

Notas de Investigación

Reconstrucción de trayectorias
educativas, laborales y de
salud en Colombia:
construcción y potencial
analítico de una base de datos
longitudinal integrada

Investigadores(as) del proyecto

Camila Galindo (Universidad de Los Andes)

Jorge Leonardo Rodríguez-Arenas (Icfes)

Junio 2026





Notas de Investigación

Edición No. 23

Junio, 2026

Edición digital

Instituto Colombiano para la Evaluación de la Educación

Oficina de Gestión de Proyectos de Investigación

Calle 26 N. 69-76, Edificio Elemento, Torre II, piso 18, Bogotá,
D. C.

Teléfono: (601) 4841410

proyectosinvestigacion@icfes.gov.co

<https://www.icfes.gov.co>

Directora general

Elizabeth Blandón Bermúdez

Jefe de Oficina de Gestión de Proyectos de Investigación

Paola Guio Veloza

Autores

Camila Galindo

Jorge Leonardo Rodríguez-Arenas

Advertencia:

El contenido de este documento es el resultado de investigaciones y obras protegidas por la legislación nacional e internacional. No se autoriza su reproducción, utilización ni explotación a NINGÚN tercero. Solo se autoriza su uso para fines exclusivamente académicos. Esta información no podrá ser alterada, modificada o enmendada.

Citar este documento en estilo APA así:

Galindo, C., & Rodríguez-Arenas, J. L. (2026). *Reconstrucción de trayectorias educativas, laborales y de salud en Colombia: construcción y potencial analítico de una base de datos longitudinal integrada* (Nota de investigación No. 23). Instituto Colombiano para la Evaluación de la Educación (Icfes).

<https://www.icfes.gov.co/investigaciones/socializacion-y-apropiacion-social-del-conocimiento/>



Reconstrucción de trayectorias educativas, laborales y de salud en Colombia: construcción y potencial analítico de una base de datos longitudinal integrada*

Camila Galindo[‡]

Jorge Leonardo Rodríguez-Arenas[§]

Resumen

Esta nota documenta la construcción de una base de datos longitudinal integrada que vincula el universo de estudiantes que presentaron el Examen de Estado para el Ingreso a la Educación Superior (Examen Saber 11°) entre 2000 y 2019 con registros administrativos de matrícula y graduación en educación superior, victimización por conflicto armado, acceso a servicios de salud, y empleo formal. La base resultante cubre 8,253,202 estudiantes y permite reconstruir la trayectoria de los egresados de la educación media en Colombia desde su graduación de secundaria hasta su desempeño en el mercado laboral. La nota describe las fuentes de datos, la estrategia de vinculación de registros, y los criterios de depuración, y discute el potencial de esta base longitudinal para la agenda de investigación del Icfes en el marco de su misión de producir información para el mejoramiento de la calidad de la educación en Colombia.

Palabras clave: datos longitudinales, registros administrativos, educación superior, trayectorias educativas, Colombia.

*Agradecemos al Icfes por brindarnos el acceso a los datos en este artículo a través del proyecto de investigación interno “OFINVES-2024-03 - Conectividad”. Las opiniones contenidas en el presente documento son responsabilidad exclusiva de sus autores y no comprometen al Icfes ni a las entidades u organizaciones a las que estos pertenecen.

[‡] Universidad de los Andes, Colombia. Correo electrónico: camilagalindo@uniandes.edu.co.

[§] Oficina de Gestión de Proyectos de Investigación, Instituto Colombiano para la Evaluación de la Educación (Icfes). Correo electrónico: jlrodrigueza@icfes.gov.co.



Reconstructing Educational, Labor, and Health Trajectories in Colombia: Construction and Analytical Potential of an Integrated Longitudinal Database

Camila Galindo Jorge Leonardo Rodríguez-Arenas

Abstract

This note documents the construction of an integrated longitudinal database linking the universe of students who sat the High School Exit Exam (Saber 11°) between 2000 and 2019 to administrative records on higher education enrollment and graduation, armed conflict victimization, health service utilization, and formal employment. The resulting database covers 8,253,202 students and enables the systematic reconstruction, at national scale, of the life trajectories of Colombian secondary school graduates from the classroom to the labor market. We describe the data sources, the record linkage strategy, and the sample restrictions, and discuss the potential of this infrastructure for the Icfes research agenda in fulfillment of its statutory mission to produce information for the improvement of education quality in Colombia.

Keywords: longitudinal data, administrative records, higher education, educational trajectories, Colombia.



Contents

1	Introducción	1
2	Datos	2
2.1	Examen Saber 11°	2
2.2	Acceso a Educación Superior: SPADIES y SNIES	3
2.3	Graduación: SNIES Graduados y Saber Pro/TyT	4
2.4	Registro Único de Víctimas: UARIV	5
2.5	Salud y Mercado Laboral: RIPS y PILA	5
3	Tendencias de matrícula	6
4	Conclusiones	9
	Referencias	11

1 Introducción

El acceso a datos longitudinales de alta calidad que permitan seguir a los individuos a lo largo del tiempo y a través de múltiples dimensiones constituye un insumo central para la investigación en economía de la educación y para el diseño de política pública basada en evidencia. En esta literatura, la vinculación de registros administrativos ha ampliado de manera sustancial la capacidad de los investigadores para estimar los efectos de largo plazo de intervenciones educativas sobre resultados en el mercado laboral, la salud y la movilidad social (Chetty et al., 2014; Currie et al., 2020). La evidencia muestra que las decisiones de formación postsecundaria generan efectos que se materializan durante décadas y que solo pueden identificarse con datos que permitan observar a los individuos a lo largo de su ciclo de vida (Card et al., 2010).

En Colombia, la investigación en economía de la educación ha crecido en las últimas décadas, con contribuciones importantes sobre el efecto del financiamiento educativo en el acceso a la educación superior (Melguizo et al., 2016; Londoño-Vélez et al., 2020, 2023), la heterogeneidad de los retornos entre tipos de programas e instituciones (Camacho et al., 2016), y la relación entre la inserción en el mercado laboral formal y las trayectorias individuales (Khanna et al., 2021). Estos estudios han utilizado, en su mayoría, fuentes administrativas de manera aislada, como la base de matrícula de SPADIES, los registros de cotizantes de la PILA o los resultados del Examen Saber 11°. Sin embargo, la mayor parte de esta literatura se concentra en cohortes o dimensiones específicas, sin reconstruir de manera integral la trayectoria de los estudiantes desde la presentación del Saber 11 hasta el ingreso y la graduación en educación superior. Aunque para algunas cohortes se han vinculado registros de Saber 11 y PILA, hasta la fecha no existen estudios que permitan seguir de manera conjunta las trayectorias educativas, laborales y de salud a nivel individual para múltiples cohortes.

La base longitudinal integrada que describe esta nota vincula el universo del Examen Saber 11° con seis fuentes administrativas complementarias y cubre dimensiones que van desde el acceso y la graduación en educación superior hasta el uso de servicios de salud y el empleo formal. La base resultante, que incluye 8,253,202 estudiantes de las cohortes

2000–2019, permitirá responder preguntas de investigación de alta relevancia para el diseño de políticas públicas que hasta ahora no habían podido abordarse por la ausencia de datos articulados a lo largo de múltiples dimensiones. Así, esta base constituye un producto de investigación con alto potencial. Aunque fue construida originalmente para responder la pregunta de investigación planteada por [Galindo & Rodríguez \(2026\)](#), su alcance trasciende ese objetivo al conformar una infraestructura de datos que habilita una agenda mucho más amplia de investigación sobre las trayectorias educativas, laborales y de salud en Colombia.

2 Datos

La base longitudinal integrada articula el universo del Examen Saber 11° con distintas fuentes administrativas a través del algoritmo de vinculación de registros estandarizado del Icfes. Este algoritmo enlaza individuos entre dos bases de datos mediante hasta 14 criterios de coincidencia exacta basados en combinaciones de número de documento de identidad, nombre completo y fecha de nacimiento, priorizando los criterios que usan el documento de identificación.

2.1 Examen Saber 11°

La fuente primaria es el universo de estudiantes que presentaron el Examen Saber 11°, administrado por el Icfes, entre 2000 y 2019. Los microdatos crudos se organizan por año y semestre y se consolidan en un panel con un universo inicial de 10,977,735 registros, que incluye la cohorte de 1999 para efectos del proceso de identificación de personas descrito a continuación. De estos registros crudos, eliminamos estudiantes que presentan el examen en calidad de evaluados individuales.¹

Uno de los principales retos para vincular estos registros con dimensiones de largo plazo, como trabajo y salud, es que una proporción de los estudiantes presenta el Saber 11° con tarjeta de identidad, el documento válido para menores de edad, mientras que, al alcanzar la mayoría de edad, su documento cambia a la cédula de ciudadanía. Para resolver esta

¹Estos registros corresponden a estudiantes que presentan el Saber 11° sin estar vinculados a una entidad educativa.

transición en los documentos de identidad, construimos un identificador único de persona a partir del cruce entre tarjeta de identidad y cédula de ciudadanía de la Registraduría Nacional del Estado Civil. Además, restringimos la base a una sola observación de Saber 11° por estudiante, correspondiente a su última presentación, y eliminamos duplicados dentro de cada cohorte.² Después de aplicar estas restricciones, contamos con 10,843,535 registros, equivalentes al 98,8% del total original.

Para homogeneizar la información de desempeño en la prueba y las variables socioeconómicas disponibles a lo largo de las cohortes 2000–2019, excluimos a los estudiantes con valores faltantes en los puntajes estandarizados de matemáticas y lectura crítica, las dos áreas disponibles de manera consistente en todos los formatos y años del examen, así como en género, nivel educativo de la madre, edad y código municipal del establecimiento educativo. Esto elimina el 5,2% de los registros. También restringimos la edad al momento de presentar el Saber 11° al rango entre 14 y 24 años (456,619 registros eliminados, equivalentes al 4,2%). Por último, eliminamos a los estudiantes de establecimientos sin identificador válido (código DANE), sin clasificación institucional (oficial o privado) o sin coordenadas georreferenciadas (1,8%). En conjunto, estas restricciones eliminan el 12,4% de los registros del universo inicial y reducen la base a 9,617,914 estudiantes únicos.

2.2 Acceso a Educación Superior: SPADIES y SNIES

Los registros de matrícula en educación superior provienen de dos sistemas administrativos del Ministerio de Educación Nacional (MEN): el Sistema para la Prevención de la Deserción en Educación Superior (SPADIES), que cubre el período 1999–2015, y el Sistema Nacional de Información de la Educación Superior (SNIES Matrícula), que cubre 2016–2024. Ambos sistemas siguen al universo de estudiantes matriculados en programas de educación superior en Colombia³. Para cada estudiante, se guarda la información de hasta cuatro programas

²El 1,2% de los estudiantes presenta el Saber 11° más de una vez. El porcentaje de duplicados dentro de una misma cohorte representa el 0,05% del universo inicial.

³Antes de la vinculación, los registros correspondientes al Servicio Nacional de Aprendizaje (SENA) se excluyen del SNIES Matrícula. El SENA no está clasificado como una institución de educación superior acreditada ante el MEN sino como una entidad de formación para el trabajo y el desarrollo humano, por lo que su inclusión distorsionaría las medidas de acceso y graduación en educación superior que se construyen en este estudio. El SENA no reporta en SPADIES.

por orden de matrícula, junto con el código de la institución, el código SNIES del programa, el primer periodo de matrícula y el municipio del campus del programa.

Del universo de 9,818,321 estudiantes del Saber 11°, el 51,5% se empareja con al menos un registro de educación superior. De estos registros emparejados, eliminamos un 0,4% que reportan primera matrícula en educación superior antes del año del examen Saber 11 y 7,8% que reporta acceder cinco años después de presentar el Saber 11. La segunda restricción nos permite homogeneizar la ventana de acceso a educación superior para todas las cohortes incluidas. Además, eliminamos 0,1% de registros para los cuales no es posible recuperar el municipio del campus que ofrece el programa.

Las siguientes restricciones son específicas al proyecto de investigación ([Galindo & Rodríguez, 2026](#)). Primero, se eliminan los programas de modalidad virtual; es decir, estudiantes que solo se hayan matriculado en programas virtuales son considerados como no matriculados en educación superior (en promedio, el 1,1% de los estudiantes). A nivel de institución, eliminamos colegios con un promedio de menos de seis estudiantes que presenten el Saber 11° por año, y municipios en los que al menos una celda municipio-año contiene menos de cinco estudiantes (0,4% de los registros). La base resultante comprende 8,253,202 estudiantes de las cohortes 2000–2019 (75,2% de los registros crudos originales).

2.3 Graduación: SNIES Graduados y Saber Pro/TyT

Los resultados de graduación se miden con dos fuentes complementarias, ambas vinculadas al universo del Saber 11° mediante el algoritmo del Icfes. El registro de graduados del SNIES (*SNIES Graduados*), construido a partir de los registros de matrícula del SNIES y SPADIES y con cobertura de graduados entre 1999 y 2024, registra a todos los estudiantes que completaron un programa de educación superior ante el MEN. El 29,8% de todos los registros del Saber 11° entre 2000–2019 pueden ser vinculados al SNIES Graduados (el 35,5% de los registros finales en la base de la Sección [2.2](#)).

Para complementar los registros de graduación del SNIES, usamos como fuente secundaria el Examen Saber Pro y el Examen Saber TyT. Estos exámenes corresponden a las pruebas estandarizadas que presentan los estudiantes al haber completado al menos el 75% de los créditos de un programa de pregrado o técnico-tecnológico respectivamente, y están

disponibles desde 2006 y 2016 hasta 2024. El 27,0% de todos los registros del Saber 11° entre 2000–2019 pueden ser vinculados al Saber Pro o al Saber TyT (el 32,1% de los registros finales en la base de la Sección 2.2).

2.4 Registro Único de Víctimas: UARIV

El Registro Único de Víctimas (RUV), administrado por la Unidad para la Atención y Reparación Integral a las Víctimas (UARIV) y con cobertura de hechos victimizantes entre 1985 y 2025, permite identificar a los estudiantes que experimentaron hechos relacionados con el conflicto armado o el desplazamiento forzado. El RUV se vincula al universo del Saber 11° mediante el algoritmo del Icfes; el 12,4% de los estudiantes se empareja con un registro de la UARIV (el 14,8% de los registros finales en la base de la Sección 2.2). Para cada estudiante emparejado se conservan el tipo de hecho victimizante, la fecha de ocurrencia y el municipio del evento.

2.5 Salud y Mercado Laboral: RIPS y PILA

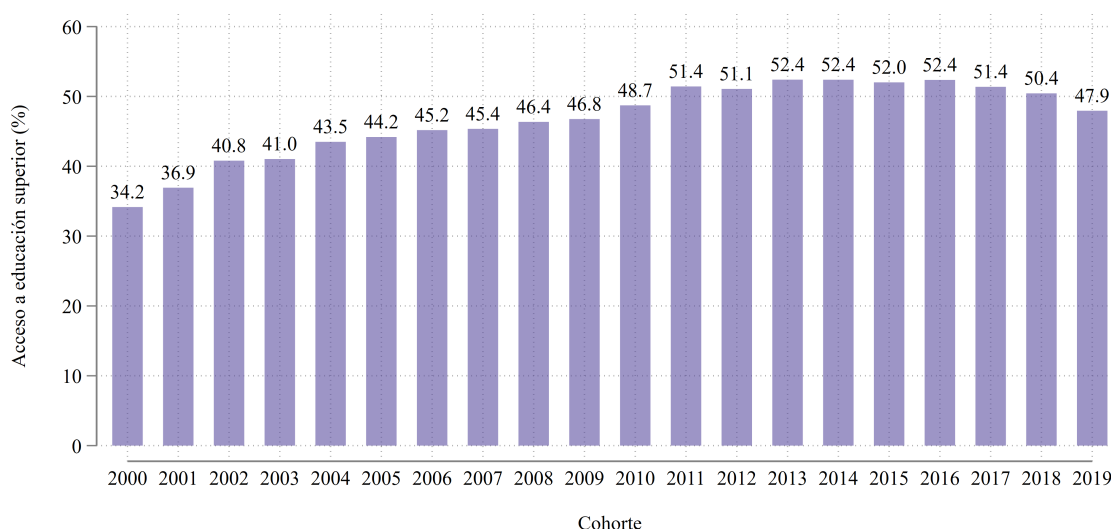
Los Registros Individuales de Prestación de Servicios de Salud (RIPS), administrados por el Ministerio de Salud y Protección Social con datos hasta 2024, capturan todos los servicios de salud prestados en el sistema formal junto con los códigos de diagnóstico CIE-10 y el municipio de prestación. Los registros RIPS están disponibles para las cohortes que presentaron el Saber 11° entre 2009 y 2019. El 93,1% de los estudiantes en la base de la Sección 2.2 cuenta con al menos un registro en RIPS.

La Planilla Integrada de Liquidación de Aportes (PILA), registro administrativo mensual de cotizaciones al sistema de seguridad social disponible hasta 2023, registra los episodios de empleo formal e incliye información sobre salarios, sector del empleador, tamaño de la firma y municipio del empleo. Los registros PILA pueden vincularse a los estudiantes que presentan el Saber 11° entre 2013 y 2019. El 55,9% de los estudiantes en la base de la Sección 2.2 se empareja con al menos un registro en PILA.

3 Tendencias de matrícula

Las siguientes figuras presentan tres tendencias descriptivas construidas a partir de la base longitudinal integrada (8,253,202 de estudiantes) que ilustran la evolución del acceso a la educación superior, la movilidad territorial de los estudiantes y la composición socioeconómica de las cohortes entre 2000 y 2019. Las medidas de acceso y migración se construyen dentro de una ventana homogénea de cinco años desde la presentación del Examen Saber 11°.

FIGURE 1: Acceso a educación superior por cohorte



Nota: La figura presenta el porcentaje de estudiantes que acceden a educación superior por cohorte. *Fuente:* Cálculos propios con base en la base longitudinal integrada del proyecto Conectividad, construida a partir del Examen Saber 11° del Icfes y los registros administrativos de matrícula en educación superior de SPADIES y SNIES Matrícula.

La Figura 1 muestra una expansión sostenida del acceso a la educación superior entre las cohortes 2000 y 2013, con la tasa de acceso pasando del 34,2% al 52,4%. A partir de 2013 la tasa se estabiliza en torno al 52%, con una leve caída en las cohortes 2018 y 2019, que puede reflejar cambios en la dinámica de acceso de las generaciones más recientes.

FIGURE 2: Migración para acceder a educación superior por cohorte



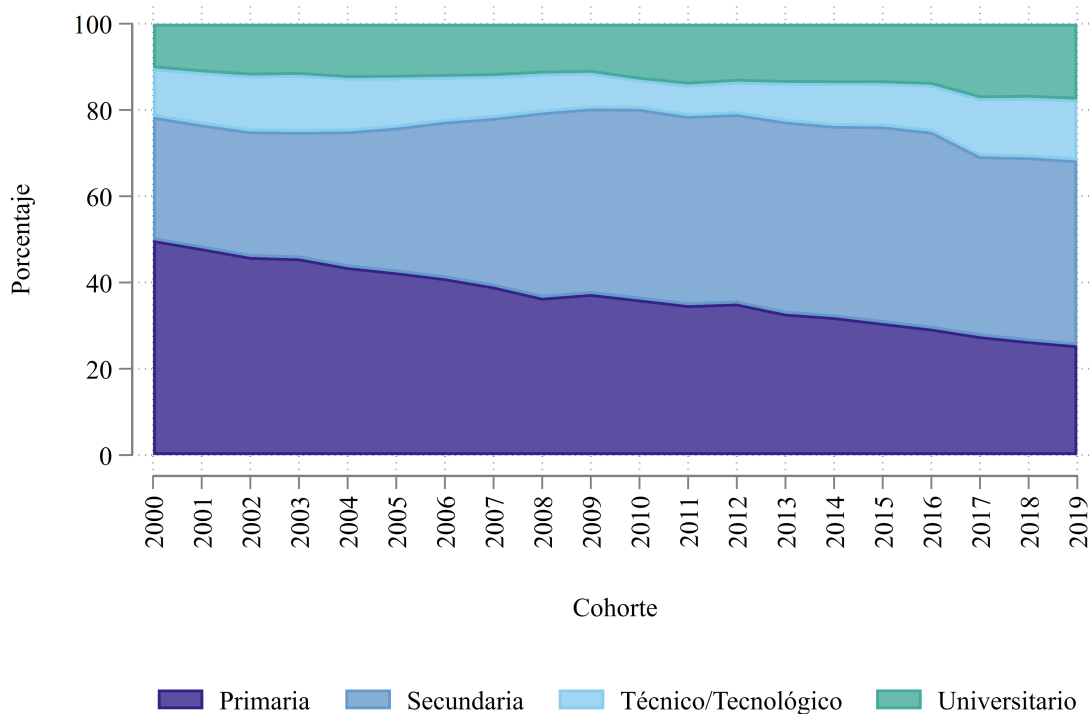
Nota: La figura presenta, entre quienes acceden a educación superior y no provienen de ciudades principales, el porcentaje de estudiantes que cursan educación superior en un municipio distinto al municipio donde presentaron Saber 11°. La medida de migración se construye comparando el municipio del establecimiento educativo de origen con el municipio del campus del primer programa de educación superior observado.

Fuente: Cálculos propios con base en la base longitudinal integrada del proyecto Conectividad, construida a partir del Examen Saber 11° del Icfes y los registros administrativos de matrícula en educación superior de SPADIES y SNIES Matrícula.

Para entender las dinámicas de migración de estudiantes que completaron secundaria fuera de las ciudades principales, donde típicamente se concentra la oferta de educación superior, calculamos la proporción de estudiantes que se matriculan en un municipio distinto al de su colegio. En principio, los estudiantes de ciudades intermedias, municipios mas pequeños, y de zonas rurales, tienen menos disponibilidad de programas. Condicional a matricularse en educación superior, la Figura 2 muestra que, la mayoría de estudiantes que no provienen de ciudades principales migra para acceder a programas de educación superior. A lo largo de todo el período analizado, más del 76% de quienes acceden a educación superior lo hacen en un municipio distinto al de su establecimiento educativo de origen. Esta cifra se mantiene relativamente estable durante la primera década, con valores cercanos al 81%, y exhibe una tendencia decreciente moderada a partir de 2014, llegando al 76,9% en 2019, tendencia que puede ser consistente con la expansión de sedes y campus satelitales de instituciones de educación superior en municipios intermedios durante ese período (Ferreyra

et al., 2022).

FIGURE 3: Distribución de la educación de la madre por cohorte



Nota: La figura presenta la distribución porcentual del máximo nivel educativo de la madre reportado por los estudiantes en el Examen Saber 11° por cohorte. Las categorías corresponden a primaria, secundaria, programa técnico o tecnológico y programa universitario. *Fuente:* Cálculos propios con base en el Examen Saber 11° del Icfes.

Por último, para caracterizar cambios de composición de los estudiantes a través del tiempo, usamos información sobre el nivel educativo de la madre. La Figura 3 revela una transformación en la composición educativa de los hogares de los estudiantes que presentan el Saber 11° entre 2000 y 2019. La proporción de estudiantes cuyas madres tienen primaria como máximo nivel educativo cae de manera continua a lo largo del período, mientras que las categorías de secundaria, técnico o tecnológico y universitario ganan participación progresivamente.

4 Conclusiones

La base de datos longitudinal documentada en esta nota permite reconstruir, por primera vez de manera sistemática y a escala nacional, la trayectoria de los egresados de la educación media en Colombia desde su desempeño en el Examen Saber 11° hasta su inserción en el mercado laboral formal, pasando por el acceso y la graduación en educación superior, la victimización por conflicto armado y el uso de servicios de salud. Esta articulación de fuentes no solo responde a las necesidades del proyecto que motivó su construcción, sino que habilita una agenda más amplia de investigación y evaluación de política pública que Colombia no ha podido abordar por falta de datos integrados a nivel individual.

En materia de política educativa, la base permite evaluar el efecto de intervenciones de acceso y financiamiento sobre resultados que trascienden la matrícula inmediata: graduación efectiva, calidad del empleo obtenido años después, salarios en el sector formal y movilidad territorial. Estudios previos han documentado los efectos de corto plazo de programas como Ser Pilo Paga y Generación E sobre el acceso a instituciones de calidad ([Londoño-Vélez et al., 2020, 2023](#)); la base aquí construida permite extender ese análisis a los resultados de largo plazo en salud y mercado laboral, y replicarlo para cohortes no cubiertas por esos estudios. De manera análoga, es posible evaluar el efecto de la apertura de nuevos campus o sedes de instituciones de educación superior sobre las decisiones de matrícula y las trayectorias subsiguientes de los estudiantes de una región, una pregunta directamente relevante para la política de descentralización de la oferta educativa del MEN.

Más allá de las intervenciones educativas en sentido estricto, la base es particularmente adecuada para estudiar cómo cambios en la conectividad y la accesibilidad territorial afectan las oportunidades de formación y las trayectorias de vida de los jóvenes. Inversiones en infraestructura de transporte, ya sea aérea, terrestre o conectividad digital, alteran el costo efectivo de acceder a la educación superior para estudiantes en municipios periféricos. La posibilidad de identificar, para cada estudiante, la distancia entre su colegio y la institución de educación superior más cercana, así como su eventual decisión de migrar para estudiar, hace que esta base sea un instrumento idóneo para evaluar ese tipo de intervenciones de conectividad en un país cuya geografía impone barreras de acceso considerables.

La vinculación con el Registro Único de Víctimas de la UARIV abre además una línea de investigación sobre los efectos del conflicto armado y el desplazamiento forzado en las trayectorias educativas y laborales, un tema central para la política de reparación y reintegración en Colombia. Y la cobertura de RIPS permite abordar preguntas sobre las consecuencias de las decisiones educativas en salud reproductiva y acceso a servicios, un vínculo que la literatura internacional ha identificado como relevante pero que rara vez puede estudiarse con datos de países en desarrollo de la magnitud y calidad de los aquí descritos. Estas preguntas centrales sobre cambios en conectividad, cambios en la oferta de educación superior, y su efecto sobre matrícula y graduación además de impactos en fertilidad, salud sexual y reproductiva, y participación y salarios en el mercado laboral formal son abordadas en [Galindo & Rodríguez \(2026\)](#).

En suma, la infraestructura de datos construida en el marco de este proyecto constituye un activo estratégico del Icfes que, en el marco de su misión legal de producir información para el mejoramiento de la calidad de la educación, permite nutrir una agenda de investigación rigurosa sobre las preguntas de política más urgentes en educación, movilidad social y bienestar en Colombia. La posibilidad de actualizar esta base con nuevas cohortes del Examen Saber 11° y con la extensión temporal de los registros de RIPS y PILA fortalece además su valor como instrumento de monitoreo y evaluación de largo plazo de los resultados del sistema educativo colombiano.

References

- Camacho, A., Messina, J., & Uribe, J. P. (2016). *The Expansion of Higher Education in Colombia: Bad Students or Bad Programs?* IDB Working Paper IDB-WP-707, Inter-American Development Bank.
- Card, D., Chetty, R., Feldstein, M., & Rouse, C. E. (2010). *Expanding Access to Administrative Data for Research in the United States*. White paper, National Science Foundation.
- Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, 104(9), 2593–2632.
- Currie, J., Kleven, H., & Zwieters, E. (2020). Technology and big data are changing economics: Mining text to track methods. *AEA Papers and Proceedings*, 110, 42–48.
- Ferreira, M. M., Galindo, C., & Urzúa, S. S. (2022). *Labor Market Effects of Short-Cycle Higher Education Programs: Lessons from Colombia*. Working Paper 30178, National Bureau of Economic Research.
- Galindo, C. & Rodríguez, J. (2026). Going the extra mile: Domestic air links and economic opportunity. Work in progress.
- Khanna, G., Medina, C., Nyshadham, A., Posso, C., & Tamayo, J. (2021). Job loss, credit, and crime in Colombia. *American Economic Review: Insights*, 3(1), 97–114.
- Londoño-Vélez, J., Rodríguez, C., & Sánchez, F. (2020). Upstream and downstream impacts of college merit-based financial aid for low-income students: Ser Pilo Paga in Colombia. *American Economic Journal: Economic Policy*, 12(2), 193–227.
- Londoño-Vélez, J., Rodríguez, C., Sánchez, F., & Álvarez Arango, L. E. (2023). *Financial Aid and Upward Mobility: Evidence from Colombia's Ser Pilo Paga*. Working Paper 31737, National Bureau of Economic Research.
- Melguizo, T., Sánchez, F., & Velasco, T. (2016). Credit for low-income students and access to and academic performance in higher education in Colombia: A regression discontinuity approach. *World Development*, 80, 61–77.