Investigación & Desarrollo
Revista periódica de difusión científica

Recibido el: 15 de enero de 2025
Aceptado el: 21 de mayo de 2025
DOI: 10.23881/idupbo.025.2-3e

Datos faltantes y análisis multidimensional con el método Alkire-Foster: Una revisión de la literatura

Missing data and multidimensional analysis with the Alkire-Foster method: A literature review

Daniel Alejandro Revilla Cabrera*
Universidad Privada Boliviana (UPB), Bolivia
danielrevilla@upb.edu
0009-0006-3979-5658

Resumen: Este artículo aborda el problema de los datos faltantes en la medición de pobreza multidimensional, con especial atención a su impacto sobre la validez axiomática del método Alkire-Foster, empleado regularmente para medir este concepto amplio de pobreza. Se sostiene que la ausencia de datos no es un problema puramente técnico y proclive a generar sesgos, sino que también puede comprometer la identificación de privaciones y vulnerar alguno de los axiomas fundamentales del método: monotonicidad, descomposición por subgrupos y descomposición dimensional. A partir de una revisión crítica de la literatura, se comparan cuatro enfoques de imputación: análisis de casos completos, imputación múltiple por ecuaciones encadenadas (MICE), técnicas de descomposición matricial (como Soft-Impute) y redes neuronales bayesianas. Se discuten sus fortalezas y limitaciones en relación con la preservación de las propiedades axiomáticas mencionadas. Además, se presentan argumentos a favor de una metodología híbrida que combine algoritmos imputativos con criterios de validación normativa para garantizar coherencia ética y solidez inferencial. El artículo concluye que la elección de métodos de imputación en contextos sociales debe considerar no solo la eficiencia estadística, sino también la compatibilidad estructural con el modelo teórico que subyace el ejercicio de medición de la pobreza. Se proponen líneas futuras para el desarrollo de marcos integradores que articulen métodos de imputación, el respeto de axiomas y el uso en política pública.

Palabras clave: Datos faltantes, pobreza multidimensional, Alkire-Foster, imputación múltiple, ética del dato, redes neuronales bayesianas.

Abstract: *This article addresses the problem of missing data in measuring multidimensional poverty, with particular attention to its impact on the axiomatic validity of the Alkire-Foster method, regularly used to measure this broad concept of poverty. It argues that the absence of data is not purely a technical problem prone to generating bias, but can also compromise the identification of deprivation and violate some of the fundamental axioms of the method: monotonicity, decomposition by subgroups, and dimensional decomposition. Based on a critical review of the literature, four imputation approaches are compared: complete case analysis, multiple imputation by chained equations (MICE), matrix decomposition techniques (such as Soft-Impute), and Bayesian neural networks. Their strengths and limitations in relation to the preservation of the aforementioned axiomatic properties are discussed. In addition, arguments are presented in favor of a hybrid methodology that combines imputation algorithms with normative validation criteria to ensure ethical consistency and inferential robustness. The article concludes that the choice of imputation methods in social contexts should consider not only statistical efficiency but also structural compatibility with the theoretical model underlying the poverty measurement exercise. Future lines of research are proposed for the development of integrative frameworks that articulate imputation methods, respect for axioms, and use in public policy.*

Keywords: *Missing data, multidimensional poverty, Alkire-Foster, multiple imputation, data ethics, Bayesian neural networks.*

*Corresponding autor

1. Introducción

El análisis de índices multidimensionales ha cobrado relevancia particular al momento de medir condiciones de vida como una alternativa más comprehensiva a las mediciones unidimensionales basadas exclusivamente en el ingreso [1]. No obstante, los problemas asociados a los datos faltantes en las encuestas de hogares que normalmente se emplean para construir estos índices, constituyen una fuente crítica de sesgo y pérdida de eficiencia, comprometiendo la validez de los resultados [24].

La medición de la pobreza ha experimentado un giro sustantivo en las últimas décadas, superando la visión restringida centrada exclusivamente en los ingresos y reconociendo que las privaciones humanas son, en esencia, múltiples, entrelazadas y contextualmente determinadas. Esta expansión conceptual ha impulsado el desarrollo de una familia diversa de métodos cuantitativos, todos orientados a capturar de manera más fiel las condiciones de privación que enfrentan los hogares y las personas. En este panorama, el método de Alkire y Foster [1] ha logrado posicionarse como una de las metodologías más utilizadas y normativamente coherentes para el estudio de la pobreza multidimensional.

El énfasis en el método Alkire-Foster (AF) en el presente trabajo no responde a un criterio de replicabilidad internacional, sino a una decisión metodológicamente informada, basada en la solidez teórica, la transparencia operativa y la potencia analítica que ofrece este enfoque en escenarios de evaluación empírica de intervenciones en contra de la pobreza. Particularmente, el método AF permite una articulación entre principios normativos (derivados de la tradición del Enfoque de las Capacidades) y exigencias prácticas en la medición, a través de una arquitectura que cumple con tres axiomas esenciales para su aplicabilidad en el diseño e implementación de políticas públicas: monotonicidad, descomposición por subgrupos y descomposición por dimensiones [1, 3].

Cuando se trata de aplicar el método AF con datos reales, la ausencia de información parcial –por ejemplo, por altas tasas de no respuesta, o por problemas en campo para recolectar información– puede comprometer el cumplimiento de estos axiomas [1]. En este contexto, resulta indispensable avanzar hacia estrategias que permitan completar la matriz de privaciones en la que se basa el método AF, garantizando que la medición resultante sea tanto empíricamente sólida como normativamente coherente.

Este artículo revisa los principales enfoques de imputación de datos faltantes aplicables en estudios que utilizan el método Alkire-Foster, con énfasis en sus fundamentos teóricos, ventajas comparativas y limitaciones prácticas. El documento se organiza en cinco secciones. La sección 2 presenta una revisión de los enfoques más comúnmente empleados en la literatura para medir pobreza multidimensional y presenta argumentos en favor de la concentración del análisis de datos faltantes en el contexto de la metodología propuesta por Alkire y Foster. La sección 3 ofrece una conceptualización crítica de los mecanismos de datos faltantes, abordando sus implicancias metodológicas y normativas en contextos de medición multidimensional. En la sección 4 se presentan y analizan comparativamente los principales métodos contemporáneos de imputación de datos (incluyendo MICE, Soft-Impute y redes neuronales bayesianas), y se introduce una propuesta metodológica híbrida que busca preservar la coherencia estructural del enfoque Alkire-Foster.

La sección 5 desarrolla una reflexión metodológica sobre la imputación como acto epistémico, explorando su relación con la construcción normativa del dato. Asimismo, en esta sección se exponen las reflexiones finales del estudio, destacando sus contribuciones principales y delineando una agenda de investigación futura orientada a la validación estructural y ética de los procesos de imputación.

2. Métodos para medir pobreza en múltiples dimensiones

El enfoque AF hereda su estructura de agregación del índice desarrollado por Foster, Greer y Thorbecke [13] ampliamente conocido como el método FGT. Este enfoque clásico ha sido central en la medición unidimensional de la pobreza, permitiendo calcular no sólo tasas de incidencia, sino también medidas de profundidad y severidad. No obstante, el FGT tradicional no está diseñado para capturar simultáneamente múltiples formas de privación. La extensión que propone el método AF, por tanto, no es meramente técnica: constituye una reestructuración conceptual del análisis de pobreza, orientada a responder a las necesidades actuales de representación, focalización y desagregación.

Diversos enfoques alternativos han sido desarrollados con el mismo objetivo general. Entre ellos, los modelos basados en funciones de bienestar [21, 27] ofrecen construcciones normativas rigurosas, pero su implementación empírica requiere supuestos de cardinalidad y sustitución entre dimensiones que suelen ser de difícil justificación. Métodos empíricos como el análisis de componentes principales (PCA) han tenido una notable difusión en salud pública y encuestas demográficas [12, 30], pero carecen de principios reflexivos y consensuados, que estén informados por las estrategias de los actores y, además, producen resultados que no pueden ser descompuestos ni interpretados normativamente. Por su parte, los modelos de clases latentes (LCA) y los enfoques de capacidades latentes [19]

capturan patrones complejos de privación, aunque con una elevada complejidad técnica y sin producir un índice agregado que cumple con las propiedades axiomáticas clásicas de una medida de pobreza [10].

Finalmente, el enfoque de dominancia estocástica multidimensional [9] constituye una alternativa robusta en términos de comparación normativa, al permitir establecer jerarquías de pobreza sin imponer una única función agregativa. Sin embargo, la carencia de una medida sintética limita severamente su utilidad en contextos de monitoreo, evaluación o asignación presupuestaria.

En resumen, el enfoque Alkire-Foster permite capturar la complejidad estructural de la pobreza desde una plataforma metodológicamente robusta y operativamente viable. Esta característica lo convierte en una opción particularmente relevante para investigaciones orientadas a incidir en política pública, como es el caso del presente trabajo.

A pesar de la solidez estructural del método Alkire-Foster, su implementación empírica se enfrenta a una limitación recurrente en contextos reales de investigación: la presencia de datos faltantes. Las encuestas de hogares, registros administrativos y censos sociales suelen presentar omisiones en variables relevantes por múltiples motivos, tales como errores de levantamiento, no respuesta voluntaria, fallas técnicas o por la propia sensibilidad de ciertas dimensiones de bienestar. Esta ausencia de información no es neutral desde el punto de vista analítico: puede introducir sesgos importantes en la medición de la pobreza si el mecanismo de omisión no es completamente aleatorio [20]. Cuando los datos faltan de manera dependiente a características observadas o no observadas (es decir, cuando responden a mecanismos MAR o MNAR según la clasificación estándar), el problema trasciende la reducción del tamaño muestral y compromete directamente la validez inferencial del índice construido [25].

En el marco del enfoque Alkire-Foster, esta problemática adquiere una dimensión crítica. Dado que el método identifica como pobre a una persona cuya carga ponderada de privaciones supera un umbral específico (k), la ausencia de información en una o varias dimensiones puede alterar injustificadamente su estatus de privación. Más aún, si esta omisión se produce de manera sistemática en ciertas dimensiones o subgrupos poblacionales, las propiedades fundamentales del índice (como la monotonicidad, la descomposición por subgrupos o la descomposición por dimensiones) pueden verse comprometidas.

En consecuencia, la forma en que se abordan los datos faltantes no puede desligarse del marco axiológico del método: no basta con restaurar la información perdida desde una lógica estadística, sino que es necesario asegurar que las soluciones adoptadas no violen los principios normativos que confieren legitimidad y utilidad al índice. Como advierten Ferreira y Lugo [11], los métodos de medición de pobreza que descansan en estructuras normativas explícitas deben prestar especial atención a las implicancias éticas y técnicas de la ausencia de datos.

Tabla 1: Comparación de métodos para medir pobreza multidimensional

Método	Fortalezas clave	Limitaciones principales
Alkire-Foster (AF)	Cumple axiomas clave (monotonicidad, descomposición por subgrupos y dimensiones). Adaptable. Índice claro y descomponible.	Requiere decisiones normativas sobre dimensiones, ponderaciones y umbrales.
FGT (Unidimensional)	Mide incidencia, profundidad y severidad. Fórmula clara. Descomponible.	No considera múltiples dimensiones. Centrado en ingreso/gasto.
Funciones de bienestar	Sólido marco teórico. Modela sustitución entre dimensiones.	Supuestos fuertes sobre preferencias. Aplicación empírica compleja.
PCA / Análisis factorial	Técnica sencilla. Usa correlaciones estadísticas.	No cumple axiomas. Pesos endógenos. Sin descomposición.
Clases latentes (LCA)	Identifica perfiles ocultos. Segmentación útil.	No produce índice agregado. Alta sensibilidad técnica.
Dominancia estocástica	Comparaciones robustas sin índice único. Independiente de ponderaciones.	No produce una medida agregada. Limitada utilidad para monitoreo o diseño de política.

Fuente: Elaboración propia.

3. Conceptualización de los datos faltantes

Siguiendo la clasificación canónica de Rubin [24], los mecanismos de datos faltantes se categorizan en tres tipos:

- Faltantes completamente al azar (MCAR)
- Faltantes al azar (MAR)
- Faltantes no al azar (MNAR)

La elección de la estrategia analítica adecuada (incluyendo la imputación o el modelado) depende críticamente de la naturaleza del mecanismo de no respuesta, el cual rara vez es plenamente identificable en estudios empíricos [20].

En los procesos de análisis empírico, especialmente aquellos que involucran grandes bases de datos provenientes de encuestas de hogares, censos de población o registros administrativos, la presencia de datos faltantes constituye una condición estructural de la evidencia disponible, más que una anomalía excepcional. Su aparición responde a una multiplicidad de causas, entre ellas la no respuesta parcial o total, errores en la recolección, fallas en la codificación, problemas de sensibilidad temática (por ejemplo, ingresos, violencia, salud reproductiva), o incluso por decisiones metodológicas ex post como la exclusión de valores extremos. Esta ausencia de información plantea un desafío crítico tanto para la validez estadística de los análisis como para su coherencia normativa cuando se trabaja con marcos como el de Alkire-Foster [1], donde la lógica de identificación y agregación descansa en una matriz completa de privaciones ponderadas.

Desde un punto de vista metodológico, los mecanismos MCAR, MAR y MNAR presentan grados diferenciados de dificultad en el tratamiento y análisis. En el caso de los datos MCAR, la probabilidad de ausencia es independiente de los valores observados o no observados; es decir, los datos faltantes se distribuyen de manera completamente aleatoria entre la población analizada. Esta condición permite el uso de análisis de casos completos sin inducir sesgo, aunque con pérdida de eficiencia [20]. Rubin [24] proporciona ejemplos en contextos de simulación donde la omisión aleatoria de respuestas no distorsiona los resultados inferenciales. Sin embargo, la literatura crítica ha señalado que esta condición es empíricamente rara y, en la práctica, suele asumirse sin verificación formal [25].

El mecanismo MAR es más común en la práctica y realista: la probabilidad de ausencia depende únicamente de variables observadas para la población analizada, pero no de aquellas que no son observadas, ni de los valores faltantes en sí. Esto permite modelar explícitamente la falta de datos a partir de información disponible bajo el supuesto de que los factores relevantes para la identificación del patrón que genera los datos faltantes se encuentran completamente medidos. Por ejemplo, en estudios de salud longitudinal, se ha documentado que la deserción de participantes está asociada a variables como edad o nivel educativo, lo cual permite tratar las ausencias bajo el supuesto MAR si tales variables son incluidas en el modelo [20].

En contraste, el mecanismo MNAR se presenta cuando la ausencia de datos depende de los propios valores faltantes o de variables no observadas, lo cual implica que ni las técnicas basadas en MCAR ni en MAR son suficientes para evitar sesgos. En estos casos, se requiere modelar conjuntamente el proceso de ausencia y el fenómeno de interés, lo cual implica supuestos fuertes e instrumentos adicionales difíciles de validar empíricamente [4, 6]. Un ejemplo común lo constituyen las encuestas de ingresos, donde las personas con mayores ingresos tienden a no responder, generando una omisión sistemática que compromete la estimación de la desigualdad [20].

En el marco del enfoque Alkire-Foster, estas distinciones adquieren una relevancia particular. En efecto, dado que el método se basa en el conteo ponderado de privaciones para la identificación de los pobres, la ausencia de datos en una dimensión puede impedir, artificialmente, la clasificación de un individuo como pobre o, inversamente, suponer una privación inexistente si se sustituye sin precaución. Además, si las ausencias afectan más frecuentemente a ciertas dimensiones o grupos poblacionales, pueden inducir sesgos que violen los axiomas fundamentales del enfoque (en particular, la monotonicidad y la descomposición por subgrupos) comprometiendo así la legitimidad ética y política de los resultados obtenidos.

La literatura ha enfatizado que, bajo marcos axiomatizados, la estrategia de tratamiento de datos faltantes no puede desligarse del marco normativo del método [11] y por tanto, cualquier decisión técnica debe ser evaluada no sólo por su consistencia estadística, sino también por su compatibilidad con la arquitectura ética de la medición.

4. Métodos de imputación: una síntesis crítica

La selección de una estrategia adecuada para el tratamiento de datos faltantes constituye una decisión metodológica de primer orden, especialmente en estudios donde la coherencia normativa del índice es tan importante como su validez estadística. En el marco del enfoque Alkire-Foster, donde la identificación de los pobres depende de la agregación de múltiples dimensiones de privación y el respeto de axiomas fundamentales como la monotonicidad o la descomposición por subgrupos, cualquier procedimiento de imputación debe ser evaluado tanto por su precisión técnica como por su compatibilidad con esta arquitectura lógica.

Siguiendo con la reflexión crítica, se propone una comparación detallada de cuatro metodologías que, por su relevancia empírica, desarrollo teórico y aplicabilidad en contextos de alta complejidad estructural, resultan especialmente pertinentes para el tratamiento de datos faltantes en análisis multidimensionales: el análisis de casos completos (complete case analysis), la imputación múltiple por ecuaciones encadenadas (Multiple Imputation by Chained Equations, MICE), las técnicas de descomposición matricial de bajo rango (como Soft-Impute), y los modelos basados en redes neuronales bayesianas (BNN). Cada uno de estos enfoques representa una tradición metodológica distinta, con fortalezas y limitaciones específicas.

El análisis de casos completos, a pesar de sus conocidas limitaciones bajo mecanismos MAR o MNAR, continúa siendo una de las prácticas más comunes en la investigación empírica aplicada debido a su simplicidad operativa [20]. Por otro lado, MICE ha ganado notable aceptación por su flexibilidad para imputar variables de diversa naturaleza en estructuras de datos complejas [7,5]. Las técnicas de descomposición matricial, como Soft-Impute, han demostrado una gran eficacia en entornos de alta dimensionalidad y escasez informativa [22].

Finalmente, los modelos basados en redes neuronales bayesianas ofrecen una vía prometedora al combinar la capacidad de aprendizaje profundo con la inferencia probabilística, permitiendo imputaciones consistentes en entornos no lineales y con alta incertidumbre. En esta última categoría destacan contribuciones como las de Hruschka et al. [18], que formalizaron el uso de redes bayesianas para clasificación con imputación [8], que aplicaron marcos bayesianos en series temporales ambientales, y Gondara y Wang [15], cuyo modelo MIDA constituye uno de los avances más influyentes en imputación automatizada con autoencoders. La sección se cierra con un apartado dedicado a discutir las bondades de enfoques híbridos, que buscan articular las virtudes de múltiples metodologías para preservar tanto la coherencia estructural del índice como su robustez empírica.

4.1. Lista completa

Uno de los enfoques más ampliamente utilizados para el tratamiento de datos faltantes, particularmente en estudios empíricos con limitaciones operativas o metodológicas, es el denominado análisis de casos completos (complete case analysis), también conocido como lista completa o case-wise deletion.

Este método consiste en excluir del análisis todas aquellas observaciones que presentan datos faltantes en una o más variables relevantes, trabajando únicamente con el subconjunto de casos totalmente observados. Su popularidad radica en su simplicidad, facilidad de implementación y compatibilidad directa con la mayoría de los algoritmos estadísticos tradicionales. Sin embargo, esta estrategia se apoya en un supuesto altamente restrictivo: la validez inferencial de los resultados depende críticamente de que los datos estén ausentes de forma completamente aleatoria (Missing Completely at Random, MCAR). Cuando este supuesto no se cumple (lo que ocurre con frecuencia en aplicaciones reales), el método introduce sesgos significativos y una pérdida innecesaria de eficiencia muestral, comprometiendo así la validez tanto de las estimaciones como de sus intervalos de confianza [25, 20].

Existen múltiples ejemplos empíricos que ilustran las consecuencias de aplicar este enfoque en condiciones que no satisfacen MCAR. En un análisis del National Health and Nutrition Examination Survey (NHANES), Allison [2] muestra que excluir casos con datos faltantes en variables de salud y comportamiento induce una subestimación sistemática de riesgos asociados con hipertensión, dado que los individuos con condiciones más graves eran menos propensos a responder completamente.

Un caso similar fue documentado por Vink [29]en el contexto de estudios longitudinales sobre el bienestar infantil, donde la omisión de observaciones incompletas condujo a una sobreestimación de la asociación entre nivel educativo de los padres y logros académicos. En ambos ejemplos, la solución metodológica consistió en reemplazar el enfoque de lista completa por técnicas de imputación múltiple, particularmente MICE, que permitieron recuperar la información ausente y restaurar la validez de las inferencias. Estos casos refuerzan la advertencia metodológica de que el uso acrítico del análisis de casos completos puede ser no solo ineficiente, sino profundamente engañoso en contextos sociales, económicos y de salud pública donde los datos faltantes rara vez son aleatorios.

4.2. Método de Imputación por Ecuaciones Encadenadas - MICE

Este método busca realizar la imputación de datos faltantes a través del supuesto de que cada variable que tenga datos faltantes puede modelarse de manera condicionada a la información con la que sí se cuenta en la base de datos, sean los indicadores mismos que se busca imputar u otras variables observables entre a población. El proceso de imputación es iterativo y responde al siguiente conjunto de ecuaciones [31].

Dado un conjunto de datos P con variables $X_1, X_2, ..., X_p$ donde algunas de ellas tienen valores faltantes, se asume que la distribución conjunta de datos sigue la siguiente regla de factorización condicional:

$$P_{MICE}(X_1, X_2, \ldots, X_p) = \prod_{j=1}^{p} P_{MICE}\,(X_j \mid X_{-j})$$

Cada variable con faltantes X_{-j} se modela condicionalmente a las demás variables X_{-j}:

$$P_{MICE}(X_j|X_{-j}) = f_i(X_{-j}) + \epsilon_j$$

donde:

- $f_{-j}(\cdot)$ es un modelo predictivo específico en función del conjunto de datos.
- X_{-j} son las demás variables del conjunto de datos, observadas o imputadas previamente.
- ϵ_j es el error en la predicción del valor faltante, incluido en la ecuación de manera aditiva.

A través de este método se iniciará la imputación de datos a cada variable que contenga datos faltantes de acuerdo con el siguiente método secuencial. En primer lugar, y tomando en cuenta el valor faltante para el individuo *i* en el indicador *j* (X_{ij}), se realiza una imputación preliminar basada en la media si la variable es continua y la moda si la variable es categórica.

$$X_{ij}^{(0)} = \begin{cases} X_{ij} & \text{, si el dato existe} \\ \dfrac{1}{N_j}\displaystyle\sum_{k=1}^{N_j} X_{kj} & \text{, si el dato es faltante} \end{cases}$$

A partir de esta imputación inicial para todos los valores faltantes en el indicador *j* (X_{ij}) se introduce un ajuste al mismo en función del tipo de dato trabajado de la siguiente manera.
Si X_j es continua se aplica una regresión lineal:

$$X_j = \beta_0 + \sum_{k \neq j} \beta_k X_k + \epsilon_j, \epsilon_j \sim N(0, \sigma^2)$$

Si X_j es categórica se aplica una regresión logística multinomial:

$$P_{MICE}(X_j = k) = \frac{e^{\beta_k X_{-j}}}{\sum_i e^{\beta_i X_j}}$$

Si X_j es binaria se aplica una regresión logística:

$$P_{MICE}(X_j = 1) = \frac{1}{1 + e^{-(\beta_0 + \sum_{k \neq j} \beta_k X_k)}}$$

Cada observación faltante X_{ij} es reemplazada por una muestra de la distribución modelo en la siguiente iteración de imputación (llamada genéricamente *t* +1):

$$X_{ij}^{(t+1)} \sim P_{MICE}(X_j|X_{-j}^{(t)})$$

El proceso previamente descrito se repite con todas las variables con datos faltantes hasta alcanzar la convergencia, normalmente medida a través de:

$$max|X_j^{(t+1)} - X_j^{(t)}| < \epsilon$$

Con base en esta situación se generan múltiples conjuntos de datos imputados, mismos que a través de una combinación de datos usando la fórmula de Rubin, permitiendo así incorporar la incertidumbre en la imputación del estadístico que se quiere reproducir, incluso en presencia de datos faltantes (denotado como $\hat{\theta}_k$). Luego de *m* iteraciones para imputar los datos faltantes, la estimación del estadístico de interés a partir de este método es $\bar{\theta}$ calculado como:

$$\bar{\theta} = \frac{1}{m}\sum_{k=1}^{m} \theta_k$$

Las fórmulas de Rubin permiten estimar parámetros como la media del estadístico de interés y su varianza, entre otras. Además, la varianza total (denotada como *T)*, se puede descomponer de la siguiente manera:

$$T = W + \left(1 + \frac{1}{m}\right) B$$

donde:

- *W* es la varianza dentro de los conjuntos imputados:

$$W = \frac{1}{m}\sum_{k=1}^{m} Var(\theta_k)$$

- *B* es la varianza entre los conjuntos imputados:

$$B = \frac{1}{m-1}\sum_{k=1}^{m} (\theta_k - \bar{\theta})^2$$

Para más detalles ver [28]. Así, este método permite realizar la imputación de los datos y su posterior verificación para poder tener el conjunto de datos completos y adecuados para trabajos correspondientes.

4.3. Método de Imputación a través de Matrices de Bajo Rango - SOFT-IMPUTE

El método Soft-Impute está basado en la descomposición de valores singulares (SVD) para la imputación de datos faltantes. Se basa en la idea de que los datos tienen una estructura de bajo rango, por lo que los valores faltantes pueden ser reconstruidos utilizando una aproximación de matriz de bajo rango.

Su fundamento se explica dado un conjunto de datos representados por una matriz de entrada $\mathbf{X} \in \mathbb{R}^{m \times n}$ donde existen datos faltantes [26]. El algoritmo estructurará el problema de

imputación como una minimización de función de costo:

$$\min_{\hat{X}} ||M \odot (X - \widehat{X})||_F^2 + \lambda \sum \sigma_i$$

donde M es una máscara binaria aplicada al conjunto de datos que muestra la presencia o ausencia de datos:

$$M_{ij} = \begin{cases} 1 & \text{, si el dato existe} \\ 0 & \text{, si el dato es faltante} \end{cases}$$

- $\| \cdot \|_F^2$ es la norma de Frobenius, misma que mide la diferencias entre la matriz original y la imputada.
- λ es un parámetro de ajuste que permite controlar el grado de suavizado aplicado a los valores singulares.
- σ_i son los valores singulares presente en la matriz X.

Debido a que existe una penalización sobre los valores singulares, se fuerza a la matriz $\hat{X}$ a tener una estructura de bajo rango. De esta manera se logra reducir el ruido y se da mayor certeza a los datos imputados.

Este método sigue los siguientes pasos:

1. Imputación inicial de valores faltantes Para cada valor faltante de la matriz X_{ij} se imputa la media de la columna a la que pertenece:

$$X_{ij}^{(0)} = \begin{cases} X_{ij} & \text{, si } M_{ij} = 0 \\ \frac{\sum_k X_{kj} M_{kj}}{\sum_k M_{kj}} & \text{, si } M_{ij} = 0 \end{cases}$$

2. Descomposición en Valores Singulares (SVD) La matriz X se descompone en tres matrices U, S y V^T. Donde:
   - $U \in \mathbb{R}^{m \times r}$ y $V \in \mathbb{R}^{n \times r}$ son matrices de componentes principales
   - $S \in \mathbb{R}^{r \times r}$ es una matriz diagonal con valores singulares σ_i

3. Penalización de valores singulares Se aplica una penalización suave:

$$S_\lambda = max(S - \lambda, 0)$$

La penalización busca eliminar los valores singulares pequeños y reduce el rango de la matriz, logrando de este modo realizar imputaciones más estables.

4. Reconstruir la matriz imputada La nueva matriz imputada es reconstruida aplicando la transformada inversa de SVD:

$$\hat{X} = US_\lambda V^T$$

5. Actualización de valores faltantes Se reemplazan los valores imputados en la matriz original a través de:

$$X^{(t+1)} = M \odot X + (1 - M) \odot \hat{X}^t$$

donde:

- M garantiza que los valores observados no tengan cambios
- (1-M) permite que solo se actualicen los valores imputados

6. Repetir hasta convergencia El proceso se repite hasta que los valores imputados converjan:

$$max|X_{ij}^{(t+1)} - X_{ij}^t| < \epsilon$$

Para más detalles, ver [33].

4.4. Método de Imputación por Redes Neuronales Bayesianas - BNN

El método de Redes Neuronales Bayesianas BNN es una variación de las redes neuronales donde los pesos y sesgos no son valores fijos, en su lugar son distribuciones de probabilidad modeladas. Estas características permiten generar valores imputados con intervalos de confianza como también capturar la incertidumbre en la predicción [18].

Una red neuronal estándar, es un modelo que aprende de un conjunto de pesos W a partir del conjunto de datos X para realizar una predicción Y:

$$P_{BNN}(Y|X, W) = f(X, W) + \epsilon$$

donde:

- $f(X, W)$ es la función no lineal que aprende la red neuronal.
- $\epsilon \sim N(0, \sigma^2)$ es un término de error.

En el caso de una red neuronal bayesiana se asume que los pesos siguen una distribución de probabilidad:

$$W \sim P_{BNN}(W)$$

Haciendo uso de la regla de Bayes para calcular la distribución a posteriori de los pesos en función de un conjunto de datos de entrenamiento D.

$$P_{BNN}(W|D) = \frac{P_{BNN}(D|W) P_{BNN}(W)}{P_{BNN}(D)}$$

donde:

- $P_{BNN}(W|D)$ es la distribución a posteriori de los pesos
- $P_{BNN}(D|W)$ es la verosimilitud de los datos
- $P_{BNN}(W)$ es la distribución previa de los pesos
- $P_{BNN}(D)$ es la evidencia

Se usan métodos de aproximación para encontrar la distribución *P(W|D)* de la siguiente manera que servirá para determinar la inferencia Bayesiana:

- Inferencia Variacional Se busca aproximar la distribución a posterior *P(W|D)* con una distribución más simple *Q(W)*, buscando así minimizar la divergencia de Kullback-Liebler (KL):

$$KL(Q(W) \,||\, P_{BNN}(W|D)) = \int Q\left(W\right) log \frac{Q(W)}{P_{BNN}(W|D)} dW$$

La función objetivo de la inferencia variacional busca maximizar el Límite Inferior Probatorio ELBO.

$$\mathcal{L}(Q) = \mathbb{E}_{Q(W)}[logP_{BNN}(D|W)] - KL(Q(W)||P_{BNN}(W))$$

- Monte Carlo Dropout En este método la aproximación se obtiene como el promedio de múltiples muestreos Monte Carlo:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f\left((W^{(t)}, x)\right)$$

donde:

- $\hat{y}$ es el valor imputado
- $W^{(t)}$es un subconjunto de pesos activados en la iteración i
- T es el número de muestreos Monte Carlo

A través de este método se genera una distribución de valores imputados que siguen la siguiente forma:

$$X_j^{imputado} = \mathbb{E}[P_{BNN}(X_j|D)]$$

Estos métodos se aplican siguiente los siguientes pasos:

1. Definición del modelo bayesiano Se debe definir distribuciones previas sobre los pesos y los sesgos

$$W \sim \mathcal{N}(0, \sigma^2 I)$$
$$b \sim \mathcal{N}(0, \sigma^2)$$

Se debe utilizar una función de activación $\phi(\cdot)$ donde la salida de la capa oculta sigue una distribución Gaussiana.

$$h = \phi(XW + b)$$

El resultado final sigue una distribución normal con media HW' y varianza de σ^2

$$y \sim \mathcal{N}(hW', \sigma^2)$$

2. Entrenamiento del modelo El modelo se debe entrenar basándose en la log-verosimilitud marginal:

$$logP_{BNN}(D) = \int P_{BNN}(D|W) P_{BNN}(W) dW$$

En este punto se debe utilizar uno de los métodos de aproximación previamente descritos.

3. Predicción de valores faltantes Se realiza la predicción de los valores faltantes

$$X_j^{imputado} = \mathbb{E}[P_{BNN}(X_j|D)$$

La incertidumbre de esta imputación es cuantificada a través de la varianza de la distribución predictiva:

$$\sigma^2_{imputado} = Var(P_{BNN}(X_j|D))$$

Permitiendo de esta manera intervalos de confianza para cada imputación realizada [8].

Esto permite que el modelo no sólo impute valores esperados, sino que ofrezca intervalos de confianza, fundamentales para evitar una representación artificialmente precisa de las privaciones, especialmente en dimensiones sensibles como salud o violencia.

4.5. Método Híbrido

A partir de la revisión realizada, queda claro que todos los métodos de imputación revisados tienen fortalezas y desafíos al momento de ser empleados en una aplicación empírica. Esto sugiere que debería privilegiarse una metodología mixta que combine las fortalezas de cada uno de los métodos de imputación y mitigue los desafíos que presenta. Para esto, es importante tener en cuenta los siguientes elementos:

- El método MICE es altamente efectivo en entornos donde los datos cumplen con características de mecanismos de datos faltantes MAR [31]. Este modelo de imputación está diseñado en función a datos MAR, es decir, la probabilidad de que un dato esté ausente depende de otras variables observadas, pero no del valor faltante en sí mismo. Debido a que la ausencia de datos depende directamente de valores no observados, y debido a que MICE no modela el mecanismo de ausencia de estos datos, tiende a subestimar los valores faltantes, especialmente aquellos donde hay autocensura u omisión selectiva [23, 32].
- El método BNN es eficiente para modelar la incertidumbre en la imputación de datos, especialmente cuando se

utilizan herramientas como el Monte Carlo Dropout [14], logra capturar las relaciones no lineales de mejor manera que los otros métodos propuestos pero lamentablemente a la hora de trabajar en escalabilidad, el costo computacional es demasiado alto [17].

- Por su parte el modelo SOFT-IMPUTE, tiene un desempeño eficiente con grandes matrices de datos, pero asumen que los datos tienen una estructura de bajo rango, situación que no siempre es cierta [22]. Sin embargo, este método, al igual que MICE tiene buen desempeño con datos faltantes que cumplen características MAR. Soft-Impute no modela el mecanismo de ausencia, y por lo tanto ignora cualquier dependencia directa entre la falta de datos y los valores faltantes [16, 26].

Ante la posibilidad de que cada uno de los métodos analizados pueda cumplir por sí solo con todas las exigencias empíricas que se pueden presentar en una aplicación con datos reales, se propone un método híbrido que utiliza cada uno de los tres métodos previamente descritos para solucionar problemas específicos. Este método busca abordar diversos tipos de variables en los datos y cumplir con los principios del método de Alkire-Foster.

Este método híbrido aprovecha el trabajo realizado por el algoritmo en la identificación y clasificación de variables, obteniendo la siguiente diferenciación de variables:

- Continuas
- Categóricas
- Binarias

Para las variables continuas se aplicará el método Soft Impute [33] previamente descrito en este documento. En el caso en el que las variables sean categóricas se aplicará el método MICE [31] y finalmente si las variables fueran binarias se aplicará el método BNN [18], ambos previamente descritos en este documento.

Un elemento importante al momento de juzgar la pertinencia de un método de imputación u otro (o una combinación de varios), es lograr verificar el cumplimiento de los principios establecidos por el método Alkire-Foster [1,2] una vez concluido el proceso de imputación. Es importante mencionar que esta verificación no depende sólo del método de imputación aplicado, sino también de las características del conjunto de datos resultante.
En este sentido, se debe evaluar si la matriz de privaciones generada permite que el modelo mantenga validez lógica y analítica conforme a los axiomas fundamentales del enfoque AF:

- Monotonicidad dimensional
- Desagregación de subgrupos
- Descomposición dimensional

Cada uno de estos axiomas tiene implicancias concretas para que la medida imputada sea útil para guiar una acción en contra de la pobreza multidimensional. La monotonicidad exige que ningún individuo pobre reduzca su número de privaciones tras la imputación, asegurando que no se oculte pobreza de manera artificial. La desagregación permite comparar resultados entre distintos grupos poblacionales o territoriales, condición necesaria para el diseño de políticas focalizadas. Finalmente, la descomposición dimensional posibilita identificar qué dimensiones contribuyen más a la pobreza, lo cual es esencial para priorizar intervenciones.

El cumplimiento de estos tres principios representa una condición necesaria para asegurar la validez del método de conteo con doble corte una vez completado el proceso de imputación. Más allá de la completitud del conjunto de datos, lo que se busca es que las privaciones reflejadas por el modelo mantengan coherencia con la estructura lógica que sustenta el enfoque. Garantizar esta consistencia axiomática permite afirmar que los valores imputados no distorsionan las reglas del método, conservan la validez de las privaciones construidas y permiten una interpretación legítima de los resultados.

El Cuadro 1 resume la compatibilidad teórica de cada técnica de imputación con los axiomas fundamentales del enfoque Alkire-Foster. Esta evaluación comparativa constituye el insumo central que justifica la propuesta metodológica híbrida que se presenta a continuación.

Cuadro 1: Compatibilidad de métodos de imputación con axiomas fundamentales del enfoque Alkire-Foster

Método	Monotonicidad	Descomposición por subgrupos	Descomposición dimensional
Lista completa	No	No	No
MICE	Parcial	Parcial	Parcial
Soft-Impute	Parcial	Parcial	Parcial
BNN	Alta	Alta	Alta

Fuente: Evaluación basada en compatibilidad teórica con los axiomas del método Alkire-Foster. Vea [2, 7, 8, 15, 18, 22, 25].

5. Reflexiones finales

Este artículo ha abordado el problema de los datos faltantes en la medición de pobreza multidimensional desde una perspectiva metodológica crítica, técnica y normativa. A partir de una revisión exhaustiva de los mecanismos de ausencia de datos y de las principales estrategias de imputación, se ha planteado una propuesta metodológica que articula técnicas algorítmicas modernas con criterios de validación axiológica.

La imputación de datos no constituye únicamente una operación estadística o computacional. En el contexto del enfoque Alkire-Foster para la medición de pobreza multidimensional, imputar datos implica tomar decisiones que afectan la representación estructural de las privaciones, su conteo y la posterior identificación de los pobres. Por ello, esta práctica debe ser entendida como un acto epistémico con consecuencias éticas, políticas y analíticas.

El presente artículo sostiene que, en marcos axiomatizados para la medición de pobreza como el de Alkire-Foster, el tratamiento de datos faltantes no puede reducirse a criterios de precisión predictiva o simplicidad operativa. Las decisiones metodológicas deben alinearse con la lógica normativa del índice, particularmente con los axiomas de monotonicidad, descomposición por subgrupos y descomposición dimensional. En este sentido, el valor de los métodos de imputación no debe juzgarse únicamente por su capacidad de reconstruir la matriz original, sino por su potencial para preservar las propiedades fundamentales del modelo y su utilidad como herramienta de política pública.

Esta perspectiva ha orientado el diseño de una propuesta metodológica híbrida, que asigna técnicas imputativas según la naturaleza de las variables y complementa los procesos computacionales con una validación estructural ex post. Lejos de representar un eclecticismo técnico, este enfoque busca restituir al proceso de imputación su dimensión normativa, permitiendo que el dato reconstruido sea también un dato normativamente válido.

La evaluación de compatibilidad axiológica post-imputación (a través de pruebas estructurales sobre la matriz reconstruida) constituye, en este marco, no solo un paso accesorio, sino una fase crítica del análisis. La propuesta es, por tanto, epistemológicamente deliberada: imputar no solo reconstituye un conjunto de datos; también reconstituye una narrativa estructurada sobre la pobreza, y como tal, debe ser tratada con la misma responsabilidad que cualquier decisión de modelado.

El enfoque híbrido desarrollado permite asignar métodos de imputación específicos según la naturaleza de cada variable (categórica, ordinal, continua o binaria), y someter el resultado a una verificación estructural posterior, basada en los principios del enfoque Alkire-Foster.

La contribución principal del artículo reside en su reivindicación de una imputación normativamente informada. Al centrar el análisis en la compatibilidad entre el dato imputado y la lógica del índice, se plantea un cambio de paradigma: dejar de considerar la imputación como un simple preprocesamiento técnico para entenderla como un componente integral del diseño metodológico. Esta perspectiva obliga a repensar la relación entre técnicas computacionales y modelos normativos, y abre un espacio para la construcción de estrategias imputativas que no sólo reproduzcan la matriz original, sino que la reconstruyan con fidelidad normativa.

Como líneas futuras de investigación, se propone avanzar en la automatización de procesos de validación axiológica post-imputación, el desarrollo de marcos de imputación adaptativos basados en simulaciones y aprendizaje estructurado, y la integración de mecanismos participativos que permitan combinar conocimiento experto con métodos algorítmicos. La imputación, desde esta perspectiva, deja de ser una operación esta-

dística subordinada para convertirse en una dimensión fundamental del proceso de construcción del dato, con implicaciones directas sobre el modo en que se mide, representa y gobierna la pobreza.

Referencias

[1] S. Alkire y J. Foster, "Counting and multidimensional poverty measurement," *J. Public Econ*., vol. 95, no. 7-8, pp. 476–487, 2011, doi: 10.1016/j.jpubeco.2010.11.006.
[2] S. Alkire, J. M. Roche, P. Ballon, J. Foster, M. E. Santos, y S. Seth, Multidimensional Poverty Measurement and Analysis. Oxford, UK: Oxford Univ. Press, 2015. [En línea]. Disponible: https://ophi.org.uk/multidimensional-poverty-measurement-and-analysis/
[3] S. Alkire y M. E. Santos, "Measuring acute poverty in the developing world: Robustness and scope of the multidimensional poverty index," *World Dev*., vol. 59, pp. 251–274, 2014, doi: 10.1016/j.worlddev.2014.01.026.
[4] P. D. Allison, Missing Data. Thousand Oaks, CA, USA: Sage Publications, 2001. [En línea]. Disponible: https://us.sagepub.com/en-us/nam/missing-data/book9438
[5] M. J. Azur, E. A. Stuart, C. Frangakis, y P. J. Leaf, "Multiple imputation by chained equations: What is it and how does it work?" *Int. J. Methods Psychiatr. Res*., vol. 20, no. 1, pp. 40–49, 2011, doi: 10.1002/mpr.329.
[6] J. Bound, C. Brown, y N. Mathiowetz, "Measurement error in survey data," en Handbook of Econometrics, vol. 5, J. J. Heckman y E. Leamer, Eds. Amsterdam, Netherlands: Elsevier, 2001, pp. 3705–3843, doi: 10.1016/S1573-4412(01)05012-7.
[7] S. van Buuren, Flexible Imputation of Missing Data. Boca Raton, FL, USA: CRC Press, 2018. [En línea]. Disponible: https://stefvanbuuren.name/fimd/
[8] E. Chen, M. S. Andersen, y R. Chandra, "Deep learning framework with Bayesian data imputation for modelling and forecasting groundwater levels," *Environ. Model. Softw.,* vol. 178, Art. no. 106072, 2024, doi: 10.1016/j.envsoft.2024.106072.
[9] J.-Y. Duclos, D. E. Sahn, y S. D. Younger, "Robust multidimensional poverty comparisons," *Econ. J*., vol. 116, no. 514, pp. 943–968, 2006, doi: 10.1111/j.1468-0297.2006.01118.x.
[10] I. Dutta, R. Nogales, y G. Yalonetzky, "Endogenous weights and multidimensional poverty: A cautionary tale," *J. Dev. Econ*., vol. 151, Art. no. 102649, 2021, doi: 10.1016/j.jdeveco.2021.102649.
[11] F. H. G. Ferreira y M. A. Lugo, "Multidimensional poverty analysis: Looking for a middle ground," World Bank Res. Obs., vol. 28, no. 2, pp. 220–235, 2013, doi: 10.1093/wbro/lks013.
[12] D. Filmer y L. H. Pritchett, "Estimating wealth effects without expenditure data—or tears: An application to educational enrollments in states of India," *Demography*, vol. 38, no. 1, pp. 115–132, 2001, doi: 10.1353/dem.2001.0003.
[13] J. Foster, J. Greer, y E. Thorbecke, "A class of decomposable poverty measures," *Econometrica*, vol. 52, no. 3, pp. 761–766, 1984, doi: 10.2307/1913475.
[14] Y. Gal y Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," en Proc. 33rd Int. Conf. Mach. Learn. (ICML), New York, NY, USA, 2016, pp. 1050–1059. [En línea]. Disponible: http://proceedings.mlr.press/v48/gal16.html
[15] L. Gondara y K. Wang, "MIDA: Multiple imputation using denoising autoencoders," en Advances in Knowledge Discovery and Data Mining (PAKDD), Melbourne, VIC, Australia, 2018, pp. 260–272, doi: 10.1007/978-3-319-93040-4_21.
[16] T. Hastie, R. Mazumder, J. D. Lee, y R. Zadeh, "Matrix completion and low-rank SVD via fast alternating least squares*," J. Mach. Learn. Res*., vol. 16, no. 1, pp. 3367–3402, 2015. [En línea]. Disponible: https://jmlr.org/papers/v16/hastie15a.html
[17] J. M. Hernández-Lobato y R. Adams, "Probabilistic backpropagation for scalable learning of Bayesian neural networks," en Proc. 32nd Int. Conf. Mach. Learn. (ICML), Lille, France, 2015, pp. 1861–1869. [En línea]. Disponible: http://proceedings.mlr.press/v37/hernandez-lobato15.html
[18] E. R. Hruschka, E. R. Hruschka, y N. F. F. Ebecken, "Bayesian networks for imputation in classification problems," *J. Intell. Inf. Syst.,* vol. 29, no. 3, pp. 231–252, 2007, doi: 10.1007/s10844-006-0004-5.
[19] J. Krishnakumar y P. Ballon, "Estimating basic capabilities: A structural equation model applied to Bolivia," *World Dev*., vol. 36, no. 6, pp. 992–1010, 2008, doi: 10.1016/j.worlddev.2007.10.008.
[20] R. J. A. Little y D. B. Rubin, Statistical Analysis with Missing Data, 3ra ed. Hoboken, NJ, USA: Wiley, 2019, doi: 10.1002/9781119482260.
[21] E. Maasoumi, "The measurement and decomposition of multi-dimensional inequality," *Econometrica*, vol. 54, no. 4, pp. 991–997, 1986, doi: 10.2307/1912845.
[22] R. Mazumder, T. Hastie, y R. Tibshirani, "Spectral regularization algorithms for learning large incomplete matrices," *J. Mach. Learn. Res.,* vol. 11, pp. 2287–2322, 2010. [En línea]. Disponible: https://www.jmlr.org/papers/v11/mazumder10a.html
[23] R. C. Pereira, P. H. Abreu, P. P. Rodrigues, y M. A. T. Figueiredo, "Imputation of data missing not at random: Artificial generation and benchmark analysis," *Expert Syst. Appl*., vol. 249, Art. no. 123654, 2024, doi: 10.1016/j.eswa.2024.123654.
[24] D. B. Rubin, "Inference and missing data," Biometrika, vol. 63, no. 3, pp. 581–592, 1976, doi: 10.1093/biomet/63.3.581.
[25] J. L. Schafer y J. W. Graham, "Missing data: Our view of the state of the art," *Psychol. Methods*, vol. 7, no. 2, pp. 147–

177, 2002, doi: 10.1037/1082-989X.7.2.147.
[26] A. Sportisse, C. Boyer, y J. Josse, "Imputation and low-rank estimation with missing not at random data," *Stat. Comput*., vol. 30, no. 6, pp. 1629–1643, 2020, doi: 10.1007/s11222-020-09963-5.
[27] K. Tsui, "Multidimensional poverty indices," *Soc. Choice Welf*., vol. 19, no. 1, pp. 69–93, 2002, doi: 10.1007/s003550100155.
[28] S. van Buuren y K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in R," *J. Stat. Softw.,* vol. 45, no. 3, pp. 1–67, 2011, doi: 10.18637/jss.v045.i03.
[29] G. Vink, L. E. Frank, J. Pannekoek, y S. van Buuren, "Predictive mean matching imputation of semicontinuous variables," Stat. Neerl., vol. 68, no. 1, pp. 61–90, 2014, doi: 10.1111/stan.12023.
[30] S. Vyas y L. Kumaranayake, "Constructing socio-economic status indices: How to use principal components analysis," *Health Policy Plan*., vol. 21, no. 6, pp. 459–468, 2006, doi: 10.1093/heapol/czl029.
[31] J. N. Wulff y L. Ejlskov, "Multiple imputation by chained equations in praxis: Guidelines and review," *Electron. J. Bus. Res*. Methods, vol. 15, no. 1, pp. 41–56, 2017. [En línea]. Disponible: https://academic-publishing.org/index.php/ejbrm/article/view/1344
[32] J. Xiao y O. Bulut, "Evaluating the performances of missing data handling methods in ability estimation from sparse data," *Educ. Psychol. Meas.,* vol. 80, no. 5, pp. 932–954, 2020, doi: 10.1177/0013164420911136.
[33] Q. Yao y J. T. Kwok, "Accelerated and inexact soft-impute for large-scale matrix and tensor completion," *IEEE Trans. Knowl. Data Eng*., vol. 31, no. 9, pp. 1665–1679, 2018, doi: 10.1109/TKDE.2018.2868112.