CONSIDERACIONES A LA IMPUTACIÓN MÚLTIPLE. UN CASO DE ESTUDIO CON DATOS PANEL

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

* y** Instituto de Investigación de Estudios Superiores Económicos y Sociales de la Universidad Veracruzana, México (IIESES-UV)
*** y*****Instituto de Estadística de la Facultad de Ciencias Económicas y de Administración, Universidad de la República
****Departamento de Neurocognición, Universidad Católica del Uruguay

*y** Luis Castelazo Ayala s/n, colonia Industrial Ánimas, C.P. 91190, Xalapa, Veracruz, México; *** y*****Eduardo Acevedo 1139. C.P. 11200, Montevideo Uruguay; **** Comandante Braga 2715, Montevideo, CP 11600, Montevideo, Uruguay
*ddelcallejo@uv.mx,**mcanal@uv.mx,
elena.vernazza@fcea.edu.uy;*alar.urruticoechea@ucu.edu.uy;
*****ramon@iesta.edu.uy
*0000-0003-4753-6577 - **0000-0002-1258-5902 - ***0000-0003-3123-2165 - ****0000-0001-6229-2633 - *****0000-0002-2505-4238

Recibido 28 de diciembre de 2020, aceptado septiembre 2021

RESUMEN

Los datos faltantes son todo un reto en los análisis estadísticos. 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 subrutinas de diversos software destinadas 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) con un 24% de datos faltantes. Con el objetivo de imputar estos datos, se utilizó un modelo de imputación múltiple y se agregaron algunas restricciones al sistema. El principal aporte de este ejercicio es mostrar que un buen proceso de imputación requiere del diagnóstico del problema, de la configuración del modelo de imputación y, finalmente, de la verificación de la calidad de los datos imputados.

Palabras clave: imputación, datos faltantes, series de tiempo, datos panel, imputación múltiple.

Códigos JEL: C1, C23, C8

CONSIDERATIONS FOR MULTIPLE IMPUTATION. CASE OF STUDY WITH PANEL DATA

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

* y** Instituto de Investigación de Estudios Superiores Económicos y Sociales de la Universidad Veracruzana, México (IIESES-UV)
*** y*****Instituto de Estadística de la Facultad de Ciencias Económicas y de Administración, Universidad de la República
****Departamento de Neurocognición, Universidad Católica del Uruguay

*y** Luis Castelazo Ayala s/n, colonia Industrial Ánimas, C.P. 91190, Xalapa, Veracruz, México; *** y*****Eduardo Acevedo 1139. C.P. 11200, Montevideo Uruguay; **** Comandante Braga 2715, Montevideo, CP 11600, Montevideo, Uruguay
*ddelcallejo@uv.mx,**mcanal@uv.mx,
elena.vernazza@fcea.edu.uy;*alar.urruticoechea@ucu.edu.uy;
*****ramon@iesta.edu.uy
*0000-0003-4753-6577 - **0000-0002-1258-5902 - ***0000-0003-3123-2165 - ****0000-0001-6229-2633 - *****0000-0002-2505-4238

Received December 28th 2020, accepted September 2021

ABSTRACT

Missing data is a challenge for statistical analysis. Imputation, as the process of replacing missing data with an estimated value, is a regular problem in any research project. There are many imputation models and packages that make this process. Nevertheless, the election of the adequate imputation model is transcendental for the results reliability. In this study we work with a Time-Series Cross-Section dataset (TSCS) and 24% of missing data. We used a multiple imputation model and aggregated some prior information to the system. The principal contribution to this exercise is to show that a good imputation requires (beside the software) a problem diagnosis, the configurations of the model imputation, and finally, the diagnostic of the quality of the data imputation.

Keywords: Imputation, missing data, time series, panel data, multiple imputation.

JEL Codes: C1, C23, C8

1 INTRODUCCIÓN

La mayoría de los métodos de análisis estadístico requieren de tablas completas, pero los datos reales tienen que lidiar con casos de datos faltantes (Honaker & King, 2010; Zhang, 2015). Muchos estudios no reportan la manera en la que éstos son tratados/imputados (Bell, Fiero, Horton & Hsu, 2014; Wood, White, & Thompson, 2004) o bien, se destaca la alternativa de borrarlos, lo cual es entendible, dado que además de las dificultades técnicas que esto implica, la prevalencia científica parece ser la de restarle importancia (Van Buuren, 2018).

Las técnicas de imputación, a grandes rasgos, se pueden dividir en dos categorías, por un lado, están los métodos de imputación simple (aleatorios y deterministas), recomendados cuando existe un patrón monótono de datos faltantes y por otro, los métodos de imputación múltiple propuestos por Rubin en 1978 (Muñoz-Rosas y Álvarez-Verdejo, 2009), indicados cuando existe un patrón arbitrario.

Los métodos de imputación múltiple han ganado fama en últimas fechas, debido a la disponibilidad de acceso a softwares que implementan estos procesos, como por ejemplo en el software R, a través de librerías como: VIM (Kowarik & Templ, 2016), MICE (Van Buuren & Groothuis-Oudshoorn, 2011), MissForest (Stekhoven and Bühlmann, 2012) y Hmisc (Harrell, 2020), así como, las que han demostrado tener mejor performance que otros métodos, como es el de la imputación por la media o la regresión (Baraldi & Enders, 2010; Cheema, 2014) y por probar que son robustos al supuesto de normalidad (Leite y Beretvas, 2010). Sin embargo, la imputación múltiple no debería ser considerada como una solución técnica para los datos faltantes, sino como un proceso que requiere del juicio científico y estadístico, que va desde el diagnóstico del problema de datos faltantes, pasando por la configuración del modelo de imputación, hasta la validación de la calidad de los datos. No es posible tomar decisiones únicamente en función de un software, es necesario ajustar el proceso de una manera apropiada (Van Buuren, 2018).

Cada situación es diferente, por lo que a priori, no es conveniente adoptar el mismo procedimiento de imputación para todas las tablas de datos (Medina y Galván, 2007). 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, el tipo de datos (categórico o numérico) y de la estructura de la tabla de datos (series de tiempo, datos panel, diseño experimental, etc.).

En este trabajo, los datos utilizados son datos panel, los cuales son un arreglo matricial de columnas y filas, donde a los individuos se les mide una o más variables a lo largo del tiempo (Arellano y Bover, 1990), de tal manera que la variación en la temporalidad y la variación en los individuos resultan igual de importantes para el estudio.

Considerando las características de datos panel y dado el patrón arbitrario que poseen los datos faltantes en este tipo de tablas, el método de imputación que se suele utilizar es el de imputación múltiple, mediante esta imputación se pretende obtener estimadores no sesgados. Este método, reemplaza cada dato faltante por un conjunto de datos aceptables (verosímiles) que representan la incertidumbre alrededor del valor real (desconocido), después de ser analizados se completa el dato faltante original con alguno de estos valores (Rubin, 1987). Este proceso está ligado a probabilidades de ocurrencia que dependen del comportamiento de las variables, es así como el valor exacto que se incluya podrá variar, pero en términos probabilísticos será esencialmente el mismo.

Honaker & King (2010) proponen un algoritmo para el ajuste de imputación de datos panel, para ello, toman el procedimiento de imputación múltiple como base y consideran los patrones de la temporalidad que puede variar drásticamente entre países y, los patrones al interior de cada país pueden variar muy suavemente. Además, el algoritmo permite que se introduzcan restricciones al sistema, las cuales sirven como información a priori que se incluirá dentro del modelo.

El objetivo de este trabajo es imputar 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).[1] El principal aporte de este ejercicio es mostrar que un buen proceso de imputación requiere, además del software, del diagnóstico del problema de datos faltantes, de la configuración del modelo de imputación y, finalmente, de la verificación de la calidad de los datos imputados.

2 METODOLOGÍA

2.1 Diagnóstico del problema de datos faltantes

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

1. El tipo de mecanismo de generación del dato faltante: Completamente al Azar (MCAR, por sus siglas en inglés); al Azar (MAR por sus siglas en inglés) y faltante no azaroso (NMAR, por sus siglas en inglés) Van Buuren (2018).
2. 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).

[1] Los datos fueron recopilados y proporcionados por el Dr. Edgar J. Saucedo-Acosta y la Mtra. Nallelly Patricia Bolaños del Instituto de Investigación de Estudios Superiores, Económicos y Sociales de la Universidad Veracruzana (IIESES-UV).

Así, la mayor dificultad de este método de imputación reside en la generación del modelo del que posteriormente se simularán los datos faltantes (Antía y Coimbra, 2009).

Entre todas las alternativas analizadas, en este trabajo se opta por la propuesta de Honaker & King (2010) cuyo modelo de estimación de datos faltantes está 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 librería desarrollada en *R-project* (R Core Team, 2019), llamada Amelia II (Honaker, King and Blackwell, 2018).

Amelia II, se basa en dos condiciones: a) el tipo de mecanismo de generación del dato faltante es MAR; y b) utilizan un algoritmo de EMB, que consiste en hacer un remuestreo de 5 iteraciones 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 (Honaker et al., 2018).

Para el caso de la tabla de datos en este trabajo, se asume que la probabilidad de datos faltantes en cualquier país es la misma, o en todo caso, que la probabilidad de datos faltantes de cualquier año es la misma, por lo que el mecanismo de generación de datos faltantes considerado es el MAR. Además, para utilizar un modelo de imputación múltiple se requiere que por lo menos exista un 5% de datos faltantes (Van Buuren, 2018) condición que se cumple al tener 24% de datos faltantes. Por último, EMB es un algoritmo adecuado para eficientizar el proceso de imputación y poder hacer pruebas comparativas, que es una de las herramientas que nos permite un diagnóstico de la calidad de los datos imputados.

2.2 Configuración del modelo de imputación

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 (2010). 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. Se configuraron dos modelos de imputación o dos modelos de respuesta considerando una sola variable en el análisis (índice de Gini). El primer modelo analizado fue sin restricciones (sin información a priori), con 5 iteraciones, que es el que ofrece la

librería de manera automática. El segundo, usando un modelo de imputación flexible con el objetivo de mejorar los resultados.

Un modelo de imputación flexible es aquel en el que se puede incorporar información a priori, el reto de integrar este tipo de información es el ajuste manual que debe de hacerse, sin embargo, la evidencia empírica sugiere que un modelo de imputación flexible es mucho mejor que un modelo automático (Murray, 2018). En el caso de la librería Amelia II (Honaker et al., 2018), es relativamente sencillo incorporar esta información. La pregunta importante aquí es ¿Qué información es relevante integrar?

De acuerdo con Clavel, Merceron & Escarguel (2014), las observaciones con un intervalo de confianza muy grande se pueden remover interactivamente tomando en cuenta algunas consideraciones como media y varianza. De manera que, para introducir restricciones al sistema, partimos de la base del análisis de los datos. Después de obtener resultados en el primer modelo y graficar las series de tiempo por países, se detectaron 4 países con problemáticas mayores: a) Guatemala, que tenía mayor cantidad de datos faltantes y los intervalos de confianza que arrojaba eran grandes; b) Japón con datos faltantes consecutivos al inicio y al final del período e intervalos de confianza grandes; y c) Suiza y Uruguay que además de tener datos faltantes consecutivos al inicio del periodo mostraba intervalos de confianza grandes. La elección de las restricciones siguió la regla de introducir información a priori del país con mayores problemáticas (Guatemala), hacer de nuevo la imputación y si aún se detectan irregularidades, se restringiría al segundo país con intervalos de confianza con mayor amplitud en la imputación (Japón), y así sucesivamente hasta encontrar el mejor modelo.

2.3 Verificación de la calidad de los datos imputados

La verificación de la calidad de datos imputados está relacionada con el ajuste del modelo de imputación, esto significa todo un reto, sin embargo, existen algunas aproximaciones (Murray, 2018). Una de ellas consiste en comparar y contrastar las densidades de los datos imputados y los datos observados. En este trabajo se presentan gráficamente ambas densidades y posteriormente se testea la igualdad mediante la prueba de Kolmogorov-Smirnov (Abayomi, Gelman & Levy, 2008). Se presenta, además, junto a cada estimación por país y tiempo, su correspondiente intervalo de confianza de las estimaciones por país.

3 RESULTADOS

Como se mencionó en párrafos anteriores, 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: 1) la comparación de las densidades de los datos observados y los estimados; y 2) la amplitud de los intervalos de confianza de las estimaciones.

Tal como se observa en la figura 1, las densidades de la imputación y de los datos reales no coinciden. En los datos observados se muestra 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).

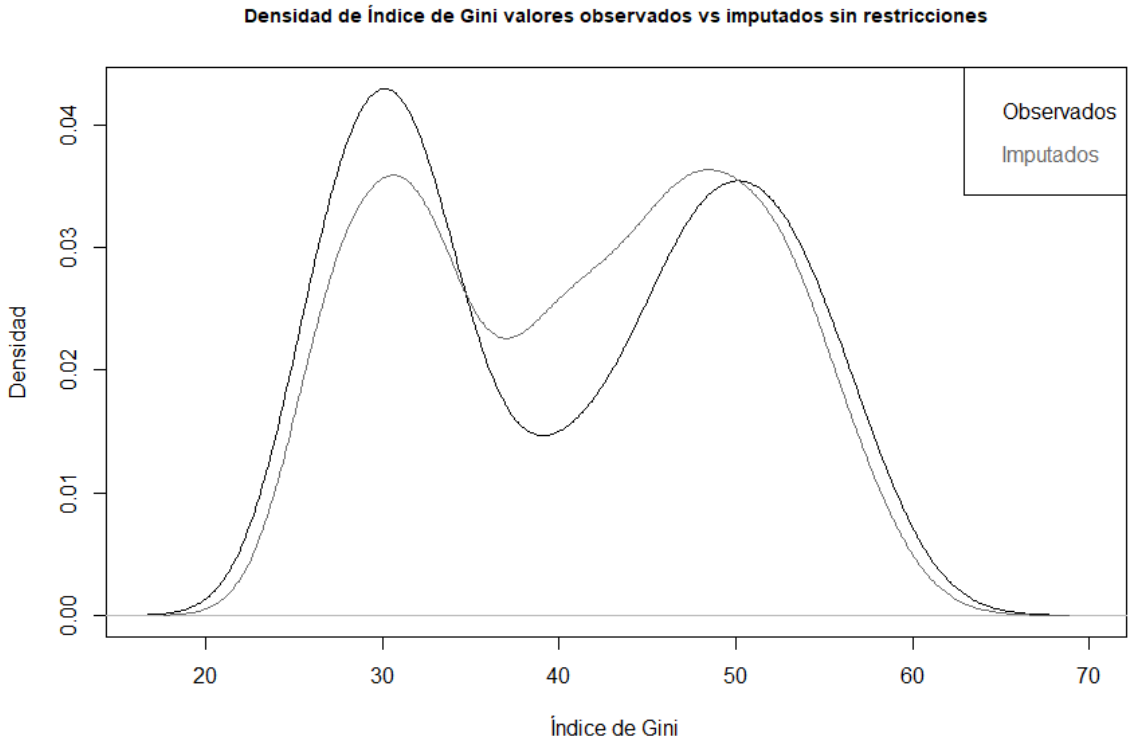

Figura 1. Densidades relativas de los datos observados vs los datos imputados 1.
Fuente: Elaboración propia, Librería Amelia II, *R-project*, versión 1.7.5

En la Figura 2, se muestra la amplitud de los intervalos de confianza de las estimaciones para cuatro países que se destacan por tener la mayor cantidad de datos faltantes e intervalos de confianza grandes: Guatemala, Japón, Suiza y Uruguay; es importante mencionar que no son los únicos datos que presentan alguna problemática, pero si son los más destacados.

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 siguió la regla de introducir información a priori de Guatemala (media y desviación estándar proveniente de

los datos observados) e imputar nuevamente. Como aún se detectaron irregularidades se procedió a restringir a Japón (media y desviación estándar) y al encontrarse un mejor modelo se detuvo la inclusión de información a priori.

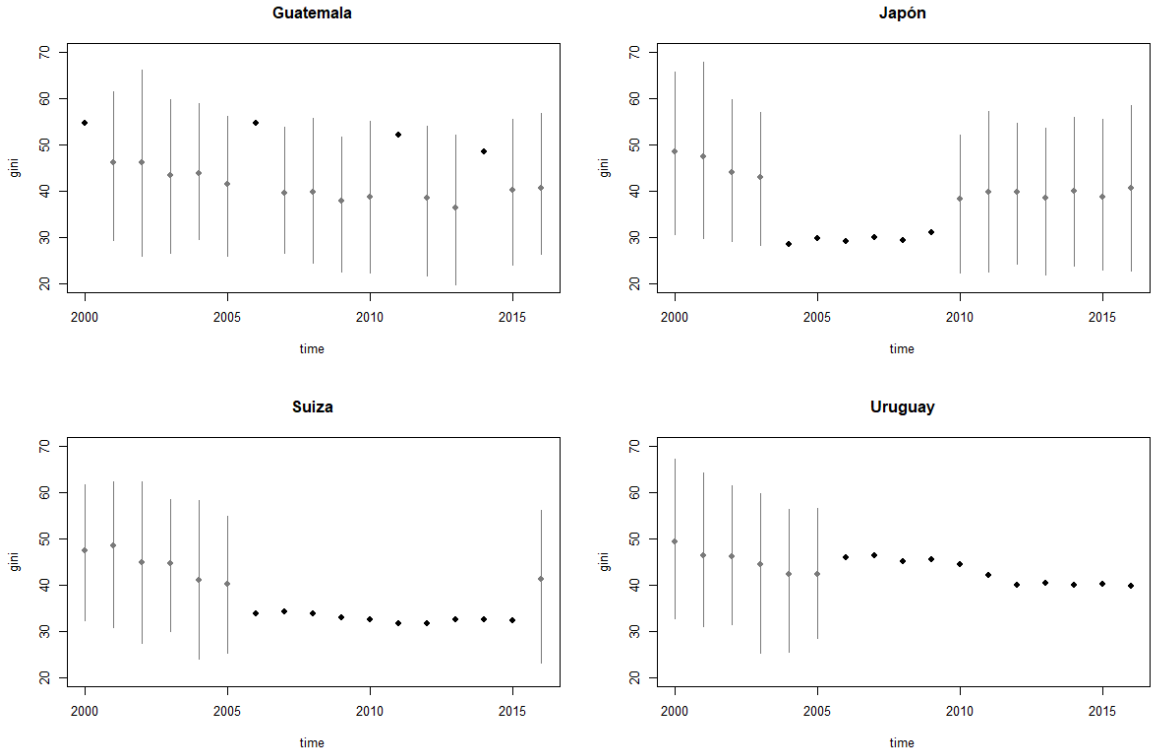

Figura 2. Intervalos de confianza para las imputaciones de 4 países.
Fuente: Elaboración propia, Librería Amelia II, *R-project*, versión 1.7.5

Con la introducción de la información a priori para Guatemala y Japón, el ajuste de la distribución del modelo de estimación es más adecuado tal como lo muestra la Figura 3, donde 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 ($p<0.9997$).

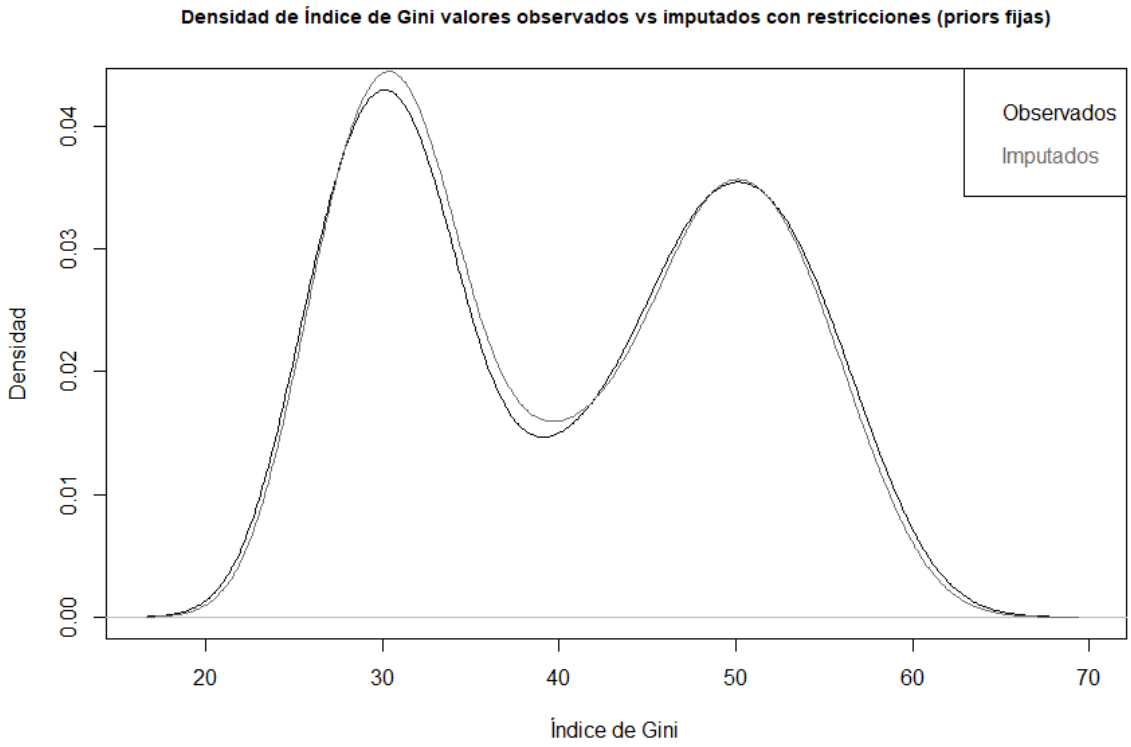

Figura 3. Densidades relativas (con restricciones al sistema) de los datos observados vs los datos imputados.
Fuente: Elaboración propia, Librería Amelia II, *R-project*, versión 1.7.5

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 (ver Figura 4). 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* (Honaker et al., 2018).

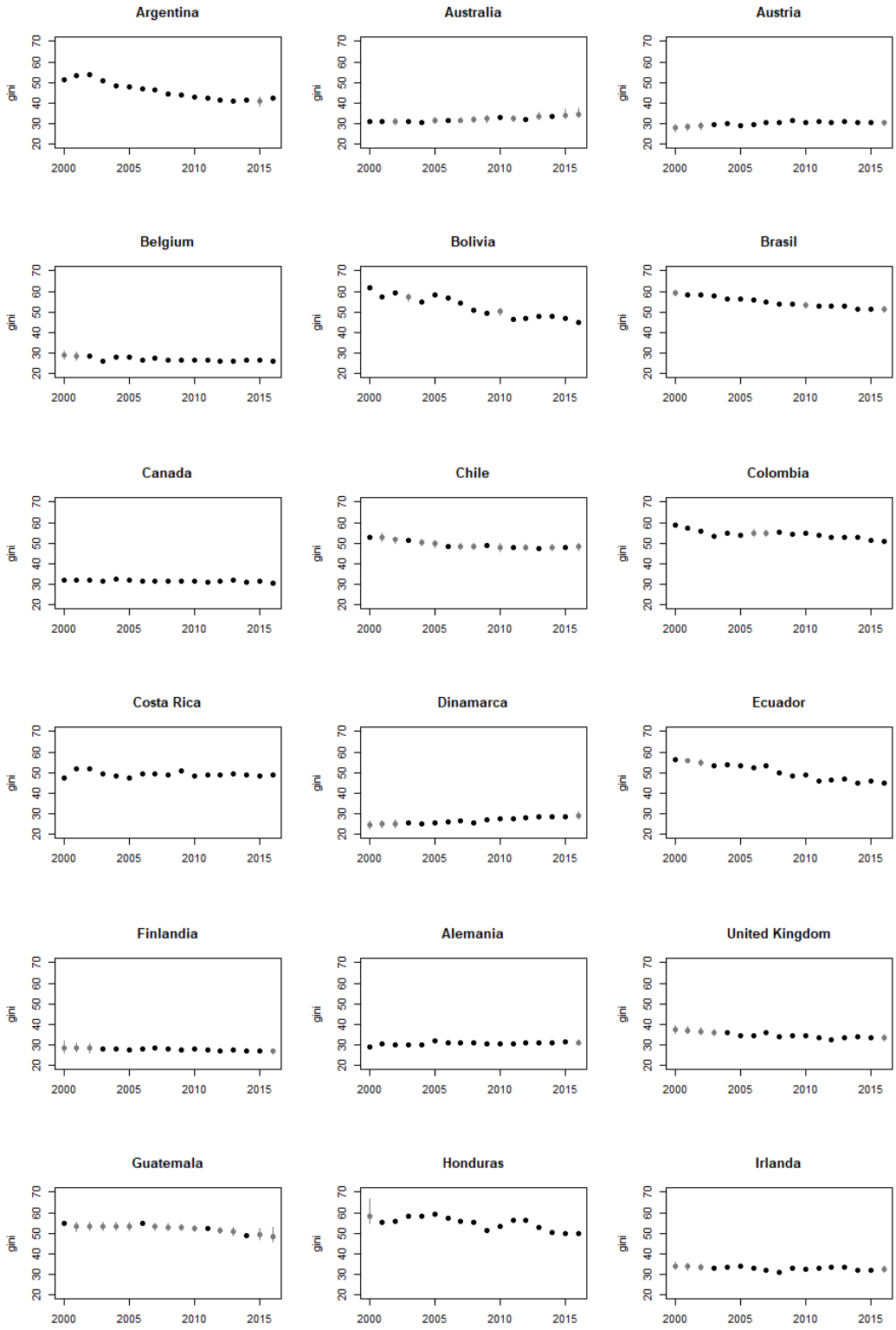

Figura 4 (continua). Datos observados vs datos estimados con dos restricciones a dos países
Fuente: Librería Amelia II, *R-project*, versión 1.7.5

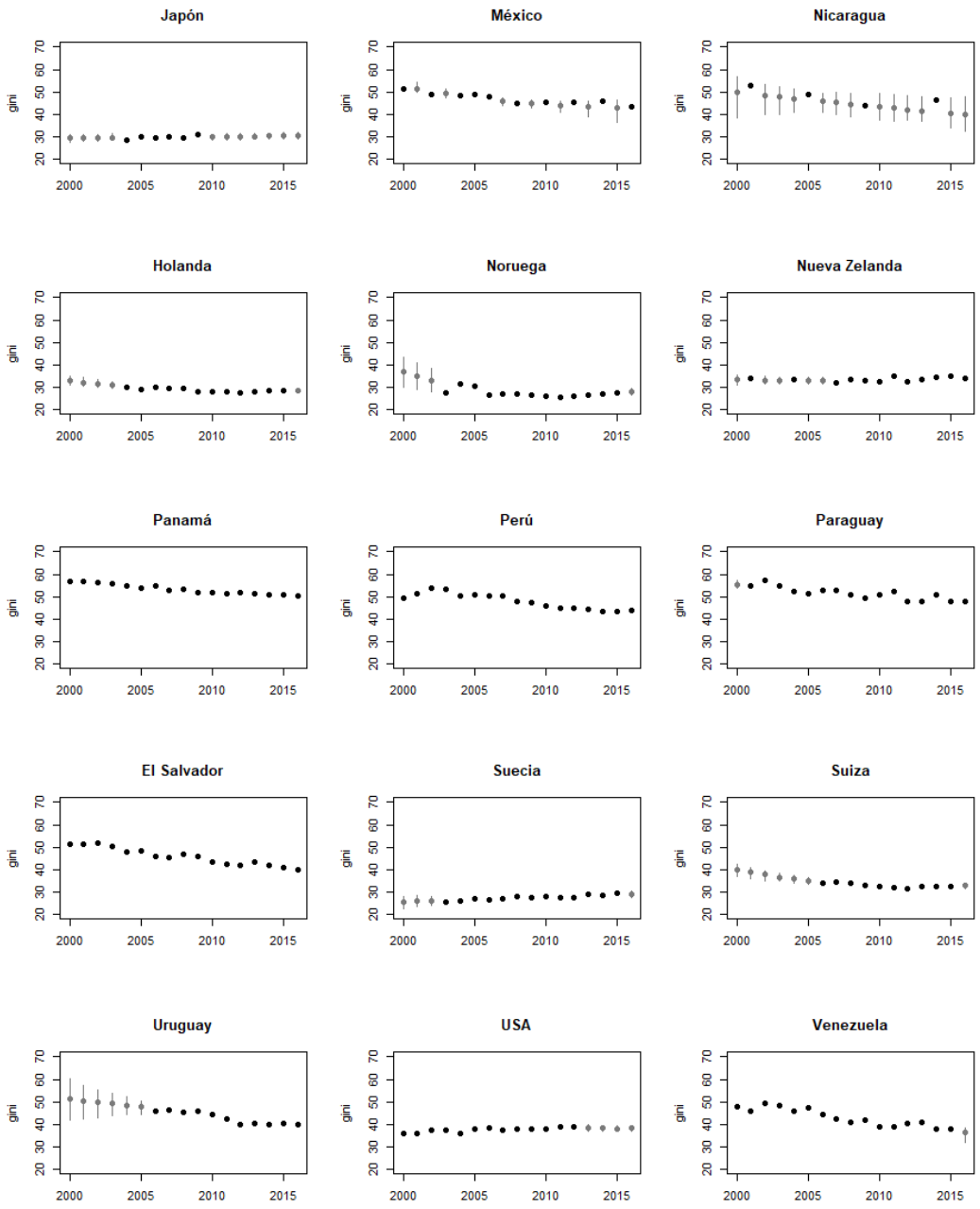

Figura 4 (continua). Datos observados vs datos estimados con dos restricciones a dos países
Fuente: Librería Amelia II, *R-project*, versión 1.7.5
Nota: En las series de tiempo los puntos negros son los valores observados y los puntos grises son la media de las distribuciones de la imputación. Las líneas grises representan el intervalo de confianza al 95% de la distribución de imputación.

4 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 y

Galván, 2007). Sus implicaciones en el análisis secundario de datos deben evaluarse con cautela, y con este trabajo se concluye que no existe el método de imputación ideal. Más bien lo que se requiere es que el investigador realice: 1) un diagnóstico del problema de datos faltantes; 2) configure el modelo de imputación de acuerdo con las necesidades de su tabla; y 3) 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; dependiendo de la aplicación particular, los métodos verosímiles pueden o no ser robustos cuando se apartan del supuesto del modelo; para algunos casos la estimación puede ser posible al apartarse del modelo pero requerirán el supuesto MCAR (Schafer & Graham, 2002). Así, conocer el mecanismo de generación de información faltante en la tabla de datos es crucial para el modelo de imputación.

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-Down, Hughes, Tilling & Heron, 2019) y Van Buuren (2018) se establece que 5% de datos faltantes ha sido sugerido como el límite mínimo para que la imputación múltiple funcione. En diversos artículos de divulgación científica se establece un intervalo de entre 10% y 40% de datos faltantes para asegurar un funcionamiento de la imputación múltiple, más de 40% de datos faltantes debería ser considerado más como una hipótesis, que como datos determinantes (Madley-Down et al., 2019). Los autores de la librería Amelia II, no establecen un porcentaje límite, solo mencionan que el bootstrapping de 5 iteraciones, es suficiente si los datos faltantes no son muchos (Honaker y 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 comparativa 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 (Kosen, Livne, Madai, Galinovic, Frey & Fiebach, 2019). Así, 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. Así que, como menciona Van Buuren no deberíamos dejar nuestras decisiones libradas únicamente a un software (2018). 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. Razón por la cual, se consideró necesario integrar información a priori, siguiendo la regla de introducir 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). Esto podría servir para futuros trabajos. 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 como proxy (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 múltiple ratio propuesto por Takahashi (2017), siempre que se contara con la información de una variable proxy para los mismos países en el mismo periodo.

En lo que respecta 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 tanto de manera gráfica como a través de la prueba de Kolmogorov Smirnov, entre mayores similitudes existan, se asume una mejor imputación. Además, en este trabajo se consideró un análisis gráfico de la serie de tiempo por país y los intervalos de confianza de los datos imputados. La combinación del análisis de ambos elementos 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

Abayomi, K., Gelman, A., & Levy, M. (2008). Diagnostics for Multivariate Imputations. *Royal Statistical Society*, *57*(3), 273-291. https://doi.org/10.1111/j.1467-9876.2007.00613.x

Arellano, M. y Bover, O. (1990). La econometría de datos panel. *Investigaciones Económicas*, *XIV* (1), 3-45.

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology*, *48*(1), 5-37. https://doi.org/10.1016/j.jsp.2009.10.001

Bell, M. L., Fiero, M., Horton, N. J., & Hsu, C.-H. (2014). *Handling missing data in RCTs; a review of the top medical journals*. BMC Med Res Methodol 14, 118, 2-8. https://doi.org/10.1186/1471-2288-14-118

Clavel, J., Merceron, G., & Escarguel., G. (2014). Missing data estimation in morphometrics: how much is too much? *Systematic biology*; 63(2), 203–18. doi: 10.1093/sysbio/syt100 PMID: 24335428

Cheema, J. R. (2014). A Review of Missing Data Handling Methods in Education Research. *Review of Educational Research*, *84*(4), 487-508. https://doi.org/10.3102/0034654314532697

Harrell, F. (2020). Nonparametric Missing Value Imputation using Random Forest: missForest. R package version 1.4.

Honaker, J., & King, G. (2010). What to Do about Missing Values in Time-Series Cross-Section Data. *America Journal of political science*, *54*(2), 561-581. https://doi.org/10.1111/j.1540-5907.2010.00447.x

Honaker, J., King, G., and Blackwell, M. (2018). AMELIA II: A Program for Missing Data. R package version 1.7.5.

Kowarik, A., & Templ, M. (2016). Imputation with the R Package VIM. *Journal of Statistical Software*, *74*(1), 1-16. https://doi.org/10.18637/jss.v074.i07

Kossen, T., Livne, M., Madai, V. I., Galinovic, I., Frey, D., & Fiebach, J. B. (2019). A framework for testing different imputation methods for tabular datasets. bioRxiv, 773762. https://doi.org/10.1101/773762

Leite, W., & Beretvas, S. (2010). The Performance of Multiple Imputation for Likert-type Items with Missing Data. *Journal of Modern Applied Statistical Methods*, 9(1),64-74 https://doi.org/10.22237/jmasm/1272686820

Medina, F. y Galván, M. (2007). *Imputación de datos: Teoría y práctica.* Naciones Unidas, CEPAL, Div. de Estadística y Proyecciones Económicas. Recuperado de http://www.cepal.org/publicaciones/xml/9/29949/LCL2772e.pdf

Madley-Down, P., Hughes, R., Tilling, K., & Heron, J. (2019). The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of clinical epidemiology*, 110,63-73. https://doi.org/10.1016/j.jclinepi.2019.02.016

Muñoz-Rosas, J. F. y Álvarez-Verdejo, E. (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.

Murray, J. (2018). Multiple Imputation: A Review of Practical and Theoretical Findings. https://arxiv.org/pdf/1801.04058.pdf

R Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Recuperado de https://www.R-project.org.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. John Wiley & Sons, New York.

Shafer, J., & Grahams, J. (2002). Missing data: our view of the state of the art. *Psychol Methods*. 7(2),147-77. PMID: 12090408.

Stekhoven, D.J. and Bühlmann, P. (2012), 'MissForest - nonparametric missing value imputation for mixed-type data', Bioinformatics, 28(1), 112-118, https://doi.org/10.1093/bioinformatics/btr597

Takahashi, M. (2017). Multiple ratio imputation by the EMB algorithm: theory and simulation. *Journal of Modern Applied Statistical Methods*, 16(1), 630-656. doi: 10.22237/jmasm/1493598840

Van Buuren, S. (2018). *Flexible Imputation of Missing Data* (Segunda). Taylor & Francis. Recuperado de https://stefvanbuuren.name/fimd/

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

Wood, A. M., White, I. R., & Thompson, S. G. (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), 368-376. https://doi.org/10.1191/1740774504cn032oa

Yusuke Y., Mai U., Kazushi M. & Masahiko G. (2020). Multiple imputation for longitudinal data in the presence of heteroscedasticity between treatment groups, *Journal of Biopharmaceutical Statistics*, 30:1, 178-196, DOI: 10.1080/10543406.2019.1632878

Zhang, Z. (2015). Missing data imputation: Focusing on single imputation. A*nnals of Translational Medicine*, 1-8. https://doi.org/10.3978/j.issn.2305-5839.2015.12.38