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

LIECD

LABORATORIO DE INVESTIGACIÓN EN ESTADÍSTICA Y CIENCIA DE DATOS

Imputación de datos faltantes.

Un caso de estudio con datos panel

LIECD
LABORATORIO DE INVESTIGACIÓN EN ESTADÍSTICA Y CIENCIA DE DATOS

- noblemática datos faltantes
- Estructura datos panel
- mputación múltiple
- Estudio de caso
- cances, limitaciones y pregun

▶mática datos faltantes

El estudio consistía en demostrar que el tipo de capitalismo modificaba la relación entre la desigualdad y la confianza en las instituciones y algunas otras variables*.

Pero, una de las variables (índice de Gini) lucía así

	A	B	C	D	E	F
1	Ind	Var 1	Var 2	Var 3	...	Var m
2	Ind 1	23	52	12		n.a.
3	Ind 2	45	61	n.a.		5
4	Ind 3	n.a.	43	15		7
5	...					
6	Ind n	34	n.a.	19		7

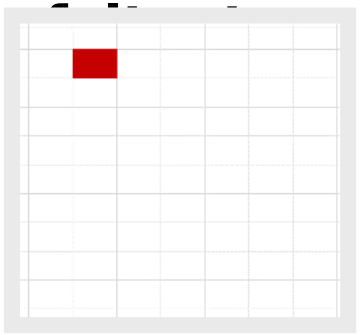
El artículo ya fue aceptado para su publicación.

▶mática datos faltantes

Existe una gran variedad de técnicas para imputar datos faltantes

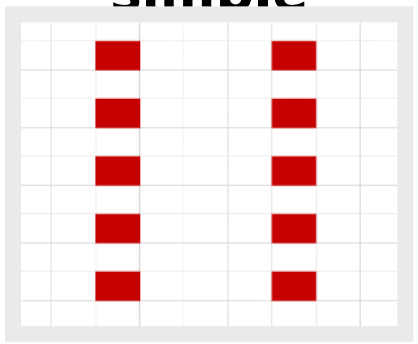
¿Cuál usar?

Borrar al individuo que contiene datos



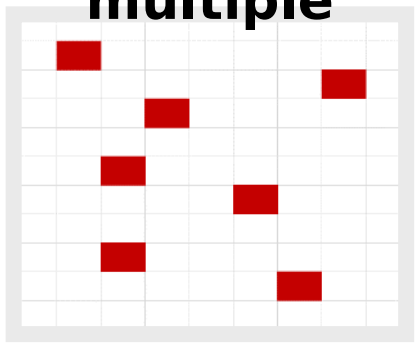
Depende de la cantidad de

Métodos de imputación simple



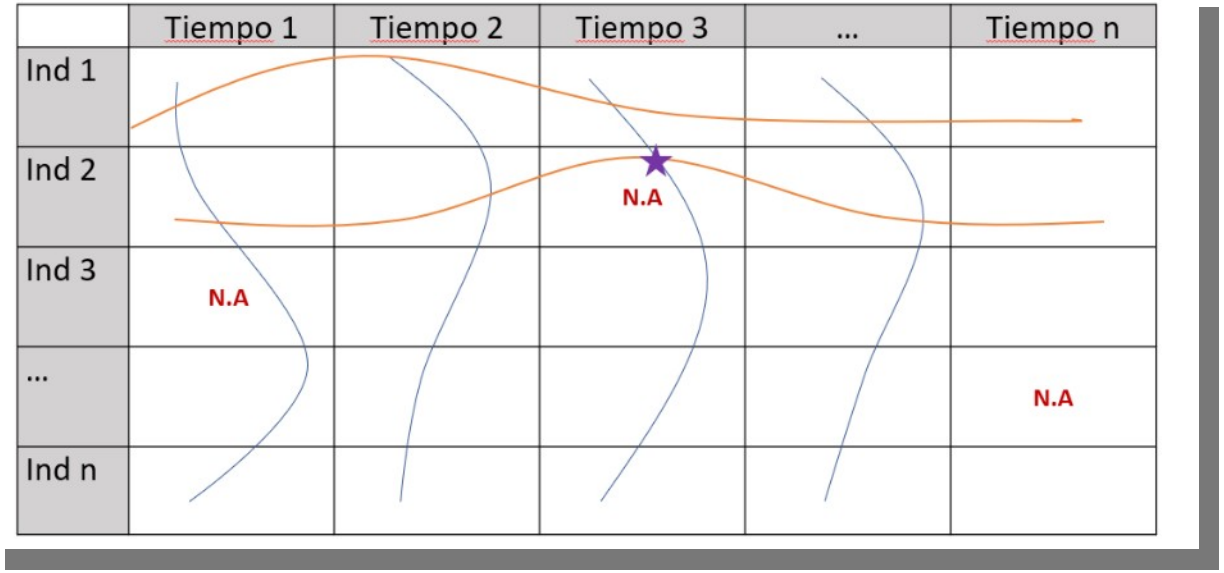
Patrones monótonos
Media, regresión, ratio

Métodos de imputación múltiple



Patrones arbitrarios

ctura datos panel



- Los datos panel son un arreglo matricial, donde a los individuos (países, estados, genes, etc.) se les mide una o más variables a lo largo del tiempo.

- La variación en la temporalidad y la variación en los individuos son igual de importantes para el estudio.
- En el caso de datos panel, dado el patrón arbitrario que suelen seguir los datos
  faltantes, se sugieren los métodos de imputación múltiple.

putación múltiple

El proceso generalizado de imputación múltiple remplaza 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 con alguno de estos valores. (Rubin, 1987)

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

putación múltiple

Los **datos verosímiles**, se calculan de acuerdo a una serie de supuestos que varían según el algoritmo usado (por eso es que hay diversidad de procesos en la imputación múltiple).

Tipos de mecanismos faltantes:

- MCAR (Missing completely at Random)
- MAR (Missing at random)
- NI (Not ingnorable)

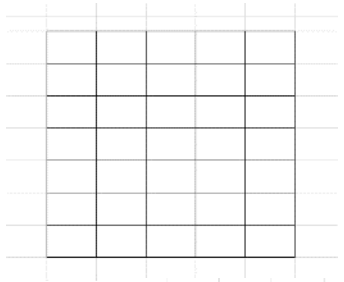
Algoritmos

- MCMC (Markov Chain Monte Carlo)- Librería norm en R y Proc Mi (SAS).
- FCS (Fully Conditional Specification)- Librería mice en R y Missing values en SPSS.
- EMB (Expectation-Maximization with Bootstraping)- Librería Amelia II en R.

putación múltiple

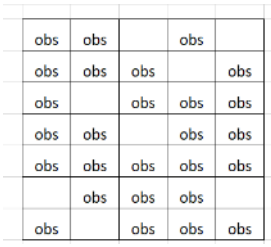
En el caso de Honaker, King, y Blackwell (2018), se utilizan los siguientes supuestos:

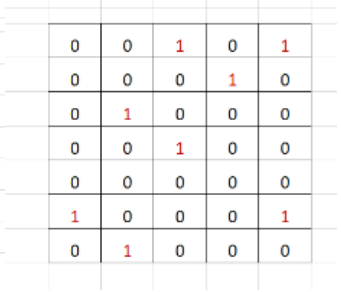
1) $D_{n,k} \sim N_k(\mu, \Sigma)$



D

Dobs





Dmis

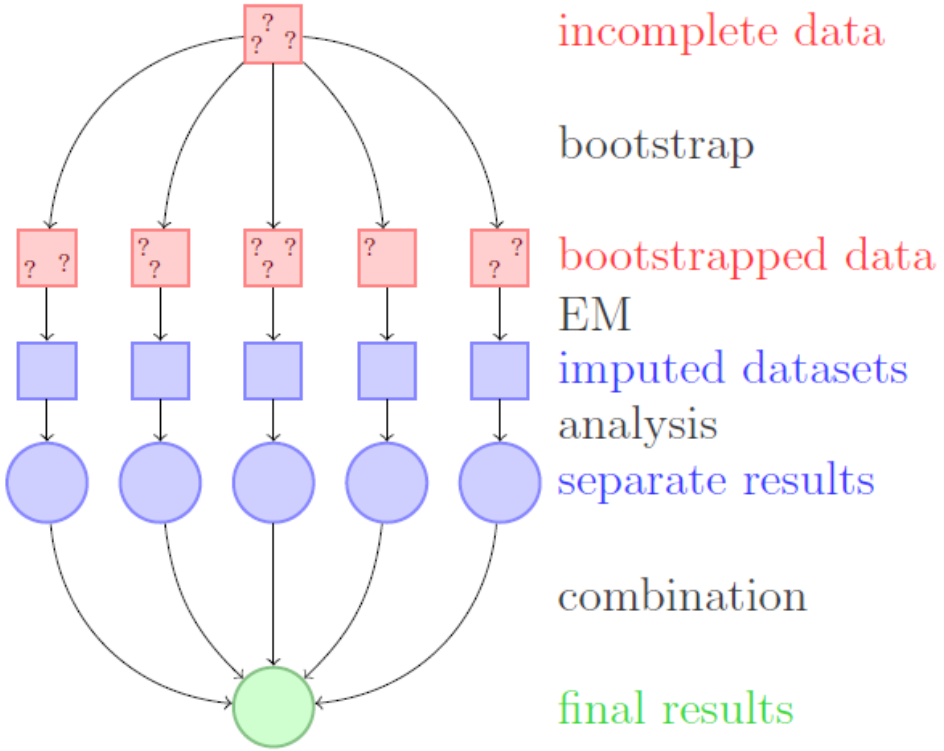
2) Los datos faltan aleatoriamente (MAR).

Los datos faltantes dependen únicamente de los datos observados.

Por lo que, la distribución a posteriori será:

$$p(D_{mis} \mid D_{obs}, \theta)$$

putación múltiple



En R

Library (Amelia)

- Especificar la variable de temporalidad.
- Especificar la variable de países.
- Permite agregar restricciones al sistema.

Figura 1. Esquema de la imputación múltiple, tomado de Honaker,King y Blackwell (2018).

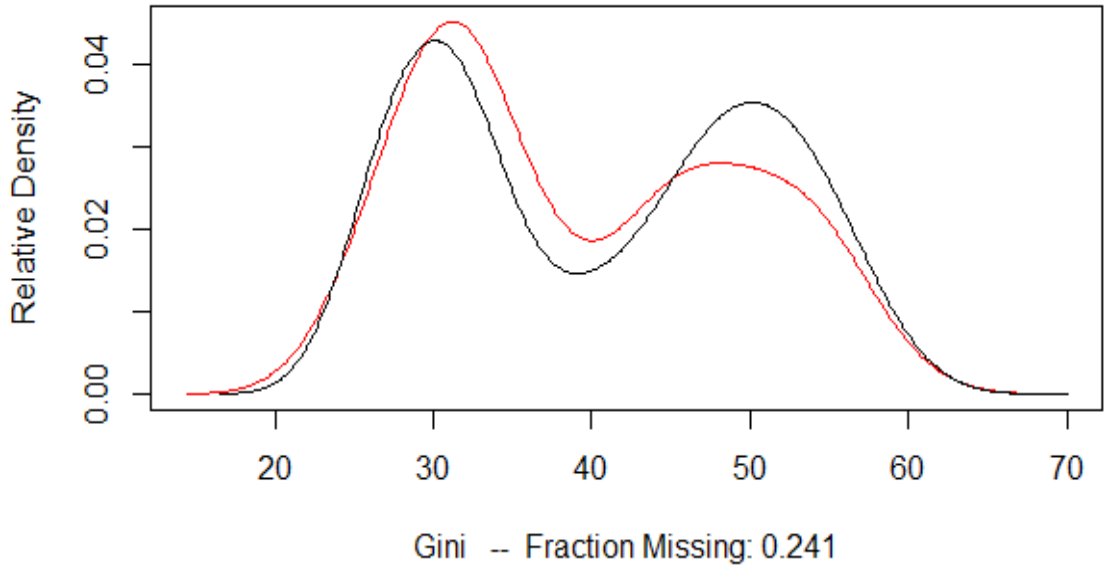
Honaker, J., King, G., y Blackwell, M. (2018). AMELIA II: A Program for Missing Data.
R package version 1.7.5.

tudio de caso

Los datos obtenidos corresponden 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%). Se utilizó m=5

Los resultados del primer ejercicio de imputación, muestran dos poblaciones al interior de la tabla.

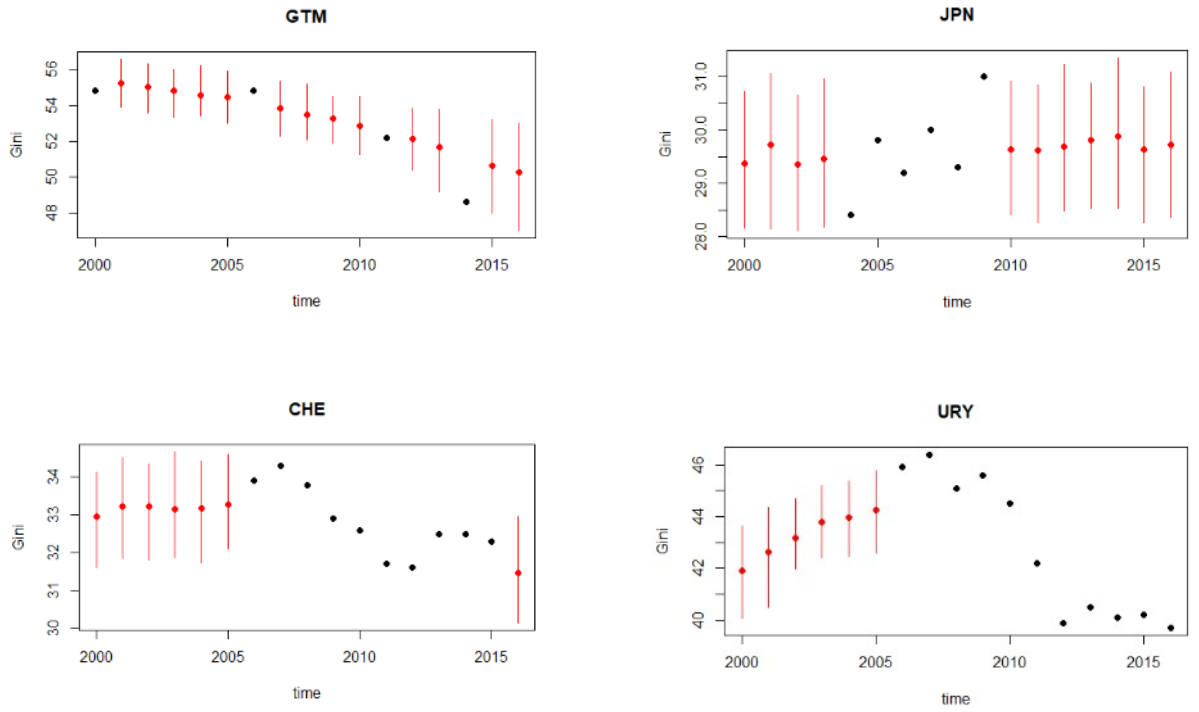
La línea negra es la distribución de los datos observados y la línea roja de los datos imputados.



*Los datos fueron recopilados y proporcionados por el Dr. Edgar J. Saucedo-Acosta y
la Mtra. Nallelly Patricia Bolaños.*

tudio de caso

Sin embargo...



Al visualizar individualmente a cada país, se destacaron los casos de Guatemala, Japón, Suiza y Uruguay, por presentar irregularidades en la estimación del coeficiente de Gini.

**Los datos fueron recopilados y proporcionados por el Dr. Edgar J. Saucedo-Acosta y la Mtra. Nallelly Patricia Bolaños.*

tudio de caso

Se decidió introducir restricciones al sistema ¿Cuáles son las más adecuadas?
Al integrar una restricción (media y desviación estándar en algún país) toda la imputación cambia.

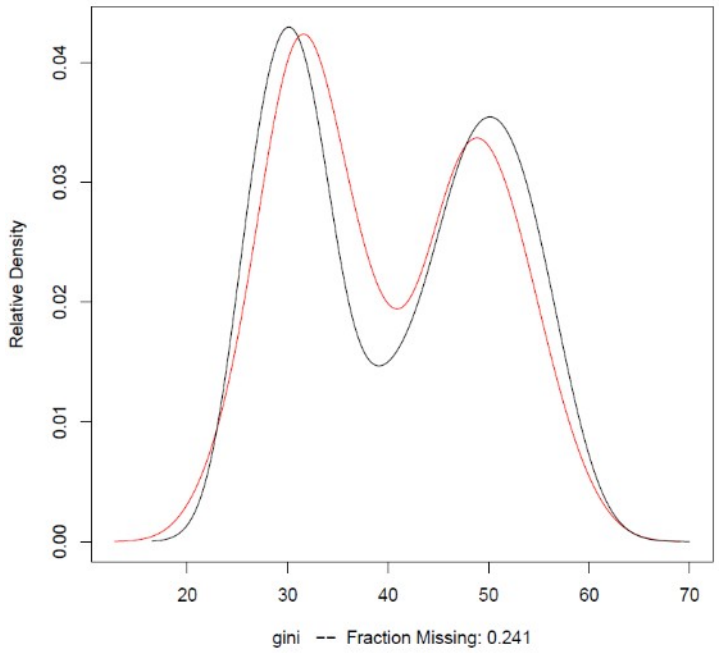
- Se inició con el primer país Guatemala, y se compararon los datos. Guatemala no presentaba problemas, pero ahora otros países agrandaban sus intervalos.

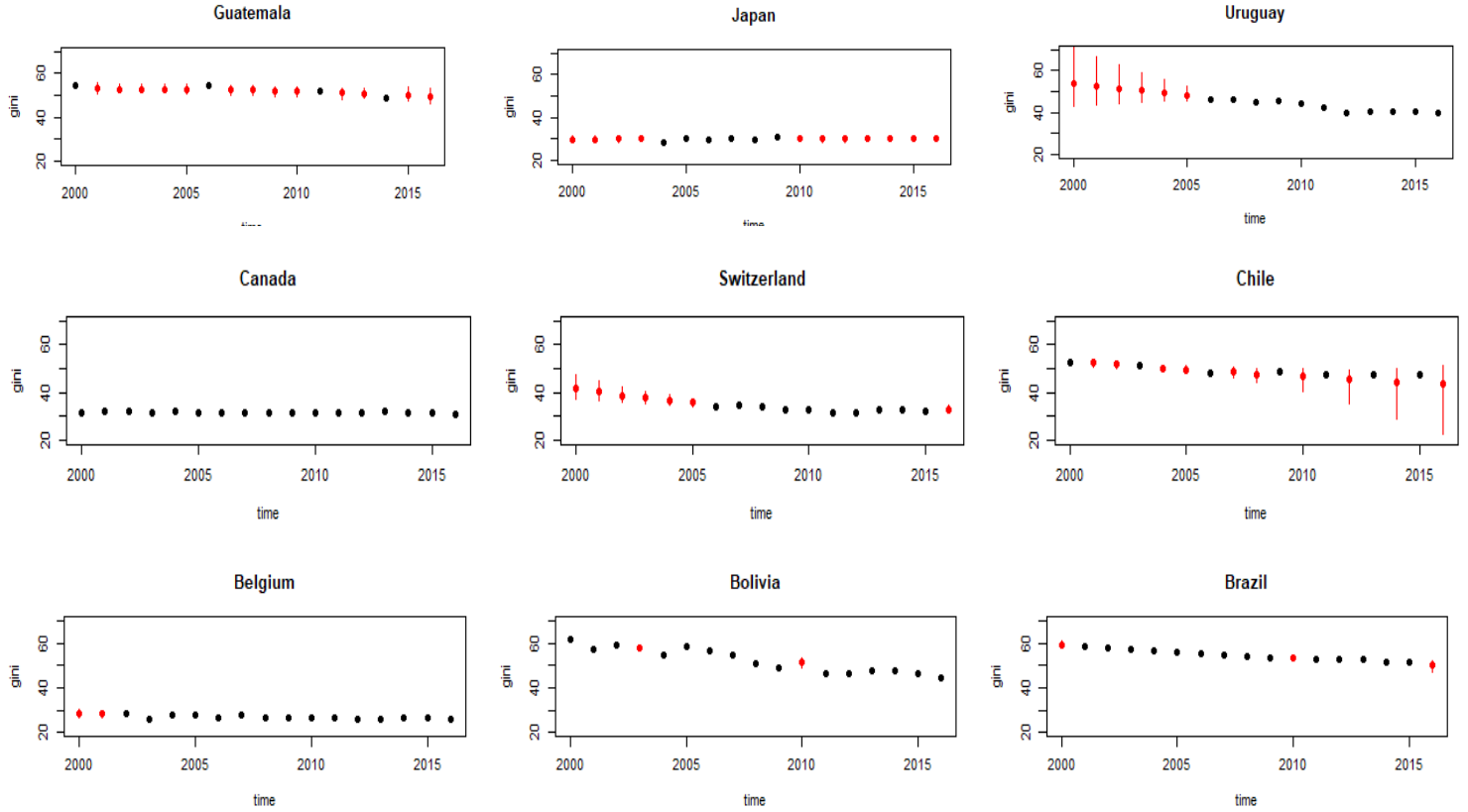
- Se integraron restricciones a Guatemala y Japón, los resultados parecían más estables que la primera vez.

- Se integraron restricciones a Guatemala, Japón y Suiza, pero los resultados empeoraban.

tudio de caso

Con las restricciones integradas a Guatemala y Japón.





**Los datos fueron recopilados y proporcionados por el Dr. Edgar J. Saucedo-Acosta y la Mtra. Nallelly Patricia Bolaños.*

ances, limitaciones y pregu

Alcances

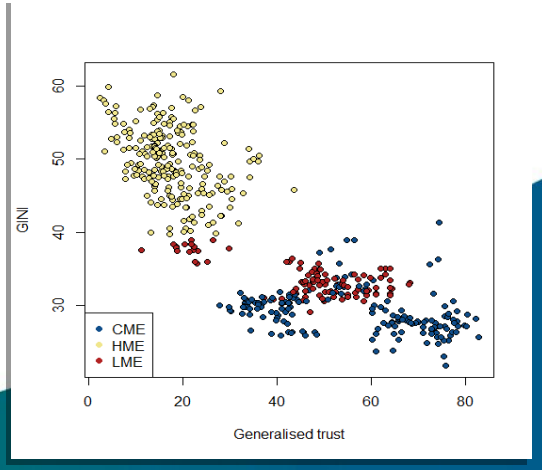
- La meta de la imputación generalmente no es encontrar el valor exacto para cada dato faltante, sino **hacer posible una inferencia estadística** (Takahashi, 2017).

- La imputación múltiple ha demostrado ser superior a métodos como el de la media, regresión (Baraldi & Enders (2010) and Cheema (2014)). Además Leite & Beretvas (2010) han probado que es robusta al supuesto de normalidad.

- La tabla de datos se completó y se utilizó para un estudio posterior

Tabla 1. Países con su coeficiente de Gini origin

País	2000	2001	2002	2003	2004	2005	2006	2007
Argentina	51.1	53.3	53.8	50.7	48.3	47.7	46.6	46.3
Australia	31	31.1	31.1	30.9	30.6	31.6	31.4	30.7
Austria	27.9	28.7	30.1	29.5	29.8	28.7	29.6	30.6
Belgium	27.1	30.1	28.3	26.1	28	27.8	26.3	27.5
Bolivia	61.6	57.4	59.3	57.6	55	58.5	56.7	54.5
Brazil	61.1	58.4	58.1	57.6	56.5	56.3	55.6	54.9
Canada	31.7	31.8	31.8	31.6	32.2	31.7	31.6	31.6
Switzerland	32.7	32.6	34	33.6	33.8	33.3	33.9	34.3
Chile	52.8	52.8	52	51.5	51.4	51	48.2	48.9
Colombia	58.7	57.2	55.8	53.4	54.8	53.7	53.9	53.4
Costa Rica	47.4	51.6	51.9	49.3	48.3	47.5	49.4	49.3



ances, limitaciones y pregu

Alcances

- La librería Amelia es veloz*, lo cual permite hacer las pruebas que uno considere necesarias, se aprovechó esta herramienta.

- La imputación de datos de una tabla de datos como esta, es de detalles finos. Se procuró que fuera lo mejor posible.

*Debido a la implementación del algoritmo que los autores proponen. Se han logrado imputar datos de tablas con hasta 240 variables y 32000 casos (p.p. 565)

ancias, limitaciones y pregu

Limitaciones

- Se utilizó una m=5, porque es la sugerida por los autores que dicen: “A menos que los datos faltantes sean ¿muchos?, m=5 (el proceso automático) es ¿probablemente? adecuado” 135 datos faltantes de 561 casos es ¿poco? ¿mucho?

- Se utilizó el gráfico de densidad como estrategia de corroboración, sin embargo ¿Qué tan eficiente es este gráfico para decidir si los datos imputados son “confiables” o no lo son?

- Además de el gráfico de densidad, se graficaron las series de tiempo por país y se evalúo caso por caso. Se comprobó con el economista experto para ver si los datos imputados parecían realistas, dijo que sí.

ancias, limitaciones y pregu

Preguntas

- ¿Qué porcentaje de datos faltantes es recomendable para hacer una imputación de esta naturaleza?
- ¿Qué regla seguir para establecer las restricciones al sistema? ¿La que seguimos es la mejor?
- Existe una propuesta de Takahashi (2017) con el nombre de “multiple ratio imputation”, donde propone utilizar una variable proxy y utilizar imputación múltiple ¿funcionaría para este caso?
- La imputación de datos faltantes es un tema especial por que es particularizado a cada tabla de datos, no hay reglas generales que se puedan aplicar.

LIECD
LABORATORIO DE INVESTIGACIÓN EN ESTADÍSTICA Y CIENCIA DE DATOS

¡Gracias !