Identificación de datos faltantes y imputación en tasas de vacunación para un estudio ecológico del cono sur de Sudamérica en cuatro países

Ramón Álvarez-Vaz, Silvia Rodríguez-Collazo, Mauro Loprete

https://doi.org/10.1590/SciELOPreprints.12147

Envíado en: 2025-06-02
Postado en: 2025-06-16 (versión 1)
(AAAA-MM-DD)

IDENTIFICATION OF MISSING DATA AND IMPUTATION IN VACCINATION RATES FOR AN ECOLOGICAL STUDY OF THE SOUTHERN CONE OF SOUTH AMERICA IN FOUR COUNTRIES

Ramón Álvarez-Vaz,
Facultad de Ciencias Económicas y de Administración,
Universidad de la República, Uruguay
Montevideo, CP 11200
ORCID:https://orcid.org/0000-0002-2505-4238
ramon.alvarez@fcea.edu.uy

Silvia Rodríguez-Collazo,
Facultad de Ciencias Económicas y de Administración,
Universidad de la República, Uruguay
Montevideo, CP 11200
ORCID:https://orcid.org/0000-0002-3871-6448
silvia.rodrigez@fcea.edu.uy

Mauro Loprete,
Facultad de Ciencias Económicas y de Administración,
Universidad de la República, Uruguay
Montevideo, CP 11200
ORCID:https://orcid.org/0000-0003-1560-0183
mauro.loprete@fcea.edu.uy

2 de junio de 2025

Título en español: Identificación de datos faltantes y imputación en tasas de vacunación para un estudio ecológico del cono sur de Sudamérica en cuatro países

Contribución de autoría: Ramón Álvarez-Vaz fue responsable de la Conceptualización, Análisis formal, Curación de datos, Investigación, Metodología, Administración del proyecto, Software, Supervisión,Validación, Visualización, Redacción borrador original y Redacción del artículo.

Silvia Rodríguez-Collazo fue responsable de la Conceptualización, Investigación, Supervisión, Validación, Visualización, Redacción Borrador original.

Mauro Loprete fue responsable de la Curación de datos, Metodología, Software, Validación, Visualización.

Declaración de fuentes de financiación y conflicto de interés Los autores declaran que para esta investigación no tuvieron fuente de financiación externa.

Agradecimientos A todo el equipo de Facultad de Ciencias Económicas y de Administración, Universidad de la República, que colaboró fielmente con la investigación.

Correspondencia Ramón Álvarez-Vaz, E-mail ramon.alvarez@fcea.edu.uy

RESUMEN

En los estudios epidemiológicos con diseño ecológicos, es práctica habitual tener que trabajar con datos a nivel país, que surgen de fuentes de datos secundarias. Por eso motivo hay que pasar por un proceso de armonización de modo, que asegure la calidad y completitud de los datos, como elemento fundamental de los que se conoce como *investigación reproducible*. Eso permite que otras investigadoras y otros investigadores, puedan replicar los resultados al acceder a los mismos repositorios, que habitualmente son portales internacionales que publican datos estadísticos del área de la salud, como la OMS, el Banco Mundial o los propios institutos de estadística o ministerios de salud de cada país. Este proceso de armonización que garantiza la reproducibilidad tiene como una etapa fundamental el transparentar el proceso de identificación de datos faltantes y cualquier modificación que se haga al imputar y permitir tener datos completos, que amplia el uso de diferentes técnicas estadísticas. En este trabajo para un grupo de tasas de vacunación se muestra todo el proceso seguido que incluye visualización y varios métodos de imputación que se adecuán a la naturaleza de los datos, para 4 países del cono sur para un período de 20 años. Los detalles de este proyecto están en la plataforma OSF en el proyecto *Análisis de diferentes indicadores demográficos, sociales, económicos y su asociación con un grupo de enfermedades imunoprevenibles para 4 países del cono sur de Sudamérica* en https://osf.io/6r3ew/.

***K*eywords** Armonización, datos faltantes, imputación, investigación reproducible, visualización

ABSTRACT

In epidemiological studies with ecological design, it is common practice to work with country-level data, which arise from secondary data sources. For this reason, a harmonization process must be carried out in order to ensure the quality and completeness of the data, as a fundamental element of what is known as reproducible research. This allows other researchers to replicate the results by accessing the same repositories, which are usually international portals that publish statistical data in the health area, such as the WHO, the World Bank or the statistical institutes or health ministries of each country. This harmonization process that guarantees reproducibility has as a fundamental stage the transparency of the process of identifying missing data and any modification that is made when imputing and allowing for complete data, which expands the use of different statistical techniques. This work shows the entire process followed for a group of vaccination rates, including visualization and various imputation methods that are appropriate to the nature of the data, for 4 countries in the Southern Cone for a period of 20 years. The details of this project are on the OSF platform in the project *Analysis of different demographic, social, and economic indicators and their association with a group of immunopreventable diseases for 4 countries in the Southern Cone of South America* at https://osf.io/6r3ew/.

***K*eywords** Harmonization, Imputation, Missing data, Reproducible research, Visualization

1. Introducción

En los estudios epidemiológicos con diseño ecológicos, es práctica habitual tener que trabajar con datos a nivel país, que surgen de fuentes de datos secundarias. Por eso motivo hay que pasar por un proceso de armonización de modo, que asegure la calidad y completitud de los datos, como elemento fundamental de los que se conoce como *investigación reproducible*. En general los datos de fuentes secundarios, no siempre tienen datos del todo completos e incluso la calidad de los mismos no siempre excelente. Parte de la explicación de esa debilidad se debe al origen de los datos que luego esos portales publican. O bien son las propias agencias que imputan o reportan datos promedios de períodos pasados o sencillamente lo que divulgan es lo que recibieron y aparecen los datos faltantes y/o anómalos. Sin importar lo que finalmente se produzca al liberar los datos debe asegurarse que al acceder a los repositorios sea transparente el proceso de armonización que garantice la reproducibilidad, [1], [2], [3],[4], [5]. En esta sección se presentan las diferentes etapas que permiten establecer un flujo de trabajo como se detalla a continuación:

- ETAPA 1 - Armonización de los datos (previo al procesamiento de los mismos) y preparación del proyecto en el software a utilizar.

- ETAPA 2 - Implementación del sistema de información con una interface gráfica para la descarga de datos de fuentes de datos secundarias.
- ETAPA 3 - Análisis exploratorio y selección de variables.
- ETAPA 4 - Visualización de datos faltantes.
- ETAPA 5 - Imputación de datos faltantes.
- ETAPA 6 - Visualización interactiva (R Shiny).

El ciclo de vida de este trabajo consta un primer avance presentado en Agosto de 2024 con un resumen extendido para la XVII Semana Internacional de la Estadística y la Probabilidad, de la Facultad de Ciencias Fisico Matemáticas de la Benemérita Universidad Autónoma de Puebla con número de preprint https://doi.org/10.5281/zenodo.13763282; se complementa posteriormente con el trabajo casi terminado previo a la elaboración de este documento que se presentó en las jornadas de Estadística de Octubre de 2024 en Montevideo, Uruguay, disponible en https://doi.org/10.5281/zenodo.14652623. Para asegurar la reproducibilidad de los resultados del análisis realizado, se dispone el código y datos utilizados en un repositorio público en la plataforma OSF al que se puede acceder a través de https://osf.io/6r3ew/. El documento está estructurado del siguiente modo: Una sección de Materiales en 2.1 en página 3, seguido de los métodos que aparecen en la sección 2.2 de la página 3, proponiendo finalmente algunas conclusiones de lo hallado y posible pasos a seguir que se consignan en la sección 3, de la página 18.

2. Material y Métodos

2.1. Materiales

En esta sección dada la naturaleza del trabajo que consiste en usar fuentes de datos secundarias se presenta los datos disponibles, donde se considera la cantidad de variables, sus definiciones de donde se extraen, como parte del fluxograma presentado en la sección anterior.

2.1.1. Bloque de vacunas

A continuación se listan todas las vacunas inicialmente consideradas en el trabajo. Hay que tener en cuenta que se trabaja con la información que corresponde a 20 años de 4 países (**Argentina, Chile, Paraguay, Uruguay**), por lo cual el bloque de vacunas originalmente debería tener 1520 registros (19*20*4). Sin embargo faltan registros, hay 1360 observaciones por lo que de ese bloque de 19, teniendo en cuenta la falta de información para este trabajo sólo se considerarán 7 patologías inmuno-prevenibles a través de las tasas que corresponden a las mismas y del cual se crea un bloque que se denomina **BRV**, por Bloque Reducido de Vacunas.

- HEPB3: Hepatitis B; tercera dosis
- HIB3: Haemophilus influenzae tipo b; tercera dosis
- POL3: Polio; tercera dosis
- DPT1: Difteria, Pertusis, Tétanus; primera dosis
- MCV1: Sarampion, primera dosis
- MCV2: Sarampion; segunda dosis
- DPT3: Difteria, Pertusis, Tétanus; tercera dosis

2.2. Metodología

La metodología que se utiliza luego de la armonización de los datos, consiste en el análisis en 3 dimensiones

- Análisis de la calidad de los datos y de datos Faltantes (ver sección 2.2.1).
- Imputación de los datos faltantes (ver sección 2.2.5).
- Análisis estadśtico de la información los datos armonizados y luego imputados mediante diferentes aproximaciones metodológicas.

2.2.1. Datos faltantes para bloque de vacunas

Se cargan los datos estructurados en formato largo, es decir que se tiene una tabla de datos con 80 observaciones correspondientes a los 4 países estudiados en el período de tiempo 2000 a 2019. En esta parte de la preparación de datos se usa el software `R` [6] con librerías como `psych`[7],la interface `Rstudio` [8], con `readxls` [9] y `knitr` [10]. Es importante recordar que en la seccción 2 se consignó que se iba a trabajar con el bloque reducido de vacunas (denominado de aquí en más **BRV**).

Se crea para el análisis el bloque reducido de vacunas y los indentificadores de países y los años. A continuación se muestran los primeros 6 registros de la tabla para 10 años, es decir parados en 2019 10 años hacia atrás.

```
#anio  pais HEPB3 HIB3 POL3 DTP1 MCV1 MCV2 DTP3
#2010 argent. 94   94   95   95  105   94   94
#2011 argent. 91   91   93   94   95   91   91
#2012 argent. 91   91   90   94   94   89   91
#2013 argent. 94   94   90   94   94   83   94
#2014 argent. 94   94   92   98   95   96   94
#2015 argent. 94   94   93   94   89   87   94
```

En cambio si se considera los 20 años se muestran los primeros 6 registros de la tabla.

```
# anio pais HepB3 Hib3 Pol3 DTP1 MCV1 MCV2 DTP3
# 2000 argent. NA   83   88   88   91   56   83
# 2001 argent. NA   83   85   92   89   56   83
# 2002 argent. 66   93   94   NA   95   NA   93
# 2003 argent. 90   96   95   NA   97   80   96
# 2004 argent. 73   98   91   96   99   93   98
# 2005 argent. 88   98   95   93  110   88   98
```

Si bien se efectúa el análisis para los 10 últimos años y para los 20 años, por un tema de extensión solo se presentan los resultados para 20 años.

2.2.2. Análisis de bloques de vacunas en los 20 años (2000-2019)

En la presente sección se analizan los datos correspondientes a los 20 años de análisis y se visualizan los datos faltantes. Se utiliza, para identificar los patrones de datos faltantes, la librería `naniar`, [11].

```
##          vars  n  media   min  max
## HEPB3       3 68  88.40   66   96
## HIB3        4 76  89.34   55   98
## POL3        5 79  89.65   71   97
## DTP1        6 77  91.32   67  100
## MCV1        7 79  91.14   66  110
## MCV2        8 57  81.60   28  110
## DTP3        9 80  89.92   72   98
```

A partir de la Figura 1 se puede decir que para el período de 20 años, nuevamente la vacuna MCV2 es la que tiene mayor cantidad de valores ausentes.

En la Figura 2 por un lado se puede ver cómo es el patrón de no respuesta por país, la vacuna **MCV2** tiene casi un $30\,\%$ de valores ausentes que le corresponden principalmente a Uruguay.

Por otro lado, se pueden ver las observaciones cuyos valores están por arriba de $100\,\%$, siendo la vacuna que tiene más valores anómalos **MCV1**.

A fin de investigar si existe un patrón para las observaciones por debajo del umbral $90\,\%$, se crea la Figura 3, la cual muestra que nuevamente Paraguay, en los 20 años, es el país con inmunización insuficiente y solo para **MCV2**, Argentina y Chile muestran baja inmunización.

Tal como muestra la Figura 4, las correlaciones entre las vacunas siguen siendo todas positivas.

2.2.3. Exploración de otros patrones para 20 años

Para el período de 2010 a 2019, Uruguay es el país con mas valores faltantes que se producen a lo largo de los 10 años.

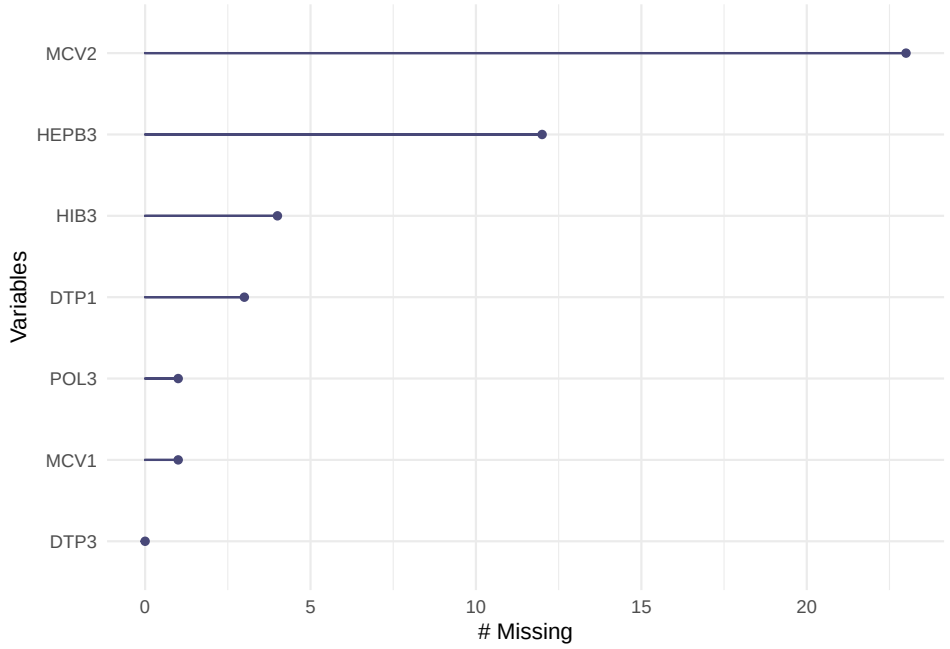

Figura 1: Datos faltantes por vacuna para 20 años.

Figura 2: Datos faltantes y valores por encima de 100 % por observación país año para 20 años.

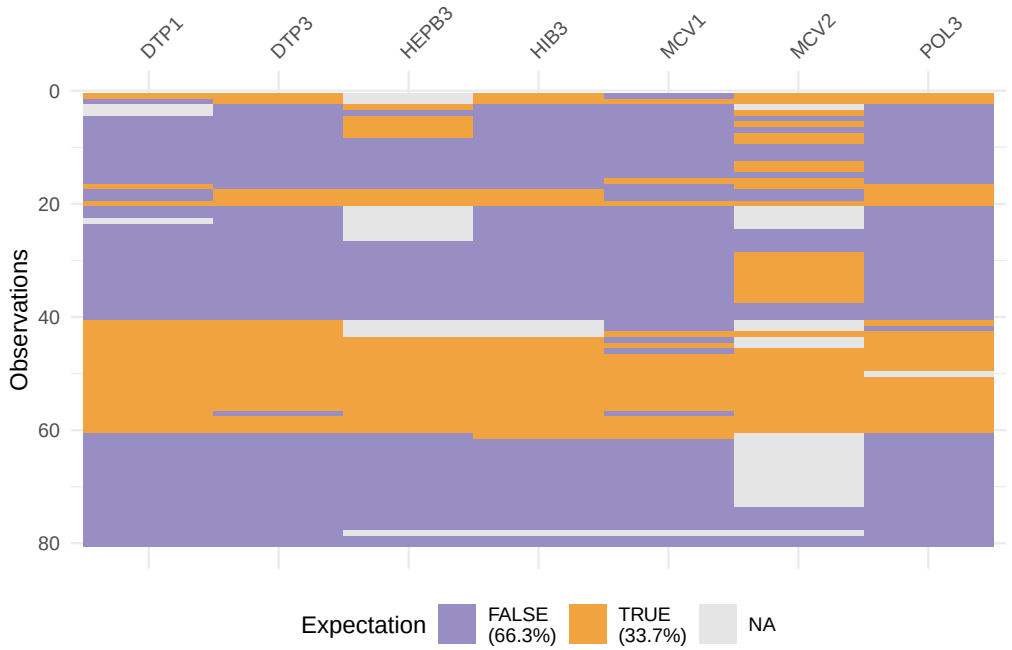

Figura 3: Datos faltantes y valores por debajo del umbral de 90 % por observación país año para 20 años.

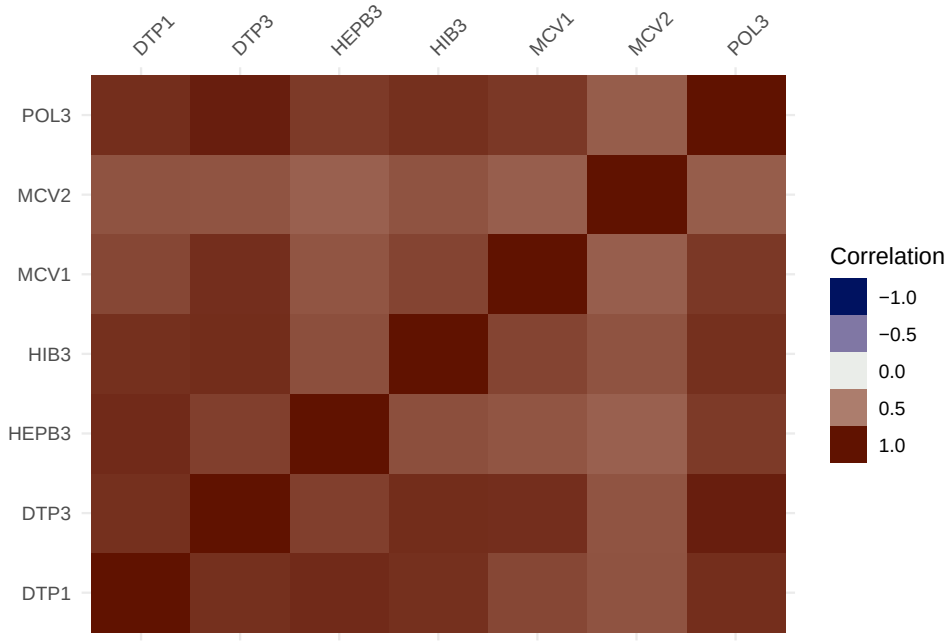

Figura 4: Correlación cruzada para variables de vacunas para 20 años.

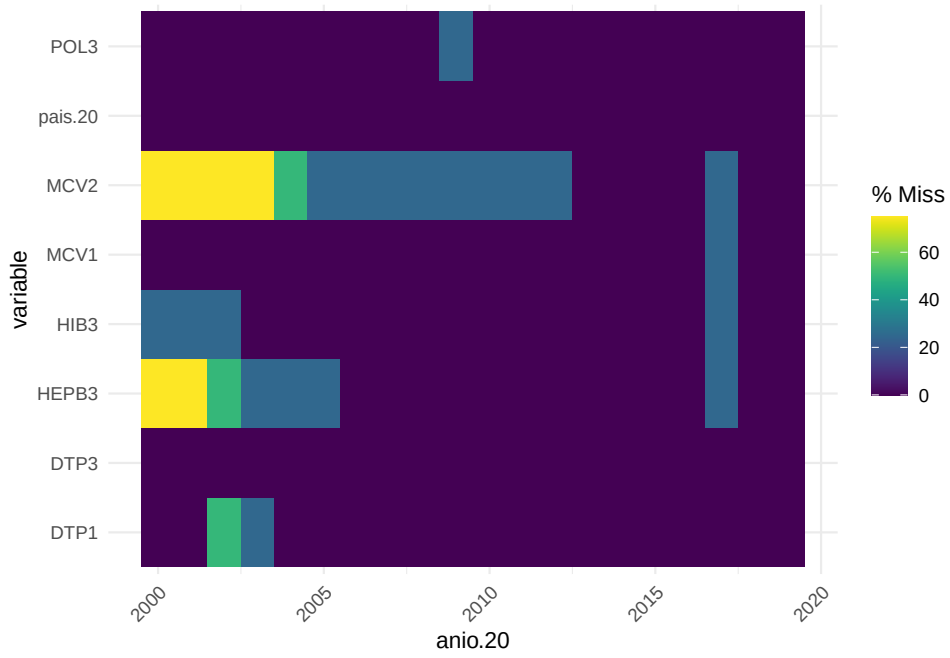

Figura 5: Patrón de respuesta por variable en el tiempo, para período 20 años.
1

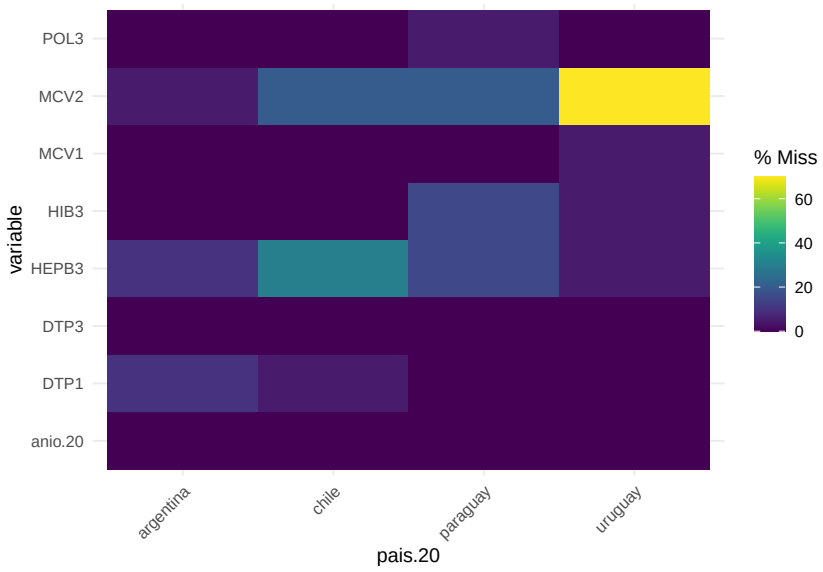

Figura 6: Patrón de respuesta por variable en cada país, para período 20 años.

Al ampliar la ventana temporal y considerar de 2000 a 2019 (Figura 7), Argentina es el país con menos valores faltantes, seguidos por Chile y Paraguay que concentra los sin datos en los primeros años, mientras que Uruguay muestra que solo en 5 años tiene valores completos, siendo esto para la última parte del período y donde la vacuna que mas valores ausentes presenta como se vió en secciones anteriores es **MCV2** .

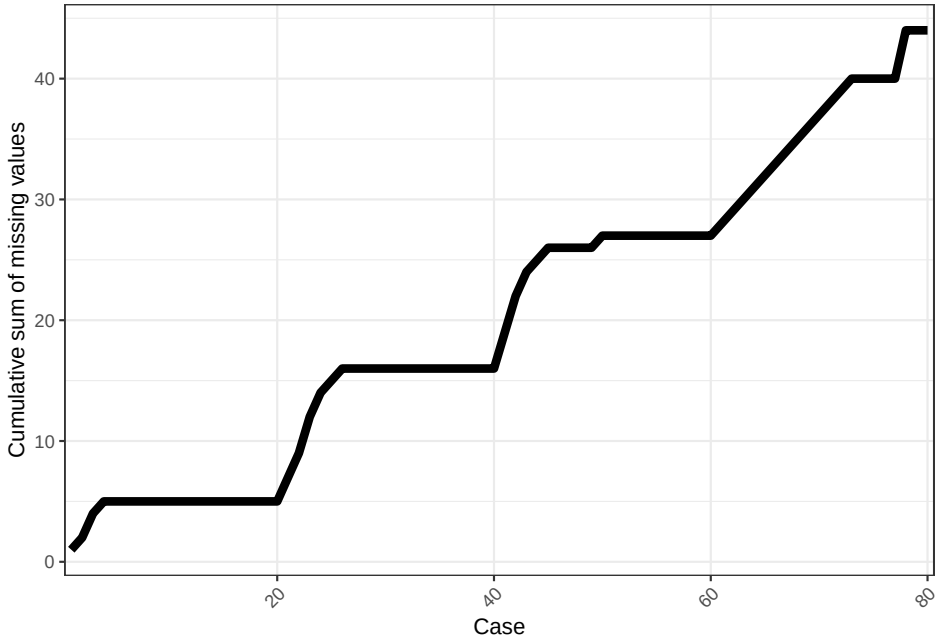

Figura 7: Casos acumulados con valores faltantes para 20 años.

En función de lo observado en las diferentes figuras se destaca que:

- En general hay más valores ausentes al considerar el período de 20 años, concentrados en $(2010 - 2019)$.
- Paraguay es el país con menor vacunación en la mayoría de las vacunas, por debajo del 90 $\%$.
- Las tasas muestran correlación lineal positiva.

A partir de los resultados encontrados se puede decir que:

- Hay que truncar a 100 las tasas por arriba de ese valor, en este caso son **MCV1**, **MCV2** y **DTP1**.
- Hay que imputar **MCV2** para varios años en Uruguay y para esa vacuna Argentina es el que tiene casi la información completa.
- Hay que imputar para los últimos 2 años de Uruguay en las variables **HEPB3, HIB3, MCV1, MCV2**, mientras que esa misma imputación hay que hacerla en Chile a comienzo de período.
- **DTP3**, **MCV1** y **POL3** no tienen datos faltantes, salvo un año en Paraguay.

2.2.4. Propuestas de Imputación

Se presentan a continuación algunas de las propuestas de como tratar los datos faltantes:

- Método univariado **Hotdeck**, estratificado por país. Es decir que la imputación de los datos faltantes en cada variable depende solamente de los valores en esa variable y en ese país, **y no considera los datos de los otros países**. El mecanismo de imputación consiste en detectar cuando se presenta para un año y país determinado, un dato o datos faltantes y recibir un donante, que es el último dato disponible. La función *hotdeck* pertenece a la libreria VIM,[12].
- Método univariado de tipo **kNN**. La función kNN de la librería VIM hace una k-Imputación del vecino más cercano basada en una variación de la distancia de Gower (Gower 1971) para variables numéricas, categóricas, ordenadas y semicontinuas. En este caso se fija un radio de tamaño k que permite considerar las observaciones que se aglomeran, de acuerdo a las distancias entre cada fila (recordar que son cada año en un país). Dado que

los resultados son muy similares a los encontrados para hotdeck no se considera esta versión de imputación, [12]. Al igual que el método hot-deck, el método de k vecinos más cercanos se basa en la observación de valores donantes, es decir que utiliza una agregación de los k valores más cercanos como valor imputado y el tipo de agregación depende del tipo de variable. El cálculo de la distancia para definir los vecinos más próximos se basa en una extensión de la distancia de Gower, siendo la distancia entre dos observaciones la media ponderada de las contribuciones de cada variable, donde el peso debe representar la importancia de la variable, por lo tanto, la distancia entre la i-ésima y la j-ésima observación puede definirse como

$$d_{i,j} = \frac{\sum_{p}^{k=1} w_k \delta_{i,j,k}}{\sum_{p}^{k=1} w_k}. \tag{1}$$

donde w_k es la ponderación y $\delta_{i,j,k}$ la contribución de la variable k. Esta última se computa como una distancia que se reescala con la siguiente expresión

$$\delta_{i,j,k} = \frac{|x_{i,k} - x_{j,k}|}{r_k}. \tag{2}$$

donde r_k es el rango de la variable k.

- Método multivariado con la función y librería `Amelia`. Este utiliza un algoritmo EM (espectativa-maximización) y bootstrap, además permite no sólo trabajar con más de una tabla de datos imputada por bloque sino también considerar los datos como si fueran de panel (que es el caso). Al ser un método multivariado no es lo mismo que se considere solamente la matriz que tiene datos faltantes por bloque, que si se considera un bloque en su totalidad, [13],[14],[15],[16]. Como resultado de la imputación al ser solución de un algoritmo de tipo EM (espectativa-maximización) y bootstrap se tiene una serie de tablas imputadas que luego deben ser promediadas. Una característica de este método es que supone distribución normal multivariada.

Los supuestos para el método multivariado son lo siguientes:

- El modelo de imputación en el que se basa el algoritmo supone que los datos completos (es decir, tanto observados como no observados) tiene una distribución gaussiana multivariada.
- $D \sim N_k(\mu, \sigma^2)$, donde $D = D_{obser}, D_{falt}$
- El algoritmo también supone, sólo observar a D_obs,y no a D que los datos faltantes tiene un mecanismo de tipo *MAR*. Esta equivale a suponer que el patrón de datos faltantes solo depende de los datos observados D_{obs}
- Si M es la matriz de faltantes, con celdas $m_{ij} = 1$ si es dato faltante o 0 si es datos observado

$$p(M|D) = p(M|D_{obs}). \tag{3}$$

$$p(D_{obs}, M|\theta) = p(M|D_{obs})p(D_{obs}|\theta). \tag{4}$$

$$L(\theta|D_{obs})p(D_{obs}|\theta), \tag{5}$$

Usando la leyes de esperanzas iteradas

$$p(D_{obs}|\theta) = \int p(D|\theta)dD_{falt} \tag{6}$$

$$p(\theta|D_{obs})p(D_{obs}|\theta) = \int p(D|\theta)dD_{falt} \tag{7}$$

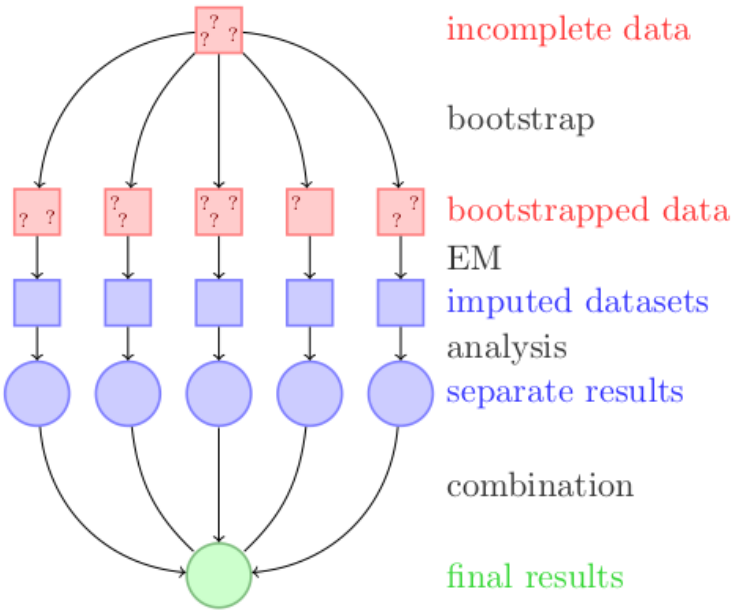

Figura 8: Un esquema del enfoque de imputación múltiple con el algoritmo EMB (extraído de [16, pag 4]).

- el algoritmo *EMB* combina el algoritmo *EM* clásico con un enfoque boostrapo para extraer muestras de la **distribución a posteriori**
- Para cada corrida, se extrae una muestra boostrapa que permite simular la incertidumbre de la estimación, y luego se ejecuta el algoritmo EM para encontrar la moda a posteriori de los datos muestreados, [15].
- Si el interéses estimar la cantidad q, una media, coeficiente de regresión, etc, lo que se puede hacer es trabajar con m conjunto de datos imputados y obtener una media de las m estimaciones hechas por separado

$$\bar{q} = \frac{1}{m}\sum_{j=1}^{m} q_j \tag{8}$$

De la cual se puede luego calcular un error estándar como

$$EE(q)^2 = \frac{1}{m}\sum_{j=1}^{m} EE(q_j)^2 + \frac{\sum_{j=1}^{m}(q_j - \bar{q})^2}{(m-1)} \tag{9}$$

2.2.5. Imputación para vacunas

A partir del estudio de los datos faltantes los posibles mecanismos de imputación son:

- Escenario 1: si falta en t, imputar por la media de $t-1; t+1$ si es posible.
- Escenario 2: imputar por le media del período.
- Escenario 3: probar algún mecanismo de imputación de tipo Missing Completely at Random (MCAR), Missing at Random (MAR), Missing not at Random (MNAR). En particular decidir si se procede a hacer imputación univariada o multivariada (con los métodos presentados en sección 2.2.4) y también considerar un mecanismo de imputación que tenga en cuenta que se trata de datos de panel.

Se considera el BRV

```
# anio  pais HEPB3 HIB3 POL3 DTP1 MCV1 MCV2 DTP3
# 2010 argent. 94   94   95   95  105   94   94
# 2011 argent. 91   91   93   94   95   91   91
# 2012 argent. 91   91   90   94   94   89   91
# 2013 argent. 94   94   90   94   94   83   94
# 2014 argent. 94   94   92   98   95   96   94
# 2015 argent. 94   94   93   94   89   87   94
```

Los valores que resumen las 7 tasas muestran 2 tasas con la mitad de los datos.

```
##       vars  n  mean min max
## HEPB3    1 39 88.54  73  96
## HIB3     2 39 88.54  73  96
## POL3     3 40 88.15  71  96
## DTP1     4 40 91.33  73 100
## MCV1     5 39 89.54  66 105
## MCV2     6 36 82.81  60 101
## DTP3     7 40 89.10  73  96
```

Si se considera el período más largo, se ve que hay años donde no hay valores para Argentina.

```
# anio  pais HEPB3 HIB3 POL3 DTP1 MCV1 MCV2 DTP3
#  2000 argent. NA   83   88   88   91   56   83
#  2001 argent. NA   83   85   92   89   56   83
#  2002 argent. 66   93   94   NA   95   NA   93
#  2003 argent. 90   96   95   NA   97   80   96
#  2004 argent. 73   98   91   96   99   93   98
#  2005 argent. 88   98   95   93  110   88   98
```

```
##       vars  n  mean min max
## HEPB3    1 68 88.40  66  96
## HIB3     2 76 89.34  55  98
## POL3     3 79 89.65  71  97
## DTP1     4 77 91.32  67 100
## MCV1     5 79 91.14  66 110
## MCV2     6 57 81.60  28 110
## DTP3     7 80 89.92  72  98
```

2.2.6. Truncamiento de datos anómalos

En esta sección se modifican los datos de vacunas que superan el 100 %, ya que no puede haber una tasa de vacunación superior a dicho número, el valor en estos casos es truncado por 100 %.

```
vars  n  media min max
HEPB3    1 68 88.40  66  96
HIB3     2 76 89.34  55  98
POL3     3 79 89.65  71  97
DTP1     4 77 91.32  67 100
MCV1     5 79 90.91  66 100
MCV2     6 57 81.40  28 100
DTP3     7 80 89.92  72  98
```

2.2.7. Identificación y Visualización de datos faltantes para BRV

Para el análisis y visualización de los datos faltantes se usa la librería `naniar` [11] y `VIM` [12].

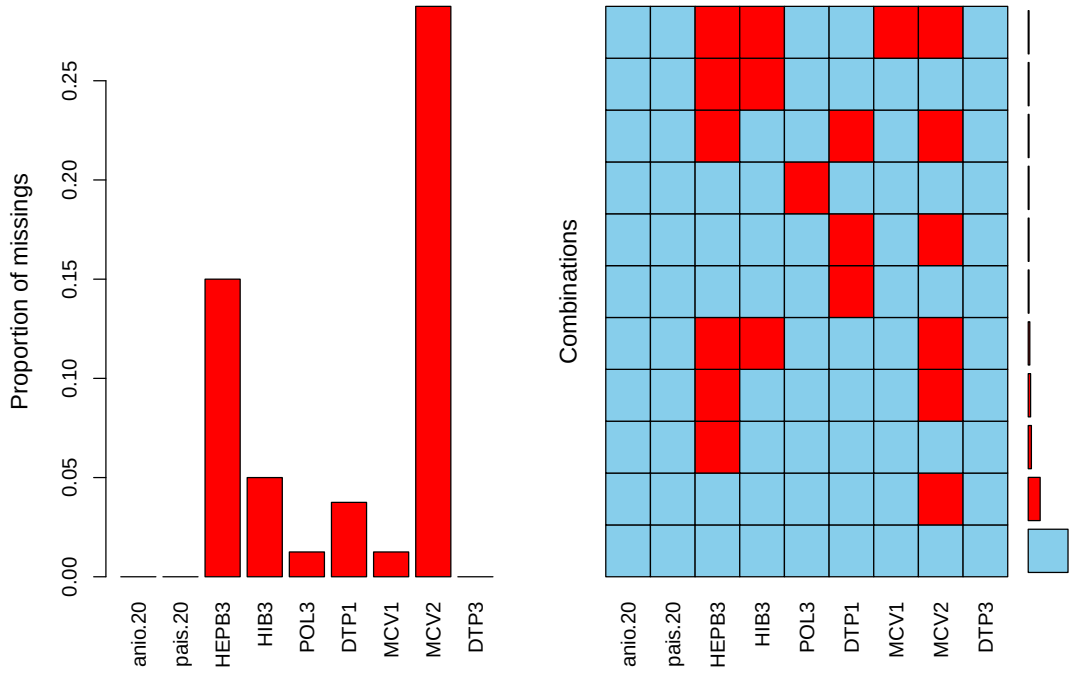

Figura 9: Patrones de respuesta para bloques de Vacunas.

La Figura 9 muestra, por un lado la proporción de datos faltantes (parte derecha) y por otro lado las distintas combinaciones de variables (vacunas en este caso) con datos faltantes y su proporción en el total (parte izquierda de la figura). A modo de entender mejor esta última, se detallan algunos casos. Hay observaciones (país-año) que no tienen ningún NA, por lo que la primer fila de la figura de la izquierda esta completa de rectángulos celestes, a su vez se observa que esto es lo que sucede en la mayoría de las observaciones ya que la barra del final de la fila es la de mayor tamaño. Le sigue la fila donde el único rectángulo rosado es en la variable MCV2, indicando que hay observaciones donde el único valor que falta es en la variable MCV2. En el otro extremo, se observa la última fila en la cual las variables HEPB3, HIB3, MCV1 y MCV2 están pintadas de rosado, lo que quiere decir que hay observaciones con datos faltantes en esas 4 variables, pero son los casos que menos se observan ya que la barra del final es la más chica.

Se muestra a continuación la cantidad de valores faltantes para cada una de las 7 vacunas. Como ya se vio en la sección de valores faltantes de vacunas, MCV2 es la que tiene más (23 observaciones).

```
##  Missings per variable:
##  Variable Count
##     HEPB3    12
##      HIB3     4
##      POL3     1
##      DTP1     3
##      MCV1     1
##      MCV2    23
##      DTP3     0
##  Missings in combinations of variables:
##  Combinations Count Percent
##  0:0:0:0:0:0:0    50   62.50
##  0:0:0:0:0:1:0    15   18.75
##  0:0:0:1:0:0:0     1    1.25
##  0:0:0:1:0:1:0     1    1.25
##  0:0:1:0:0:0:0     1    1.25
##  1:0:0:0:0:0:0     4    5.00
##  1:0:0:0:0:1:0     3    3.75
##  1:0:0:1:0:1:0     1    1.25
##  1:1:0:0:0:0:0     1    1.25
##  1:1:0:0:0:1:0     2    2.50
##  1:1:0:0:1:1:0     1    1.25
```

Esta última tabla, refleja lo mostrado en la Figura anterior pero agrega la información de cuantas observaciones tienen valores faltantes en determinado conjunto de vacunas.

Hay 50 observaciones (país-año) que no tienen NA en ninguna observación. Le siguen 15 observaciones que sólo tienen NA en la columna 8, teniendo en cuenta la figura anterior, se ve que corresponde a la vacuna MCV2.

Aquí se muestra para cada variable, el promedio de datos consecutivos con NA. En HEPB3, en promedio hay 3 observaciones seguidas sin dato.

```
#    HEPB3   3
#    HIB3    2
#    POL3    1
#    DTP1    1.5
#    MCV1    1
#    MCV2    3.83
#    DTP3    0
```

2.2.8. Imputación por Hotdeck para BRV

Partiendo de los datos, donde ya se identificaron los valores faltantes, luego de aplicar el método de *hotdeck* los resultados son:

Cuadro 1: Valores imutados por método hotdeck.

Tasa	n	media	min	max	Estado
HEPB3	68	88.40	66	96	Orig.
HEPB3	80	88.67	66	96	Impu
HIB3	76	89.34	55	98	Orig.
HIB3	80	89.06	55	98	Impu
POL3	79	89.65	71	9	Orig.
POL3	80	89.61	71	97	Impu
DTP1	77	91.32	67	10	Orig.
DTP1	80	91.30	67	100	Impu
MCV1	79	90.91	66	10	Orig.
MCV1	80	90.97	66	100	Impu
MCV2	57	81.40	28	10	Orig.
MCV2	80	83.89	28	100	Impu
DTP3	80	89.92	72	9	Orig.
DTP3	80	89.92	72	98	Impu

2.2.9. Imputación por kNN

Se imputa por el método **kNN**, fijando un radio de 5 para el tamaño de los grupos, es decir se consideran las distancias entre los 4 vecinos más cercanos.Al igual que el método hot-deck, el método de k vecinos más cercanos se basa en la observación de valores donantes, es decir que utiliza una agregación de los k valores más cercanos como valor imputado.

Nuevamente, en la Tabla 2 se puede ver algunas estadísticas de las variables vacunas antes y después del proceso de imputación pero esta vez mediante *kNN*, donde se vuelve a observar que en general las medias no se modifican sustancialmente.

Cuadro 2: Valores imputados por método kNN.

Tasa	n	media	min	max	Estado
HEPB3	68	88.40	66	96	Orig.
HEPB3	80	88.79	66	96	Impu
HIB3	76	89.34	55	98	Orig.
HIB3	80	89.24	55	98	Impu
POL3	79	89.65	71	9	Orig.
POL3	80	89.45	71	97	Impu
DTP1	77	91.32	67	10	Orig.
DTP1	80	91.54	67	100	Impu
MCV1	79	90.91	66	10	Orig.
MCV1	80	90.96	66	100	Impu
MCV2	57	81.40	28	10	Orig.
MCV2	80	83.06	28	100	Impu
DTP3	80	89.92	72	9	Orig.
DTP3	80	89.92	72	98	Impu

En la Figura 10 que sigue se puede ver la comparación de la imputación para la tasa **MCV2** usando 2 métodos univariados.

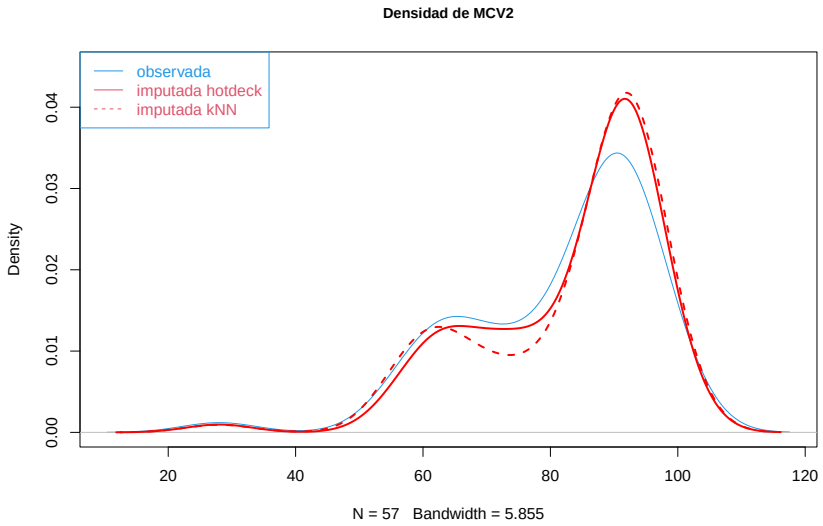

Figura 10: Comparación de densidades globales para **MCV2** imputación univariada.

En la Figura 11 que sigue se puede ver la comparación de la imputación para la tasa **HEPB3** usando 2 métodos univariados.

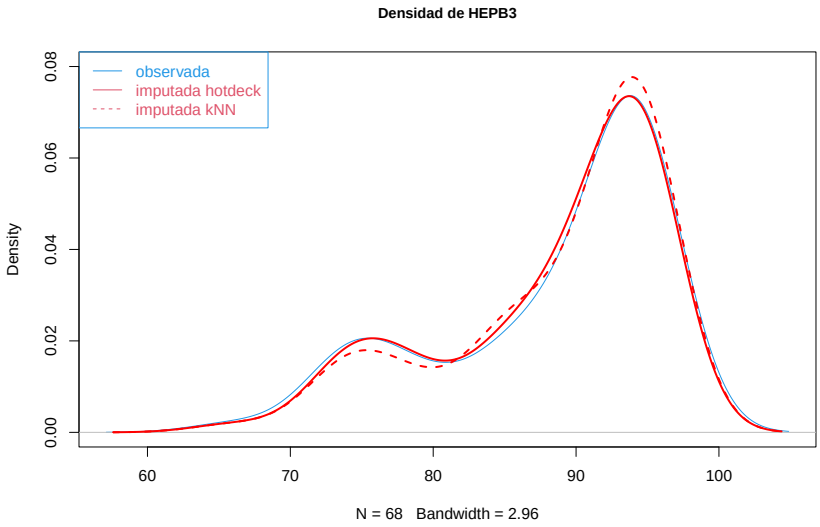

Figura 11: Comparación de densidades globales para **HEPB3** imputación univariada.

2.2.10. Imputación multivariada para MCV2

Antes de hacer el procedimiento de imputación multivariada con la libreria Amelia, hay que fijar la restricción de acotar las tasas al recorrido $[0, 100]$.

En la Figura 12 se muestra la densidad observada (color azul) y estimada (color rojo) para la variable *MCV2* que era la que tenía mayor cantidad de datos a imputar. Luego, se observan los gráficos de los valores de los datos observados (color negro) y los imputados junto con su intervalo de confianza (color rojo), en cada uno de los 4 países, para la variable mencionada, *MCV2*.

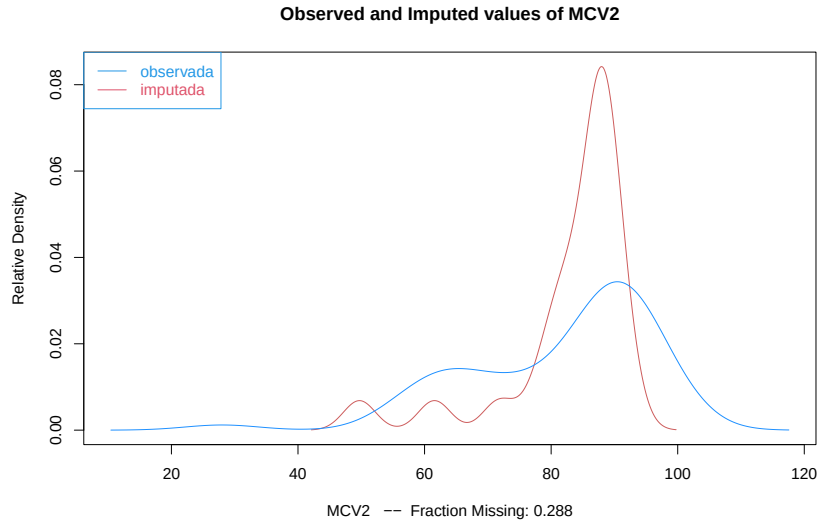

Figura 12: Comparación de densidades globales para **MCV2** imputación múltiple.

Recordando que las tasas de vacunación que componen el bloque *BRV* se pueden ver como datos de panel, es importante ver el resultado que produce la imputación multivariada, por lo cual se muestran en lugar de las densidades, los valores imputados para cada país, pero tomando como que las 7 tasas son una variable aleatoria multivariada de dimensión $p = 7$ y cuya estructura se incorpora al paso del tiempo. Para eso cuando hay una imputación aparece un segmento rojo que no solo representa cuando hay un valor imputado, sino una cota del error de imputación a través de un intervalo de confianza, como aparece en las figuras 13, 14, 15.

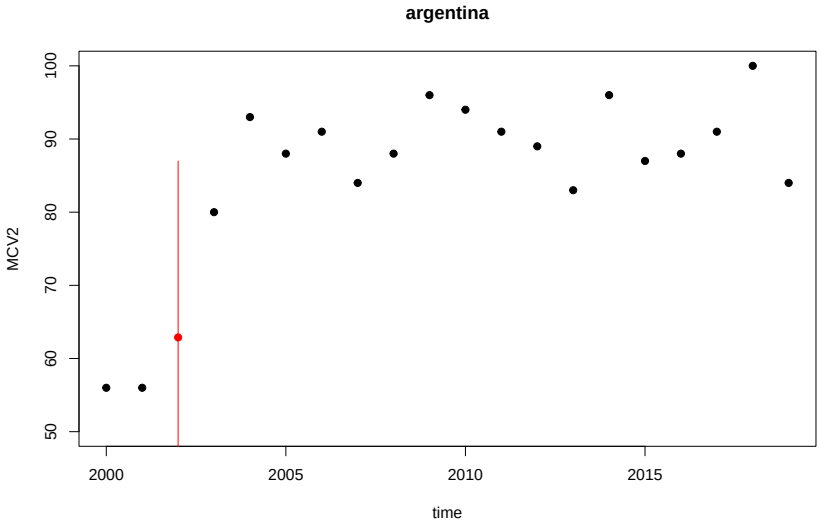

Figura 13: Imputación para **MCV2** para Argentina.

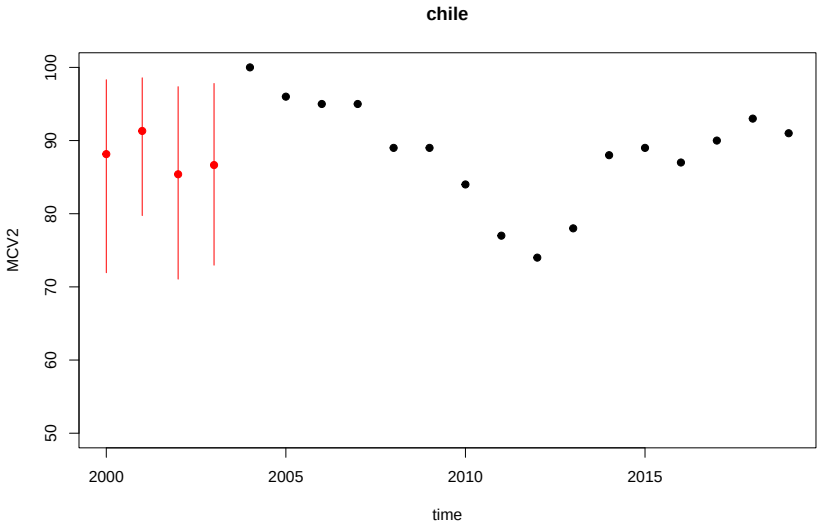

Figura 14: Imputación para **MCV2** para Chile.

Finalmente, se muestran las densidades observadas (color azul) y estimadas (color rojo) para las restantes 6 variables imputadas del bloque vacunas.

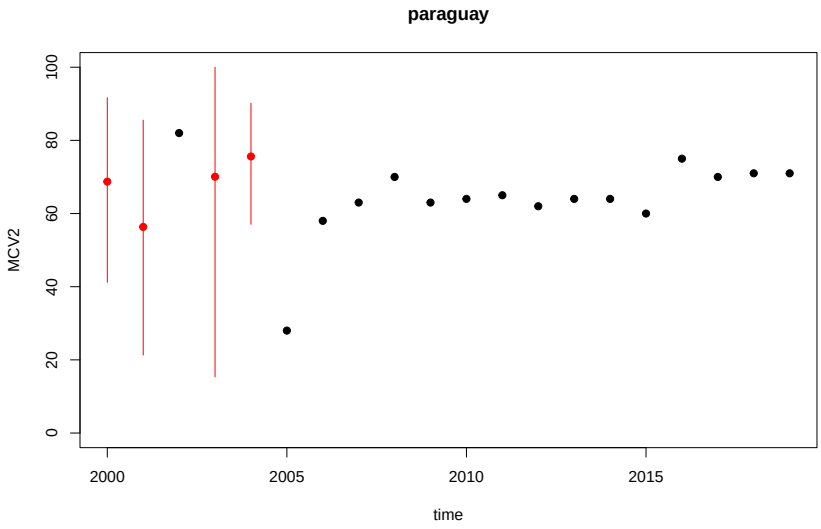

Figura 15: Imputación para **MCV2** para Paraguay.

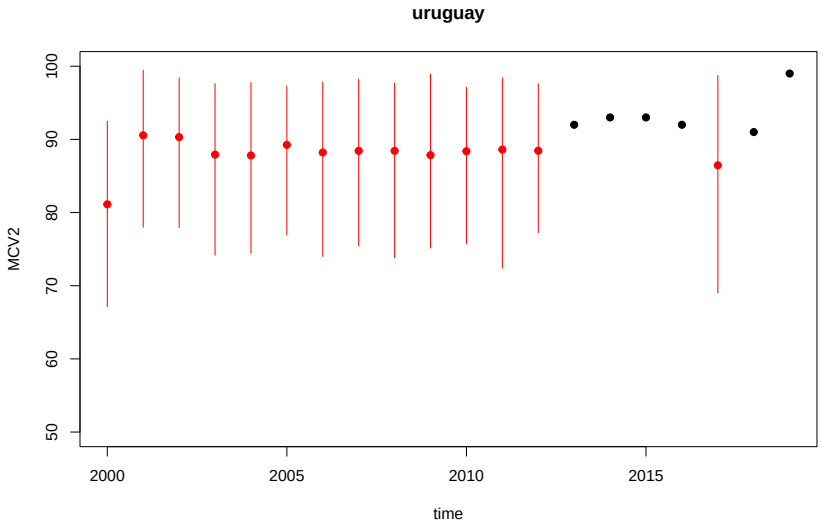

Figura 16: Imputación para **MCV2** para Uruguay

2.2.11. Imputación multivariada para HEPB3

Dado que *HEPB3* es la segunda vacuna, junto con *MCV2*, con más cantidad de datos faltantes, se evalúa hacer la imputación multivariada de la misma en la presente sección. En este caso puede verse en la Figura 17 que la densidad de la variable imputada, sigue siendo bimodal con un corrimiento hacia el centro, donde se ve el peso que tiene Paraguay.

Antes de decidir con cual de los métodos trabajar, se presentan las comparaciones de las tasas imputadas por Hotdeck y kNN, en las figuras 22 y 23 para MCV2 y HEPB3 respectivamente.

Estas a su vez se deben comparar con las figuras 13, 14, 15 y 16 para MCV2 y 18, 19, 20 y 21 para HEPB3, las cuales muestran la imputación multivariada.

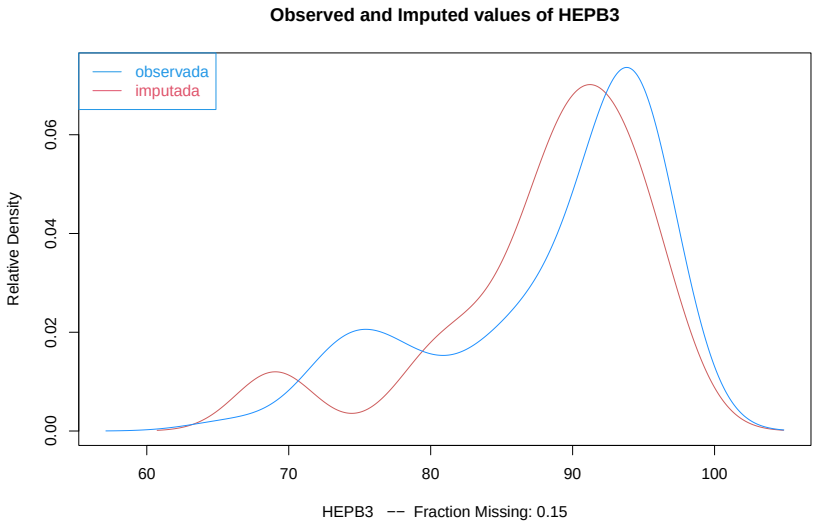

Figura 17: Comparación de densidades globales para **HEPB3** imputación múltiple.

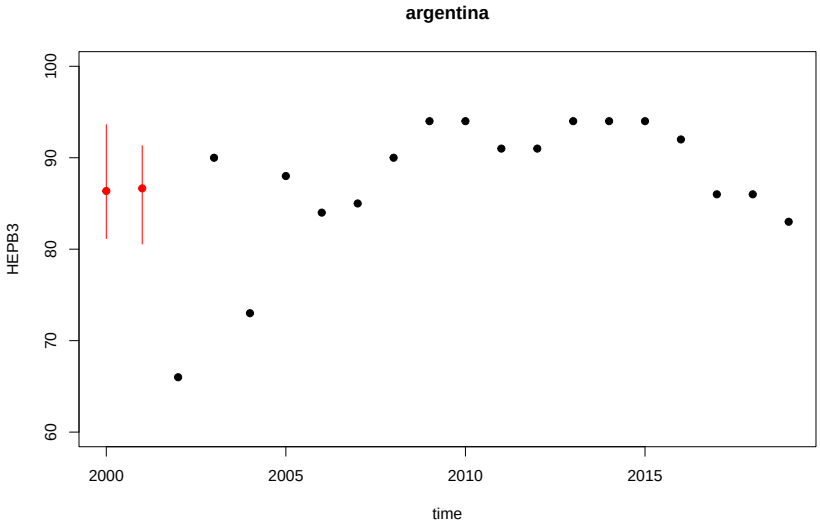

Figura 18: Imputación para **HEPB3** para Uruguay.

3. Conclusiones y pasos a seguir con la imputación de BRV

Teniendo en cuenta las densidades de las imputaciones univariadas, que aparecen en las figuras 10 y 11 hechas con los métodos hotdeck y kNN, se encuentra que:

- Para el caso de **HEPB3** el ajuste es bastante bueno, mientras que para **MCV2** (variable con mas cantidad de datos faltantes) ambos métodos tienden a acentuar los 2 modos de la densidad.
- Cuando se evalúa la performance del método multivariado se ve en las Figuras 13 y 16 que el ajuste entre lo observado y lo imputado es más bajo. Sobre todo para el caso de **MCV2** que tiende a concentrar a nivel global los datos en un rango de $[80, 100]$. Si se evalúa como queda a nivel de cada país para el caso de **MCV2**, dado que son en los extremos de la serie y a comienzos de la misma para Chile y Paraguay, los valores imputados oscilan con mucha variabilidad en el rango $[70, 100]$ y donde Uruguay tiene un valor imputado en media que casi no varía en los primeros 10 años.
- Por otra parte los valores imputados mediante el método multivariado quedan posicionados en cada serie, rompiendo para alguna tasa la tendencia con la que viene fluctuando, tal es el caso para Argentina para *HEPB3*.

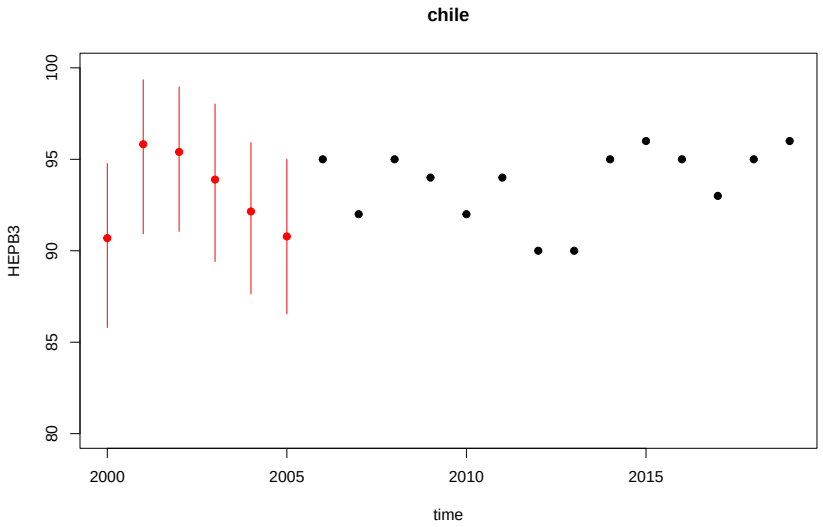

Figura 19: Imputación para **HEPB3** para Chile.

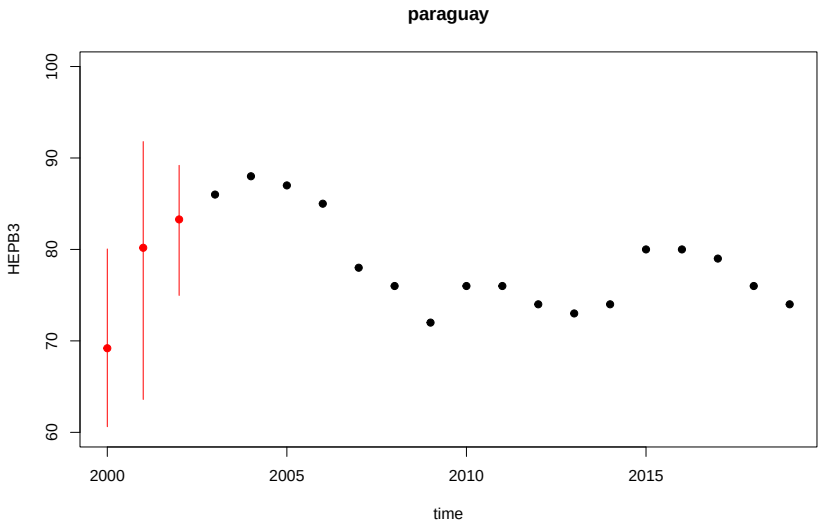

Figura 20: Imputación para **HEPB3** para Paraguay.

- Para el caso de **HEPB3** al ser menor la cantidad de datos faltantes, la variabilidad local (por país) también baja y donde el aspecto a rescatar es que en los extremos de la serie al comienzo del período, en la imputación para Paraguay crece la tasa, mientras que en Chile se ve un descenso de la tasa.

Los resultados para el mecanismo de imputación multivariada, pueden deberse a estas posibles causas:

- El método descansa en el supuesto de multinormalidad de las variables.
- Es insuficiente la cantidad de observaciones (n = 20) por país.
- Dado que en este caso las variables a imputar son porcentajes claramente no pueden conformar una distribución *multinormal*, salvo que se haga una transformación.

Por estos motivos y teniendo en cuenta las comparaciones que surgen de las tasas imputadas, se opta por considerar la siguiente estrategia:

- Comenzar imputando mediante el mecanismo de kNN, que es el que parece generar menos oscilaciones en el período.
- Luego imputar mediante Hotdeck y Amelia.

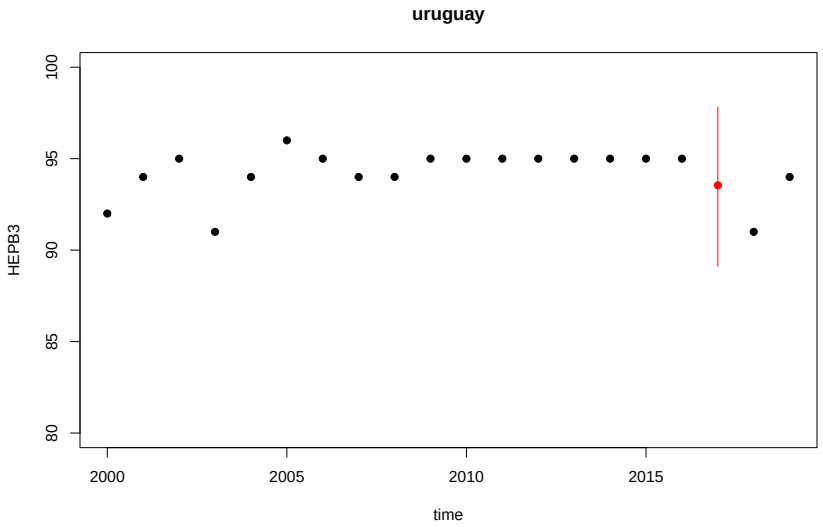

Figura 21: Imputación para **HEPB3** para Uruguay.

- Finalmente comparar los resultados de los clústeres obtenidos según cada uno de los 3 método propuestos.

- Otra opción a considerar es generar una matriz de datos producto de la combinación de métodos, es decir utilizar el método que mejor se ajuste en cada variable y país.

.

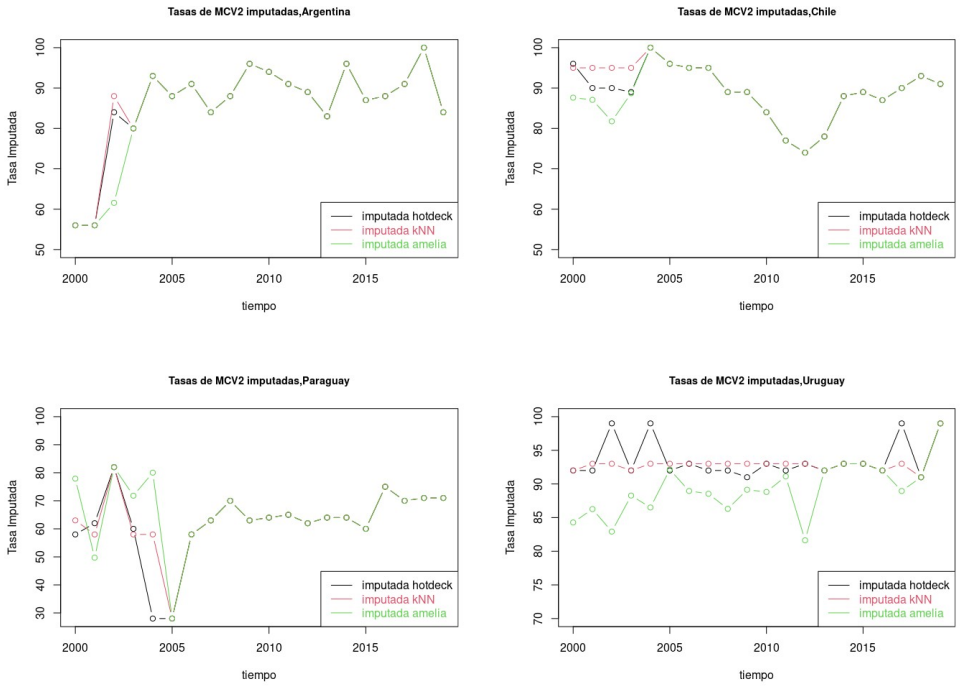

Figura 22: Comparación de series imputadas de MCV2

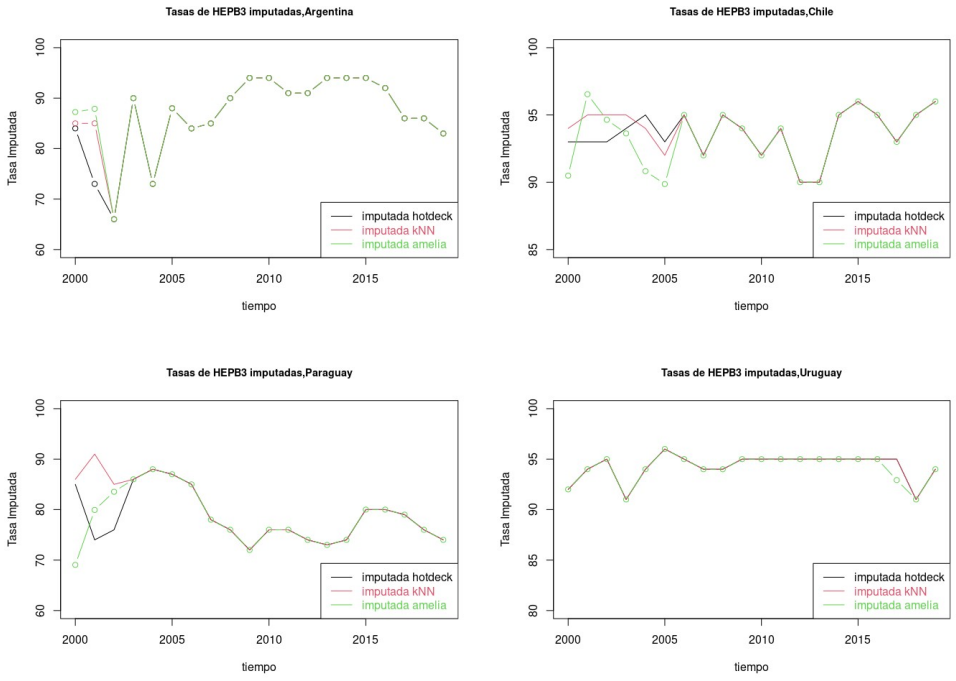

Figura 23: Comparación de series imputadas de HEPB3

Referencias

[1] Peng, R. D. 2009 . Reproducible research and Biostatistics, *Biostatistics* **10**(3): 405–408.
https://doi.org/10.1093/biostatistics/kxp014

[2] Gandrud, C. 2015 n.d. . *Reproducible Research with R and R Studio*.

[3] Gandrud, C., 2019 Reproducible research with r and rstudio.
https://github.com/christophergandrud/Rep-Res-Book

[4] Glennie, R. , 2021 Reproducible research with r and rstudio (3rd edition) by christopher gandrud, *Journal of Agricultural, Biological and Environmental Statistics* **26**(1): 128–129.
https://doi.org/10.1007/s13253-020-00418-y

[5] Kohrs, F. E.and Auer, S., Bannach-Brown, A., Fiedler, S., Haven, T. L., Heise, V. Weissgerber, T. L. 2023 . Eleven strategies for making reproducible research and open science training the norm at research institutions., *techreport*.

[6] R Core Team 2022 . *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
https://www.R-project.org/

[7] Revelle, W. 2022 . *psych: Procedures for Psychological, Psychometric, and Personality Research*, Northwestern University, Evanston, Illinois. R package version 2.2.9.
https://CRAN.R-project.org/package=psych

[8] RStudio Team 2020 . *RStudio: Integrated Development Environment for R*, RStudio, PBC., Boston, MA.
http://www.rstudio.com/

[9] Wickham, H. Bryan, J. 2022 . *readxl: Read Excel Files*. R package version 1.4.1.
https://CRAN.R-project.org/package=readxl

[10] Xie, Y. 2015 . *Dynamic Documents with R and knitr*, 2nd edn, Chapman and Hall/CRC, Boca Raton, Florida. ISBN 978-1498716963.
https://yihui.org/knitr/

[11] Tierney, N., Cook, D., McBain, M. Fay, C. 2021 . *naniar: Data Structures, Summaries, and Visualisations for Missing Data*. R package version 0.6.1.
https://CRAN.R-project.org/package=naniar

[12] Kowarik, A. Templ, M. 2016 . Imputation with the R package VIM, *Journal of Statistical Software* **74**(7): 1–16.

[13] Dempster, A., Laird, N. M. Rubin, D. 1977 . Maximum likelihood estimation from incomplete data via the em algorithm, *Journal of the Royal Statistical Society* **39**: 1?38.

[14] King, G., Tomz, M. Wittenberg, J. 2000 . Making the most of statistical analyses: Improving interpretation and presentation, *American Journal of Political Science* **44**(2): 341?55.

[15] Honaker, J. King, G. 2010 . What to do about missing values in time series cross-section data, *American Journal of Political Science* **54**(2): 561?81.

[16] Honaker, J., King, G. Blackwell, M. 2011 . Amelia II: A program for missing data, *Journal of Statistical Software* **45**(7): 1–47.

Este preprint fue presentado bajo las siguientes condiciones:

- Los autores declaran que son conscientes de que son los únicos responsables del contenido del preprint y que el depósito en SciELO Preprints no significa ningún compromiso por parte de SciELO, excepto su preservación y difusión.
- Los autores declaran que se obtuvieron los términos necesarios del consentimiento libre e informado de los participantes o pacientes en la investigación y se describen en el manuscrito, cuando corresponde.
- Los autores declaran que la preparación del manuscrito siguió las normas éticas de comunicación científica.
- Los autores declaran que los datos, las aplicaciones y otros contenidos subyacentes al manuscrito están referenciados.
- El manuscrito depositado está en formato PDF.
- Los autores declaran que la investigación que dio origen al manuscrito siguió buenas prácticas éticas y que las aprobaciones necesarias de los comités de ética de investigación, cuando corresponda, se describen en el manuscrito.
- Los autores declaran que una vez que un manuscrito es postado en el servidor SciELO Preprints, sólo puede ser retirado mediante solicitud a la Secretaría Editorial deSciELO Preprints, que publicará un aviso de retracción en su lugar.
- Los autores aceptan que el manuscrito aprobado esté disponible bajo licencia Creative Commons CC-BY.
- El autor que presenta el manuscrito declara que las contribuciones de todos los autores y la declaración de conflicto de intereses se incluyen explícitamente y en secciones específicas del manuscrito.
- Los autores declaran que el manuscrito no fue depositado y/o previamente puesto a disposición en otro servidor de preprints o publicado en una revista.
- Si el manuscrito está siendo evaluado o siendo preparando para su publicación pero aún no ha sido publicado por una revista, los autores declaran que han recibido autorización de la revista para hacer este depósito.
- El autor que envía el manuscrito declara que todos los autores del mismo están de acuerdo con el envío a SciELO Preprints.