Educación | Icfes

Notas de Investigación

Análisis exploratorio de causalidad en variables medidas en pruebas Saber

Investigadores (as) del proyecto
Juan David Correa Granada
Carolina Márquez Narváez
Sebastián Quiñones Arredondo
Valentina Marín Galvis
Santiago Jiménez Villegas
David Cardona Franco

Asesores Icfes
Alexander Villegas Mendoza
Camilo Gaitán Cardozo

Proyecto Convocatorias de investigaciones del Icfes 2024

Análisis exploratorio de causalidad en variables medidas en pruebas Saber[*]

Juan David Correa Granada[2]
Carolina Márquez Narváez[3]
Sebastián Quiñones Arredondo[4]
Valentina Marín Galvis[5]
Santiago Jiménez Villegas[6]
David Cardona Franco[7]

Resumen

Esta investigación explora relaciones causales entre las variables medidas en el examen Saber 11° durante los periodos comprendidos entre el 2010 y el 2023. Se emplea un enfoque causal no paramétrico, técnicas de aprendizaje de modelos causales y estimación para estimar el efecto de distintas variables en el desempeño de los estudiantes.

Se parte de la recopilación, limpieza y normalización de datos, abarcando factores sociodemográficos y académicos de estudiantes, y entidades educativas. Luego, se emplean

[*] Las ideas, opiniones, tesis y argumentos expresados son de propiedad exclusiva de los autores y no representan el punto de vista del Icfes. Este proyecto de investigación recibió apoyo financiero por parte del Icfes en el marco de la estrategia para el fomento a la investigación: Convocatoria Piloto de Investigaciones Cortas para Grupos de Investigación 2024.
[2] Universidad Autónoma de Manizales, jcorrea@autonoma.edu.co.
[3] Universidad Autónoma de Manizales, carolina.marquezn@autonoma.edu.co.
[4] Universidad Autónoma de Manizales, sebastian.quinonesa@autonoma.edu.co.
[5] Universidad Autónoma de Manizales, valentina.maring@autonoma.edu.co.
[6] Universidad Autónoma de Manizales, santiago.jimenezv@autonoma.edu.co.
[7] Universidad Autónoma de Manizales, david.cardonaf@autonoma.edu.co.

algoritmos de aprendizaje de estructura para identificar relaciones causales y estadísticas entre variables. El modelo resultante se ajusta con conocimiento experto.

A partir de los datos y el modelo aprendido, se estiman parámetros causales de interés discriminados por variables relacionadas con el territorio, el estrato socioeconómico y el género de las personas. Se evidencian hallazgos que dan cuenta del impacto de factores como el acceso a Internet, la educación de los padres y las características de la institución educativa en el desempeño de los estudiantes en la prueba Saber 11°.

Palabras claves: Causalidad; DataIcfes; Pruebas Saber; Inferencia.

1. Introducción

Tradicionalmente, el análisis de datos se limita a la determinación de correlación entre variables (Chica-Gómez, 2010; Timarán-Pereira et al., 2019). Mientras que, estudios previos que buscan identificar relaciones causales están usualmente enmarcados en el uso de modelos lineales o logísticos combinados con supuestos difíciles de escrutar (Abadía, 2017; Barrios-Aguirre, 2021; Celis, 2012; ICFES, 2021).

A pesar de lo anterior, la identificación de relaciones de causalidad no está limitada a ningún modelo particular y la elección de supuestos para el análisis puede justificarse a partir de los mismos datos y conocimiento experto (Pearl, 2000). Dichas relaciones permiten una toma más robusta y efectiva de decisiones, así como la formulación de políticas efectivas. Es por esto que, en esta investigación se explora el uso de métodos de análisis causal, no paramétrico y con suposiciones guiadas por los datos, para soportar la toma de decisiones en pro de la calidad en educación, aún en ausencia de controles experimentales.

Específicamente, se exploran datos disponibles de los exámenes Saber 11° (Icfes, 2024), considerando los periodos 2010-2023. El análisis no sólo abarca los resultados de los estudiantes y las instituciones educativas, sino también datos personales y socioeconómicos, con el fin de establecer relaciones (causales) con los primeros.

Esta es la primera etapa de una investigación de largo plazo y plantea el acercamiento a modelos causales que incluyan variables como edad, lugar de residencia, etnia, condiciones económicas, condiciones familiares o instituciones de formación y el aprendizaje de los estudiantes, a partir de evidencia consignada en DataIcfes.

El análisis parte de la caracterización de los datos y la formulación de modelos causales de grafo (Pearl, 2000) a partir de técnicas de aprendizaje de estructura causal a partir de datos observacionales (Spirtes et al., 1995). Dichos modelos permiten formular y resolver preguntas sobre el efecto de algunas variables relevantes en el resultado de las pruebas Saber 11°, teniendo en cuenta factores territoriales, diferenciales e interseccionales.

2. Análisis

El análisis realizado se enmarca en dos tareas clásicas de inferencia causal: el *descubrimiento de la estructura* causal del problema y la *estimación de los efectos* causales que unas variables tienen en otras (Pearl, 2000; Spirtes et al., 1995). Ambas combinan un análisis cuantitativo de los datos y conocimiento cualitativo en forma de un modelo causal.

Se desarrollaron actividades de consecución, procesamiento y limpieza de los datos disponibles en el repositorio DataIcfes utilizando un algoritmo para aprender la estructura de un modelo causal. Este proceso ofrece un modelo parcial que es refinado a partir de conocimiento experto.

Finalmente, se definen parámetros causales de interés y se estiman utilizando técnicas de inferencia apropiadas.

2.1 Consecución de datos

Se descargan los datos de los Exámenes Saber 11° del repositorio DataIcfes (Icfes, 2024) para los periodos 2010-1 hasta 2023-2 con aproximadamente 7.866.000 registros.

2.2 Limpieza y procesamiento de datos

Uno de los retos que se presentan al procesar los datos objeto de estudio es la variabilidad del formato y las variables medidas. En esta etapa se realizaron las siguientes actividades:

- Detección de codificación y separador de los archivos.
- Inventario de las variables medidas en cada periodo.
- Normalización y homologación de valores en el dominio de las variables (incluyendo remoción de acentos, estandarización de la capitalización).
- Identificación de datos ausentes.
- Separación de datos indexables (como nombres de instituciones).
- Consolidación de columnas que cambiaron de nombre entre pruebas pero significan lo mismo (ej: tiene_serviciotv).

- Recodificación de valores largos.
- Combinación de los datos de los diferentes periodos.
- Siguiendo la metodología del análisis de valor agregado (ICFES, 2021), se utilizan las variables de recalificación de la prueba saber de los periodos 2012-1 a 2014-1 para ajustar modelos y recalificar los periodos 2010-1 a 2013-2 y hacerlos comparables con periodos posteriores.
- Descartar variables a partir de un análisis de porcentaje de datos faltantes en el contexto de todos los periodos considerados.

Estas transformaciones permiten reducir el tamaño de la base de datos de Saber 11° de ~13GB a ~3.1GB, manteniendo 71 variables que incluyen información sobre el colegio, información demográfica del estudiante, información de la familia y situación económica del estudiante, y los puntajes en la prueba.

2.3 Efectos de interés

A partir de una revisión bibliográfica se identifican variables que han sido previamente asociadas con el desempeño de los estudiantes en su programa educativo. Entre estas se encuentran variables como: nivel educativo de los padres, ingresos de los padres y nivel socieconómico, género, institución educativa y valor de la pensión, tecnología y acceso a internet, y ubicación territorial.

De manera correspondiente, planteamos preguntas que intentarán estimar a partir de los datos del repositorio y la metodología propuesta:

- ¿Cuál es el efecto del acceso a las TIC en los resultados de los estudiantes, según el contexto territorial?
- ¿Cuál es el efecto del acceso a internet en los resultados de las pruebas Saber 11° para las zonas rurales frente a las urbanas?
- ¿Cómo influyen los factores familiares en el desempeño de los estudiantes en las pruebas Saber 11°, según el género del estudiante y su condición socioeconómica?
- ¿Cuáles son las principales variables que explican las diferencias entre los resultados obtenidos por estudiantes de colegios privados y públicos en zonas urbanas?

2.4 Aprendizaje de la estructura y refinamiento de modelos causales

Para responder a estas preguntas, partimos la de construcción de modelos causales de grafo que representan la estructura causal de las variables, creados con base en la aplicación de algoritmos especializados para inferir dicha estructura a partir de datos observacionales. Específicamente utilizamos *Fast Causal Inference* (FCI) (Spirtes, 2001) para inferir la manera en que las variables de estudio pueden relacionarse causalmente. Dichas relaciones se representan a través de un grafo.

En línea con las preguntas planteadas, se estudia la estructura de los siguientes conjuntos de variables medidas en la prueba Saber 11°:

- **Variables incluidas en todos los modelos**: punt_c_naturales, punt_ingles, punt_lectura_critica, punt_matematicas, punt_sociales_ciudadanas, estu_genero, periodo, fami_estratovivienda, cole_area_ubicacion, cole_cod_depto_ubicacion y cole_cod_mcpio_ubicacion.
- **Modelo relacionado con acceso a tecnología**: fami_tienecomputar y fami_tieneinternet.
- **Modelo con variables relacionadas con el colegio**: cole_area_ubicacion, cole_bilingue, cole_calendario, cole_cod_depto_ubicacion, cole_cod_mcpio_ubicacion, cole_genero, cole_jornada, cole_naturaleza estu_nse_individual y estu_nse_establecimiento.
- **Modelo con variables de la situación familiar**: fami_educacionmadre, fami_educacionpadre, fami_estratovivienda, fami_ocupacionmadre, fami_ocupacionpadre, estu_nse_individual.

Los modelos parciales obtenidos con FCI se toman como base para la formulación de los modelos a utilizar en el proceso de inferencia. La Figura 1 muestra el modelo de grafo asociado con las variables que indican si un estudiante tenía acceso a internet y a un computador al momento de presentar la prueba. El modelo también incorpora covariables asociadas a estas dos variables y otras que dan cuenta del territorio en el que se encuentra el colegio y el género de quien presenta la prueba. La variable U representa otros factores

exógenos no medidos que afectan tanto el estrato de la vivienda como el área (urbana o rural) en la que se encuentra la institución educativa. El grafo muestra explícitamente la variable de resultado de la competencia en Matemáticas, pero el mismo modelo se utiliza para analizar las demás competencias.

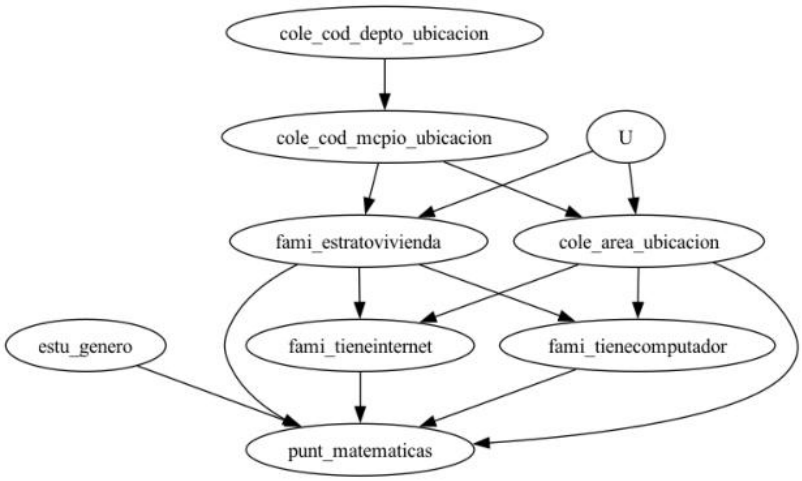

Figura 1. Modelo causal de grafo sobre acceso a internet y computador.

La Figura 2 presenta el modelo que da cuenta de factores familiares relevantes para los puntajes obtenidos. La variable U representa factores no medidos que causan la educación de los padres. Se asume que la educación de los padres afecta sus respectivas ocupaciones, y estas a su vez determinan el nivel socio-económico del estudiante. Se asume, en principio, que todas las variables pueden afectar el resultado obtenido en cada competencia.

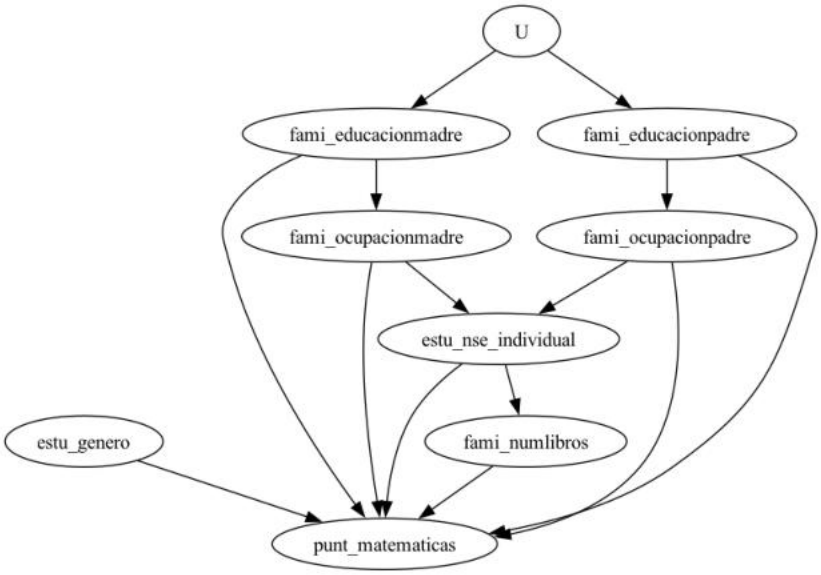

Figura 2. Modelo causal de grafo que relaciona las variables que representan factores familiares.

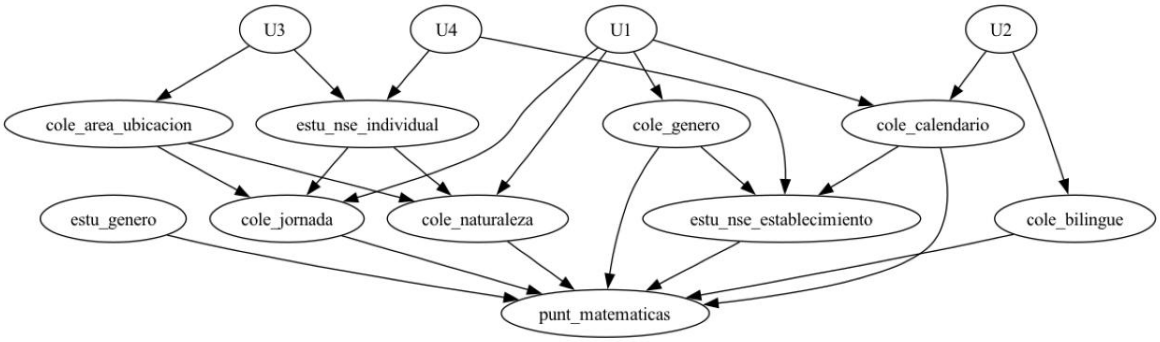

Figura 3. Modelo causal de grafo para las variables relacionadas con la institución educativa.

La Figura 3 presenta un modelo con variables relacionadas con la institución educativa y cómo estas se relacionan causalmente. La presencia de las variables *U1, U2, U3* y *U4*, indica factores no medidos (confusores) que afectan conjuntamente a las variables a las que apuntan.

2.5 Estimación de parámetros de interés

En esta sección se presentan efectos causales en términos del incremento esperado del puntaje de cada competencia ante la presencia de una variable. Para estimar estos efectos se utilizó

la librería DoWhy (Sharma & Kiciman, 2020), específicamente las técnicas de Propensity Score (PS) (Hernan & Robins, 2006) aproximaciones con regresión lineal.

Respecto a la pregunta planteada del efecto del acceso a la tecnología, la Figura 4 muestra los incrementos atribuibles al acceso a internet. Se puede observar que dicho acceso es un factor positivo en todas competencias, y el beneficio no es significativamente distinto entre personas de género masculino y femenino.

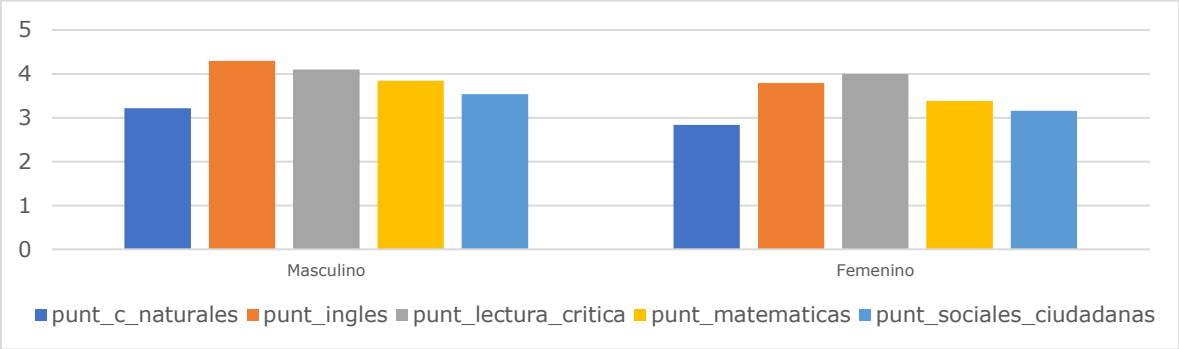

Figura 4. Incremento esperado en el puntaje de las competencias atribuible al acceso a internet.

En la Figura 5 se describen los incrementos atribuibles a tener acceso a un computador, discriminados por departamento. Se puede observar que el beneficio de tener computador no se aprovecha de manera uniforme en todo el territorio. Es decir, aún para aquellas personas que tienen acceso a un computador en departamentos como Chocó o Guaviare, este resulta menos beneficioso que para una persona en Atlántico o Nariño.

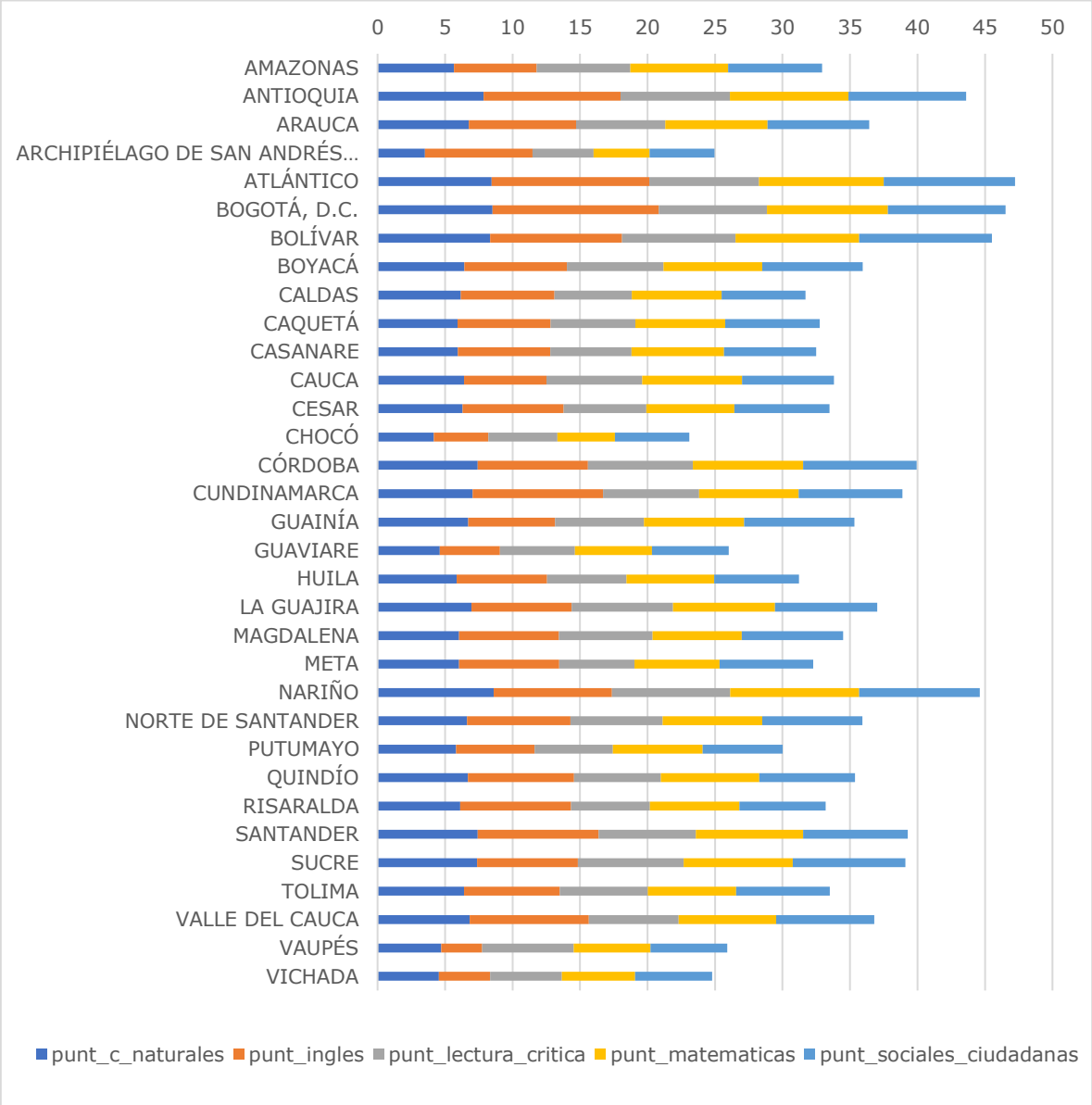

Figura 5. Incremento esperado en el puntaje de las competencias atribuible al acceso a computador, discriminado por departamento.

En cuanto a variables familiares, se estimó el incremento esperado en los puntajes atribuible a la educación de la madre y el padre (Figura 6). Se encontró que los años de formación de la madre tienen un efecto mayor a los del padre. Y que la competencia que más se beneficia de la educación de los padres es Inglés.

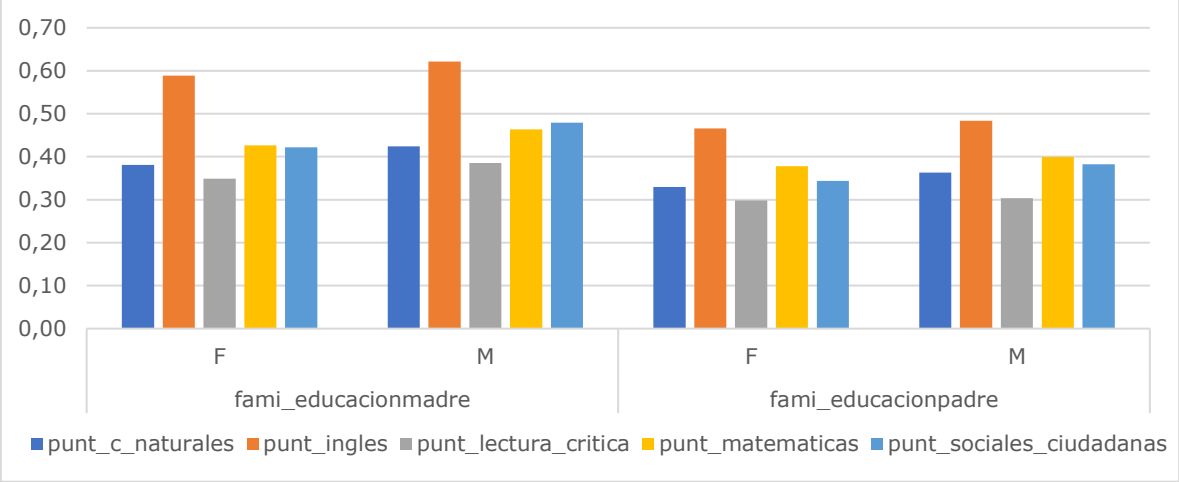

Figura 6. Incremento esperado en el puntaje de las competencias atribuible a cada año de educación recibida por parte de la madre/padre, discriminado por el género del estudiante.

De manera similar, se encontró (Figura 7) que la competencia de Inglés es la que más se beneficia de un mayor nivel socioeconómico (NSE) de la persona. También se evidencia que el efecto del NSE no varía con el género.

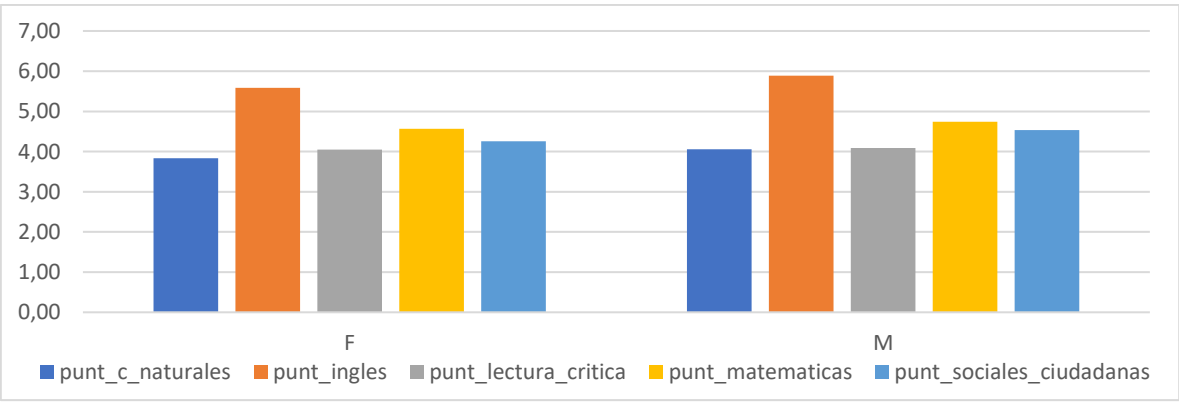

Figura 7. Incremento esperado en el puntaje de las competencias atribuible a cada nivel socioeconómico adicional (nse_individual), discriminado por el género del estudiante.

Para la pregunta relacionada con los factores asociados a las instituciones educativas en las áreas rurales y urbanas, se evaluó el impacto atribuible a variables como si la institución es mixta o no, si tiene calendario distinto a A o si es oficial o no. La Figura 8 relaciona el impacto de dichas variables discriminado por área.

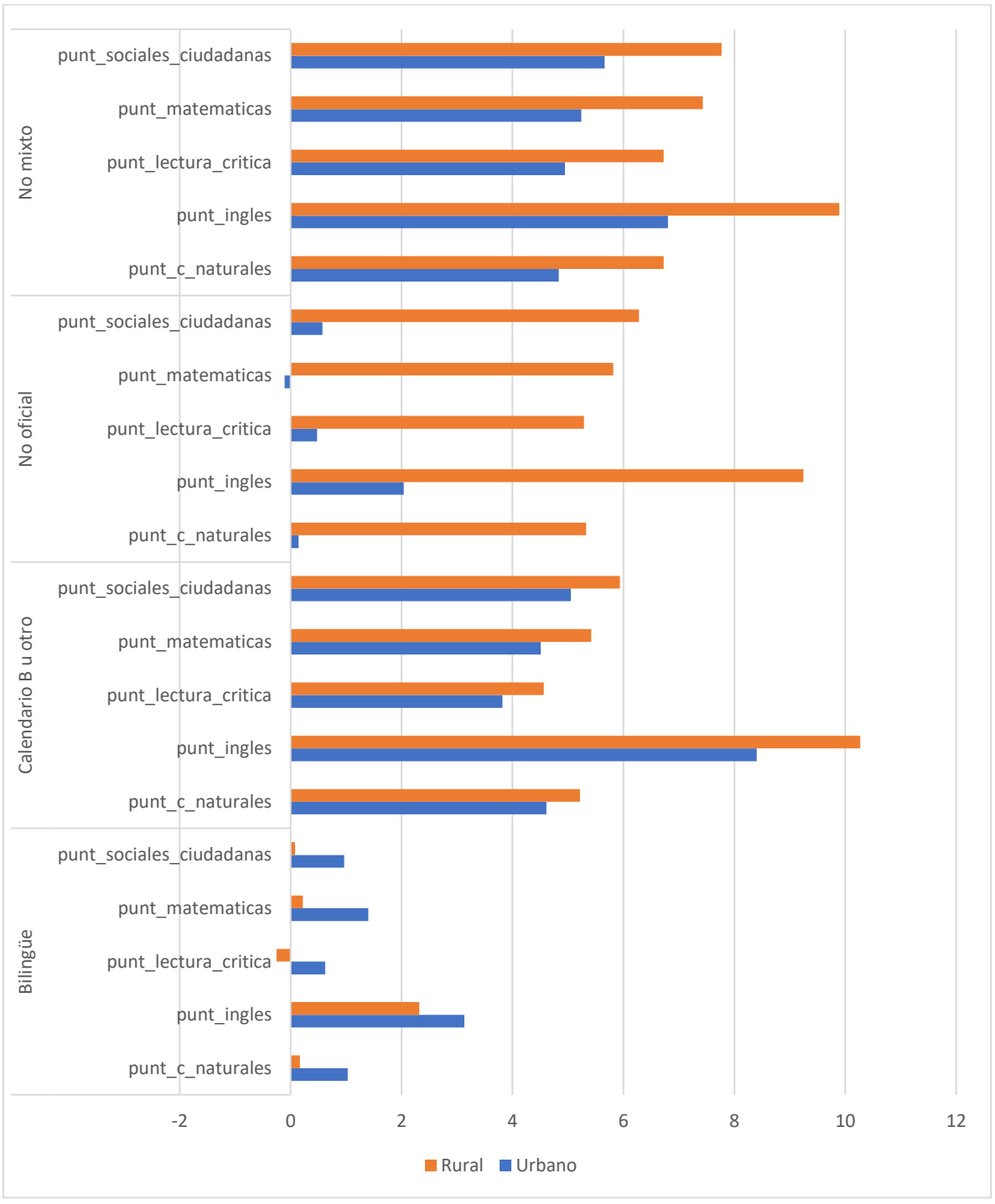

Figura 8. Incremento esperado en el puntaje de las competencias atribuible a variables del colegio, discriminados por zonas rurales y urbanas.

3. Reflexiones finales

En esta investigación corta hemos explorado el uso de técnicas de inferencia causal para obtener información valiosa sobre el efecto que variables como el acceso a internet, la educación de los padres o factores asociados a la institución educativa tienen en el desempeño de los estudiantes que presentan la prueba Saber 11°.

Varios de estos efectos pueden medirse teniendo en cuenta el territorio, el área (rural o urbana) y factores como el género y el estrato socioeconómico. Este ejercicio apenas constituye una prueba de concepto y deja ver la posibilidad de incorporar otros factores medidos en las pruebas como condiciones de discapacidad, étnicas, y otras variables asociadas al territorio, que puedan brindar información más detallada y útil para quienes toman decisiones en los territorios con el fin de mejorar la calidad de la educación.

Además de realizar un estudio más amplio en términos de variables, queda la puerta abierta para medir otro tipo de parámetros como los efectos directos e indirectos. También, se considera el uso de técnicas que añadan robustez a la estimación, como lo son aquellas que buscan refutar los modelos propuestos. Finalmente, este ejercicio y las mejorar posteriores pueden ser aplicados a datos de otras pruebas como Saber Pro y Saber TyT.

4. Referencias

Abadía, L. K., & Bernal, G. (2017). A widening gap? A gender-based analysis of performance on the Colombian high school exit examination. Revista de economía del Rosario, 20(1), 5-31.

Amit Sharma, Emre Kiciman. DoWhy: An End-to-End Library for Causal Inference. 2020. https://arxiv.org/abs/2011.04216.

Barrios Aguirre, F., Forero, D. A., Castellanos Saavedra, M. P., & Mora Malagón, S. Y. (2021). The impact of computer and internet at home on academic results of the Saber 11° national exam in colombia. SAGE Open, 11(3), 21582440211040810.

Celis, M. T., Jimenez, O., & Jaramillo, J. F. (2012). ¿Cuál es la brecha de la calidad educativa en Colombia en la educación media y en la superior? Estudios sobre calidad de la educación en Colombia, 67-98.

Chica Gómez, S. M., Galvis Gutiérrez, D. M., & Ramírez Hassan, A. (2010). Determinantes del rendimiento académico en Colombia. Pruebas ICFES-Saber 11°o, 2009. Revista Universidad EAFIT, 46(160), 48-72.

Chickering, D. M. (2002). Optimal structure identification with greedy search. Journal of machine learning research, 3(Nov):507–554.

Hernán, M. A., Lanoy, E., Costagliola, D., & Robins, J. M. (2006). Comparison of dynamic treatment regimes via inverse probability weighting. Basic & clinical pharmacology & toxicology, 98(3), 237-242.

Instituto Colombiano para la Evaluación de la Educación – Icfes (2021). Medición de los efectos de la educación superior en Colombia sobre el aprendizaje estudiantil - Informe Técnico. Bogotá D.C.

Instituto Colombiano para la Evaluación de la Educación – Icfes (marzo de 2024). DataIcfes: Repositorio de Datos Abiertos del Icfes. 04. Saber 11° [Conjunto de datos]. Recuperado de https://rb.gy/w4bo58.

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. Psychological methods, 15(4), 309.

Pearl, J. (2009). Causality. Cambridge university press.

Spirtes, Peter; Meek, Christopher; and Richardson, Thomas. Causal inference in the presence of latent variables and selection bias. In Proceedings of the Eleventh conference on Uncertainty in artificial intelligence, pages 499–506, 1995.

Spirtes, P. (2001, January). An anytime algorithm for causal inference. In *International Workshop on Artificial Intelligence and Statistics* (pp. 278-285). PMLR.