EasyChair Preprint

№ 9007

# How to Implement Some Multiple Imputation Models for Panel Data

Ramón Álvarez-Vaz, Diana Del-Callejo-Canal,
Margarita Edith Canal-Martínez, Elena Vernazza and
Alar Urruticoechea



October 5, 2022

# Como implementar algunos modelos de imputación múltiple para datos de panel

*Ramón Álvarez-Vaz , Diana Del-Callejo-Canal , Margarita Edith Canal-Martínez , Elena Vernazza , Alar Urruticoechea*

**Abstract** Los datos faltantes son todo un reto en los análisis estadísticos porque los resultados que arrojan tienen limitaciones. La imputación, entendida como el proceso de reemplazar los datos faltantes con un valor estimado, es un problema regular en los proyectos de investigación. Existen muchos modelos y paqueterías destinada para este proceso, sin embargo, la selección del modelo de imputación adecuado al tipo de datos disponibles es trascendental para la fiabilidad del resultado. En este estudio se trabaja con una tabla de datos cruzada que involucran series de tiempo (datos panel) para 33 países y 17 variables (Índice de Gini anual para período 2000-2016), con un 24 % de datos faltantes. Con el objetivo de imputar estos datos, se utilizó un modelo de imputación múltiple propuesto por Honaker y King (2010) y se agregaron algunas restricciones al sistema. Los principales resultados obtenidos, conducen a la siguiente interrogante Se puede confiar en la imputación? Todos los archivos necesarios para reproducir los resultados presentados están disponibles en: https://gitlab.com/iesta.fcea.udelar/como-implementar-algunos-modelos-de-imputacion-mulltiple-para-datos-de-panel .

**Palabras clave:** Datos faltantes - Datos panel - Imputación

## Introducción

La mayoría de los métodos de análisis estadístico requieren de tablas completas pero muchos de los datos reales tienen que lidiar con casos de datos faltantes (J. Honaker and King 2010),(Zhang 2016). Aunque muchos estudios no reportan la manera en la que éstos son tratados/imputados (Bell et al. 2014),(Wood, White, and Thompson 2004), es de suma importancia conocer el método de imputación utilizado ya que de estas imputaciones depende la confiabilidad de los resultados. Las técnicas de imputación pueden dividirse en dos grandes categorías, por un lado las que se basan en métodos de imputación simple (aleatorias y deterministas), recomendados cuando existe un patrón monótono y por otro las que usan métodos de imputación múltiples, como lo que se encuentran en (Rubin 1987),(Muñoz-Rosas and Álvarez-Verdejo 2009), recomendados cuando existe un patrón arbitrario. La elección del modelo de imputación dependerá, por una parte de las características de la tabla de datos y por otra de la información disponible alrededor de los datos faltantes; el patrón de los datos faltantes, del tipo de datos (categórico o numérico) y el mecanismo de generación de los datos o la estructura de los mismos (series de tiempo, datos panel, diseño experimental, etc.). Los datos utilizados en este trabajo son datos de panel, los cuales pueden verse como arreglo matricial de columnas y renglones, donde a los individuos (países en este caso) se les mide una o más variables a lo largo del tiempo (Arellano and Bover 1990), de tal manera que la variación en la temporalidad y la variación en los individuos resultan igual de importantes para el estudio.

## Objetivo

El objetivo de este trabajo es imputar adecuadamente el índice de Gini en el 24 % de los datos faltantes de una matriz con estructura de datos panel con las siguientes características: 33 países (individuos) y 17 variables (índice de Gini anual para el período 2000-2016).

## Metodología

Los métodos actuales de estimación múltiple, varían dependiendo de dos condiciones:

- El tipo de mecanismo faltante, que pueden ser tres formas, Completamente al Azar (MCAR, por sus siglas en inglés); al Azar (MAR por sus siglas en inglés) y de Información no Ignorable (NI por sus siglas en inglés).
- El tipo de algoritmo utilizado, Monte Carlo con Cadenas de Markov (MCMC, por sus siglas en inglés); Especificación Totalmente Condicional (FCS por sus siglas en inglés); Esperanza-Maximización (EM en sus siglas en inglés) y muy recientemente Esperanza-Maximización con Bootstraping (EMB por sus siglas en inglés).

Así la mayor dificultad de este método de imputación reside en la generación del modelo del que posteriormente se simularán los datos faltantes.

(J. Honaker and King 2010) proponen un modelo de estimación de datos faltantes pensado específicamente para estructuras de series de tiempo en tablas cruzadas (Time-Series Cross-Section Data), dicha propuesta incluye un modelo de

estimación que considera los cambios en los individuos y las tendencias a lo largo del tiempo simultáneamente. Además, los mismos autores implementan su propuesta en una paquetería en R-project, llamada Amelia II (H. Honaker, King, and Blackwell 2018).

Para ello se basan en dos condiciones: a) el tipo de mecanismo faltante es MAR y b) utilizan un algoritmo de EMB, que consiste en hacer un remuestreo de 5 para la tabla de datos incompleta, imputar mediante Esperanza-Maximización a los valores faltantes de cada muestra, separar los resultados de dicha imputación y analizarlos para finalmente obtener un valor, con sus intervalos de confianza (H. Honaker, King, and Blackwell 2018).

Así a grandes rasgos, la propuesta de Honaker y King, consiste en 1) extraer la información relevante de las proporciones de los datos observados y construir un modelo de estimación, 2) a partir de ese modelo completar los datos faltantes, y 3) a partir de ello construir un nuevo modelo con los datos "completos". Es un proceso iterativo hasta encontrar el punto de convergencia. Finalmente, el proceso ofrece un único dato de imputación con su correspondiente intervalo de confianza al 95 %.

Los datos utilizados en este trabajo corresponden a 33 países, con el coeficiente de Gini registrado desde el año 2000 hasta el 2016. En total son 561 casos, de los cuales 135 son datos faltantes (24 %)1. Se utilizó un modelo de imputación sin restricciones, el cual presentaba algunos problemas, posteriormente y después de analizar los datos se utilizaron restricciones para Guatemala y Japón, a través de la media y la desviación estándar de cada país.

## Resultados

El primer modelo de imputación fue realizado bajo el supuesto MAR, sin imponer ninguna restricción, con un remuestreo de 5. Los resultados fueron analizados teniendo la consideración en forma conjunta y complementaria:

- la comparación de las densidades de los datos observados y los estimados
- la amplitud de los intervalos de confianza de las estimaciones.

Los resultados muestran que las densidades de la imputación y de los datos reales no coinciden. En los datos observados se verifica la presencia de dos poblaciones, una que contempla a países con el índice de Gini entre los 20 y 40 puntos (en escala de 1 a 100) y otra población de países con el índice de Gini en escala de 40 a 70. En el caso de los datos imputados, se presenta también una estructura de dos poblaciones, sin embargo, se observa un claro desfase entre ambas densidades. Además, los resultados de la prueba de contraste por Kolmogorov-Smirnov, indican que existe suficiente evidencia estadística para rechazar la hipótesis de igualdad entre ambas densidades (p-valor<0.0404). Con respecto a la amplitud de los intervalos de confianza de las estimaciones para cuatro países, que se destacan por tener la mayor cantidad de datos faltantes Guatemala, Japón, Suiza y Uruguay, los intervalos de confianza grandes.

A partir de estos últimos resultados, se decide integrar restricciones al sistema y volver a estimar. En particular, en este trabajo la elección de las restricciones sigue la regla de introducir información a priori de Guatemala (con la media y desviación estándar proveniente de los datos observados) e imputar nuevamente. Como aún persisten irregularidades se procede a restringir a Japón (media y desviación estándar) y al encontrarse un mejor modelo se detiene la inclusión de información a priori. Con la introducción de la información a priori para Guatemala y Japón, el ajuste de la distribución del modelo de estimación es más adecuadodonde la densidad de la imputación refleja (sin desfase) las dos poblaciones que hay en los datos observados. Además, los resultados de la prueba de contraste de Kolmogorov-Smirnov establecen que no existe suficiente evidencia estadística para rechazar la hipótesis de igualdad entre ambas densidades (p<0.9997). Por último, es posible observar que las estimaciones de la media y los intervalos de confianza para los datos faltantes estimados, presentan resultados por países con intervalos de confianza de menor amplitud que en el modelo sin restricciones. Se entiende, por lo tanto, que el segundo modelo funciona mejor y que las estimaciones son más confiables que las del primero. Todos los resultados fueron calculados utilizando la librería Amelia II de R-project (H. Honaker, King, and Blackwell 2018).

## Conclusiones y Discusión

Disponer de una tabla de datos completos es ideal, pero aplicar métodos de imputación inapropiados para lograrlo, puede generar más problemas de los que resuelve (Medina and Galván 2007). Sus implicaciones en el análisis secundario de datos deben evaluarse con cautela, y en este trabajo surgen evidencias de que no existe el método de imputación ideal. Más bien lo que se requiere es que el investigador realice:

- un diagnóstico del problema de datos faltantes;
- configure el modelo de imputación de acuerdo a las necesidades de su tabla;
- verifique la calidad de los datos imputados.

Respecto al diagnóstico del problema de datos faltantes, en este caso de estudio se utilizaron tres elementos: 1) el tipo de mecanismo de generación de datos faltantes; 2) el porcentaje de datos faltantes; y 3) la estructura de la tabla de datos.

En lo referente al mecanismo de generación de la información faltante, la estimación por máxima verosimilitud se basa en supuestos cruciales: la muestra debe tener tamaño suficiente para que las estimaciones sean aproximadamente insesgadas y normalmente distribuidas para algunos casos la estimación puede ser posible al apartarse del modelo pero requerirán el supuesto MCAR (Schafer and Graham 2002). Además del mecanismo de generación de la información faltante, es necesario considerar el porcentaje de datos faltantes. En el caso de estudio el 24 % de los datos eran faltantes y la primer interrogante que surgió fue £Es esto mucho o poco? De acuerdo con Shafer, 1999 (como se cita en (Madley-Dowd et al. 2019) ) y (Stef Van Buuren 2018) se establece que un 5 % de datos faltantes es el límite mínimo para que la imputación múltiple funcione. Los autores de la librería Amelia II, no establecen un porcentaje límite, solo mencionan que el bootstrap con 5 iteraciones, es suficiente si los datos faltantes no son muchos (J. Honaker and King 2010). Sin embargo, en un estudio realizado por Clavel et al. (2014) se muestra una comparación de 7 librerías del software R y establecen que Amelia II y Norm tuvieron las mejores estimaciones y que Amelia II funciona mejor con menos de 25 % de datos faltantes -con las condiciones establecidas en ese estudio particularmente-.

La estructura de la tabla de datos es un factor que influye en la toma de decisiones sobre el modelo de imputación. No es lo mismo usar una tabla de datos transversales, longitudinales, o panel, como tampoco es lo mismo si los datos provienen de un diseño de experimentos que de estudios observacionales, o si los datos son categóricos o numéricos. Por ejemplo, un reciente estudio sobre datos longitudinales en un diseño de experimentos con presencia de heteroscedasticidad muestra que al separar la imputación por grupos ésta resultó ser menos sesgada y más precisa que al hacer una imputación simultánea (Yusuke, Mai, Kazushi & Masahiko, 2020). También, en una comparación de datos categóricos bajo el supuesto MCAR, el porcentaje más bajo de error de clasificación lo obtuvo la imputación simple de la media, mientras que para datos numéricos bajo el supuesto MAR, la Esperanza-Maximización (EM) mostró el menor sesgo (Kossen et al. 2019). De este modo, la elección de una librería como Amelia II, que fue creada para la estructura de datos tipo panel es algo que se tiene que incluir como parte del diagnóstico del problema de datos faltantes.

Sobre la configuración del modelo de imputación, los resultados arrojados por los dos modelos estudiados en este ejercicio son distintos. En la imputación automática (sin ninguna restricción) cuatro de los 33 países presentaban intervalos de confianza muy grandes: Guatemala, Japón, Suiza y Uruguay y el resto presentaba algunos inconvenientes. Por esta razón, se consideró necesario considerar información a priori del país con mayores problemáticas (Guatemala) e imputar nuevamente. Al continuar detectando irregularidades se incorporó la restricción sobre el segundo país con intervalos de confianza con mayor amplitud en la imputación (Japón). Con la introducción de la información a priori al modelo se mejoró la distribución de la imputación.

Algunas de las sugerencias para mejorar la imputación múltiple es la inclusión de tantas variables como sea posible (Murray 2018) o bien la inclusión de una variable auxilar (Takahashi 2017). En el caso de la aplicación presentada en este trabajo, se hizo la imputación con una sola variable, aunque a futuro sería interesante hacer comparaciones con el método de imputación de múltiple ratio propuesto por (Takahashi 2017), siempre que se contara con la información de información auxiliar para los mismos países en el mismo periodo.

Con respecto a la verificación de la calidad de los datos, la literatura sugiere que se utilice la comparación de las densidades de los datos observados y los imputados, considerando que si hay mayore similitud, se asume una mejor imputación. En este trabajo se consideró un análisis gráfico combinado con los intervalos de confianza de los datos imputados aspecto que ofrece información útil para la verificación de la calidad de la imputación.

Por último es importante resaltar el aporte de la tabla completa en sí misma (con los valores imputados), como una contribución para su utilización en estudios posteriores.

## Referencias

10 Arellano, Manuél, and Olympia Bover. 1990. "La Econometría de Datos Panel." *Investigaciones Económicas* XIV (1): 3–45. https://www.fundacionsepi.es/investigacion/revistas/paperArchive/Ene1990/v14i1a1.pdf.

Bell, Melanie L, Mallorie Fiero, Nicholas J Horton, and Chiu-Hsieh Hsu. 2014. "Handling Missing Data in RCTs; a Review of the Top Medical Journals," 8.

Honaker, H., G. King, and M. Blackwell. 2018. "AMELIA II."

Honaker, James, and Gary King. 2010. "What to Do about Missing Values in TimeSeries CrossSection Data." *America Journal of Political Science* 54 (2): 561–81. https://doi.org/https://doi.org/10.1111/j.1540-5907.2010.00447.x.

Kossen, Tabea, Michelle Livne, Vince I Madai, Ivana Galinovic, Dietmar Frey, and Jochen B Fiebach. 2019. "A Framework for Testing Different Imputation Methods for Tabular Datasets." *bioRxiv*, January, 773762. https://doi.org/10.1101/773762.

Leite, W., and N. Beretvas. 2010. "The Performance of Multiple Imputation for Likert-Type Items with Missing Data." *Journal of Modern Applied Statistical Methods* 9: 64–74. https://doi.org/10.22237/jmasm/1272686820.

Madley-Dowd, P., R. Hughes, K. Tilling, and J. Heron. 2019. "The Proportion of Missing Data Should Not Be Used to Guide Decisions on Multiple Imputation." *Journal of Clinicical Epidemiology* 110: 63–73. https://doi.org/https://doi.org/10.1016/j.jclinepi.2019.02.016.

Medina, F., and M. Galván. 2007. "Imputación de Datos: Teoría y Práctica." Publicación de las Naciones Unidas: CEPAL. https://repositorio.cepal.org/bitstream/handle/11362/4755/1/S0700590_es.pdf.

Muñoz-Rosas, Juan Francisco, and Encarnación Álvarez-Verdejo. 2009. "Métodos de Imputación Para El Tratamiento de Datos Faltantes: Aplicación Mediante R/Splus." *Revista de Métodos Cuantitativos Para La Economía y La Empresa*, 3–30. http://www.upo.es/RevMetCuant/art25.pdf.

Murray, Jared S. 2018. "Multiple Imputation: A Review of Practical and Theoretical Findings." *Statistical Science* 33 (2): 142–59. https://doi.org/10.1214/18-STS644.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. USA: John Wiley & Sons, Inc. https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470316696.fmatter.

Schafer, Joseph L., and John W. Graham. 2002. "Missing Data: Our View of the State of the Art." *Psychological Methods* 7 (2): 147–77.

Takahashi, M. 2017. "Multiple Ratio Imputation by the EMB Algorithm: Theory and Simulation." *Journal of Modern Applied Statistical Methods* 16 (1): 630–56. https://doi.org/10.22237/jmasm/1493598840.

Van Buuren, S., and K Groothuis-Oudshoorn. 2011. "Mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software, Articles* 45 (3): 1–67. https://doi.org/10.18637/jss.v045.i03.

Van Buuren, Stef. 2018. *Flexible Imputation of Missing Data*. Segunda. Florida, USA: Taylor & Francis. https://stefvanbuuren.name/fimd/.

Wood, Angela M, Ian R White, and Simon G Thompson. 2004. "Are Missing Outcome Data Adequately Handled? A Review of Published Randomized Controlled Trials in Major Medical Journals." *Clinical Trials: Journal of the Society for Clinical Trials* 1 (4): 368–76. https://doi.org/10.1191/1740774504cn032oa.

Yamaguchi, Yusuke, Mai Ueno, Kazushi Maruo, and Masahiko Gosho. 2020. "Multiple Imputation for Longitudinal Data in the Presence of Heteroscedasticity Between Treatment Groups." *Journal of Biopharmaceutical Statistics* 30 (1): 178–96. https://doi.org/10.1080/10543406.2019.1632878.

Zhang, Zhongheng. 2016. "Missing Data Imputation: Focusing on Single Imputation." *Annals of Translational Medicine* 4 (1): 9. https://doi.org/10.3978/j.issn.2305-5839.2015.12.38.

*Ramón Álvarez-Vaz*
*Universidad de la República,FCEA,IESTA*
ramon.alvarez@fcea.edu.uy

*Diana Del-Callejo-Canal*
*Universidad Veracruzana,México*

*Margarita Edith Canal-Martínez*
*Universidad Veracruzana,México*

*Elena Vernazza*
*Universidad de la República,FCEA,IESTA*

*Alar Urruticoechea*
*Universidad Católica del Uruguay, Uruguay*