4.0 %, 2010 de 4.5 %, 2011 de 4.7 % y 2023 de 4.8 %. Bajo NN(1) y Caliper, los mejores resultados también se observan en 2023, con un sesgo promedio de 5.2 %, aunque en términos generales estos métodos muestran mayor variabilidad y más años con deterioro del balance.

En conjunto, la evidencia comparativa sugiere que Kernel constituye la especificación metodológicamente más estable dentro de las alternativas implementadas en esta investigación. Aunque no elimina completamente los problemas de balance residual y no garantiza mejoras uniformes en todas las covariables ni en todos los años, el algoritmo exhibe un comportamiento comparativamente más consistente, menos sensible a incrementos del sesgo promedio. En este sentido, NN(1) y Caliper cumplen una función útil como ejercicios de robustez, pero sus resultados deben interpretarse con cautela, especialmente en los años donde el sesgo posterior aumenta o permanece por encima del umbral de 10 %.

3.2.3. Alcances del balance y criterios para interpretar el ATT

Los resultados de balance presentados en los apartados anteriores permiten evaluar qué tan comparables son los grupos de tratamiento y control después del emparejamiento. En términos generales, el objetivo del emparejamiento por puntajes de propensión es construir un grupo de comparación lo más parecido posible al grupo tratado en sus características observables. Por ello, cuando el sesgo estandarizado promedio disminuye después del emparejamiento y las covariables dejan de presentar diferencias estadísticamente significativas, se cuenta con mayor evidencia de que la comparación entre ambos grupos es más adecuada que antes del emparejamiento.

Sin embargo, es importante señalar que el PSM no elimina todos los posibles problemas de selección. Su principal limitación es que solo puede controlar diferencias en características observables, es decir, aquellas incluidas en el modelo de selección. Si existen factores no observados que influyan simultáneamente en la probabilidad de formalizar y en el ingreso, podría persistir cierto sesgo residual. Por esta razón, los resultados deben interpretarse como estimaciones condicionadas al conjunto de covariables observadas.

A partir de los diagnósticos obtenidos, el método Kernel se adopta como especificación principal del análisis. Esta decisión no se fundamenta en una superioridad teórica absoluta, sino en su mejor desempeño empírico dentro de la muestra analizada. A diferencia de Vecino Más Cercano y Caliper, Kernel no depende de un único control para cada unidad tratada, sino que utiliza información de varios controles, asignando mayor peso a aquellos con puntajes más cercanos. Esta característica lo hace menos sensible a casos individuales y más flexible cuando existen diferencias entre tratados y controles.

La comparación con los métodos Vecino Más Cercano y Caliper refuerza esta elección. Ambos métodos funcionan como ejercicio de robustez, pero muestran un desempeño menos estable que Kernel. En particular, presentan varios años en los que el sesgo promedio no mejora de forma suficiente e incluso en algunos casos empeoran después del emparejamiento. Además, las pruebas t por covariable muestran que, bajo estos métodos, persisten diferencias estadísticamente significativas en algunos años. Por tanto, NN(1) y Caliper se utilizan como contraste metodológico, pero no como especificaciones principales del análisis.

No obstante, la superioridad relativa de Kernel no debe interpretarse como absoluta. Aunque fue el método con mejor desempeño general, la calidad del balance varía entre años. En algunos periodos el emparejamiento produjo una comparación muy sólida entre tratados y controles; en otros, el balance fue favorable pero menos robusto; y en algunos casos persistieron señales de desbalance residual. Es por eso, que la interpretación de los resultados debe distinguir distintos niveles de confianza según la calidad del balance alcanzado en cada año.

Con base en el sesgo estandarizado promedio posterior al emparejamiento y en la presencia o ausencia de diferencias significativas en las pruebas t por covariable, los años pueden organizarse en tres grupos interpretativos. El primer grupo correspondiente a los años con balance fuerte: 2005, 2008, 2009, 2010, 2011, 2022 y 2023. En estos casos, Kernel reduce claramente el sesgo promedio y lo ubica en valores cercanos o inferiores al umbral de referencia de 5 %. Además, las pruebas t posteriores al emparejamiento muestran menor presencia de diferencias estadísticamente significativas entre ambos grupos. Estos años constituyen los escenarios donde la comparación entre grupos es más sólida.

Un segundo grupo corresponde a los años con balance intermedio: 2006, 2012, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2024 y 2025. En estos periodos, el método de emparejamiento por Kernel mejora el balance respecto a la situación previa al emparejamiento, pero el sesgo promedio posterior se mantiene por encima del 5 % o persisten diferencias significativas en algunas covariables. En consecuencia, las estimaciones del ATT para estos años pueden interpretarse, pero con una cautela mayor que en los años de balance fuerte.

Finalmente, los años 2007, 2013 y 2021 presentan las mayores señales de desbalance residual bajo Kernel. Estos casos no implican que los resultados deban descartarse, pero sí deben leerse con mayor prudencia. En particular, 2007 registra el mayor sesgo promedio posterior bajo Kernel, superior a 10 %, además de diferencias significativas persistentes en covariables como sexo, edad, años de educación, soltero y jefatura del hogar. En 2021 también se observan diferencias significativas en variables específicas, lo que sugiere un balance incompleto. Los ATT estimados para estos años deben interpretarse como evidencia menos robusta que la correspondiente a los años anteriores.

Esta clasificación no implica excluir años del análisis, sino distinguir el grado de confianza con el que deben interpretarse los efectos estimados. Un ATT obtenido en un año con balance fuerte se apoya en una comparación más limpia entre tratados y controles. En cambio, un ATT estimado en un año con balance intermedio o débil no necesariamente debe descartarse, pero sí debe interpretarse con mayor cautela, porque los grupos comparados aún pueden conservar diferencias observables relevantes.

Adicionalmente, para la inferencia estadística de las estimaciones del ATT se empleó un procedimiento de remuestreo bootstrap con 200 repeticiones y semilla fijada en 42. Esta estrategia permite aproximar la variabilidad muestral del efecto estimado y obtener errores estándar asociados al ATT. La fijación de una semilla garantiza la reproducibilidad de los resultados, mientras que el uso de bootstrap resulta pertinente debido a que el emparejamiento introduce una etapa adicional de estimación en la construcción del grupo de comparación.

En resumen, el análisis de balance respalda el uso del emparejamiento por puntajes de propen-

sión como estrategia para mejorar la comparabilidad entre egresados universitarios formales e informales. No obstante, también muestra que la calidad del emparejamiento no es homogénea a lo largo del tiempo. Los resultados deben entenderse como estimaciones válidas bajo el supuesto de que las diferencias relevantes entre grupos están capturadas por las covariables observadas. En consecuencia, las estimaciones del ATT deben interpretarse como evidencia empírica condicionada al modelo, al soporte común y a la calidad del balance alcanzado en cada año.

Con estos criterios, la siguiente sección presenta las estimaciones del ATT. La interpretación de dichos resultados considera no solo el signo y la magnitud del efecto estimado, sino también su significancia estadística, calculada mediante bootstrap, y la calidad del balance alcanzada en cada año.

3.3. Resultados de las estimaciones del ATT

3.3.1. Efecto promedio del tratamiento sobre los tratados bajo Kernel

En esta sección se presentan las estimaciones del efecto promedio del tratamiento sobre los tratados (ATT) mediante Kernel para el periodo 2005–2025. El tratamiento corresponde a transitar de la informalidad a la formalidad, mientras que el grupo de control está conformado por egresados universitarios que permanecieron en la informalidad. La variable de resultado es el logaritmo del ingreso mensual en el trimestre cinco (t_5).

La interpretación de los coeficientes se realiza en términos porcentuales mediante la transformación $(e^{ATT} - 1) \times 100$. Asimismo, la inferencia estadística se obtuvo mediante bootstrap. Los resultados se presentan en la tabla 3.2

Tabla 3.2: ATT estimados mediante emparejamiento Kernel con bootstrap ($B = 200$, semilla 42), 2005–2025.

Año	N	ATT		SE	z	p > z	IC 95 % [Inf., Sup.]	Prima (%)
		Unmatched	Matched	bootstrap				
2005	460	0.4879 ***	0.4688	0.0679	6.90	0.000	[0.3357, 0.6019]	59.81 %
2006	453	0.3656 ***	0.3870	0.0739	5.24	0.000	[0.2423, 0.5318]	47.26 %
2007	441	0.4221 ***	0.3688	0.0760	4.85	0.000	[0.2199, 0.5177]	44.60 %
2008	462	0.3809 ***	0.3902	0.0760	5.13	0.000	[0.2412, 0.5391]	47.73 %
2009	474	0.3724 ***	0.3211	0.0689	4.66	0.000	[0.1860, 0.4563]	37.86 %
2010	501	0.2904 ***	0.2828	0.0602	4.70	0.000	[0.1648, 0.4008]	32.68 %
2011	483	0.3641 ***	0.3589	0.0645	5.57	0.000	[0.2325, 0.4854]	43.18 %
2012	496	0.3991 ***	0.3856	0.0585	6.60	0.000	[0.2710, 0.5002]	47.05 %
2013	466	0.4667 ***	0.4217	0.0694	6.07	0.000	[0.2856, 0.5578]	52.46 %
2014	505	0.4967 ***	0.5156	0.0572	9.01	0.000	[0.4035, 0.6278]	67.46 %
2015	502	0.3600 ***	0.3746	0.0601	6.24	0.000	[0.2569, 0.4924]	45.44 %
2016	477	0.4010 ***	0.3767	0.0571	6.60	0.000	[0.2649, 0.4886]	45.75 %
2017	508	0.4105 ***	0.4076	0.0654	6.23	0.000	[0.2795, 0.5358]	50.32 %
2018	515	0.4002 ***	0.4047	0.0658	6.15	0.000	[0.2758, 0.5337]	49.89 %
2019	590	0.4689 ***	0.4630	0.0566	8.19	0.000	[0.3522, 0.5739]	58.88 %
2020	230	0.4539 ***	0.4760	0.0906	5.25	0.000	[0.2984, 0.6536]	60.96 %
2021	434	0.4700 ***	0.4476	0.0629	7.12	0.000	[0.3244, 0.5709]	56.46 %
2022	617	0.3643 ***	0.3298	0.0538	6.14	0.000	[0.2245, 0.4352]	39.07 %

Continúa en la siguiente página...

Tabla 3.2 – Continuación

Año	<i>N</i>	ATT Unmatched	ATT Matched	SE bootstrap	<i>z</i>	<i>p</i> > <i>z</i>	IC 95 % [Inf., Sup.]	Prima (%)
2023	736	0.4572 ***	0.4358	0.0455	9.58	0.000	[0.3466, 0.5250]	54.62 %
2024	637	0.4607 ***	0.4414	0.0588	7.50	0.000	[0.3262, 0.5567]	55.49 %
2025	681	0.4829 ***	0.4278	0.0567	7.54	0.000	[0.3166, 0.5389]	53.39 %

Variable de resultado: $\ln(\text{ingreso mensual})$.

Fuente: elaboración propia con datos de la ENOE (INEGI, 2005–2025).

Los resultados muestran un patrón consistente durante todo el periodo 2005–2025. En todos los años, el ATT estimado bajo Kernel es positivo y estadísticamente significativo. Esto indica que, después de comparar egresados universitarios con características similares, quienes transitan de la informalidad a la formalidad registraron ingresos mensuales superiores a los de aquellos que permanecieron en la informalidad.

La magnitud de los efectos estimados es económicamente relevante. Los ATT oscilan entre 0.2828 y 0.5156 puntos logarítmicos, lo que equivale aproximadamente a diferencias de ingreso mensual entre 32.7 % y 67.4 %. El efecto más bajo se observa en 2010, uno de los años con mejor balance, mientras que el efecto más alto corresponde a 2014, año donde el balance se muestra favorable respecto al equilibrio, aunque no alcanza el umbral de aceptación. A pesar de esta variación, el signo positivo del efecto se mantiene en todos los años, lo que sugiere una prima salarial persistente asociada a la formalización.

La interpretación de las estimaciones anuales debe considerar la calidad del balance alcanzado. Los años 2005, 2008, 2009, 2010, 2011, 2022 y 2023 presentan la base más sólida para interpretar el ATT, por lo que las estimaciones en estos años: 0.4688 (59.81 %), 0.3902 (47.73 %), 0.3211 (37.86 %), 0.2828 (32.68 %), 0.3589 (43.18 %), 0.3298 (39.07 %) y 0.4358 (54.62 %), ofrecen mayor confianza. En estos casos, el emparejamiento logró reducir de manera favorable el sesgo a niveles aceptables, además de no mostrar diferencias conjuntas significativas entre tratados y controles después del emparejamiento, a excepción de 2022.

Años como 2006, 2012, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2024 y 2025, presentan resultados favorables. En estos casos, el balance posterior al emparejamiento fue bueno, aunque el emparejamiento no logró completamente reducir el sesgo estandarizado a niveles adecuados, estos se ubicaron cerca de los umbrales recomendados reduciendo sesgo, aunque los resultados de estos años no son del todo favorables, podemos interpretar las estimaciones ATT con una confianza moderada. Correspondiéndole un ATT de 0.3870 (47.26 %) al año 2006, 0.3852 (47.05 %) a 2012, 0.5156 (67.46 %) a 2014, 0.3746 (45.44 %) a 2015, de 0.3767 (45.75 %) al año 2016, de 0.4076 (50.32 %) al año 2017, 0.4047 (49.89 %) al año 2018, 0.4630 (58.88 %) al año 2019, 0.4760 (60.96 %) en 2020, de 0.4414 (55.49 %) al año 2024 y de 0.4278 (53.39 %) al año 2025.

En contraste, los años restantes 2007, 2013 y 2021 requieren una lectura más cautelosa, aunque el emparejamiento logró reducciones, estas se muestran alejadas de los umbrales aceptables. Aunque los ATT son positivos y estadísticamente significativos, los diagnósticos de balance muestran mayor persistencia de desbalance. Esto no implica descartar los resultados, pero sí reconocer que la comparación entre tratados y controles no fue tan limpia como en los años con mejor

desempeño del emparejamiento. Estos años estiman efectos de 0.3688 (44.60 %) en 2007, 0.4217 (52.46 %) en 2013 y 0.4476 (56.46 %) en 2021, siendo este uno de los años con mayor aumento.

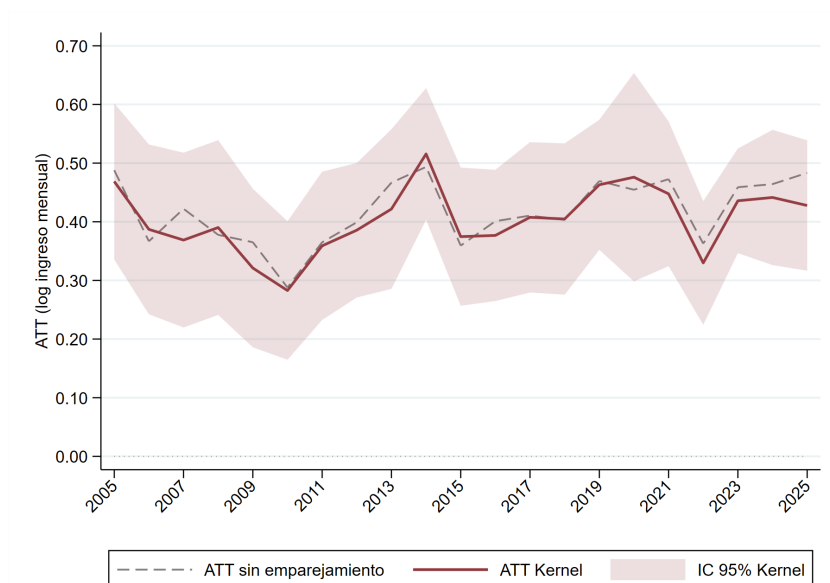


Figura 3.2: Comparación del ATT antes y después del emparejamiento Kernel. Egresados universitarios, México 2005–2025

Fuente: Elaboración propia con datos de la ENOE (INEGI, 2005–2025).

En resumen, el ATT estimado bajo Kernel es positivo y estadísticamente significativo en todos los años del periodo 2005–2025, con una prima salarial que supera el 30 % en cada año analizado. Pasar de la informalidad a la formalidad se asocia con mayores ingresos mensuales.

3.3.2. Comparación de resultados con Vecino Más Cercano y Caliper

Con el objetivo de evaluar la sensibilidad de las estimaciones al algoritmo de emparejamiento utilizado, se compararon los resultados obtenidos mediante Kernel con dos métodos alternativos: Vecino Más Cercano con un vecino, sin reemplazo, y Caliper. Esta comparación permite analizar si la relación positiva entre la transición de la informalidad a la formalidad y el ingreso mensual se mantiene al modificar las reglas de emparejamiento.

Los resultados detallados por año y método se encuentran disponibles en la aplicación desarrollada para esta investigación, específicamente en la sección *PSM Resultados*. Este tablero permite consultar las estimaciones del ATT para los métodos Kernel, Vecino Más Cercano y Caliper durante el periodo 2005–2025.

En términos generales, los tres métodos de emparejamiento producen estimaciones del ATT positivas y estadísticamente significativas en todos los años del periodo analizado. La coincidencia en el signo y la significancia indica que la conclusión principal no depende del algoritmo específico

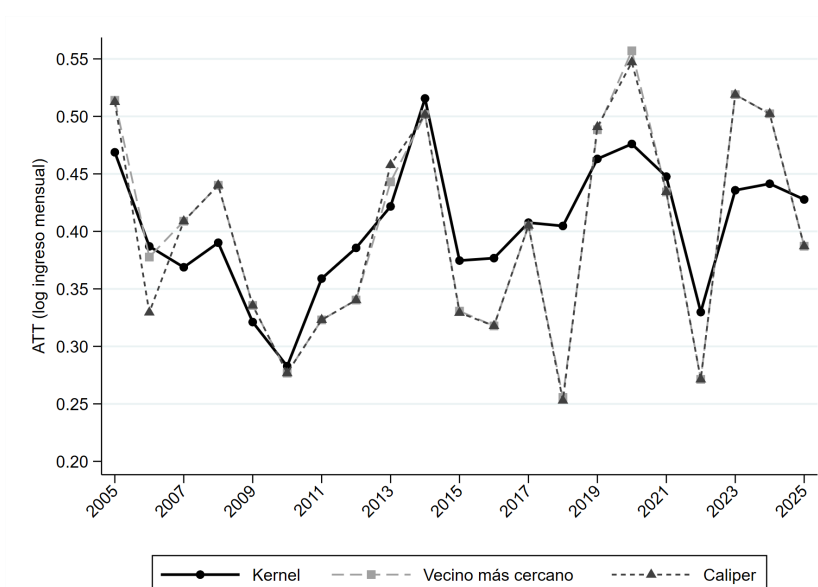


Figura 3.3: Comparación del ATT estimado por método de emparejamiento. Egresados universitarios, México 2005–2025.

Fuente: Elaboración propia con datos de la ENOE (INEGI, 2005–2025).

utilizado. No obstante, la magnitud estimada varía entre métodos, lo que refleja diferencias en la forma en que cada uno construye el contrafactual.

Asimismo, las estimaciones obtenidas mediante Vecino Más Cercano y Caliper resultan estadísticamente significativas en todos los años del periodo analizado. Esto refuerza la evidencia de una asociación positiva entre la formalización laboral y el ingreso mensual. No obstante, aunque la dirección del efecto se mantiene, la magnitud estimada varía entre métodos. En algunos años, los métodos alternativos producen estimaciones cercanas a Kernel, como ocurre en 2010, 2014, 2017, 2019 y 2021. En otros casos, las diferencias son más marcadas, como en 2018, 2020, 2022, 2023 y 2024, lo que muestra que la magnitud del ATT puede ser sensible al algoritmo de emparejamiento empleado, dada la calidad del balance.

Esta sensibilidad es esperable debido a la forma en que cada método construye el contrafactual. Mientras Kernel utiliza información ponderada de múltiples unidades de control, asignando mayor peso a aquellas con puntajes de propensión más cercanos, Vecino Más Cercano selecciona una única unidad de control para cada tratado. Por su parte, Caliper restringe el emparejamiento a controles ubicados dentro de una distancia máxima estipulada respecto al puntaje de propensión del tratado. En consecuencia, los métodos alternativos pueden generar estimaciones más variables, especialmente en años donde la muestra es más reducida o donde el balance inicial entre tratados y controles es más débil.

La comparación también debe leerse junto con los diagnósticos de balance presentados previamente. Aunque Vecino Más Cercano y Caliper confirman la dirección positiva y la significancia

estadística del efecto, en varios años muestran un desempeño menos favorable que Kernel en términos de sesgo estandarizado promedio y balance por covariable. En particular, estos métodos presentan algunos años en los que el sesgo promedio posterior al emparejamiento permanece elevado o incluso aumenta respecto a la situación previa. Esto ocurre, por ejemplo, en 2005, 2006, 2018, 2020, 2021 y 2022 para Vecino Más Cercano, y en 2005, 2006, 2015, 2018, 2020, 2021 y 2022 para Caliper.

En conjunto, Vecino Más Cercano y Caliper respaldan la conclusión principal del análisis: confirman la dirección positiva del efecto, pero dado su menor desempeño en los diagnósticos de balance, sus estimaciones no se adoptan como especificación principal.

3.3.3. Síntesis de resultados e implicaciones para el mercado laboral

Las estimaciones del ATT presentadas en las secciones anteriores muestran un patrón consistente: durante el periodo 2005–2025, los egresados universitarios que transitaron de la informalidad a la formalidad registraron ingresos mensuales superiores a los de egresados observacionalmente similares que permanecieron informales. Esta consistencia se mantiene bajo los tres algoritmos de emparejamiento y a lo largo de todos los años analizados, con mayor solidez en los periodos de balance fuerte.

Desde una perspectiva laboral, estos resultados son más consistentes con una interpretación de segmentación de mercado de trabajo que con una visión de plena movilidad entre formalidad e informalidad. Si ambos sectores fueran espacios equivalentes para egresados universitarios comparables, no se esperaría observar una prima salarial positiva y persistente asociada a la formalización. En cambio, la evidencia sugiere que la formalidad constituye un segmento con mejores retornos económicos, mientras que la permanencia en la informalidad puede reflejar restricciones de acceso, dificultades para encontrar empleos formales, diferencias en la calidad del empleo o barreras estructurales.

Esta interpretación dialoga con la literatura sobre informalidad en México y América Latina. Por un lado, los resultados son compatibles con estudios que identifican barreras de entrada al empleo formal y segmentación laboral (Alcaraz et al. 2015). Por otro lado, no descartan por completo la visión más matizada de la informalidad como una posible decisión voluntaria en ciertos contextos (Maloney, 2004). Sin embargo, para el grupo específico analizado en esta investigación, egresados universitarios, la evidencia favorece a una lectura en la que transitar hacia la formalidad se asocia con mejores resultados salariales, aun después de comparar individuos observacionalmente similares.

También es importante señalar que la solidez de los resultados no es homogénea en todos los años analizados. Los periodos con mejor balance posterior al emparejamiento ofrecen una base más sólida, para interpretar el ATT, mientras que los años con mayor sesgo residual o diferencias significativas persistentes en covariables deben leerse con mayor cautela. Esta distinción no invalida el patrón general, pero sí permite reconocer que la calidad del emparejamiento varía a lo largo del tiempo.

En términos de implicaciones, los resultados sugieren que la formalización laboral puede asociar-

se con mejoras relevantes en el bienestar económico de los egresados universitarios, entendido aquí como incrementos en el ingreso monetario mensual. La formalidad parece operar como un mecanismo que permite traducir de manera más efectiva la educación superior en mayores ingresos.

En resumen, la transición de la informalidad a la formalidad se asocia con una prima salarial positiva, estadísticamente significativa y económicamente relevante para los egresados universitarios en México durante el periodo 2005–2025. Esta evidencia sugiere que la formalidad constituye un segmento donde el capital humano universitario recibe una mayor recompensa monetaria.

3.3.4. Alcances y limitaciones del estudio

La interpretación de los resultados debe mantenerse dentro de los alcances del método utilizado. El Emparejamiento por Puntajes de Propensión mejora la comparabilidad entre tratados y controles observacionalmente similares, pero no elimina necesariamente todas las fuentes de sesgo. Su validez depende del supuesto de selección sobre observables, es decir, de que las covariables incluidas capturen los principales factores tanto en la probabilidad de transitar hacia la formalidad como en el ingreso posterior.

Si el acceso a la formalidad depende de factores no observados, como redes de contacto, habilidades individuales, calidad de la institución educativa, área de estudio o experiencia laboral previa no medida, parte de la brecha estimada podría seguir reflejando diferencias estructurales no captadas por el modelo. En este caso, el ATT estimado absorbería no solo el efecto de la formalización, sino también la ventaja preexistente de quienes lograron acceder al empleo formal. Asimismo, la transición hacia la formalidad no implica necesariamente que todos los empleos formales sean de alta calidad ni que la formalidad resuelva por sí sola otros problemas del mercado laboral, como la subocupación, la precariedad contractual, las diferencias regionales o la baja productividad.

Por tanto, los resultados no deben interpretarse como una prueba experimental definitiva, sino como evidencia empírica consistente con una prima salarial positiva asociada a la formalización laboral, bajo los supuestos del modelo, el soporte común y la calidad del balance alcanzado.

Capítulo 4

Conclusiones

Esta investigación tuvo como objetivo analizar el impacto que tiene la formalización laboral sobre el ingreso mensual de los egresados universitarios en México durante el periodo 2005–2025. En particular, se buscó responder si transitar de la informalidad a la formalidad se asocia con incrementos en el ingreso monetario mensual y si la brecha salarial observada entre grupos puede interpretarse como una diferencia atribuible a la formalización una vez controladas las características observables de los egresados.

El punto de partida fue reconocer que una comparación simple entre egresados formales e informales puede estar afectada por sesgo de selección. Los individuos no se distribuyen aleatoriamente entre la formalidad y la informalidad, sino que pueden diferir tanto en características observables, como en características no observables. Por esta razón, la tesis recurrió al enfoque de inferencia causal y al emparejamiento por puntajes de propensión, con el propósito de construir grupos más comparables y aproximarse de manera más rigurosa al efecto asociado a la transición hacia la formalidad.

Los resultados muestran evidencia consistente de una prima salarial asociada a la formalización laboral. Bajo el método principal, Kernel, el efecto promedio del tratamiento sobre los tratados fue positivo y estadísticamente significativo en todos los años del análisis, con efectos que oscilan entre 0.2828 y 0.5156 puntos logarítmicos, equivalentes a incrementos en el ingreso mensual de entre 32.7% y 67.4%. La confianza en estos resultados se apoya en los diagnósticos de balance y en los ejercicios de robustez: Kernel se mantuvo como especificación principal por presentar el desempeño más estable, mientras que Vecino Más Cercano y Caliper confirmaron la dirección positiva de efecto. No obstante, la calidad del balance no fue homogénea entre años, por lo que la fuerza interpretativa del ATT debe entenderse como variable entre periodos y condicionada al balance alcanzado.

Estos hallazgos son consistentes con la hipótesis de segmentación laboral. Si la formalidad y la informalidad fueran espacios laborales integrados para egresados universitarios comparables, no se esperaría observar una prima salarial positiva y persistente asociada a la transición hacia la formalidad. La evidencia sugiere, en cambio, que la formalidad constituye un segmento del mercado de trabajo con mejores retornos económicos, donde el capital humano universitario recibe una

mayor recompensa. Así, la educación superior incrementa el capital humano, pero su traducción en mayores ingresos depende también del segmento laboral en el que se inserta el trabajador. Esta lectura es coherente con la literatura que documenta la persistencia de la informalidad entre trabajadores calificados en México y las diferencias entre los segmentos formal e informal (Alcaraz et al. 2015; Huesca Reynoso y Padilla Arriola, 2012). No obstante, debe mantenerse matizada: como señala Maloney (2004), la informalidad también puede responder en ciertos casos a decisiones voluntarias o preferencias por flexibilidad, posibilidad que los resultados de esta investigación no descartan, pero para el grupo analizado la evidencia respalda con mayor fuerza la hipótesis de segmentación. A esta limitación se suma que la calidad balance varió entre años, por lo que los resultados deben interpretarse con distintos niveles de confianza según el periodo analizado. En consecuencia, futuras investigaciones podrían incorporar variables adicionales relacionadas con el área de estudio, el tipo de institución educativa, la calidad del empleo, las prestaciones laborales, el tamaño de la empresa, el tipo de contrato o la experiencia laboral acumulada, así como analizar si la prima salarial varía por sexo, región, rama de actividad o cohorte laboral, y explorar métodos complementarios que aborden de forma más directa la heterogeneidad no observada.

En conjunto, la pregunta que motivó esta investigación, si la formalización representa una mejora salarial atribuible al cambio de estado laboral, encuentra una respuesta afirmativa y consistente en los datos: para los egresados universitarios mexicanos que transitaban al empleo formal entre 2005 y 2025, dicha transición se asoció con incrementos en el ingreso mensual, incluso después de comparar individuos observacionalmente similares. La educación superior, por sí sola, no garantiza esos mayores ingresos; es la inserción en el segmento formal del mercado laboral lo que determina en qué medida el capital humano se traduce en bienestar económico, siempre bajo los supuestos del emparejamiento y la calidad del balance alcanzado.

Referencias

- Alcaraz, C., Chiquiar, D., & Salcedo, A. (2015). *Informality and Segmentation in the Mexican Labor Market* (Working Paper). Banco de México.
- Antón, A., Hernández, F., & Levy Algazi, S. (2013). *The End of Informality in Mexico?: Fiscal Reform for Universal Social Insurance*. Inter-American Development Bank.
- Austin, P. C. (2009). Balance Diagnostics for Comparing the Distribution of Baseline Covariates Between Treatment Groups in Propensity-Score Matched Samples. *Statistics in Medicine*, 28(25), 3083-3107.
- Austin, P. C. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research*, 46(3), 399-424.
- Becker, G. S. (1964). *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. National Bureau of Economic Research.
- Bosch, M., & Maloney, W. F. (2008). *Cyclical Movements in Unemployment and Informality in Developing Countries* (Policy Research Working Paper N.º 4648). The World Bank.
- Bosch, M., & Maloney, W. F. (2010). Comparative Analysis of Labor Market Dynamics Using Markov Processes: An Application to Informality. *Labour Economics*, 17(4), 621-631.
- Caliendo, M., & Kopeinig, S. (2008). Some Practical Guidance for the Implementation of Propensity Score Matching. *Journal of Economic Surveys*, 22(1), 31-72.
- Conover, E., Khamis, M., & Pearlman, S. (2022). Job Quality and Labour Market Transitions: Evidence from Mexican Informal and Formal Workers. *The Journal of Development Studies*, 58(7), 1332-1348.
- De Hoyos, R., Martínez de la Calle, J. M., & Székely, M. (2010). Educación y movilidad social en México. En J. Serrano & F. Torche (Eds.), *Movilidad social en México: Población, desarrollo y crecimiento* (pp. 135-164). Centro de Estudios Espinosa Yglesias.
- Esquivel Hernández, G. (2008). ¿Es la política social una causa de la informalidad en México? *Ensayos Revista de Economía*, 27(1), 1-32.
- Fields, G. S. (1975). Rural-Urban Migration, Urban Unemployment and Underemployment, and Job-Search Activity in LDCs. *Journal of Development Economics*, 2(2), 165-187.

-
- Gallardo Del Ángel, R. (2019). An Analysis on Gross Flows of Workers in México. *Economía: Teoría y Práctica*, (50), 145-171.
- Gasparini, L., & Tornarolli, L. (2009). Labor Informality in Latin America and the Caribbean: Patterns and Trends from Household Survey Microdata. *Desarrollo y Sociedad*, (63), 13-80.
- Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., & Vermeersch, C. M. J. (2016). *Impact Evaluation in Practice* (2.^a ed.). Inter-American Development Bank; World Bank. <https://doi.org/10.1596/978-1-4648-0779-4>
- Gong, X., & van Soest, A. (2002). Wage Differentials and Mobility in the Urban Labour Market: A Panel Data Analysis for Mexico. *Labour Economics*, 9(4), 513-529.
- Gujarati, D. N., & Porter, D. C. (2009). *Basic Econometrics* (5.^a ed.). McGraw-Hill/Irwin.
- Harris, J. R., & Todaro, M. P. (1970). Migration, Unemployment and Development: A Two-Sector Analysis. *The American Economic Review*, 60(1), 126-142.
- Heinrich, C., Maffioli, A., & Vázquez, G. (2011). *A Primer for Applying Propensity-Score Matching* (Technical Note). Inter-American Development Bank.
- Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396), 945-960.
- Huesca Reynoso, L., & Padilla Arriola, N. (2012). Empleo, escolaridad y sector informal en la Frontera Norte de México y Chihuahua: expectativas de ocupación en la crisis. *Ensayos Revista de Economía*, 31(2), 57-86.
- Imbens, G. W., & Wooldridge, J. M. (2009). Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature*, 47(1), 5-86. Consultado el 29 de mayo de 2026, desde <http://www.jstor.org/stable/27647134>
- Karlin, S., & Taylor, H. M. (1975). *A First Course in Stochastic Processes* (2.^a ed.). Academic Press.
- Leuven, E., & Sianesi, B. (2003). PSMATCH2: Stata Module to Perform Full Mahalanobis and Propensity Score Matching, Common Support Graphing, and Covariate Imbalance Testing.
- Levy, S. (2008). *Good Intentions, Bad Outcomes: Social Policy, Informality, and Economic Growth in Mexico*. Brookings Institution Press.
- Levy, S., & Székely, M. (2016). ¿ Más escolaridad, menos informalidad? Un análisis de cohortes para México y América Latina. *El trimestre económico*, 83(332), 499-548.
- Maloney, W. F. (2004). Informality Revisited. *World Development*, 32(7), 1159-1178.
- Mendoza, J. E. (2010). El mercado laboral en la frontera norte de México: estructura y políticas de empleo. *Estudios Fronterizos*, 11(21), 9-42.

- Mincer, J. A. (1974). The Human Capital Earnings Function. En *Schooling, Experience, and Earnings* (pp. 83-96). National Bureau of Economic Research.
- Mortensen, D. T., & Pissarides, C. A. (1994). Job Creation and Job Destruction in the Theory of Unemployment. *The Review of Economic Studies*, 61(3), 397-415.
- Neyman, J. S. (1990). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9 [Translated and edited by D. M. Dabrowska and T. P. Speed from the Polish original published in 1923]. *Statistical Science*, 5(4), 465-472.
- OCDE. (2025). *Ampliar la protección social y combatir la informalidad en América Latina*. OECD Publishing.
- Perry, G. E., Maloney, W. F., Arias, O. S., Fajnzylber, P., Mason, A. D., & Saavedra-Chanduvi, J. (2007). *Informalidad: escape y exclusión*. Banco Mundial. <https://openknowledge.worldbank.org/handle/10986/6730>
- Pratap, S., & Quintin, E. (2006). Are Labor Markets Segmented in Developing Countries? A Semiparametric Approach. *European Economic Review*, 50(7), 1817-1841.
- Rauch, J. E. (1991). Modelling the Informal Sector Formally. *Journal of Development Economics*, 35(1), 33-47.
- Rodríguez Pérez, R. E., Castro Lugo, D., & Mendoza López, M. (2019). Desigualdad salarial y trabajo informal en regiones de México. *Región y Sociedad*, 31.
- Rosenbaum, P. R., & Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70(1), 41-55.
- Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 66(5), 688-701.
- Rubin, D. B. (2008). For Objective Causal Inference, Design Trumps Analysis. *The Annals of Applied Statistics*, 2(3), 808-840.
- Wooldridge, J. M. (2010). *Introducción a la econometría: un enfoque moderno*. Cengage Learning.