Badler, Clara
Alsina, Sara
Beltrán, Celina
Puigsubirá, Cristina
Vitelleschi, Ma. Susana
Instituto de Investigaciones Teóricas y Aplicadas, Escuela de Estadística.

INFLUENCIA DEL MECANISMO, DE LA ESTRUCTURA DE LOS DATOS Y DEL TRATAMIENTO DE LA INFORMACIÓN INCOMPLETA EN EL ANALISIS ESTADÍSTICO DE LA EPH[1]

1. INTRODUCCIÓN

La información incompleta es un problema que se presenta al trabajar con bases de datos. Ha sido demostrado que el conocimiento, o ausencia del conocimiento de los mecanismos que hacen que ciertos valores estén perdidos, es fundamental en la elección de un tratamiento apropiado y una adecuada interpretación de los resultados del análisis.

El cumplimiento del supuesto que los datos faltantes han sido perdidos completamente al azar o simplemente al azar, según las definiciones propuestas por Little y Rubin, garantizan la posibilidad de aplicar exitosamente muchas técnicas estadísticas sin recurrir a modelos complicados con datos faltantes. Sin embargo, debe tenerse en cuenta, además, que el mecanismo y el tratamiento aplicado para solucionar el problema de falta de información pueden afectar la estructura de los datos de manera tal de alterar los resultados.

La consideración de diferentes situaciones resultantes puede realizarse mediante la aplicación de procedimientos de simulación de pérdidas y el uso posterior de técnicas estadísticas para describir e interpretar un determinado fenómeno.

En este trabajo se comparan los resultados obtenidos al realizar un análisis mediante la técnica de componentes principales a variables registradas en la EPH del Aglomerado Gran Rosario para las ondas Octubre de 1997 y 1998 en las que se simulan y tratan pérdidas originadas por distintos mecanismos.

2. METODOLOGÍA

2.1. Conjuntos de datos y Esquemas de Pérdida

Una matriz Y_{nxp} correspondiente a una base de datos de observaciones sobre n individuos muestreados a partir de una población de tamaño N que presenta información

[1] Proyectos: PICT N° 0200095-01996 de la ANPCyT y PID N° 19/E045 de la SECyT.

incompleta en algunas de las p variables en estudio, puede representarse en forma esquemática como lo indica la figura 1, visualizando en la zona punteada la población en estudio y en la rayada los valores observados de las variables en estudio.

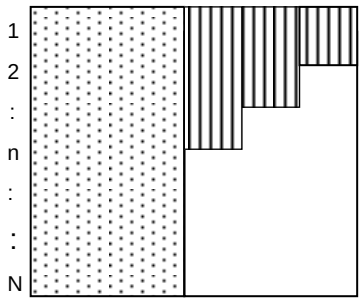

Figura 1: Esquema de información incompleta

Se denomina esquema de pérdida a la forma que toma esta última área al representar a las observaciones para las que no se ha registrado información en cada variable.

La técnica de simulación aplicada a la generación de pérdidas permite construir diferentes esquemas.

2.2. Mecanismo de pérdida

Mediante simulación se obtienen esquemas de pérdida en distintas variables con el objeto de describir la importancia de su consideración al realizar el tratamiento de la información.

Cuando se supone que la probabilidad de respuesta es independiente de las variables observadas completamente y de las observadas en forma parcial, se dice que los datos perdidos están perdidos completamente al azar (MCAR).

Cuando se supone que la probabilidad de respuesta depende de las variables observadas completamente pero no de las observadas en forma parcial, se dice que los datos perdidos están perdidos al azar (MAR).

Cuando la probabilidad de respuesta depende de las variables observadas en forma parcial y probablemente de las observadas completamente, el mecanismo de pérdida no es ni MAR ni MCAR.

Para el tratamiento de datos con información faltante, en general, las técnicas clásicas requieren el supuesto MCAR. Para evaluarlo, se utiliza un test propuesto por Little basado en la estadística del cociente de verosimilitud, cuya distribución es compleja para esquemas generales de pérdidas pero se simplifica para el monótono.

En un esquema de pérdida monótono con tres variables, en el que se observa:

- Y_1 en forma completa (n unidades),
- Y_2 en n_2 ($<n$) unidades e
- Y_3 en n_3 ($<n_2$) unidades,

se calcula la estadística:

$$d^2 = \frac{SSB_1}{MST_1} + \frac{SSB_{2.1}}{MST_{2.1}} = \frac{(n-1)\,2\,F_1}{(n-3)+2\,F_1} + \frac{(n_2-1)\,F_{2.1}}{(n_2-2)+F_{2.1}}$$

siendo:

- SSB_1 y $SSB_{2.1}$ las sumas de cuadrados entre grupos para las variables Y_1 e Y_2 ajustando por Y_1, respectivamente,
- MST_1 y $MST_{2.1}$ los cuadrados medios totales para las variables Y_1 e Y_2 ajustando por Y_1, respectivamente,
- F_1 y $F_{2.1}$ las estadísticas F del análisis de la variancia de Y_1 sobre el esquema de pérdida y la estadística F del análisis de covariancia de la variable Y_2 sobre los grupos determinados por los esquemas donde Y_2 es observada, ajustando por Y_1, respectivamente.

La estadística d^2 tiene distribución chi-cuadrado con tres grados de libertad bajo el supuesto de normalidad de las variables y su distribución es la misma, asintóticamente, independientemente de la de las variables.

2.3. Estudio de la Estructura de los datos

Es de interés estudiar la estructura de los datos con respecto a las relaciones entre las variables, sobre todo cuando la técnica utilizada para el tratamiento y análisis de los datos con información incompleta se basa en dichas relaciones. Si bien es posible realizar este análisis mediante distintos procedimientos, se opta por el estudio de las correlaciones.

Para ello se procede de la siguiente manera:

- Se divide el conjunto de datos originales en grupos con respecto a una determinada variable (Y_j). En este caso, la variable de interés presenta pérdida de información simulada.
- Se utiliza como punto de corte para dicha división los percentiles de la variable Y_j.
- Se calcula la correlación entre las variables que presentan información incompleta en cada uno de los subconjuntos de datos resultantes de la división y la matriz de correlaciones para el conjunto completo de datos.
- Se analiza la tendencia que presentan las correlaciones en los distintos grupos.
- Se investiga de qué forma se ven afectados los resultados obtenidos de la aplicación de técnicas estadísticas basadas en las correlaciones.

2.4. Tratamiento de información faltante

Casos Completos

Consiste en utilizar sólo las unidades que son observadas en forma completa y descartar del análisis aquellas que presentan valores perdidos en alguna de las variables.

Si el supuesto MCAR se satisface, el conjunto de unidades con información completa puede ser considerado como una submuestra aleatoria de los datos originales y las estimaciones obtenidas a partir de ellos serán insesgadas. Caso contrario, las estimaciones

muestrales pueden resultar sesgadas y la naturaleza de los sesgos dependerá del mecanismo de pérdida.

Si bien ésta es una técnica muy fácil de aplicar, presenta además la desventaja de la disminución del tamaño muestral.

Imputación por el método de Buck

Este método propone la estimación de los valores perdidos mediante predicciones por regresión.

Si las variables Y_1, Y_2,...,Y_p tienen una distribución normal multivariada con media μ y matriz de variancias y covariancias Σ, las variables con información faltante se pueden escribir como funciones de regresión lineal de las variables observadas, donde los coeficientes son funciones conocidas de μ y Σ.

El método de Buck propone la estimación de μ a través de la media muestral y de Σ a través de la matriz de variancias y covariancias, ambas obtenidas a partir de las unidades completamente observadas. Con estas estimaciones se calculan las regresiones de las variables con información incompleta sobre las variables observadas en cada unidad y los valores resultantes son utilizados como estimación de los perdidos:

$$\hat{y}_{ij.per} = \hat{\beta}' \mathbf{Y_{obs.i}}$$

donde $y_{ij.per}$ es el valor faltante de la variable Y_j en el i-ésima unidad, $Y_{obs.i}$ es el vector de los valores observados en dicha unidad y $\hat{\beta}$ es el vector de coeficientes estimados de regresión.

Dado que este método proyecta los valores de las variables sobre la ecuación de regresión, se supone que la relación entre las mismas es lineal. Pero, si se realiza extrapolación de las unidades completas, este supuesto resulta débil.

Los estimadores obtenidos a partir del conjunto que ahora incluye los valores así imputados, son consistentes bajo el supuesto MCAR; bajo ciertas condiciones también lo son cuando el mecanismo de pérdida es MAR.

Este método puede ser aplicado también a variables categóricas reemplazándolas por tantas variables dummy como categorías menos una.

2.5. Componentes Principales

Con este procedimiento se trata de explicar la variación de un conjunto de p variables (Y) a través de un número reducido de variables no observables (CP) llamadas componentes, que son combinaciones lineales de las variables originales y no están correlacionadas entre sí. Se obtienen tantas componentes como el número de variables originales:

$$CP_j = a_{1j} Y_1 + a_{2j} Y_2 + ... + a_{pj} Y_p \qquad j=1,2,...,p$$

siendo a_{ij} los coeficientes de los autovectores normalizados correspondientes a los autovalores (λ_j) de la matriz de variancias y covariancias o la matriz de correlaciones.

Las componentes son dispuestas en orden decreciente al autovalor que la originó, el cual es la variancia de dicha componente:

$$Var(CP_j) = \lambda_j \qquad j=1,2,...,p$$

Para decidir cuántas componentes serán retenidas en el análisis no existe un único criterio sino que hay que tener en cuenta varios de ellos y decidir en función de los resultados obtenidos.

3. MATERIAL

Se trabaja con información de los desocupados con ocupación anterior detectados en la EPH del Aglomerado Gran Rosario para las ondas de un mismo mes, octubre, de los años 1997 y 1998.

De los dos archivos que configuran la información de la encuesta, se selecciona el correspondiente a "personas" y las variables:

- EDAD: años cumplidos a la fecha del relevamiento.
- TTOA: "Tiempo transcurrido desde que dejó la ocupación anterior" (en días). Esta variable se construye mediante una combinación lineal de las variables relevadas referidas al tiempo en años, meses y días transcurridos desde que dejó la ocupación anterior.
- TBE: "Tiempo de búsqueda de empleo" (en días). Esta variable se construye mediante una combinación lineal de las variables relevadas "tiempo en meses que busca empleo" y "tiempo en días que busca empleo".

4. RESULTADOS

4.1. Simulación de pérdida de información

Mediante el programa estadístico SAS, se generan pérdidas en las variables TBE y TTOA para la obtención del esquema monótono presentado en la figura 2, según los siguientes mecanismos:

- MCAR: se seleccionan al azar las unidades que tendrán información faltante en las variables TTOA y TBE.
- MAR: se seleccionan al azar unidades consideradas como "potenciales pérdidas" en TTOA y TBE. Se consideran como perdidas aquellas en las que la variable EDAD supera un cierto valor. Los valores perdidos dependen de la variable EDAD pero no de las variables sujetas a no respuesta.
- NO ALEATORIO-valores máximos: se seleccionan las unidades con los mayores valores de TTOA y se generan pérdidas en TTOA y TBE. De los restantes valores de TTOA, se seleccionan los mayores y se generan pérdidas en TBE.
- NO ALEATORIO-valores mínimos: se seleccionan las unidades con los menores valores de TTOA y se generan pérdidas en TTOA y TBE. De los restantes valores de TTOA, se seleccionan los menores y se generan pérdidas en TBE.

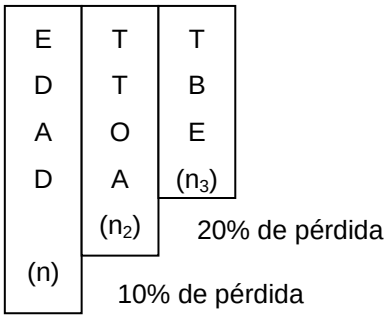

Figura 2: Esquema de pérdida monótono con tres variables.

4.2. Evaluación del mecanismo de pérdida

Para los distintos mecanismos en las dos ondas se aplica el test propuesto por Little para evaluar la hipótesis que los datos perdidos están perdidos completamente al azar.

Cuadro 1: Test para evaluar el supuesto MCAR.

ONDA	MECANISMO DE PERDIDA	F_1	$F_{2.1}$	d^2 (GL=3)	PROB	DECISIÓN*
Octubre 1997 (N=171)	MCAR	1.1636	0.639	2.96	0.40	No se rechaza
	MAR	35.9433	0.031	50.975	$4.95E^{-11}$	Se rechaza
	Valores máx.	0.7435	182.77	85.02	$2.57E^{-18}$	Se rechaza
	Valores mín.	0.01	8.13	7.9	0.04	Se rechaza
Octubre 1998 (N=142)	MCAR	0.0518	0.456	0.56	0.90	No se rechaza
	MAR	25.65	2.346	40.33	$9.06E^{-9}$	Se rechaza
	Valores máx.	2.82	400.28	102.09	$5.52E^{-22}$	Se rechaza
	Valores mín.	0.1035	0.759	0.97	0.81	No se rechaza

* α=0.05

En la onda de octubre de 1997 se rechaza dicha hipótesis en todos los mecanismos que no son MCAR, a diferencia de la onda octubre de 1998 en la que no se rechaza la hipótesis cuando el mecanismo genera pérdidas en las unidades con valores mínimos de la variable TTOA.

Se consideran para cada onda y para cada mecanismo, dos conjuntos de datos con:

a) las unidades que presentan información en las tres variables (Casos Completos)

b) las unidades que presentan información en las tres variables y las unidades que fueron imputadas mediante el método de Buck (Imputación Buck).

En cada caso, se estudia la estructura de los datos y se aplica la técnica de componentes principales basada en la matriz de correlaciones.

4.3. Estudio de la estructura de los datos.

Los cuadros 2 y 3 presentan las matrices de correlaciones de los datos originales y los distintos conjuntos de datos obtenidos luego de generar pérdidas, en las dos ondas en estudio.

Cuadro 2: Matrices de correlaciones según mecanismo de pérdida y tratamiento para la onda octubre de 1997.

MECANISMO	DATOS	MATRIZ DE CORRELACIONES			
SIN PERDIDAS	**ORIGINALES**		EDAD	TTOA	TBE
		EDAD	1	0.1	0.19
		TTOA	0.1	1	0.43
		TBE	0.19	0.43	1
MCAR	**CASOS COMPLETOS**		EDAD	TTOA	TBE
		EDAD	1	0.09	0.15
		TTOA	0.09	1	0.40
		TBE	0.15	0.40	1
	IMPUTACION BUCK		EDAD	TTOA	TBE
		EDAD	1	0.10	0.16
		TTOA	0.10	1	0.40
		TBE	0.16	0.40	1
MAR	**CASOS COMPLETOS**		EDAD	TTOA	TBE
		EDAD	1	0.11	0.17
		TTOA	0.11	1	0.40
		TBE	0.17	0.40	1
	IMPUTACION BUCK		EDAD	TTOA	TBE
		EDAD	1	0.10	0.20
		TTOA	0.10	1	0.40
		TBE	0.20	0.40	1
VALORES MAXIMOS	**CASOS COMPLETOS**		EDAD	TTOA	TBE
		EDAD	1	0.15	0.20
		TTOA	0.15	1	0.85
		TBE	0.20	0.85	1
	IMPUTACION BUCK		EDAD	TTOA	TBE
		EDAD	1	0.15	0.19
		TTOA	0.15	1	0.92
		TBE	0.19	0.92	1
VALORES MINIMOS	**CASOS COMPLETOS**		EDAD	TTOA	TBE
		EDAD	1	0.12	0.24
		TTOA	0.12	1	0.35
		TBE	0.24	0.35	1
	IMPUTACION BUCK		EDAD	TTOA	TBE
		EDAD	1	0.12	0.26
		TTOA	0.12	1	0.36
		TBE	0.26	0.36	1

Cuadro 3: Matrices de correlaciones según mecanismo de pérdida y tratamiento para la onda octubre de 1998.

MECANISMO	DATOS	MATRIZ DE CORRELACIONES			
SIN PERDIDAS	**ORIGINALES**		EDAD	TTOA	TBE
		EDAD	1	0.18	0.11
		TTOA	0.18	1	0.02
		TBE	0.11	0.02	1
MCAR	**CASOS COMPLETOS**		EDAD	TTOA	TBE
		EDAD	1	0.15	0.09
		TTOA	0.15	1	0.01
		TBE	0.09	0.01	1
	IMPUTACION BUCK		EDAD	TTOA	TBE
		EDAD	1	0.15	0.11
		TTOA	0.15	1	0.01
		TBE	0.11	0.01	1
MAR	**CASOS COMPLETOS**		EDAD	TTOA	TBE
		EDAD	1		0.14
		TTOA	0.18	1	0.07
		TBE	0.14	0.07	1
	IMPUTACION BUCK		EDAD	TTOA	TBE
		EDAD	1	0.17	0.17
		TTOA	0.17	1	0.09
		TBE	0.17	0.09	1
VALORES MAXIMOS	**CASOS COMPLETOS**		EDAD	TTOA	TBE
		EDAD	1	0.01	0.04
		TTOA	0.01	1	0.78
		TBE	0.04	0.78	1
	IMPUTACION BUCK		EDAD	TTOA	TBE
		EDAD	1	-0.01	-0.01
		TTOA	-0.01	1	0.93
		TBE	-0.01	0.93	1
VALORES MINIMOS	**CASOS COMPLETOS**		EDAD	TTOA	TBE
		EDAD	1	0.20	0.14
		TTOA	0.20	1	-0.01
		TBE	0.14	-0.01	1
	IMPUTACION BUCK		EDAD	TTOA	TBE
		EDAD	1	0.19	0.15
		TTOA	0.19	1	-0.01
		TBE	0.15	-0.01	1

En las dos ondas, se observa en la matriz de correlaciones que con el mecanismo NO ALEATORIO, por el cual se pierden los valores máximos, la correlación entre TTOA y TBE se ve sobrestimada, considerablemente, tanto en los datos imputados como en los casos completos, debiéndose a que la relación entre las variables no es la misma para los que tienen valores bajos y altos de TTOA.

Para detectar características de la estructura de las relaciones entre las variables en las cuales se generaron pérdidas, se divide la muestra original de cada onda en grupos, utilizando como puntos de corte los percentiles 20, 40, 60 y 80 de la distribución de la variable TTOA y se calcula la correlación entre las variables TTOA y TBE en cada uno de ellos.

Cuadro 4: Correlación de TTOA y TBE en cada grupo para la onda octubre 1997.

GRUPO DE TTOA	n	COEF. DE CORRELACION TTOA*TBE
1	34	0.93
2	34	0.96
3	34	0.83
4	34	0.57
5	35	-0.16

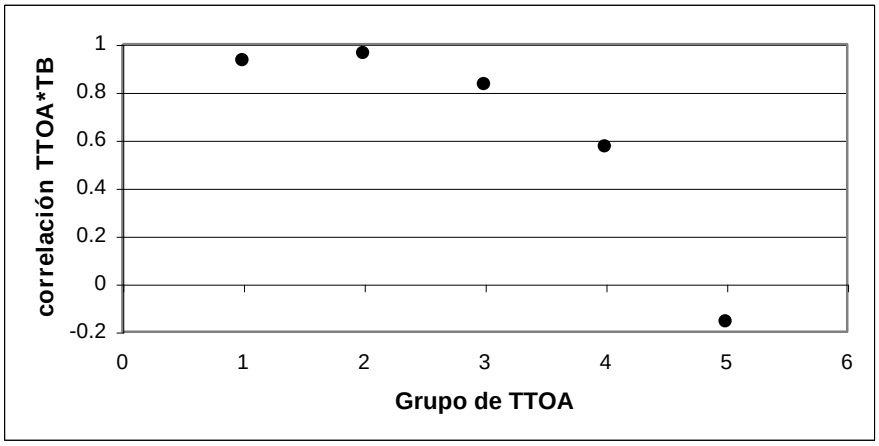

Figura 3: Correlación entre TTOA y TBE según grupo para la onda de octubre 1997.

Cuadro 5: Correlación de TTOA y TBE en cada grupo para la onda octubre 1998.

GRUPO DE TTOA	n	COEF. DE CORRELACION TTOA*TBE
1	28	0.92
2	28	0.9
3	28	0.72
4	29	0.38
5	29	-0.45

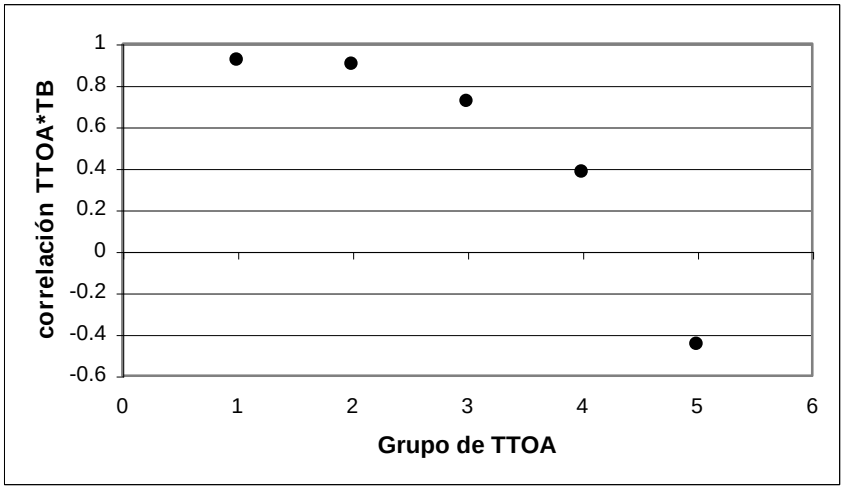

Figura 4: Correlación entre TTOA y TBE según grupo para la onda de octubre 1998.

Se puede observar en las dos ondas que a medida que aumenta el valor de TTOA, la correlación entre las variables disminuye.

Cuando se pierden los valores más altos de TTOA, se utiliza la información de las restantes unidades que presenta una correlación entre las variables, mayor que la de la muestra original. En cambio, cuando las pérdidas corresponden a los valores más bajos de TTOA, no disminuye notablemente porque en las unidades próximas, la correlación se mantiene elevada.

4.4. Aplicación de Componentes Principales

Para las dos ondas se presentan los resultados obtenidos de la aplicación de la técnica de componentes principales en los conjuntos de datos en los que se generaron pérdidas según mecanismo MCAR y según mecanismo NO ALEATORIO en las unidades con los valores más altos de TTOA, ya que en estas situaciones los resultados presentan diferencias.

Componentes Principales con el conjunto de datos originales, casos completos e imputados por Buck, con mecanismo MCAR.

El cuadro 6 presenta los autovalores de la matriz de correlaciones sobre la cual se realizó el análisis y los porcentajes de variancia explicada, para cada conjunto de datos de la onda de octubre 1997, cuando el mecanismo de pérdida es MCAR.

Cuadro 6: Autovalores y porcentajes de variancia explicada, onda octubre 1997.

Conj. De datos	Componente Nro.	Autovalor	% de variancia explicada acumulada
DATOS ORIGINALES	1	1.51	50.3
	2	0.93	81.2
	3	0.56	100
CASOS COMPLETOS	1	1.46	48.7
	2	0.94	80.1
	3	0.60	100
IMPUTACIÓN BUCK	1	1.47	49.0
	2	0.94	80.2
	3	0.59	100

El porcentaje de variancia explicada por las dos primeras componentes es aproximadamente el 80% en todos los casos.

Cuadro 7: Coeficientes de las dos primeras componentes, onda octubre 1997.

Variable	COEFICIENTES					
	DATOS ORIGINALES		METODO DE BUCK		CASOS COMPLETOS	
	CP1	CP2	CP1	CP2	CP1	CP2
EDAD	0.37	0.92	0.36	0.93	0.35	0.93
TBE	0.67	-0.15	0.67	-0.17	0.67	-0.17
TTOA	0.64	-0.37	0.65	-0.34	0.65	-0.32

En el cuadro 7 se observa que los coeficientes de las componentes, para el conjunto con datos imputados y de casos completos, no difieren notablemente de las obtenidas con el conjunto de datos originales, por lo tanto las interpretaciones obtenidas de los dos conjuntos de datos son similares.

En la onda de 1997, la primera componente está caracterizada por las tres variables mientras que la segunda contrapone la edad con TTOA.

Para la onda de 1998, los resultados se presentan en los cuadros 8 y 9.

Cuadro 8: Autovalores y porcentajes de variancia explicada, la onda octubre 1998.

Conj. De datos	Componente Nro.	Autovalor	% de variancia explicada acumulada
DATOS ORIGINALES	1	1.22	40.6
	2	0.98	73.3
	3	0.80	100
CASOS COMPLETOS	1	1.18	39.4
	2	0.99	72.4
	3	0.83	100
IMPUTACION BUCK	1	1.19	39.6
	2	0.99	72.5
	3	0.83	100

Las dos primeras componentes para esta onda representan más del 70% de la variancia total presente en los datos, para todos los conjuntos de datos. El cuadro 9 presenta los coeficientes de las mismas.

Cuadro 9: Coeficientes de las dos primeras componentes, onda octubre 1998.

Variable	COEFICIENTES					
	DATOS ORIGINALES		METODO DE BUCK		CASOS COMPLETOS	
	CP1	CP2	CP1	CP2	CP1	CP2
EDAD	0.69	-0.05	0.69	-0.03	0.70	-0.03
TBE	0.40	0.86	0.43	0.81	0.39	0.85
TTOA	0.60	-0.51	0.58	-0.58	0.60	-0.52

En el cuadro 9 se observa que los coeficientes de las componentes principales, obtenidos del conjunto con datos imputados y de casos completos, son similares a los provenientes de los datos originales, por lo tanto se llegará a las mismas interpretaciones.

Para la onda octubre 1998, la primera componente está caracterizada por las tres variables mientras que la segunda contrapone TBE con TTOA.

En las dos ondas, cuando el mecanismo de pérdida es MCAR no se altera la interpretación de las componentes principales.

Componentes Principales con el conjunto de datos originales, casos completos e imputados por Buck, con mecanismo no aleatorio – valores máximos.

El cuadro 10 presenta los autovalores de la matriz de correlaciones sobre la cual se realizó el análisis y los porcentajes de variancia explicada, para cada conjunto de datos de la onda octubre 1997, cuando las pérdidas simuladas corresponden a las unidades con valores altos de la variable TTOA.

Cuadro 10: Autovalores y porcentajes de variancia explicada, onda octubre 1997.

Conj. De datos	Componente Nro.	Autovalor	% de variancia explicada acumulada
DATOS ORIGINALES	1	1.51	50.3
	2	0.93	81.2
	3	0.56	100
CASOS COMPLETOS	1	1.91	63.8
	2	0.94	95.0
	3	0.15	100
IMPUTACION BUCK	1	1.98	66.0
	2	0.94	97.5
	3	0.08	100

En los tres conjuntos de datos, las dos primeras componentes representan más del 80% de la variancia total presente en los datos. Sin embargo se puede observar que en los datos imputados y los casos completos este porcentaje es mayor. El cuadro 11 presenta los coeficientes de las dos primeras componentes.

Cuadro 11: Coeficientes de las dos primeras componentes, onda octubre 1997.

Variable	COEFICIENTES					
	DATOS ORIGINALES		METODO DE BUCK		CASOS COMPLETOS	
	CP1	CP2	CP1	CP2	CP1	CP2
EDAD	0.37	0.92	0.24	0.97	0.26	0.97
TBE	0.67	-0.15	0.69	-0.15	0.69	-0.15
TTOA	0.64	-0.37	0.68	-0.19	0.68	-0.21

Se observa que aplicando el método de imputación de Buck o utilizando los casos completos, casi todos los coeficientes de las componentes son similares a los obtenidos con el conjunto de datos originales.

Para la onda de octubre 1998, los resultados correspondientes a la aplicación de la técnica de componentes principales se presentan en los cuadros 12 y 13.

Cuadro 12: Autovalores y porcentajes de variancia explicada, onda octubre 1988.

Conj. De datos	Componente Nro.	Autovalor	% de variancia explicada acumulada
DATOS ORIGINALES	1	1.22	40.6
	2	0.98	73.3
	3	0.80	100
CASOS COMPLETOS	1	1.78	59.4
	2	1.0	92.7
	3	0.22	100
IMPUTACION BUCK	1	1.94	64.5
	2	1.0	97.8
	3	0.07	100

En los casos completos e imputados, las dos primeras componentes representan más del 90% de la variancia total presente en los datos. No obstante, el porcentaje correspondiente al de los datos originales es mucho menor, presentándose en el cuadro 13 los coeficientes de las mismas.

Cuadro 13: Coeficientes de las dos primeras componentes, onda octubre 1998.

Variable	COEFICIENTES					
	DATOS ORIGINALES		METODO DE BUCK		CASOS COMPLETOS	
	CP1	CP2	CP1	CP2	CP1	CP2
EDAD	0.69	0.05	-0.02	0.99	0.05	0.99
TBE	0.40	-0.86	0.71	0.02	0.71	-0.01
TTOA	0.60	0.51	0.71	0.01	0.71	-0.05

Se observa que, en este caso, los coeficientes de las componentes difieren mucho de los obtenidos con el conjunto de datos originales, ya sea en el conjunto de datos imputados o utilizando los casos completos.

En los datos con imputación y los casos completos, la primer componente está relacionada positivamente con las variables TTOA y TBE mientras que la segunda, con la EDAD únicamente. Esto significa que el asumir erróneamente que el mecanismo que produjo las pérdidas es MCAR, puede llevar a interpretaciones incorrectas.

5. DISCUSIÓN

En la onda de octubre de 1997, al generar pérdida de información y completar los datos mediante imputación o utilizar los casos completos, la interpretación de los resultados obtenidos al aplicar la técnica de componentes principales no varía, independientemente, del mecanismo de pérdida. Es decir que la estimación incrementada del coeficiente de correlación entre las variables referidas al tiempo de desocupación no afecta el análisis. Esto se debe a que este valor era alto en la muestra original.

En cambio, en la onda de octubre 1998, se observa que cuando el mecanismo conduce a perder los valores más altos de la variable TTOA, el coeficiente de correlación estimado luego de las imputaciones resulta mucho mayor que el proveniente de los datos originales. Esto distorsiona la estructura de las relaciones entre las variables afectando la interpretación de los resultados de la aplicación de la técnica de componentes principales.

En las dos ondas, cuando la estructura de los datos no se modifica luego de aplicar la técnica de tratamiento para los valores perdidos, los resultados de la aplicación de la técnica de componentes principales coinciden con los obtenidos del conjunto original de datos.

Por lo tanto, al analizar bases de datos con información incompleta debe considerarse:

- la naturaleza de las pérdidas,
- los supuestos sobre las características del mecanismo que las produjo,
- el tratamiento aplicado para “completar” los valores perdidos,

- la estructura de los datos originales y la resultante de la aplicación del tratamiento seleccionado.

BIBLIOGRAFÍA

1. Badler, C.; Alsina, S.; Puigsubirá, C.; Vitelleschi, M.; Arnesi, N. (1998) "Datos perdidos en encuestas de hogares. un aporte metodológico". Primera reunión sobre estadística pública del Instituto Interamericano de Estadística. Encuestas a Hogares: reformulación de la Encuesta Permanente de Hogares de Argentina. Página web http://www.-indec.mecon.ar/servicio/iasi_obj.htm.
2. Johnson, R. A. and D.W. Wichern. (1988). "Applied Multivariate Statistical Analysis". Prentice Hall. New Jersey.
3. Little, R. J. and D. B. Rubin. (1987). "Statistical Analysis with Missing Data". John Wiley & Sons. New York.
4. Little, R. J. (1988). "A Test of Missing Completely at Random for Multivariate Data with Missing Values". Journal of the Royal Statistical Society. Vol. 83, Nº 404.
5. Molenberghs, A; Goetghebeur, E. (1997). "A Simple Fitting Algorithm for Incomplete Categorical Data". Journal of the Royal Statistical Society". Vol. 59, Nº 2, Serie B.