Emparejamiento selectivo: Medición y tendencias en América Latina y el Caribe

Albina, Iván

Enero, 2026

Apéndice Online

Tabla A.1 Censos disponibles por país – década.

País	Década 1960	Década 1970	Década 1980	Década 1990	Década 2000	Década 2010
Argentina	-	1970	1980	1991	2001	2010
Bolivia	-	1976	-	1992	2001	2012
Brasil	1960	1970	1980	1991	2000	2010**
Chile	1960*	1970	1982	1992	2002	2017
Colombia	1964*	1973	1985	1993	2005	-
Costa Rica	1963*	1973	1984	-	2000	2011
Dom. Rep.	1960*	-	1981	-	2002	2010
Ecuador	1962*	1974	1982	1990	2001	2010
El Salvador	-	-	-	1992	2007	-
Guatemala	1964	1973	1981	1994	2002	-
Haití	-	1971	1982	-	2003	-
Honduras	1961*	1974	1988	-	2001	-
Jamaica	-	-	1982	1991	2001	-
México	1960*	1970		1990/1995	2000/2005	2010/2015
Nicaragua	-	1971	-	1995	2005	-
Panamá	1960	1970	1980	1990	2000	2010
Paraguay	1962	1972	1982	1992	2002	-
Perú	-	-	-	1993	2007	2017
Puerto Rico	-	1970	1980	1990**	2000**/2005**	2010
Trinidad y Tobago	-	-	1980	1990	2000	2011
Uruguay	1963	1975	1985	1996	2006	2011**
Venezuela	-	1971	1981	1990	2001	-
Total	6	17	17	17	21	11

Nota: la tabla presenta el año de censo correspondiente a cada década y país. Los años marcados con * son censos en los cuales no es posible vincular a los habitantes del hogar. Aquellos marcados con ** no tienen información de los años de escolaridad.

Fuente: elaboración propia en base a IPUMS (2020).

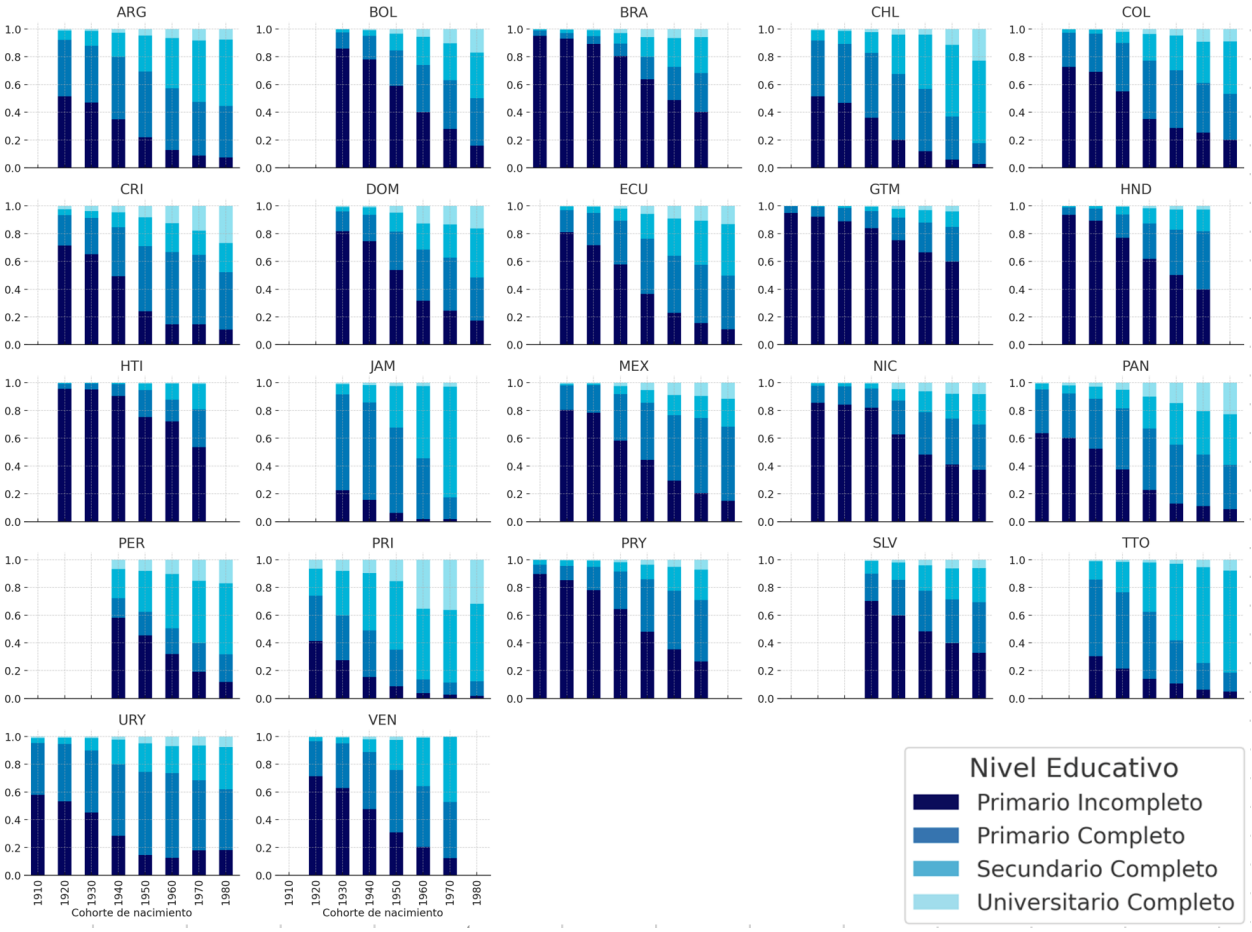
Tabla A2 – Estadísticas descriptivas período completo

Panel A. En pareja												
Medias	1960		1970		1980		1990		2000		2010	
muestrales	Mujeres	Hombres	Mujeres	Hombres	Mujeres	Hombres	Mujeres	Hombres	Mujeres	Hombres	Mujeres	Hombres
Share (%)	64.92	64.06	63.86	65.63	61.02	63.07	60.13	61.31	58.08	59.07	52.88	51.42
Edad	34.16	35.13	34.21	35.10	33.95	34.77	34.20	34.99	34.73	35.48	35.11	35.71
Años de escolaridad	3.08	3.53	3.84	4.51	5.89	6.42	6.72	7.24	7.69	7.95	10.25	10.14
PRII (%)	78.14	76.02	68.32	63.29	46.43	42.01	37.37	32.45	31.15	27.69	13.42	12.10
PRIC (%)	18.30	18.84	24.19	26.58	37.05	39.76	40.73	44.26	36.92	40.27	34.21	37.02
SEC (%)	3.14	3.49	6.16	7.25	13.77	13.57	18.24	17.88	26.38	25.81	37.27	38.06
UNIC (%)	0.42	1.65	1.33	2.88	2.75	4.66	3.67	5.41	5.54	6.23	15.11	12.82
Obs.	1424584	1421867	2802242	2778991	3613742	3660534	3954365	3853054	5873289	5636849	3119801	2920138
Panel B. Solteros												
Medias	1960		1970		1980		1990		2000		2010	
muestrales	Mujeres	Hombres	Mujeres	Hombres	Mujeres	Hombres	Mujeres	Hombres	Mujeres	Hombres	Mujeres	Hombres
Share (%)	35.08	35.94	36.14	34.37	38.98	36.93	39.87	38.69	41.92	40.93	47.12	48.58
Edad	33.70	32.30	33.64	32.26	33.06	31.80	33.14	31.99	33.55	32.48	33.64	32.92
Años de escolaridad	3.62	3.69	4.26	4.63	6.49	6.53	7.56	7.58	8.65	8.35	10.93	10.38
PRII (%)	72.30	74.13	63.66	61.88	41.26	40.77	31.20	30.12	25.53	26.01	11.42	12.47
PRIC (%)	21.99	20.30	26.70	27.02	37.25	39.48	38.72	42.45	33.04	37.13	28.18	32.42
SEC (%)	4.85	3.77	7.95	8.15	17.50	15.15	24.32	21.41	32.75	29.25	41.37	40.77
UNIC (%)	0.87	1.80	1.69	2.95	3.99	4.59	5.75	6.02	8.69	7.61	19.03	14.33
Obs.	528257	519202	1190928	1150934	1627503	1457335	2012951	1818742	3377285	3021956	2198090	1966669

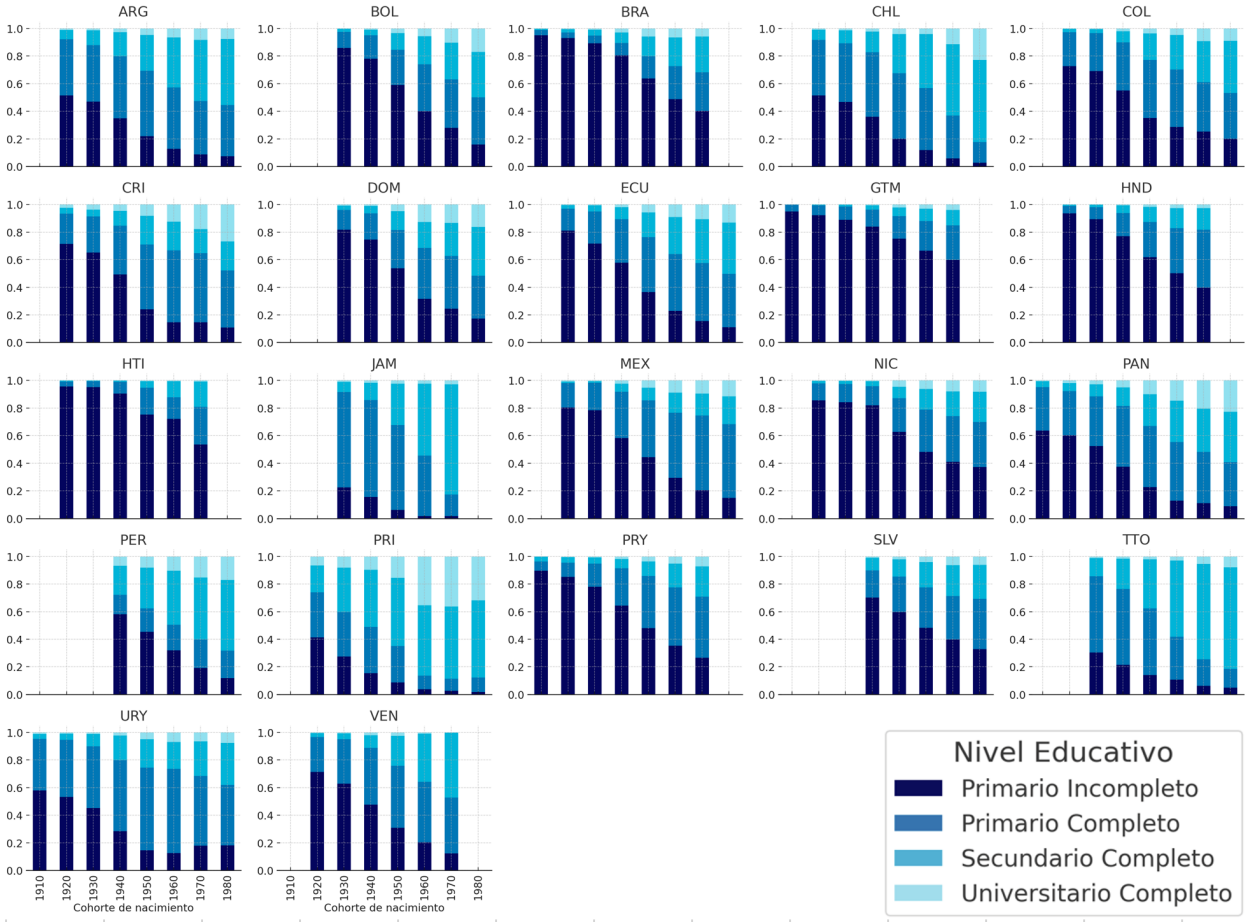
Nota: La tabla presenta el promedio regional de medias muestrales de edad y educación, así como también el total de observaciones para cada década censo. Se divide a la muestra entre individuos solteros y aquellos en pareja. Es importante destacar que los países incluidos en cada década censo pueden variar dependiendo de la disponibilidad de datos para cada década y país.
Fuente: IPUMS-I (2020)

<u>Figura A1</u> - Distribución educativa de hombres por país y cohorte de nacimiento.

Panel A. Hombres

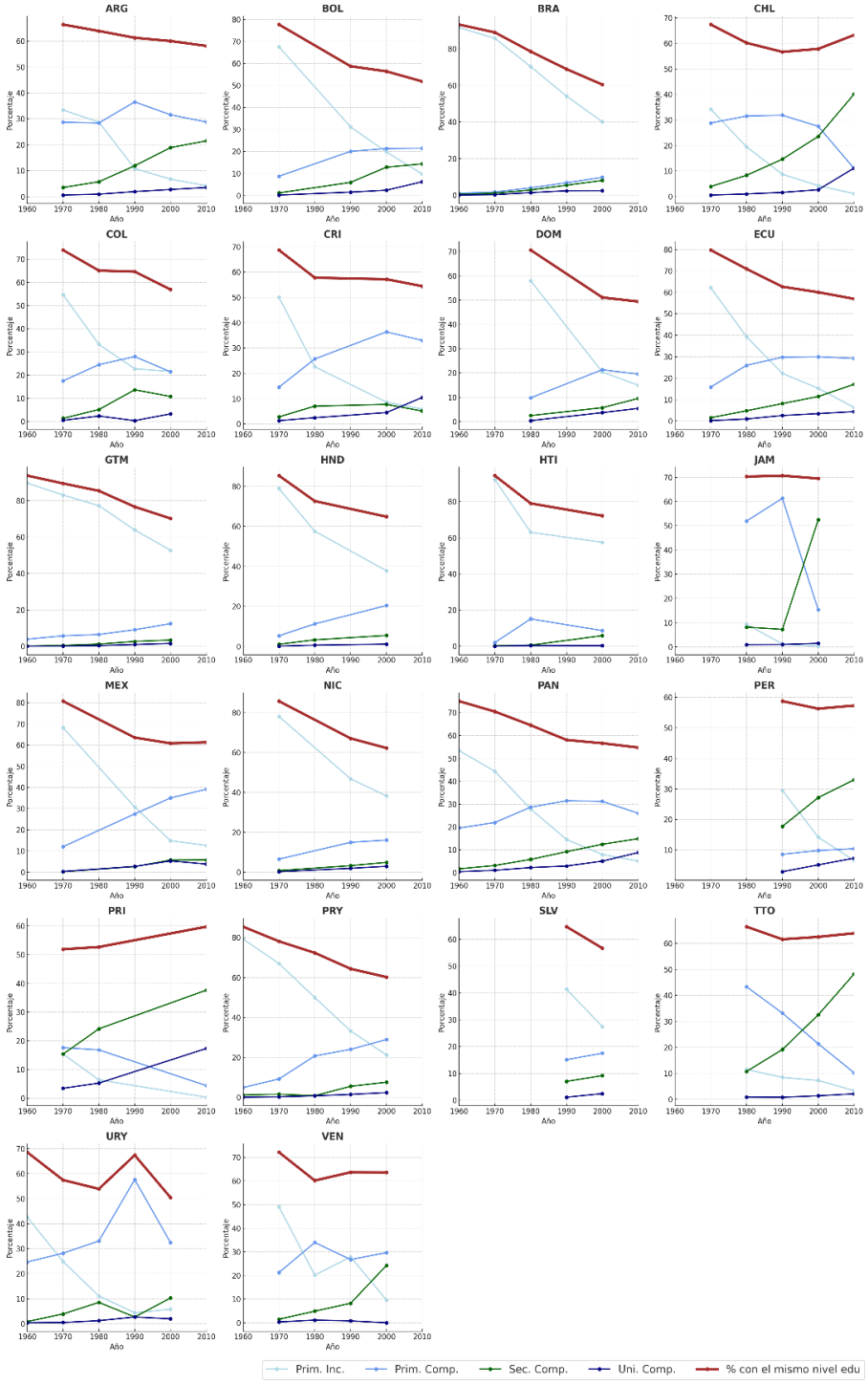


Panel B. Mujeres



Nota: cada panel muestra el porcentaje de hombres según máximo nivel educativo alcanzado, por país por década de censo. En el primer panel, se incluyen hombres de entre 25 y 45 años; en el segundo, mujeres de ese mismo rango etario.
Fuente: IPUMS-I (2020)

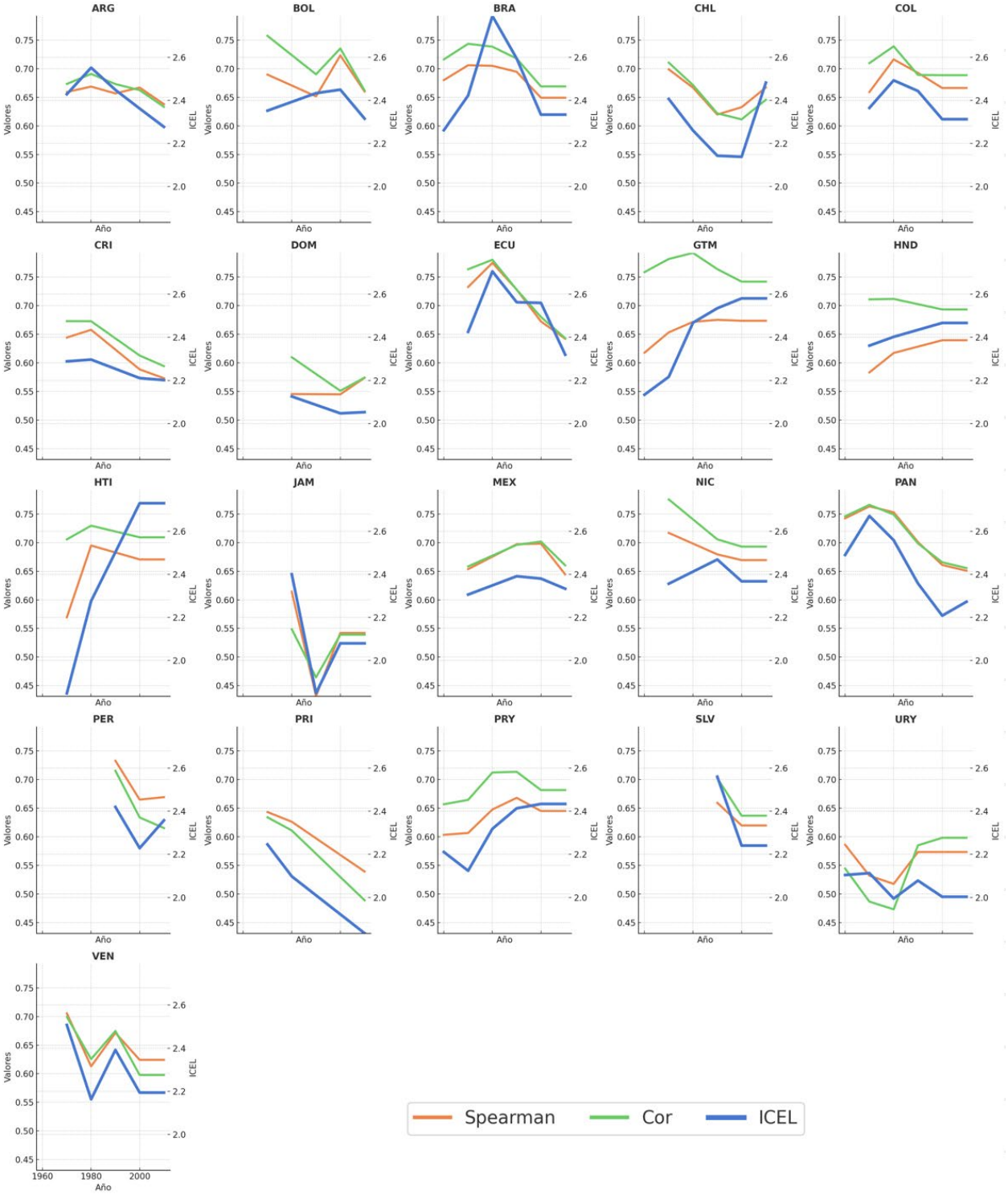
Figura A2. Porcentaje de parejas por nivel educativo – por país y por década de censo



Nota: esta figura presenta el porcentaje de parejas con el mismo nivel educativo, desglosado por nivel educativo y para el total de la población. Los países incluidos pueden variar según la disponibilidad de censos.

Fuente: IPUMS-I (2020).

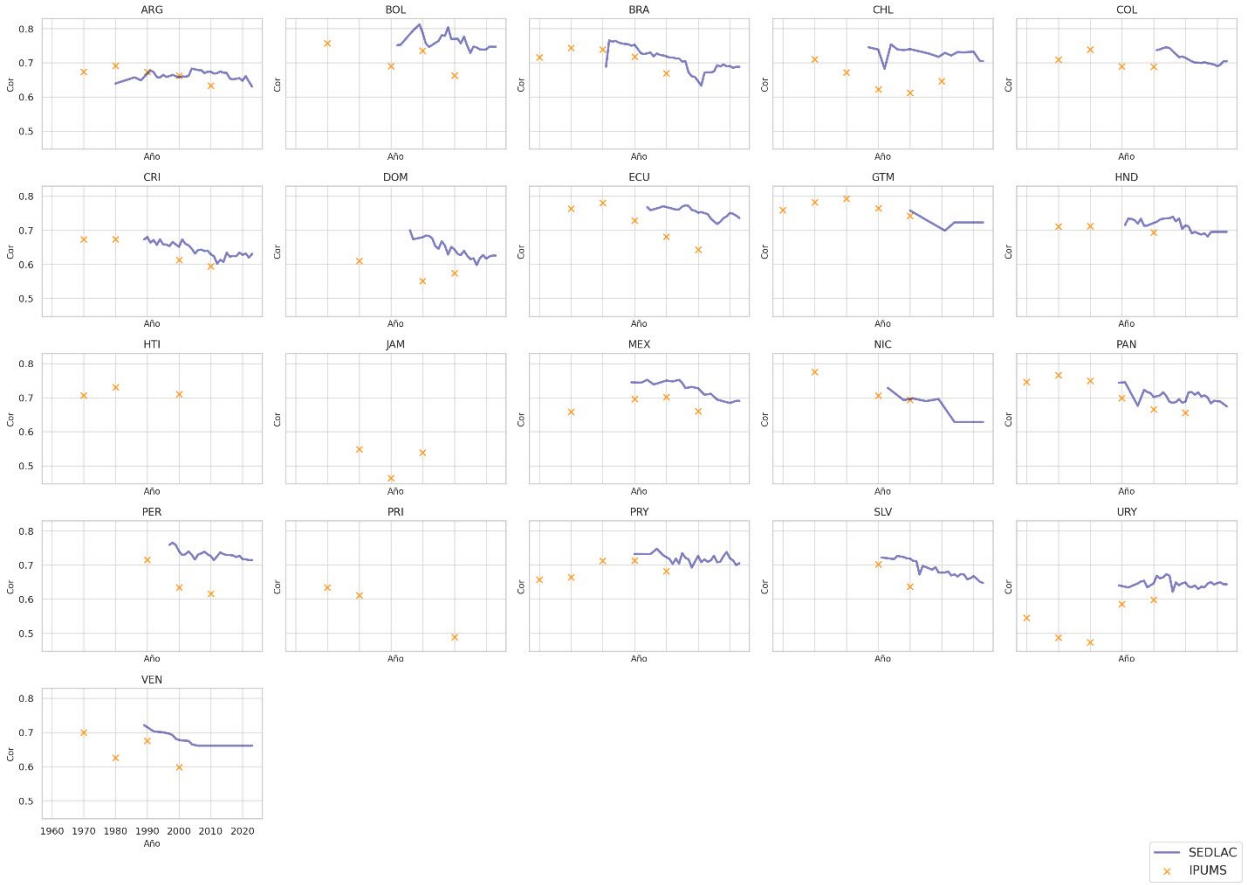
Figura A3. Evolución del ES según indicadores alternativos:



Nota: Cada panel del gráfico muestra la evolución de los indicadores de correlación de Spearman, Pearson y del ICEL para cada uno de los años con información disponible.

Fuente: Elaboración propia en base a IPUMS (2020).

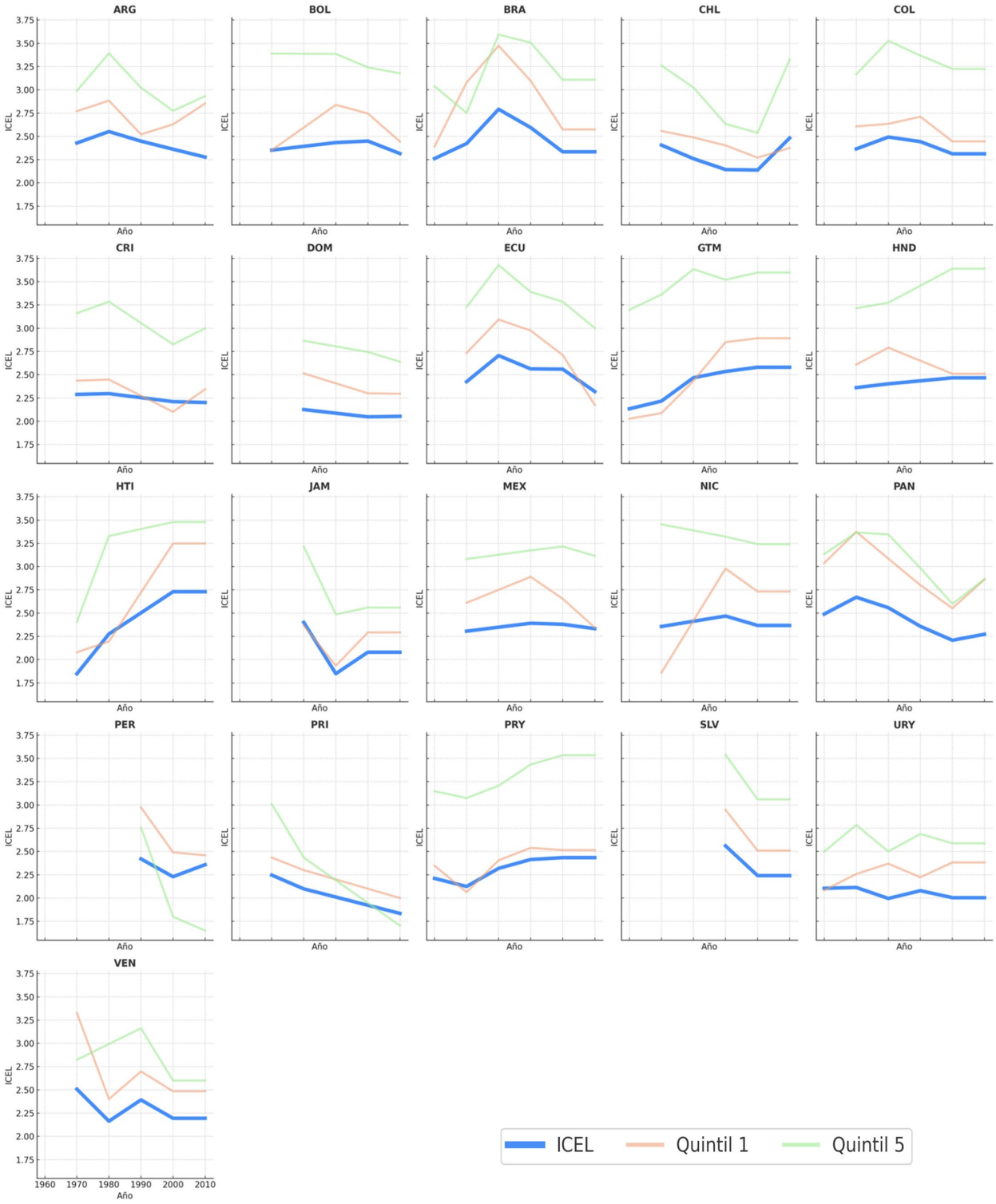
Figura A4. Coeficiente de correlación de Pearson según fuentes alternativas



Nota: este gráfico presenta los valores del Coeficiente de Correlación Lineal de Pearson utilizando estimaciones provenientes de dos fuentes de datos alternativas. Los puntos representan los valores provenientes de las estimaciones de IPUMS. Las líneas, los valores de SEDLAC (CEDLAS y Banco Mundial).

Fuente: Elaboración propia en base IPUMS (2020) y a SEDLAC (2024).

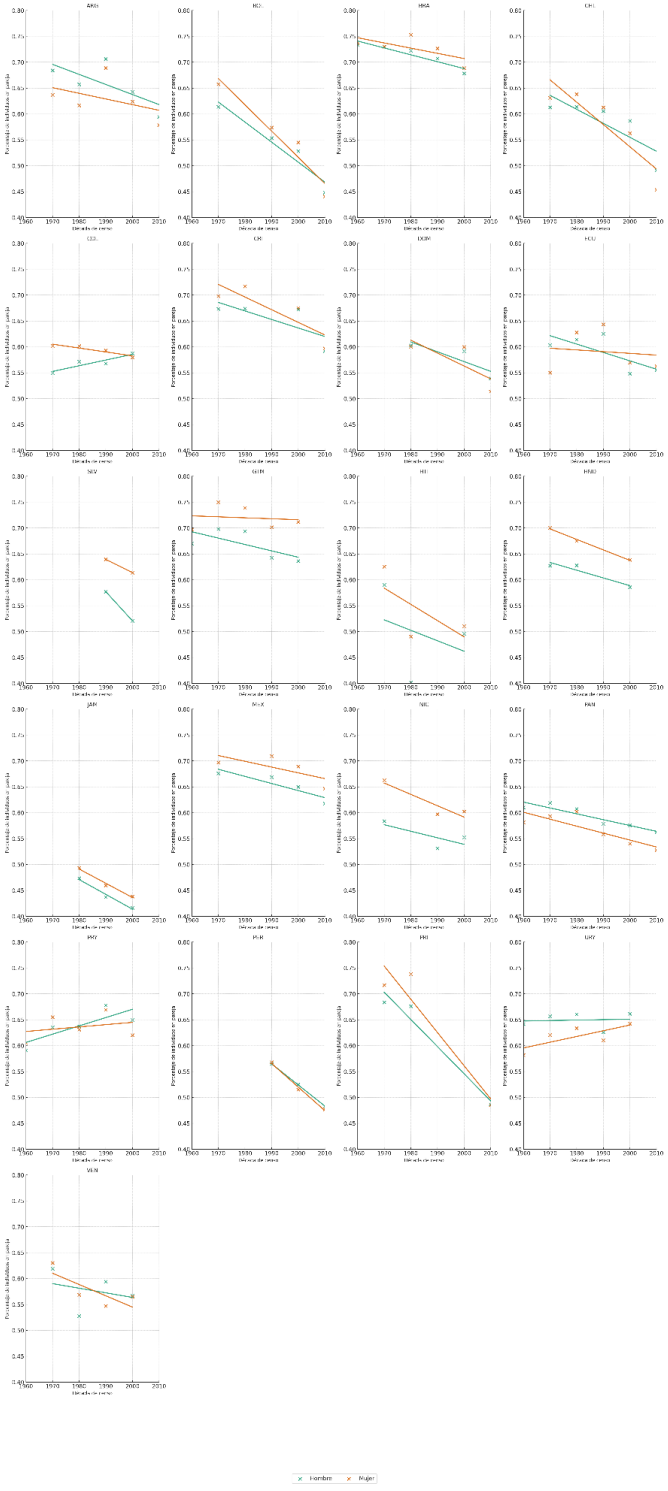
Figura A5. Evolución del ICEL



Nota: Cada panel del gráfico muestra la evolución de ICEL agregado, así como el ICEL Q1 y Q5 para cada uno de los años con información disponible.

Fuente: Elaboración propia en base a IPUMS (2020).

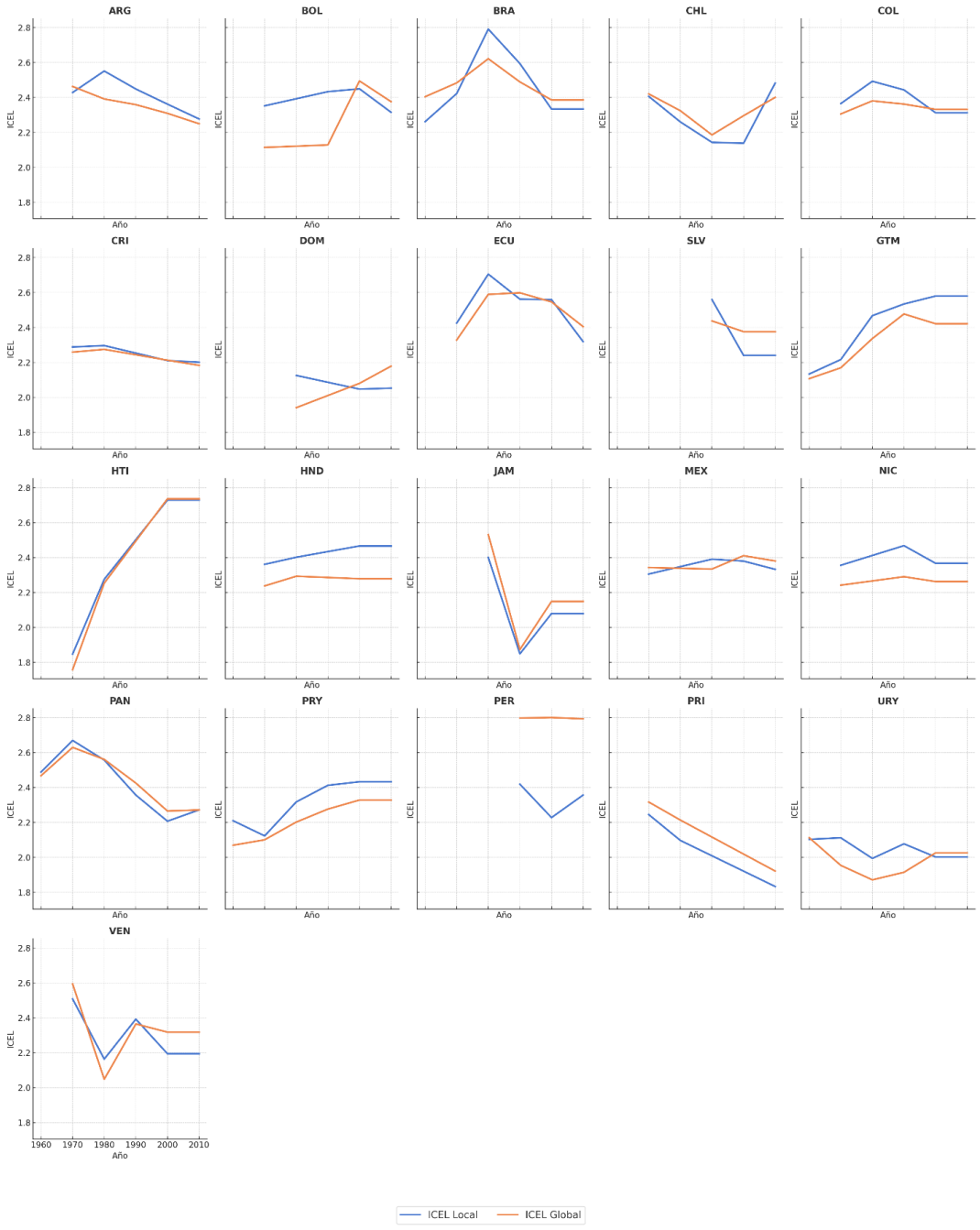
Figura A6. Porcentaje de individuos en pareja, por país y censo.



Nota: la figura presenta el share de individuos de entre 25 y 45 años en pareja en relación al total de la población.

Fuente: elaboración propia en base a IPUMS (2020)

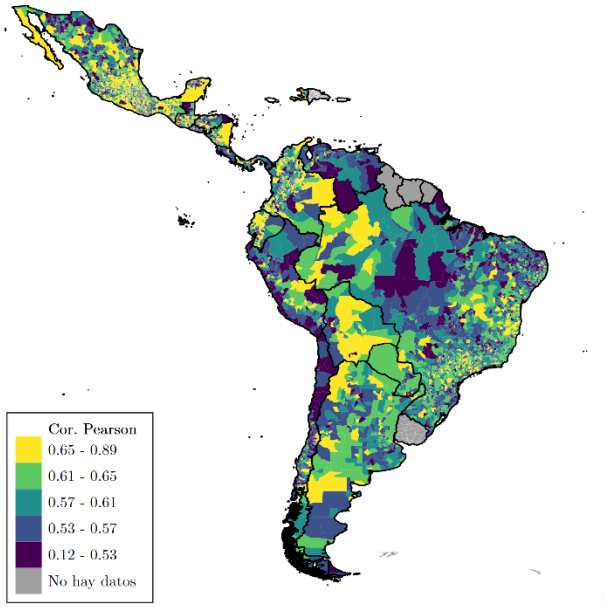
Figura A7. Evolución ICEG vs ICEL



Nota: la figura presenta la evolucón del ICEL y del ICEG agregado por década censo.

Fuente: elaboración propia en base a IPUMS (2020)

Figura A8. Coeficientes de correlación lineal educativa por Geolevel 2 –LAC, Década 2000



Nota: el gráfico presenta los valores que toma el coeficiente de correlación de Pearson de años educativos de hombre y mujeres en pareja, por unidad subnacional de segundo orden en la década de los 2000.

Fuente: elaboración propia en base a IPUMS (2020)

Apéndice I - Identificación de parejas

Para poder calcular las medidas de emparejamiento selectivo es necesario observar la educación de los ambos miembros de una pareja. Para eso, es necesario identificar a los miembros que conforman cada pareja.

Si bien esta fuente de datos cuenta con una variable denominada "SPLOC", que permite identificar diferentes parejas dentro de un hogar, esta no se encuentra disponible en algunos de los censos, motivo por el que no se utilizará. En su lugar, se utilizará la variable "RELATE", la cual identifica al jefe de hogar y todas las relaciones vinculadas con este. Si bien este criterio no permite establecer conexiones de parejas entre el resto de los miembros del hogar diferentes al jefe, creemos que es el más adecuado para un mejor estudio de las tendencias, ya que esta variable se encuentra disponible en todos los censos utilizados en el análisis.

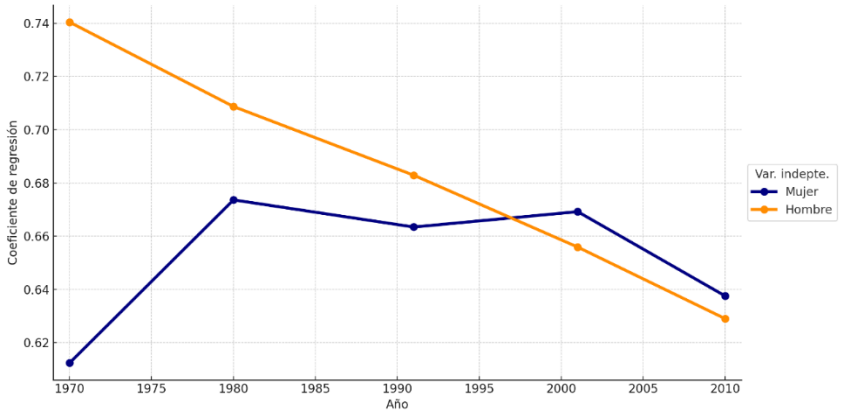
Apéndice II – Regresión y regresión inversa

El modelo que se estima consiste en una regresión en la cual la variable dependiente son los años de educación de la esposa (Yedu_esposa) y la independiente, los años de educación del esposo (Yedu_esposo): $Yedu_{esposa} = \alpha + \beta X edu_{esposo} + u$.

Por definición, el coeficiente de regresión es: $\beta = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{\sigma_y}{\sigma_x}\rho$. Donde σ es el desvío estándar de cada una de las distribuciones y ρ es el coeficiente de correlación de Pearson. Esto implica que un aumento en el coeficiente de regresión puede reflejar un aumento en el coeficiente de correlación de Pearson o un aumento en la varianza del logro educativo de las mujeres en relación con la varianza de la educación de los hombres (si se considera la especificación anterior).

Para mostrar la invalidez de este indicador para medir la evolución del ES, a continuación, se estudiará la evolución de los coeficientes de regresión a lo largo del tiempo para Argentina, considerando alternativamente como variables independientes la educación de esposos y de esposas, respectivamente. La Figura A2.1 muestra las tendencias que surgirían si se considerara la educación de la esposa, además de la típicamente computada educación del hombre, como variable independiente.

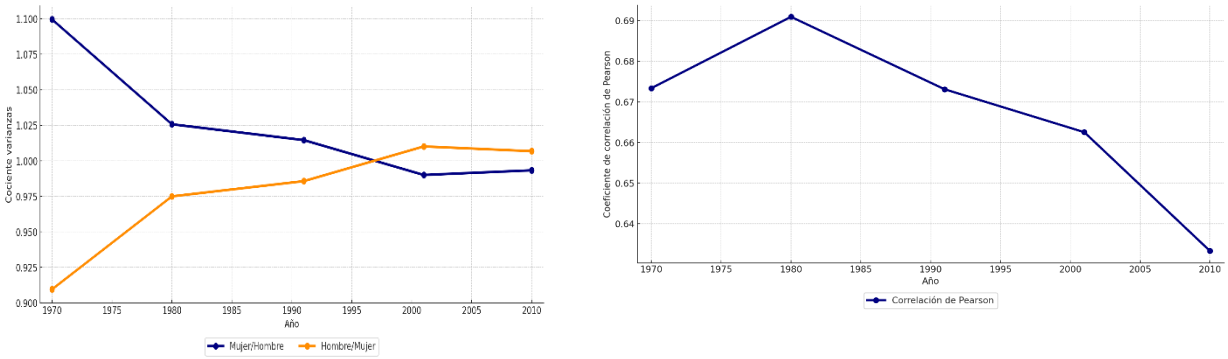
Figura A2.1- Estimación de emparejamiento selectivo a través de MCO cambiando la variable independiente en Argentina.



Nota: el gráfico presenta los resultados de una regresión en la cual la variable dependiente es la educación del hombre (mujer) y la independiente la de la mujer (hombre), por año de censo disponible. Se consideran parejas en las cuales al menos uno de los cónyuges tiene entre 25 y 45 años.
Fuente: elaboración propia en base a IPUMS (2020)

Mientras que el coeficiente de la educación de la esposa disminuye con el tiempo, el coeficiente de del esposo aumenta durante el mismo período. El motivo que explica estas diferencias se vincula con que la varianza en los años de educación entre los hombres cayó sustancialmente durante el período en consideración, en comparación con la de los años de educación entre las mujeres.

Figura A2.2. Evolución de los componentes del coeficiente de regresión en Argentina.
Panel A. Cociente de varianzas Panel B. Coeficiente de correlación de Pearson



Nota: el panel A presenta el cociente de desvíos estándar entre la distribución educativa de mujeres y hombres, alternando el numerador y el denominador. El panel B presenta la evolución del cociente de correlación lineal de Pearson.
Fuente: elaboración propia en base a IPUMS (2020)

La comparación de los resultados de las Figuras A2.1 y A2.2 muestra cómo los cambios en las distribuciones marginales de la educación, si se ignoran, pueden conducir a conclusiones injustificadas sobre la evolución del emparejamiento educativo selectivo.

<u>Apéndice III.</u> Indicadores de segunda generación

Ratio de probabilidad y modelos log-lineales

El ratio de probabilidad es una medida utilizada para evaluar el emparejamiento selectivo en términos de la probabilidad de que dos individuos con características similares formen una pareja, en comparación con aquellos que poseen características diferentes. Esta métrica permite identificar si existe una tendencia a que personas con niveles educativos semejantes se emparejen con mayor frecuencia de lo esperado por azar. El ratio de probabilidad (Odds Ratio, OR) se define como el cociente entre:

- La probabilidad de formar una pareja homogámica (es decir, una pareja en la que ambos miembros tienen el mismo nivel educativo).
- La probabilidad de formar una pareja no homogámica (una pareja en la que los miembros tienen niveles educativos diferentes).

El ratio de probabilidad se puede expresar mediante la siguiente fórmula:

$$IO(a,b,c,d) = \ln\left(\frac{a \cdot d}{b \cdot c}\right)$$

El Odds Ratio es ampliamente utilizado en la literatura demográfica, ya que puede derivarse directamente del enfoque log-lineal (véanse, por ejemplo, Mare (2001), Mare y Schwartz (2005), Bouchet-Valat (2014)). En economía, ha sido empleado por Siow (2015) y por trabajos como Chiappori et al. (2017), Ciscato y Weber (2020), y Chiappori, Costa-Dias, Crossman y Meghir (2020), entre otros.

El modelo log lineal es ampliamente utilizado en la sociología para medir el Emparejamiento selectivo. La hipótesis nula del modelo log-lineal es la independencia estadística completa (es decir, emparejamiento aleatorio). En particular, sea i el nivel educativo del esposo y j el nivel educativo de la esposa, y sea μ_{ijkl} el número de parejas donde el esposo tiene educación i y la esposa tiene educación j. Entonces, la independencia estadística significa que:

$$\mu_{ij} = A \cdot \alpha_i \cdot \beta_j$$

Donde $A, \alpha_i,$ y β_j son parámetros a ajustar. Usualmente, los investigadores toman logaritmos de la ecuación anterior, de manera que:

$$\log\left(\pi_{ij}\right) = A' + \alpha_i' + \beta_j'$$

Donde nuevamente, A', α_i', and β_j' son parámetros a ajustar, pero ahora pueden ser ajustados. Un término residual adicional añadido a la Ecuación 1.4 captura cuán assortativos son los emparejamientos, netos de lo que se predice por emparejamiento aleatorio. Por ejemplo, para medir las tendencias agregadas, Schwartz y Mare (2005) añaden una variable dummy igual a 1 si la educación del esposo y la de la esposa son

iguales y cero en caso contrario (el "modelo de homogamia", ecuación (2)). Para aislar partes particulares de la distribución educativa que están impulsando las tendencias agregadas, estiman un "modelo de cruces", donde añaden varias variables dummy que estiman la dificultad relativa de cruzar cada límite educativo (Schwartz y Mare (2005), ecuación (3)).

Propuesta de Eika:

La propuesta de Eika, et al. (2019) sugiere medir el Emparejamiento Selectivo educativo a través de la comparación de tablas de contingencia. Plantea que, a partir de estas tablas, es posible medir el emparejamiento marital entre los niveles de educación del esposo e_o y la esposa e_a como la probabilidad observada de que un esposo con nivel de educación e_o esté casado con una esposa con nivel de educación e_a, en relación con la probabilidad que surgiría bajo emparejamiento aleatorio. Formalmente, se define al indicador $s(e_o, e_a)$ como:

$$s(e_o, e_a) = \frac{P(E_o = e_o, E_a = e_a)}{P(E_o = e_o)\mathrm{P}\,(E_a = e_a)},$$

donde $E_o(E_a)$ denota el nivel educativo del esposo (esposa). El emparejamiento selectivo positivo se produce entonces cuando el indicador $s(e_o, e_a)$ toma valores mayores a 1, lo cual significa que hombres y mujeres con niveles de educación e_o y e_a se casan con más frecuencia de lo que se esperaría bajo un patrón de emparejamiento aleatorio en términos de educación. Para obtener una medida del emparejamiento selectivo educativo agregado, se calcula el promedio ponderado de los parámetros de emparejamiento a lo largo de la diagonal principal.

Ejercicio de simulación

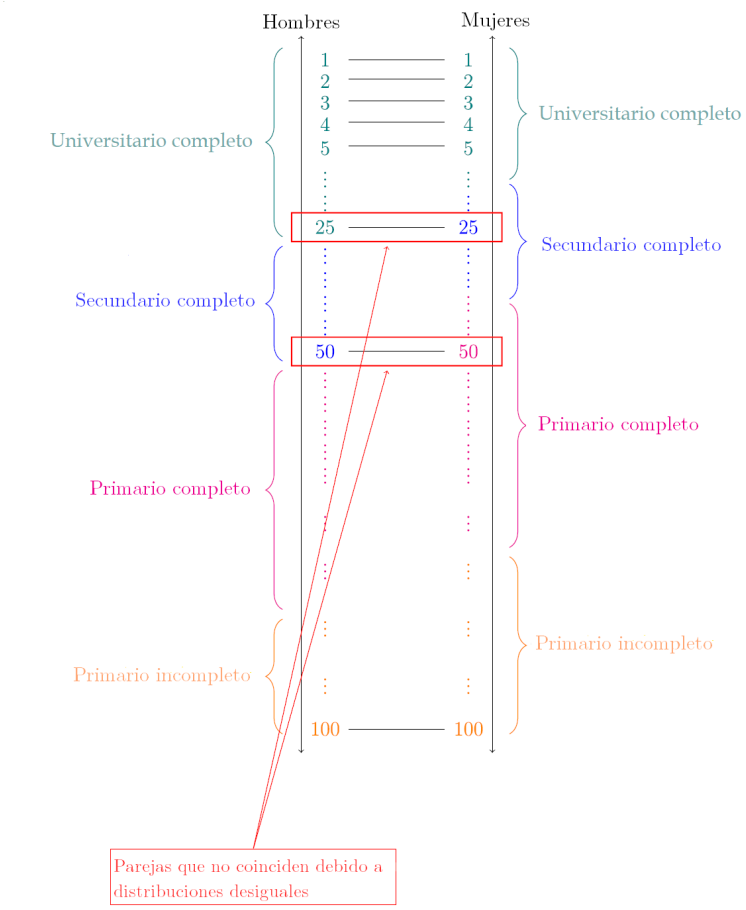
Una manera de evaluar la capacidad que tiene el indicador propuesto por Eika et al., (2019) para controlar completamente los cambios en las distribuciones educativas de hombres y mujeres es mediante la estimación del mismo en diferentes escenarios, en los cuales se asume un emparejamiento selectivo perfecto, pero se varían los supuestos en relación a las distribuciones educativas. En esta sección, se presentará un ejercicio de simulación basado en el trabajo de Shen (2019), adaptado a los datos de Argentina[1], y se incluirán nuevos escenarios, que incorporan las críticas de Bratsberg et al., (2018). A continuación, se evaluará el indicador mediante su funcionamiento en cuatro escenarios alternativos: el observado y tres simulados.

En la primera simulación, se tomará la distribución educativa de hombres y mujeres, dividida en los cuatro niveles presentados en la sección anterior, para cada uno de los años bajo análisis, y se asumirá un emparejamiento perfecto. En cada año t, un número igual de hombres y mujeres se rankean por separado según su nivel educativo, de modo que hay un

[1] Si bien se podría haber elegido cualquiera de los 22 países disponibles en los censos, la información para Argentina presenta la ventaja de que se encuentra disponible para cinco décadas a lo largo del tiempo, lo cual permite estudiar las tendencias sobre la evolución del fenómeno. Además, dicho país, cuenta con un número de observaciones considerable (15 millones en total) pero no excesivo, que permiten realizar los ejercicios de simulación sin demandar grandes costos computacionales (como sí podría suceder en los casos de México o Brasil, por ejemplo).

hombre “top” y una mujer “top”. Cada año, todos forman una pareja y la pareja siempre será perfecta en términos de emparejamiento selectivo: el hombre “top" y la mujer “top" siempre coinciden, al igual que el hombre y la mujer en segundo lugar, y así sucesivamente. Esto es, las personas siempre se emparejarán con alguien que ocupe la misma posición en el ranking de educación correspondiente a su género, sin importar si dicha persona tenga o no el mismo nivel educativo. En la figura A.3.1 se presenta la representación visual de dicho ordenamiento.

Figura A.3.1. Mecanismo de emparejamiento



Fuente: elaboración propia en base a Shen (2019)

Si bien se supone un emparejamiento perfecto, en la primera simulación se permitirá que los tamaños de las categorías educativas difieran según el género y cambien en cada década de acuerdo con las distribuciones educativas observadas en la realidad.

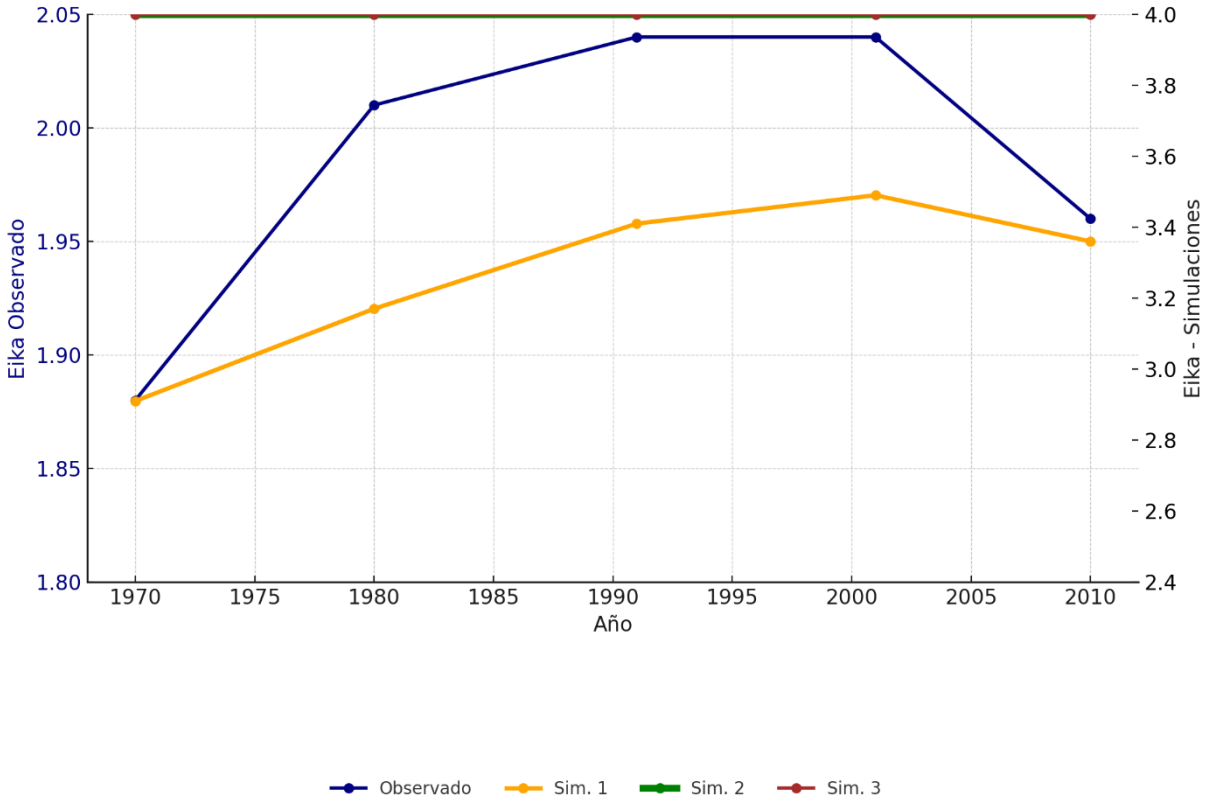
En la segunda simulación, se supondrá un emparejamiento selectivo perfecto y se permitirá que los tamaños de las categorías educativas varíen en el tiempo, pero se asumirá que no existen brechas de género en términos educativos.

En la tercera simulación, además de suponer que existe un Emparejamiento Selectivo perfecto y que no existen brechas de género, se mantendrá constante el tamaño relativo de

dos de las cuatro categorías educativas en cada una de las distribuciones a lo largo del tiempo. En particular, se supondrá que los individuos con universitario completo siempre representan una proporción igual al 5% del total, mientras que aquellos con primario completo representan un 45%. A su vez, la evolución del grupo de individuos con primario incompleto se mantendrá en su estado observado en la realidad, mientras que el porcentaje de individuos con secundario completo responderá a los cambios en el último grupo.

Se evaluará el desempeño del indicador en los cuatro escenarios. Dado que el emparejamiento, por construcción, nunca cambia en los tres escenarios simulados, la métrica debería permanecer constante, incluso cuando cambien las distribuciones de educación de hombres y mujeres. En la Figura A.3.2, se presentan los resultados obtenidos del ejercicio de simulación.

Figura A.3.2. Tendencia del indicador Eika por categorías – Resultados de las simulaciones



Notas: la figura presenta las estimaciones del indicador de Eika para cuatro escenarios alternativos: el observado en la realidad (eje izquierdo) y tres simulados (eje derecho). Es decir, para cada categoría educativa, se calculan las probabilidades (en relación a la coincidencia aleatoria) de observar una pareja en la que ambos miembros tienen ese nivel educativo y se calcula un promedio ponderado de estas probabilidades, donde los ponderadores son la proporción de parejas de educación similar que caen en esa celda. Los datos corresponden a los censos de Argentina.

Fuente: IPUMS-I (2020)

Los resultados de la primera simulación (representados por la línea amarilla) exhiben un aumento en el nivel de emparejamiento selectivo entre los años 1970 y 2010, a pesar de que,

se esperaría que este indicador se mantuviera constante a lo largo de todo el período. De manera llamativa, estos resultados muestran tendencias similares a las observadas en el escenario real.

Por otro lado, en las representaciones en verde y rojo se exhiben los valores del indicador cuando se eliminan las brechas de género a lo largo del tiempo. En estos casos, el indicador funciona de manera correcta, dado que se mantiene constante a lo largo de las décadas y alcanza su nivel máximo[2]. Este ejemplo proporciona evidencia de que, si bien el indicador propuesto por Eika logra controlar una parte de las variaciones en la distribución, no logra abordar las variaciones en las brechas de género (Shen, 2019).

Esta limitación se puede explicar considerando el rango de valores que el indicador puede tomar. Este indicador adquiere valores inferiores a 1 cuando no existe emparejamiento selectivo positivo y alcanza un valor igual a 4 cuando el patrón de emparejamiento es perfecto, y todos los valores de la tabla de contingencia se encuentran únicamente en la diagonal principal. Es posible, entonces, que exista un emparejamiento perfecto en términos educativos, pero que el indicador no alcance su valor máximo debido a la existencia de brechas educativas entre hombres y mujeres. En consecuencia, a medida que estas brechas se reduzcan, el valor del indicador aumentará, a pesar de que los patrones de emparejamiento se mantengan constantes. La explicación proporcionada por Shen para este fenómeno está alineada con este razonamiento, ya que sugiere que a medida que las distribuciones educativas de hombres y mujeres evolucionaron y se volvieron más similares, la probabilidad de encontrar un cónyuge con el mismo nivel educativo aumentó. Por esta razón, el indicador resultante de la primera simulación muestra una tendencia similar a la del mundo real y refleja variaciones derivadas de cambios en las brechas educativas.

Para entender la intuición, se presenta el siguiente ejemplo, representado en la Tabla A.3.1. Considere una población de hombres y mujeres cuyas distribuciones educativas se dividen en dos niveles (alta y baja) y son observadas en tres períodos. En el primer período, los hombres se distribuyen de igual manera entre el nivel bajo y alto de educación (50% en cada nivel), mientras que el 70% de las mujeres tiene educación baja y el 30% restante, alta. En ese caso, si existiera emparejamiento perfecto, el 30% de mujeres con educación alta se emparejaría con el 30% de los hombres con mayor educación. Luego, quedaría un 20% de hombres educados que se emparejaría con mujeres de nivel educativo bajo, dado que no encontrarían parejas con el mismo nivel. Por último, dado que el 20% de mujeres de educación baja se emparejó con el nivel más alto, quedaría un 50% de mujeres que se emparejaría con el 50% de hombres restante.

El problema del indicador propuesto radica en que la matriz de contrafactuales utilizada para computarlo tiene en consideración el total de parejas que podrían emparejarse dadas las distribuciones educativas de cada género, pero no considera que, si hay brechas de género, habrá un porcentaje de individuos (en este caso mujeres) que se emparejará con el residual de individuos del otro género (hombres) que no se emparejaron con personas de su mismo

[2] El valor máximo que puede tomar el indicador de Eika es igual al total de categorías educativas en las cuales se divide la distribución educativa. En este caso, si las distribuciones de hombres y mujeres fueran iguales, la matriz aleatoria debería contener valores iguales a (0.25*0.25=0.0625), y si el emparejamiento fuera perfecto, las celdas de la diagonal principal deberían ser todas iguales a (0.25/0.0625=4)

nivel educativo, dado que los géneros no tienen las mismas proporciones. De esa manera, si bien el emparejamiento selectivo es perfecto, el indicador es menor a 2.

Por otra parte, en el segundo período, el porcentaje de mujeres educadas aumenta en 10 puntos porcentuales. En ese caso, la brecha de género se achica, y el indicador Eika se acerca a 2. Finalmente, en el tercer período, la brecha se anula y únicamente en ese caso, el indicador reporta que el emparejamiento es perfecto, a pesar de que el patrón se mantuvo a lo largo de los tres períodos, es decir, a pesar de que los individuos pertenecientes al mismo ranking de cada distribución se emparejaron en los tres casos.

Tabla A.3.1. Crítica Indicador Eika

M1. Mat obs (Perfecta)

H\M	Alta	Baja	Total
Alta	0.50	0.20	**0.70**
Baja	0.00	0.30	**0.30**
Total	**0.50**	**0.50**	

M2. Matriz aleatoria

H\M	Alta	Baja
Alta	0.35	0.35
Baja	0.15	0.15

M3. Matriz Eika

H\M	Alta	Baja
Alta	1.43	0.57
Baja	0.00	2.00

Indicador Eika: 1.64

M1. Mat obs (Perfecta)

H\M	Alta	Baja	Total
Alta	0.50	0.10	**0.60**
Baja	0.00	0.40	**0.40**
Total	**0.50**	**0.50**	

M2. Matriz aleatoria

H\M	Alta	Baja
Alta	0.30	0.30
Baja	0.20	0.20

M3. Matriz Eika

H\M	Alta	Baja
Alta	1.67	0.33
Baja	0.00	2.00

Indicador Eika: 1.81

M1. Mat obs (Perfecta)

H\M	Alta	Baja	Total
Alta	0.50	0.00	**0.50**
Baja	0.00	0.50	**0.50**
Total	**0.50**	**0.50**	

M2. Matriz aleatoria

H\M	Alta	Baja
Alta	0.25	0.25
Baja	0.25	0.25

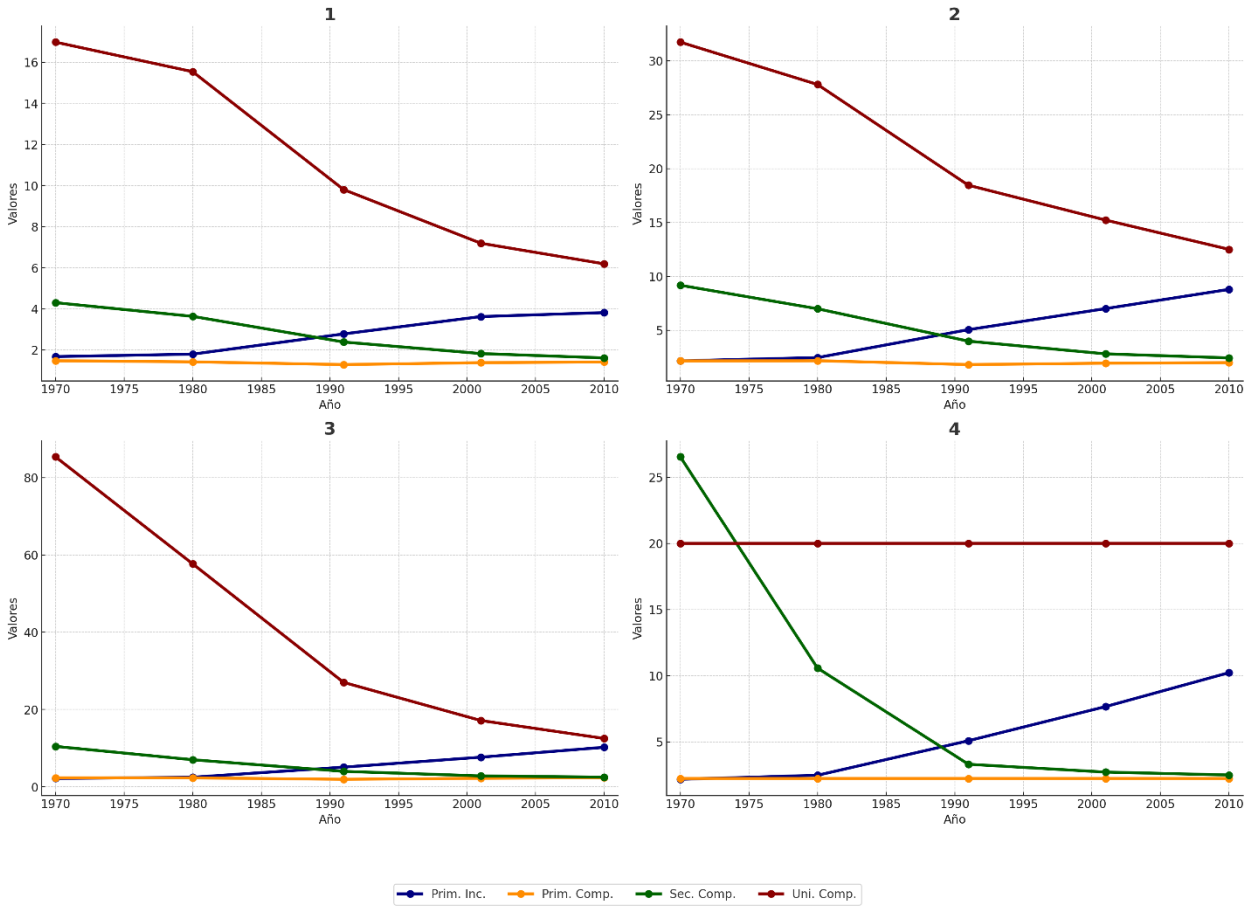
M3. Matriz Eika

H\M	Alta	Baja
Alta	2	0
Baja	0	2

Indicador Eika: 2

A pesar de que este problema parece resolverse cuando se controlan las brechas de género (escenarios 3 y 4), al examinar la evolución de cada categoría por separado en relación con un escenario contrafactual aleatorio, los desafíos persisten. La Figura A.3.3 presenta los resultados de la evolución de las categorías de la diagonal principal de la tabla de contingencia a lo largo del tiempo.

Figura A.3.3. Tendencia del indicador Eika por categorías – Resultados de las simulaciones



Notas: la figura presenta las estimaciones del indicador de Eika por categorías para cuatro escenarios alternativos: el observado en la realidad y tres simulados. Es decir, para cada categoría educativa, se calculan las probabilidades (en relación a la coincidencia aleatoria) de observar una pareja en la que ambos miembros tienen ese nivel educativo. Los datos corresponden a los censos de Argentina.
Fuente: IPUMS-I (2020)

Los resultados evidencian que, incluso al considerar el control de las brechas de género, las categorías correspondientes al escenario observado y a las dos primeras simulaciones (escenarios 1, 2 y 3) muestran tendencias similares. Sin embargo, en la última simulación, donde se controla el tamaño de los grupos con UNIC y PRIC, se observa que las tendencias de dichas categorías se mantienen estables a lo largo del tiempo. Bratsberg et al. (2018) explican que estos resultados surgen como consecuencia de un efecto de composición de los grupos.

La composición de los grupos educativos ha cambiado necesariamente como resultado de la expansión educativa en todas las cohortes. Si bien los cambios mecánicos resultantes en las probabilidades de emparejamiento pueden controlarse normalizando las frecuencias de emparejamiento observadas con las que se habrían aplicado en el emparejamiento aleatorio, dicha normalización no se ocupa de los posibles cambios asociados en la composición de los grupos educativos.

Para entender la intuición, se presenta el siguiente ejemplo, representado en la tabla A.3.2. Considere una población de hombres y mujeres, divididos en dos niveles educativos (alto y bajo) en dos momentos diferentes del tiempo. En ambos casos, el patrón de emparejamiento es perfecto y no existen brechas de género en educación. Sin embargo, entre el primero y segundo período, se da una gran expansión educativa. Mientras que, en el primer caso, el 90% de los integrantes de cada género son poco educados, en la segunda distribución solamente el 50% es poco educado. Dado que el emparejamiento es perfecto y que no existen brechas de género, el indicador agregado de Eika es igual a 2. Sin embargo, cuando uno observa los valores de la matriz Eika por categoría, se observa que, dependiendo el tamaño de las distribuciones observadas, el contrafactual aleatorio hará que los valores finales de la matriz Eika crezcan de manera cuadrática a medida que disminuye el tamaño de los grupos.

En pocas palabras, si el logro educativo aumenta entre las cohortes, esto puede causar que el logro bajo identifique un grupo más pequeño y más homogéneo y el logro alto identifique un grupo más grande y heterogéneo en términos de alguna otra característica que puede ser el factor de emparejamiento real. Si este fuera el caso, observaríamos un emparejamiento selectivo educativo creciente para el grupo de baja educación y un apareamiento selectivo decreciente para el grupo de educación alta, incluso si los patrones de apareamiento subyacentes no cambiaran por completo.

En resumen, si bien el indicador propuesto por Eika et al. (2019) pretende controlar los cambios en las distribuciones mediante una normalización basada en un contrafactual aleatorio, no tiene en cuenta los posibles problemas que pueden surgir debido a los cambios en las brechas de género, ni considera las variaciones en la composición de cada grupo a medida que evolucionan las distribuciones.

Tabla A.3.2. Crítica Indicador Eika por categorías

M. Perfecta

H\M	**Alta**	**Baja**	**Total**
Alta	0.90	0.00	0.90
Baja	0.00	0.10	0.10
Total	**0.90**	**0.10**	

Matriz aleatoria

H\M	**Alta**	**Baja**
Alta	0.81	0.09
Baja	0.09	0.01

Matriz Eika

H\M	**Alta**	**Baja**
Alta	1.10	0.00
Baja	0.00	10.00

Indicador Eika: 2

M. Perfecta

H\M	**Alta**	**Baja**	**Total**
Alta	0.50	0.00	**0.50**
Baja	0.00	0.50	**0.50**
Total	**0.50**	**0.50**	

Matriz aleatoria

H\M	**Alta**	**Baja**
Alta	0.25	0.25
Baja	0.25	0.25

Matriz Eika

H\M	**Alta**	**Baja**
Alta	2.00	0.00
Baja	0.00	2.00

Indicador Eika: 2

Apéndice IV – **Funcionamiento y crítica al indicador de Shen (2019)**

Shen propone una métrica denominada "Normalización Perfecta-Aleatoria" (PRN, por sus siglas en inglés), que se define como una normalización que toma un valor de cero si el patrón de emparejamiento observado es aleatorio y un valor de uno si es perfecto. Luego, se evalúa dónde se encuentran los emparejamientos observados en relación con estos dos límites[3]. El indicador se computa a partir de tablas de contingencia. Formalmente, sean E_o y E_a dos variables aleatorias para el nivel de educación de un hombre y una mujer, respectivamente, y sean e_o y e_a dos niveles de educación cualesquiera. Entonces, se denota $p_{\text{sim}}(e_o, e_a)$ donde $\text{sim} \in \{\text{Perfecta, Aleatoria, Observada}\}$ como la probabilidad de observar un emparejamiento con $E_o = e_o$ y $E_a = e_a$ en cada. La normalización perfecta aleatoria es:

$$s_{Shen}(e_o, e_a) = \frac{p_{\text{observada}}\,(E_o = e_o, E_a = e_a) - p_{\text{aleatoria}}\,(E_o = e_o, E_a = e_a)}{p_{\text{perfecta}}\,(E_o = e_o, E_a = e_a) - p_{\text{aleatoria}}\,(E_o = e_o, E_a = e_a)}$$

La simulación de emparejamiento aleatorio lo es en un sentido estadístico, es decir, asume que E_o y E_a son variables aleatorias independientes, entonces $p_{\text{aleatoria}}\,(e_o, e_a) = P(E_o = e_o) * P(E_a = e_a)$[4]. El problema del indicador de Shen radica en que puede fallar en situaciones donde la brecha de género es tan significativa en un nivel educativo que no es compensada por el residuo del siguiente.

El indicador de Shen tiene un buen funcionamiento, en el sentido de que los valores de la métrica se encuentran entre -1 y 1, para la mayoría de los países y los años considerados. Sin embargo, este parece mostrar fallas en algunos casos en los cuales las distribuciones educativas presentan grandes brechas de género. Un ejemplo de dicha falla se presenta al considerar el censo de Brasil correspondiente a la década de 1960. Brasil es un país cuyo desarrollo educativo ha sido muy pobre durante muchos años, principalmente a mediados del siglo XX. La distribución educativa de hombres y mujeres, al dividir en cuatro categorías, presenta la siguiente forma:

[3]Si las parejas observadas están por debajo del contrafactual de emparejamiento aleatorio (ES negativo), entonces se trata al emparejamiento aleatorio como el límite superior (normalizado para que sea igual a 0) y se simula un escenario de emparejamiento perfectamente negativo, es decir, el hombre "top" se empareja con N-ésima mujer de la distribución educativa, el hombre 2 coincide con N-1, etc. Luego, se considera al contrafactual de emparejamiento perfectamente negativo como límite inferior, normalizado para ser -1, y la Normalización perfecta-aleatoria (para el ES negativo) pregunta dónde, entre estos dos límites, se encuentran los emparejamientos observados.

[4] La expresión analítica para $p_{\text{perfecta}}(e, e')$ depende de cómo coincidan los hombres y las mujeres "sobrantes".

Tabla A4.1. Matriz observada distribución educativa parejas – Brasil, 1960.

Mujer/Hombre	PRII	PRIC	SECC	UNIC	Total
PRII	92.16	2.05	0.78	0.31	**95.30**
PRIC	1.04	1.06	0.57	0.51	**3.19**
SECC	0.35	0.29	0.41	0.33	**1.38**
UNIC	0.01	0.02	0.02	0.08	**0.13**
Total	**93.56**	**3.43**	**1.78**	**1.24**	**100.00**

En base a dicha tabla de contingencia observada, es posible calcular el contrafactual de emparejamiento perfecto de la manera en que lo propone Shen. A continuación, se presenta la tabla correspondiente a dicho escenario hipotético.

Tabla A.4.2. Matriz de emparejamiento perfecto – Brasil, 1960.

Mujer/Hombre	PRII	PRIC	SECC	UNIC	Total
PRII	93.56	1.74	0.00	0.00	**95.30**
PRIC	0.00	1.69	1.50	0.00	**3.19**
SECC	0.00	0.00	**0.28**	1.11	**1.38**
UNIC	0.00	0.00	0.00	0.13	**0.13**
Total	**93.56**	**3.43**	**1.78**	**1.24**	**100.00**

Se puede observar que en un caso como el presentado, en el cual las distribuciones educativas presentan grandes brechas en algunos niveles educativos, el valor observado en la diagonal principal es mayor al valor del contrafactual perfecto. Esto significa que el indicador de Shen tomará valores mayores a 1.

Al ver la tabla de contingencia del escenario contrafactual, se observa que la brecha en el nivel universitario entre hombres y mujeres es tal que los hombres universitarios se juntan con mujeres universitarias, y luego, el residual de hombres se junta con mujeres con educación secundaria completa. El problema es que es tan grande la diferencia entre universitarios (y tan parecida al porcentaje de individuos con secundario), que el residual de mujeres que queda para emparejarse con hombres es menor al porcentaje observado.

Apéndice V – Simulación Gihleb y Lang.

Siguiendo el enfoque presentado por Gihleb y Lang (2020), se examina el rendimiento de estas diferentes métricas, incluyendo la correlación lineal de Pearson, utilizando los datos de censos para parejas en Argentina. Para llevar a cabo el análisis, se simulan dos escenarios de emparejamiento perfecto. En el primero, se divide la distribución de años de educación en rangos de 0 a 18, teniendo en cuenta la disponibilidad de datos. En el segundo, se divide la distribución de educación en cuatro categorías educativas estandarizadas según el proyecto IPUMS. Los escenarios de emparejamiento perfecto se simulan de la misma manera que se describió en la subsección anterior, y a continuación se presentan los resultados correspondientes para la distribución de años educativos (los resultados relacionados con la estimación basada en niveles educativos se encuentran en la Tabla A4).

Tabla 2. Resultados de las diferentes métricas bajo emparejamiento perfecto.

Año	ρ	r	γ	τ_a	τ_b
1970	0.98	0.98	1.00	0.81	0.95
1980	0.99	0.99	1.00	0.87	0.97
1991	1.00	1.00	1.00	0.87	0.99
2001	0.99	0.99	1.00	0.82	0.96
2010	0.99	0.99	1.00	0.85	0.96

Nota: la tabla presenta los resultados que surgen al estimar diferentes métricas de correlación en un escenario en el que se simula emparejamiento perfecto, utilizando datos de Argentina para diferentes décadas de censo.
Fuente: IPUMS-I (2020)

Los resultados presentados en la Tabla 2 revelan que, al considerar los años de educación como variable de interés, tanto el coeficiente de correlación de Pearson como el coeficiente de correlación de Spearman muestran un desempeño similar y prácticamente perfecto. Sin embargo, cuando se examinan las cuatro categorías educativas, se observa que el coeficiente de correlación de rangos de Spearman exhibe un mejor rendimiento, seguido de cerca por la Tau B de Kendall y el coeficiente de correlación de Pearson. En contraste, la Tau A de Kendall presenta un desempeño deficiente en ambos casos. Es importante destacar que, debido al diseño del "experimento", el indicador de Goodman y Kruskal es igual a 1 en todos los períodos.

Apéndice VI. Transformación de la variable educativa

La definición de la variable educativa transformada y de los cuantiles que surgen de esta se realiza teniendo en consideración dos aspectos fundamentales de las distribuciones de educación. En primer lugar, la educación se encuentra limitada en un rango entre 0 y 18 años, con puntos de acumulación correspondientes a la finalización de cada nivel educativo (primario, 6 años; secundario, 12; universitario, 17/18 años). Si se definieran cuantiles en función de esta variable, se correría el riesgo de incluir los mismos valores de años de educación en diferentes cuantiles. En segundo lugar, dada la expansión educativa en la región, existe una relación positiva entre el año de nacimiento y la educación promedio de los individuos. Si además se considera que las edades de los cónyuges son similares en la práctica, el hecho de no controlar por estos cambios podría generar distorsiones en las métricas de ES de cada censo.

Para definir cuantiles en base a una variable continua y permitir la comparación entre rankings de individuos de diferentes edades, se sigue la metodología propuesta por Halliday et al. (2021). Esta implica definir cuantiles en función de los residuos de una regresión de los años de educación en relación con la variable de edad. De manera formal, para cada censo disponible, se estima la siguiente regresión de Mínimos Cuadrados Ordinarios:

$$aedu_{id} = \alpha + \sum_{a \varepsilon A} \beta^{a} a\tilde{n}o\_nac\ {}^{a}_{id} + \epsilon_{id}\ \ para\ g[1,2],$$

donde $aedu_{id}$ representa los años de escolaridad del individuo i perteneciente al censo relevado en la década d. La variable independiente $a\tilde{n}o\_nac\ {}^{a}_{id}$ representa el año de nacimiento a del individuo i perteneciente al censo relevado en la década d. La regresión se corre para cada país y para cada género, g, por separado. Luego de correr cada regresión, se obtienen los residuos de las mismas y se definen cuantiles en base a estos.

Ventajas del ICEL

El Indicador Cuantílico Educativo Local, además de ser insensible a los cambios en las distribuciones marginales de educación, presenta una serie de ventajas que lo hacen apropiado para el estudio del ES. En primer lugar, permite controlar los cambios en las distribuciones de educación entre diferentes censos y entre cohortes de nacimiento dentro de un mismo censo, además de los cambios en las brechas de género y las variaciones en la composición de los grupos. En segundo lugar, la división permite estudiar el emparejamiento en diferentes partes de la distribución educativa. En tercer lugar, la interpretación del indicador es más intuitiva que, por ejemplo, los índices tradicionales de correlación. El cuarto punto se refiere al esquema de clasificación de las categorías educativas que se elija. Esta elección puede ser arbitraria y basarse en divisiones tradicionales, pero no implica que sea la única o la más adecuada. Además, diferentes esquemas de clasificación pueden dar lugar a diferentes tendencias (Gihleb & Lang, 2020).[5] Al considerar cuantiles de la

[5] Un contrapunto importante es que, al separar las categorías educativas de la manera tradicional, se reconoce que la educación también es una señal. Por ejemplo, tener 11 o 12 años de educación puede no parecer una diferencia

distribución, esta métrica omite dicha decisión. Por último, es posible generar extensiones directas a partir del mismo indicador, las cuales serán presentadas en las próximas secciones.

Apéndice VII

La literatura sobre emparejamiento selectivo ha estudiado de manera recurrente cómo cambiaría la desigualdad de ingresos de los hogares si las parejas se formaran aleatoriamente en lugar de seguir los patrones observados (Eika et al., 2019; Greenwood et al., 2014; Hryshko et al., 2017, para Estados Unidos; Leal, 2015; Pereira y Santos, 2017 para países latinoamericanos). En estos ejercicios, la aleatorización genera pseudohogares, es decir, hogares contrafactuales construidos al reasignar parejas manteniendo fija la población y (según el caso) ciertas características observables.

Para abordar este punto, se han empleado dos enfoques alternativos de aleatorización conocidos como "aditivo" y "de imputación", según la terminología propuesta por Harmenberg (2014). Ambos enfoques se orientan hacia la aleatorización de la formación de hogares, aunque divergen en la forma en que asignan los ingresos al nuevo hogar formado aleatoriamente –denominado "pseudohogar". El enfoque aditivo calcula los ingresos de los pseudohogares sumando los ingresos individuales, mientras que el enfoque de imputación establece los ingresos del pseudohogar asumiendo que siguen la misma distribución que los ingresos reales de los hogares con características semejantes. Formalmente, sean x_i, x_j características individuales (edad, educación, etc.) de hombres y mujeres, y sean y_i, y_j los ingresos individuales de hombres y mujeres, respectivamente. El ingreso familiar de un hogar $(i,j), y_{ij}$ se define naturalmente como $y_{ij} := y_i + y_j$. Sea $dF(z)$ la distribución (verdadera) de una variable dada z. El enfoque aditivo, manteniendo constantes los atributos x_i, x_j en parejas, equivale a calcular la distribución:

$$\left(dF(y_i \mid x_i) * dF(y_j \mid x_j)\right) dF(x_i, x_j)$$

donde $*$ denota el operador de convolución. Es decir, se supone que a los agentes con características (i,j) se les asigna aleatoriamente una pareja, de modo que $y_i \mid x_i$ y $y_j \mid x_j$ son independientes. Bajo la aleatorización incondicional, los atributos que se mantienen constantes, x_i, x_j, están vacíos, por lo que esto se reduce a:

$$dF(y_i) * dF(y_j)$$

El enfoque de imputación, utilizando las características observables x_i, x_j, en su lugar calcula:

$$\mathrm{dF}(y_{ij} \mid x_i, x_j) dF(x_i) dF(x_j)$$

Bajo este esquema, el ingreso del hogar se imputa a partir de las características observables del mismo. Estos dos enfoques de aleatorización son conceptualmente distintos y, en general, tiene poco sentido comparar los resultados que surgen de cada uno de ellos. El beneficio de la aleatorización por adición es que toma la

significativa en términos de años, pero puede representar una distinción importante en términos de poseer o no un diploma de secundaria completa.

población existente y la aleatoriza, con lo cual no hay pérdida de información. El gran inconveniente es que dicho método considera que la oferta de trabajo y los ingresos son exógenos a la formación del hogar. El beneficio de la aleatorización de la imputación es que sí tiene en cuenta que la oferta laboral y los ingresos son endógenos a la formación del hogar. El inconveniente de dicha metodología en la práctica es que los atributos disponibles dan una imputación menos que perfecta.