Análisis de la pobreza multidimensional con datos del Banco Mundial

Josué Orión Nava Ponce

12 de junio del 2024

Índice

Introducción

La pobreza multidimensional es un fenómeno complejo que trasciende la mera carencia de ingresos, abarcando múltiples dimensiones que afectan el bienestar humano, como la educación, la salud, la vivienda y el acceso a servicios básicos. Este proyecto, titulado "Análisis de la pobreza multidimensional con datos del Banco Mundial", se propone explorar y analizar estas dimensiones utilizando métodos estadísticos multivariados. La pobreza multidimensional se mide a través de diversos indicadores, proporcionando una visión más completa y precisa de las privaciones que enfrentan las personas en diferentes contextos.

El Banco Mundial, con su amplia base de datos y su enfoque en el desarrollo global, ofrece una plataforma ideal para este análisis. Los datos utilizados en este proyecto abarcan 149 países y seis indicadores clave: pobreza monetaria, nivel educativo, escolarización, acceso a electricidad, saneamiento y agua potable. Estos indicadores proporcionan una representación integral de las condiciones de vida y las carencias que afectan a las poblaciones. Sin embargo, es necesario realizar algunas consideraciones previas que den soporte a la validez estadística del presente estudio, como los años en los que se registraron los indicadores y la eliminación de valores atípicos.

El objetivo general de este proyecto es realizar un análisis de la pobreza multidimensional mundial, descomponiendo los datos en componentes principales, factores y clusters para identificar patrones y relaciones significativas. Específicamente, se busca presentar estadísticas descriptivas, realizar análisis factorial, y llevar a cabo un análisis de clustering para comparar grupos de países según diferentes formas de regionalización mundial.

La importancia de este análisis radica en su capacidad para informar políticas públicas más efectivas y focalizadas. Al entender mejor las múltiples dimensiones de la pobreza, los gobiernos y las organizaciones pueden diseñar intervenciones que aborden las causas subyacentes de la pobreza en lugar de solo sus síntomas. Además, este enfoque integral permite un seguimiento más preciso del progreso hacia la reducción de la pobreza, lo que es esencial para ajustar estrategias y garantizar que se alcancen los objetivos de desarrollo.

Objetivos

Objetivo general. Realizar un análisis de la pobreza multidimensional mundial.

Objetivos específicos:

- Presentar estadísticas generales sobre pobreza multidimensional en un conjunto de países.
- Realizar un análisis factorial.
- Realizar un análisis de clusters.
- Proporcionar una herramienta para proponer políticas que disminuyan los niveles de pobreza a nivel mundial.

Antecedentes

Definición de pobreza y pobreza multimensional

Conforme al *Diccionario de la Lengua Española*, el vocablo "pobreza" proviene de "pobre", cuyo adjetivo es "necesitado", que no tiene lo necesario para vivir" de donde es común vincularla a la ausencia de bienes.

La Organización de las Naciones Unidas (ONU) en su informe de desarrollo humano 2000, define a la pobreza humana como "el empobrecimiento en múltiples dimensiones: la privación en cuanto a una vida larga y saludable, en cuanto a conocimiento, en cuanto a un nivel decente de vida, en cuanto a participación".

Por otra parte, el Banco Mundial ha señalado que aun cuando la pobreza se ha definido como la falta de lo necesario para asegurar el bienestar material, ésta también comprende la ausencia de otros recursos, criterio similar al de la ONU.

> La pobreza se define comúnmente como la falta de lo necesario para asegurar el bienestar material, en particular alimentos, pero también vivienda, tierras y otros activos. En otras palabras, la pobreza entraña una carencia de muchos recursos que da lugar al hambre y a las privaciones físicas.

La Comisión Económica para América Latina (CEPAL), que trabaja en esta región del mundo con los fundamentos que sustentan las acciones de la ONU, las cuales se han plasmado en el Programa de las Naciones Unidas para el Desarrollo (PNUD), define a la pobreza, en términos generales, como:

> ... la incapacidad de las personas de vivir una vida tolerable. Entre los aspectos que la componen se menciona llevar una vida larga y saludable, tener educación y disfrutar de un nivel de vida decente, además de otros elementos como la libertad política, el respeto de los derechos humanos, la seguridad personal, el acceso al trabajo productivo y bien remunerado y la participación en la vida comunitaria.

La Medida de Pobreza Multidimensional (MPM) busca comprender la pobreza más allá de una simple dimensión monetaria al incluir el acceso a la educación y la infraestructura básica junto con el índice de recuento monetario en la línea de pobreza de $1.90. La medida del Banco Mundial se inspira y orienta en otras medidas multidimensionales destacadas, en particular el Índice de Pobreza Multidimensional (IPM) elaborado por el PNUD y la Universidad de Oxford, pero difiere de ellas en un aspecto importante: incluye la pobreza monetaria (medida como un consumo diario inferior al $1.90 en PPA de 2011) como una de las dimensiones (El Banco Mundial, 2024).

Si bien la pobreza monetaria está fuertemente correlacionada con las privaciones en otros ámbitos, esta correlación dista mucho de ser perfecta. El informe Pobreza y prosperidad compartida 2020 (Banco Mundial, 2020) muestra que más de un tercio de quienes experimentan pobreza multidimensional no están incluidos en el índice de recuento monetario, en consonancia con las conclusiones de

la edición anterior del informe (Banco Mundial, 2018). El MPM de un país es al menos tan alto o mayor que la pobreza monetaria, lo que refleja el papel adicional de las dimensiones no monetarias en el aumento de la pobreza multidimensional y su importancia para el bienestar general (El Banco Mundial, 2024).

El Banco Mundial

Conformado por 189 países miembros, con personal de más de 170 países, y oficinas en más de 130 lugares, el Grupo Banco Mundial es una asociación mundial única: las cinco instituciones que lo integran trabajan para reducir la pobreza y generar prosperidad compartida en los países en desarrollo.

Colaboración con los Gobiernos

El BIRF y la AIF conforman el Banco Mundial, el que proporciona financiamiento, asesoría sobre políticas y asistencia técnica a los Gobiernos de los países en desarrollo. La AIF se concentra en los países más pobres del mundo, en tanto que el BIRF otorga asistencia a los países de ingreso mediano y los países pobres que tienen capacidad crediticia.

Colaboración con el sector privado

IFC, MIGA y CIADI se concentran en el fortalecimiento del sector privado en los países en desarrollo. A través de estas instituciones, el Grupo Banco Mundial proporciona financiamiento, asistencia técnica, seguros contra riesgos políticos y mecanismos de solución de diferencias a las empresas privadas, incluidas las instituciones financieras.

Un solo Grupo Banco Mundial

Si bien las cinco instituciones que conforman el Grupo Banco Mundial tienen sus propios países miembros, órganos directivos y convenios constitutivos, todas ellas trabajan al unísono para brindar servicios a sus países clientes. Los desafíos de desarrollo actuales solo se pueden encarar con la participación del sector privado. Pero el sector público sienta las bases para facilitar la inversión del sector privado y permitirle a este prosperar. Las funciones complementarias de las cinco instituciones permiten al Grupo Banco Mundial tener la capacidad única de conectar los recursos financieros internacionales con las necesidades de los países en desarrollo.

Importancia de medir la pobreza multidimensional

Medir la pobreza multidimensional es fundamental para abordar de manera efectiva los desafíos que enfrenta la población en situación de pobreza. A diferencia de las medidas tradicionales que se

centran únicamente en el ingreso monetario, la pobreza multidimensional reconoce que la privación puede ocurrir en varias dimensiones, tales como la salud, la educación, las condiciones de vida y la participación social. Esta perspectiva integral permite una comprensión más profunda y precisa de las diversas formas en que la pobreza afecta a las personas y comunidades.

La importancia de medir la pobreza multidimensional radica en varios aspectos clave:

1. Identificación precisa de necesidades: Al considerar múltiples dimensiones, se pueden identificar con mayor precisión las áreas específicas donde las personas están experimentando privaciones. Esto es crucial para diseñar intervenciones que realmente aborden las causas subyacentes de la pobreza en lugar de solo sus síntomas.

2. Políticas públicas más efectivas: Con una medición multidimensional, los gobiernos y las organizaciones pueden diseñar políticas y programas más específicos y efectivos. Por ejemplo, si se descubre que la falta de acceso a servicios de salud es una dimensión crítica de la pobreza en una región particular, se pueden asignar recursos y desarrollar programas dirigidos específicamente a mejorar la atención médica.

3. Evaluación y monitoreo: La medición multidimensional permite un seguimiento más completo del progreso hacia la reducción de la pobreza. Esto es fundamental para evaluar la efectividad de las políticas implementadas y ajustar las estrategias según sea necesario para asegurar que se logren los objetivos de desarrollo.

4. Enfoque en el bienestar holístico: La pobreza no solo se trata de la falta de ingresos, sino también de la falta de oportunidades y derechos. Medir la pobreza de manera multidimensional permite un enfoque más holístico y centrado en el bienestar humano, reconociendo la importancia de factores como la educación, la salud y la participación en la comunidad.

5. Empoderamiento de las comunidades: Al visibilizar las múltiples dimensiones de la pobreza, se empodera a las comunidades afectadas, proporcionando una mejor comprensión de sus propias necesidades. Esto puede fomentar una mayor participación en la toma de decisiones y en la implementación de soluciones, lo que puede llevar a resultados más sostenibles y equitativos.

En resumen, medir la pobreza multidimensional es esencial para una comprensión completa y precisa de la pobreza. Esta medición permite desarrollar políticas y programas más efectivos, garantizar una mejor asignación de recursos y, en última instancia, mejorar el bienestar de las personas y las comunidades en todo el mundo. La adopción de este enfoque integral es crucial para abordar de manera efectiva las diversas y complejas formas en que la pobreza impacta la vida de millones de personas.

Métodos

Información general sobre el conjunto de datos

El conjunto de datos contiene medidas en porcentaje de 6 diferentes indicadores de tasas de privación para 149 países. La definición de los indicadores y los umbrales de privación son los siguientes (World Bank Group, 2018):

- **Pobreza monetaria.** Un hogar está privado si los ingresos o gastos, en dólares estadounidenses de paridad de poder adquisitivo de 2011, son inferiores a 1.90 USD por persona y día.
- **Nivel educativo.** Un hogar sufre privaciones si ningún adulto (de edad equivalente al noveno grado o superior) ha completado la educación primaria.
- **Escolarización.** Un hogar sufre privaciones si al menos un niño (de edad equivalente a 8º curso o inferior) no está escolarizado.
- **Electricidad.** Un hogar sufre carencias si no tiene acceso a la electricidad.
- **Saneamiento.** Un hogar sufre carencias si no tiene acceso a un nivel de saneamiento, aunque sea limitado.
- **Agua potable.** Un hogar sufre carencias si no tiene acceso a agua potable, aunque sea de calidad limitada.

Los datos se refieren a la proporción de personas que viven en hogares con carencias según cada indicador.

Selección del periodo de tiempo para el análisis

Considerando el año de informe se observa que el mínimo es 2009 y el máximo es 2022. La mediana es 2018 y la media es 2017.584, habiendo una desviación estándar de 3.04.

Los indicadores macroeconómicos y de pobreza suelen cambiar lentamente, pero pueden ser influenciados por políticas gubernamentales, eventos económicos globales y otras variables. Por lo tanto, se debe elegir un rango de años que sea suficientemente largo para ser representativo pero no tan largo como para incluir cambios significativos.

Ravallion (1992) discute la necesidad de elegir periodos temporales que sean suficientemente largos para ser representativos, pero no tan largos que incluyan cambios estructurales importantes. Por lo tanto, seleccionamos una vecindad de 3 años alrededor de la media redondeada y la mediana, es decir 2018 ± 3 años que, de acuerdo con la literatura, parece ser una selección razonable.

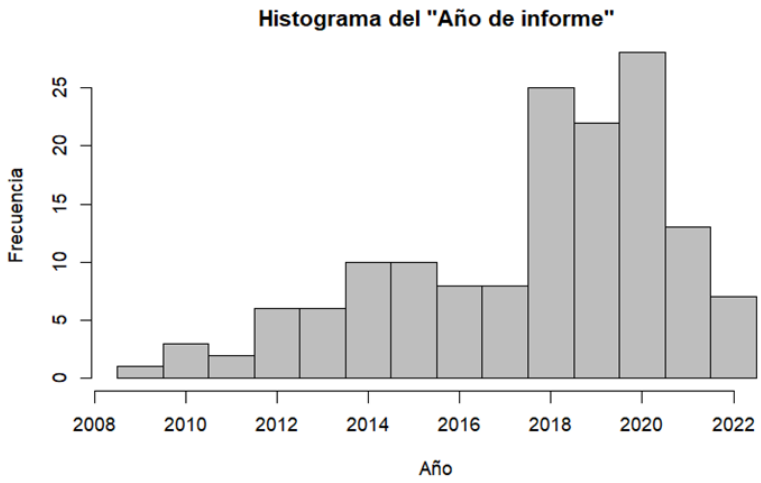

Figura 1: Gráfico de histograma de la variable del "Año de informe". Se observa que la variable no se distribuye normalmente.

Tomar esta vecindad corresponde a tomar la primera desviación estándar, sin embargo, es importante señalar que en este caso no se aplica la regla empírica de la distribución normal (que dice que la primera desviación contiene al 66 % de los datos), ya que los datos no tienen distribución normal, como lo muestra el histograma en la Figura 1. Más bien, corresponde al 76.51 % (114 países) de los datos (ver Anexo 1), en cuyo caso es mejor.

Análisis de datos faltantes

Se tiene un conjunto de datos de 6 variables medidas en 114 países, por lo tanto, se deberían tener 684 datos. Sin embargo, hay 73 datos perdidos (ver Anexo 2).

El Cuadro 1 muestra el número de datos perdidos por columna. Las variables de pobreza monetaria y logro educativo no tienen datos perdidos, y en el caso de electricidad y agua potable no parece ser alarmante; las dos variables alarmantes son la de matrícula educativa y saneamiento, respectivamente. El Cuadro 2 muestra el número de datos perdidos por fila: 72 filas tienen datos completos, a 13 filas les falta un dato, a 27 filas les faltan dos datos y a 2 filas les faltan tres datos (ver Anexo 2).

Se reconoce que el número de datos faltantes es en general alarmante. Sin embargo, no podemos eliminar las columnas que resultan alarmantes en este sentido porque son parte esencial en la medición de la pobreza multidimensional. Según Little y Rubin (2002), eliminar filas con muchos datos faltantes puede ser beneficioso para la calidad del análisis estadístico; por lo tanto, en cuanto a las filas, que representan a los diferentes países, consideramos adecuado conservar aquellas a las que les faltan como mucho dos del total de los seis datos. Aunque esto último pueda parecer no muy viable, si consideramos solo a aquellas que tienen un dato faltante o no tienen ninguno, podría haber mucho sesgo en la clusterización, ya que probablemente los países con sus registros

completos son países con no mucha pobreza multimendional; por lo tanto, lo que hacemos con esta estrategia es tratar de minimizar el sesgo.

Cuadro 1: Datos faltantes por columna.

Monetaria	L. educativo	M. educativa	Electricidad	Saneamiento	A. potable
0	0	37	3	23	10

Cuadro 2: Datos faltantes por fila.

Fila	D. P.	Fila	D. P.	Fila	D. P.	Fila	D. P.	Fila	D. P.	Fila	D. P.	Fila	D. P.
1	0	17	0	33	0	49	0	65	0	81	0	97	0
2	1	18	0	34	0	50	2	66	0	82	0	98	0
3	0	19	0	35	2	51	2	67	2	83	0	99	0
4	0	20	2	36	2	52	0	68	0	84	0	100	0
5	2	21	2	37	0	53	0	69	2	85	2	101	0
6	2	22	1	38	0	54	0	70	0	86	0	102	0
7	2	23	0	39	0	55	3	71	0	87	0	103	0
8	0	24	2	40	2	56	1	72	0	88	2	104	1
9	0	25	0	41	2	57	2	73	0	89	1	105	0
10	2	26	0	42	3	58	0	74	0	90	0	106	0
11	2	27	2	43	0	59	2	75	2	91	0	107	0
12	0	28	2	44	1	60	0	76	1	92	2	108	1
13	0	29	0	45	2	61	0	77	0	93	0	109	0
14	0	30	2	46	1	62	0	78	0	94	0	110	1
15	1	31	0	47	0	63	2	79	0	95	0	111	0
16	1	32	2	48	0	64	0	80	2	96	1	112	2
												113	0
												114	0

El Cuadro 3 muestra el número de países por región en el conjunto de datos original y en el conjunto de datos filtrados. Se observa que la cantidad de países eliminados no es proporcional al número de países que contenía cada región. Las regiones más afectadas fueron la de SAR y EAP, seguidas de LAC. Los nombres de las regiones son:

- EAP: Este de Asia y el Pacífico.
- ECA: Europa y Asia Central.
- LAC: América Latina y el Caribe.
- MNA: Medio Oriente y Norte de África.
- OHI: Resto del mundo.
- SAR: Asia del Sur.
- SSA: Africa Sub-Sahariana.

Cuadro 3: Número de países por región considerados en los conjuntos de datos.

Conjunto de datos original						
EAP	ECA	LAC	MNA	OHI	SAR	SSA
20	43	18	11	7	6	44
Conjunto de datos filtrado						
EAP	ECA	LAC	MNA	OHI	SAR	SSA
12	40	11	6	6	3	34

Se eliminaron las dos filas a las que les faltaban tres datos; por lo tanto, nos quedamos con 112 filas o países (ver Anexo 3). El Cuadro 4 muestra los países que se tomaron en cuenta para el análisis, con base en las consideraciones anteriores.

Cuadro 4: Países considerados en el análisis.

Angola	Albania	Argentina	Armenia
Australia	Austria	Belgium	Benin
Burkina Faso	Bulgaria	Belarus	Bolivia
Brazil	Botswana	Switzerland	Chile
Côte d'Ivoire	Colombia	Cabo Verde	Cyprus
Czech Republic	Germany	Djibouti	Denmark
Dominican Republic	Egypt, Arab Rep.	Spain	Estonia
Ethiopia	Finland	Fiji	France
Gabon	Georgia	Ghana	Guinea
Gambia	Guinea-Bissau	Greece	Honduras
Croatia	Hungary	Iran, Islamic Rep.	Iceland
Israel	Italy	Kazakhstan	Kenya
Kyrgyz Republic	Kiribati	Korea, Rep.	Lao PDR
Liberia	Sri Lanka	Lesotho	Luxembourg
Latvia	Moldova	Maldives	Mexico
Marshall Islands	North Macedonia	Mali	Malta
Myanmar	Mongolia	Mauritius	Malawi
Malaysia	Namibia	Niger	Nigeria
Netherlands	Norway	Pakistan	Panama
Peru	Philippines	Poland	Portugal
West Bank and Gaza	Romania	Russian Federation	Rwanda
Senegal	Sierra Leone	Serbia	South Sudan
São Tomé and Príncipe	Slovak Republic	Slovenia	Sweden
Eswatini	Seychelles	Chad	Togo
Thailand	Tajikistan	Tonga	Tunisia
Türkiye	Taiwan, China	Tanzania	Uganda
Ukraine	Uruguay	United States	Viet Nam
Vanuatu	Kosovo	Zambia	Zimbabwe

Imputación multivariada de datos faltantes

a. Imputación usando el algoritmo EM

El conjunto de datos debería contener un total de 672 datos, pero le hacen falta 67 (ver Anexo 4). Lerdo (2014) sugiere utilizar el algoritmo EM para tratar los datos faltantes; y, dado que no es método de imputación directa, se puede implementar a la par de métodos estocásticos para generar resultados más precisos. Sin embargo, en general se asume que el patrón de datos faltantes es aleatorio y que la población proviene de una distribución normal multivariada para poder implementar el algoritmo (Lerdo, 2014).

En la Figura 2 (ver código en el Anexo 4) se observa evidencia suficiente para rechazar la normalidad multivariada de los datos y, a continuación, una prueba de Mardia apoya lo mencionado. Las filas que no tenían datos completos fueron eliminadas para la construcción de la gráfica y la prueba. Por lo tanto, es necesario transformar los datos a la normalidad para poder usar el algoritmo EM como método de imputación.

```
# Impresión de la prueba de Mardia.
Test                Statistic   p-value     Result
Mardia Skewness     243.801     2.83e-25    NO
Mardia Kurtosis     11.389      0           NO
MVN                 NA          NA          NO
```

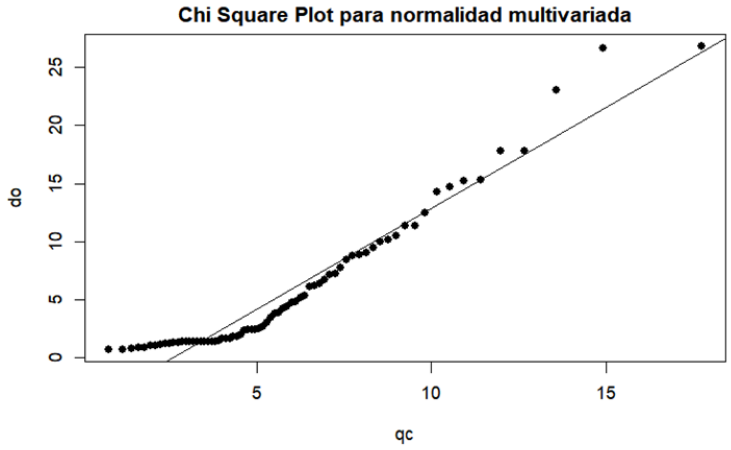

Figura 2: Gráfico chi cuadrado para verificar normalidad multivariada de las filas con datos completos. Se observa que hay evidencia suficiente para rechazar la hipótesis de la normalidad multivariada.

Siguiendo la metodología de Box-Cox para las transformaciones hacia la casi normalidad, la elección $\lambda = 0.1$ parece ser razonable basados en la función a maximizar, como muestra la Figura 3. Por lo tanto, tomamos la transformación de Box-Cox (ver código en Anexo 5):

$$y_i^{(\lambda)} = \frac{y_i^{\lambda} - 1}{\lambda}, \qquad \lambda = 0.1$$

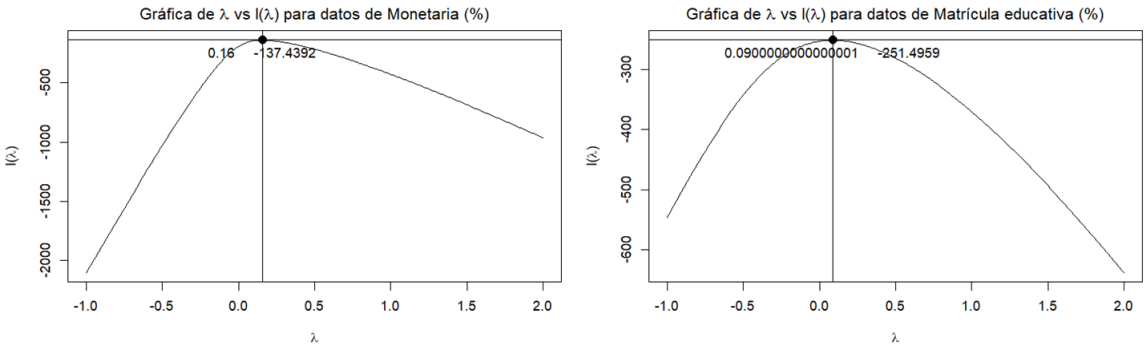

Figura 3: Máximización de la función de Box-Cox para la elección de λ para la transformación. Se muestran solo las gráficas para dos variables, sin embargo, el resto de variables obtuvieron resultados similares.

La Figura 4 muestra que la transformación no fue suficiente para garantizar la normalidad, de hecho, ni siquiera pareció mejorarla; la prueba de Mardia a continuación rechaza la normalidad (ver

código en el Anexo 6). En la gráfica se observa que la observación 72 parece estar considerablemente alejada del resto de puntos, aunque hay una tendencia al respecto.

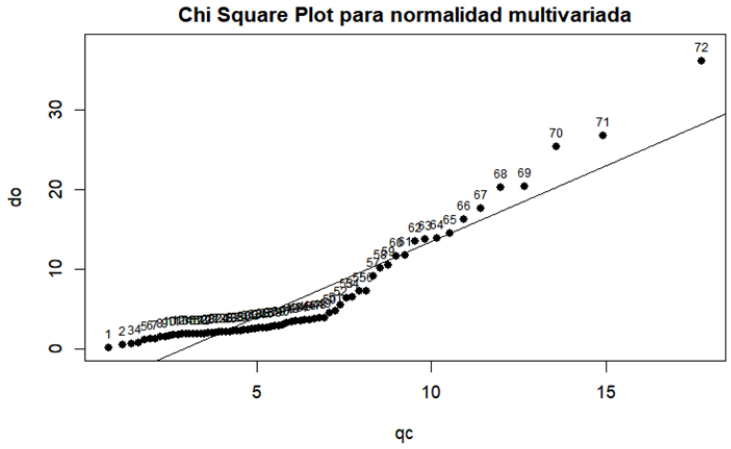

Figura 4: Gráfico chi cuadrado para verificar normalidad multivariada de los datos transformados usando el método de Box-Cox con $\lambda = 0.1$.

```
# Impresión de la prueba de Mardia.
Test                Statistic               p-value                 Result
Mardia Skewness     375.45429231442         7.69795283037624e-49    NO
Mardia Kurtosis     16.2221584256859        0                       NO
MVN                 NA                      NA                      NO
```

La transformación de Box-Cox está diseñada para aproximar los datos a una distribución normal, pero no garantiza que los datos transformados serán perfectamente normales. En algunos casos, aplicar una transformación de Box-Cox puede empeorar la aproximación a la normalidad, especialmente si los datos ya están cerca de ser normales o si tienen una estructura compleja que la transformación no puede abordar adecuadamente. Osborne (2010) discute cómo la transformación puede no siempre mejorar la normalidad y, en algunos casos, puede tener efectos adversos.

Ya que los datos originales no parecían tan alarmantemente alejados de la normalidad, intuitivamente se aplicó una transformación más sencilla, que consiste en la raíz cuadrada. La Figura 5 muestra que la normalidad de los datos ha mejorado, aunque la prueba de Mardia sigue indicando que los datos no son normales (ver código en Anexo 7). Se observa que la observación 72 parece ser un valor atípico, aunque nuevamente se observa una tendencia.

```
# Impresión de la prueba de Mardia.
Test                Statistic               p-value                 Result
Mardia Skewness     86.8006561841091        0.0052012090496009      NO
Mardia Kurtosis     1.64913370582488        0.0991202457538136      YES
MVN                 NA                      NA                      NO
```

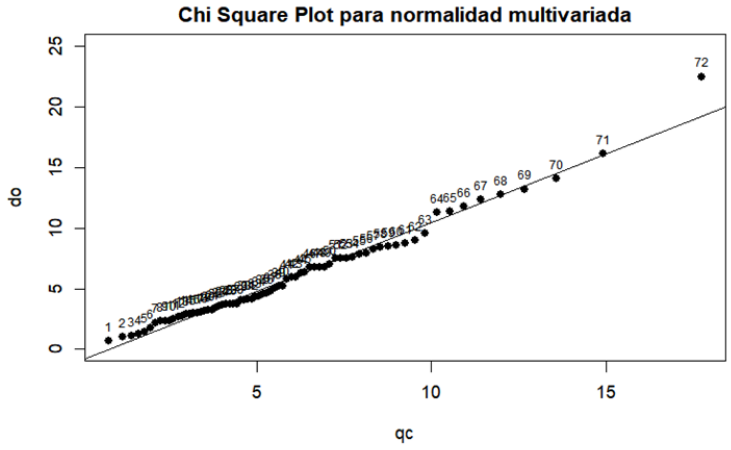

Figura 5: Gráfico chi cuadrado para verificar normalidad multivariada de los datos transformados por la raíz cuadrada. Esto se hizo con el conjunto de datos en el cual todavía no se eliminaban las 4 filas mencionadas. Una raíz de mayor orden alejaba más los datos de la normalidad.

Si se elimina la observación que parece ser atípica y se vuelve a hacer el análisis de normalidad, resulta que ahora la observación 71 parece atípica (ver código en el Anexo 8). Si se elimina esta última y de nuevo se hace el análisis de normalidad, resulta que ahora la observación 70 parece ser atípica; y así sucesivamente. En conclusión, este método no parece ser bueno para la identificación de outliers.

Retomando el objetivo inicial, aplicando el algoritmo EM en R mediante la paquetería `Amelia` al conjunto de datos transformados por la raíz cuadrada, se obtiene un conjunto de datos completo: `dc1` (ver código en el Anexo 9). Ahora, se aplica una destransformación del conjunto de datos que consiste en elevar los elementos al cuadrado. El Cuadro 5 muestra las primeras 10 filas del conjunto de datos completo y destransformado.

b. Imputación usando FSC

Galarza (2013) muestra que el método de imputación múltiple es más robusto que otros métodos y es más preciso ante la ausencia de un número no pequeño de datos faltantes.

Buuren (2007) sugiere que los datos continuos se imputen como continuos y los datos discretos como discretos, siendo la especificación condicional la forma más conveniente de hacerlo; concluye, pues, que a pesar de sus debilidades teóricas, el método de Especificación Totalmente Condicional (FCS) es una alternativa útil y flexible al método de imputación conjunta por modelado (JM) cuando la distribución conjunta de los datos no se especifica fácilmente.

Ya que el patrón de datos faltantes es aleatorio (véase Buuren, 2007), y dado que los datos no siguen una distribución normal, se usará el método de imputación múltiple por especificación totalmente condicional (FCS). Aplicamos el método FSC en R mediante la paquetería `mice`, y

obtenemos un nuevo conjunto de datos completo `dc2` (ver código en el Anexo 10). El Cuadro 5 muestra las primeras 10 filas del conjunto de datos completo.

c. Imputación usando regresión lineal

El método `norm.predict` en la paquetería `mice` usa regresión lineal estándar para imputar valores faltantes. `norm.predict` ajusta un modelo de regresión lineal a cada variable con datos faltantes, usando las otras variables como predictores, y luego predice los valores faltantes a partir de este modelo. Según Van Buuren y Groothuis-Oudshoorn (2011), la imputación por regresión es útil para preservar las relaciones lineales entre variables, lo cual es crucial en estudios multivariados.

El método de regresión no asume que los datos tengan distribución normal, pero sí los errores. Por comodidad, e intentando mejorar los supuestos del modelo, aplicamos la misma transformación que cuando usamos el algorimto EM. Aplicando el método mencionado en R se obtiene un tercer conjunto de datos completo llamado `dc3` (ver código en el Anexo 11). El Cuadro 5 muestra las primeras 10 filas del conjunto de datos completo.

Comparación de los resultados de imputación

El Cuadro 5 muestra las primeras 10 filas de los tres conjuntos de datos completados mediante los métodos de imputación mencionados en las subsecciones anteriores. Se puede ver que, en general, los datos imputados son diferentes (iluminados en color gris), habiendo algunas imputaciones muy similares.

La inconsistencia de las imputaciones puede deberse a las distintas suposiciones y algorimos usados en cada método. El método que usa el algoritmo EM asume que los datos siguen una distribución normal multivariante; algo que no se cumplió en principio y que la transformación mejoró, pero no garantizó; cuando los datos no siguen distribución normal, los resultados pueden ser sesgados. El método de regresión lineal no es adecuado para relaciones no lineales y puede imputar valores fuera del rango de los datos observados, además, asume que los errores tienen distribución normal; aspectos que no se revisaron previo a su implementación.

Cuadro 5: Primeras 10 filas de los tres conjuntos de datos completos.

Monetaria (%)	Logro educativo (%)	Matrícula educativa (%)	Electricidad (%)	Saneamiento (%)	Agua potable (%)
Imputación usando el algoritmo EM					
31.1220	29.7534	27.4431	52.6395	53.6375	32.1065
0.0481	0.1924	1.5340	0.0603	6.5798	9.5950
0.9588	1.0853	0.7314	0.0000	0.1940	0.3640
0.5235	0.0000	1.7930	0.0000	0.3977	0.6601
0.5169	1.7119	0.7676	0.0000	0.0000	0.4003
0.6880	0.1159	3.1710	0.0000	0.5982	0.0000
0.1600	0.4992	0.2334	0.0000	2.3758	0.0000
20.0822	50.2220	31.4680	54.2726	80.0258	22.0691
31.2038	56.4374	50.9392	47.1978	69.6388	19.6883
0.2479	0.5945	0.4359	0.0000	1.9850	0.0000
Imputación usando Especificación Totalmente Condicional (FSC)					
31.1220	29.7534	27.4431	52.6395	53.6375	32.1065
0.0481	0.1924	2.9728	0.0603	6.5798	9.5950
0.9588	1.0853	0.7314	0.0000	0.1940	0.3640
0.5235	0.0000	1.7930	0.0000	0.3977	0.6601
0.5169	1.7119	1.1421	0.0000	0.0000	0.4288
0.6880	0.1159	1.7930	0.0000	4.5385	0.0000
0.1600	0.4992	1.7930	0.0000	4.5385	0.0000
20.0822	50.2220	31.4680	54.2726	80.0258	22.0691
31.2038	56.4374	50.9392	47.1978	69.6388	19.6883
0.2479	0.5945	0.3796	0.0000	0.0670	0.0000
Imputación usando regresión lineal					
31.1220	29.7534	27.4431	52.6395	53.6375	32.1065
0.0481	0.1924	2.0214	0.0603	6.5798	9.5950
0.9588	1.0853	0.7314	0.0000	0.1940	0.3640
0.5235	0.0000	1.7930	0.0000	0.3977	0.6601
0.5169	1.7119	1.2467	0.0000	0.0000	0.2681
0.6880	0.1159	0.2909	0.0000	0.6492	0.0000
0.1600	0.4992	0.5018	0.0000	0.8920	0.0000
20.0822	50.2220	31.4680	54.2726	80.0258	22.0691
31.2038	56.4374	50.9392	47.1978	69.6388	19.6883
0.2479	0.5945	0.5212	0.0000	0.9059	0.0000

Cuadro 6: Medias obtenidas con cada método.

Monetaria (%)	Logro educativo (%)	Matrícula educativa (%)	Electricidad (%)	Saneamiento (%)	Agua potable (%)
Imputación usando el algoritmo EM					
9.521	12.138	7.936	15.261	23.096	8.900
Imputación usando Especificación Totalmente Condicional (FSC)					
9.521	12.138	8.182	15.025	23.005	8.880
Imputación usando regresión lineal					
9.521	12.138	7.860	15.209	23.061	8.866

Cuadro 7: Matrices de varianzas y covarianzas de los tres conjuntos de datos.

Monetaria (%)	Logro educativo (%)	Matrícula educativa (%)	Electricidad (%)	Saneamiento (%)	Agua potable (%)
		Imputación usando el algoritmo EM			
249.323	194.615	122.112	358.992	361.878	130.935
194.615	329.670	176.515	362.498	431.422	144.679
122.112	176.515	161.211	230.871	256.264	95.051
358.992	362.498	230.871	632.327	644.745	219.414
361.878	431.422	256.264	644.745	862.005	272.093
130.935	144.679	95.051	219.414	272.093	133.866
		Imputación usando Especificación Totalmente Condicional (FSC)			
249.323	194.615	119.859	342.609	362.728	131.087
194.615	329.670	173.693	354.616	432.136	144.926
119.859	173.693	158.067	211.907	251.777	93.110
342.609	354.616	211.907	592.770	627.220	218.023
362.728	432.136	251.777	627.220	865.818	273.313
131.087	144.926	93.110	218.023	273.313	134.334
		Imputación usando regresión lineal			
249.323	194.615	122.805	355.863	362.193	131.254
194.615	329.670	177.433	361.026	431.776	145.084
122.805	177.433	162.134	229.092	258.133	96.007
355.863	361.026	229.092	623.464	641.744	219.665
362.193	431.776	258.133	641.744	863.399	273.220
131.254	145.084	96.007	219.665	273.220	134.420

El Cuadro 6 muestra las medias para conjunto de datos completado. Las medias de las primeras dos columnas coinciden porque eran columnas completas. En general, las medias son muy parecidas, pero hay una mayor similitud entre los métodos 1 y 3 (EM y regresión).

El Cuadro 7 muestra las matrices de varianzas y covarianzas para cada método. Nuevamente, las varianzas de las variables 1 y 2 permanecen iguales porque eran columnas con datos completos. Por otra parte, se observa que las varianzas y covarianzas son similares, siendo nuevamente los métodos 1 y 3 (EM y regresión) los que tienen mayor similitud. Se observa también que hay mayor diferencia entre las varianzas de la variable con mayor cantidad de datos ausentes (variables 3), pero no es el caso de la segunda variable con mayor datos ausentes (variable 5).

El Cuadro 8 muestra las matrices de correlaciones para los tres conjuntos de datos. Las correlaciones son muy parecidas entre los diferentes métodos, llegado a coincidir en algunos casos. Nuevamente, las mayores similitudes las encontramos entre el método que usa el algoritmo EM y el de regresión.

En general, las estadísticas de resumen son prácticamente iguales. Esto es bueno para el análisis descriptivo de la pobreza multidimensional, ya que cualquier método daría las mismas conclusiones.

Cuadro 8: Matrices de correlaciones de los tres conjuntos de datos.

Monetaria (%)	Logro educativo (%)	Matrícula educativa (%)	Electricidad (%)	Saneamiento (%)	Agua potable (%)
		Imputación usando el algoritmo EM			
1.000	0.679	0.609	0.904	0.781	0.717
0.679	1.000	0.766	0.794	0.809	0.689
0.609	0.766	1.000	0.723	0.687	0.647
0.904	0.794	0.723	1.000	0.873	0.754
0.781	0.809	0.687	0.873	1.000	0.801
0.717	0.689	0.647	0.754	0.801	1.000
		Imputación usando Especificación Totalmente Condicional (FSC)			
1.000	0.679	0.604	0.891	0.781	0.716
0.679	1.000	0.761	0.802	0.809	0.689
0.604	0.761	1.000	0.692	0.681	0.639
0.891	0.802	0.692	1.000	0.876	0.773
0.781	0.809	0.681	0.876	1.000	0.801
0.716	0.689	0.639	0.773	0.801	1.000
		Imputación usando regresión lineal			
1.000	0.679	0.611	0.903	0.781	0.717
0.679	1.000	0.767	0.796	0.809	0.689
0.611	0.767	1.000	0.721	0.690	0.650
0.903	0.796	0.721	1.000	0.875	0.759
0.781	0.809	0.690	0.875	1.000	0.802
0.717	0.689	0.650	0.759	0.802	1.000

Para un análisis más complejo, donde ya se considera importante cada observación, usaremos los resultados obtenidos con el método de Especificación Totalmente Condicional (FSC), ya que, como se argumentó, es un método robusto y no parece tener problemas con los supuestos para su implementación, como los otros dos métodos usados.

Identificación y eliminación de valores atípicos

Una forma de identificar los valores atípicos es haciendo gráficos de dispersión, donde los puntos deben aproximarse a formar un elipse o elipsoide y los puntos más alejados se consideran outliers; sin embargo, esta técnica asume que los datos tienen distribución normal multivariada, pero no es este el caso. Otra técnica es hacer gráficos de control de calidad, sin embargo, en sus versiones simples también asume normalidad de los datos.

Otra forma de identificar valores atípicos es mediante el análisis de componentes principales. Los datos transformados por componentes principales se grafican y se observa si hay puntos muy alejados del resto.

El Cuadro 9 muestra la importancia de los componentes principales. Se observa que los dos primeros componentes principales explican el 87.1 % de la varianza total.

Cuadro 9: Importancia de los componentes.

	PC1	PC2	PC3	PC4	PC5	PC6
Desviación estándar	2.1788	0.69308	0.55777	0.50098	0.37280	0.26697
Proporción de varianza	0.7912	0.08006	0.05185	0.04183	0.02316	0.01188
Proporción acumulada de varianza	0.7912	0.87128	0.92313	0.96496	0.98812	1.00000

La Figura 6 muestra la gráfica de los dos primeros componentes principales para los datos transformados (ver código en el Anexo 13). Esta gráfica no resulta tan ilustrativa para la identificación de valores atípicos, por lo tanto, se procederá a emplear otros métodos.

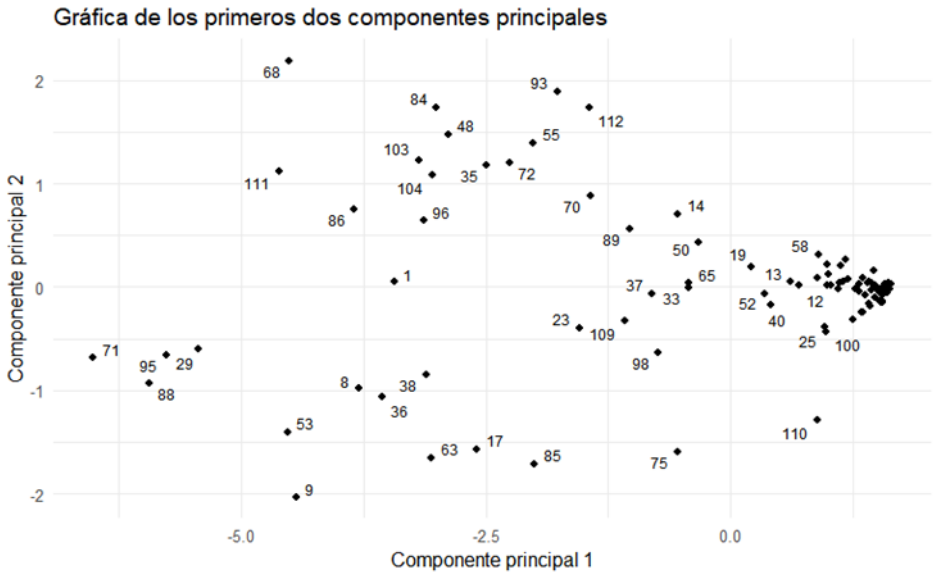

Figura 6: Gráfico de los dos primeros componentes principales. No resulta especialmente clara para la identificación de valores atípicos.

El método Local Outlier Factor (LOF) es utilizado para identificar outliers basándose en la densidad local de los datos. Los outliers son observaciones que tienen una densidad significativamente menor que sus vecinos. LOF es especialmente útil para detectar outliers en datasets de alta dimensión y con formas complejas (Breunig, 2000).

La distancia de Mahalanobis es una medida de distancia multidimensional que toma en cuenta las correlaciones entre variables. Se utiliza para identificar outliers en datos multivariados y es eficaz cuando los datos siguen una distribución aproximadamente normal.

Isolation Forest es un método de detección de outliers basado en la idea de aislar puntos de datos mediante la construcción de árboles de partición aleatoria. Los outliers son más fáciles de aislar y, por lo tanto, tienden a tener caminos de aislamiento más cortos en comparación con los puntos normales.

El método de clustering por K-medias se puede utilizar para detectar outliers basándose en la distancia de las observaciones a los centroides de los clusters. Los puntos que se encuentran a una gran distancia de cualquier centroide pueden considerarse outliers. Aunque este método no está específicamente diseñado para la detección de outliers, puede ser efectivo en algunos contextos.

El Cuadro 10 muestra los outliers arrojados por cada uno de los métodos mencionados (ver código en el Anexo 14). Para la consideración final de los outliers usaremos el criterio de eliminar aquellos que aparezcan en al menos dos de las pruebas y además estén considerablemente alejados del resto de puntos según la Figura 7; dichos puntos son: 63, 68, 71, 84, 88, 95 y 98.

Cuadro 10: Resultados de 4 métodos distintos en la identificación de outliers.

Local Outlier Factor	11	22	40	43	49	59	62	81	94	98	110
Distancia de Mahalanobis	35	53	63	68	70	84	88	95	98	111	
Insolation Forest	29	68	71	88	95	98					
K-medias	63	68	71	84	88	98					

Eliminando las filas correspondientes a los puntos que fueron identificados como outliers (ver código en el Anexo 15) se obtiene finalmente un conjunto de datos de 105 países con mediciones en 6 variables. Los países considerados finalmente para el análisis se muestran en el Cuadro 11.

Cuadro 11: Países considerados en el análisis.

Angola	Albania	Argentina	Armenia
Australia	Austria	Belgium	Benin
Burkina Faso	Bulgaria	Belarus	Bolivia
Brazil	Botswana	Switzerland	Chile
Côte d'Ivoire	Colombia	Cabo Verde	Cyprus
Czech Republic	Germany	Djibouti	Denmark
Dominican Republic	Egypt, Arab Rep.	Spain	Estonia
Ethiopia	Finland	Fiji	France
Gabon	Georgia	Ghana	Guinea
Gambia	Guinea-Bissau	Greece	Honduras
Croatia	Hungary	Iran, Islamic Rep.	Iceland
Israel	Italy	Kazakhstan	Kenya
Kyrgyz Republic	Kiribati	Korea, Rep.	Lao PDR
Liberia	Sri Lanka	Lesotho	Luxembourg
Latvia	Moldova	Maldives	Mexico
Marshall Islands	North Macedonia	Malta	Myanmar
Mongolia	Mauritius	Malaysia	Namibia
Nigeria	Netherlands	Norway	Pakistan
Panama	Peru	Philippines	Poland
Portugal	West Bank and Gaza	Romania	Russian Federation
Senegal	Sierra Leone	Serbia	São Tomé and Príncipe
Slovak Republic	Slovenia	Sweden	Eswatini
Seychelles	Togo	Thailand	Tonga
Tunisia	Türkiye	Taiwan, China	Tanzania
Uganda	Ukraine	Uruguay	United States
Viet Nam	Vanuatu	Kosovo	Zambia
Zimbabwe			

Como se puede observar en el Cuadro 12, en esta última eliminación de datos la región SSA

(África Sub-Sahariana) fue la más afectada.

Cuadro 12: Número de países por región considerados en los conjuntos de datos.

Conjunto de datos original						
EAP	ECA	LAC	MNA	OHI	SAR	SSA
20	43	18	11	7	6	44
Conjunto de datos filtrado						
EAP	ECA	LAC	MNA	OHI	SAR	SSA
12	40	11	6	6	3	34
12	39	11	6	6	3	28

Estadísticas descriptivas

La tabla de abajo muestra las medias de cada indicador considerando el conjunto de los 105 países. La pobreza suele asociarse vulgarmente a la pobreza monetaria; no obstante, la tabla muestra que, en términos de pobreza multidimensional, son más preocupantes las carencias de saneamiento y electricidad en los hogares.

Indicador	**Media (%)**
Monetaria	7.370
Nivel educativo	9.692
Escolarización	6.742
Electricidad	12.017
Saneamiento	20.652
Agua potable	7.696

En la matriz de varianzas y covarianzas muestral se observa que hay una mayor variabilidad en los indicadores de saneamiento y electricidad, respectivamente. Se observa una menor variabilidad en la escolarización y el agua potable. Resulta interesante señalar que las variables con mayor varianza son también los de las medias más altas, y las de menor varianza son las de medias más bajas.

$$\mathbf{S} = \begin{bmatrix} 155.425 & 112.460 & 79.791 & 224.830 & 272.394 & 110.323 \\ 112.460 & 216.460 & 127.116 & 229.971 & 324.755 & 109.882 \\ 79.791 & 127.116 & 113.291 & 159.979 & 195.048 & 71.967 \\ 224.830 & 229.971 & 159.979 & 425.686 & 499.283 & 184.154 \\ 272.394 & 324.755 & 195.048 & 499.283 & 759.051 & 251.645 \\ 110.323 & 109.882 & 71.967 & 184.154 & 251.645 & 113.904 \end{bmatrix}$$

En la matriz de correlaciones muestral (ver código en el Anexo 16) se observa que las variables tienen una buena asociación lineal directa entre sí , lo cual es esperable. Se destaca que las carencias de electricidad y saneamiento tienen la correlación más alta (0.878), seguido por saneamiento y agua potable (0.856). De nuevo, vulgarmente se suele pensar que la pobreza monetaria tiene la mayor

asociación con cualquier otro tipo de pobreza, sin embargo, la matriz de correlaciones muestra que la carencia de saneamiento la supera en todas las correlaciones con las demás variables. La correlación más baja se da entre pobreza monetaria y escolarización, y aún así es considerable (0.601).

$$\mathbf{R} = \begin{bmatrix} 1.000 & 0.613 & 0.601 & 0.874 & 0.793 & 0.829 \\ 0.613 & 1.000 & 0.812 & 0.758 & 0.801 & 0.700 \\ 0.601 & 0.812 & 1.000 & 0.728 & 0.665 & 0.634 \\ 0.874 & 0.758 & 0.728 & 1.000 & 0.878 & 0.836 \\ 0.793 & 0.801 & 0.665 & 0.878 & 1.000 & 0.856 \\ 0.829 & 0.700 & 0.634 & 0.836 & 0.856 & 1.000 \end{bmatrix}$$

Se destaca que las asociaciones entre las variables son positivas, lo cual es de esperarse ya que las variables tienen el mismo sentido: todas miden un cierto nivel de carencias o pobreza en un determinado ámbito.

Un análisis previo que consideró a los 149 países arrojó medias considerablemente más altas en todas las variables, esto pudo deberse a no haber eliminado países con todas las consideraciones hechas en este análisis: vecindad temporal, número de datos ausentes, outliers, etc. Sin embargo, resulta intuitivo que los países con mayor pobreza multidimensional hagan estudios de pobreza no tan frecuentemente, carezcan de algunos registros en los indicadores o resulten tener mediciones muy atípicas. Esto, desde luego que representa un sesgo en la presente investigación, sin embargo, se ha justificado en cada paso la toma de decisiones, donde hemos priorizado la validez estadística.

Análisis factorial

En esta sección no se pretende realizar un análisis factorial exhaustivo, sino más bien, dar una herramienta estadística más para la interpretación de los datos, en este caso, para identificar factores. En este sentido, se implementaron los métodos de máxima verosimilitud y componentes principales para obtener las cargas factoriales estimadas rotadas. Se usó la matriz de correlaciones muestral **R**, ya que **S** presenta varianzas considerablemente distintas.

Un análisis de componentes principales muestra que los primeros dos componentes explican un porcentaje de varianza acumulada del 89.76 %, y los tres primeros explican el 94.1 % (ver Cuadro 13). Y una *Scree plot* en la figura de abajo muestra que dos componentes paracen ser adecuados para el análisis (ver código en el Anexo 17).

Cuadro 13: Importancia de los componentes.

	PC1	PC2	PC3	PC4	PC5	PC6
Desviación estándar	2.1915	0.76322	0.5103	0.40597	0.33394	0.27913
Proporción de varianza	0.8005	0.09708	0.0434	0.02747	0.01859	0.01299
Proporción acumulada de varianza	0.8005	0.89756	0.9410	0.96843	0.98701	1.00000

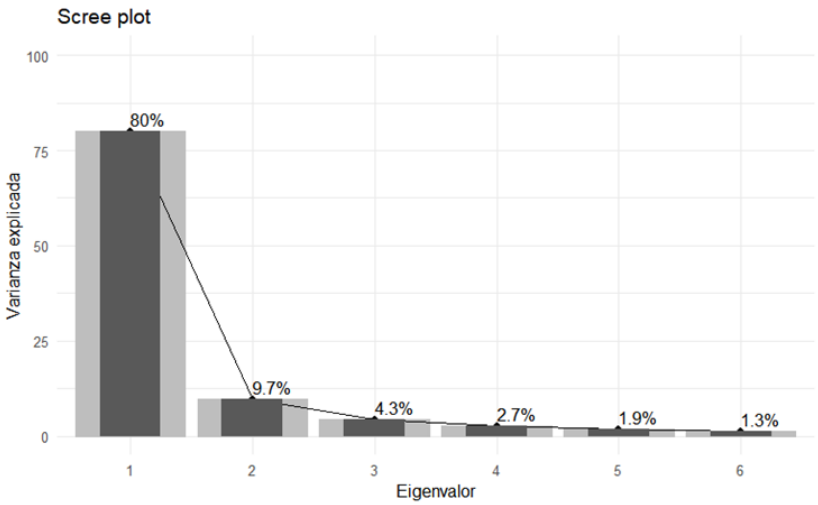

Figura 7: Scree plot para eigenvalores de la matriz de correlaciones **R**.

El Cuadro 14 presenta las cargas rotadas para un modelo con $m = 2$ factores usando los métodos mencionados (ver código en el Anexo 18). Las cargas más grandes en el primer factor son de pobreza monetaria y carencias de electricidad, saneamiento y agua potable, mientras que en el segundo factor las cargas más grandes son las correspondientes a educación. Podríamos nombrar a estos factores respectivamente como *factor material* y *factor educativo*. En esta interpretación ambos métodos son consistentes, sin embargo, algunas variables siguen teniendo un peso no pequeño en aquellos factores que en los que no se mencionaron como cargas significativas. Hay una buena proporción de varianza explicada en ambos casos, siendo el de componentes principales el que explica más.

Cuadro 14: Cargas factoriales para un modelo con $m = 2$ factores.

	Máxima verosimilitud		Componentes principales	
Monetaria	0.873	0.316	0.907	0.278
Nivel educativo	0.367	0.928	0.435	0.846
Escolarización	0.429	0.705	0.339	0.888
Electricidad	0.814	0.495	0.813	0.502
Saneamiento	0.723	0.578	0.788	0.511
Agua potable	0.782	0.445	0.856	0.384
SC cargas	2.877	2.234	3.141	2.242
Prop. var. exp.	0.479	0.372	0.524	0.374
Prop. var. exp. ac.	0.479	0.852	0.524	0.897

El Cuadro 15 muestra las cargas rotadas para un modelo con $m = 3$ factores usando nuevamente ambos métodos (ver código en el Anexo 18). En el primer factor las cargas más grandes son las de pobreza monetaria y carencias de electricidad y agua potable, similarmente al caso del modelo con $m = 2$ factores. En el segundo factor las cargas más altas están dadas por las variables de carencias educativas. Por último, en el tercer factor las cargas más grandes son para saneamiento y nivel educativo. Resulta un poco más complicado interpretar estos factores, sin embargo, abusando un poco de los resultados, los podríamos identificar respectivamente como *factor material*, *factor educativo* y *factor de salubridad y nivel educativo*. En esta interpretación ambos métodos son consistentes. Un modelo con $m = 3$ factores explica más del 90 % de la varianza.

Cuadro 15: Cargas factoriales para un modelo con $m = 3$ factores.

	Máxima verosimilitud			Componentes principales		
Monetaria	0.907	0.284	0.289	0.914	0.291	-0.185
Nivel educativo	0.299	0.667	0.525	0.291	0.644	-0.675
Escolarización	0.308	0.898	0.229	0.342	0.896	-0.241
Electricidad	0.667	0.460	0.478	0.767	0.439	-0.376
Saneamiento	0.515	0.369	0.764	0.640	0.303	-0.667
Agua potable	0.637	0.354	0.520	0.760	0.251	-0.498
SC cargas	2.124	1.805	1.494	2.613	1.650	1.382
Prop. var. exp.	0.354	0.301	0.249	0.435	0.275	0.230
Prop. var. exp. ac.	0.354	0.655	0.904	0.435	0.710	0.941

En general se pudo observar una distinción entre *factor material* y *factor educativo*, y en el modelo con $m = 3$ factores se asomó el *factor de saneamiento* (aunque estaba conectado con la variable de nivel educativo).

Análisis de clustering

El método del codo (elbow method) es una técnica visual para determinar el número óptimo de clusters. Este método evalúa la suma total de los errores cuadráticos dentro de los clusters (Within-Cluster Sum of Squares, WSS) para diferentes números de clusters k. La WSS mide la compactación de los clusters, es decir, la suma de las distancias al cuadrado de cada punto a su centroide dentro de un cluster.

La Figura 8 muestra que considerar 3 clusters para el análisis parece ser adecuado. Y la Figura 9 muestra el resultado de la clusterización usando el método de las k-medias con $k = 3$.

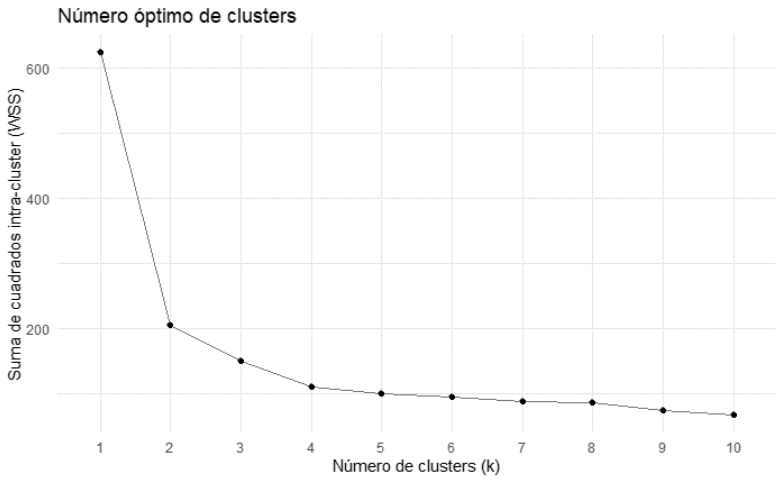

Figura 8: Gráfica del método del codo para la determinación del número óptimo de clusters.

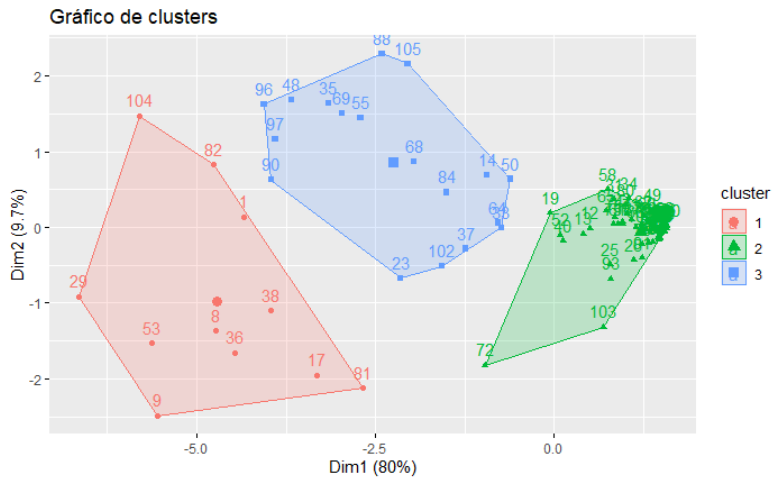

Figura 9: Gráfico de clusters usando el método de las k-medias con $k = 3$.

El primer cluster (marcado con color rojo en la Figura 9) consta de 11 países, donde los 11 son países del África Sub-Sahariana (SSA). El segundo cluster (marcado con color verde en la Figura 9) consta de 76 países e incluye una mezcla de países de las 6 regiones mencionadas anteriormente, pero prevalece Europa y Asia Central (ECA), y en menor medida América Latina y el Caribe (LAC) y el Este de Asia y el Pacífico (EAP). El tercer cluster (marcado con color azul en la Figura 9) consta de 18 países, en los que predominan países del África Sub-Sahariana (SSA), con algunos del Este de Asia y el Pacífico (EAP).

Cuadro 16: Países que conforman cada cluster.

Cluster	**Países**
1	Angola, Benin, Burkina Faso, Côte d'Ivoire, Ethiopia, Guinea, Guinea-Bissau, Liberia, Senegal, Sierra Leone, Zambia
2	Albania, Argentina, Armenia, Australia, Austria, Belgium, Bulgaria, Belarus, Bolivia, Brazil, Switzerland, Chile, Colombia, Cabo Verde, Cyprus, Czech Republic, Germany, Denmark, Dominican Republic, Egypt, Spain, Estonia, Finland, Fiji, France, Georgia, Greece, Honduras, Croatia, Hungary, Iran, Iceland, Israel, Italy, Kazakhstan, Kyrgyz Republic, Korea, Lao PDR, Sri Lanka, Luxembourg, Latvia, Moldova, Maldives, Mexico, Marshall Islands, North Macedonia, Malta, Mongolia, Mauritius, Malaysia, Netherlands, Norway, Pakistan, Panama, Peru, Philippines, Poland, Portugal, West Bank and Gaza, Romania, Russian Federation, Serbia, Slovak Republic, Slovenia, Sweden, Seychelles, Thailand, Tonga, Tunisia, Türkiye, Taiwan, Ukraine, Uruguay, United States, Viet Nam, Kosovo
3	Botswana, Djibouti, Gabon, Ghana, Gambia, Kenya, Kiribati, Lesotho, Myanmar, Namibia, Nigeria, São Tomé and Príncipe, Eswatini, Togo, Tanzania, Uganda, Vanuatu, Zimbabwe

El Cuadro 18 muestra que el cluster 1 tiene las mayores medias de niveles de pobreza en las 6 dimensiones consideradas en este estudio. Siendo más alarmante cualquier dimensión de la pobreza antes que la dimensión monetaria. Resulta muy grave el nivel de carencias de saneamiento en los

Cuadro 17: Número de países por región en cada cluster.

Cluster	EAP	ECA	LAC	MNA	OHI	SAR	SSA
1	0	0	0	0	0	0	11
2	9	39	11	5	6	3	3
3	3	0	0	1	0	0	14

hogares, de las cuales padecen dos terceras partes de la población, que no tienen acceso a un nivel de saneamiento en sus hogares, aunque sea limitado. El nivel educativo también es alarmante, e indica que en este cluster el porcentaje de personas que viven en hogares en donde ningún adulto ha completado la educación primaria es de 43.5 %. En general, en este cluster las medias de cada indicador resultan alarmanetes.

El cluster 2 tiene las medias de los indicadores más bajos, interpretándose esto en que tienen en promedio el menor nivel de pobreza multidimensional. Por ejemplo, la media de hogares con carencia de energía eléctrica es muy baja (si observamos el conjunto de datos original notaremos que hay países que tienen 0 en este indicador). El nivel educativo y la escolarización también son muy bajos, estando relativamente cerca del 0. Y la pobreza monetaria tampoco parece ser significativa en este cluster.

El cluster 3 tiene cierto parecido con el cluster 1 en cuanto a medias alarmantes de pobreza en sus dimensiones de saneamiento, electricidad, agua potable y monetaria, respectivamente, sin embargo, difiere de ella en que sus medias referentes a las dimensiones de nivel educativo y escolarización son considerablemente más pequeñas que en el cluster 1.

Cuadro 18: Medias de cada cluster.

Cluster	Monetaria	Nivel educativo	Escolarización	Electricidad	Saneamiento	Agua potable
1	25.455318	43.596301	32.871950	52.6418107	67.643241	26.258995
2	1.182167	2.940417	2.538754	0.8607708	4.994838	2.330259
3	22.444653	17.479996	8.517838	34.2923461	58.040735	19.007656

El Cuadro 19 presenta las varianzas de cada variable en cada cluster. Las mayores varianzas las tiene el cluster 1, y las menores, el cluster 2. Medias más altas en el Cuadro 18 coinciden con varianzas más altas en cluster e índice en el Cuadro 19, y medias más bajas coinciden con varianzas más bajas.

Cuadro 19: Varianzas en cada cluster.

Cluster	Monetaria	Nivel educativo	Escolarización	Electricidad	Saneamiento	Agua potable
1	195.799963	204.33549	109.56015	342.640495	258.75714	68.97773
2	4.237679	21.87477	17.99830	4.262406	58.57004	12.60998
3	193.538751	95.72072	25.10741	234.156082	228.08326	113.45796

Resulta intuitivo que los países con niveles de ingresos más bajos y niveles de pobreza más altos, que resultan ser también los políticamente más inestables, sean los que tengan diferencias más marcadas entre ellos, como se puede ver en la Figura 9: el cluster 1, con menor cantidad de países es el que forma el polinomio de mayor área y tiene los puntos más alejados entre sí. En este sentido, esto es intuitivo en el contexto de los problemas que se viven en la región Sub-Sahariana

del África. Sin embargo, no resulta intuitivo la consideración del cluster 1, donde se nos presentan países de Europa y países de América Latina dentro del mismo grupo. Considerando esto, resulta interesante hacer un nuevo análisis de clustering solo para el cluster 2.

La Figura 10 muestra que considerar 3 clusters para el análisis parece ser adecuado. Y la Figura 11 muestra el resultado de la clusterización usando el método de las k-medias con $k = 3$.

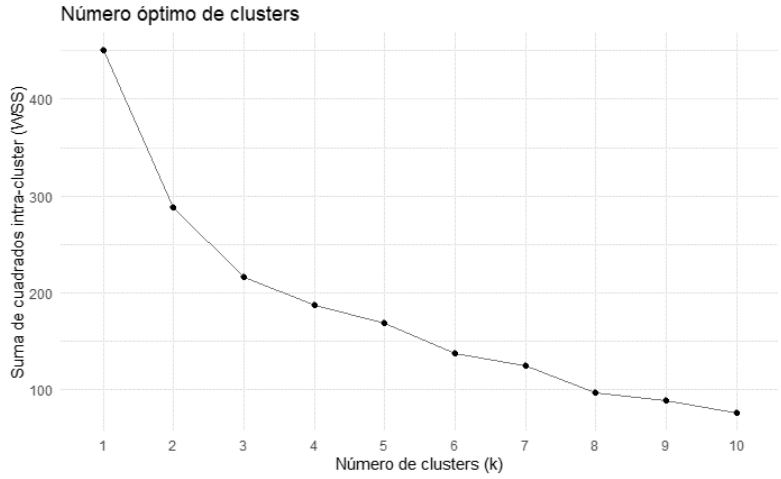

Figura 10: Gráfica del método del codo para la determinación del número óptimo de clusters.

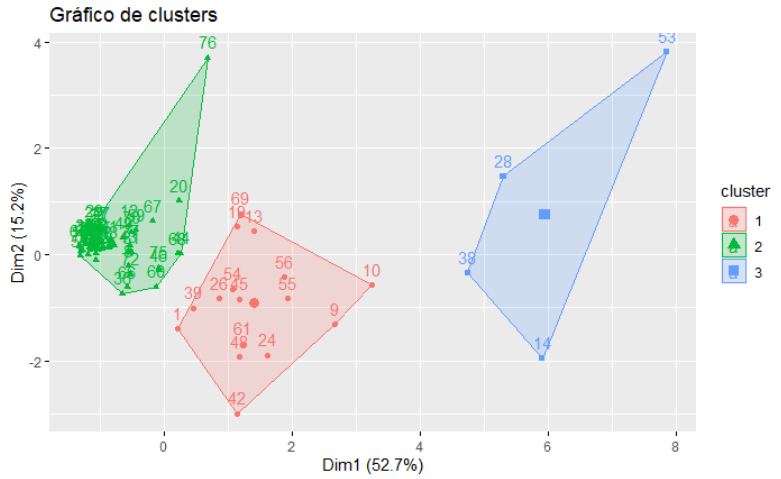

Figura 11: Gráfico de clusters usando el método de las k-medias con $k = 3$.

El primer cluster (marcado con color rojo en la Figura 11), consta de 16 países; lo constituyen principalmente países de América Latina y el Caribe (LAC), Este de Asia y el Pacífico (EAP) y Europa y Asia Central (ECA). El segundo cluster (color verde en la Figura 11) consta de 56 países; lo constituyen principalmente países de Europa y Asia Central (ECA), y además incluye algunos países de todas las demás regiones. El tercer cluster (color azul en la Figura 11) consta de 4 países; y resulta también una mezcla de varias regiones.

Cuadro 20: Países que conforman cada cluster.

Cluster	**Países**
1	Albania, Bolivia, Brazil, Colombia, Dominican Republic, Fiji, Georgia, Sri Lanka, Moldova, Marshall Islands, Mongolia, Panama, Peru, Philippines, Russian Federation, Tunisia
2	Argentina, Armenia, Australia, Austria, Belgium, Bulgaria, Belarus, Switzerland, Chile, Cyprus, Czech Republic, Germany, Denmark, Egypt, Arab Rep., Spain, Estonia, Finland, France, Greece, Croatia, Hungary, Iran, Islamic Rep., Iceland, Israel, Italy, Kazakhstan, Kyrgyz Republic, Korea, Rep., Luxembourg, Latvia, Maldives, Mexico, North Macedonia, Malta, Mauritius, Malaysia, Netherlands, Norway, Poland, Portugal, West Bank and Gaza, Romania, Serbia, Slovak Republic, Slovenia, Sweden, Seychelles, Thailand, Tonga, Türkiye, Taiwan, China, Ukraine, Uruguay, United States, Viet Nam, Kosovo
3	Cabo Verde, Honduras, Lao PDR, Pakistan

Cuadro 21: Número de países por región en cada cluster.

Cluster	EAP	ECA	LAC	MNA	OHI	SAR	SSA
1	4	4	6	1	0	1	0
2	4	35	4	4	6	1	2
3	1	0	1	0	0	1	1

El Cuadro 22 muestra que el tercer cluster tiene las medias de indicadores más alarmantes (aunque no resultan tan extremas como en la clusterización anterior). Las medias de los índices del nivel educativo en este cluster son parecidas al cluster 3 de la clusterización anterior, pero hay una diferencia más significativa en cuanto a las medias de los indicadores de pobreza monetaria y carencias de electricidad, saneamiento y agua potable.

El cluster 2 tiene las medias más pequeñas, estando todas muy cercanas de 0, especialmente la que refiere a pobreza monetaria y carencias de electricidad y agua potable. El primer cluster parece ser algo intermedio entre el cluster 1 y el 3, con una pobreza monetaria y escolar relativamente baja, pero carencias de saneamiento y agua potable significativas.

En el Cuadro 23 se observa que el cluster 3 tiene las varianzas más grandes casi en todos los indicadores, lo cual es esperable; sin embargo, en cuanto a nivel educativo y agua potable el cluster 1 tiene mayores varianzas a pesar de tener menores medias, lo cual indica que los países que conforman este cluster tienen mayores diferencias en cuanto a estos indicadores que los otros clusters.

Cuadro 22: Medias de cada cluster.

Cluster	Monetaria	Nivel educativo	Escolarización	Electricidad	Saneamiento	Agua potable
1	1.9575263	5.222022	2.269537	1.7835364	11.925138	6.7267818
2	0.5221638	1.503500	1.954971	0.1142748	1.882721	0.6857135
3	7.3207769	13.930834	11.788579	7.6206524	20.843276	7.7677970

Cuadro 23: Varianzas en cada cluster.

Cluster	Monetaria	Nivel educativo	Escolarización	Electricidad	Saneamiento	Agua potable
1	4.566149	38.659412	1.945445	4.0821492	78.01848	17.382054
2	0.415989	5.573759	9.898322	0.2271174	10.44681	1.618042
3	13.903818	24.028726	137.933723	6.1141623	110.78965	5.677924

En general, se observa de la clusterización que no es posible generalizar una región al decir que todos los países que la conforman tienen un cierto nivel de pobreza multidimensional. En cambio, se observa diversidad en las regiones. Aunque hay regiones que resultan ser más homogéneas en este sentido, y hablar de tendencias permite obtener conclusiones generales.

Discusión

El análisis de la pobreza multidimensional en este estudio ha revelado una serie de hallazgos significativos que proporcionan una visión profunda sobre las diversas dimensiones de la pobreza en diferentes regiones del mundo. Este enfoque integral permite identificar no solo las regiones que tienen mayores niveles de pobreza, sino las dimensiones específicas en que esta se manifiesta.

El análisis de los clusters ha permitido identificar patrones regionales significativos. Los países del África Sub-Sahariana (SSA) muestran niveles alarmantes de carencias en todas las dimensiones consideradas, especialmente en saneamiento y nivel educativo. Esto coincide con informes recientes del Banco Mundial y otras organizaciones internacionales que destacan la necesidad urgente de mejorar las infraestructuras básicas y los sistemas educativos en esta región.

En contraste, los países del cluster 2, que incluye principalmente a naciones de Europa y Asia Central, muestran los niveles más bajos de pobreza multidimensional. Esto refleja los avances significativos en desarrollo económico y social que estas regiones han experimentado en las últimas décadas. Sin embargo, es importante señalar que, a pesar de los bajos niveles de pobreza, aún existen áreas donde se puede mejorar, como la integración social y el acceso a servicios de calidad. Al respecto de esto, la segunda clusterización arrojó más luz sobre cuáles son esas regiones que requieren una mayor atención.

Medir la pobreza multidimensional es muy importante para captar la estructura de la pobreza (es decir, en sus diferentes contextos). Tradicionalmente la pobreza suele asociarse con la pobreza monetaria; sin embargo, este enfoque reconoce que la privación puede ocurrir en varias dimensiones, tales como la salud, la educación, las condiciones de vida y la participación social. Esto permite una comprensión más precisa de las necesidades específicas de las poblaciones de cada región y facilita la elaboración de políticas más efectivas.

El análisis también revela varios desafíos en la medición y el tratamiento de la pobreza multidimensional. La falta de datos completos y actualizados es un obstáculo significativo, especialmente en regiones con altos niveles de pobreza. La imputación de datos faltantes, aunque útil, puede introducir sesgos que afectan la presición de los resultados. En este sentido, es esencial seguir mejorando la recopilación de datos para obtener una imagen más clara y precisa de la pobreza.

Además, la identificación de valores atípicos es un paso crítico para garantizar la validez del análisis estadístico. Los métodos de análisis, comoi el algoritmo EM y la imputación por FSC han demostrado ser útiles en este estudio, pero también es necesario seguir explorando y desarrollando técnicas más robustas para manejar los datos faltantes y atípicos.

Los hallazgos de este estudio tienen implicaciones importantes para la formulación de políticas públicas. Los gobiernos y las organizaciones internacionales deben centrarse en abordar las dimensiones específicas de la pobreza identificadas en cada región. Por ejemplo, en el África Sub-

Sahariana, las políticas deberían priorizar la mejora de las infraestructuras de saneamiento y el acceso a la educación básica. En regiones como Europa y Asia Central, las políticas pueden enfocarse en garantizar la calidad y la accesibilidad de los servicios.

Además, resulta necesario adoptar políticas integrales que consideren la interrelación entre diferentes dimensiones de la pobreza simultáneamente. Estas intervenciones suelen tener un impacto duradero en la reducción de la pobreza. Por ejemplo, programas integrados que combinen mejoras en salud, educación y condiciones de vida de manera integral.

Conclusiones

Tanto en economía como en filosofía política se dice que la mejor inversión de un país consiste en invertir en la educación, argumentando que si mejora la educación de los ciudadanos, estos mejorarán su calidad de vida, lo cual se traduce en que sus niveles de pobreza en general disminuyen. En este sentido, el análisis factorial permitió diferenciar entre un *factor educativo* y un *factor material* y, siguiendo el razonamiento anterior, con tan solo mejorar el factor educativo de la pobreza multidimensional, en el largo plazo se disminuiría cualquier otro tipo de pobreza como, por ejemplo, la pobreza material.

Al inicio del presente estudio se dijo que el objetivo del Banco Mundial y las demás instituciones que le colaboran es reducir la pobreza y generar prosperidad compartida en los países en desarrollo. En este sentido, los resultados de este estudio son útiles para proponer políticas que reduzcan la pobreza multidimensional. Por ejemplo, vimos que en una parte de la región del África Sub-Sahariana los niveles de pobreza son muy extremos en todos los sentidos, pero no lo son en otra parte de la misma región, donde aunque la pobreza material es alta, la pobreza educativa no lo es.

En este sentido, y siguiendo la teoría política y económica, se debería impulsar la educación en una parte de la región del África Sub-Sahariana; esto disminuiría en el corto plazo la pobreza educativa y, en el largo plazo, la pobreza material. Sin embargo, esta propuesta solo considera los resultados del estudio de la pobreza multidimensional; una propuesta más completa debería considerar aspectos como la disponibilidad de recursos, la intervención de países ajenos en su política y el número de habitantes, por ejemplo.

No es posible decir que todos los países incluidos en una región tienen un nivel de pobreza alto o bajo, sino que más bien suele haber diversidad. Sin embargo, sí hay tendencias en algunas regiones, como es el caso del África Sub-Sahariana, cuyos países se agruparon en los clusters con mayores medias en sus niveles de pobreza; o el caso de Europa y algunos países de Asia Central, cuyos países se agruparon en los clusters con menores medias en sus niveles de pobreza.

Se observó también que la brecha entre los países de América Latina con los de Europa no es tan grande como la brecha que hay entre los países de América Latina con los del África Sub-Sahariana, los niveles de pobreza en esta región resultaron ser extremos, como se pudo observar en el cluster 1 de la primera clusterización.

Finalmente, este estudio destaca la relevancia de la pobreza multidimensional como una herramienta clave para comprender y abordar las diversas formas en que la pobreza afecta a las personas y las comunidades. Al adoptar este enfoque, podemos avanzar hacia un futuro donde todas las personas tengan la oportunidad de vivir una vida digna y satisfactoria, libre de privaciones y con acceso a las oportunidades necesarias para su desarrollo personal y social.

Bibliografía

(2018). *Piecing together: The Poverty Puzzle*. World Bank Group.

(2010). *Pobreza, marginación y vulnerabilidad conforme a la Ley General de Desarrollo Social y su reglamento*. Instituto de Investigaciones Jurídicas de la Universidad Nacional Autónoma de México (1a. ed.). Suprema Corte de Justicia de la Nación.

Ravallion, M. (1992). *Poverty comparisons: A Guide to Concepts and Methods*. The World Bank.

Rubin, D. (1976). *Inference and Missing Data*. Oxford Journals, 63(3), 581-592.

Lerdo, M. (2014). *Estimación de datos faltantes con el Algoritmo EM*. Universidad Nacional Autónoma de México.

Box, G., Cox, D. (1964). *An Analysis of Transformations*. Journal of the Royal Statistical Society, 26(2), 2011-252.

Osborne, J. (2010). *Improving your data transformations: Applying the Box-Cox transformation*. Practical Assessment, Research, and Evaluation, 15(12).

Galarza, L. (2013). *Comparación mediante simulación de los métodos em e imputación múltiple para datos faltantes*. Universidad Nacional Mayor de San Marcos.

Buuren, S. (2007) *Multiple imputation of discrete and continuous data by fully conditional specification*. SAGE Publications, 16, 219-242.

Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). *mice: Multivariate Imputation by Chained Equations in R*. Journal of Statistical Software, 45(3), 1-67.

Breunig, M., Kriegel, H., Ng, R., & Sander, J. (2000). *LOF: Identifying Density-Based Local Outliers*. ACM SIGMOD Record, 29(2), 93-104.

De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. (2000). *The Mahalanobis distance*. Chemometrics and Intelligent Laboratory Systems, 50(1), 1-18.

Liu, F., Ting, K., & Zhou, Z. (2008). *Isolation Forest.* 2008 Eighth IEEE International Conference on Data Mining, 413-422.

Likas, A., Vlassis, N., & Verbeek, J. (2003). *The global k-means clustering algorithm.* Pattern Recognition, 36(2), 451-461.

Anexos

Anexo 1. Selección del periodo de tiempo para el análisis.

```
library(readxl)
pob_mult <- read_excel("C:\\Users\\orion\\Documents\\6. Semestre VI
\\Métodos multivariados II\\Proyecto\\MPM-Data-AM2023.xlsx")
nom <- c("Monetaria (%)", "Logro educativo (%)", "Matrícula educativa (%)",
"Electricidad (%)", "Saneamiento (%)", "Agua potable (%)")
report_year <- as.numeric(unlist(pob_mult[3:151,4]))
hist(report_year, breaks = seq(min(report_year) - 0.5, max(report_year) + 0.5, by = 1),
     main = 'Histograma del "Año de informe"',
     xlab = "Año",
     ylab = "Frecuencia",
     col = "gray")
mean(report_year)
median(report_year)
sd(report_year)
min(report_year)
max(report_year)
report_year_filt <- report_year[report_year >= 2015 & report_year <= 2021]
length(report_year_filt)/length(report_year)
length(report_year_filt)
```

Anexo 2. Análisis de datos perdidos.

```
pob_mult_fil <- pob_mult[pob_mult[, 4] >= 2015 & pob_mult[, 4] <= 2021, ]
datos <- pob_mult_fil[2:115,]
datos2 <- datos[,10:15]
# Total de datos.
6*114
# Total de datos ausentes.
sum(sapply(datos2, function(x) sum(x == "-")))
# Datos ausentes por columna.
apply(datos2, 2, function(x) sum(x == "-"))
# Datos ausentes por fila.
lost_row <- apply(datos2, 1, function(x) sum(x == "-"))
table(cut(lost_row, breaks = c(-Inf, 0, 1, 2, 3), labels = c("0", "1", "2", "3")))
```

Anexo 3. Países considerados en el análisis y número de países por región.

```
# Calcular el número de datos faltantes en las columnas 10 a 15 para cada fila.
lost_data <- apply(pob_mult_fil[, 10:15], 1, function(x) sum(x == "-"))
# Filtrar las filas con a lo mucho dos valores perdidos en las columnas 10 a 15.
cd <- pob_mult_fil[lost_data <= 2, ][2:113,]
```

```
# Países considerados en el análisis después del filtro.
as.vector(cd[,3])
pob_mult
# Número de países considerados en el conjunto de datos original y en el filtrado.
table(as.vector(unlist(pob_mult[3:151,1])))
table(as.vector(unlist(cd[,1])))
```

Anexo 4. Análisis de normalidad.

```
cdm <- as.matrix(cd[, 10:15])
cdm[cdm == "-"] <- NA
cdm <- apply(cdm, 2, function(x) as.numeric(x))
datos <- as.data.frame(cdm)

# Gráfico chi cuadrado.
datos2 <- as.matrix(na.omit(datos))
j <- c(1:length(as.matrix(datos2[,2])))
qc <- qchisq(((j-(1/2))/length(as.matrix(datos2[,2]))), 6)
xbar <- colMeans(datos2); S <- var(datos2); d = c()
for (i in 1:length(as.matrix(datos2[,2]))){
d <- append(d, t(datos2[i,]-xbar) %*% solve(S) %*% (datos2[i,]-xbar))}
do <- sort(d)
plot(qc, do, pch = 16, main="Chi Square Plot para normalidad multivariada")
abline(lm(do ~ qc))

# Prueba de Mardia.
library(MVN)
library(ggplot2)
res_mardia <- mvn(data = datos2, mvnTest = "mardia")
print(res_mardia$multivariateNormality)
```

Anexo 5. Maximización de la función de Box-Cox.

```
for (i in 1:6) {
  datos <- as.numeric(unlist(datos_comp[, i]))
  # Agregar corrección ya que hay valores que son ceros.
  datos <- datos + 0.0000001
  vl <- seq(-1, 2, by = 0.01)
  n <- length(datos)
  v <- c()

  for (k in 1:length(vl)) {
    lambda <- vl[k]
    vxt <- c()
    if (lambda != 0) {
      for (j in 1:n) {
        vxt <- append(vxt, ((((datos[j])^lambda) - 1) / lambda))
```

```
    }
  } else {
    for (j in 1:n) {
      vxt <- append(vxt, log(datos[j]))
    }
  }
  suma <- 0
  suma2 <- 0
  for (j in 1:n) {
    suma <- suma + ((vxt[j] - mean(vxt))^2)
  }
  for (j in 1:n) {
    suma2 <- suma2 + log(datos[j])
  }
  v <- append(v, (-(n / 2) * log(suma / n)) + ((lambda - 1) * suma2))
}
plot(vl, v, type = "l", xlab = expression(lambda), ylab = expression(paste("l(", lambda, ")")),
     main = bquote("Gráfica de " * lambda * " vs l(" * lambda * ") para datos de " * .(nom[i])))
lm <- which.max(v)
points(vl[lm], v[lm], pch = 16, col = "black", cex = 1.5)
c <- paste(vl[lm], "   ", round(v[lm], 4))
text(vl[lm], v[lm], c, pos = 1)
abline(v = vl[lm], h = v[lm])
}
```

Anexo 6. Transformación de los datos y análisis de normalidad.

```
datos <- as.data.frame(cdm)
library(MVN)
library(ggplot2)

# Transformación de lo datos.
lambda <- 0.1
datos2 <- (datos^lambda - 1) / lambda

# Gráfico chi cuadrado.
datos2 <- as.matrix(na.omit(datos2))
j <- c(1:length(as.matrix(datos2[,2])))
qc <- qchisq(((j-(1/2))/length(as.matrix(datos2[,2]))), 6)
xbar <- colMeans(datos2); S <- var(datos2); d = c()
for (i in 1:length(as.matrix(datos2[,2]))){
d <- append(d, t(datos2[i,]-xbar) %*% solve(S) %*% (datos2[i,]-xbar))}
do <- sort(d)
plot(qc, do, pch = 16, ylim=c(0,38), main="Chi Square Plot para normalidad multivariada")
abline(lm(do ~ qc))
text(qc, do, labels = j, pos = 3, cex = 0.7, col = "black")

# Prueba de Mardia.
res_mardia <- mvn(data = datos2, mvnTest = "mardia")
print(res_mardia$multivariateNormality)
```

Anexo 7. Transformación raíz cuadrada.

```
# Transformación de lo datos.
datos <- as.data.frame(cdm)
datos2 <- datos^(1/2)

# Gráfico chi cuadrado.
datos2 <- as.matrix(na.omit(datos2))
j <- c(1:length(as.matrix(datos2[,2])))
qc <- qchisq(((j-(1/2))/length(as.matrix(datos2[,2]))), 6)
xbar <- colMeans(datos2); S <- var(datos2); d = c()
for (i in 1:length(as.matrix(datos2[,2]))){
d <- append(d, t(datos2[i,]-xbar) %*% solve(S) %*% (datos2[i,]-xbar))}
do <- sort(d)
plot(qc, do, pch = 16, ylim=c(0,25), main="Chi Square Plot para normalidad multivariada")
abline(lm(do ~ qc))
text(qc, do, labels = j, pos = 3, cex = 0.7, col = "black")

# Prueba de Mardia.
library(MVN)
library(ggplot2)
res_mardia <- mvn(data = datos2, mvnTest = "mardia")
print(res_mardia$multivariateNormality)

shapiro.test(datos5)
```

Anexo 8. Eliminación de las observación 72 y nuevo análisis de normalidad.

```
# Eliminar la fila 72.
cdm <- cdm[-72, ]

# Transformación de los datos.
datos <- as.data.frame(cdm)
datos2 <- datos^(1/2)

# Gráfico chi cuadrado.
datos2 <- datos2[-72, ]
datos2 <- as.matrix(na.omit(datos2))
j <- c(1:length(as.matrix(datos2[,2])))
qc <- qchisq(((j-(1/2))/length(as.matrix(datos2[,2]))), df = ncol(datos2))
xbar <- colMeans(datos2)
S <- var(datos2)
d <- c()
for (i in 1:length(as.matrix(datos2[,2]))) {
  d <- append(d, t(datos2[i,]-xbar) %*% solve(S) %*% (datos2[i,]-xbar))}
do <- sort(d)
plot(qc, do, pch = 16, ylim=c(0,25), main="Chi Square Plot para normalidad multivariada")
abline(lm(do ~ qc))
text(qc, do, labels = j, pos = 3, cex = 0.7, col = "black")
```

Anexo 9. Imputación de datos usando el algoritmo EM.

```
datos1<- datos^(1/2)
library(Amelia)
# Realizar imputación.
di <- amelia(datos1, m = 10)
# Crear un marco de datos vacío para almacenar el conjunto de datos final.
dc1r <- datos1
# Iterar sobre cada columna para promediar los valores imputados.
for (col in names(dc1r)){dc1r[[col]] <- rowMeans(sapply(di$imputations, function(x) x[[col]]))}
dc1r
```

Anexo 10. Imputación múltiple de los datos usando FCS.

```
library(mice)
# Establecer el método Predictive Mean Matching (pmm) para todas las columnas completas.
meth <- make.method(datos)
meth[complete.cases(meth)] <- "pmm"
datos_imput <- mice(datos, method = meth, m = 10, maxit = 5, seed = 500)
dc2 <- complete(datos_imput)
round(dc2, 4)
```

Anexo 11. Imputación usando el método de predicción normal.

```
library(mice)
# m = 5 significa que generará 5 conjuntos de datos imputados diferentes
imputaciones <- mice(datos1, method = "norm.predict", m = 10)
# Resumen de las imputaciones
summary(imputaciones)
# Obtener el primer conjunto de datos imputado
dc3r <- complete(imputaciones, action = 1)
dc3 <- dc3r^2
round(dc3,4)
```

Anexo 12. Estadísticas de resumen de los tres conjuntos de datos completos.

```
round(colMeans(dc1),3)
round(colMeans(dc2),3)
round(colMeans(dc3),3)
round(var(dc1),3)
round(var(dc2),3)
round(var(dc3),3)
round(cor(dc1),3)
round(cor(dc2),3)
round(cor(dc3),3)
```

Anexo 13. Análisis de componentes principales para la identificación de outliers.

```
library(ggplot2)
library(factoextra)
library(ggrepel)
dc2s <- scale(dc2)
cp <- prcomp(dc2s, center = TRUE, scale. = TRUE)
summary(cp)
scores <- as.data.frame(cp$x)
scores$observation <- 1:nrow(scores)
ggplot(scores, aes(x = PC1, y = PC2)) +
  geom_point() +
  geom_text_repel(aes(label = observation), size = 3) +
  theme_minimal() +
  labs(title = "Gráfica de los primeros dos componentes principales",
       x = "Componente principal 1",
       y = "Componente principal 2")
```

Anexo 14. Cuatro métodos para la identificación de outliers.

```
# Método LOF.
library(dbscan)
md <- as.matrix(dc1)
lof_scores <- lof(md, k = 5)
outliers_LOF <- which(lof_scores > 2) # Usar un umbral comúnmente aceptado.
outliers_LOF

# Distancia de Mahalanobis.
library(mvoutlier)
md <- mahalanobis(dc2, colMeans(dc2), cov(dc2))
threshold <- qchisq(0.99, df = ncol(dc2)) # Nivel de confianza del 99%.
outliers_mal <- which(md > threshold)
outliers_mal

# Isolation Forest.
library(isotree)
iso_forest <- isolation.forest(dc2)
anomaly_scores <- predict(iso_forest, dc2)
threshold <- quantile(anomaly_scores, 0.95) # Umbral del 95%.
outliers_IF <- which(anomaly_scores > threshold)
outliers_IF

# K-medias.
library(stats)
set.seed(123)
kmed_res <- kmeans(dc2, centers = 3)
distances <- sqrt(rowSums((dc2 - kmed_res$centers[kmed_res$cluster, ])^2))
threshold <- quantile(distances, 0.95) # Umbral del 95%.
outliers_kmed <- which(distances > threshold)
```

```
outliers_kmed
```

Anexo 15. Eliminación de outliers.

```
v <- c(63, 68, 71, 84, 88, 95, 98)
cdf <- cd[-v, ]
as.vector(cdf[,3])
table(as.vector(unlist(cdf[,1])))
```

Anexo 16. Estadísticas descriptivas.

```
d <- dc2[-v, ]
round(colMeans(d),3)
round(var(d),3)
round(cor(d),3)
```

Anexo 17. Análisis de componentes principales para la determinación del número de factores.

```
library(factoextra)
ds <- scale(d)
cp <- prcomp(ds, center = TRUE, scale. = TRUE)
summary(cp)

fviz_eig(cp, addlabels = TRUE, ylim = c(0, 100), ylab = "Varianza explicada", xlab = "Eigenvalor",
barcolor = "grey", barfill = "gray", ggtheme = theme_minimal()) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  theme_minimal() +
  geom_bar(stat = "identity", width = 0.5)
```

Anexo 17. Análisis factorial para modelos con $m = 2$ y $m = 3$ factores.

```
# Para un modelo con m = 2 factores.
library(psych)
fa(R, nfactors=2, rotate="varimax", fm="ml", covar=FALSE)$loadings
ev <- -eigen(R)$vectors[,1:2]
Lcp <- ev %*% diag(sqrt(eigen(R)$values[1:2]))
round(varimax(Lcp)$loadings, 3)

# Para un modelo con m = 3 factores.
library(psych)
fa(R, nfactors=3, rotate="varimax", fm="ml", covar=FALSE)$loadings

ev <- -eigen(R)$vectors[,1:3]
Lcp <- ev %*% diag(sqrt(eigen(R)$values[1:3]))
round(varimax(Lcp)$loadings, 3)
```

Anexo 18. Clusterización.

```
library(factoextra)
library(ggplot2)

# Estandarizar los datos.
de <- scale(d)
# Determinación del número óptimo de clusters.
fviz_nbclust(de, kmeans, method = "wss") +
  labs(title = "Número óptimo de clusters",
       x = "Número de clusters (k)",
       y = "Suma de cuadrados intra-cluster (WSS)") +
  theme_minimal() +
  geom_line(color = "black") +
  geom_point(color = "black")
# Método de las k-medias.
set.seed(123)
clust_res <- kmeans(de, centers = 3, nstart = 25)
print(clust_res)
fviz_cluster(clust_res, data = de, main="Gráfico de clusters")

# Obtener información detallada sobre los clusters usando datos originales.
d$cluster <-  clust_res$cluster
# Medias de los clusters.
clust_med <- aggregate(d[, -ncol(d)], by=list(cluster=d$cluster), FUN=mean)
# Varianzas de los clusters.
clust_var <- aggregate(d[, -ncol(d)], by=list(cluster=d$cluster), FUN=var)
# Número de observaciones en cada cluster.
clust_tam <- table(d$cluster)
# Resultados.
print(clust_med)
print(clust_var)
print(clust_tam)

nom <- as.vector(unlist(cdf[,3]))
reg <- as.vector(unlist(cdf[,1]))
d$country <- nom
d$region <- reg

# Filtrar los países por cada cluster.
clust1 <- d[d$cluster == 1,]
clust2 <- d[d$cluster == 2,]
clust3<- d[d$cluster == 3,]

# Países correspondientes a cada cluster.
cat("Países en el Cluster 1:\n")
print(clust1$country)
cat("\nPaíses en el Cluster 2:\n")
print(clust2$country)
cat("\nPaíses en el Cluster 3:\n")
print(clust3$country)
```

```
df_clust <- rbind(countries_cluster_1, countries_cluster_2, countries_cluster_3)
# Tabla de número de países por región por cluster.
tab <- table(df_clust$cluster, df_clust$region)
print(tab)
```

Anexo 19. Segunda clusterización.

El código es similar a la primera clusterización; basta con reemplazar el conjunto de datos.