Trabajo Final

Análisis Multivariante y Cluster de la Pobreza y la Desigualdad a Nivel Mundial

AUTORES:

MANUEL CALLEJO GARCÍA

CARLOS DE MANUEL VICENTE

HUGO MUGÜERZA MENDOZA

30 de ENERO de 2025

ÍNDICE

1. INTRODUCCIÓN

El análisis de la pobreza y la desigualdad es un desafío crucial en la investigación económica y social, con implicaciones directas para el cumplimiento de los Objetivos de Desarrollo Sostenible (ODS) de las Naciones Unidas, como el ODS 1 (Fin de la pobreza) y el ODS 10 (Reducción de las desigualdades) [1]. Estos conceptos son multidimensionales, ya que abarcan aspectos como el ingreso, el acceso a servicios básicos, las oportunidades económicas y la distribución equitativa de recursos. Sin embargo, la falta de una definición universalmente aceptada y la complejidad inherente de estos fenómenos dificultan su medición y análisis [2].

El gran volumen y diversidad de variables necesarias para abordar estos temas genera desafíos analíticos significativos. Por un lado, incrementa la complejidad computacional y, por otro, dificulta la interpretación de los resultados para la formulación de políticas públicas [3]. Ante esta problemática, se hace indispensable la aplicación de métodos que permitan sintetizar los datos en un conjunto reducido de dimensiones explicativas sin perder la esencia de la información.

Este trabajo tiene como objetivo principal realizar un análisis multivariante de un conjunto de datos global sobre pobreza y desigualdad, abarcando diferentes países y periodos temporales. Para ello, se implementarán técnicas de reducción de dimensionalidad y clustering, integrando información geoespacial para identificar agrupaciones coherentes. Este enfoque busca no solo comprender patrones subyacentes en los datos, sino también proporcionar información práctica y accionable para la toma de decisiones en el ámbito socioeconómico.

2. OBJETIVOS DEL TRABAJO

El presente trabajo tiene por tanto el siguiente objetivo general: *"Realizar un análisis multivariante del conjunto de datos global sobre pobreza y desigualdad para reducir su dimensionalidad e identificar agrupaciones significativas de países desde una perspectiva socioeconómica y geoespacial".*

Para ello, contamos con los siguientes objetivos específicos que guiarán la ruta de trabajo:

- **O1:** Reducir la dimensionalidad del conjunto de datos mediante técnicas estadísticas adecuadas para identificar un subconjunto reducido de variables con alta capacidad explicativa.
- **O2:** Implementar y evaluar métodos de agrupación que permitan identificar clústeres significativos de países, optimizando el número de grupos e interpretando sus características.
- **O3:** Incorporar información espacial y restricciones geográficas en el análisis clúster para mejorar la coherencia de los resultados y proporcionar una visión integral tanto socioeconómica como geográfica.

3. ANTECEDENTES

El estudio de la pobreza y la desigualdad ha ocupado un lugar central en la agenda de investigación económica y social a lo largo de las últimas décadas, siendo temas clave para organismos internacionales como el Banco Mundial y Naciones Unidas. Estas problemáticas han sido abordadas desde diversas perspectivas, utilizando métricas ampliamente aceptadas como el "poverty headcount", el coeficiente de Gini y el Palma ratio, entre otras [4]. Dichas métricas permiten medir tanto la magnitud de la pobreza como las disparidades en la distribución del ingreso o la riqueza.

A pesar de su utilidad, estas medidas presentan limitaciones importantes, ya que su sensibilidad al contexto regional y temporal puede dificultar comparaciones directas entre países. Además, los enfoques metodológicos varían considerablemente según las fuentes de datos, por ejemplo, en el uso de líneas de pobreza basadas en ingreso o consumo, o en las metodologías de recolección de encuestas. Esto genera un entorno de análisis caracterizado por la heterogeneidad y la multidimensionalidad, lo que plantea retos significativos para la interpretación y generalización de los resultados.

Los análisis multivariantes se han convertido en herramientas fundamentales para superar estas limitaciones, permitiendo sintetizar la información y descubrir patrones subyacentes [5]. De manera similar, los métodos de clustering se utilizan para identificar agrupaciones de países o regiones con características comunes. Sin embargo, los estudios que integran estas técnicas con información geoespacial siguen siendo escasos, lo que subraya la importancia de explorar enfoques innovadores que combinen dimensiones socioeconómicas y geográficas.

Este trabajo busca aportar a este campo mediante la aplicación de técnicas multivariantes y de agrupamiento sobre un conjunto de datos global, ofreciendo una perspectiva integral para el análisis de la pobreza y la desigualdad. Este enfoque no solo permite capturar las características socioeconómicas de los países, sino también explorar cómo los factores espaciales pueden influir en los patrones de pobreza y desigualdad.

4. DATOS

El conjunto de datos empleado en este estudio proviene de la Plataforma de Pobreza y Desigualdad (PIP) del Banco Mundial, una herramienta que proporciona estimaciones globales y desagregadas sobre pobreza y desigualdad. Esta base de datos ha sido recopilada y puesta a disposición en Kaggle bajo el nombre *Global Poverty and Inequality Dataset* [6].

Este conjunto de datos se compone de 4877 observaciones donde cada instancia corresponde a una encuesta en un país para un año determinado. Además, disponemos de 108 columnas con información para determinar tanto la información temporal y regional como la socioeconómica, permitiendo un análisis tanto regional como temporal de las condiciones de pobreza y desigualdad en el mundo

Con respecto a las variables categóricas, encontramos las siguientes:

- **País ('country')** y **continente ('continent')**: identifican al país y el continente en el que se ubica.
- **Nivel de reporte ('reporting_level'):** indica si la encuesta se ha hecho a nivel nacional, urbano o rural.
- **Tipo de bienestar ('welfare_type'):** indica si los datos de la encuesta se refieren a los ingresos del hogar o a los gastos de consumo
- **Comparabilidad de las encuestas ('survey_comparability')**: refleja la calidad y consistencia de las mediciones entre países.

Del mismo modo, entre las variables numéricas encontramos las siguientes:

- **Año de encuesta ('year'):** año para el cual se ha realizado la encuesta.
- **Poverty headcount (indicador de pobreza)**: porcentaje de la población que vive por debajo de una determinada línea de pobreza. Se reporta bajo distintas líneas de pobreza: internacional (umbrales fijados por el Banco Mundial, como 1.90 USD/día o 3.20 USD/día en términos de Paridad de Poder Adquisitivo (PPA) y mediana baja y alta (umbrales adaptados a contextos nacionales específicos)
- **Income poverty gap (indicador de pobreza)**: brecha promedio entre los ingresos de la población pobre y la línea de pobreza, expresada como porcentaje de esta.
- **Total shortfall (brecha de ingreso)**: cantidad total de ingresos que le falta a la población pobre para alcanzar la línea de pobreza. Este indicador se expresa en términos absolutos y ayuda a dimensionar la magnitud del problema.
- **Average shortfall (brecha de ingreso)**: el *total shortfall* dividido entre el número de personas pobres, proporcionando una medida del déficit de ingresos promedio entre quienes viven en pobreza.
- **Deciles (deciles de ingreso):** medidas de distribución del ingreso entre distintos segmentos poblacionales para analizar si el ingreso está equitativamente distribuido o si está concentrado en ciertos grupos.
- **Coeficiente de Gini (indicador de desigualdad)**: mide la desigualdad en la distribución del ingreso. Su valor varía entre 0 y 1, donde 0 indica igualdad perfecta (todos tienen el mismo ingreso) y 1 indica desigualdad absoluta (una sola persona concentra todo el ingreso). Valores más altos reflejan sociedades más desiguales.
- **Palma ratio(indicador de desigualdad)**: relación entre el ingreso del 10% más rico de la población y el 40% más pobre. Un valor alto sugiere una concentración desproporcionada del ingreso en la parte superior de la distribución.
- **Mean Log Deviation o MLD (indicador de desigualdad)**:: varía entre 0 (igualdad perfecta) e infinito, dando más peso a las diferencias en los ingresos bajos.
- **Polarization (indicador de desigualdad)**: mide la concentración de ingresos en los extremos. Un valor bajo indica una distribución homogénea, mientras que valores altos reflejan una disminución de la clase media y una mayor separación entre ricos y pobres.

- **p90/p10 ratio (indicador de desigualdad)**: cociente entre el percentil 90 (ingreso de los más ricos) y el percentil 10 (ingreso de los más pobres). Este indicador ayuda a capturar diferencias extremas en la distribución del ingreso.
- **p50/p10 ratio (indicador de desigualdad)**: relación entre la mediana del ingreso (p50) y el percentil 10. Es útil para evaluar la desigualdad en la parte baja de la distribución del ingreso.

La base de datos integra información de diferentes versiones (2011 y 2017). La diferencia entre estas 2 versiones reside en la actualización de valores y porcentajes con respecto a la inflación, el límite de pobreza actual, etcétera. Así pues, en el estudio se priorizará la versión más reciente para garantizar la actualidad y consistencia de los registros.

Es importante señalar que, aunque este *dataset* es una fuente rica y detallada, existen ciertas limitaciones inherentes, como la posible falta de homogeneidad en las metodologías de recolección entre países y años, o la ausencia de datos completos para algunas regiones. Estas características serán consideradas a la hora de realizar el estudio.

5. METODOLOGÍA ESTADÍSTICA

El enfoque metodológico utilizado en este trabajo combina varias técnicas estadísticas para abordar los objetivos planteados. A continuación, se describen los pasos principales seguidos:

1. Preparación de los datos

Normalización Z-score: para garantizar que todas las variables tengan una escala comparable. La normalización estandariza los datos restando la media y dividiendo por la desviación estándar, de modo que las variables tengan una media de 0 y una desviación estándar de 1.

Transformación Box-Cox: en variables con fuertes asimetrías o sesgos para mejorar su ajuste a una distribución normal. La normalidad es un supuesto clave en muchas técnicas estadísticas, como la reducción de dimensionalidad y el clustering.

MICE (Multiple Imputation by Chained Equations): método de imputación múltiple que permite estimar valores perdidos a partir de relaciones entre las variables existentes, aplicando predicciones iterativas, en nuestro caso de un modelo de regresión lineal. MICE mantiene la variabilidad inherente a los datos, mejorando la robustez del análisis posterior.

2. Reducción de la dimensionalidad

Análisis de Componentes Principales (PCA): para reducir el número de variables originales a un conjunto más pequeño de componentes principales. Estas componentes explican la mayor parte de la varianza presente en los datos originales lo que facilita tanto el análisis como interpretación de los datos.

3. Análisis espacial

Autocorrelación global (Moran's I): mide el grado de dependencia espacial de las observaciones en todo el conjunto de datos. Un valor positivo y significativo indica una fuerte agrupación espacial (patrones homogéneos), mientras que un valor negativo sugiere una dispersión espacial. Se utiliza para evaluar si los países con características similares en pobreza y desigualdad tienden a agruparse geográficamente, proporcionando información clave sobre la estructura espacial del fenómeno analizado.

Autocorrelación local (Local Moran's I – LISA): permite detectar clusters espaciales y puntos atípicos dentro de los datos. Identifica regiones con alta o baja pobreza rodeadas por áreas similares (H-H o L-L clusters) o regiones atípicas con valores extremos diferentes a sus vecinos (H-L o L-H outliers). Se representa visualmente mediante un Moran plot.

4. Agrupación o *clustering*

K-means: método de clustering no jerárquico que asigna cada observación a uno de los K grupos predefinidos, minimizando la variabilidad dentro de cada cluster. La selección del k-óptimo se realiza con métricas complementarias de validación.

Método de Ward: método de clustering jerárquico que minimiza la varianza intragrupo en cada paso de la agrupación. Su principal ventaja es que genera clusters más homogéneos y permite visualizar las relaciones jerárquicas entre países mediante un dendrograma.

Skater: método de regionalización (clustering con restricciones espaciales) que garantiza que los países dentro de un mismo cluster sean geográficamente contiguos. A diferencia de K-means y Ward, que pueden formar clusters dispersos, Skater optimiza la agrupación considerando tanto la similitud socioeconómica como la proximidad geográfica, proporcionando una segmentación más realista.

5. Validación del clustering

Silhouette Score: mide la coherencia interna de los clusters evaluando la distancia media entre observaciones dentro de un mismo cluster y comparándola con la distancia a otros clusters. Un valor cercano a 1 indica clusters bien definidos, mientras que valores cercanos a 0 sugieren una asignación poco clara.

Calinski-Harabasz Index: evalúa la calidad del clustering mediante la relación entre la dispersión intergrupo e intragrupo. Valores más altos indican una mejor separación entre clusters y una estructura más definida, proporcionando un criterio adicional para la validación del número óptimo de grupos.

6. RESULTADOS Y DISCUSIÓN

Durante la preparación del entorno de trabajo, seleccionamos todas aquellas instancias pertenecientes a la versión PIP de 2017, ya que la de 2011 cuenta con las mismas instancias pero desactualizadas y con un nivel de análisis nacional, ya que los niveles urbano y rural son muy poco frecuentes y no presentes en todos los continentes. Además, realizamos una agrupación de variables en 6 subgrupos para simplificar los posteriores análisis: 'Headcount', 'Average_Shortfall', 'Income_Poverty_Gap', 'Deciles' y 'Inequality_Metrics', estos grupos se realizan en base a los nombres y características de las variables.

6.1. Preparación de los datos

Realizamos un análisis descriptivo para entender la distribución de nuestras variables. Por subgrupos, estimamos con un *Kernel Density Estimation (KDE)* Gaussiano la densidad de probabilidad de las variables, normalizadas mediante una z-score para que sean comparables. En la *Figura 1* vemos como la distribución de las variables del subgrupo 'Total_Shortfall', las cuales muestran en su totalidad una asimetría positiva.

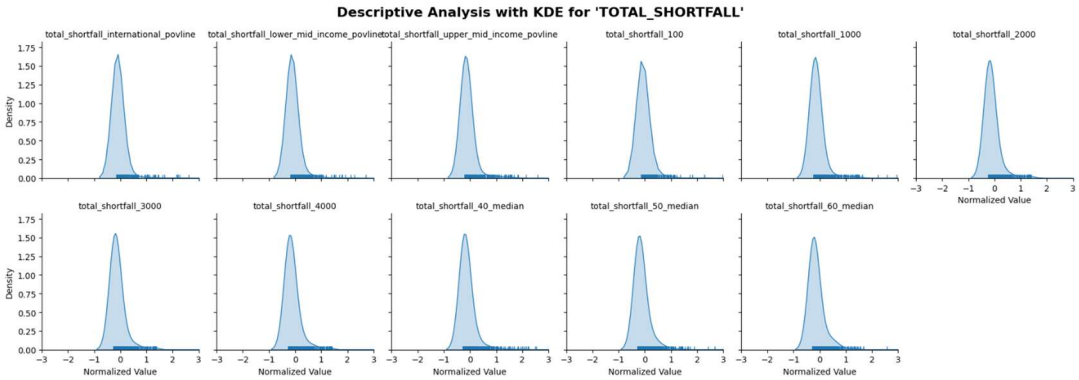

Figura 1. *KDE normalizado de las variables del subgrupo 'Total Shortfall'*

Obtenemos en la *Figura 2* una muestra similar pero para las métricas de desigualdad.

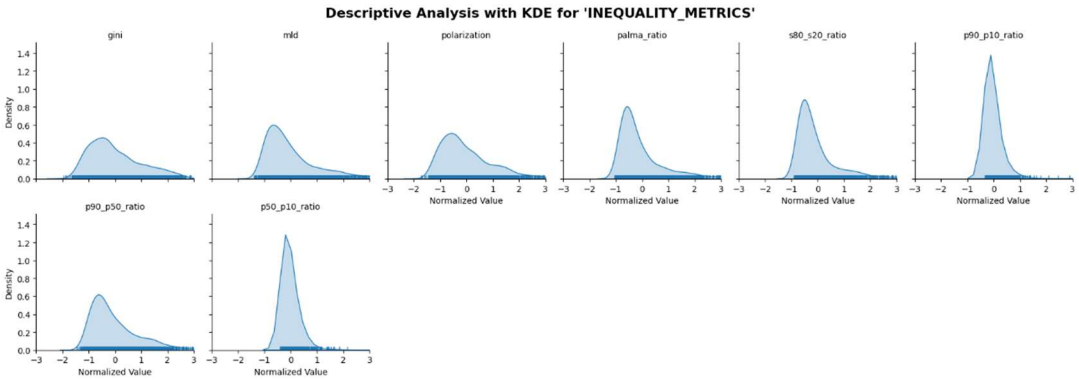

Figura 2. *KDE normalizado de las variables del subgrupo 'Inequality Metrics'*

Las asimetrías existentes en estos grupos de variables son generalizadas, por lo que hacemos, manualmente seleccionamos aquellas variables que indudablemente son asimétricas y les aplicamos una transformación Box-Cox para que cumplan normalidad. En la *Figura 3* vemos el resultado para las variables del subgrupo brecha de ingreso total "(total_shortfall"):

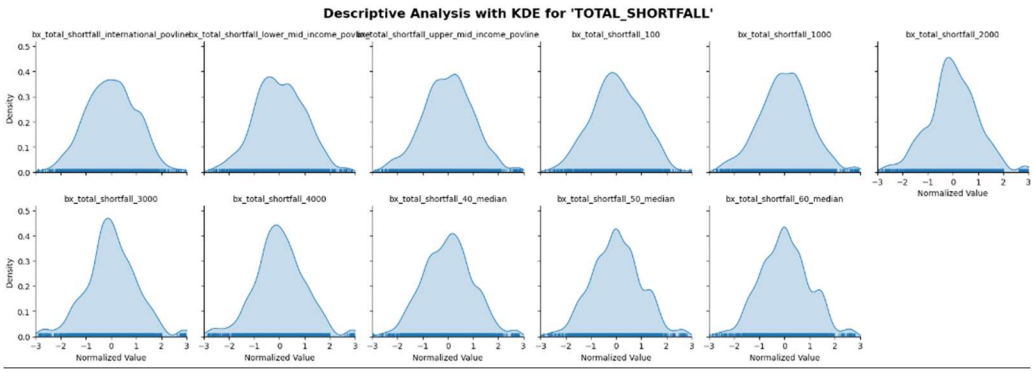

Figura 3. *KDE normalizado de las variables del subgrupo 'Total Shortfall' tras una transformación Box-Cox.*

Con respecto a la completitud de los datos, en la *Figura 4* apreciamos como hay variables con una cantidad muy grande de datos perdidos:

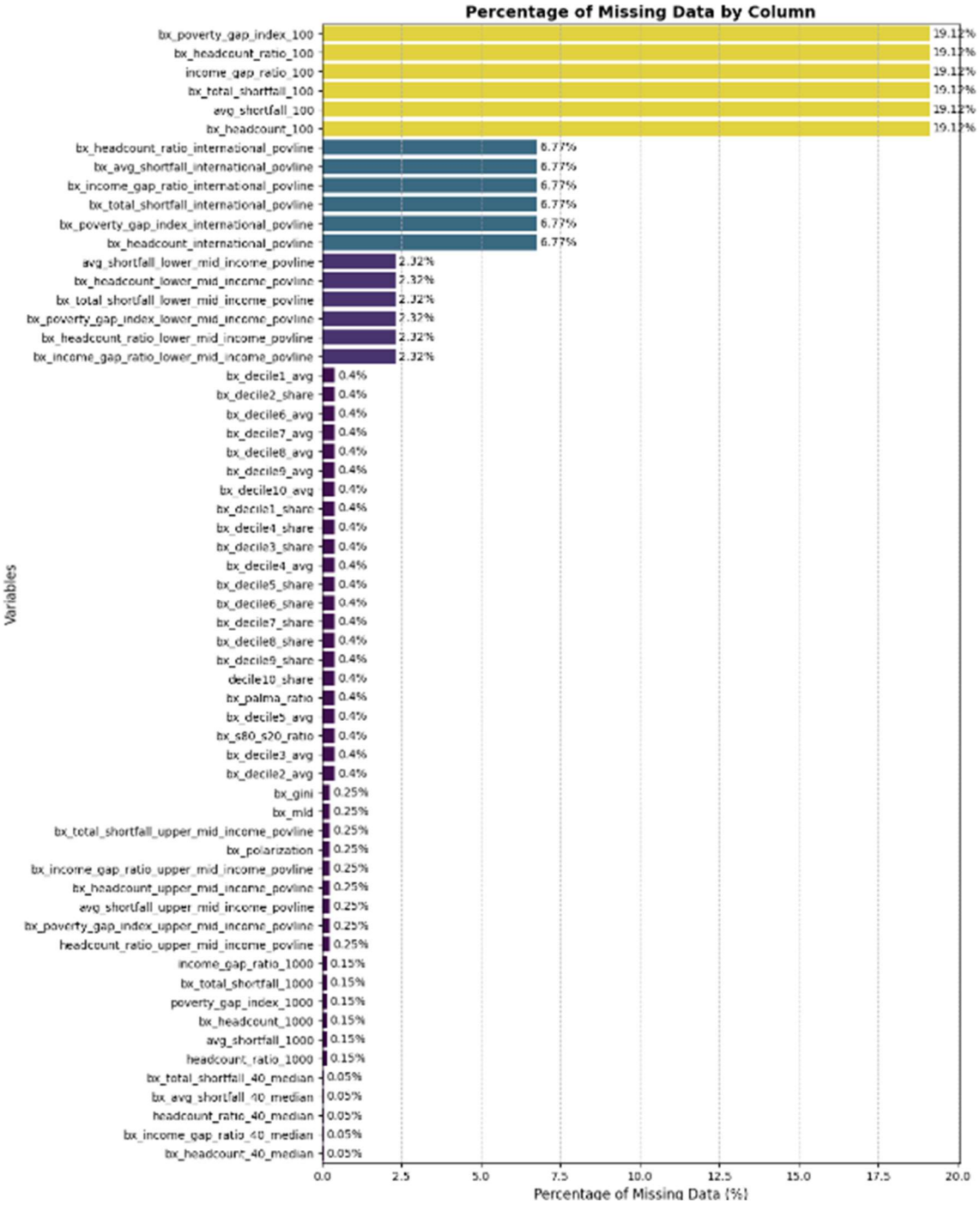

Figura 4. *Porcentaje de datos perdidos por variable*

Puesto que los porcentajes son muy similares, intuimos que debe de haber instancias defectuosas. Así pues, eliminamos todas aquellas instancias correspondientes a países donde en la encuesta no se obtuvieron suficientes datos y, con los pocos datos restantes utilizamos un método MICE con un modelo de regresión lineal con 10 iteraciones, de modo que consigamos la completitud de todo el conjunto de datos.

En este caso debido a las fuertes asimetrías y con el objetivo de no eliminar información valiosa de las colas de las distribuciones, no abordamos la existencia de datos anómalos univariantes ni multivariantes.

6.2. Análisis de correlaciones y reducción de la dimensionalidad

El objetivo de esta etapa es encontrar un subconjunto óptimo para realizar el clustering. Así pues, para cada subgrupo de variables socioeconómicas calculamos la matriz de correlaciones y la representamos su valor absoluto en un mapa de calor para identificar colinealidades entre variables visualmente. Además, realizamos un PCA donde escogemos todas aquellas componentes principales del subgrupo con lambda > 1, de modo que se excluirán aquellas sin suficiente poder explicativo. Adicionalmente, un gráfico de cargas permitirá interpreta los resultados del PCA.

El subgrupo de 'Headcount' es el más numeroso, aunque en el mapa de calor de la *Figura 5* apreciamos como muchas variables están correlacionadas. Así pues, se seleccionan 4 componentes principales (*Figura 6*). En la *Figura 7* apreciamos como la primera componente está, de forma muy similar para todas, inversamente relacionada con todas las variables del subgrupo, como si fuera un promedio general de las tasas de pobreza.

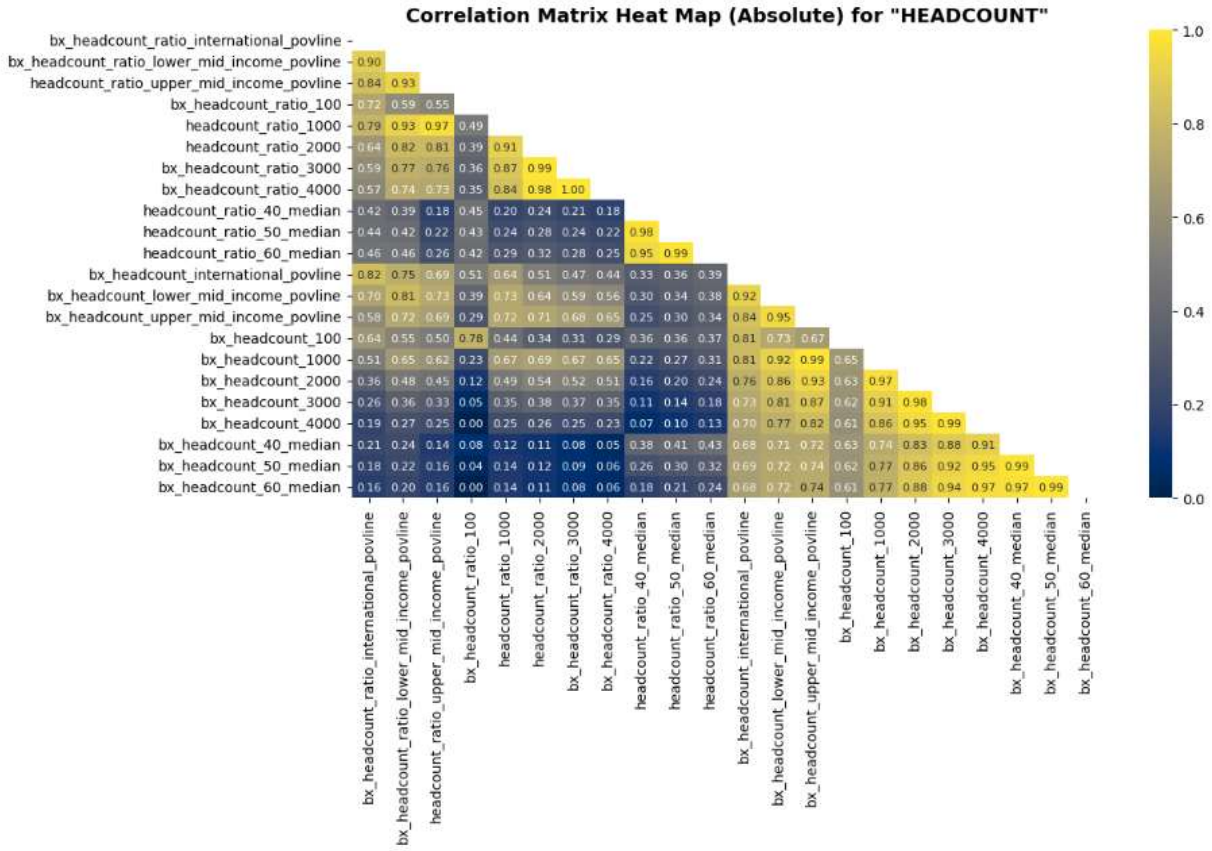

Figura 5. *Mapa de calor de la matriz de correlaciones en valor absoluto para el subgrupo 'Headcount'.*

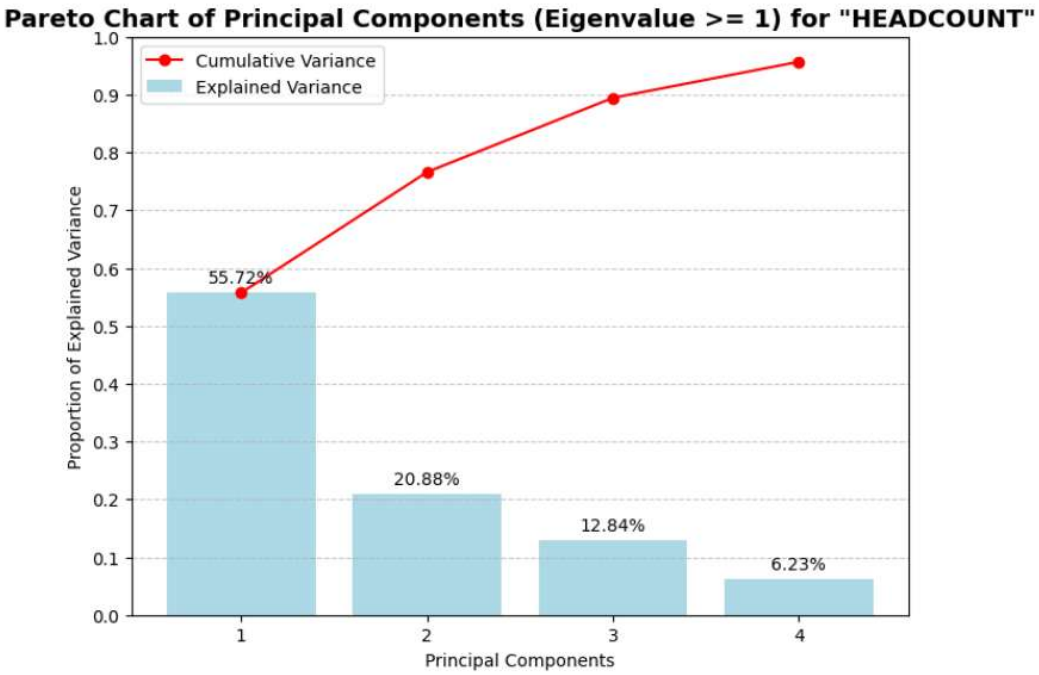

Figura 6. *Gráfico de Pareto de las componentes principales con λ>1 para el subgrupo 'Headcount'.*

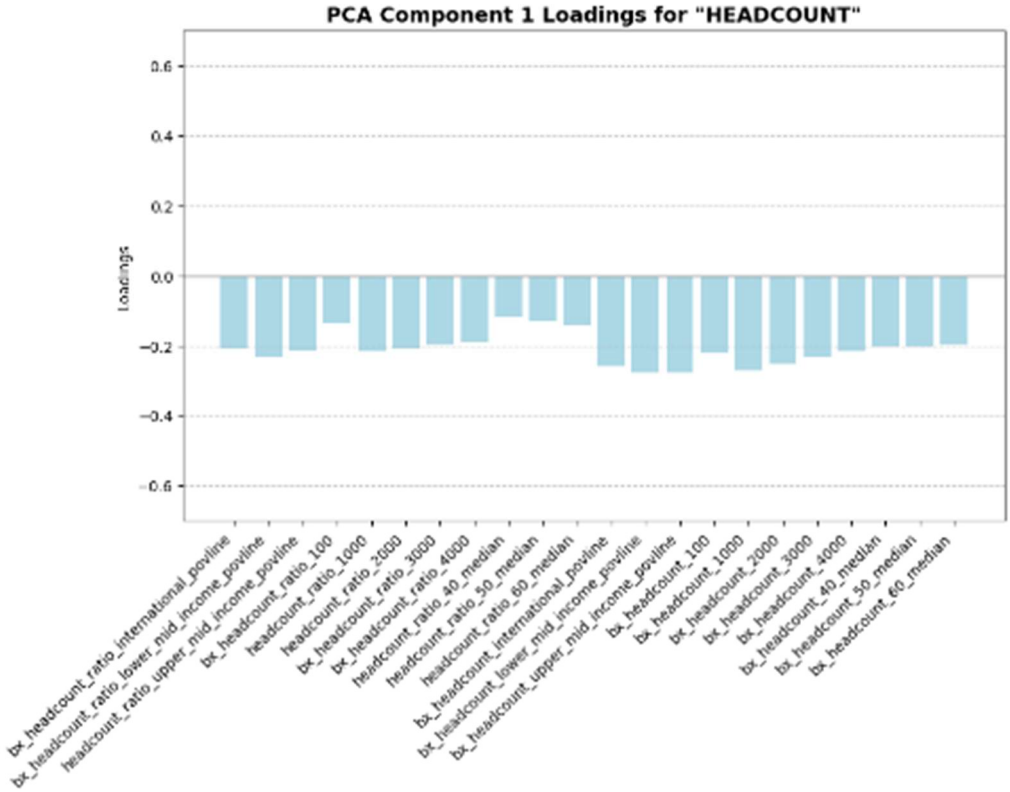

Figura 7. *Gráfico de cargas de la primera componente principal del subgrupo 'Headcount'.*

De un modo similar, graficamos e interpretamos el resto de los subgrupos. Destacamos que para el subgrupo de deciles, las correlaciones son extremadamente altas (*Figura 8*), lo cual ocasiona que una única componente principal sea capaz de agrupar el 74.1% de la varianza del subgrupo.

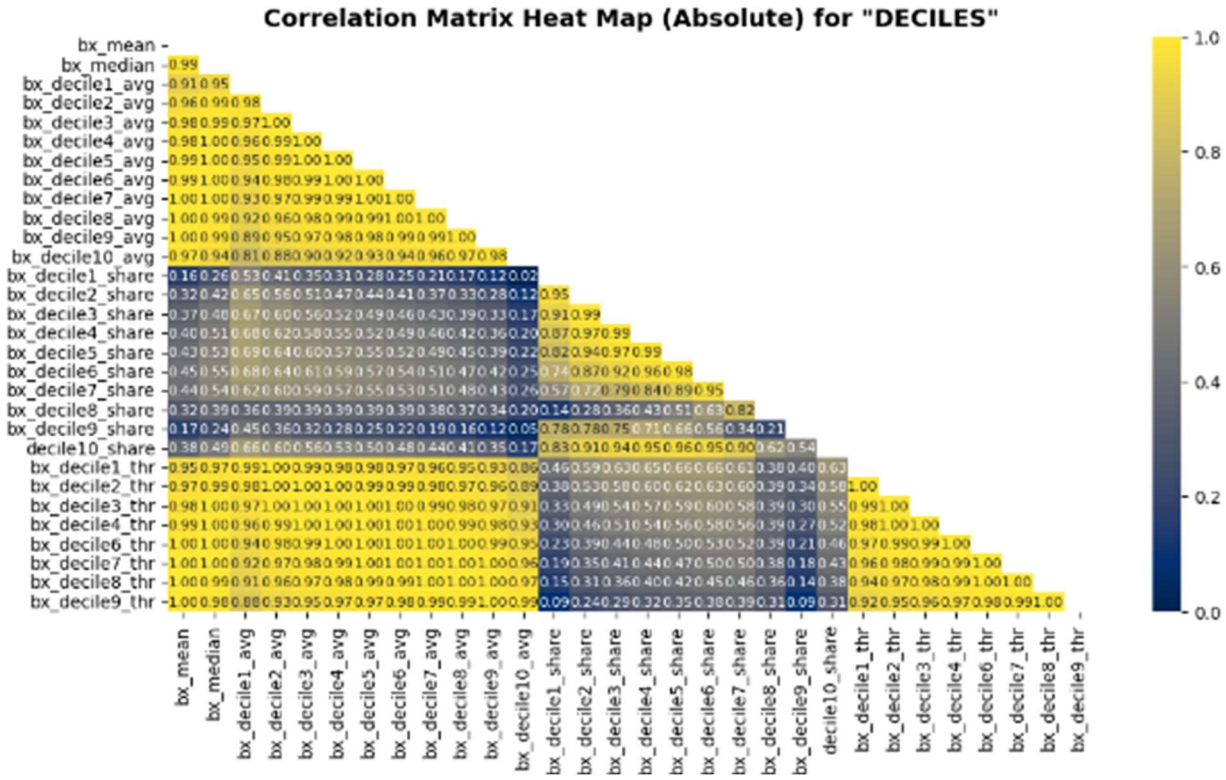

Figura 8. *Mapa de calor de la matriz de correlaciones en valor absoluto para el subgrupo 'Deciles'.*

Así pues, tras la reducción de la dimensionalidad, de las 102 variables numéricas nos quedamos con 14 componentes principales agrupadas de la siguiente forma y con la siguiente interpretación:

- **Headcount – 3 Componentes**: la primera componente se identifica como un promedio general de las tasas de pobreza, representando una visión agregada de las diferentes líneas de pobreza. La segunda componente destaca diferencias específicas entre líneas de pobreza más altas, como las correspondientes a umbrales de 4000. Finalmente, las componentes tres y cuatro capturan características más sutiles y residuales, enfocadas en particularidades de las líneas de pobreza menores, complementando el análisis global.

- **Average Shortfall – 2 Componentes**: la primera componente reflejaba principalmente la proporción promedio del déficit en todas las líneas de pobreza, funcionando de nuevo como una medida general del nivel de pobreza. Por otro lado, la segunda componente destaca las diferencias entre las líneas de pobreza más bajas y las más altas, proporcionando una visión más específica de las disparidades entre distintos umbrales de pobreza.

- **Total Shortfall – 2 Componentes**: de forma muy similar al subgrupo de 'Average Shortfall', la primera componente se centró en el déficit total promedio y la segunda en resaltar las diferencias entre las líneas de pobreza más bajas y altas.

- **Income Poverty Gap – 3 Componentes**: la primera componente principal captura un patrón global de pobreza y desigualdad de ingresos. La segunda y tercera componente diferencian de manera más específica entre las líneas intermedias y las más altas.

- **Deciles – 3 Componentes**: la primera componente capturaba de manera extremadamente clara las medias generales de los ingresos por decil, proporcionando una visión global de la distribución. La segunda componente resalta las diferencias entre los deciles más bajos y más altos, subrayando las desigualdades extremas. Por último, la tercera atiende a los patrones residuales vinculados a los ingresos de los deciles intermedios.
- **Inequality metrics – 1 Componentes**: el gráfico de cargas mostró que todas las métricas contribuyen de manera similar a una única componente principal (explica el 86'78% de la variabilidad), reflejando la coherencia en cómo estas capturan la desigualdad.

Finalmente, la matriz de correlaciones de todas las componentes principales quedaría de la siguiente forma (*Figura 9*):

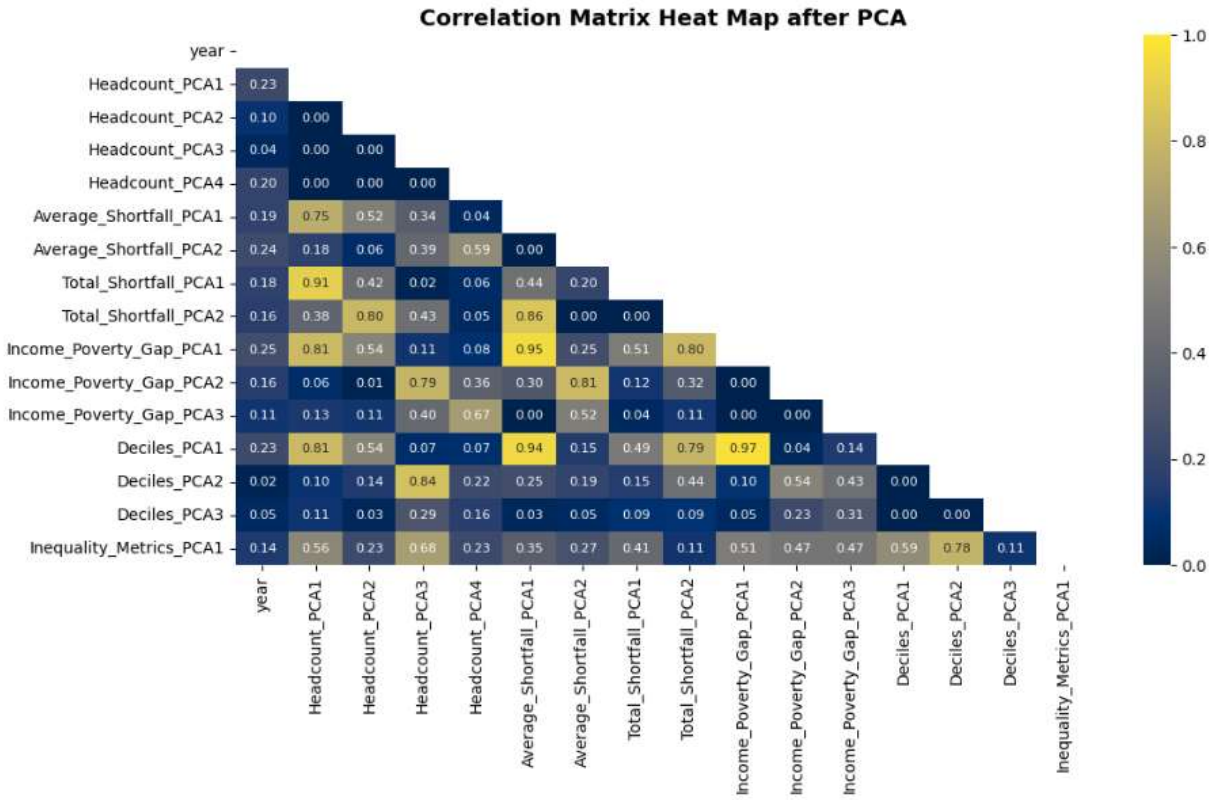

Figura 9. *Mapa de calor de la matriz de correlaciones de todas las componentes principales de los subgrupos.*

Apreciamos como todas las componentes pertenecientes al mismo subgrupo son ortogonales y por tanto con correlación 0. Sin embargo, apreciamos información redundante ya que hay altas correlaciones entre 'Average Shortfall', 'Deciles' e 'Income_Poverty_Gap'. Pese a existir correlaciones, continuamos el estudio con el conjunto completo para intentar encontrar algún patrón relevante dentro de estas variables.

6.3. Análisis espacial

Puesto que un análisis espacial otorga mayor comprensión, cargamos los datos espaciales de cada territorio para complementar el estudio. En la *Figura 10* apreciamos la cantidad de observaciones por año. Para poder homogeneizar las características de cada territorio y asegurar la mayor cantidad de presencia, tomamos los años 2004-2018, ya que tienen la mayor cantidad de observaciones, y para simplificar la variabilidad temporal, agrupamos cada territorio con el promedio de sus variables durante ese periodo:

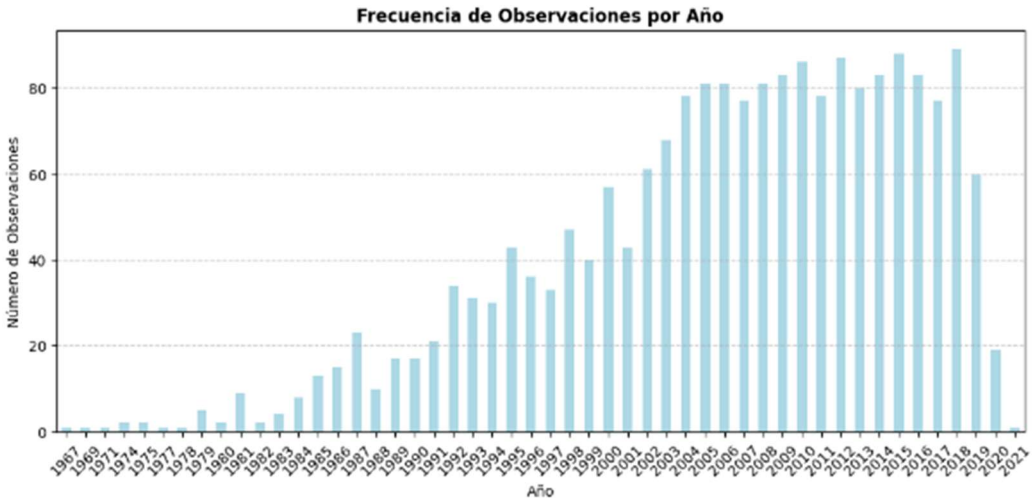

Figura 10. *Frecuencia de instancias según el año de encuesta.*

Con este conjunto de datos analizamos las correlaciones espaciales de los países para cada componente principal. Para todas ellas, el estadístico Moran's I otorga un p-valor ≈ 0, con lo que existe evidencia estadística suficiente para afirmar con un 95% que las características de los países están correlacionadas. Si graficamos el Moran Plot (*Figura 11*), apreciamos como hay una gran concentración de puntos y a lejados del origen (0, 0) para la sección superior derecha (valores altos o HH) y inferior izquierda (valores bajos o LL).

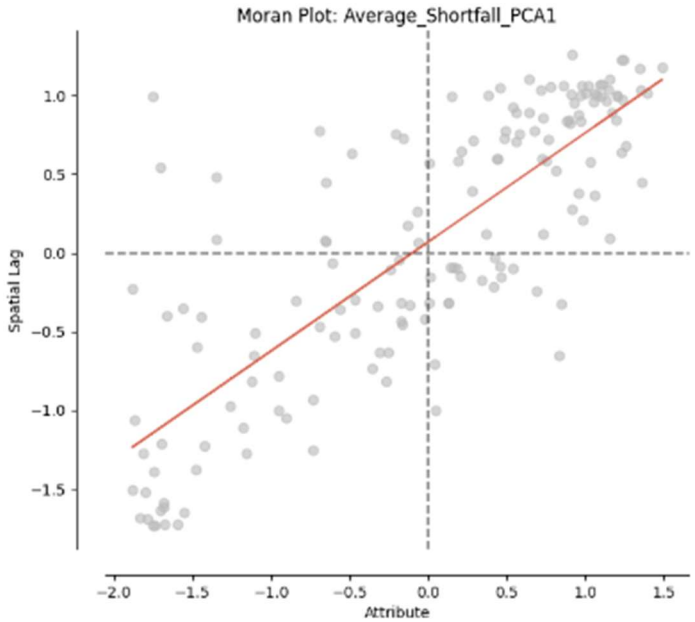

Figura 11. *Moran Plot de los datos 2004-2018 para la primera componente principal del subgrupo Average Shortfall*

Así pues, y graficando los países con un estadístico LISA significativo, observamos como los países de Europa, al contrario que los de África o alguna región de Asia, están rodeados de zonas con bajos valores de Average_Shortfall_PCA1, con lo que Europa parece tener un bajo nivel de pobreza en comparación con África o China y la India que están especialmente agrupadas con altos niveles de pobreza (*Figura 12*).

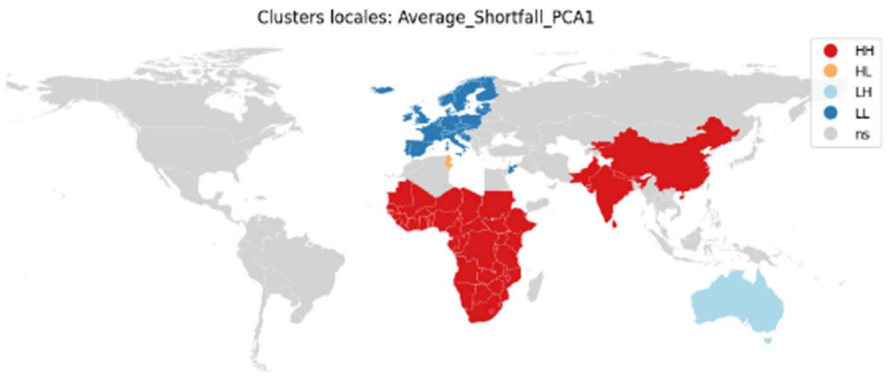

Figura 12. *Mapa de correlaciones locales significativas para la primera componente principal del subgrupo Average Shortfall*

Para las métricas de desigualdad tendríamos un análisis similar, donde, atendiendo directamente al gráfico de autocorrelaciones locales significativas de la *Figura 13*, vemos como Europa sigue manteniendo bajos valores, aunque en este caso de desigualdad, mientras que otras regiones como el sur de África o gran parte de América son más desiguales.

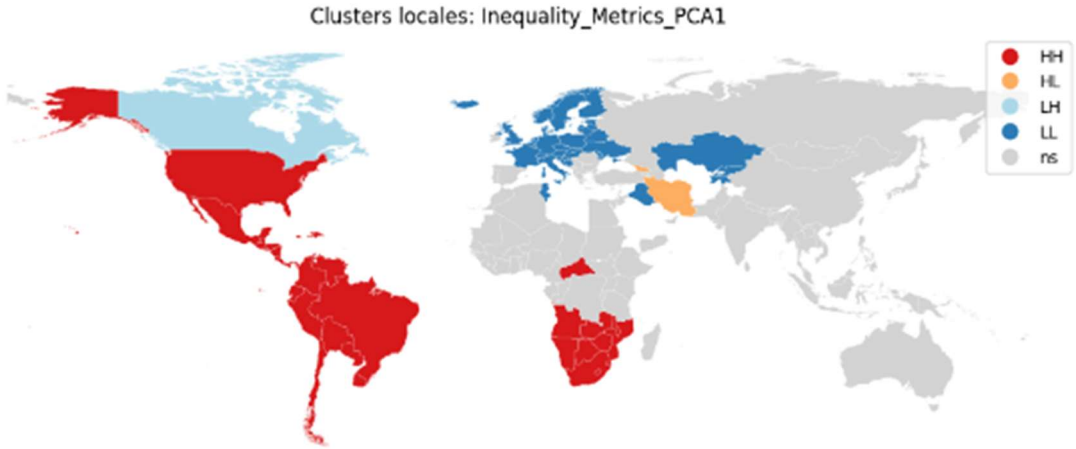

Figura 13. *Mapa de correlaciones locales significativas para la primera componente principal del subgrupo Inequality Metrics*

6.4. Análisis cluster

Pasamos pues con esta información a realizar un análisis cluster. Tras realizar diversas iteraciones y en vista de las correlaciones entre las componentes principales mostradas en la matriz de correlaciones y en el análisis espacial, decidimos que el clustering será únicamente realizado con las variables 'Average_Shortfall_PCA1' e 'Inequality_Metrics_PCA1', las cuales medirán, a grandes rasgos, la pobreza y la desigualdad respectivamente.

Realizamos diversas iteraciones para determinar el número óptimo de clusters con un método no jerárquico (K-meas) y uno jerárquico (Ward), siendo la métrica de bondad de ajuste el Silhouette score. En la *Figura 14* observamos como el número de clusters óptimos parece ser de 7 para k-means y de 8 para el método de Ward. Cabe destacar que ambos métricos ofrecen puntuaciones no muy altas del entorno de 0.42:

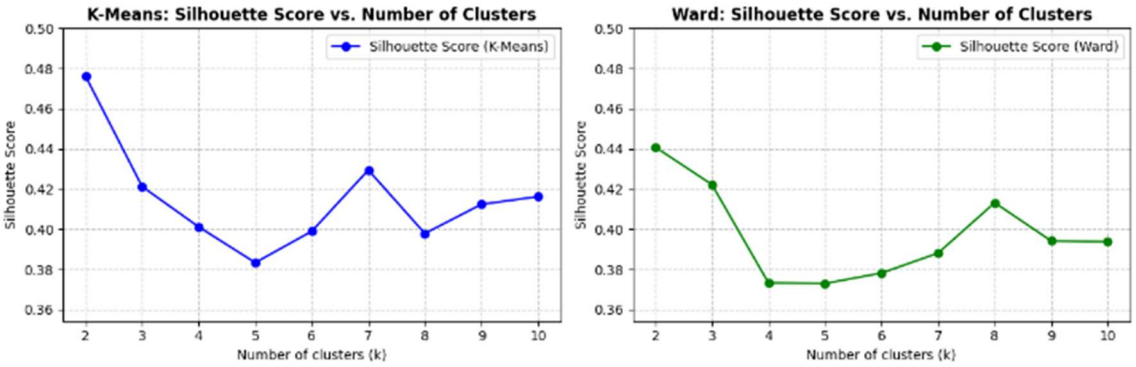

Figura 14. *Silhouette score según diferentes números de clúster para los métodos K-means y Ward*

Con los números de clúster óptimos, probamos graficar el resultado del clustering de cada método (*Figura 15*).

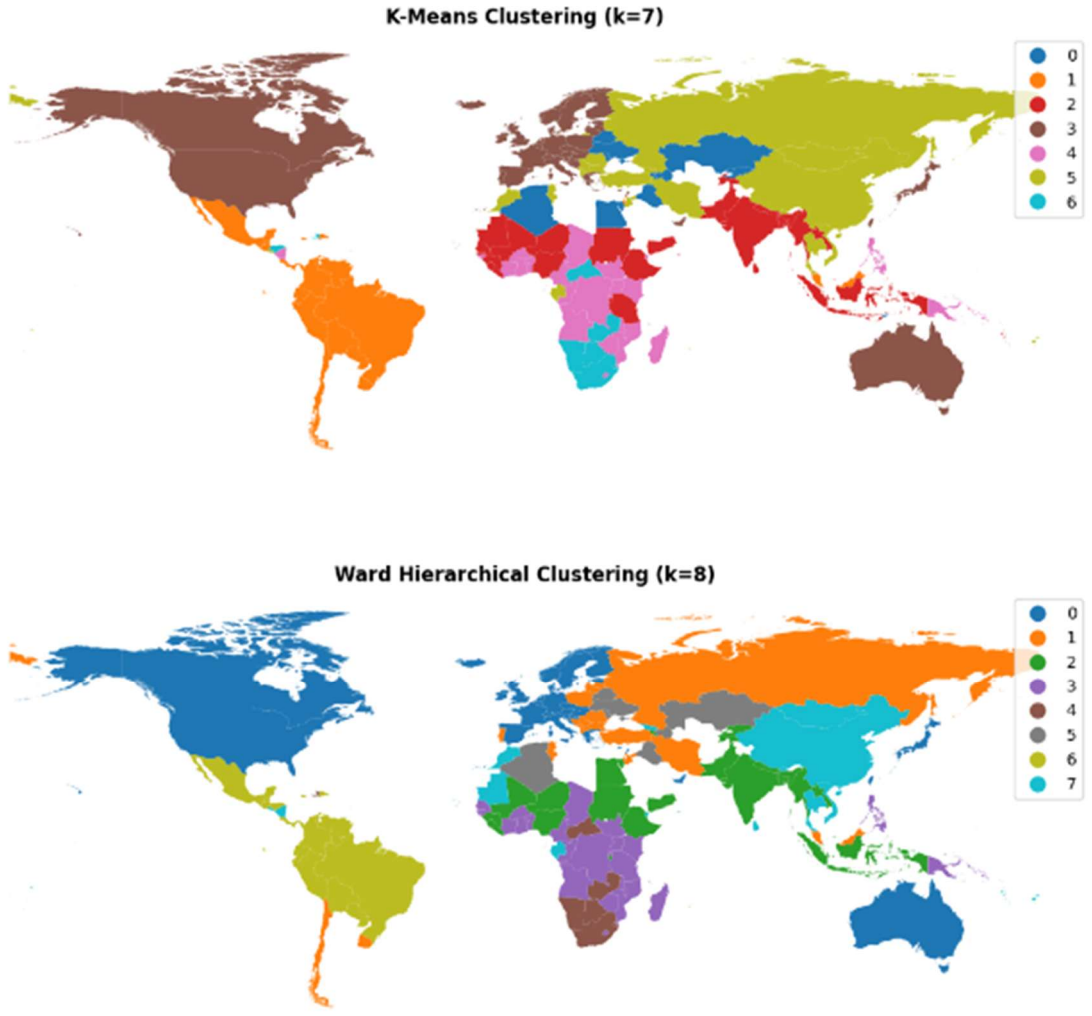

Figura 15. *Resultados del clustering con K-means (k=7) y el método de Ward (k=8)*

De la figura anterior parece complicado extraer conclusiones, ya que obviando cualquier métrica de bondad de ajuste, los resultados son claramente caóticos. Por tanto, partiendo del número de clusters óptimos establecido por k-means de 7, aplicamos un clustering espacial con Skater, obteniendo los resultados de la *Figura 16:*

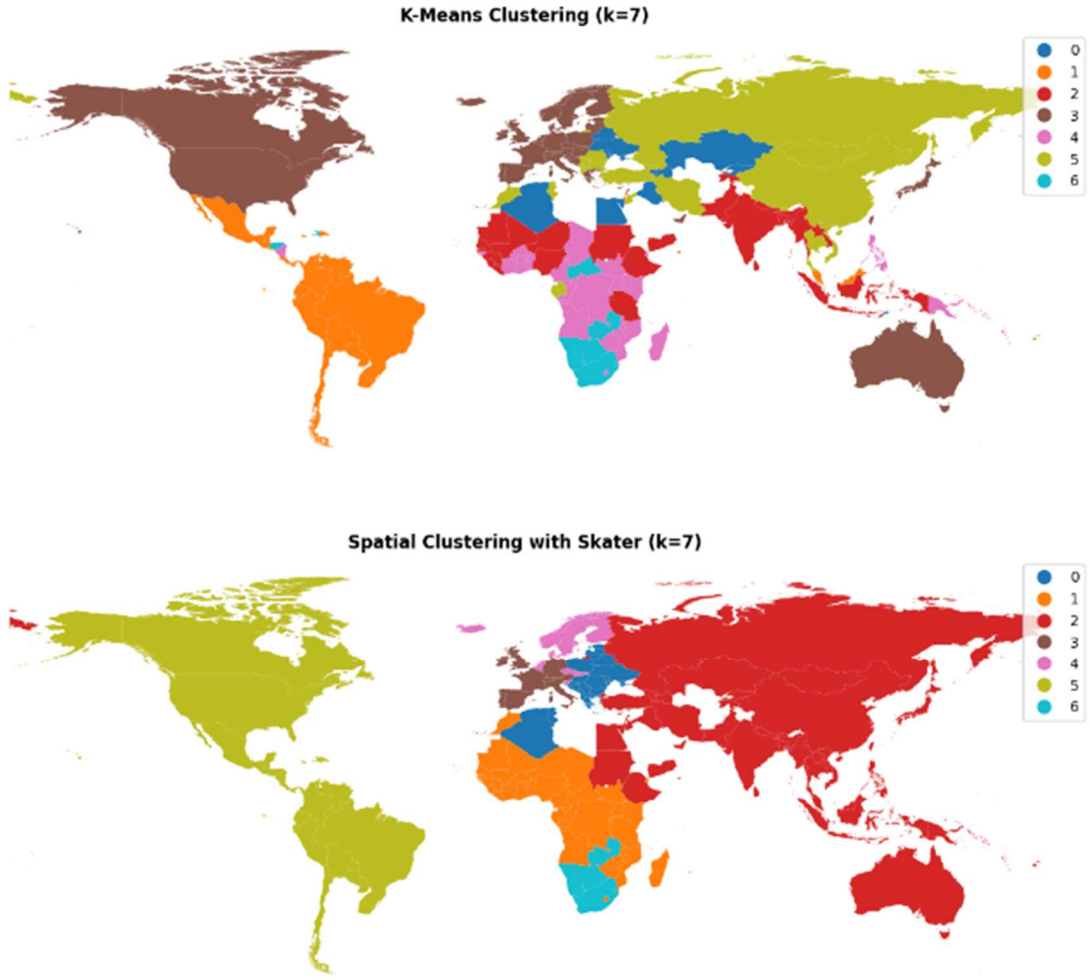

Figura 16. *Resultados del clustering con K-means y Skater con (k=7)*

A primera vista los resultados parecen mucho más interpretables para el método de Skater. Sin embargo, tal y como vemos en la *Tabla 1*, la bondad de ajuste es ridículamente menor a la obtenida por K-means o el método de Ward:

Tabla 1. *Métricas de bondad de ajuste para los distintos métodos de clustering y k=7.*

	K-means	Método de Ward	Skater
Silhouette Score	0.429	0.389	0.118
Calinski-Harabasz Index	171.2	160.95	51.38

Puesto que, tal y como se ha explicado en la metodología, un mayor Silhouette Score y Calinski-Harabasz Index indican un mejor ajuste, podemos afirmar que el método que mejor clasifica es k-means. Sin embargo, el problema está en la interpretación del clustering, aspecto que es además el más relevante a la hora de realizar agrupaciones.

Destacamos también que Skater funciona mejor para otros tamaños de cluster, ya que las restricciones de continuidad entre territorios pueden variar su k-óptimo con respecto a los métodos K-means y Ward. De hecho, para un número k=4 el Silhouette Score asciende a 0.174 y el Calinski-Harabasz Index a 67.53 o para k=3 ascenderían a 0.228 y 66.08 respectivamente. En la *Figura 17* observamos este nuevo resultado de clustering:

Figura 17. *Resultados del clustering con Skater con (k=3)*

Un total de 3 clúster resulta sencillamente interpretable con las variables que hemos empleado para realizar la agrupación. En la *Tabla 2* apreciamos los valores de los centroides para cada uno de los clústers, donde tal y como hemos destacado a partir del gráfico de cargas de las variables de PCA, valores altos de Inequality_Metrics_PCA1 se vinculan a mayor desigualdad, y valores altos de Average_Shortfall_PCA1 con niveles altos de pobreza.

Tabla 2. *Valores de los centroides resultantes del clustering con Skater y k=3.*

Clúster	Inequality_Metrics_PCA1	Average_Shortfall_PCA1
0 (Europa)	1.74	2.03
1 (África, Asia y Oceanía)	0.48	2.38
2 (América)	2.52	0.36

Así pues, a partir de la tabla anterior intuimos que el Clúster 1 es el más igualitario pero con mayor pobreza. Mientras que América tendría una mayor desigualdad pero menor pobreza, y Europa tendría unos valores medios de desigualdad y pobreza, lo cual no parece alinearse con los hallazgos de autocorrelación local.

7. CONCLUSIONES

El estudio ha cumplido su objetivo general al analizar de forma multivariante la pobreza y desigualdad global, reduciendo la dimensionalidad de los datos e identificando agrupaciones significativas de países desde una perspectiva socioeconómica y geoespacial.

Mediante el Análisis de Componentes Principales (PCA), se ha logrado reducir la dimensionalidad de 102 variables a 14 componentes principales con interpretaciones socioeconómicas relativamente claras. Estas componentes sintetizan información clave sobre pobreza y desigualdad sin pérdida significativa de datos, permitiendo una representación eficiente de los patrones subyacentes.

En cuanto a los métodos de agrupación, tanto el clustering jerárquico como el no jerárquico han mostrado desempeños similares, con valores de bondad de ajuste relativamente bajos. Sin embargo, el clustering espacial ha presentado un rendimiento considerablemente peor, con métricas significativamente más bajas. Esto sugiere que, si bien la inclusión de restricciones geográficas puede mejorar la coherencia interpretativa, también puede limitar la capacidad de los algoritmos para encontrar estructuras óptimas en los datos.

Las métricas de bondad de ajuste obtenidas, en particular el Silhouette Score y el Calinski-Harabasz Index, han sido bajas en general, lo que plantea dudas sobre la fiabilidad de los clústeres identificados. Aunque los resultados permiten ciertas interpretaciones, la baja calidad del ajuste indica que las agrupaciones podrían no reflejar patrones reales de manera robusta. Un mayor volumen de datos, con una mayor cobertura geográfica y temporal, podría mejorar la estabilidad y precisión del análisis, permitiendo obtener clústeres más representativos y menos sensibles a las limitaciones de la muestra actual.

Pese a las limitaciones en la calidad del ajuste, la incorporación del análisis espacial ha permitido identificar tendencias geográficas relevantes, como la menor pobreza en Europa en comparación con África o Asia, y las diferencias en desigualdad entre continentes. No obstante, estos hallazgos deben tomarse con precaución, ya que las métricas de bondad de ajuste sugieren que el modelo aún puede mejorarse con datos adicionales y metodologías más sofisticadas.

En conclusión, la reducción de dimensionalidad ha cumplido su función de optimizar el conjunto de datos y el clustering ha permitido esbozar agrupaciones con sentido socioeconómico y geográfico. Sin embargo, la baja calidad de ajuste indica que se deben explorar enfoques adicionales, como la incorporación de más datos o el uso de técnicas híbridas, para obtener resultados más fiables y representativos.

8. REFERENCIAS

[1] Naciones Unidas. (2015). *Transformar nuestro mundo: La Agenda 2030 para el Desarrollo Sostenible.* ONU. https://unctad.org/system/files/official-document/ares70d1_es.pdf

[2] Atkinson, A. B. (2019). *Measuring Poverty Around the World*. Princeton University Press.

[3] Alkire, S., & Foster, J. (2011). *Counting and multidimensional poverty measurement*. Journal of Public Economics, 95(7-8), 476-487. https://www.sciencedirect.com/science/article/abs/pii/S0047272710001660

[4] Red Europea de Lucha contra la Pobreza y la Exclusión Social en el Estado Español (EAPN). (2015). *Guía básica de indicadores de desigualdad, pobreza y exclusión social*. https://www.eapn.es/ARCHIVO/documentos/documentos/1446118622_guia_basica_indicadores_desigualdad_pobreza_y_exclusion_social.pdf

[5] López, R., & Pérez, M. (2020). *Enfoques, definiciones y estimaciones de pobreza y desigualdad en América Latina y el Caribe: un análisis crítico de la literatura*. https://www.researchgate.net/publication/346095692_Enfoques_definiciones_y_estimaciones_de_pobreza_y_desigualdad_en_America_Latina_y_el_Caribe-_un_analisis_critico_de_la_literatura

[6] Kaggle. (s.f.). *Global Poverty and Inequality Dataset*. https://www.kaggle.com/datasets/utkarshx27/global-poverty-and-inequality-dataset/data?select=pip_codebook.csv