# IMPUTACIÓN MÚLTIPLE EN VARIABLES CATEGÓRICAS USANDO DATA AUGMENTATION Y ÁRBOLES DE CLASIFICACIÓN

Jorge Bacallao Guerra* y Jorge Bacallao Gallestey**
*Instituto de Cibernética, Matemática y Física, La Habana
**Centro de Investigaciones y Referencia de Aterosclerosis de La Habana

**RESUMEN**
Se presenta un método para resolver problemas de valores faltantes en datos categóricos. Es una variante de la Imputación Múltiple, que combina la acción de los Árboles de Clasificación (CA) y la implementación del algoritmo Data Augmentation para datos categóricos. Se describe el método y se compara a niveles teórico y práctico con los métodos que se utilizan actualmente en la literatura, a través del desarrollo de un ejemplo con una base ficticia utilizando la implementación en lenguaje R del método propuesto.

**ABSTRACT**
It is a modication of common multiple imputation algorithms which combines classification trees (CT) and data augmentation for categorical data. We describe the rationale of the method and compare it, on theoretical and practical grounds, with two of the most frequently used methods. We use a fictitious base and an “ad hoc" R-based software.

**KEY WORDS:** Multiple Imputation, categorical data, Data Augmentation, classification trees.

**MSC**: 62G08

## 1. INTRODUCCIÓN

En los últimos años, los procedimientos de imputación múltiple han demostrado ser la mejor opción para resolver problemas de datos faltantes. Dentro de los métodos de imputación múltiple, los de naturaleza bayesiana juegan un papel trascendental porque incorporan de manera natural la incertidumbre al proceso de inferencia. El problema fundamental de estos métodos consiste en su poca robustez y en una excesiva rigidez en la selección de las variables que intervienen en el análisis.

La inmensa mayoría de estos métodos están implementados sobre un modelo multinormal y han demostrado funcionar muy bien cuando los supuestos sobre el modelo se cumplen pero, en otras instancias, no son suficientes para resolver aspectos de vital importancia, como por ejemplo, conservar la estructura de los datos cuando intervienen variables modificadoras de efecto y/o confusoras. En especial, en contextos de datos categóricos, de poco sirven los métodos y softwares más conocidos. La propia naturaleza de los datos hace que no sean factibles alternativas como la sustitución del dato ausente por la media o la sustitución por un estimado de regresión. Ocurre también que los escasos métodos que son factibles en situaciones de datos categóricos, acusan una naturaleza univariada, y esto los hace solo utilizables en aquellas situaciones en que se pueda asegurar que el patrón de ausencias es MCAR (Missing Completly at Random)

El método que proponemos es una variante de imputación múltiple aplicable a datos categóricos. Es un procedimiento multivariado que utiliza como herramientas principales a los Árboles de Clasificación [3] y al algoritmo Data Augmentation [10,11], implementado para una variable categórica. Nuestro método tiene supuestos distribucionales mínimos, una buena estrategia de selección de las variables adecuadas y una razonable complejidad computacional.

En la Sección § 2 se resumen brevemente el algoritmo Data Augmentation y los Árboles de Clasificación. En la Sección § 3 se describe en detalle el método que proponemos y en la Sección § 4 se resuelve un ejemplo con una base de datos ficticia y se comparan los resultados con algunos métodos tradicionales.

## 2. PRELIMINARES

**2.1. Imputación múltiple (IM) [1, 2, 4, 7, 8, 9]**

La imputación múltiple es la alternativa en general más favorecida por la literatura moderna sobre datos faltantes. Las ventajas de la IM se asocian con su eficiencia, su versatilidad en relación con las inferencias estadísticas post-imputación, y sobre todo, con el hecho de que incorpora la incertidumbre en el proceso de imputación.

A grandes rasgos, la IM remplaza cada valor ausente por un conjunto de ***m*** (***m*** > 1) valores plausibles, es decir, completa la base de datos original, de ***m*** maneras. A partir de la base de datos original (incompleta) se generan ***m*** nuevas bases completas.

Estos ***m*** juegos de datos "completos" se someten al análisis estadístico a que habría sido sometida la base original si no hubiese contenido observaciones faltantes, y los resultados de las inferencias (estimadores con sus errores estándar) se combinan mediante reglas simples [7] para obtener la solución deseada, que entraña un elemento de incertidumbre producto de la combinación de los estimadores y de las técnicas de simulación propias del método.

**2.2. Árboles de Regresión y Clasificación (ARC)**

Un modelo ARC describe la distribución condicional de la variable respuesta *y* dado un conjunto de predictores $X = (X1, X2,\ldots, Xp)$. El modelo tiene dos principales componentes: un árbol *T* con *b* nodos terminales y un vector $\mu = (\mu 1, \mu 2,\ldots, \mu b)$ que especifica la distribución de *Y* en cada nodo terminal.

Así, el par $(T, \mu)$ determina un modelo y si *X* está en la región correspondiente al **m**-ésimo nodo, entonces *Y/X* tiene distribución $f(Y/\mu m)$. El árbol es de regresión o clasificación de acuerdo a si la variable *Y* es cuantitativa o cualitativa. Esto da paso a un algoritmo iterativo de creación del árbol. En una base de datos, el árbol se forma teniendo en cuenta la creación de grupos de sujetos tan diferentes entre si como se pueda. Cada paso en la formación del árbol necesita escoger una entre las variables independientes y un punto de corte para ella de tal manera que se encuentre la mayor diferencia posible entre los grupos en cuanto a la variable dependiente. Supongamos que tenemos dividido el espacio en *M* regiones o nodos terminales $R1, R2,\ldots, RM$. Se modela *y* como constante *cm* en cada región *Rm*

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m) \qquad (1)$$

El primer paso consiste en definir cómo medir la "diferencia" entre grupos. En el caso de los árboles de clasificación existen tres medidas fundamentales, que miden específicamente la impureza de los nodos:

TQ Error de mala clasificación: $\frac{1}{N_m}\sum_{i\in R_m} I(y_i \neq k(m))$

UQ Índice de Gini: $\sum_{k\neq t} \hat{p}_{mk}\hat{p}_{mt}$

VQ Entropía cruzada: $\sum_{k=1}^{K} \hat{p}_{mk}\log\hat{p}_{mk}$

teniendo en cuenta que *Nm* es la cantidad de elementos en el nodo ***m***, $\hat{p}_{mk} = \frac{1}{N_m}\sum_{x_i\in R_m} I(y_i = k)$

es la proporción de observaciones de clase *k* en el nodo ***m*** y $k(m) = \arg\max_k(\hat{p}_{mk})$ es la clase predominante en el nodo ***m***. De esta manera, si denotamos la impureza por $Q_m(T)$ en cada paso de formación del árbol se busca la variable y el punto de corte que arrojen los dos próximos nodos menos impuros y este procedimiento se reitera hasta que se consiguen árboles con nodos terminales de muy pocos elementos [3]. Entonces se poda el árbol aplicando el criterio *costo-complejidad*, o sea, se trata de minimizar

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha|T| \tag{2}$$

en donde *T* es el árbol, $|T|$ la cantidad de nodos terminales y $\alpha$ un parámetro que mide el equilibrio entre el árbol más largo (cada nodo con un solo integrante) y el más chico (aquel de un solo nodo). Es necesario decir que una variable dependiente puede ser seleccionada varias veces, incluso varias veces seguidas, como la variable óptima para separar en dos nodos. Así mismo pueden existir variables en la base que el propio método no incluya en la formación del árbol. En cualquier caso, el resultado serán los nodos terminales que definen una agrupación de los sujetos de la base a partir de condiciones establecidas por los puntos de corte en las variables independientes.

**2.3. Data Augmentation**

El término Data Augmentation hace referencia a métodos para la construcción de algoritmos iterativos a través de la introducción de datos no observados o variables latentes y fue popularizado en la literatura estadística por Tanner y Wong en 1987 [11], con el artículo Data Augmentation Algorithm For Posterior Sampling. Los esquemas de Data Augmentation fueron usados por dichos autores para conseguir simulaciones factibles y simples. La construcción de esquemas de Data Augmentation que deriven en algoritmos sencillos es cuestión de arte, en donde estrategias exitosas varían de acuerdo a los datos observados y al modelo considerado.

La naturaleza del algoritmo es motivada por la siguiente representación de una densidad a posteriori, en donde todas las integrales referidas son integrales de Riemann:

$$p(\theta / y) = \int_Z p(\theta / z, y)\, p(z / y) dz \tag{3}$$

donde:

- $p(\theta / y)$ es la densidad a posteriori de los parámetros $\theta$ dados los datos $y$
- $p(z / y)$ es la densidad de los datos no observados *z* condicionada a los observados
- $p(\theta / z, y)$ es la densidad de $\theta$ dados los datos aumentados $x = (z, y)$
- $Z$ es el espacio muestral de los datos no observados

La densidad de los datos no observados condicionada a los observados puede escribirse como:

$$p(z / y) = \int_\Theta p(z / \phi, y)\, p(\phi / y) d\phi \tag{4}$$

donde $\Theta$ simboliza el espacio paramétrico de $\theta$.

Ahora, substituyendo 2 en 1 e intercambiando el orden de integración se puede observar que $p(\theta / y)$ debe satisfacer la siguiente ecuación:

$$g(\theta) = \int_{\Theta} K(\theta,\phi) g(\phi) d\phi \quad \text{donde} \quad K(\theta,\phi) = \int_{Z} p(\theta / z, y)\, p(z / \phi, y) dz \tag{5}$$

Sea *T* la transformación integral que convierte cualquier función integrable *f* en otra función integrable *Tf* a través de:

$$Tf(\theta) = \int_{\Theta} K(\theta,\phi)\, f(\phi) d\phi \tag{6}$$

Entonces 3 es equivalente a $Tg(\theta) = g(\theta).$ Esta ecuación sugiere un método iterativo para el cálculo de $p(\theta / y)$, o sea:

$$g_{i+1}(\theta) = (Tg_i)(\theta) \tag{7}$$

Tanner y Wong [11] demostraron que bajo condiciones de suavidad las $g_i$ calculadas de esta manera, siempre convergen a la distribución a la "a posteriori" deseada $p(\theta / y)$. Si la transformación integral 5 puede ser calculada analíticamente, la implementación de este método es directa. Desafortunadamente, en la mayoría de los casos no ocurre así y las integrales de los casos 1 y 2 son imposibles de resolver de manera analítica. Suele recurrirse entonces a los métodos de integración Monte Carlo. La ecuación 1 motiva el siguiente esquema iterativo:

Dada la actual aproximación $g_i$ de $p(\theta / y)$

1. Generar una muestra $z^{(1)}, \mathrm{K}\, , z^{(m)}$ de la actual aproximación $p(\theta / y)$ . Esto puede ser hecho como sigue:

• Generar $\theta$ de $g_i(\theta)$

• Generar *z* de $p(z / \phi, y)$ donde $\phi$ es el valor obtenido en el item anterior

2. Actualizar la presente aproximación a $p(\theta / y)$ como la mezcla de las densidades condicionales de $\theta$ dados los datos aumentados generados en el paso anterior. Esto es:

$$g_{i+1}(\theta) = m^{-1} \sum_{j=1}^{m} p(\theta / z^j, y) \tag{8}$$

Como se ha visto, el primer paso del algoritmo requiere generar múltiples valores de los datos no observados *z* muestreando a partir de la densidad condicional de *z* dado *y*. A este proceso se le llama Paso Imputación. El paso siguiente requiere computar la distribución "a posteriori" de $\theta$ basándose en los conjuntos de datos aumentados. Esto recibe el nombre de Posterior. El algoritmo consiste en alternar entre los dos pasos iterativamente.

**3. NUESTRO MÉTODO**

Supongamos ahora que tenemos una base de datos categóricos con ausencias. Después de hacer todas las consideraciones iniciales, como tener en cuenta si alguna variable o sujeto presentan una proporción de ausencias tan grande que no deban ser considerados, o tratar de discernir el patrón de ausencias, se puede proceder así:

Tomar la primera variable con ausencias y dicotomizarla. A saber, los datos presentes 1, y los ausentes 0. Inmediatamente, se encuentra el árbol de clasificación que producirá nodos o grupos de casos de sujetos homogéneos entre sí. Dicho de otro modo, tenemos la variable original particionada de manera tal que las proporciones de sus valores originales en los grupos están determinadas por otras variables de la base.

- Ahora, en cada grupo o nodo se implementa Data Augmentation con las siguientes características:

1. La verosimilitud es $p(X/\theta)=\prod_{j=1}^{k} p_j^{n_j}$ (verosimilitud de una variable categórica con k categorías y $n_j$ observaciones en la categoría j)

2. Utilizamos una distribución a priori Dirichlet con vector de parámetros $(1/2,1/2,...,1/2)$ para el vector $\theta$

- Se repite el proceso para cada variable con ausencias, utilizando la información rellenada en variables anteriores[1]
- Se obtienen las imputaciones múltiples (5-10) y las nuevas bases llenas. Se efectúan en ellas los análisis estadísticos requeridos y se combinan los resultados con las reglas de Rubin [6, 7]

Formalmente, la aplicación de Data Augmentation en una variable categórica con ausencias se puede escribir como: observaciones $x_1, x_2, ..., x_n$ i.i.d, que toman los valores $k_1, k_2, ..., k_K$ con probabilidades $\theta_{k_1}, \theta_{k_2}, ..., \theta_{k_K}$ de lo que se deriva la verosimilitud

$$p(X/\theta)=\prod_{j=1}^{K} \theta_j^{n_j} \tag{9}$$

donde $n_j$ es la cantidad de observaciones que tomaron el valor *j*. Llamemos $Y$ a los datos observados y $Z$ a los faltantes. Entonces los datos completos se pueden expresar $X=(Z,Y)$. Son necesarias las distribuciones condicionales $p(Z/\theta,Y)=\prod_{j=1}^{K} \theta_j^{n_{Zj}}$ con $n_{Zj}$ como la cantidad de faltantes por categoría, y $p(\theta/Y,Z) \propto p(X/\theta)p(\theta)$. Esto es:

$$p(\theta/Y,Z) \propto \frac{\Gamma\left(\frac{K+1}{2}\right)}{\Gamma^{K+1}\left(1/2\right)}\left(1-\sum_{l=1}^{K}\theta_l\right)^{-1/2}\prod_{l=1}^{K}\theta_l^{-1/2}\prod_{l=1}^{K}\theta_l^{n_j} \tag{10}$$

Cuando se aplica el Teorema de Bayes se obtiene el Kernel de otra distribución Dirichlet que será la distribución "a posteriori" buscada.

**4. EJEMPLO DE APLICACIÓN**

Para ilustrar el método propuesto, escogimos un fragmento de base de datos de 200 casos [5] donde se resumen los resultados ficticios de una encuesta que intenta medir las preferencias de los estudiantes universitarios a la hora de buscar pareja. Fueron seleccionadas tres variables categóricas:

Importancia de una buena posición económica de la pareja, 3 valores: 1-Muy importante, 2-Importante, 3-No importante
Importancia de la Apariencia Física, 4 valores: 1-Mucha, 2-Regular, 3-Poca, 4-Ninguna
Importancia del éxito en la docencia, 3 valores: 1-Mucha, 2-Regular, 3-Ninguna

[1] En este sentido, las variables deben rellenarse comenzando por la que más ausencias tiene hasta llegar a la de menor número de ausencias

A la base original se le provocaron ausencias de tipo MAR (Missing at Random) en la variable Posición Económica. Específicamente, se removieron casos según una variable Bernoulli con parámetro $p = 0.96$ cuando la variable Apariencia Física tomaba valores menores que 3. También se provocaron ausencias de tipo MCAR a la variable Éxito en la docencia, de acuerdo a una Bernoulli con parámetro $p = 0.95$. La Tabla 1 muestra los valores de los cálculos de las proporciones del valor 3 en la variable "posición económica" y del valor 2 en la variable "éxito en docencia", de acuerdo a la base original, a la solución que brinda nuestro método, a la eliminación total (eliminar todos los casos con ausencias) y a una sustitución simple de tipo Hotdeck[2].

**Tabla 1. Proporciones calculadas a partir de la aplicación de distintos métodos de imputación**

| Método de solución | Proporción de 1 en Pos Ec | Proporción de 1 en Ex Doc |
|---|---|---|
| Base completa | 0.42 | 0.29 |
| Nuestro método | 0.44 | 0.26 |
| Listwise | 0.35 | 0.26 |
| Hotdeck | 0.45 | 0.25 |

Los resultados que muestra la tabla son bastante favorables a nuestro método. Es lógico que funcione mejor que el método "listwise" (eliminación de un caso si contiene algún dato faltante), sobre todo en la variable que tiene un patrón de ausencias MAR, donde dichas ausencias dependen de otra variable de la base: allí donde "listwise" elimina, a un tiempo obvia relaciones importantes. El método "hotdeck" (imputación a partir de una muestra de vecinos próximos) trabaja de manera bastante aceptable, pero aún así es superado por nuestro método que solo es superado a su vez en estrecho margen en la complejidad computacional. Es necesario puntualizar que no existe aún un software debidamente implementado que ejecute nuestro método. La aplicación se llevó a cabo utilizando, para la parte de árboles de regresión, la versión 14 del SPSS y para el algoritmo Data Augmentation, se implementaron algoritmos en R, solo con fines investigativos, pues no está a punto para ser utilizado por usuarios.

## 5. DISCUSIÓN

Hemos propuesto un método para resolver situaciones de ausencia de datos en bases con variables de tipo categórico. El método emplea dos recursos relativamente modernos, que se justifican plenamente en esta área de aplicación. El método propuesto puede ser utilizado en bases de datos no categóricos, y aún en bases mixtas. Esto lo hace sumamente versátil y debe esta versatilidad a la flexibilidad de las herramientas sobre las cuales descansa su filosofía.

Se comporta bien preservando las relaciones de importancia entre las variables, y no está enfocado a devolver íntegra una base con ausencias, sino a reproducir aceptablemente los resultados particulares de las inferencias que se puedan querer hacer sobre tal o cual base, en relación a los que podrían hacerse si la base no tuviera ausencias.

**RECEIVED JANUARY 2009**
**REVISED SEPTEMBER 2009**

## REFERENCIAS

[1] ALLISON, P. D. (2000): Multiple imputation for missing data: A cautionary tale. **Sociological Methods and Research**, 28:301-309.

[2] ANDERSON, A. B., BASILEVSKY, A. and DEREK, P. J. (1983): Missing data: a review of the literature. En: Rossi P.H., Wright J.D., Anderson A.B. **Handbook of Survey Research**. Academic Press , New York.

[2] En específico, se sustituyó por un valor elegido al azar entre los 10 vecinos más cercanos a cada ausencia

[3] BREIMAN, L., FRIEDMAN, J.H, OLSHEN, R.A., STONE C.J. (1984): **Classification and regression trees**. The Wadsworth Statistics/Probability Series. New York, Chapman & Hall/CRC.

[4] LITTLE, R. J. A. and RUBIN, D. B. (1987): **Statistical analysis with missing data**. Wiley, New York, 1987.

[5] ROTH P. L. (1994): Missing data: A conceptual review for applied psychologists. **Personnel Psychology**, 47, 537-560.

[6] RUBIN D. B. (1976): Inference and missing data. **Biometrika**, 63, 467-474.

[7] RUBIN D. B. (1987): **Multiple imputation for non-response in surveys**. Wiley, New York.

[8] SCHAFER J. L. (1997): **Analysis of incomplete multivariate data**. Chapman and Hall, London.

[9] SCHEFFER J. (2002): Dealing with missing data. **Res Lett Inf Math Sci,** 3, 153-160.

[10] TANNER M. A. (1996): **Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions**. (3rd. Ed.) Springer-Verlag, New York.

[11] TANNER, M. A. and Wong, W. H. (1987): The calculation of posterior distributions by data augmentation. **J Am Stat Assoc.**, 82; 528-550.