Bussi, Javier
Hernández, Lucia
Marí, Gonzalo
Méndez, Fernanda
Mitas, Gerardo
Instituto de investigación Teórica y Aplicada, Escuela de Estadística

VISUALIZACIÓN Y MÉTODOS DE IMPUTACIÓN DE DATOS FALTANTES EN LA ENCUESTA DE GASTO DE LOS HOGARES[1]

Resumen:

La presencia de no respuesta es una de las principales dificultades que se presentan en las encuestas. La no respuesta puede ser total o parcial, existiendo una variedad de soluciones dependiendo de la situación. Si la no respuesta es parcial, la imputación de los valores perdidos es una opción usualmente utilizada. En el año 2018, se propuso en el Instituto Nacional de Estadística y Censos (INDEC), una revisión de los métodos de imputación aplicados a la Encuesta Nacional de Gastos de los Hogares (ENGHo). El método missForest (Stekhoven y Bühlmann, 2012) es un método de imputación no paramétrico cuyo algoritmo consiste en un proceso iterativo que asigna valores iniciales a los datos perdidos, construye un Forest ajustado el cual permite predecir nuevos datos imputados para cada una de las variables involucradas, y repite este procedimiento hasta su convergencia. En este trabajo se compara este método de imputación con otros métodos sugeridos en la bibliografía aplicados a los datos obtenidos en la ENGHo 2017-2018. Los métodos incluidos en la comparación son: Random Hot Deck (RHD), Vecino más cercano (VMC), Algoritmo Expectation-Maximization (EM), Amelia y Mice. Se determinó que la pérdida podía ser considerada completamente al azar, siendo este patrón uno de los escenarios planteados. Por otra parte, se consideró otro esquema de pérdida en la variable de interés basado en la variable estrato de áreas. Bajo ambos patrones de pérdida, se consideraron distintos porcentajes de valores perdidos. En todos los escenarios planteados, el método iterativo missForest presentó valores de Error Cuadrático Medio Normalizado (NRMSE) inferiores a los competidores, siendo el método Mice el que obtuvo valores similares, si bien en todos los casos levemente superiores. Con respecto a los tiempos de procesamiento, este último método presentó tiempos promedios muy superiores al resto de los métodos, siendo el missForest claramente el segundo método con tiempos promedios de cómputo más altos, pero aun así notablemente inferiores a los del Mice.

[1] Este trabajo se elaboró en el marco del Proyecto 1ECO199 titulado “Métodos Estadísticos en el Ámbito Oficial”, dirigido por Gonzalo Marí

Palabras claves: Encuesta de Gastos de los Hogares, Imputación, MissForest

Abstract:

The Non-Response in surveys is one of their major issues. Non-Response could be total or partial, with a variety of solutions depending on each situation in particular. If the Non-Response is partial, imputation of missing data is a method widely used. In 2018, the National Institute of Statistics and Censuses (INDEC) proposed a revision of the methods applied to the Household Expenditure Survey (ENGHo). The missForest is a nonparametric method of imputation in which the algorithm used is an iterative process that assigns initial values to the missing data, fits a random forest for each variable based on the observed values predicting new imputed observations until convergence. In this work, this method is compared to other methods reccomended for imputation in the bibliography. These methods are applied to data from the ENGHo 2017-1018. The methods considered for the comparison were: Random Hot Deck (RHD), Nearest Neighbor (NN), Expectation-Maximization Algorithm (EM), Amelia and Mice. It was determined that the values could be missing completely at random, and this type of pattern was one of the two scenarios considered for the comparison. In the second scenario considered, the probability of missing data depends on the stratum where the unit belongs. In both scenarios the Normal Root Square Mean Error (NRSME) for the missForest method was lower in comparison to all the competitors, being the Mice method the one that produced similar values but always slightly higher. With respect to computational processing times, the Mice method presented much higher average values in comparison to the other methods, being the missForest the second method with higher average processing times, but nonetheless, notably lower than those of Mice.

Keywords: Household Expenditure Survey, Imputation, missForest

1. Introducción

La presencia de no respuesta es una de las principales dificultades que se presentan en las encuestas, cuya magnitud parece haberse incrementado en los últimos años, por ejemplo, en algunos países de Europa (Beullens et al 2018). La no respuesta puede ser total o parcial, existiendo una variedad de soluciones dependiendo de la situación. En el caso de no respuesta total, una solución posible sería el ajuste de los ponderadores muestrales. Si la no respuesta es parcial, la imputación de los valores perdidos es una opción usualmente utilizada, que disminuye el impacto de la presencia de datos perdidos. Entre los procedimientos que usualmente se utilizan, se pueden mencionar los métodos probabilísticos, como el hot-deck, que producen valores imputados distintos en cada repetición del proceso de imputación, o los métodos determinísticos, como por ejemplo utilizando modelos de regresión, los cuales producen los mismos valores ante la repetición del proceso). En ambos casos, es usual que estos métodos sean aplicados en grupos de imputación compuestos por unidades con características similares. Los mismos pueden surgir a partir del uso de variables de

clasificación, por ejemplo, características demográficas en el caso de personas, o determinados por métodos de clasificación basados en procedimientos que dependen del tipo de variables que se incluyen en el análisis. Un ejemplo de estos métodos se da con el uso de árboles de clasificación o de regresión. En el año 2018, se propuso en el Instituto Nacional de Estadística y Censos (INDEC) una revisión de los métodos de imputación aplicados a la Encuesta Nacional de Gastos en los Hogares (ENGHo). En la revisión bibliográfica referida a las propuestas más recientes sobre los distintos métodos que presentaran características superadoras se prestó especial atención a un método iterativo de imputación llamado MissForest propuesto por D. J. Stekhoven y P. Bühlmann (2012), que se basa en la metodología de los Random Forests (Breiman 2001). Este método no paramétrico consiste en un proceso iterativo que asigna valores iniciales a los datos perdidos, construye un Forest ajustado el cual permite predecir nuevos datos imputados para cada una de las variables involucradas, y repite este procedimiento hasta su convergencia. Entre las principales ventajas de esta metodología, se encuentra el hecho de que la misma permite trabajar en forma simultánea tanto con variables cualitativas como cuantitativas. En una etapa inicial se pudo estudiar las cualidades de esta metodología (Bussi et al., 2018) aplicada a datos de la ENGHo 2004-2005 y determinar su potencial como método de imputación en encuestas similares en el futuro. El objetivo de este trabajo es comparar este método de imputación con otros métodos sugeridos en la bibliografía aplicados a los datos obtenidos en la ENGHo 2017-2018. Los métodos incluidos en la comparación son: Random Hot Deck (RHD), Vecino más cercano (VMC), Algoritmo Expectation-Maximization (EM), Amelia y Mice.

2. Metodología

Se presenta una descripción de la metodología del método missForest, y una breve descripción de cada uno de los métodos de imputación que se utilizarán para evaluar el desempeño del mismo.

Uno de las herramientas que permite la imputación de un conjunto de valores perdidos es la propuesta por Stekhoven y Bühlmann (2012), denominada *missForest*. El mismo es un procedimiento de imputación iterativo que se basa en la técnica *Random Forest* (Breiman, 2001), la cual consiste en la combinación de predictores basados en árboles de clasificación o de regresión que se construyen a partir de muestras independientes seleccionadas del conjunto de datos inicial. A continuación, se realiza una breve introducción de la técnica *Random Forest.*

2.1. Random Forest (RF)

La idea por detrás del método es promediar un conjunto de modelos aproximadamente insesgados, pero con cierta inestabilidad de forma tal de reducir la variancia. En particular, los árboles son ejemplos ideales de dichos modelos dado que logran captar la estructura compleja de los datos, tienen un gran nivel de desagregación y poseen un sesgo pequeño. Por otra parte, la inestabilidad de los mismos constituye una desventaja cuyo efecto se ve reducido por el promedio de un número grande de ellos.

El algoritmo consta de los siguientes pasos:

1) Para $b = 1, \dots, B$:
   a) Seleccionar una muestra bootstrap $\mathbf{Z}^*$ de tamaño N de los datos de entrenamiento
   b) Crear un árbol T_b del *random forest* a los datos bootstrap, repitiendo los siguientes pasos en cada nodo terminal del árbol, hasta que se alcance n_{min}, tamaño mínimo de nodo.
      i) Seleccionar m variables al azar de las p variables originales
      ii) Escoger el mejor punto de división entre las m variables
      iii) Dividir el nodo en dos nodos hijos
2) Obtener el conjunto de árboles $\{T_b\}_{b=1}^{B}$

Una predicción para un x dado se realiza de la siguiente forma:

Regresión: $\hat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$

Clasificación: $\hat{C}_{rf}^{B}(x) = modo\{\hat{C}_b(x)\}_{b=1}^{B}$, donde $\hat{C}_b(x)$ es la clase de predicción del *b* árbol del *random forest*.

Debido a que cada uno de los árboles generados se distribuyen idénticamente, la esperanza del promedio de *B* árboles es la misma que la esperanza de cada uno de ellos, por lo tanto, el sesgo del promedio es similar al sesgo de cada árbol individual, con lo cual el método asegura la reducción de la variancia. Por otra parte, no se puede asegurar la independencia entre los pares de árboles, a menos que el número de variables a seleccionar en cada división de un nodo, *m*, sea un número pequeño.

2.2. Procedimiento missForest

Este procedimiento de imputación hace uso del método *random forest*, lo cual permite trabajar con cualquier tipo de variables, tanto cuantitativas como cualitativas, disminuye los supuestos a realizar sobre las estructuras de los conjuntos de datos. Es un procedimiento iterativo que permite predecir valores perdidos hasta lograr la convergencia una vez que no se observen diferencias entre los distintos pasos del proceso. La herramienta RF es la elegida debido a su precisión y robustez, además de permitir la estimación de tasas de error *out-of-bag* (OOB) que permiten evaluar el procedimiento sin necesidad de contar con conjuntos de datos completos para realizar esa evaluación.

Sea un conjunto de datos $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_p)$**,** una matriz de datos de dimensión $n \times p$. Se divide la misma en cuatro partes teniendo en cuenta la existencia de valores perdidos en una variable arbitraria $\mathbf{X}_s$ en las posiciones $\mathbf{i}_{mis}^{(s)} \subseteq \{1,2, \dots, n\}$

1. $\mathbf{y}_{obs}^{(s)}$: los valores observados de la variable $\mathbf{X}_s$**.**
2. $\mathbf{y}_{mis}^{(s)}$: los valores perdidos de la variable $\mathbf{X}_s$**.**

3. $\mathbf{x}_{obs}^{(s)}$: variables distintas de $\mathbf{X}_s$ con observaciones $\mathbf{i}_{obs}^{(s)} = \{1,2,\dots,n\} \backslash \mathbf{i}_{mis}^{(s)}$.
4. $\mathbf{x}_{mis}^{(s)}$: variables distintas de $\mathbf{X}_s$ con observaciones $\mathbf{i}_{mis}^{(s)}$.

Luego, el procedimiento *missForest* se puede sintetizar en los siguientes pasos:

1. Realizar una primera imputación utilizando el método por la media o cualquier otro método, generando $\mathbf{X}^{\boldsymbol{imp}}$
2. Ordenar las variables $X_s, s = 1,\dots,p$, de acuerdo a la cantidad de valores perdidos, comenzando por las cantidades más pequeñas
3. Asignar en $\mathbf{X}_{old}^{\boldsymbol{imp}}$ la matriz de datos imputados
4. Para cada variable X_s ajustar un RF con respuesta $y_{obs}^{(s)}$ y predictores $x_{obs}^{(s)}$
5. Predecir los valores perdidos $y_{mis}^{(s)}$ aplicando el RF a $x_{mis}^{(s)}$
6. Actualizar $\mathbf{X}^{\boldsymbol{imp}}$ reemplazando los valores imputados anteriores por $y_{mis}^{(s)}$, generando $\mathbf{X}_{new}^{\boldsymbol{imp}}$
7. Repetir los pasos 3 a 6 hasta que se verifique el criterio de parada.

Se considera que el criterio de parada se cumple si la diferencia entre la matriz de datos imputados nueva y la anterior aumenta por primera vez con respecto a ambos tipos de variables, o sea, se debe verificar simultáneamente un aumento en las siguientes medidas

$$\Delta_{\mathrm{N}} = \frac{\sum_{j \in \mathbf{N}} \left(\mathbf{X}_{new}^{imp} - \mathbf{X}_{old}^{imp}\right)^2}{\sum_{j \in \mathbf{N}} \left(\mathbf{X}_{new}^{imp}\right)^2}$$

para el conjunto de **N** variables continuas, y

$$\Delta_{\mathrm{F}} = \frac{\sum_{j \in \mathbf{F}} \sum_{i=1}^{n} \mathbf{I}_{\mathbf{X}_{new}^{imp} \neq \mathbf{X}_{old}^{imp}}}{\#NA}$$

para las **F** variables categóricas, donde $\#NA$ es el número de valores perdidos en las variables categóricas.

2.3. Imputación de donantes

En la imputación del donante, el valor faltante en un registro se reemplaza con un valor observado que se copia de otro registro similar. El registro del cual se copia el valor se denomina registro "donante", de ahí, el nombre del método.

La ventaja de la imputación de donantes, comparada con la imputación basada en modelos, es que el valor imputado es siempre un valor realmente existente (observado). Con los modelos estadísticos siempre se corre el riesgo de predecir un valor que no es (físicamente) posible, especialmente al extrapolar más allá del rango de valores observado. La desventaja de la imputación del donante es que, a pesar de su amplia aplicación, la base teórica no es tan fuerte como para los métodos basados en modelos. Además, Andridge y Little (2010) concluyen en su extensa revisación que no existe consenso sobre la mejor manera de aplicar los métodos de

imputación de hot deck, y notan que "muchos métodos multivariados de hot deck parecen relativamente ad hoc". Sin embargo, los métodos hot-deck se han aplicado durante mucho tiempo en áreas relacionadas con estadísticas oficiales y en menor medida en medicina o entornos epidemiológicos.

En el método de Imputación hot deck aleatorio, el donante es seleccionado de un grupo de donantes. Específicamente, el conjunto de datos es separado en celdas de imputación para las cuales al menos una variable auxiliar tiene el mismo valor para el donante y el registro con valores perdidos. A medida que aumenta el número de variables que se utilizan para definir las celdas de imputación, decrece el tamaño de los grupos de donantes. Los donantes pueden ser elegidos aleatoriamente o se puede asignar una probabilidad de selección a cada uno. Las probabilidades serán reescaladas tanto como sea necesario dependiendo del agrupamiento y la especificación del grupo de donantes. En el método de imputación hot deck secuencial, el dataset se ordena de acuerdo a los valores de una o más variables y los valores perdidos en un registro se toman del primer registro anterior o posterior que tiene valor observado: LOCF, last observation carried forward o NOCB next observation carried backward. A su vez, cada uno de estos métodos puede ser ejecutado de forma univariada o multivariada.

El método de vecinos más cercanos utiliza una medida de similaridad para encontrar los k donantes de un registro que contiene valores perdidos. Luego, el valor donante se determina seleccionando aleatoriamente uno de los k vecinos o eligiendo el valor mayoritario, en el caso de variables categóricas. Una medida de similaridad particularmente conocida es la de Gower. Dados dos registros r y s, cada uno con p variables que pueden ser numéricas o categóricas (perdidas o no), la medida de similaridad de Gower puede ser expresada como:

$$d_g = \frac{\sum_{i=1}^{n} w_j \delta(r_j, s_j)}{\sum_{i=1}^{n} w_j}$$

Los valores w_j y δ dependen del tipo de variable. Si la j-ésima variable es numérica, entonces

$$\delta(r_j, s_j) = \begin{cases} 1 - \dfrac{|r_j - s_j|}{rango(j)} & \text{si } r_j \text{ y } s_j \text{ son observadas} \\ 0 & \text{en otro caso} \end{cases}$$

En este caso, $rango(j)$ es el rango observado de la variable j. Si la $j-$ésima variable es categóricas, entonces

$$\delta(r_j, s_j) = \begin{cases} 1 & \text{si } r_j = s_j \text{ y ambas son observadas} \\ 0 & \text{en otro caso} \end{cases}$$

Tanto para variables numéricas como categóricas, la importancia de los pesos w_j puede ser elegida como se quiera pero en general se utiliza 0 o 1 donde $w_j = 0$ equivale a excluir la j-ésima variable del cálculo del índice. Para variables dicotómicas, δ se define de otra forma:

$$\delta(r_j, s_j) = \begin{cases} r_j \wedge s_j & \text{si } r_j \text{ y } s_j \text{ son observadas} \\ 0 & \text{en otro caso} \end{cases}$$

$$w_j = r_j \vee s_j$$

Mientras que los pesos son definidos como

$$w_j = \begin{cases} r_j \vee s_j & \text{si } r_j \text{ y } s_j \text{ son observadas} \\ 0 & \text{en otro caso} \end{cases}$$

Las variables dicotómicas sólo suman a la similaridad si ambas son verdaderas. Si una sola de las dos variables es verdadera, ella suma al peso en el denominador.

2.4. Imputación basada en el algoritmo Expectation-Maximization (EM)

El objetivo del algoritmo EM es encontrar lo estimadores máximo-verosímiles para los parámetros de una distribución de probabilidad multivariada en la presencia de datos perdidos. Como tal, no es un método de imputación. Las imputaciones pueden generarse calculando el valor esperado para los datos perdidos sobre los datos observados o seleccionando una muestra de la distribución condicional multivariada. A diferencia de los métodos de imputación basados en modelos, en el algoritmo EM, no hay una distinción fija entre las variables predictoras y las predichas (se utilizan todas las disponibles para estimar los parámetros de alguna distribución). Esto significa que se necesita asumir una forma distribucional (por ejemplo, normal multivariada) para las variables en el conjunto de datos.

Las ventajas del algoritmo EM incluyen que es un método simple y bien comprendido que en principio converge. Además, dado que el algoritmo proporciona una estimación para la distribución multivariada completa, tiene una buena chance de corregir por el mecanismo aleatorio (MAR). Una desventaja es que el cálculo puede tomar muchas iteraciones ya que el criterio de convergencia (medido en términos de la diferencia en las estimaciones entre iteraciones) disminuye aproximadamente linealmente con el número de iteraciones (Schafer, 1997). La convergencia puede ser especialmente lenta cuando la fracción de valores perdidos es alta o cuando la distribución utilizada es una descripción deficiente de la distribución real de los datos. Por otra parte, el algoritmo EM no proporcionar una estimación de la variancia de las estimaciones.

En el año 2011, Honaker et al. propusieron un esquema para aleatorizar los parámetros de una distribución normal multivariada que denominaron EMB (bootstrapped EM). La idea es crear un conjunto de M bases de datos mediante remuestreo con reemplazo del conjunto de datos original, de manera que cada conjunto de datos en el conjunto tenga el mismo número de registros que el original. Luego, el algoritmo EM se aplica en cada uno de los M conjuntos de datos para encontrar las estimaciones máximo-verosímiles de los parámetros de la distribución (vector medio y matriz de covariancias). Condicional a la validez del esquema bootstrap de remuestreo, el ensemble resultante de parámetros normales multivariados estima su distribución muestral. Para cada base de datos del ensemble, cada registro puede ser

completado mediante un muestreo de la distribución normal multivariada condicional en los valores observados en el registro.

Los métodos de imputación basados en el algoritmo EM, ya sea una única imputación imputaciones múltiples, dependen de la capacidad de formular una distribución multivariada para los datos tratados de los cuales se seleccionan las imputaciones. Como alternativa, se puede formular un modelo de probabilidad separado para cada variable a imputar. Este método recibe el nombre de *multivariate imputation with chained equations* (Mice). Los parámetros del modelo y los valores para imputación son generados secuencialmente y aleatoriamente condicional a las variables conocidas (posiblemente previamente imputadas). La especificación puede ser totalmente condicional, en la que cada variable excepto la que actualmente se imputa, forme parte del modelo predictivo o no. La secuencia de imputaciones se repite hasta que ciertas propiedades de distribución han convergido. Luego se logra una imputación múltiple repitiendo todo este procedimiento M veces para crear M conjuntos de datos imputados. El término ecuaciones encadenadas se refiere al hecho de que en una secuencia donde las variables se imputan aleatoriamente, una por una, y condicionadas a las variables imputadas previamente, este condicionamiento introduce una cadena de dependencias en las distribuciones de probabilidad.

3. Materiales

Se cuenta con una base de datos preliminar correspondiente a la edición 2017/18 de la Encuesta de Gasto de los Hogares (ENGHo) que se está llevando a cabo bajo la coordinación del INDEC. Se determinó trabajar con la provincia de Santa Fe, y con el objetivo particular de imputar el ingreso de la actividad principal de los asalariados. Cabe destacar que en esta base preliminar, 805 personas se declaran como asalariados, siendo 66 el número de unidades que poseen un valor perdido en dicha variable.

Se determinan un conjunto de variables explicativas con las cuales se analizará el desempeño de los distintos métodos de imputación bajo los diversos escenarios que se plantean. A continuación, se brinda un detalle de las mismas:

Variables Cualitativas

- cp03_: Grupo de edad (en años cumplidos)
- cp04: Relación de parentesco
- cp13: Sexo
- sitconyugal: Situación Conyugal
- estado: Condición de actividad (EPH)
- cp36: tipo de negocio/institución (estatal/privado/otro)
- cp37: tipo de empleador
- cp49: tamaño del establecimiento
- nestrato: estrato de área

- tcomed2: condición de afiliado y tipo de cobertura médica
- catoc1: categoría ocupacional de la ocupación principal
- pubnived: nivel educativo
- califica: calificación ocupacional
- jer_ocup: jerarquía ocupacional
- rama_engh: rama de la actividad principal
- propauto: propietario de auto

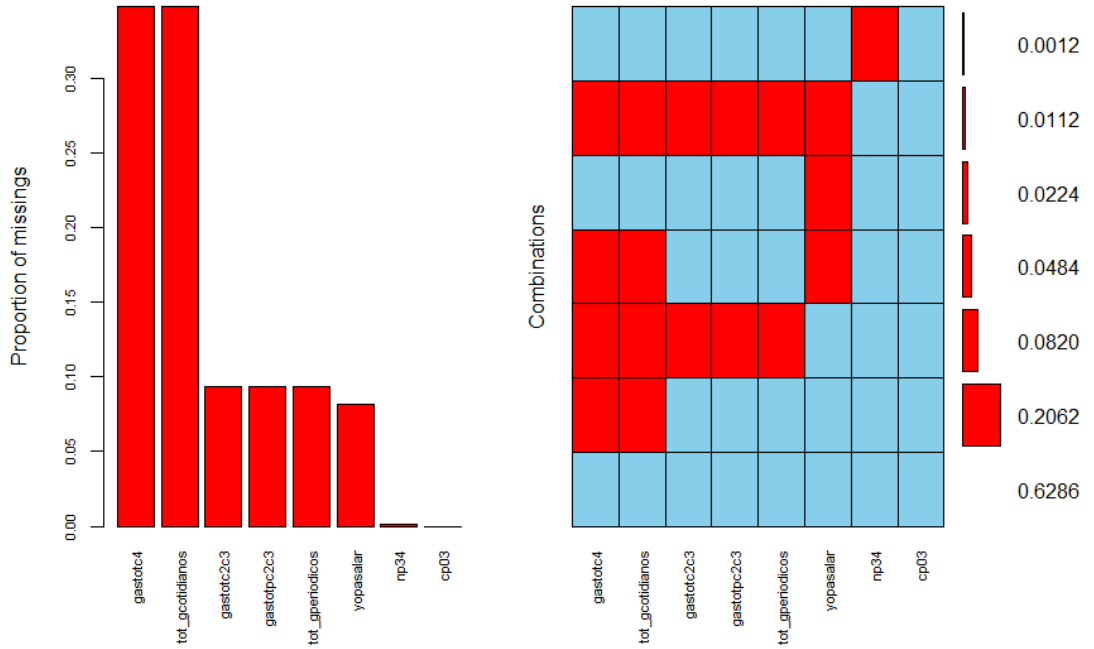
Variables Cuantitativas

- np34: Horas de trabajo de la ocupación principal
- cp36: tipo de negocio/institución (estatal/privado/otro)
- gastotc4: gasto total mensualizado del Cuestionario 4
- gastotc2c3: gasto total mensualizado del hogar del Cuestionario 2 y 3
- gastotpc2c3: gasto total mensualizado del hogar del Cuestionario 2 y 3 per cápita
- tot_gperiodicos: total gastos mensualizados periódicos del Cuestionario 3 de hogar
- tot_gcotidianos: total gastos mensualizados cotidianos del conjunto de Cuestionario 4
- cantmiem: cantidad de miembros del hogar
- cantperc: cantidad de perceptores del hogar

4. Visualización de datos faltantes en la ENGHo

El gráfico 1 muestra que las variables cuantitativas gastotc4 y tot_gcotidianos son las que presentan el mayor porcentaje de valores faltantes, mientras que la combinación gastotc4 y tot_gcotidianos es el patrón de pérdida más frecuente en la base de datos.

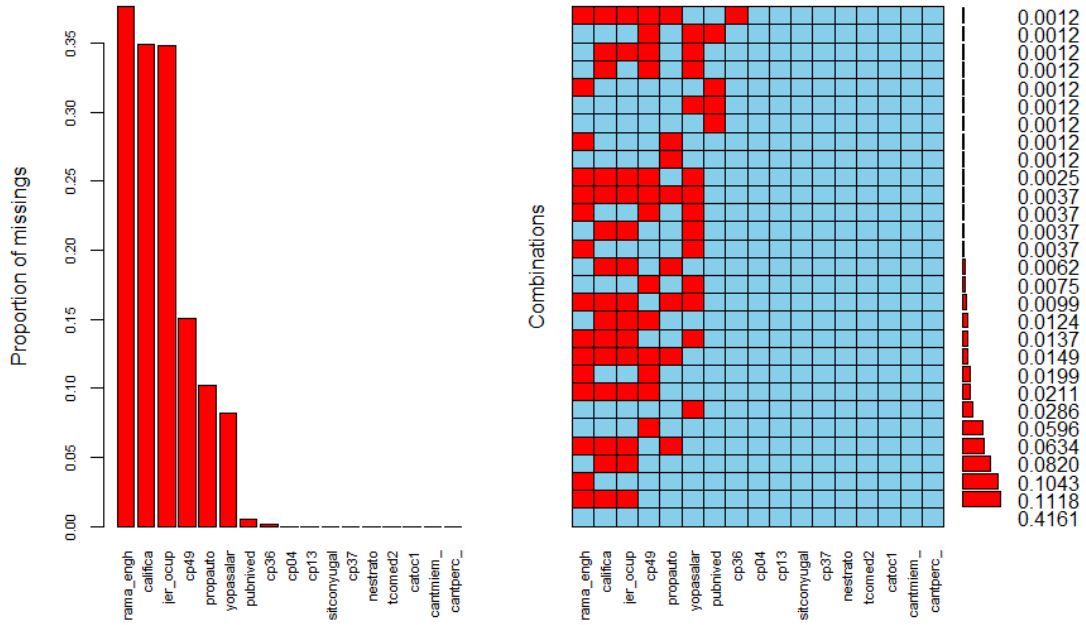
Gráfico 1 Proporción de valores *missing* en cada variable cuantitativa (izquierda) y ocurrencia de patrones de datos faltantes (derecha).



Fuente: INDEC.

Las variables cualitativas que presentan mayor porcentaje de valores ausentes son rama de la ocupación principal, calificación y jerarquía ocupacional.

Gráfico 2 Proporción de valores *missing* en cada variable cualitativa (izquierda) y ocurrencia de patrones de datos faltantes (derecha).



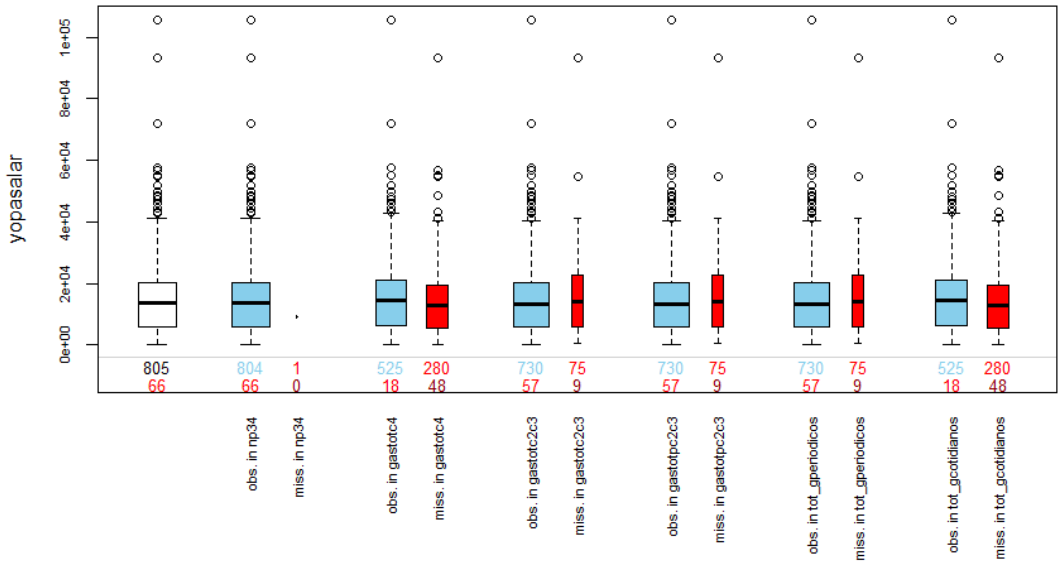
Fuente: INDEC.

En los gráficos 3 y 4 se compara la distribución del ingreso de la ocupación principal (variable yopasalar*)* con respecto a la ausencia o no de todas las otras variables.

El *boxplot* de la izquierda muestra la distribución del ingreso de la ocupación y los números indican que en la base de datos hay 805 observaciones de las cuales 66 son faltantes. Los otros diagramas de caja, que aparecen en pares, comparan las distribuciones de yopasalar, divididas de acuerdo con la pérdida o no en la otra variable.

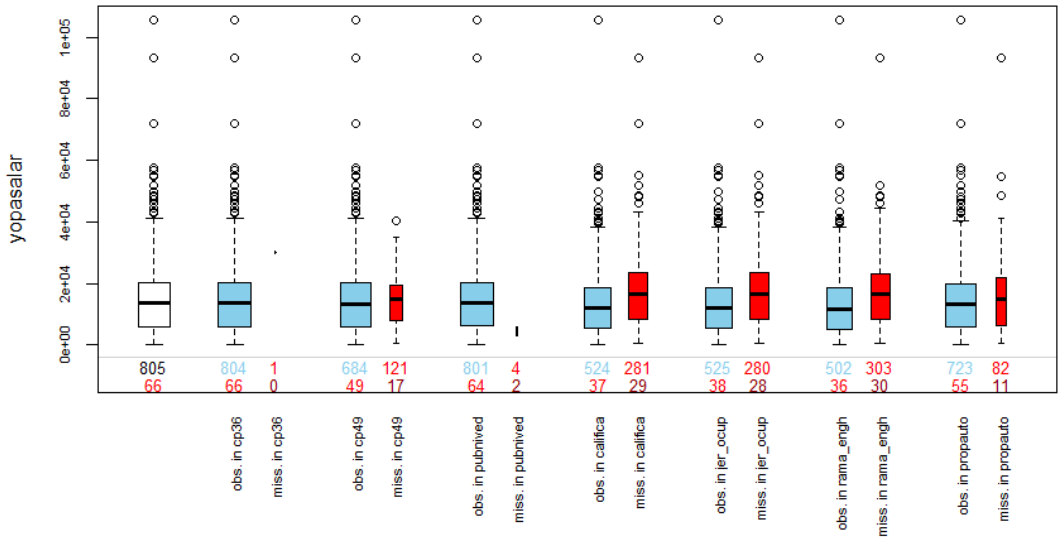
Los anchos de los diagramas de caja indican el número de observaciones utilizadas para realizar el boxplot. El número de observaciones (arriba) y los valores faltantes (abajo) por grupo se presentan debajo de las cajas.

Gráfico 3 Diagramas de cajas del ingreso de la ocupación principal condicionado a la presencia o no de valores faltantes en otras variables.



Fuente: INDEC.

Gráfico 4 Diagramas de cajas del ingreso de la ocupación principal condicionado a la presencia o no de valores faltantes en otras variables.



Fuente: INDEC.

Para confirmar o rechazar la sospecha de que las medianas de las distribuciones no difieren significativamente, se realiza la prueba de Kruskal-Wallis. Los valores de probabilidad asociada mostrados en la tabla 1 (variables cuantitativas) indican que el ingreso de la ocupación principal mediano en el caso en que se observa el gasto no parece diferir del caso en que falta el gasto.

Tabla 1

Variable	gastotc4	gastotc2c3	gastotpc2c3	tot_gperiodicos	tot_gcotidianos
p-value	0.1275	0.4909	0.4909	0.4909	0.1275

Fuente: INDEC.

Sin embargo, para algunas variables cualitativas, la presencia de valores perdidos depende de la magnitud de los valores de la variable ingreso de la ocupación principal (gráfico 4 y tabla 2). Se observa que los valores faltantes en las variables rama de la ocupación principal, calificación y jerarquía ocupacional ocurren predominantemente para individuos de mayor ingreso de la ocupación principal. Esta situación deberá ser estudiada con mayor detenimiento en estudios futuros.

Tabla 2

Variable	cp49	pubnived	califica	jer_ocup	rama_engh	propauto
p-value	0.7022	0.1102	0.0000	0.0000	0.0000	0.3362

Fuente: INDEC.

5. Estudio por Simulación

Para evaluar el método de imputación missForest, el cual está basado en la metodología Random Forest, se consideró el conjunto de observaciones de la base preliminar de la ENGHO 2017/18 formado por aquellas personas con ingreso de la ocupación principal declarado, o sea, sin observación missing. Cabe aclarar que el resto de las variables consideradas en el estudio pueden o no contenter valores perdidos.

Se consideraron dos esquemas de pérdida, uno considerando que la misma es completamente al azar (MCAR), y la otra reproduciendo los porcentajes de missing hallados en la base para la variable *nestrato* que corresponde al estrato de las áreas, que conforman un conglomerado de viviendas y que constituyen una de las unidades de muestreo. Cabe destacar que en teoría estos estratos discriminan a la población de áreas de acuerdo al nivel socioeconómico de la población.

Para cada uno de los esquemas, se consideraron distintos porcentajes de perdidas, 5%, 10%, 15% y 20%. Se repitió la generación de valores perdidos 100 veces en forma independiente para cada uno de los distintos porcentajes en la variable *ingreso de la actividad principal*. En cada una de las repeticiones y para cada uno de los métodos, se calculó el tiempo de procesamiento y la raiz cuadrada del error cuadrático medio normalizado (NRMSE) definido como

$$NRMSE = \sqrt{\frac{media((x^{true} - x^{imp})^2)}{var(x^{true})}}$$

donde x^{true} es el valor de la variable ingreso observada y x^{imp} es el valor imputado para la misma variable. Para realizar la evaluación y comparación entre los métodos, se consideró el promedio de los 100 tiempos y los 100 NRMSE calculados para cada método en cada uno de los escenarios.

Las variables consideradas como explicativas en cada uno de los métodos dependen de las posibilidades de los mismos. De esta forma, para el RHD se tuvo en cuenta la edad categorizada en 3 grupos con igual frecuencia en cada uno de ellos, el sexo y la cantidad de miembros del hogar (1, 2, 3, 4, 5 y más personas), lo que define 30 celdas de imputación. En el caso de los métodos missForest, VMC y Mice, se utilizaron como variables explicativas el

conjunto de variables cuantitativas y cualitativas descripto en la sección anterior, mientras que para los métodos EM y Amelia, por las carácterísticas de ambos métodos, sólo se utilizaron variables cuantitativas.

A continuación, se presentan los resultados de las simulaciones realizadas para cada uno de los esquemas de pérdida con los distintos porcentajes considerados, y para los métodos considerados.

Tabla 3. NRMSE promedio para distintos porcentajes de pérdida y métodos de imputación en esquema de pérdida MCAR

% Missing	NRMSE					
	missForest	**RHD**	**VMC**	**EM**	**Amelia**	**Mice**
5%	0.1486	0.2864	0.2049	0.1647	01685	0.1527
10%	0.2142	0.4038	0.2929	0.2407	0.2458	0.2187
15%	0.2617	0.5020	0.3648	0.2889	0.2964	0.2675
20%	0.3112	0.5905	0.4248	0.3419	0.3492	0.3168

Tabla 4. Tiempo promedio en segundos para distintos porcentajes de pérdida y métodos de imputación en esquema de pérdida MCAR

% Missing	Tiempo (en segundos)					
	missForest	**RHD**	**VMC**	**EM**	**Amelia**	**Mice**
5%	32.9227	0.0153	0.0121	0.0251	0.5448	246.4913
10%	32.1862	0.0238	0.0184	0.0226	0.5405	247.8758
15%	33.9386	0.0177	0.0244	0.0248	0.5379	249.3228
20%	32.1146	0.0210	0.0344	0.0259	0.5431	248.0310

Tabla 5. NRMSE promedio para distintos porcentajes de pérdida y métodos de imputación en esquema de pérdida por estrato

% Miss-ing	NRMSE					
	missForest	**RHD**	**VMC**	**EM**	**Amelia**	**Mice**
5%	0.1453	0.2765	0.1939	0.1636	0.1667	0.1498
10%	0.2166	0.4091	0.2989	0.2357	0.2407	0.2200
15%	0.2664	0.5016	0.3622	0.2895	0.2964	0.2696
20%	0.2987	0.5860	0.4099	0.3274	0.3346	0.3057

Tabla 6. Tiempo promedio en segundos para distintos porcentajes de pérdida y métodos de imputación en esquema de pérdida por estrato

% Miss-ing	Tiempo (en segundos)					
	missForest	**RHD**	**VMC**	**EM**	**Amelia**	**Mice**
5%	31.5734	0.0154	0.0128	0.0302	0.5429	246.6057
10%	32.5246	0.0172	0.0186	0.0236	0.5327	247.1970
15%	32.6942	0.0175	0.0268	0.0234	0.5364	248.4104
20%	31.5602	0.0182	0.0321	0.0264	0.5536	249.6488

Como se puede observar no existen diferencias entre los resultados encontrados para ambos esquemas de pérdida, con lo cual las conclusiones son válidas para ambas situaciones. Como es de esperar, en todos los métodos existe un aumento en los promedios de NRMSE, no así se observa cambios significativos en los tiempos promedios con la excepción del método VMC donde los tiempos aumentan considerablemente a medida que aumenta el porcentaje de valores perdidos. De todas formas, el tiempo de procesamiento del VMC no resulta ser un inconveniente ni aun con el mayor porcentaje de pérdida considerado.

En todos los escenarios planteados, el método missForest presenta valores promedios de NRMSE menores al resto de los métodos. El método que presenta valores cercanos, pero siempre superiores es el Mice. Los métodos basados en la generación de datos a partir de una distribución multivariada para variables cuantitativas (EM y Amelia) presentan resultados similares entre si, pero mayores que los métodos mencionados. Por último, cabe mencionar que la imputación por VMC y por RHD son los de valores de error más altos, siendo el último el que presenta los resultados más desfavorables.

Con respecto a los tiempos, los métodos que consideran una imputación en cadena de las variables consideradas (missForest y Mice) presentan tiempos muy superiores a los competidores, siendo el Mice el que presenta los mayores tiempos superando los 4 minutos de procesamiento, en promedio. El resto de los métodos se ejecutan en tiempos promedios inferiores al segundo.

6. Conclusiones

En la presente investigación se realizó el estudio de los distintos métodos de imputación aplicados a la variable ingreso de la actividad principal en una base preliminar de la Encuesta Nacional de Gasto de los Hogares 2017/18. El objetivo es el estudio de las bondades del método iterativo missForest y comparar el desempeño con métodos de imputación tradicionales bajo diversos escenarios. Para determinar los mismos, se estudió la distribución de las distintas variables bajo la presencia de valores perdidos, y se determinó que la pérdida podía ser considerada completamente al azar, siendo este patrón uno de los escenarios planteados. Por otra parte, se consideró otro esquema de pérdida en la variable de interés basado en la variable estrato de áreas. Bajo ambos patrones de pérdida, se consideraron distintos porcentajes de valores missing. En todos los escenarios planteados, el método iterativo missForest presentó valores de Error Cuadrático Medio Normalizado (NRMSE) inferiores a los competidores, siendo el método Mice el que obtuvo valores similares, si bien en todos los casos levemente superiores. Con respecto a los tiempos de procesamiento, éste último método presenta tiempos promedios muy superiores al resto de los métodos. Teniendo en cuenta ambos aspectos, se puede concluir que el método missForest resulta ser el más apropiado para las variables consideradas, quedando el estudio en el futuro de otros esquemas de pérdida y otras regiones del país donde se desarrolla la ENGHo.

REFERENCIAS BIBLIOGRÁFICAS

Andridge, R.R., Little, R.J. (2010). A Review of Hot Deck Imputation for Survey Non-response. *International Statistical Review,* 78, 40-64.

Beullens, K., Loosveldt G., Vandenplas C., Stoop I. (2018). Response Rates in the European Social Survey: Increasing, Decreasing, or a Matter of Fieldwork Efforts? *Survey Methods: Insights from the Field.* Recuperado de https://surveyinsights.org/?p=9673.

Breiman,L. (2001). Random forests*. Machine Learning, 45, 5-32.*

Bussi, J., Hernández, L., Marí, G., Méndez, F., Mitas, G. (2018). Imputación de la No Respuesta Utilizando el Procedimiento Random Forest. XLVI Coloquio Argentino de Estadística. 31 de Julio al 3 de Agosto de 2018. Río Cuarto, Córdoba, Argentina.

Honaker, J., King, G., Blackwell, M. (2011). Amelia II: A Program for Missing Data. *Journal of*

Statistical Software, 45, 1-47.

Misztal, M. (2013) Some Remarks on the Data Imputation using “Missforest” Method. *Acta Universitatis Lodziensis. Folia Oeconomica 285.*

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. CRC Press.

Stekhoven,D.; Bühlmann, P. (2012) MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics, 28, 112-118.*

Stekhoven,D. (2013). missForest: Nonparametric Missing Value Imputation using Random Forest. *R package version 1.4.*

Tang, F.; Ishwaran,H. (2017). Random Forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal, 10, 363-377.*

van der Loo, M; de Jonge, E. (2018). *Statistical Data Cleaning with Applications in R*. Wiley & Sons.

FUENTES

Instituto Nacional de Estadística y Censos (INDEC). Encuesta Nacional de Gasto de los Hogares 2017-18.