Imputación de valores faltantes de variables de razón y proporción mediante modelos de regresión Beta cero-uno inflados

Adely Margarita Parada
Director: Mario José Pacheco López

Tesis de grado presentada como requisito parcial
para optar al título de Estadística

10 de agosto de 2023

Resumen

En los datos recogidos en las últimas décadas es habitual encontrar la presencia de valores faltantes por múltiples razones que pueden llevar a los investigadores a usar estrategias incorrectas para tratarlos teniendo como resultado inferencias incorrectas. Este problema es mucho más común en variables de razón y proporción dado que la mayoría de métodos de imputación múltiple no consideran la naturaleza de este tipo de variables. En la literatura existen modelos para este tipo de variables como los modelos de regresión beta que se pueden emplear para construir modelos de imputación múltiple. Dado esta necesidad en el presente trabajo se propone presentar el desarrollo de una función en el software R que utilice métodos de imputación múltiple con modelos de regresión Beta cero-uno inflados que permitan la implementación correcta de imputación para variables de proporción y razón.

1. Introducción

El análisis de datos faltantes se emplea con el fin de obtener inferencias a partir de un conjunto de datos que presenta variables con un determinado porcentaje de información ausente. De acuerdo a Arteaga (2009) esta falta de información puede tener varias razones,

como errores a la hora de diligenciar, abandono del estudio, datos no disponibles, entre otros.

Para el manejo de manera eficiente de información faltante, se han desarrollado varios métodos estadísticos que permitan abordar de manera adecuada su presencia en los datos. Algunos de los métodos más comunes, son imputación por la media, imputación por vecinos cercanos, imputación por regresión, o imputación múltiple. También se encuentra la imputación por el método MICE (Multiple Imputation by Chained Equations) que tiene como enfoque el generar múltiples imputaciones por módelos condicionales específicos para cada variable con datos faltantes.van Buuren (2011)

Como señala Johnson (2016), los modelos Beta se usan para modelar variables restringidas generalmente entre cero y uno, usada en procesos de entrada en simulación estocástica, cálculos de costos esperados en proyectos de ingeniería civil o industrial, entre otras. También se presenta el modelo Beta inflado que se propone para aquellas variables dependientes en el intervalo [0, 1] con probabilidad positiva en los valores extremos 0 y 1, que permiten estimaciones más precisas y un mejor ajuste para este tipo de variables. Un ejemplo de uso de esta variante de la distribución beta, es la aplicación propuesta por Burch (2019), en el que se busca analizar la relación entre la importancia o relevancia de una palabra y su dispersión.

En este trabajo se propone su uso en el Índice de Mujeres, Paz y Justicia (WPS) para aplicar la función de imputación propuesta en las variables de razón y proporción, centrandose en la comprensión y evaluación de 27 países para ampliar el alcance y la comprensión del WPS más allá de los países que cuentan con toda la información disponible.

En un mundo cada vez más consciente de la importancia de la igualdad de género y el desarrollo de una paz sostenible, el Índice de Mujeres, Paz y Justicia (WPS por sus siglas en inglés) se ha convertido en una herramienta que ha permitido comprender y abordar las disparidades a las que se enfrentan las mujeres alrededor del mundo. El índice fue desarrollado por el Instituto Georgetown para las Mujeres, la Paz y la Seguridad y el Instituto de la Paz de Estocolmo (SIPRI) y se basa en la Resolución 1325 del Consejo de Seguridad de las Naciones Unidas, ofreciendo una visión más amplia y rigurosa del estado de las mujeres.

En la construcción del WPS el GIWPS and PRIO (2021) considera varios aspectos, como la participación política de las mujeres, inclusión ecónomica, acceso a educación, seguridad y protección de los derechos humanos de las mujeres, entre otros, construyendo así un marco que permite evaluar las políticas y medidas implementadas para promover la igualdad de género y la participación de las mujeres en la paz y la justicia.

2. Imputación de datos faltantes

De acuerdo a Little et al. (2012), los datos faltantes se definen como valores no disponibles pero que serían significativos para el análisis si se observaran. Por tal razón para su tratamiento es fundamental entender el por qué su ausencia y sus implicaciones entendiendo el mecanismo de datos ausentes, que permite explicar la relación de estos con los datos observados. Xiao-Hua (2019).

2.1. Mecanismo de datos faltantes

Como indica Arteaga (2009) el mecanismo de datos faltantes puede clasificarse en tres categorías:

- Datos faltantes completamente aleatorios (MCAR): no existe relación entre los valores de las variables (observadas y faltantes) y la probabilidad de que falten. Los elementos que faltan son simplemente una muestra aleatoria de los datos observados.
- Datos faltantes aleatorios (MAR): la falta depende solo de los datos observados de en la variable y no de los valores que faltan.
- Datos faltantes no aleatorios (NMAR): la probabilidad de que falte un elemento depende del valor no observado de los elementos faltantes

2.2. Tratamiento de datos faltantes

Dentro del tratamiento de datos faltantes se pueden encontrar múltiples tratamientos, a continuación se nombran algunos de ellos Xiao-Hua (2019):

- Casos completos: también conocido como eliminación total, excluye todas las observaciones que cuentan con datos faltantes en alguna de las variables de interés. A pesar de ser uno de los métodos más usados presenta una gran desventaja y es la pérdida de precisión debido a la reducción del tamaño de la muestra, llevando a posibles sesgos cuando las observaciones omitidas difieren de los casos completos, no cumpliendo la hipótesis MCAR.
- Imputación de media incondicional: este método de imputación reemplaza los valores faltantes de una variable por la media de los valores observados de esa variable, sin embargo este método subestima la variabilidad.

- Imputación de media condicional: aquí se utilizan modelos de regresión con datos completos, sin embargo este método también subestima la variabilidad.
- Casos disponibles: se reduce la pérdida de información para cada variable realizando una estimación con los valores observados por medio de subconjuntos.
- Imputación por Regresión Estocástica: se emplean modelos de regresión para estimar los valores faltantes de cada variable a partir de los datos observados de las variables restantes para usarla como explicativas.
- Imputación Hot-Deck: utiliza los valores observados de individuos con características similares para la imputación.
- Imputación Múltiple: realiza un proceso iterativa para obtener subconjuntos de datos completos, obteniendo los resultados que se usan para la estimación final.

2.3. Algoritmo MICE

Otro de los métodos utilizados es la imputación multivariada por medio de ecuaciones encadenadas (MICE, por sus siglas en inglés), llamada también 'especifición totalmente condicional' o 'imputación múltiple secuencial de regresión', que crea multiples predicciones para cada valor perdido, sin embargo si la calidad de los valores observados para el valor faltante es baja, se tendrá como resultado una imputación de baja calidad. Azur (2011)

Los pasos generales para la imputación son:

- 1. Se crea un valor temporal, que corresponde a la media de los datos observados para cada $y_i j$.
- 2. En la variable incompleta $Y_i j$, se estima un modelo de regresión con los datos observados, donde $Y j$ es la variable de respuesta y las demás son usadas como explicativas y estos 'nuevos' valores reemplazan los valores generados en el paso anterior.
- 3. Los datos generados en los pasos anteriores, son repetidos para cada variable que tiene valores faltantes. El ciclo en cada variable consiste en una iteración o ciclo. Al final de cada ciclo todos los valores faltantes son reemplazados con predicciones que reflejan las relaciones observadas en los datos.
- 4. Los pasos 2-4 se repiten durante varios ciclos, actualizándose las imputaciones en cada ciclo.

2.4. Reglas de Rubin

Sea θ el parámetro de interés, sea $\hat{\theta}$ el estimador del parámetro de interes, con varianza estimada $\hat{v}\hat{\theta}$, y sea $\hat{\theta}^{(1)}$, $\hat{\theta}^{(2)}$,...,$\hat{\theta}^{(m)}$, $\hat{v}\hat{\theta}^1$, $\hat{v}\hat{\theta}^2$, ...,$\hat{v}\hat{\theta}^m$ las estimaciones y sus varianzas a partir de cada una de las m imputaciones, entonces de acuerdo a las reglas de Rubin ver van Buuren (2018), se puede obtener la estimación conjunta de θ y su varianza como:

$$\bar{\theta} = \frac{1}{m}\sum_{j=1}^{m}\hat{\theta}^{(j)}$$

$$\hat{v}(\bar{\theta}) = \frac{1}{m}\sum_{j=1}^{m}\hat{v}(\hat{\theta}^{(j)}) + (1+\frac{1}{m})B$$

donde

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

Como medidas de severidad producto del proceso de imputación se puede calcular:

- Proporción de la varianza debida a los datos faltantes:

$$\lambda = \frac{(1+\frac{1}{m})B}{\hat{v}(\bar{\theta})}$$

- Incremento relativo de la varianza:

$$r = \frac{(1+\frac{1}{m})B}{\bar{v}(\theta)}$$

- Fracción de información faltante sobre WPS:

$$\gamma = (1+r)^{-1}\left[r + \frac{2}{\nu+3}\right]$$

siendo $\nu = \frac{n(n-1)}{(n+2)}(1-\lambda)$.

3. Modelos de regresión Beta

El modelo de regresión se apoya en el supuesto de que la variable respuesta (y) tiene una distribución beta restringida al intervalo (0,1). Esta distribución proporciona la flexibilidad para modelar proporciones dada la versatilidad de su función de densidad, dado que puede tomar diferentes formas dependiendo de los valores de los parámetros α y β. Esta función de densidad está dada por:

$$\pi(y;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

Donde $\alpha > 0$, $\beta > 0$ y $\Gamma(\cdot)$ se refiere a la función gamma.

Ferrari (2013) aplica una regresión a la variable de interés a partir de un conjunto de variables exógenas, por este motivo se propone la aplicación de un modelo de regresión beta que redefine la función de densidad de la siguiente manera, buscando modelar la media e incluyendo un parámetro ϕ como parámetro de precisión:

$$\pi(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, 0 < y < 1$$

Donde $0 < \mu < 1$ y $\phi > 0$. Entre mayor valor tenga ϕ, el valor de la varianza para y será más pequeño.

Después de obtener la función de densidad, en la que se asume que $y_{\mathrm{i}}, ..., y_{\mathrm{n}}$ es una variable independiente, con media μ_{t} y ϕ desconocido, se llega a:

$$E(y) = \mu$$

$$Var(y) = \frac{\mu(1-\mu)}{\phi+1}$$

y se asume que el modelo se obtiene con la hipótesis que la media de y_{t} puede ser escrito:

$$g(\mu_{\mathrm{t}}) = \sum_{i=1}^{k} x_{\mathrm{ti}}\beta_{\mathrm{i}} = \eta_{\mathrm{t}}$$

donde $\beta = (\beta_1, ..., \beta_{\mathrm{k}})^T$ es un vector de parámetros desconocidos y $x_{\mathrm{t1}}, ..., x_{\mathrm{tk}}$ son observaciones en k covariables ($k < n$), que son fijas y conocidas. Y $g(\cdot)$ es una función de enlace estrictamente monótona y dos veces diferenciable que mapea (0, 1).

Para la función de enlace $g(\cdot)$ se toma la función logit de la siguiente manera:

$$g(\mu) = log\left(\frac{\mu}{1-\mu}\right)$$

Cuando las variables beta incluyen ceros y uno, la distribución beta usual no brinda una descripción adecuada de los datos. En estos casos, Ospina (2012), propone el manejo de modelos cero o uno inflados con un poco más de flexibilidad cuando la variable se encuentra entre cero y uno:

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, y \in (0,1)$$

Donde $\Gamma(.)$ es la función gamma. Si y $\mathbf{B}(\mu,\phi)$, entonces E(y)= μ y Var(y) = $\frac{\mu(1-\mu)}{\phi+1}$. Por lo tanto, μ es la media de la distribución y ϕ es el parámetro de precisión.

4. Datos

Los datos contienen la información del Índice de Mujeres, Paz y Seguridad 2021/2022 en su tercera edición para 197 países. Este indicador fue creado entre el Instituto Georgetown para Mujeres, Paz y Seguridad (GIWPS, por sus siglas en inglés) y el Instituto de Investigación de la Paz de Oslo (PRIO, por sus siglas en inglés).

GIWPS and PRIO (2021) construyen este índicador como una medida integral que construida a partir de tres grandes dimensiones:

- Inclusión: Económica, Social y Política
- Justicia: Leyes formales y discriminación informal
- Seguridad: A nivel familiar, comunitario y social.

La información recogida proviene de fuentes oficiales, como departamentos nacionales de estadística, organizaciones pertenecientes a la ONU, Peace Research Institute Oslo, entre otras. Los datos están fundamentados en la información de la población o muestras representativas, omitiendo de esta manera el juicio de expertos para calificar el desempeño.

Las dimensiones nombradas con anterioridad se construyen a partir de otros indicadores, que se detallan a continuación:

- Inclusión: se mide por los logros de las mujeres en educación, empleo, representación parlamentaria, y acceso a telefónos celulares y servicios financieros.
    - Educación: promedio de años de educación de las mujeres de 25 años o más. Fuente: Instituto de Estadística de la UNESCO
    - Empleo: porcentaje de mujeres de 25 años o más que están empleadas. Fuente: base de datos ILOSTAT
    - Uso del teléfono celular: porcentaje de mujeres de 15 años o más que informan tener un teléfono móvil que usan para hacer y recibir llamadas. Fuente: Encuesta mundial Gallup 2018
    - Inclusión financiera: porcentaje de mujeres de 15 años o más que informan tener una cuenta individual o conjunta en un banco o institución financiera y con uso en el último año. Fuente: Base de datos Global Findex del Banco Mundial.
    - Parlamento: porcentaje de escaños ocupados por mujeres en las cámaras baja y alta del parlamento nacional. Fuente: Unión Interparlamentaria
- Justicia: mide el alcance de la discriminación contra la mujer en el sistema legal, al igual que el sesgo a favor de los hijos varones y la exposición a normas discriminatorias, especificamente contra las mujeres en oportunidades económicas y trabajo remunerado.
    - Normas discriminatorias: porcentaje de hombres de 15 años o más que no estuvierón con la proposición: “Es perfectamente aceptable que cualquier mujer de su familia tenga un trabajo remunerado fuera del hogar si lo desea”. Fuente: Gallup y Organización Internacional del Trabajo
    - Sesgo de hijo: proporción de sexos al nacer (relación entre el número de niños nacidos y el número de niñas nacidas) supera la tasa demográfica natural de 1.05. Fuente: Departamento de Asuntos Sociales y Económicos de las Naciones Unidas

    - Discriminación Jurídica: puntaje agregado de las leyes y regulaciones que limitan la capacidad de las mujeres para participar en la sociedad y la economía o que diferencian entre hombres y mujeres. Fuente: Banco Mundial, Women Business, and the Law

- Seguridad: se mide a nivel familiar, comunitario y social, incluida la violencia de pareja durante toda la vida.

    - Violencia organizada: número total de muertes en batalla por conflictos estatales, no estatales y unilaterales por cada 100,000. Fuente: Programa de datos sobre conflictos de Uppsala
    - Percepción de la seguridad de la comunidad: porcentaje de mujeres de 15 años o más que informan que "se sienten seguras caminando solas de noche en la ciudad o el área donde vive".Fuente: Encuesta mundial Gallup 2019
    - Violencia de pareja: porcentaje de mujeres que sufrieron violencia física o sexual cometida por su pareja íntima en los 12 meses anteriores. Fuente: Base de datos mundial de ONU Mujeres sobre la violencia contra la mujer.

Después de obtener la información anteriormente nombrada se realiza una normalización para que la información sea comparable, reescalando los valores originales de 0 a 1 (0 a 100), siendo 0 el peor desempeño y 1 o 100 el mejor desempeño. Para aquellos indicadores que no se encuentran en porcentajes como educación y representación parlamentaria se establecen valores máximos aspiracionales, como 15 años para la media de años de escolaridad y 50 % para la representación parlamentaria.

$$Normalizacion = \frac{Valoractual - Limiteinferior}{Limitesuperior - Limiteinferior}$$

Después de tener la normalización se asigna el mismo peso a cada una de las dimensiones anteriormente nombradas, quedando de la siguiente manera:

$$Inclusion = \frac{(Educacion + I.financiera + Empleo + Celular + R.parlamentaria)}{5}$$

$$Justicia = \frac{(D.Legal + Sesgohijo + NormasDiscriminatorias)}{3}$$

$$Seguridad = \frac{(Violenciaorganizada + Percepcionseguridad + ViolenciaPareja)}{3}$$

Teniendo estas tres dimensiones, se calcula una media geométrica, que permite darle el miso peso a cada dimensión:

$$WPS = (X_1 * X_2 * X_3)^{\frac{1}{3}}$$

Siendo $X_1 = Inclusión, \quad X_2 = Justicia \quad y \quad X_3 = Seguridad$

Una aproximación de la varianza del WPS se puede calcular como:

$$\hat{V}(W\bar{P}S) = (W\bar{P}S)^2 \frac{1}{n^2} \sum_{i=1}^{n} \ln\left(\frac{X_i}{W\bar{P}S}\right)^2$$

Para acceder a los datos se puede realizar desde el siguiente enlace: https://giwps.georgetown.edu/wp-content/uploads/2021/10/WPS-Index-2021-Data.csv.

5. Metodología

La imputación por regresión con la función MICE en un primer momento genera una media muestral a partir de los valores observados para los datos ausentes dentro de cada variable. Después de este primer paso en la variable incompleta Y_ij, se estima un modelo de regresión con los datos observados, donde Yj es la variable de respuesta y las demás son usadas como explicativas y estos nuevos valores reemplazan los valores generados en el paso anterior en un proceso iterativo que es modificable con el argumento maxit en la función. van Buuren (2011)

Utilizando el software R, se propone una función que permita la imputación de variables con datos faltantes de razón y proporción mediante módelos de regresión Beta cero-uno inflados con la siguiente metodología:

La función cuenta con los siguientes argumentos que el usuario debe indicar: *dataset*, base de datos con las variables a imputar; *beta_var*, vector con la posición de las columnas con distribución beta; *formulas_beta*, vector con las fórmulas propuestas por el usuario para el modelo de imputación, puede ser NULL; *formulas_x*, vector con las fórmulas propuestas por el usuario para imputar las variables que no tienen distribución beta, puede ser NULL; *maxit*, número de iteraciones, por default es 5; *m*, número de bases a imputar, por default es 5; *seed*, número entero que se usa como semilla para que los resultados seán reproducibles, puede ser NULL.

Después de contar con la información descrita, se inicia el proceso de imputación iniciando con la carga de los paquetes que se usarán: *dplyr*, *mice* y *gamlss*. A continuación se realiza una selección de las variables y que son aquellas que el usuario señaló como las variables con distribución beta y las variables x, aquellas que no tienen distribución beta. Al terminar esta selección se verifican las fórmulas indicadas por el usuario para los modelos, en caso de que los argumentos sean NULL se construiran a partir de las variables x y y. Para el argumento *seed* también se hace una verificación, en caso que sea NULL se generara de manera aleatoria y se informara al usuario al finalizar el proceso.

Después de las validaciones anteriores se da inicio a la imputación del número de bases indicadas en el argumento m realizando una imputación inicial para toda la base de datos con la función *MICE* utilizando el método sample, que selecciona de forma aleatoria valores observados de la variable y los usa para la imputación de los valores faltantes. Con la base obtenida en el paso anterior, se inicia un proceso iterativo en el que para las variables no beta se utiliza la función *MICE* con el método pmm, que utiliza una coincidencia de la media predictiva, es decir para cada valor faltante, el método forma un pequeño conjunto de donantes de los datos observados que tienen valores previstos más cercanos al valor previsto para la entrada que falta. van Buuren (2018).

Por último, se realiza la imputación de las variables beta con los resultados obtenidos

en el paso previo. Para la imputación de estas variables se toma en cuenta el valor mínimo y máximo de la variable ya que estos valores determinaran la familia de la distribución a usar en el modelo, es decir cuando la variable esten en el rango (0, 1) se usara la distribución beta, cuando este entre [0, 1) se usara la distribución beta inflada 0 (BEINF0), cuando este entre (0, 1] se usara la distribución beta inflada 1 (BEINF1) y por último para aquellas variables entre [0, 1] se usara la distribución beta inflada (BEINF). Todas estas familias disponibles en la función gamlss, función con la que se calcula el modelo. Con los resultados obtenidos, se predicen los valores ausentes en la variable y estos resultados se almacenan junto con los imputados con la función *MICE* en una nueva base de datos completa que será la que obtenga el usuario.

Al obtener la función se realiza el análisis del conjunto de datos del Índice de Mujeres, Paz y Seguridad 2021/2022, en el que se aplicara la función propuesta con el fin de imputar la información de 27 países que presentan datos faltantes por distintas condiciones y que conlleva a no poder calcular el índice.

Para iniciar con el análisis se descarga la información del índice y del iso3, este último permite abreviar los nombres de la base original, para la cual antes de realizar la unión con el iso3 se hace una limpieza de datos a nivel general, es decir, nombres de columnas, nombres de países y exclusión de columnas que no se usaran en el análisis. Otro de los cambios que se realizan en la base de datos original, es dejar las columnas con distribución beta en el rango [0, 1] de acuerdo a la información que se encuentre en cada una. Este reescalamiento se hace dividiendo la variable entre 100. Las variables a las que se aplicó este proceso fueron: inclusión financiera, empleo, uso del celular, parlamento, discriminación jurídica, normas discriminatorias, violencia de pareja, percepción de seguridad y violencia organizada.

Después de tener esta base se usa la función propuesta para la imputación de los datos faltantes, pidiendo a la función 15 bases imputadas con 50 iteraciones con el fin de conseguir convergencia en la imputación obtenida.

Al obtener los resultados, se construye un PCA para cada una de las 15 bases imputadas con el fin de visualizar la información de las variables usadas en la construcción del índice y su relación. A partir de los PCA de cada base, se construye un PCA Pool, que se construye a partir de un promedio de los resultados para cada base imputada. Para obtener esta información se utiliza la función PCA del paquete FactoMineR, y para la visualización se usa la función fviz_pca_var del paquete factoextra.

Adicional al ACP, con las bases imputadas se construye el Índice WPS para cada una de las 15 bases, utilizando la misma metodología usada por sus autores, es decir, se realiza la normalización de cada variable con el límite inferior y el límite superior. Teniendo esto, se construyen las tres dimensiones, Inclusión, Justicia y Seguridad que dan como resultado por medio de una media geométrica, el WPS para cada base. Después de obtener esta

información para cada base se construye el WPS Pool, que es un promedio simple de los 15 índices construidos.

Adicional a las funciones y/o paquetes nombrados con anterioridad, también se hizo el uso del paquete ggplot2 para la visualización, janitor para la limpieza de datos y funciones como select, mutate, left_join con el fin de transformar la base de datos para el análisis.

6. Resultados

Patrones de datos faltantes

En la Figura 1 en la parte izquierda del gráfico se encuentra el número de patrones de datos faltantes en la base de datos, que en este caso es 17. En la parte derecha, se puede visualizar el número de variables con datos faltantes en cada patrón, mientras que en la parte inferior se ven el número de observaciones que tienen valores faltantes en la variable señalada en la parte superior del gráfico.

Una de las variables que más faltantes tiene es la de inclusión financiera, con 27 observaciones sin esta información, seguido por percepción de seguridad, uso del celular, normas de discriminación con 26, seguidas por educación.

La siguiente tabla muestra un resumen del Promedio, Desviación Estándar y el % de datos faltantes en cada variable.

Cuadro 1: Datos faltantes, Promedio y Desviación de la base de datos original

Variable	% Datos Faltantes	Promedio	Desviación Estandár
Educación	9.64	8.50	3.42
Inclusión financiera	13.71	0.56	0.28
Empleo	7.61	0.51	0.17
Uso de celular	13.20	0.87	0.17
Parlamento	1.02	0.24	0.12
Discriminación Jurídica	3.55	0.76	0.17
Sesgo de hijo	5.58	1.05	0.02
Normas discriminatorias	13.20	0.16	0.13
Violencia de pareja	8.12	0.13	0.09
Percepción de seguridad	13.20	0.57	0.17
Violencia organizada	0.00	0.01	0.08

Las variables con mayor porcentaje de valores faltantes son Inclusión financiera, uso del celular y normas discriminatorias, todas con más del 13 %, seguidas por educación. La única variable con información completa es la de Violencia organizada seguida por

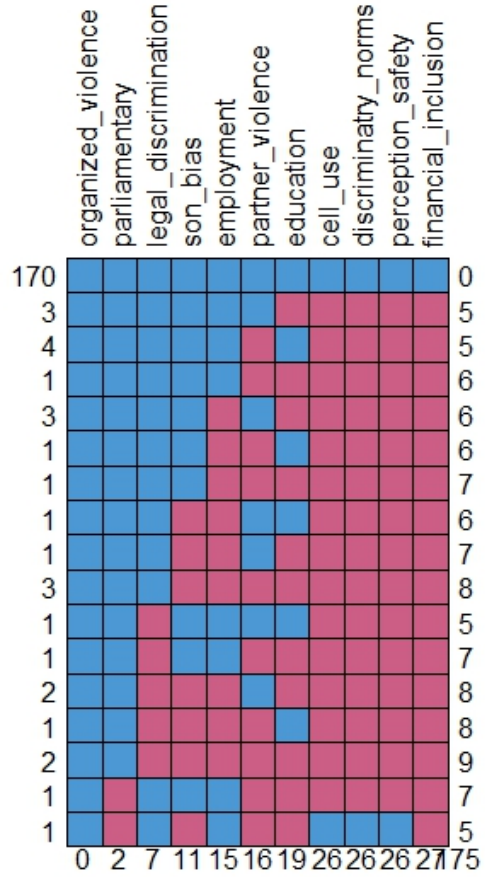

Figura 1: Patrones de datos faltantes.

Parlamento con el 1 % de ausencia de información. El promedio educativo para mujeres mayores de 25 años es de tan solo 8 años, lo que muestra que las mujeres nisiquiera alcanzan a tener un nivel básico educativo en educación bachiller, llevando esto a menores oportunidades laborales de calidad. La discriminación jurídica también muestra un promedio alto para los países que tienen esta información, ya que el promedio de es de 78 % mostrando esto el alto porcentaje de leyes que limitan la capacidad de las mujeres para ser participes de la sociedad y la economía.

En la figura 2, se muestra el ranking de los 27 países de interes de acuerdo al Índice de Mujeres, Paz y Justicia, en los que se ve como Taiwan tiene unos buenos resultados dejandolo en el puesto 17, estos resultados pueden estar apoyados por el trabajo que ha realizado el país para contribuir a la igualdad de género y asi mismo ofrecer más oportunidades de desarrollo a las mujeres.Gender® (2007)

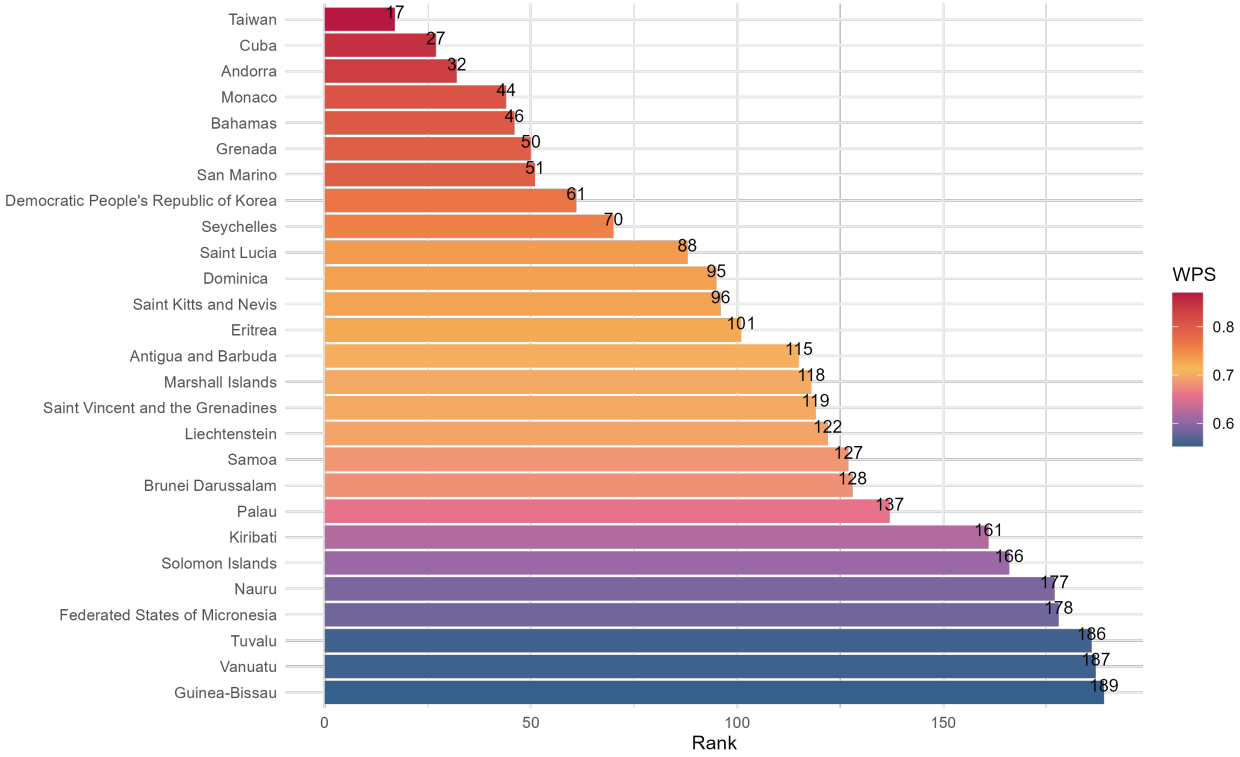

Figura 2: Ranking países de interés.

Relación del WPS con otros índices

En la figura 3 a pesar de no ser una relación lineal es posible ver que aquellos países con mejor WPS tienen menos posibilidades de ser afectados de forma significativa en las pandemias. GIWPS and PRIO (2021)

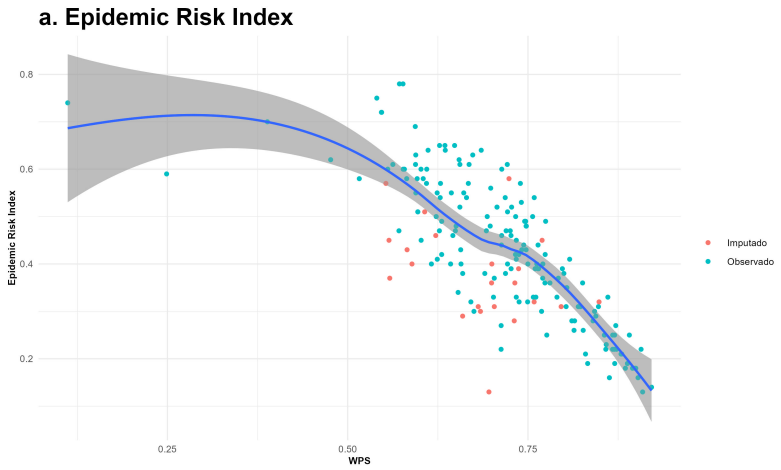

Figura 3: Relación del Índice de Epidemia

Otros indicadores que aunque no presentan una relación lineal presentan un buen ajuste respecto al WPS son el Índice de Estados Frágiles que de acuerdo a for Peace (2017) determina el nivel de vulnerabilidad y estabilidad de un país, que muestra que aquellos países que tienen mejores resultados en el WPS tienen mejores resultados en este indicador. El Índice Desempeño Ambiental (EPI por sus siglas en inglés) diseñado por Wolf et al. (2022) proporciona una medida que recoje la información de varios indicadores a nivel ambiental , también muestra un buen ajuste al WPS ya que entre un mejor resultado en el WPS mejores resultados se obtienen también en el EPI. El mismo comportamiento se ve en el Índice Global de Adaptación al Cambio Climático (GAIN por sus siglas en inglés), desarrollado por University of Notre Dame from Washington que evalúa y clasifica el nivel de adaptación de los países al cambio climatico.

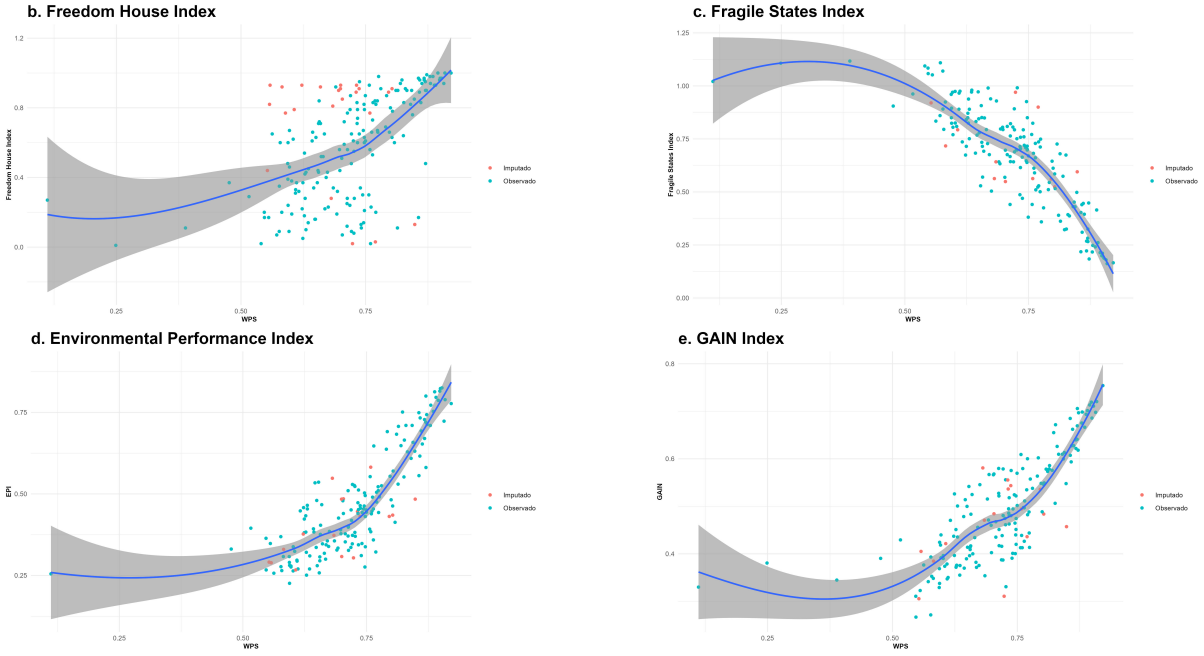

Figura 4: Indicadores relacionados con el WPS

En general, en los países imputados en la figura 4 no se encuentra un patrón en sus resultados, al contrario se encuentran dispersos al relacionarlos con otros indicadores.

Como resultado de los ACP, en la figura 5 se obtiene que las variables mejor representadas son las normas de discriminación, educación, inclusión financiera, discriminación legal seguidas por violencia de pareja y empleo. La variable con menos representación en la dimensión 1 y 2 es violencia organizada. Las dimensiones 1 y 2 representan el 48 % de la varianza total. Llama la atención que las variables normas discriminatorias y discriminación juridica a pesar de estar en la misma dimensión, Justicia presentan una correlación negativa, lo mismo para las variables violencia de pareja y percepción de la seguridad pertenecientes a la dimensión de seguridad.

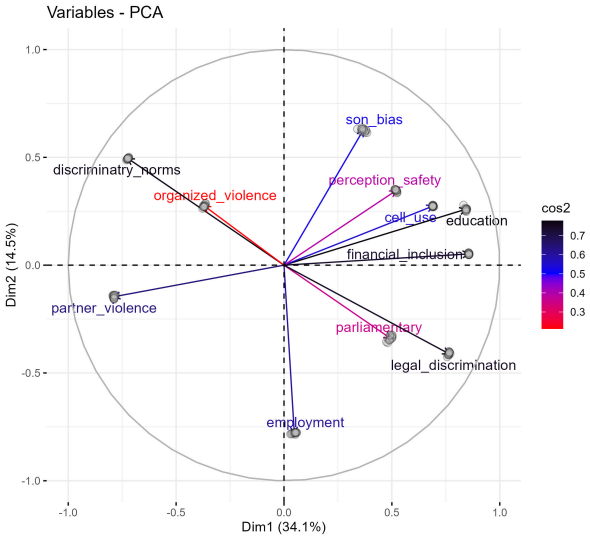

Figura 5: Gráficos de las variables del ACP

En la figura 6 en la que visualiza que los países imputados no muestran alguna particularidad que los haga agruparse de manera especifica en alguna de las dos dimensiones, al contrario, estos países muestran una amplia dispersión a lo largo del gráfico propuesto.

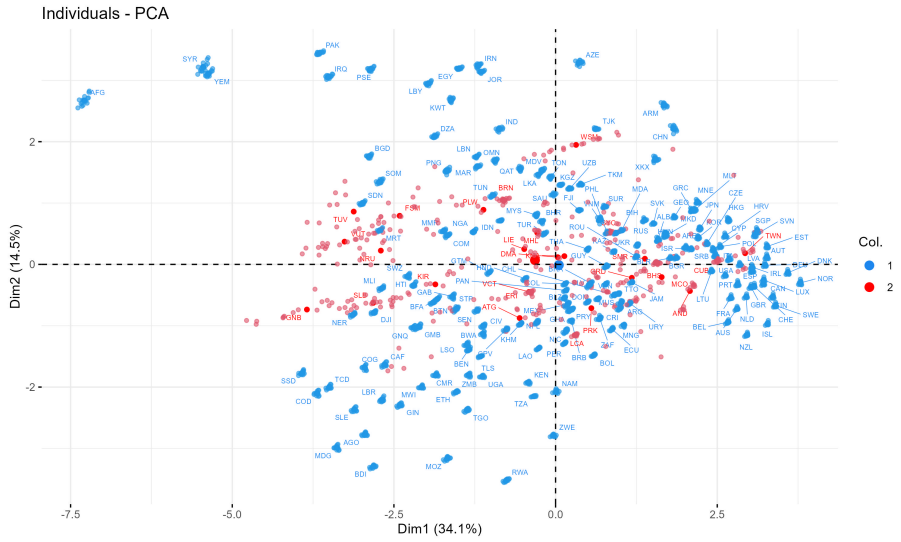

Figura 6: Gráficos de los individuos del ACP

En la tabla que se encuentra a continuación muestra con color azul aquellos países que fueron imputados, es importante anotar que a pesar de no tener toda la información disponible, estos pueden tener resultados sobresalientes, algunos de ellos como Taiwan, Cuba y Andorra que muestran buenos resultados. Otros países, especialmente aquellos que se encuentran en el continente africano muestran resultados en los que las políticas de igualdad de género aún no son las suficientes.

Cuadro 2: Desempeño y clasificación del país en el Índice e indicadores de paz y seguridad de las mujeres

PAÍS	RANK	WPS
Norway	1	0.921
Finland	2	0.909
Iceland	3	0.907
Denmark	4	0.903
Luxembourg	5	0.9
Switzerland	6	0.898
Sweden	7	0.895
Austria	8	0.891
United Kingdom	9	0.888
Netherlands	10	0.885
Germany	11	0.88
Canada	12	0.879
New Zealand	13	0.873
Spain	14	0.872
Singapore	15	0.87
France	16	0.87
Taiwan	17	0.87
Slovenia	18	0.87
Portugal	19	0.867
Ireland	20	0.867
Estonia	21	0.863
United States	22	0.861
Belgium	23	0.859
Latvia	24	0.858
United Arab Emirates	25	0.856
Australia	26	0.856
Cuba	27	0.849
Croatia	28	0.848
Israel	29	0.844
Italy	30	0.842
Poland	31	0.84

Cuadro 2: Desempeño y clasificación del país en el Índice e indicadores de paz y seguridad de las mujeres

PAÍS	RANK	WPS
Andorra	32	0.834
Lithuania	33	0.833
Czechia	34	0.83
Hong Kong SAR China	35	0.829
South Korea	36	0.827
Serbia	37	0.826
Japan	38	0.823
Cyprus	39	0.82
Malta	40	0.815
Belarus	41	0.814
Slovakia	42	0.811
Georgia	43	0.808
Monaco	44	0.807
Bulgaria	45	0.804
Bahamas	46	0.803
Montenegro	47	0.803
Jamaica	48	0.8
North Macedonia	49	0.798
Grenada	50	0.796
San Marino	51	0.796
Greece	52	0.792
Hungary	53	0.79
Costa Rica	54	0.78
Uruguay	55	0.776
Bolivia	56	0.774
Ecuador	57	0.774
Argentina	58	0.774
Trinidad & Tobago	59	0.771
Russia	60	0.77
North Korea	61	0.77
Mongolia	62	0.769
Romania	63	0.765
Guyana	64	0.764
Bosnia & Herzegovina	65	0.764
Albania	66	0.762
Kazakhstan	67	0.761
Turkmenistan	68	0.76
Philippines	69	0.759

Cuadro 2: Desempeño y clasificación del país en el Índice e indicadores de paz y seguridad de las mujeres

PAÍS	RANK	WPS
Seychelles	70	0.759
Chile	71	0.757
Nicaragua	72	0.756
Moldova	73	0.75
Mauritius	74	0.749
Rwanda	75	0.748
South Africa	76	0.748
El Salvador	77	0.747
Ukraine	78	0.747
Ghana	79	0.747
Venezuela	80	0.746
Dominican Republic	81	0.745
Thailand	82	0.744
Uzbekistan	83	0.741
Laos	84	0.741
Tanzania	85	0.739
Barbados	86	0.738
Paraguay	87	0.737
St. Lucia	88	0.737
Kosovo	89	0.737
Brazil	90	0.734
Fiji	91	0.734
Suriname	92	0.734
Peru	93	0.733
Panama	94	0.733
Dominica	95	0.732
St. Kitts & Nevis	96	0.731
Zimbabwe	97	0.727
Armenia	98	0.727
Tajikistan	99	0.726
Mexico	100	0.725
Eritrea	101	0.724
Colombia	102	0.722
China	103	0.722
Kenya	104	0.721
Belize	105	0.72
Tonga	106	0.719
Cambodia	107	0.718

Cuadro 2: Desempeño y clasificación del país en el Índice e indicadores de paz y seguridad de las mujeres

PAÍS	RANK	WPS
Nepal	108	0.714
Namibia	109	0.713
Kyrgyzstan	110	0.713
Bahrain	111	0.713
Qatar	112	0.713
Timor-Leste	113	0.707
Indonesia	114	0.707
Antigua & Barbuda	115	0.703
Saudi Arabia	116	0.703
Malaysia	117	0.702
Marshall Islands	118	0.7
St. Vincent & Grenadines	119	0.7
Honduras	120	0.698
Sri Lanka	121	0.698
Liechtenstein	122	0.696
Turkey	123	0.693
Vietnam	124	0.692
Cape Verde	125	0.69
Uganda	126	0.685
Samoa	127	0.684
Brunei	128	0.681
Oman	129	0.675
Mozambique	130	0.673
Maldives	131	0.671
Ethiopia	132	0.668
Benin	133	0.667
Guatemala	134	0.665
Zambia	135	0.661
Tunisia	136	0.659
Palau	137	0.659
Botswana	138	0.657
São Tomé & Príncipe	139	0.657
Togo	140	0.656
Senegal	141	0.656
Côte d'Ivoire	142	0.654
Kuwait	143	0.653
Lesotho	144	0.65
Iran	145	0.649

Cuadro 2: Desempeño y clasificación del país en el Índice e indicadores de paz y seguridad de las mujeres

PAÍS	RANK	WPS
Cameroon	146	0.648
Jordan	147	0.646
Malawi	148	0.644
Bhutan	149	0.643
Nigeria	150	0.635
Burundi	151	0.635
Lebanon	152	0.63
Azerbaijan	153	0.63
Myanmar (Burma)	154	0.629
Comoros	155	0.628
Burkina Faso	156	0.627
Egypt	157	0.627
Morocco	158	0.624
Equatorial Guinea	159	0.624
Gabon	160	0.623
Kiribati	161	0.622
Algeria	162	0.616
Haiti	163	0.611
Mali	164	0.61
Angola	165	0.609
Solomon Islands	166	0.607
Papua New Guinea	167	0.605
Eswatini	168	0.602
Guinea	169	0.602
India	170	0.597
Gambia	171	0.597
Libya	172	0.596
Djibouti	173	0.594
Liberia	174	0.594
Bangladesh	175	0.594
Niger	176	0.594
Nauru	177	0.589
Micronesia (Federated States of)	178	0.582
Congo - Brazzaville	179	0.582
Madagascar	180	0.579
Mauritania	181	0.577
Central African Republic	182	0.577
Somalia	183	0.572

Cuadro 2: Desempeño y clasificación del país en el Índice e indicadores de paz y seguridad de las mujeres

PAÍS	RANK	WPS
Palestinian Territories	184	0.571
Sierra Leone	185	0.563
Tuvalu	186	0.558
Vanuatu	187	0.557
Sudan	188	0.556
Guinea-Bissau	189	0.553
Chad	190	0.547
Congo - Kinshasa	191	0.547
South Sudan	192	0.54
Iraq	193	0.516
Pakistan	194	0.476
Yemen	195	0.389
Syria	196	0.249
Afghanistan	197	0.112

En figura 7 es posible ver que la mayoría de los países de la región europea presenta buenos resultados en el indicador, lo que muestra políticas establecidas claras para trabajar por la equidad de género y la inclusión de la mujer a nivel social y económico. Mientras que países como Afganistan, Siria, Yemen y Pakistan tienen una mayor brecha en sus resultados indicando esto una oportunidad de mejora en sus políticas de igualdad.

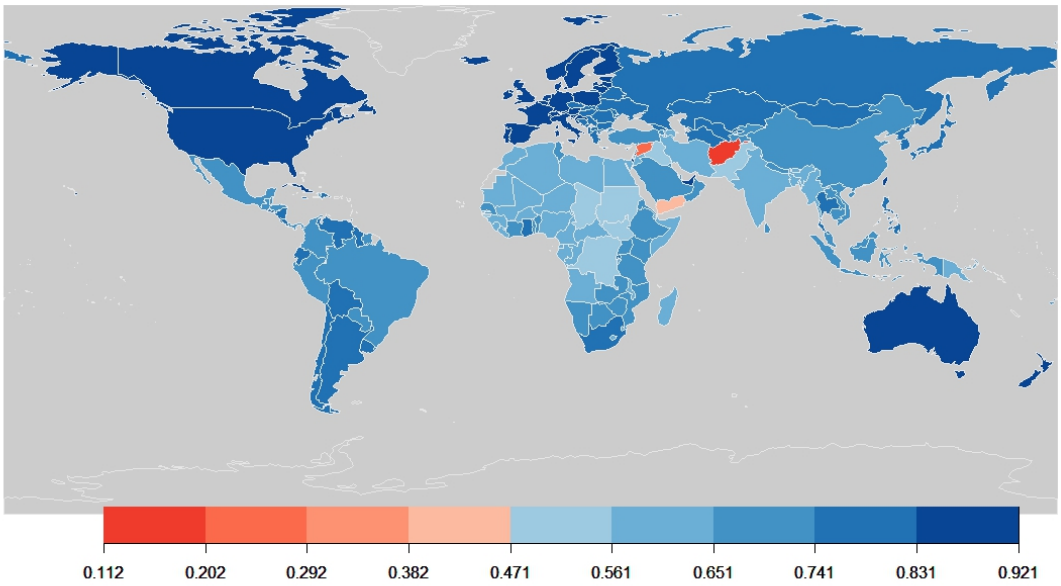

Figura 7: WPS a nivel mundial

El error estimado que se obtiene para los países imputados muestra para la mayoría de ellos un ee bajo, sin embargo para Micronesia, Tuvalu y Nauru superior al 0.09 mientras que para Cuba, Andorra y Monaco muestra un error más bajo.

Cuadro 3: Resumen error estimado y límites inferior y superior de los países imputados

PAÍS	RANK	WPS	ee	LI	LS
Taiwan	17	0.87	0.042	0.788	0.952
Cuba	27	0.849	0.032	0.786	0.911
Andorra	32	0.834	0.037	0.762	0.907
Monaco	44	0.807	0.044	0.721	0.893
Bahamas	46	0.803	0.04	0.725	0.881
Grenada	50	0.796	0.054	0.69	0.902
San Marino	51	0.796	0.045	0.708	0.884
North Korea	61	0.77	0.062	0.648	0.891
Seychelles	70	0.759	0.056	0.649	0.868
St. Lucia	88	0.737	0.053	0.633	0.841
Dominica	95	0.732	0.043	0.647	0.816
St. Kitts & Nevis	96	0.731	0.06	0.613	0.849

Cuadro 3: Resumen error estimado y límites inferior y superior de los países imputados

PAÍS	RANK	WPS	ee	LI	LS
Eritrea	101	0.724	0.065	0.596	0.851
Antigua & Barbuda	115	0.703	0.044	0.617	0.79
Marshall Islands	118	0.7	0.055	0.592	0.808
St. Vincent & Grenadines	119	0.7	0.049	0.603	0.796
Liechtenstein	122	0.696	0.075	0.549	0.843
Samoa	127	0.684	0.063	0.561	0.808
Brunei	128	0.681	0.059	0.565	0.797
Palau	137	0.659	0.081	0.501	0.818
Kiribati	161	0.622	0.096	0.434	0.81
Solomon Islands	166	0.607	0.073	0.464	0.75
Nauru	177	0.589	0.096	0.401	0.777
Micronesia (Federated States of)	178	0.582	0.104	0.378	0.786
Tuvalu	186	0.558	0.101	0.36	0.756
Vanuatu	187	0.557	0.093	0.375	0.74
Guinea-Bissau	189	0.553	0.057	0.441	0.665

A continuación se presenta un resumen de las medidas de severidad de los países imputados, en la que λ es la Proporción de varianza debida a los datos faltantes, r es el Incremento relativo de la varianza y γ es la Fracción de información faltante sobre el WPS y ν los grados de libertad. Se destacan en rojo y negrita los países que tuvieron resultados en la Proporción de varianza e Incremento superior al 30 %, Rubin recomienda que estas medidas no sean mayor a este valor, sin embargo estos resultados muestran que el nivel de datos faltantes es tan severo que la proporción de varianza debida al proceso de imputación también se incrementa.

Cuadro 4: Resumen medidas de severidad de los países imputados

PAIS	WPS	λ	r	ν	γ
Andorra	0,834	14,499	16,958	165,898	0,155
Antigua & Barbuda	0,703	**55,024**	**122,34**	87,267	0,56
Bahamas	0,802	0,032	0,032	193,968	0,01
Brunei	0,680	0,058	0,058	193,918	0,011
Cuba	0,848	0,034	0,034	193,965	0,01
North Korea	0,769	**39,533**	**65,381**	117,323	0,405
Dominica	0,731	**55,149**	**122,962**	87,024	0,561
Eritrea	0,723	**39,709**	**65,864**	116,982	0,407
Micronesia (Federated States of)	0,582	8,619	9,432	177,307	0,096

Cuadro 4: Resumen medidas de severidad de los países imputados

PAIS	WPS	λ	r	ν	γ
Grenada	0,796	12,899	14,809	169,002	0,139
Guinea-Bissau	0,552	22,886	29,677	149,625	0,239
Kiribati	0,622	14,184	16,528	166,51	0,152
Liechtenstein	0,695	**58,951**	**143,609**	79,648	0,599
Marshall Islands	0,699	2,909	2,996	188,387	0,039
Monaco	0,806	15,975	19,013	163,033	0,17
Nauru	0,589	16,521	19,791	161,974	0,175
Palau	0,659	17,099	20,626	160,853	0,181
St. Kitts & Nevis	0,730	**33,369**	**50,08**	129,284	0,344
St. Lucia	0,736	0,016	0,016	193,999	0,01
St. Vincent & Grenadines	0,699	0,038	0,038	193,956	0,011
Samoa	0,684	11,572	13,086	171,577	0,126
San Marino	0,795	8,019	8,718	178,471	0,09
Seychelles	0,758	0,008	0,008	194,014	0,01
Solomon Islands	0,606	4,736	4,971	184,841	0,058
Taiwan	0,869	6,822	7,322	180,793	0,078
Tuvalu	0,558	8,361	9,124	177,807	0,094
Vanuatu	0,557	7,613	8,241	179,258	0,086

7. Conclusiones

La propuesta del presente trabajo fue la de proponer una función que permitiese imputar valores faltantes en el intervalo (0, 1), [0, 1), (0, 1] y [0, 1], haciendo uso de la distribución beta o distribuciones beta cero-uno infladas de acuerdo a la información de la variable.

Como resultado se obtiene que la función propuesta, aplicada en la base de datos del WPS Index 2021/2022 permite la imputación de manera adecuada de acuerdo a la información de cada variable, haciendo uso de otras funciones en el software y del algoritmo MICE.

La base de datos contaba con valores faltantes para 27 países, afectando las posibles conclusiones para estos, al aplicarle la función fue posible proporcionar un indicador que permitió ver que estos paises tienen en algunos casos mejores resultados que aquellos países para los cuales hay información observada.

Dado los resultados obtenidos, se espera que la función propuesta sirva de apoyo para posteriores investigaciones en los que las variables de interes presenten las distribuciones nombradas con anterioridad.

Referencias

Arteaga, F. Ferrer-Riquelme, A., 2009. 3.06 - missing data, in: Brown, S.D., Tauler, R., Walczak, B. (Eds.), Comprehensive Chemometrics. Elsevier, Oxford, pp. 285–314. URL: https://www.sciencedirect.com/science/article/pii/B9780444527011001253, doi:https://doi.org/10.1016/B978-044452701-1.00125-3.

Azur, M. J. Stuart, E.A..F.C..L.P.J., 2011. Multiple imputation by chained equations: what is it and how does it work? International Journal of Methods in Psychiatric Research , 40–49doi:https://doi.org/10.1002/mpr.329.

Burch, B. Egbert, J., 2019. Zero-inflated beta distribution applied to word frequency and lexical dispersion in corpus linguistics. Journal of Applied Statistics 47, 337–353. URL: https://doi.org/10.1080/02664763.2019.1636941, doi:10.1080/02664763.2019.1636941.

van Buuren, S., .G.O.K., 2011. mice: Multivariate imputation by chained equations in r. Journal of Statistical Software 45, 1–67. doi:10.18637/jss.v045.i03.

van Buuren, S., 2018. Flexible Imputation of Missing Data, Second Edition. Chapman and Hall/CRC.

Ferrari, S. Cribari-Neto, F., 2013. Beta regression for modelling rates and proportions. Journal of Applied Statistics , 799–815doi:http://dx.doi.org/10.1080/0266476042000214501.

Gender®, G., 2007. Progress of women's rights: Taiwan women, an incredible century. http://www.globalgender.org/en-global/status_page/index/2. Accedido en junio de 2023.

GIWPS, PRIO, 2021. Women, peace, and security index 2021/22: Tracking sustainable peace through inclusion, justice, and security for women. washington .

Johnson, R.W., 2016. Applications of the beta distribution part 1: Transformation group approach arXiv:1307.6437.

Little, R.J., D'Agostino, R., Cohen, M.L., Dickersin, K., Emerson, S.S., Farrar, J.T., Frangakis, C., Hogan, J.W., Molenberghs, G., Murphy, S.A., Neaton, J.D., Rotnitzky, A., Scharfstein, D., Shih, W.J., Siegel, J.P., Stern, H., 2012. The prevention and treatment of missing data in clinical trials. New England Journal of Medicine 367, 1355–1360. URL: https://doi.org/10.1056/NEJMsr1203730, doi:10.1056/NEJMsr1203730, arXiv:https://doi.org/10.1056/NEJMsr1203730.

Ospina, R. Ferrari, S., 2012. A general class of zero-or-one inflated beta regression models. Computational Statistics Data Analysis 56, 1609–1623. doi:https://doi.org/10.1016/j.csda.2011.10.005.

for Peace, T.F., 2017. Fragile states index methodology and cast framework. https://fragilestatesindex.org/wp-content/uploads/2017/05/FSI-Methodology.pdf. Accedido en junio de 2023.

University of Notre Dame from Washington, D., . Nd-gain center - university of notre dame. https://gain.nd.edu/about/. Accedido el 29 de junio de 2023.

Wolf, M.J., Emerson, J.W., Esty, D.C., de Sherbinin, A., Wendling, Z.A., et al., 2022. 2022 environmental performance index. New Haven, CT: Yale Center for Environmental Law Policy. epi.yale.edu.

Xiao-Hua, Z., 2019. Challenges and strategies in analysis of missing data. Biostatistics Epidemiology 4, 15–23. doi:https://doi.org/10.1080/24709360.2018.1469810.

8. Anexos

- Script en R: función propuesta para la imputación de variables de razón y proporción beta_impute.
- Script en R: Análisis de la base propuesta para la aplicación de la función propuesta.
- Carta de habeas data