Clara Badler
Sara Alsina
Daniel Wojdyla
Fernando Fracchia
Instituto de Investigaciones Teóricas y Aplicadas, Escuela de Estadística.

TRATAMIENTO DE BASES DE DATOS CON INFORMACIÓN FALTANTE EN SAS®.[1]

1. INTRODUCCIÓN

En el proceso de una investigación, un inconveniente comunmente hallado en el análisis estadístico de los datos está dado por la forma en la cual los datos han sido almacenados. Frecuentemente existen diferencias entre los formatos de almacenamiento de las bases de datos originales y los formatos de trabajo de los programas estadísticos. Por tal motivo, es habitual que se realice un procedimiento de importación de los datos en el programa estadístico.

En el proceso de importación es fundamental que se considere el problema que puede causar la presencia de valores perdidos en la base de datos original dado que los distintos programas estadísticos los interpretan de diferentes maneras. Aspectos a tener en cuenta cuando se produce este proceso incluyen el formato de la base de datos original, la forma en la cual cada programa almacena los valores perdidos y la distinción entre los diferentes tipos de variable.

Una vez que la base de datos ha sido importada, se debe conocer cómo las transformaciones aplicadas a los datos y los diferentes procedimientos estadísticos tratan a los datos perdidos.

El objetivo de este trabajo es analizar algunas situaciones que se pueden presentar al importar datos en SAS a partir de formatos de bases de datos comunmente encontrados en la práctica. La elección de este programa obedece a que dentro de los programas estadísticos, SAS es uno de los más utilizados debido a su flexibilidad para el manejo de grandes masas de datos y a la riqueza de procedimientos estadísticos con la que cuenta. Todos los comentarios se refieren a la versión 6.12 bajo Windows y suponen que los productos SAS/BASE y SAS/ACCESS han sido licenciados.

Se brindan algunas sugerencias que pueden ayudar a enfrentar los problemas ocasionados por los valores perdidos ilustrándolas con datos de la Encuesta Permanente de Hogares (EPH) perteneciente al Gran Rosario, onda Agosto 1998, la cual presenta características propias que deben tenerse en cuenta previo a su análisis.

[1] Proyecto PICT Nº 0200095-01996 de la ANPCyT.

2. TRATAMIENTO EN SAS

2.1. El Proceso de Importación de Datos.

Se examinan cuatro formatos en los cuales las bases de datos habitualmente estan almacenadas:

- Archivos DBase (extensión: .dbf)
- Archivos Microsoft Excel (extensión: .xls)
- Archivos de texto con campos delimitados por tabulaciones (normalmente con extensión .txt)
- Archivos de texto con campos delimitados por blancos (normalmente con extensión .txt)

Existen otros formatos encontrados en la práctica pero en la mayoría de los casos el tratamiento es similar a alguno de los formatos presentados.

Dos procedimientos diferentes se pueden seguir para importar una base de datos en SAS. El primero consiste en la utilización de un asistente ("wizard") de importación el cual guía al usuario paso a paso en el proceso de importación. Este método acepta como formatos de los datos originales a cualquiera de los presentados anteriormente. El segundo método se basa en la escritura de un programa SAS y permite leer archivos en cualquiera de los formatos de texto sin la necesidad de contar con el producto SAS/ACCESS.

Además, las bases con formato Microsoft Excel pueden ser importadas utilizando DDE (intercambio dinámico de información entre aplicaciones Windows) sin la necesidad de contar con el producto SAS/ACCESS.

Es importante destacar que el producto SAS/ACCESS y en particular el procedimiento "access" permite la utilización de otros métodos de importación de datos incluyendo la creación de enlaces dinámicos a través de ODBC con las bases de datos originales. ODBC es una interface de programación que permite que las aplicaciones accedan a datos almacenados en sistemas de manejo de bases de datos, que utilizan el lenguaje de consulta estructurada (SQL) como un estándar para el acceso de datos. Formatos compatibles con este tipo de datos incluye: Microsoft Access, Microsoft Excel, FoxPro y DBase entre otros.

2.2. Valores perdidos.

El primer aspecto a considerar en el proceso de importación al analizar bases de datos con valores perdidos, es el tipo de variable. Básicamente existen dos tipos de variables: numérica y caracter. Si bien las variables de tipo fecha pueden considerarse como un tercer tipo en muchos programas de base de datos, en SAS las variables de tipo fecha son almacenadas internamente como numéricas y por lo tanto las convenciones válidas para variables numéricas son aplicables también a ellas.

Los blancos o puntos en las variables de tipo caracter son representados en SAS con blancos. Especial atención se debe prestar al proceso de importación de un archivo de texto con campos delimitados por blancos para evitar problemas con la presencia de valores perdidos.

Para variables de tipo numérico con blancos o puntos, SAS utiliza el punto decimal como indicador de un valor perdido. Además, es posible especificar a través de la sentencia "missing" ciertos valores (alfabéticos) a ser considerados como valores perdidos cuando se leen los datos.

Internamente, SAS almacena los valores perdidos como valores negativos muy grandes. Especificamente, un valor numérico perdido en SAS es menor que cualquier valor no perdido. Esto debe ser considerado cuando se realizan comparaciones entre variables con el objeto de recodificar variables. La falla en la consideración de este detalle puede provocar que valores perdidos se transformen en valores no perdidos en variables derivadas o transformadas. Al realizar un ordenamiento de las observaciones (a través del PROC SORT) se debe tener en cuenta que si la variable criterio para el ordenamiento incluye valores perdidos en algunas observaciones las mismas serán ubicadas al comienzo de la base ordenada.

Además, dentro de los valores perdidos es posible establecer un orden que va desde ._ (el menor de todos los valores perdidos), continua con . , luego .A hasta .Z (el mayor de todos los valores perdidos).

2.3. Transformaciones y Procedimientos Estadísticos.

Un aspecto de importancia es el que se denomina propagación de los valores perdidos. En SAS, una operación entre variables que contienen valores perdidos genera datos perdidos. Más especificamente si se realiza la suma de variables y una de ellas tiene un valor perdido para una observación determinada el resultado será un valor perdido. Es decir,

```
T = A + B + C + D;
```

generará un valor perdido para una observación si alguna de las variables tiene un valor perdido para dicha observación. Una forma de solucionar este problema es emplear la función "sum". Por ejemplo, la sentencia anterior se puede escribir alternativamente como

```
sum(A, B, C, D);
```

Esta sentencia solamente producirá un valor perdido si todas las variables involucradas tienen valores perdidos. En cualquier otra situación considerará al valor perdido como 0, lo cual en algunos casos tiene sentido como puede ser en las casos de valores perdidos producidos por diseño.

Un criterio conveniente que permite evaluar la calidad de la información y la dimensión del problema de los valores perdidos consiste en la realización de los siguientes pasos inmediatamente después del proceso de importación.

- Ejecutar un PROC CONTENTS y verificar que los formatos de las variables en el archivo SAS coinciden con el formato de las variables originales.
- Ejecutar un PROC FREQ para las variables categóricas y un PROC UNIVARIATE para las variables continuas de forma tal que sea posible identificar categorías y valores que han sido asignados como datos perdidos.
- Corregir los formatos de las variables y asignar los valores perdidos en un paso DATA.

Una vez depurados los datos es necesario considerar cómo los diferentes procedimientos estadísticos tratan a los valores perdidos. No existe un procedimiento específico dentro de SAS para el tratamiento de valores perdidos. En la mayoría de ellos la presencia de un valor perdido en una variable provoca que la observación sea excluída del análisis. Es decir, solamente los casos completos son considerados en el análisis. Existen otros procedimientos que debido a sus características consideran los casos disponibles en el análisis (por ejemplo PROC CORR).

Ciertos procedimientos tales como PRINQUAL, STANDARD y TRANSREG, posibilitan la transformación de variables y proveen técnicas que pueden ser utilizadas como métodos simples de imputación para datos perdidos. Sin embargo, se debe tener mucha precaución al utilizarlos con estos fines dado que estos procedimientos no fueron diseñados específi-

camente para tales tarea y los autores reconocen que se pueden presentar inconsistencias en algunos casos.

Por último, todos los procedimientos cuentan con ayuda online, la cual describe la forma en que los valores perdidos son tratados.

3. APLICACIÓN A LA EPH.

Se utilizó la base ampliada del Gran Rosario, Onda Agosto 1998 de la Encuesta Permanente de Hogares. Esta base contiene 151 variables y 3567 registros.

Con fines de ilustración se creó una base con variables seleccionadas y la misma cantidad de registros, de forma tal de disponer de casos para los diferentes tipos de variables y para las distintas formas en que se presenta la categoría para datos faltantes. La base de trabajo incluyó las siguientes variables:

Nombre	Tipo	Longitud	Comentarios
Fecha de Nacimiento (H11)	Fecha	8	Contiene 6 valores perdidos.
Condición de Actividad (ESTADO)	Numérica	1	No tiene valores perdidos.
Cuanto gana en esa ocupación (P21)	Numérica	10	No contiene valores perdidos.
Combinación de 2 variables que representan la tenencia o no de ingreso individual (CODINGRE)	Numérica	1	No contiene celdas vacías pero tiene una categoría asignada a los valores perdidos.
Monto ingresos totales (P47T)	Numérica	10	No contiene celdas vacías pero con fines didácticos se reemplazan los ceros con blancos en la base original.
Carrera que cursa o cursó (P57)	Caracter	60	Tiene celdas vacías
Monto de Ingreso Total Familiar (ITF)	Numérica	12.2	No contiene celdas vacías pero contiene ceros que no son todos valores perdidos.
Monto de Ingreso Per Cápita Familiar (IPCF)	Numérica	12.2	No contiene celdas vacías pero con fines didácticos se reemplazan los ceros con blancos.
Decil relativo al Ingreso Total Individual (DECIND)	Caracter	2	Variable de tipo caracter pero con contenido exclusivamente numérico. No contiene valores perdidos.

Decil relativo al Ingreso de la Ocupación Principal (DECOCU)	Caracter	2	Variable de tipo caracter con contenido exclusivamente numérico. Contiene valores perdidos.
Rama de Actividad (RAMA)	Numérica	2	Contiene celdas vacías y tiene categorías asignadas a clases especiales que pueden ser tomadas como valores perdidos.

Todas las transformaciones realizadas para crear la base reducida se realizaron en FoxPro. Los tipos de archivos con los que se trabajó y que fueron creados a partir de FoxPro fueron: Dbase (formato nativo), texto delimitado por blanco y texto delimitado por caracteres. Dado que FoxPro no graba archivos en formato Microsoft Excel, el archivo de la base reducida en formato Dbase fue abierto en Excel y grabado en dicho formato. Se comprobó que el proceso de importación de la base desde Dbase a Excel no produce ninguna alteración o distorsión a los datos originales.

Las distintas bases fueron importadas en SAS utilizando el asistente de importación. Los resultados obtenidos para cada uno de los 4 formatos fueron los siguientes:

- Dbase: se observó que la estructura de la base es exactamente igual a la original. La única diferencia es que SAS asigna un punto (.) para los valores perdidos en las variables numéricas. Tanto los valores perdidos originales como los creados en ciertas variables numéricas son reconocidos.
- Microsoft Excel: La estructura de la base sufre un cambio en algunas extensiones de los campos, aunque no se detecta un criterio específico para los cambios. Los valores no sufren ninguna modificación, ya sean blancos o no.
- Texto Delimitado por Tabulaciones: Un primer detalle a considerar es que SAS permite leer archivos que incluyan o no los nombres de los campos en la primer fila del archivo de texto. Si se los incluye estos deben estar separados por tabulaciones y deben respetar las convenciones para los nombres de variables en SAS. Si no se incluye en la primer fila los nombres de las variables, SAS asigna los nombres VAR1, VAR2, VAR3, etc. En cuanto a la variable de tipo fecha, se pueden presentar problemas cuando las fechas son almacenadas en formato numérico sin ningún tipo de separador. Una forma de solucionar tal problema es convertir a un formato caracter a la variable original y luego aplicarle el formato de fecha correspondiente. Por otro lado no se presentan problemas para reconocer los datos faltantes.
- Texto Delimitado por Blancos: Con características similares a las del caso anterior, con la excepción que cuando se tienen variables de tipo caracter que contienen un espacio (blanco) intermedio, SAS asigna como valor solamente a la primer palabra, tomando la segunda palabra como valor de la variable siguiente, con la consecuencia lógica de errores en la lectura de los datos. Se puede corregir este problema eliminando los blancos en la variable caracter o reemplanzándolos por caracteres no blancos. Otro problema surge cuando se tienen variables de tipo caracter con contenido exclusivamente numérico. En tal caso la variable será leída con formato numérico y será necesario cambiarle el formato en un paso DATA.

Adicionalmente las bases con formato texto fueron importadas con un programa SAS que incluye la sentencia INFILE, la cual permite abrir el archivo y especificar el tipo de delimitador utilizado. El proceso anterior se puede ejecutar con las siguientes sentencias:

```
data Name;
     infile 'Path' delimiter='09'x missover dsd;
run;
```

La opción DELIMITER permite especificar el delimitador utilizado. En este caso '09'x representa delimitación por tabulaciones. Otras alternativas son ' ' y ' , ' que corresponden a delimitados por blancos y comas respectivamente.

La opción MISSOVER sirve para que SAS considere valores perdidos en la última variable evitando que lea un valor de la línea siguiente. La opción DSD sirve para que SAS considere a dos delimitadores consecutivos como un valor perdido.

Conviene aclarar que posterior a la sentencia INFILE se puede incluir una sentencia INFORMAT para que SAS reconozca a las variables con el formato correspondiente. Esto es de especial importancia en el caso de variables de tipo fecha y las variables de tipo caracter con contenido exclusivamente numérico que no se encuentran diferenciadas en la base importada.

En el caso de la base delimitada por blancos se debe tener la misma precaución que al importarlos a través del asistente especialmente con los blancos intermedios.

Otra alternativa para importar datos en formato Excel consiste en la utilización de DDE. Para la aplicación de este método el archivo a importar debe estar abierto en Excel y se puede llevar a cabo a través de las siguientes sentencias SAS:

```
filename datos DDE 'Excel|c:\base.xls!f1c1:f3567c11';
data ejemplo;
     infile datos dlm='09'x notab dsd missover;
run;
```

Donde "base.xls" es la planilla Excel que contiene la base y "f1c1:f3567c11" hace referencia al rango de valores (identificado por filas y columnas) a importar y la opción NOTAB es necesaria para que SAS no convierta los tabulados, que existen originalmente en Excel, en blancos. El DLM es una abreviatura de la opción DELIMITER. Otra opción consiste en copiar el contenido de Excel en el “Clipboard” y utilizar la siguiente sentencia, la cual es equivalente a la anterior sentencia FILENAME.

```
filename datos DDE 'clipboard';
```

Posterior al proceso de importación, para evaluar la calidad de los datos importados se aplican los procedimientos CONTENTS y FREQ (o UNIVARIATE).

La siguiente tabla muestra el resultado del PROC CONTENTS para los cuatro formatos importados a través del asistente.

Formato Original: DBase

#	Variable	Type	Len	Pos	Format	Informat
1	H11	Num	8	0	MMDDYY8.	YYMMDD8.
2	ESTADO	Num	8	8	1.	1.
3	P21	Num	8	16	10.	10.
4	CODINGRE	Num	8	24	1.	1.
5	P47T	Num	8	32	10.	10.
6	P57	Char	60	40	$60.	$60.
7	ITF	Num	8	100	12.2	12.2
8	IPCF	Num	8	108	12.2	12.2
9	DECIND	Char	2	116	$2.	$2.
10	DECOCU	Char	2	118	$2.	$2.
11	RAMA	Num	8	120	2.	2.

Formato Original: Microsoft Excel

#	Variable	Type	Len	Pos	Format	Informat
1	H11	Num	8	0	MMDDYY8.	MMDDYY8.
2	ESTADO	Num	8	8	1.	1.
3	P21	Num	8	16	10.	10.
4	CODINGRE	Num	8	24	1.	1.
5	P47T	Num	8	32	10.	10.
6	P57	Char	60	40	$60.	$60.
7	ITF	Num	8	100	12.2	12.2
8	IPCF	Num	8	108	10.2	10.2
9	DECIND	Char	2	116	$2.	$2.
10	DECOCU	Char	10	118	$10.	$10.
11	RAMA	Num	8	128	10.	10.

Formato Original: Texto Delimitados con Tabulaciones

#	Variable	Type	Len	Pos	Format	Informat
1	VAR1	Num	8	0	BEST12.	BEST32.
2	VAR2	Num	8	8	BEST12.	BEST32.
3	VAR3	Num	8	16	BEST12.	BEST32.
4	VAR4	Num	8	24	BEST12.	BEST32.
5	VAR5	Num	8	32	BEST12.	BEST32.
6	VAR6	Char	32	40	$32.	$32.
7	VAR7	Num	8	72	BEST12.	BEST32.
8	VAR8	Num	8	80	BEST12.	BEST32.
9	VAR9	Char	4	88	$4.	$4.
10	VAR10	Char	4	92	$4.	$4.
11	VAR11	Num	8	96	BEST12.	BEST32.

Formato Original: Texto Delimitados con Blancos

#	Variable	Type	Len	Pos	Format	Informat
1	FECHA	Num	8	0	BEST12.	BEST32.
2	ESTADO	Num	8	8	BEST12.	BEST32.
3	P21	Num	8	16	BEST12.	BEST32.
4	CODINGRE	Num	8	24	BEST12.	BEST32.
5	P47T	Num	8	32	BEST12.	BEST32.
6	P57	Char	24	40	$24.	$24.
7	ITF	Num	8	64	BEST12.	BEST32.
8	IPCF	Num	8	72	BEST12.	BEST32.
9	DECIND	Num	8	80	BEST12.	BEST32.
10	DECOCU	Num	8	88	BEST12.	BEST32.
11	RAMA	Num	8	96	BEST12.	BEST32.

Como se observa en las tablas anteriores, el proceso de importación a través del asistente no presenta inconvenientes al importar archivos Dbase reconociendo todos los formatos de la base original.

Los resultados son similares cuando la base tiene formato Microsoft Excel con pequeñas diferencias en la longitud de ciertos campos (IPCF y DECOCU).

En el caso de los archivos con formato de texto es donde se presentaron los mayores problemas. En caso de no incluir una primera fila con los nombres de las variables, SAS asigna los nombres VAR1,..., VAR11. Todos los formatos numéricos, incluida la fecha, son leídos con el formato BEST32 el cual es el que utiliza SAS por defecto. Este formato es el que proporciona mayor información acerca del valor, de acuerdo a la información disponible acerca del ancho del campo del que proviene. Si bien los valores numéricos no generarían problemas los valores con formato fecha no son leídos correctamente. Una forma de solucionar este problema es a través de manipulaciones en un paso DATA de los valores leídos. Las siguientes sentencias permiten transformar una variable numérica en una variable fecha:

```
Dia = input(substr(put(fecha, $8.),1,2),2.);
Mes = input(substr(put(fecha, $8),3,2),2.);
Año = input(substr(put(fecha, $8.)5,2),2.);
NvaFecha = mdy(Mes,Dia,Año);
```

Las primeras tres sentencias extraen los valores correspondientes a día, mes y año de la fecha original. El proceso consiste en convertir a la variable numérica en variable caracter (put), extraer el valor deseado (substr) y convertirlo nuevamente en variable numérica (input). La cuarta sentencia transforma valores de día, mes y año en formato numérico a valores de fecha.

Otra alternativa que se presenta frecuentemente en la práctica está dada por la lectura de valores con formato fecha que estan almacenados con formato caracter y que incluyen los símbolos de división, tales como '/' o '-'. Se puede leer un formato fecha a partir de una variable con formato caracter con la siguiente sentencia, en un paso DATA:

```
NvaFecha = input(fecha, ddmmyy8.);
```

O

```
NvaFecha = input(fecha, ddmmyy10.);
```

Otro inconveniente con los archivos con formato de texto se presenta en la longitud de las variables caracter. En ninguno de los dos casos el asistente de importación reconoció la longitud correcta de la variable P57 (60 caracteres). Además en el formato de texto delimitado con blancos interpretó las variables caracter con contenido exclusivamente numérico como variables numéricas. Se puede solucionar este problema mediante la siguiente sentencia:

```
DECOCUN = input(DECOCU, $2.);
```

En cuanto a la interpretación de los valores faltantes no se presentaron muchos problemas. Dentro de las variables numéricas se pueden presentar dos casos: que exista un código para valor perdido (por ejemplo 9 o 99) o que se le haya asignado un blanco o punto en

el archivo original. En el primer caso habrá que utilizar una sentencia para recodificar el código asignado como valor perdido, como por ejemplo:

```
If codingre = 9 then codingre = .;
```

En el segundo caso el proceso de importación no genera problemas. En las variables caracter el blanco es considerado como valor faltante y su presencia es reconocida en cualquiera de los cuatro formatos, siempre que se tomen en cuenta los comentarios presentados anteriormente.

Otro punto a considerar es cuando se recodifican variables numéricas que contienen valores perdidos. Por ejemplo, si la variable EDAD contiene valores perdidos y se ejecuta la siguiente sentencia para categorizar la variable EDAD:

```
If edad < 30 then edadcat = 1;
```

Los valores perdidos de la variable EDAD serán incluidos también en dicha categoría dado que los valores perdidos toman un valor negativo muy pequeño. La forma correcta para dicha codificación es:

```
If 0 < edad < 30 then edadcat = 1;
```

Como se puede observar, ciertos problemas se presentan cuando los datos estan almacenados en formato de texto, pero los mismos pueden solucionarse con sentencias en un paso DATA posterior a la lectura de los datos.

4. CONCLUSIÓN.

Al utilizar información proveniente de bases de datos se debe tener en cuenta que el programa utilizado considere las unidades que en la base original presentan falta de información como tales, ya que su significado pudo ser alterado en alguna etapa del procesamiento, dificultando su identificación y en consecuencia pudiendo afectar el análisis posterior. SAS provee de recursos para que al usuario no se le presenten mayores inconvenientes en la lectura de bases de datos y para solucionar problemas en el tratamiento de bases de datos con información faltante.

Es de destacar que estas conclusiones se basan en la experiencia en la importación de bases de la Encuesta Permanente de Hogares. Diferentes problemas se pueden presentar con bases con formatos más complejos.

5. BIBLIOGRAFÍA.

Aster, R. and Seidman, R. “Professional SAS Programming Secrets.” McGraw-Hill. 1997.

Cody, R. and Pass, R. "SAS® Programming by Example." Cary, NC: SAS Institute Inc., 1995.

SAS Institute Inc., SAS® Language: Reference, Version 6, First Edition, Cary, NC: SAS Institute Inc., 1990. 1042 pp.

SAS Institute Inc., SAS® Procedures Guide: Version 6, Third Edition, Cary, NC: SAS Institute Inc., 1990. 705 pp.