Depuración automática de datos estadísticos

1. Antecedentes

Con este artículo se pretende poner de manifiesto la forma de ayudar a resolver algunos problemas de tipo práctico, que surgen al implementar algoritmos basados en la teoría desarrollada por FELLEGI&HOLT [1] para optimizar la depuración automática de datos estadísticos. Por ello, nos limitaremos a la exposición de dichos problemas y a ilustrar con determinados ejemplos la forma propuesta para solucionarlos, respetando en todo momento la terminología y conceptos establecidos por FELLEGI&HOLT, y añadiendo algún nuevo concepto cuando sea indispensable para clarificar la forma propuesta para resolver el problema considerado.

Con cierta frecuencia la hipótesis establecida por FELLEGI&HOLT sobre la disponibilidad del Conjunto Completo de Edits (CCE), que es totalmente aceptable desde un punto vista teórico (según prueban los teoremas establecidos al efecto por FELLEGI&HOLT), no es aceptable desde una óptica práctica u operativa, puesto que se presentan casos reales de imputación caracterizados por concurrir en ellos las siguientes circunstancias:

- los expertos en el tema considerado establecen un Conjunto de Edits Explícitos (CEE) que incluye un gran número de edits.
- el nº de variables participantes (activas) en la definición de los edits del CEE es también muy amplio.
- las variables están muy interrelacionadas.

En tales casos suele producirse un crecimiento exponencial en el número de posibles combinaciones de edits, cuya capacidad para generar un Edit Esencialmente Nuevo (EEN) debe ser comprobada.

Para evitar en la medida de lo posible el precipitado crecimiento, es necesario desarrollar complejos sistemas de filtro que disminuyan drásticamente el número de combinaciones de edits a comprobar, de tal manera que sea posible generar el CCE en tiempo útil.

De la imposibilidad práctica de generar el CCE en tiempo útil, observada en la práctica en situaciones de gran complejidad, se derivan dos circunstancias indeseables:

- los expertos descomponen el CEE en varios subconjuntos parciales y a partir de cada uno de ellos tratan de generar su correspondiente Conjunto de Edits Implícitos (CEI), con lo cual se dispone de varios CCE, en lugar de solamente uno como se presupone teóricamente. La descomposición del CEE no es, ni mucho menos, una tarea baladí dado que no existe ningún sistema que la realice de forma automática, los expertos han de realizar engorrosas pruebas sucesivas de descomposición, tratando de minimizar el número de CCEs.
- el proceso de imputación, al operar con varios CCEs, se realiza en varias pasadas sucesivas. En cada pasada se declaran como no imputables (fijas) una o varias variables que fueron susceptibles de imputación en pasadas anteriores. La fijeza de variables lesiona uno de los principios esenciales del sistema propuesto por FELLEGI&HOLT: el principio del "cambio mínimo = respeto máximo", consistente en que se debe imputar el menor número posible de variables, respetando así al máximo posible los datos originales. La vulneración del principio del "cambio mínimo" puede producirse si ocurre, por ejemplo, lo siguiente:

- el registro a imputar, R°, falla dos edits [e(m) y e(n)].
- El CEE se descompone en dos subconjuntos, CEE (1) y CEE (2), a partir de los cuales se obtienen los correspondientes CCE (1) y CCE (2).
- e(m) $\in$ CCE (1) y e(n) $\in$ CCE (2).
- La variable X está activa en ambos edits.
- En la primera pasada, para evitar el fallo de e(m), el sistema decide la imputación de la variable X, lo que hace que, en la segunda pasada, la variable X se considere fija, lo que implica su exclusión del Conjunto Mínimo (X$\notin$CM).
- En la segunda pasada, para evitar el fallo de e(n), el sistema decide la imputación de una variable (Z) distinta de X, por ser ésta ahora fija.
- Si hubiera habido solamente un CCE, en lugar de dos, quizá podrían evitarse ambos errores imputando un valor adecuado solamente a la variable X.

Una forma natural de resolver el problema de imputación, respetando escrupulosamente los principios esenciales del sistema propuesto por FFLLEGI&HOLT, es operar, en lugar de con el CCE, con el CCE -R° (Conjunto Completo de Edits a nivel de registro) consistente en lo siguiente:

Cuando un registro determinado (R°) ha de ser imputado por haber fallado algún edit, se determina si el CEE es suficiente para permitir su imputación respetando el principio del cambio mínimo, en cuyo caso la imputación exitosa de R° no exige la derivación de ningún edit implicado por los edits del CEE (este caso es el más frecuente en la práctica). En caso contrario, es necesario generar dinámicamente el CEI requerido por R° (CEI-R°), con lo cual se obtiene el correspondiente CCE -R°=CEE+CEI-R°, lo que permite imputar R° de forma análoga a como se hubiera hecho haciendo uso del CCE.

En el APARTADO 2 se presenta la forma de generar, en caso necesario, el CEI-R° correspondiente a un CEE de tipo lógico (edits cualitativos aplicables al caso de variables cualitativas), y se comparan los resultados obtenidos al operar con el CCE y con el CCE -R°.

El APARTADO 3 se dedica a ilustrar la forma de, respetando los principios esenciales del sistema FELLEGI&HOLT, abordar la imputación cuando los errores observados no son aleatorios sino sistemáticos, en cuyo caso se hace uso de un nuevo concepto: Edit de Imputación Determinística (EID).

El APARTADO 4 se dedica a la problemática planteada por los edits aritméticos de tipo lineal (aplicables al caso de variables cuantitativas). En su artículo del año 1976, FELLEGI&HOLT anuncian que en un próximo artículo completarían algunos aspectos no considerados en ese momento. La publicación de dicho artículo es desconocida por el autor de estas notas, lo que no implica que no haya sido publicado.

Por último, el APARTADO 5 se dedica a los EDITS MIXTOS.

Por razones de brevedad, en lo que sigue, llamaremos edit derivado a todo edit que sea generado en base a varios edits preexistente, los cuales pueden ser:

- Todos explícitos

- Todos implícitos
- Una mezcla de explícitos e implícitos

2. Generación del CEI-R°

En el cuadro siguiente aparecen los datos que, después de convertir en tiras binarias los correspondientes edits explícitos, usaremos para ilustrar la generación del CEI-R°. Los valores observados (expresados por el correspondiente bit 1) en el registro actual (R°) determinan el fallo de dos de los 4 edits componentes del CEE. También se detalla las variables activas (en ellas aparece un bit 0 en alguno de sus valores válidos) en cada uno de los edits.

EDIT	VARIABLES														EDIT FALLADO (F)	VARIABLES ACTIVAS
	A			B				C				D				
	1	2	3	1	2	3	4	1	2	3	4	1	2	3		
E1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	F	A B
E2	0	1	1	1	1	1	1	0	1	1	0	1	1	1	-	A C
E3	1	0	1	1	1	1	1	0	0	1	1	0	1	0	-	A C D
E4	1	1	1	1	1	1	0	1	1	1	1	**0**	0	1	F	B D
R°		***1***				***1***					***1***			***1***	-------	------

2.1 Determinación del conjunto mínimo (CM) de variables a imputar

La determinación del CM se basa en el conocimiento de los edits Fallados (F).

El registro actual (R°) falla un edit si a cada bit 1 de R° corresponde en el edit considerado un bit 1. Así pues, se tiene que los edits fallados por R° son el 1 y el 4, F=(1 4).

La determinación del CM es un proceso iterativo que se repite(r) hasta conseguir que las variables incluidas en el CM cubran todos los edits de F.

Una variable candidata a ser incluida en el CM es aquella que cumple las siguientes condiciones (AND):

- no está todavía incluida en el CM
- está activa en algún edit perteneciente al FPC(r) actual.

Al empezar la reiteración 1, el conjunto de edits Fallados Pendientes de Cubrir (FPC) coincide con F. Es decir, FPC(r=1)= (1 4).

Las variables activas en cada uno de los edits fallados, y por tanto candidatas para cubrirlos, son las siguientes:

- E1 $\rightarrow$ A B
- E4 $\rightarrow$ B D

La variable A está activa en 1 de los 2 edits fallados (es decir, cubre 1 de los 2 edits fallados).

La variable B está activa en los 2 edits fallados (es decir, cubre 2 de los 2 edits fallados).

La variable D está activa en 1 de los 2 edits fallados.

La selección entre las variables candidatas se decide mediante un Índice de Sospecha (IS) cuyo valor se determina, fundamentalmente, en base al nº de edits fallados en los que está activa la variable considerada. Los datos correspondientes a las variables sospechosas (las que están activas en algún edit fallado) son los siguientes:

- variable A: F/A =1/3, la variable A falla 1 de los 3 edits en los que está activa.
- variable B: F/A =2/2, la variable B falla 2 de los 2 edits en los que está activa.
- variable D: F/A =1/2, la variable D falla 1 de los 2 edits en los que está activa.
- Por otra parte, se tiene:
- REITERACIÓN 1: el sistema selecciona la variable B por ser la más sospechosa.
- la B CUBRE los edits 1 y 4 $\rightarrow$ se eliminan, con lo cual se actualiza el conjunto FPC.
- al empezar la reiteración 2 se comprueba que FPC es un conjunto vacío.

En consecuencia, se establece el CM = (B).

2.2 Eliminación de edits que no pueden fallar

En el cuadro siguiente se clasifican los edits según el conjunto al cual pertenecen:

CLASIFICACIÓN DE LOS EDITS SEGÚN FALLO Y PERTENENCIA DE SUS VARIABLES ACTIVAS AL CM

R° FALLA EL EDIT (S/N)	ALGUNA DE LAS VARIABLES DEL EDIT PERTENECE (S/N) AL CM		
	S	N	TOTAL
S	F	∅	F
N	NFPF	NFNPF	NF
TOTAL	PF = F + NFPF	NPF	TE

Siendo:

- TE= TOTAL EDITS (los que haya al comenzar la determinación del CM)
- F= edits Fallados por el registro original (R°)
- NF= edits No Fallados por el registro original (R°)
- NFPF= edits No Fallados que Pueden Fallar, dependiendo de la imputación realizada (R*).
- NFNPF= edits No Fallados que No Pueden Fallar, cualquiera que sea la imputación realizada.

- PF= edits que Pueden Fallar = F + NFPF.
- NPF= edits que No Pueden Fallar, cualquiera que sea la imputación realizada.

Para facilitar el proceso de imputación del R° a depurar, se consideran, además de los conjuntos de edits precitados, los siguientes:

- AV(X) = edits en los cuales está Activa la Variable X.
- EFE(X) = Edits cuyo Fallo debe ser Evitado mediante la imputación de un valor adecuado a la variable X.

Un edit e pertenece al conjunto NPF si cumple alguna de las siguientes condiciones (OR):

- Ninguna de las variables pertenecientes al CM está activa en e.
- Alguna variable no perteneciente al CM (en cuyo caso no será imputada) tiene un valor que impide el fallo del edit e.

Por otra parte, el edit e pertenece al conjunto EFE (X) si cumple las siguientes condiciones (AND):

- $e \in$ PF.
- $e \in$ AV(X).
- $e \in$ NAV(α) $\rightarrow$ es decir, el edit e pertenece al conjunto de Edits NAV(α) si ninguna de las variables activas en e será imputada después de la variable X .Dicho de otra forma equivalente , todas las variables activas en el edit e tienen actualmente un valor que no variará , que será :
- el valor fijo original registrado en Rº , si tales variables no se incluyeron en el CM.
- el valor fijado mediante imputación previa a la actual imputación de la variable X, si tales variables se incluyeron en el CM.

Dado que los edits pertenecientes al conjunto NAV(α) se caracterizan por no ser activos en ninguna variable cuya imputación será posterior a la imputación de la variable X, es evidente que si, después de imputar X, se falla algún edit perteneciente al conjunto EFE (X), las variables a imputar posteriormente no evitaran su fallo, lo que implica que el registro imputado (R*) seguiría fallando algún edit del EFE (X),

Así pues, en base a los datos de nuestro ejemplo, se tiene lo siguiente:

- TE = (1 2 3 4).
- NPF = (2 3) $\rightarrow$ Los valores observados en variables no pertenecientes al CM

(C° = 4 y D° =3) nos permiten eliminar previamente los edits 2 y 3.

- PF = (1 4).

2.3 Imputación de la variable B

Un edit e pertenece al conjunto EFE(B) si cumple las siguientes condiciones (AND):

- $e \in PF$ = (1 4)
- $e \in AV(B)$ = (1 4)
- $e \in NAV(\alpha) \rightarrow$ cuando α es un conjunto vacío (como ocurre ahora puesto que B es la última variable a imputar), se ignora esta condición.

El resultado de la intersección de conjuntos es EFE(B) = (1 4).

Cuando el conjunto EFE(B) tiene varios edits (como ocurre en nuestro caso), el conjunto de valores admisibles como imputación de la variable B viene dado por el resultado obtenido al efectuar con los edits pertenecientes al EFE(B) la operación de UNION en las columnas de bits correspondientes a la variable B. Así pues, se tiene:

- UNION/B (1 4) = (1 1 1 1)

Cuando ocurre, como ahora, que el vector UNION resultante no tiene ningún bit 0, decimos que B es una variable vacía porque no existe ningún valor de B que evite el fallo de todos los edits del conjunto EFE(B). La variable B es vacía porque el conjunto actual de edits no es completo respecto al registro en curso de imputación. Por tanto, en ese momento se ejecuta una llamada al procedimiento generador, para que genere un nuevo edit en base a, precisa y exclusivamente, lo siguiente:

- edits contribuyente, los del conjunto EFE(B).
- variable generadora, la B.

El procedimiento generador utilizado es el especificado por FELLEGI&HOLT, consistente en lo siguiente:

1. intersección ($\cap$) de los edits EFE(B), en todas las variables excepto en la vacía (B).
2. unión ($\cup$) de los edits EFE(B), en la variable B.

El resultado obtenido es el siguiente:

Nº de EDIT	Campo Generador y edits contribuyentes	VARIABLES														FALLADO (F)	VARIABLES ACTIVAS
		A			B				C				D				
		1	2	3	1	2	3	4	1	2	3	4	1	2	3		
1	-	1	1	0	0	0	1	1	1	1	1	1	1	1	1	F	A B
4	-	1	1	1	1	1	1	0	1	1	1	1	**0**	0	1	F	B D
5	B (1 4)	**1**	**1**	**0**	1	1	1	1	1	1	1	1	0	0	1	F	**A D**
R°			***1***				***1***					***1***			***1***	----	------

El resultado es la generación de un Edit Esencialmente Nuevo (EEN) que, si se hubiese generado el CCE (Conjunto Completo de Edits), habría formado parte del mismo, según puede verse al final de este APARTADO. Así pues, dicho edit será llamado indistintamente:

- edit nº 5.
- edit B (1 4), para indicar que fue generado, en base al campo generador B, por los edits contribuyentes 1 y 4.

Antes de actualizar el conjunto de edits disponibles, el procedimiento generador comprueba que la nueva tira binaria es realmente un edit. Para ello ha de ocurrir (AND):

1. la desactivación de la variable vacía, es decir, el resultado de la operación UNION debe ser un vector unitario (U(X)/EFE(X)= "11111 ... 111")
2. ha de haber al menos 2 variables activas (su tira binaria debe ser una mezcla de 0/1)
3. no debe haber ninguna variable con una tira binaria nula (Vector Nulo → 000000...0000).

Al ser vacía la variable B, la generación del nuevo edit implica obtener una versión actualizada del CM, en base al conjunto actual de edits. Por tanto, se repiten, si procede, los pasos anteriores, es decir:

Determinación del nuevo CM

Al añadir un nuevo edit puede variar el conjunto de edits fallados (F), supuesto que falle el nuevo edit

En todo caso, lo que siempre varía es el conjunto de edits PF. En consecuencia, es necesario determinar cual es el nuevo CM y comprobar si son vacías sus variables, utilizando para ello el nuevo PF.

Los datos correspondientes a las variables sospechosas (las que están activas en algún edit fallado) son los siguientes:

variable A: F/A =2/4, la variable A falla 2 de los 4 edits en los que está activa.

variable B: F/A =2/2, la variable B falla 2 de los 2 edits en los que está activa.

variable D: F/A =2/3, la variable D falla 2 de los 3 edits en los que está activa.

Por otra parte, se tiene:

REITERACIÓN 1: el sistema selecciona la variable B por ser la más sospechosa.

La B CUBRE los edits 1 y 4 → se eliminan, con lo cual se actualiza el conjunto FPC.

Al empezar la reiteración 2 se tiene FPC=(5). Entre las variables activas en el edit 5, el sistema selecciona la más sospechosa (D).

Al empezar la reiteración 3 se comprueba que FPC es un conjunto vacío.

En consecuencia, se establece el CM = (B D).

Eliminación de edits que no pueden fallar

Actualmente se tiene los siguientes conjuntos de edits:

- TE =(1 2 3 4 5)
- NPF = (2 3) $\rightarrow$ Los valores (A° = 2 y C° =4), observados en variables no pertenecientes al CM, nos permiten eliminar previamente los edits 2 y 3.
- PF = (1 4 5)

La variable D es la única que ha de ser imputada después de la variable B. Así pues, se tiene:

NAV(α) = NAV(D) = (2 3 5).

Imputación de la primera variable del CM (B)

Un edit e pertenece al conjunto EFE(B) si cumple las siguientes condiciones (AND):

- e $\in$ PF = (1 4 5)
- e $\in$ AV(B) = (1 4)
- e $\in$ NAV(D) = (1 2)

El resultado de la intersección de conjuntos es EFE(B) = (1).

En general, cuando el conjunto EFE(B) tiene varios edits, el conjunto de valores admisibles como imputación de la variable B viene dado por el resultado obtenido al efectuar con los edits pertenecientes al EFE(B) la operación de UNION en las columnas de bits correspondientes a la variable B. En nuestro caso, al haber solamente un edit en EFE(B), no tiene sentido dicha operación, ya que el conjunto de valores admisibles como imputación viene dado por los bits 0 que haya en el edit 1.

La tira binaria correspondiente a la variable B en el edit 1 es (0 0 1 1). Se supone que el módulo de imputación decide imputar el valor B* =1, el cual evita el fallo del edit 1.

Así pues, se produce la siguiente actualización:

PF = (4 5).

Imputación de la segunda variable del CM (D)

Un edit e pertenece al conjunto EFE(D) si cumple las siguientes condiciones (AND):

- e $\in$ PF = (4 5)
- e $\in$ AV(D) = (3 4 5)
- e $\in$ NAV (α) $\rightarrow$ cuando α es es un conjunto vacío (como ocurre ahora puesto que D es la última variable a imputar), se ignora esta condición.

El resultado de:

- la intersección de conjuntos, es EFE(D) = (4 5).
- UNION/D (4 5) = (0 0 1).

Se supone que el módulo de imputación decide imputar el valor D*=2, el cual evita el fallo de los edits 4 y 5.

La tira binaria correspondiente a la variable B en el edit 4 es (1 1 1 0). Así pues, la única imputación admisible es B*= 4, con lo cual estaríamos en la siguiente situación:

- Dado que D*= 2 desactiva los edits (4 5), PF será un conjunto vacío, es decir, el par de imputaciones (B*= 1; D*=2), junto con los valores originales A°=2 y C°= 4, garantizan que el registro R* no fallará ningún edit.
- CCE -R° = CEE + CEI- R°. La imputación exitosa ha exigido la generación de solamente un edit implícito.
- se ha respetado el principio del cambio mínimo.
- no ha sido necesaria la generación del CCE, que está constituido por los 9 edits siguientes (5 implícitos):

CCE CORRESPONDIENTE AL CEE

Nº DE EDIT	Campo Generador y Edits Contribuyentes	VARIABLES														EDIT FALLADO (F)	VARIABLES ACTIVAS
		A			B				C				D				
		1	2	3	1	2	3	4	1	2	3	4	1	2	3		
1	explícito	1	1	0	0	0	1	1	1	1	1	1	1	1	1	F	A B
2	explícito	0	1	1	1	1	1	1	0	1	1	0	1	1	1	-	A C
3	explícito	1	0	1	1	1	1	1	0	0	1	1	0	1	0	-	A C D
4	explícito	1	1	1	1	1	1	0	1	1	1	1	**0**	0	1	F	B D
5	B (1 4)	1	1	0	1	1	1	1	1	1	1	1	0	0	1	F	A D
6	A (1 2)	1	1	1	0	0	1	1	0	1	1	0	1	1	1	F	B C
7	B (4 5)	1	1	1	1	1	1	1	0	1	1	0	0	0	1	F	C D
8	A (2 3)	1	1	1	1	1	1	1	0	0	1	0	0	1	0	-	C D
9	A (1 3)	1	1	1	0	0	1	1	0	0	1	1	0	1	0	-	B C D

En nuestro ejemplo hemos hecho deliberadamente que R° falle el 50% de los edits del CEE. En situaciones reales:

- la tasa de fallos no es naturalmente tan elevada como en el ejemplo.
- la diferencia entre los tamaños del CCE (cuando se ha podido generar) y del CCE -R° es normalmente muy grande.
- se comprueba fácilmente que si el registro a imputar fuese, por ejemplo, R°=(A=1, B=2, C= 1, D=3), ocurriría lo siguiente:
- el conjunto de edits (del CEE) fallados sería F =(4).

- las variables candidatas a formar el CM son B o D.
- tanto si se decide que CM=(B), como si se selecciona CM =(D), la imputación de la variable elegida se realiza sin tener que generar ningún edit implicado.
- la imputación respeta el principio del cambio mínimo.
- es decir, para este R° los edits originales (CEE) se comportan de modo análogo al CCE.

Por último:

- al depurar ficheros estadísticos reales de cierta complejidad (tamaño del CEE = 1500 edits; Nº de variables 146; nº total de valores válidos = 3521), el nº medio de edits generados por registro erróneo (por valores inválidos o inconsistentes) ha sido de 3,2 edits implícitos.
- al operar con el CEE, el exceso de recursos informáticos (memoria, CPU) exigidos por la generación dinámica de los edits implícitos, se compensa sobradamente con el mayor consumo que se exigiría de dichos recursos si operásemos con el CCE, dado que entonces el tamaño de las matrices de edits sería mucho mayor.

3. Imputación en el caso de errores sistemáticos

Dentro de los errores ajenos al muestreo (non-sampling errors), que son los aquí considerados, podemos distinguir dos tipos de errores:

- Errores aleatorios (***Ramdom/Aleatory errors***) son aquellos que se producen sobre un conjunto de datos estadísticos, en cualquiera de sus fases de elaboración, y tienen las siguientes características:
  - o Se producen fundamentalmente por falta de cuidado.
  - o pueden ocurrir en cualquier momento y afectar a cualquier variable (uniformidad en su distribución).
- Errores sistemáticos (***Systematic errors***) son aquellos que se producen:
  - o al no comprender bien las preguntas, conceptos, definiciones o instrucciones, tanto por parte de quien responde como por parte de los agentes que intervienen en las distintas etapas del proceso estadístico.
  - o de forma intencionada por quien responde (para proteger su intimidad, por desconfianza de que su información pueda usarse con fines fiscales, policiales, etc.)

La metodología "pura" propuesta por FELLEGI&HOLT es totalmente satisfactoria para el tratamiento de los errores aleatorios. Sin embargo, no es adecuada para la depuración de los errores sistemáticos, tales como los producidos en el caso del Censo de Edificios realizado en España en el año 1981 y descritos en la referencia bibliográfica (2). Por ello, el Instituto Nacional de Estadística incluyó en el sistema DIA (basado en la metodología FELLEGI&HOLT y especializado en la Depuración e Imputación Automáticas de datos cualitativos) un módulo orientado al tratamiento de los errores sistemáticos, los cuales son eliminados en base a las llamadas RIDs (Reglas de Imputación Determinística).

3.1 Expresión de una RID

Una RID es una regla que combina la detección y la imputación, mediante dos partes:

- la parte condicional (que está a la izquierda del " ="), expresiva del error sistemático.
- la parte determinante de la imputación a realizar (que está a la derecha del "="), si se cumple la parte condicional.

Un ejemplo de RID sería: A (1) $\cap$ B (2) $\cap$ C (3) = B (blanco).

Se supone que los expertos han comprobado previamente que:

- se produce un error sistemático si un registro tiene los valores A=1, B=2 y C=3.
- en tal caso, lo más razonable es imputar a la variable B el valor "blanco".

El primer miembro de la RID es realmente la forma normalizada por FELLEGI&HOLT para expresar un error, es decir, es un edit (llamado EDR en el sistema DIA: Edit Derivado de RID).

En el sistema DIA:

- hay un analizador de EDITS-RID, que informa y/o resuelve posibles conflictos entre ambos.
- la entrada al módulo generador del CCE es el (CEE + el conjunto de EDR).
- en una primera fase se ejecutan las RIDs que procedan (imputaciones determinadas por los expertos).
- en una fase posterior se aplican las imputaciones decididas en base a la metodología FELLEGI&HOLT.

El desfase entre ambos tipos de imputaciones puede provocar la llamada REIMPUTACION, que se produce cuando una variable (X) es imputada en la primera fase por una RID, y luego en la segunda fase el sistema decide la inclusión de la variable X en el CM, de tal manera que el valor imputado finalmente difiere de lo determinado en la RID.

Una forma de evitar:

- por una parte, el tratamiento inadecuado de los errores sistemáticos, al aplicar estrictamente el sistema FELLEGI&HOLT.
- y por otra, la posible reimputación, al utilizar las RIDs.

sería utilizar un tipo especial de edit, que llamaremos EID (Edit de Imputación Determinística), que se obtiene a partir de una RID.

3.2 Conversión de una RID en un EID

En una RID hay dos tipos de variables:

- hay solamente una variable que aparece en los dos miembros de la RID. A tal variable la llamaremos genéricamente variable ID, por ser la única variable objeto de la Imputación Determinística. Así pues, la variable ID aparece, en el primer miembro, expresando una de las condiciones exigidas para que se produzca el error sistemático, y en el segundo, determinando la imputación requerida.
- las demás variables de la RID, que aparecen solamente en su primer miembro y expresan el resto de las condiciones exigidas para que se produzca el error sistemático, son las que llamaremos variables condicionales distintas de la ID.

En nuestro anterior ejemplo de RID, serían:

- B, la variable ID.
- A y C, las variables condicionales distintas de B=ID.

Las reglas para convertir una RID en un EID (cuya comprensión se facilita al aplicarlas posteriormente a la RID del ejemplo anterior), son las siguientes:

1. La condición de fallo impuesta en la RID a una variable condicional distinta de la variable ID, se expresa de igual forma en el EID.

2. La condición de fallo impuesta en la RID a la variable ID, se convierte en el EID en la condición de fallo impuesta a la variable IMA-ID, que es la imagen de la variable ID.

3. El complemento (¬) a la imputación determinada en la RID para la variable ID, se convierte en el EID en la condición de fallo impuesta a la variable ID.

La RID anterior [A (1) ∩ B (2) ∩ C (3) = B (blanco)] la convierte el sistema en el siguiente EID:

A (1) ∩ IMA- B (2) ∩ C (3) ∩ B (¬blanco).

Como puede verse, el EID resultante de la conversión anterior responde perfectamente a la estructura exigida por FELLEGI&HOLT a un edit. Lo único que se requiere para volver a operar prácticamente de conformidad con lo exigido por el sistema FELLEGI&HOLT, con las ventajas que ello comporta, es lo siguiente:

- Si hay RIDs, el sistema amplia la longitud del registro para añadir al final el espacio requerido para registrar una copia (imagen) de los valores que tienen las variables ID que aparecen en las distintas RIDs. Una vez acabado el proceso de imputación, se eliminan las variables imagen para que el registro imputado tenga la longitud original.

- La matriz de edits consta de tres submatrices:

  - La que registra el CEE. Tiene un número fijo de edits.

  - La que registra el conjunto de EIDs. También tiene un número fijo de EIDs.

  - La que registra el conjunto de edits generados dinámicamente (CEI-R°) para poder imputar adecuadamente el R°.Tiene un número variable de edits derivados.

El sistema determina, como de costumbre, el conjunto de edits fallados (F), pero, por haber EIDs, hace lo siguiente:

- Las variables imagen (IMA-X), que son meramente auxiliares, no se imputan nunca (no se incluyen nunca en el CM).

- Si falla un EID, la correspondiente variable ID se incluye en el CM, lo que implica que se respetará siempre la imputación exigida en la RID correspondiente al EID fallado. Así pues, en nuestro ejemplo de RID, el sistema incluiría la variable B en el CM y el módulo de imputación solamente puede imputar el valor "blanco", que es lo exigido por la RID.

- Una vez acabada la imputación exigida por los EIDs fallados, comienza el proceso habitual de imputación de las variables incluidas en el CM para corregir los fallos que se hayan producido en el CEE.

- Dado que en el CM no puede haber variables repetidas, es imposible que una variable se impute dos veces, es decir, desaparece la posibilidad de REIMPUTACIÓN

4. Edits aritméticos

Se supone que el usuario ha establecido los siguientes edits aritméticos, de tipo lineal:

- $A+B+C \neq T$
- $C-B \geq 0$

el registro a depurar (R°) tiene los datos siguientes:

- A = 14
- B = 5
- C = 5
- T = 32

El primero de los edits anteriores no respeta la normalización establecida para los edits aritméticos de tipo lineal, consistente en que deben ser expresados mediante un Tipo de Desigualdad estricta (TD >0) o débil (TD ≥ 0). Para normalizarlo, lo descomponemos en las dos desigualdades siguientes:

- $A+B+C>T \rightarrow A+B+C-T>0$
- $A+B+C<T \rightarrow A+B+C-T<0$

La dirección de la 2ª desigualdad no está normalizada. Al normalizarla, tenemos: --

- (A+B+C-T)>0. Así pues, el conjunto de normalizado de edits aritméticos es el siguiente.

- $e(1) \rightarrow A+B+C-T>0$
- $e(2) \rightarrow T-(A+B+C)>0$
- $e(3) \rightarrow C-B \geq 0$

En base a los edits anteriores, se obtiene la correspondiente matriz de coeficientes:

MATRIZ DE COEFICIENTES (M1)

EDITS (i)	VARIABLES (j)				TD(i)
	A	B	C	T	
1	1	1	1	-1	>
2	-1	-1	-1	1	>
3	0	-1	1	0	≥

Para abordar el problema de la imputación debemos considerar dos hipótesis:

- HIPÓTESIS 1: SE DISPONE DEL CCE.
- HIPÓTESIS 2: NO SE DISPONE DEL CCE.

4.1 Imputación cuando se dispone del CCE

Se supone que a partir del CEE, facilitado por los expertos en el tema considerado, se ha generado el correspondiente CCE, aplicando de forma reiterada el sistema descrito por FELLEGI&HOLT (combinación de pares de edits cuyos coeficientes en la variable generadora son ambos distintos de cero y de distinto signo).

Es decir, a partir de los siguientes edits originales:

$e(1) \rightarrow A+B+C- T \ >0$

$e(2) \rightarrow T- (A+B+C) >0$

$e(3) \rightarrow C-B \geq 0$

se han generado los siguientes edits derivados (implicados por los edits originales):

- $e(4) \rightarrow A+2C -T \geq 0$, con edits contribuyentes (originales 1 y 3) y variable generadora B.
- $e(5) \rightarrow -A-2B +T \geq 0$, con edits contribuyentes (originales 2 y 3) y variable generadora C.

Posteriormente (en el apartado 4.3) se detalla la forma de obtener el CCE de la forma propuesta por FELLEGI&HOLT.

Al igual que en el caso de los edits lógicos, para determinar el conjunto mínimo (CM) de campos a imputar ha de determinarse previamente el conjunto de edits fallados (F) por R°.

EDITS FALLADOS POR R°

En base a los datos de la matriz M1 se calculan los valores de su matriz asociada (M2), haciendo lo indicado a continuación:

MATRIZ M1

EDITS (i)	VARIABLES (j)				TD(i)
	A	B	C	T	
1	1	1	1	-1	>
2	-1	-1	-1	1	>
3	0	-1	1	0	≥
4=(1,3,B)	1	0	2	-1	≥
5=(2,3,C)	-1	- 2	0	1	≥
DATOS DEL REGISTRO ORIGINAL					
R°	14	5	5	32	--

MATRIZ M2

EDITS (i)	Productos de coeficientes por valores de R°: p(i, j) = c(i, j).R°(j)				$\sum$ p (i, j)	FALLADO (F)	Variables Activas
	A	B	C	T			
1	14	5	5	-32	- 8 < 0	-	A B C T
2	-14	-5	-5	32	8 > 0	F	A B C T
3	0	-5	5	0	0 ≥ 0	F	B C
4=(1,3,B)	14	0	10	-32	- 8 < 0	-	A C T
5=(2,3,C)	-14	-10	0	32	8 ≥ 0	F	A B T

El conjunto de edits fallados por R° se determina en base al valor (no negativo) obtenido en la columna "$\sum$ p(i, j)" de la matriz M2. Así pues, vemos que los edits fallados son: F = (2 3 5).

DETERMINACIÓN DEL CONJUNTO MINIMO (CM) DE CAMPOS A IMPUTAR

En el CM ha de figurar al menos una de las variables activas en cada uno de los edits fallados. Tales variables son las siguientes:

- en el edit 2 son activas las variables (A B C T)
- en el edit 3 son activas las variables (B C)
- en el edit 5 son activas las variables (A B T)

La única variable que cubre los tres edits fallados es la B.

Así pues, se tiene el conjunto CM =(B).

IMPUTACIÓN DE LA VARIABLE B

Igual que en el caso de los edits cualitativos, se tiene los siguientes conjuntos:

- TE = (1 2 3 4 5)
- NPF = (4): un edit pertenece al conjunto NPF cuando ninguna de sus variables activas pertenece al CM.
- PF = (1 2 3 5)

Por otra parte, el edit e pertenece al conjunto EFE (B) si cumple las siguientes condiciones (AND):

- e $\in$ PF.
- e $\in$ AV(B).
- e $\in$ NAV(α) = Edits No Activos en α, siendo α un cierto conjunto (no vacío) de variables.

En nuestro caso, dado que después de imputar la variable B no existe ninguna variable pendiente de imputación, α es un conjunto vacío. Así pues, ignorando la condición "e $\in$ NAV(α)", puesto que solamente es aplicable cuando α no es un vacío, se tiene los siguientes conjuntos de edits:

- AV(B) = (1 2 3 5): edits en los cuales está Activa la variable X (su coeficiente es distinto de cero).
- EFE(B) = Edits que cumplen las siguientes condiciones (AND):
  - e $\in$ PF = (1 2 3 5)
  - e $\in$ AV(B) = (1 2 3 5)

En definitiva, se tiene EFE (B) = (1 2 3 5)

Un valor de B será adecuado como imputación si impide el fallo de todos los edits pertenecientes al conjunto EFE (B).

VALORES DE B QUE EVITAN EL FALLO DE LOS EDITs DEL CONJUNTO EFE (B)

- Para evitar el fallo del e(1) debe ocurrir que: A +B + C - T $\leq$ 0.

  Dados los valores fijos (14 5 32) que tienen las variables (A C T), para evitar el fallo del e(1), debe ser: B $\leq$ 13.

- Para evitar el fallo del e(2) debe ocurrir que: T - A -B – C $\leq$ 0.

  Dados los valores fijos (14 5 32) que tienen las variables (A C T), para evitar el fallo del e(2), debe ser: B $\geq$ 13.

- Para evitar el fallo del e(3) debe ocurrir que: -B + C < 0.

  Dado el valor fijo (5) que tiene la variable (C), para evitar el fallo del e(3), debe ser: B > 5.

- Para evitar el fallo del e(5) debe ocurrir que: - A -2B + T < 0.

  Dados los valores fijos (14 32) que tienen las variables (A T), para evitar el fallo del e(5), debe ser: B > 9.

DETERMINACIÓN DE COTAS

En este caso tenemos tres Cotas Inferiores (CI):

- CI(2,B) $\geq$ 13
- CI(3,B) >5
- CI(5,B) >9

Por tanto será: CImxa (B) $\geq$13

Por otra parte, tenemos solamente una Cota Superior (CS):

- CS (1, B) $\leq$13.

Por tanto será: CSmna (B) ≤13.

Dado que Clmxa (B) ≤ Csmna (B), el IVA(B) (es decir, el Intervalo de Valores Admisibles para la variable B) viene dado por: 13 ≤ B* ≤ 13.

Así pues, la única imputación posible será B* = 13, con la cual se determina un registro sin errores: R* = (14 13 5 32).

4.2 Imputación cuando no se dispone del CCE

Si no se dispone del CCE, se opera con el CEE y se tiene la siguiente matriz de coeficientes:

MATRIZ M1

EDITS(i)	VARIABLES (j)				TD(i)
	A	B	C	T	
1	1	1	1	-1	>
2	-1	-1	-1	1	>
3	0	-1	1	0	≥
DATOS DEL REGISTRO ORIGINAL					
R°	14	5	5	32	--

MATRIZ M2

EDITS (i)	productos de coeficientes por valores de R°: p(i, j) = c(I, j).R°(j)				∑ p(i, j)	FALLADO (F)	Variables Activas
	A	B	C	T			
1	14	5	5	-32	- 8 < 0	-	A B C T
2	-14	-5	-5	32	8 > 0	F	A B C T
3	0	-5	5	0	0 ≥ 0	F	B C

DETERMINACIÓN DEL CONJUNTO MÍNIMO (CM) DE CAMPOS A IMPUTAR

En el CM ha de figurar al menos una de las variables activas en cada uno de los edits fallados. Tales variables son las siguientes:

- en el edit 2 son activas las variables (A B C T)
- en el edit 3 son activas las variables (B C)

las variables B y C cubren los dos edits fallados. Así pues, el CM podría estar constituido por solamente una cualquiera de ellas.

Supuesto que se decide aleatoriamente que CM =(B), se procede a determinar su imputación del modo siguiente:

IMPUTACIÓN DE LA VARIABLE B

En nuestro caso, tenemos los siguientes conjuntos de edits:

- TE = (1 2 3).
- NPF = (∅) → la variable B está activa en los tres edits.
- PF = (1 2 3).

Por otra parte, el edit e pertenece al conjunto EFE (B) si cumple las siguientes condiciones (AND):

- e ∈ PF = (1 2 3).
- e ∈ AV(B) = (1 2 3).

En definitiva, se tiene EFE (B) = (1 2 3)

Un valor de B será adecuado como imputación si impide el fallo de todos los edits pertenecientes al conjunto EFE (B).

VALORES DE B QUE EVITAN EL FALLO DE LOS EDITs DEL CONJUNTO EFE (B)

- Para evitar el fallo del e(1) debe ocurrir que: A + B + C- T ≤ 0

  Dados los valores fijos que tienen las variables A, C y T, para evitar el fallo del e(1), debe ser: B ≤ 13.

- Para evitar el fallo del e(2) debe ocurrir que: T – A - B – C ≤ 0.

  Dados los valores fijos que tienen las variables A, C y T, para evitar el fallo del e(2), debe ser: B ≥ 13.

- Para evitar el fallo del e(3) debe ocurrir que: - B + C < 0 → B > 5

DETERMINACIÓN DE COTAS

En este caso tenemos las siguientes cotas inferiores:

CI (2, B) = 13(≥)

CI (3, B) = 5 (>)

Por tanto será: CImxa (B) = 13(≥).

Por otra parte, solamente tenemos una cota superior:

- CS (1, B) = 13(≤).

Por tanto será: CSmna(B) = 13(≤)

Dado que CImxa (B) ≤ CSmna(B), el IVA(B) (es decir, el Intervalo de Valores Admisibles para la variable B) viene dado por: 13 ≤ B* ≤ 13. Así pues, la única

imputación posible será B* = 13, con lo cual se determina un registro sin errores: R* = (14 13 5 32).

¿Qué hubiera ocurrido si en lugar de decidir aleatoriamente la inclusión de la variable B en el CM, se hubiese seleccionado la variable C?. Lo vemos a continuación:

IMPUTACIÓN DE LA VARIABLE C

En nuestro caso, tenemos los siguientes conjuntos de edits:

- TE = (1 2 3).
- NPF = ($\varnothing$) $\rightarrow$ la variable C está activa en los tres edits.
- PF = (1 2 3)

Por otra parte, el edit e pertenece al conjunto EFE (C) si cumple las siguientes condiciones (AND):

- e $\in$ PF = (1 2 3).
- e $\in$ AV(C) = (1 2 3).

En definitiva, se tiene EFE (C) = (1 2 3).

Un valor de C será adecuado como imputación si impide el fallo de todos los edits pertenecientes al conjunto EFE (C).

VALORES DE C QUE EVITAN EL FALLO DE LOS EDITs DEL CONJUNTO EFE (C)

- Para evitar el fallo del e(1) debe ocurrir que: A + B + C- T $\leq$ 0.

  Dados los valores fijos que tienen las variables A, B y T, para evitar el fallo del e(1), debe ser: C $\leq$ 13.

- Para evitar el fallo del e(2) debe ocurrir que: T – A - B – C $\leq$ 0.

  Dados los valores fijos que tienen las variables A, B y T, para evitar el fallo del e(2), debe ser: $\geq$ 13.

- Para evitar el fallo del e(3) debe ocurrir que: - B + C < 0.

  Dado el valor fijo que tiene la variable B, para evitar el fallo del e(3), debe ser

  C < 5.

DETERMINACIÓN DE COTAS

En este caso tenemos solamente una cota inferior:

CI (2, C) = 13 ($\geq$).

Por tanto será: CImxa (C) = 13($\geq$).

Por otra parte, tenemos las siguientes cotas superiores:

- CS (1, C) = 13 (≤).
- CS (3, C) = 5 (<).

Por tanto será: CSmna(C) = 5 (<)

Dado que CImxa (C) > CSmna(C), no es posible imputar a C un valor que evite el fallo de todos los edits pertenecientes al conjunto EFE(C). La variable C es una variable vacía.

La inexistencia de una imputación adecuada para la variable C se deriva de que el CEE no es un Conjunto Completo de Edits para R° (CCE -R°). Para completarlo, se genera dinámicamente un edit derivado en base a, precisa y exclusivamente, lo siguiente:

- edits contribuyentes, los del conjunto EFE(C).
- variable generadora, la C.

Los coeficientes de la variable generadora en los edits contribuyentes (que ahora son tres, en lugar de los dos exigidos por el sistema generador propuesto por FELLEGI&HOLT) son los siguientes:

- c(1, C) = 1
- c(2, C) = - 1
- c(3, C) = 1
- todos ellos no nulos, lógicamente, por ser también todos ellos activos en la variable generadora C.

Nota: dado que, en general, el nº de edits contribuyentes puede ser mayor de dos, y que el sistema generador propuesto por FELLEGI&HOLT solamente admite que sean dos, se ha definido y utilizado un nuevo sistema generador para obviar este problema. Puede comprobarse fácilmente que el nuevo sistema, cuando el nº de edits contribuyentes es 2, genera el mismo edit derivado que el generado por el sistema FELLEGI&HOLT.

En este caso, por ser $\sum c(i, g) \neq 0$, será $k(i) = \{ c(i, g)/ \sum c(i, g)^2 \} - 1/ \sum c(i, g)$.

Los valores correspondientes a cada uno de los k(i) se obtienen a partir de los coeficientes de la variable generadora, de la forma indicada en el cuadro siguiente:

EDITS CONTRIBUYENTES (i)	VARIABLE GENERADORA (g = C)		
	c(i,g)	$c(i,g)^2$	k(i) = {c(i,g) /T2} - 1/T1
1	1	1	k(1) = 1/3 -1 = - 2/3
2	-1	1	k(2) = -1/3 -1 = - 4/3
3	1	1	k(3) = 1/3 -1 = - 2/3
TOTAL	T1=$\sum$c(i, g)=1	T2=$\sum c(i, g)^2$=3	----------

En base a los datos de nuestro caso, se tiene lo siguiente:

MATRIZ M1, AMPLIADA CON LOS COEFICIENTES K(i)

EDITS (i)	VARIABLES (j)				k(i)
	A	B	C	T	
1	1	1	1	-1	- 2/3
2	-1	-1	-1	1	- 4/3
3	0	-1	1	0	- 2/3
DATOS DEL REGISTRO ORIGINAL					
R°	14	5	5	32	----------

Multiplicando, para cada edit, sus coeficientes c(i,g) por su correspondiente k(i), se obtiene la siguiente matriz M3:

MATRIZ PRODUCTO (M3)

EDITS (i)	Valores de celda = k(i).c (i, g)				TD (i) Ver nota (a)
	A	B	C	T	
1	- 2/3	- 2/3	- 2/3	2/3	<
2	4/3	4/3	4/3	- 4/3	<
3	0	2/3	-2/3	0	≤
TOTAL	2/3	4/3	0	- 2/3	$\leq \rightarrow$ TD no normalizada

Nota (a): por ser negativos los k(i), en los edits contribuyentes varía el sentido de su TD(i).

El TD correspondiente al edit derivado representado por la fila TOTAL no está normalizado, pues su expresión sería: $(2/3)\,A + (4/3)B - (2/3)T \leq 0$.

Multiplicando por (-1), para normalizarlo, y dividiendo por 2/3, se tiene la siguiente expresión normalizada y simplificada para el nuevo edit:

$4=(1, 2, 3, C) \rightarrow - A - 2B + T \geq 0$

Obsérvese que el ahora numerado como edit 4 coincide con el edit 5 que se obtiene al generar, mediante el sistema propuesto por FELLEGI&HOLT, el conjunto completo de edits implicados.

Debe tenerse presente que la normalización de los TD(i) solamente puede realizarse si todos los k(i) son del mismo signo (positivos o negativos). En otro caso, la normalización provocaría el incumplimiento de la condición (necesaria) exigida para que la expresión lineal generada sea un edit esencialmente nuevo, consistente en que $\sum_i k(i)c(i, g) = 0$, es decir, que en el nuevo edit debe ser nulo el coeficiente correspondiente a la variable generadora.

En definitiva, si no todos los k(i) son del mismo signo, la expresión generada no es un edit. En tal caso, se genera un nuevo CM forzando la exclusión (no selección) en

el mismo de la variable potencialmente generadora, pero realmente estéril. La exclusión:

- es aplicable solamente para el CM actual, de tal manera que en un CM posterior (supuesto que lo hubiese) la variable considerada recobraría su condición de candidata a imputable.
- se justifica para evitar la repetición de la generación de un "no edit", que podría ocurrir si la variable considerada fuese la candidata más sospechosa, o si fuese seleccionada aleatoriamente entre un conjunto de candidatas igualmente sospechosas.

Dado que al añadir el edit 4 varía el conjunto total de edits, antes de pasar al proceso de imputación debe comprobarse si el edit derivado es fallado por R°.

EDITS FALLADOS POR R°

En base a los datos de la matriz M1 se calculan los valores de su matriz asociada (M2), haciendo lo indicado a continuación:

MATRIZ M1

EDITS (i)	VARIABLES (j)				TD (i)
	A	B	C	T	
1	1	1	1	-1	>
2	-1	-1	-1	1	>
3	0	-1	1	0	≥
4=(1,2,3,C)	-1	- 2	0	1	≥
DATOS DEL REGISTRO ORIGINAL					
R°	14	5	5	32	--

MATRIZ M2

EDITS (i)	productos de coeficientes por valores de R°: p(i, j) = c(i, j).R°(j)				$\sum$ p(i,j)	FALLADO (F)	Variables Activas
	A	B	C	T			
1	14	5	5	-32	- 8 < 0	-	A B C T
2	-14	-5	-5	32	8 > 0	F	A B C T
3	0	-5	5	0	0 ≥ 0	F	B C
4=(1, 2, 3, C)	-14	- 10	0	32	8 ≥ 0	F	A B T

Así pues, vemos que los edits fallados son: F = (2 3 4).

DETERMINACIÓN DEL CONJUNTO MINIMO (CM) DE CAMPOS A IMPUTAR

En el CM ha de figurar al menos una de las variables activas en cada uno de los edits fallados. Tales variables son las siguientes:

- en el edit 2 son activas las variables (A B C T)
- en el edit 3 son activas las variables (B C)
- en el edit 4 son activas las variables (A B T)

la única variable que cubre los tres edits fallados es la B.

Así pues, el sistema decide que: CM = (B).

IMPUTACIÓN DE LA VARIABLE B

Ahora se tiene los siguientes conjuntos de edits:

- TE = (1 2 3 4).
- NPF = ($\varnothing$) $\rightarrow$ la variable B está activa en todos los edits.
- PF = Edits que Puede Fallar = (1 2 3 4). El registro original R° no falla alguno de estos edits, pero puede fallar cualquiera de ellos dependiendo de los valores que se decida imputar a la variable B.
- AV(B) = (1 2 3 4)
- EFE(B) = Edits que cumplen las siguientes condiciones (AND):
  - o $e \in$ PF = (1 2 3 4)
  - o $e \in$ AV(B) = (1 2 3 4)

En definitiva, se tiene la siguiente intersección de conjuntos: EFE (B) = (1 2 3 4).

Un valor de B será adecuado como imputación si impide el fallo de todos los edits pertenecientes al conjunto EFE (B).

VALORES DE B QUE EVITAN EL FALLO DE LOS EDITs DEL CONJUNTO EFE (B)

- Para evitar el fallo del e(1) debe ocurrir que: $A + B + C - T \leq 0$.

  Dados los valores fijos (14 5 32) que tienen las variables (A C T), para evitar el fallo del e(1), debe ser: $B \leq 13$.

- Para evitar el fallo del e(2) debe ocurrir que: $T - A - B - C \leq 0$.

  Dados los valores fijos (14 5 32) que tienen las variables (A C T), para evitar el fallo del e(2), debe ser: $B \geq 13$.

- Para evitar el fallo del e(3) debe ocurrir que: $-B + C < 0$.

  Dados el valor fijo (5) que tiene la variable (C), para evitar el fallo del e(3), debe ser: $B > 5$.

- Para evitar el fallo del e(4) debe ocurrir que: - A -2B + T < 0.

  Dados los valores fijos (14 32) que tienen las variables (A T), para evitar el fallo del e(4), debe ser: B > 9.

DETERMINACIÓN DE COTAS

En este caso tenemos tres cotas inferiores:

- CI(2,B) =13 (≥)
- CI(3,B) = 5(>)
- CI(4,B) = 9 (>)

Por tanto será: CImxa (B) = 13(≥)

Por otra parte, tenemos solamente una cota superior:

- CS (1, B) = 13 (≤)

Por tanto será: CSmna(B) = 13(≤).

Dado que CImxa (B) ≤ CSmna(B), el IVA(B) (es decir, el Intervalo de Valores Admisibles para la variable B) viene dado por: 13 ≤ B* ≤ 13. Así pues, la única imputación posible será B* = 13, con lo cual se determina un registro sin errores: R* = (14 13 5 32).

Del resultado obtenido cuando el sistema decide aleatoriamente que sea CM= (C), y ocurre que C es una variable vacía, se concluye lo siguiente:

1. se respeta el principio del cambio mínimo (solamente se imputa una variable: la variable B).
2. se obtiene el mismo resultado que cuando el sistema decide aleatoriamente que sea CM= (B), que a su vez coincide con el obtenido al operar con el CCE.
3. mediante un proceso iterativo que combina la determinación del CM adecuado con la imputación pertinente, se ha generado dinámicamente un edit derivado (4). Así pues, para nuestro R° se ha determinado que CCE–R°= (1 2 3 4)= CEE + CEI-R°.

Por último, aunque el subsistema encargado de procesar los edits cuantitativos no está todavía operativo, dada su similitud funcional con el subsistema especializado en el proceso de edits lógicos, lo normal es que lo dicho respecto a éstos al final del apartado 2 sea totalmente aplicable a los edits cuantitativos, en situaciones reales de gran complejidad.

4.3 Generación del CCE mediante el sistema propuesto por Fellegi&Holt

El proceso generador de edits derivados se realiza ejecutando una serie repetitiva de ciclos generadores (CG) hasta que en uno de ellos no sea posible generar ningún edit realmente nuevo (es decir, no redundante de uno ya existente).

En cada ciclo generador se intenta la generación en base a cada una de las variables componentes del registro a depurar (todas ellas son, en principio, variables potencialmente generadoras).

Para cada edit derivado se registra cuales fueron sus edits contribuyentes y cual fue su variable generadora.

En el 1º de los ciclos generadores (CG =1) se intenta la generación solamente con pares de edits originales (a los cuales convenimos en asignarles el CG = 0, indicando con ello que no han sido generados en ningún CG) que cumplan las dos condiciones exigidas por FELLEGI&HOLT:

- Los coeficientes correspondientes a una misma variable en ambos edits son no nulos.
- Dichos coeficientes son de signo distinto.

En el 2º de los ciclos generadores (CG =2) se intenta la generación con pares de edits que, además de cumplir las dos condiciones exigidas, sean:

- Uno de ellos, original; el otro, derivado en el ciclo generador anterior (CG =1).
- O bien, dos edits derivados en el ciclo generador anterior.

En general, al empezar un nuevo ciclo generador (por ejemplo, el CG = n), se intenta la generación apareando todos los edits existentes antes de empezar el CG = n, con cada uno de los edits generados en el CG = (n –1).

EDITS DERIVADOS EN EL CG =1

Edit derivado en base a:

- Los edits contribuyentes (1 2).
- La variable generadora A

Los coeficientes de la variable generadora en los edits contribuyentes son:

- c(1, A) = 1
- c (2, A) = -1
- ambos no nulos y de distinto signo.

El edit derivado será una combinación lineal de los edits contribuyentes.

En general, se tiene la siguiente combinación lineal:

$$e(d) = \sum k(i).\ e(i) \quad \text{(siendo i un edit contribuyente)}$$

El valor de k(i) se determina en función de $\sum c(i, g)$:

- Si $\sum c(i, g) = 0$, es decir, si para los edits contribuyentes es nula la suma de todos los coeficientes de la variable generadora, será k(i) = 1.
- Si $\sum c(i, g) \neq 0$, será $k(i) = \{ c(i, g) / \sum c(i, g)^2 \} – 1/ \sum c(i, g)$.

NOTA: la fórmula anterior es diferente de la propuesta por FELLEGI&HOLT, pero es la que se ha determinado y utilizado para averiguar los valores de los coeficientes de la combinación lineal de errores, k(i), porque tiene las siguientes propiedades:

- Determina para el edit derivado los mismos coeficientes c(i,g) que los obtenidos aplicando el sistema propuesto por FELLEGI&HOLT, para el caso de dos edits contribuyentes.
- Es aplicable a cualquier número de edits contribuyentes.
- Su aplicación dinámica es muy sencilla puesto que k(i) = 1 cuando $\sum$ c(i, g) = 0, caso muy frecuente.
- Los valores de k(i) determinados del modo indicado hacen que sea $\sum$ k(i).c(i, g) = 0. Es decir, que en el edit derivado no sea activa la variable generadora por ser nulo su coeficiente (como comprobaremos posteriormente), que es la condición necesaria exigida por FELLEGI&HOLT para que el edit derivado sea un edit esencialmente nuevo.

En el caso actual, por ser $\sum$ c(i, g) = 0, se tiene que k(i) = 1.

Una vez que hemos determinado los valores de k(i), multiplicando cada fila (i) de coeficientes de M1 por su correspondiente k(i), obtenemos una nueva matriz (que llamaremos matriz producto M3) cuya fila de totales, $\sum$k(i).c(i, g), representa los coeficientes que corresponden a cada variable en el edit derivado e(d).

En base a los datos del caso actual, se tiene lo siguiente:

MATRIZ M1, AMPLIADA CON LOS COEFICIENTES K(i)

EDITS (i)	VARIABLES (j)				K (i)
	A	B	C	T	
1	1	1	1	-1	1
2	-1	-1	-1	1	1

MATRIZ PRODUCTO (M3)

EDITS (i)	Valores de celda = k(i).c(i, g)				TD (i)
	A	B	C	T	
1	1	1	1	-1	>
2	-1	-1	-1	1	>
TOTAL	0	0	0	0	....

Se observa que:

- El total obtenido para la variable generadora (A) es nulo, lo que pone de manifiesto que en el hipotético edit derivado no estaría activa dicha variable por ser nulo su coeficiente.
- Al no haber al menos dos variables activas, la suma de los edits 1 y 2 no es un edit. Por tanto, en lo que sigue no se intenta la generación en base al par de edits (1 2).

Así pues, se intenta la generación de un edit derivado en base a:

- Los edits contribuyentes (1 3).
- La variable generadora B

Los coeficientes de la variable generadora en los edits contribuyentes son:

- c(1, B) = 1
- c(3, B) = -1
- ambos no nulos y de distinto signo.

También en este caso, por ser $\sum$ c(i, g) = 0, se tiene que k(i) = 1.

MATRIZ M1, AMPLIADA CON LOS COEFICIENTES
K (i)

EDITS (i)	VARIABLES (j)				K (i)
	A	B	C	T	
1	1	1	1	-1	1
3	0	-1	1	0	1

MATRIZ PRODUCTO (M3)

EDITS (i)	Valores de celda = k(i).c(i, g)				TD (i)
	A	B	C	T	
1	1	1	1	-1	>
3	0	-1	1	0	≥
TOTAL	1	0	2	- 1	≥

Se observa que el total obtenido para la variable generadora (B) es nulo, lo que pone de manifiesto que en el edit derivado no estará activa dicha variable por ser nulo su coeficiente.

En definitiva, la desigualdad expresiva del edit derivado será la siguiente:

$4 = (1, 3, B) \rightarrow \quad A + 2\,C - T \geq 0.$

Edit derivado en base a:

- Los edits contribuyentes (2 3).
- La variable generadora C

Los coeficientes de la variable generadora en los edits contribuyentes son:

- c(2, C) = - 1
- c(3, C) = 1
- ambos no nulos y de distinto signo.

También en este caso, por ser $\sum c(i, g) = 0$, se tiene que k(i) = 1.

MATRIZ M1, AMPLIADA CON LOS COEFICIENTES K(i)

EDITS(i)	VARIABLES (j)				K (i)
	A	B	C	T	
2	-1	-1	-1	1	1
3	0	-1	1	0	1

MATRIZ PRODUCTO (M3)

EDITS (i)	Valores de celda = k(i).c(i, g)				TD (i)
	A	B	C	T	
2	-1	-1	-1	1	>
3	0	-1	1	0	≥
TOTAL	- 1	- 2	0	1	≥

En definitiva, la expresión del edit derivado será la siguiente:

$5 = (2, 3, C) \rightarrow \quad - A - 2\,B + T \geq 0.$

EDITS DERIVADOS EN EL CG = 2

Una vez acabado el CG = 1, se inicia el 2º ciclo generador teniendo en cuenta que los nuevos edits derivados determinan actualmente la siguiente matriz M1:

MATRIZ DE COEFICIENTES DE EDITS-VARIABLES (M1)

EDITS (i)	VARIABLES (j)				TD (i)
	A	B	C	T	
EDITS ORIGINALES					
1	1	1	1	-1	>
2	-1	-1	-1	1	>
3	0	-1	1	0	≥
Edits derivados en el primer ciclo generador (solamente se combinaron los edits originales)					
4 =(1, 3, B)	1	0	2	-1	≥
5 =(2, 3, C)	-1	- 2	0	1	≥

Edit derivado en base a:

- Los edits contribuyentes (1 5).
- La variable generadora A

En base a los datos del caso actual, se tiene lo siguiente:

MATRIZ M1, AMPLIADA CON LOS COEFICIENTES
K (i)

EDITS (i)	VARIABLES (j)				K (i)
	A	B	C	T	
1	1	1	1	-1	1
5	-1	-2	0	1	1

MATRIZ PRODUCTO (M2)

EDITS (i)	Valores de celda = k(i).c(i, g)				TD (i)
	A	B	C	T	
1	1	1	1	-1	>
5	-1	-2	0	1	≥
TOTAL	0	-1	1	0	≥

Así pues, la expresión del edit derivado es:

- B + C ≥ 0, pero se descarta por ser redundante (igual al edit 3).

Edit derivado en base a:

- Los edits contribuyentes (2 4).
- La variable generadora A

En base a los datos de nuestro caso actual, se tiene lo siguiente:

MATRIZ M1, AMPLIADA CON LOS COEFICIENTES K (i)

EDITS (i)	VARIABLES (j)				K (i)
	A	B	C	T	
2	-1	-1	-1	1	1
4	1	0	2	-1	1

MATRIX PRODUCTO (M3)

EDITS (i)	Valores de celda = k(i).c(i, g)				TD (i)
	A	B	C	T	
2	-1	-1	-1	1	>
4	1	0	2	-1	≥
TOTAL	0	-1	1	0	≥

Así pues, la expresión del edit derivado es:

$-B + C \geq 0$, pero se descarta por ser igual al edit 3.

Edit derivado en base a:

- Los edits contribuyentes (4 5).
- La variable generadora A

Con los datos de nuestro caso actual, se tiene lo siguiente:

MATRIZ M1, AMPLIADA CON LOS COEFICIENTES K (i)

EDITS (i)	VARIABLES (j)				K (i)
	A	B	C	T	
4	1	0	2	-1	1
5	-1	-2	0	1	1

MATRIZ PRODUCTO (M3)

EDITS (i)	Valores de celda = k(i).c(i, g)				TD (i)
	A	B	C	T	
4	1	0	2	-1	≥
5	-1	-2	0	1	≥
TOTAL	0	-2	2	0	≥

La expresión del edit derivado es: $-2B + 2C \geq 0$.

Normalizando dicha expresión para que su primer coeficiente sea (-1), la expresión simplificada del edit derivado es:

$-B + C \geq 0$, pero se descarta por ser igual al edit 3.

Edit derivado en base a:

- Los edits contribuyentes (1 5).
- La variable generadora B

Los coeficientes de la variable generadora en los edits contribuyentes son:

- c(1, B) = 1
- c(5, B) = - 2
- ambos no nulos y de distinto signo.

En este caso, por ser $\sum c(i, g) \neq 0$, será $k(i) = \{c(i, g) / \sum c(i, g)^2\} - 1/\sum c(i, g)$.

Los valores correspondientes a cada uno de los k(i) se obtienen a partir de los coeficientes de la variable generadora B, de la forma indicada en el cuadro siguiente:

EDITS CONTRIBUYENTES (i)	VARIABLE GENERADORA (g = B)		
	c(i,g)	$c(i,g)^2$	k(i) = {c(i,g) /T2 }- 1/T1
1	1	1	1/5 + 1 = 6/5
5	-2	4	- 2/5 + 1 = 3/5
TOTAL	$T1=\sum c(i, g) = -1$	$T2=\sum c(i, g)^2 = 5$	----------

En base a los datos de nuestro caso actual, se tiene lo siguiente:

MATRIZ M1, AMPLIADA CON LOS COEFICIENTES
K(i)

EDITS (i)	VARIABLES (j)				K (i)
	A	B	C	T	
1	1	1	1	- 1	6/5
5	-1	-2	0	1	3/5

MATRIZ PRODUCTO (M3)

EDITS(i)	Valores de celda = k(i).c(i, g)				TD(i)
	A	B	C	T	
1	6/5	6/5	6/5	- 6/5	>
5	- 3/5	- 6/5	0	3/5	≥
TOTAL	3/5	0	6/5	-3/5	≥

En definitiva, la expresión simplificada del edit derivado será la siguiente:

A + 2C - T ≥ 0, pero se descarta por ser igual al edit 4.

Edit derivado en base a:

- Los edits contribuyentes (2 4).
- La variable generadora C

Los coeficientes de la variable generadora en los edits contribuyentes son:

- c(2, C) = -1
- c(4, C) = 2
- ambos no nulos y de distinto signo.

En este caso, por ser $\sum c(i, g) \neq 0$, será $k(i) = \{c(i, g) / \sum c(i, g)^2\} - 1/ \sum c(i, g)$.

Los valores correspondientes a cada uno de los k(i) se obtienen a partir de los coeficientes de la variable generadora C, de la forma indicada en el cuadro siguiente:

EDITS CONTRIBUYENTES (i)	VARIABLE GENERADORA (g = C)		
	c(i,g)	$c(i,g)^2$	k(i) = {c(i,g) /T2 }- 1/T1
2	-1	1	-1/5 -1 = - 6/5
4	2	4	2/5 –1 = - 3/5
TOTAL	$T1=\sum c(i, g) = 1$	$T2=\sum c(i, g)^2 = 5$	----------

En base a los datos de nuestro caso actual, se tiene lo siguiente:

MATRIZ M1, AMPLIADA CON LOS COEFICIENTES K (i)

EDITS (i)	VARIABLES (j)				K (i)
	A	B	C	T	
2	-1	-1	-1	1	- 6/5
4	1	0	2	-1	- 3/5

MATRIZ PRODUCTO (M3)

EDITS (i)	Valores de celda = k(i).c(i, g)				TD (i) Ver nota (a)
	A	B	C	T	
2	6/5	6/5	6/5	- 6/5	<
4	- 3/5	0	- 6/5	3/5	≤
TOTAL	3/5	6/5	0	-3/5	≤ (no normalizado)

Nota (a): por ser negativos los k(i), se invierte el sentido de los TD (i).

Como el TD correspondiente al edit derivado (≤) no está normalizado, para normalizarlo se multiplica la fila TOTAL por (- 1), con lo cual se tiene: - 3/5 A – 6/5 B + 3/5 T ≥ 0.

En definitiva, la expresión simplificada del edit derivado será la siguiente:

- A - 2B + T ≥ 0, pero se descarta por ser igual al edit 5.

Edit derivado en base a:

- Los edits contribuyentes (1 5).
- La variable generadora T

En base a los datos de nuestro caso actual, se tiene lo siguiente:

MATRIZ M1, AMPLIADA CON LOS COEFICIENTES
K (i)

EDITS (i)	VARIABLES (j)				K (i)
	A	B	C	T	
1	1	1	1	-1	1
5	-1	-2	0	1	1

MATRIZ PRODUCTO (M3)

EDITS (i)	Valores de celda = k(i).c(i, g)				TD (i)
	A	B	C	T	
1	1	1	1	-1	>
5	-1	-2	0	1	≥
TOTAL	0	-1	1	0	≥

En definitiva, la expresión simplificada del edit derivado es:

- B + C ≥ 0, pero se descarta por ser igual al edit 3.

Edit derivado en base a:

- Los edits contribuyentes (2 4).
- La variable generadora TEn base a los datos de nuestro caso actual, se tiene lo siguiente:

MATRIZ M1, AMPLIADA CON LOS COEFICIENTES K(i)

EDITS (i)	VARIABLES (j)				K (i)
	A	B	C	T	
2	-1	-1	-1	1	1
4	1	0	2	-1	1

MATRIZ PRODUCTO (M3)

EDITS (i)	Valores de celda = k(i).c(i, g)				TD (i)
	A	B	C	T	
2	-1	-1	-1	1	>
4	1	0	2	-1	≥
TOTAL	0	-1	1	0	≥

En definitiva, la expresión del edit derivado es:

- B + C ≥ 0, pero se descarta por ser igual al edit 3.

Edit derivado en base a:

- Los edits contribuyentes (4 5).
- La variable generadora T

En base a los datos de nuestro caso actual, se tiene lo siguiente:

MATRIZ M1, AMPLIADA CON LOS COEFICIENTES K (i)

EDITS (i)	VARIABLES (j)				K (i)
	A	B	C	T	
4	1	0	2	-1	1
5	-1	-2	0	1	1

MATRIZ PRODUCTO (M3)

EDITS (i)	Valores de celda = k(i).c(i, g)				TD (i)
	A	B	C	T	
4	1	0	2	-1	≥
5	-1	-2	0	1	≥
TOTAL	0	-2	2	0	≥

En definitiva, la expresión simplificada del edit derivado es:

- B + C ≥ 0, pero se descarta por ser igual al edit 3.

EDITS DERIVADOS EN EL CG = 3

Una vez acabado el CG = 2, se comprueba que en él no se ha generado ningún edit realmente nuevo y entonces finaliza el proceso generador.

Así pues, el CCE queda representado mediante la siguiente matriz de coeficientes:

MATRIZ M1 CORRESPONDIENTE AL CCE

EDITS (i)	VARIABLES (j)				TD (i)	CG
	A	B	C	T		
1	1	1	1	-1	>	0
2	-1	-1	-1	1	>	0
3	0	-1	1	0	≥	0
4=(1, 3, B)	1	0	2	-1	≥	1
5=(2, 3, C)	-1	- 2	0	1	≥	1

5. EDITS MIXTOS

5.1 EDITS MIXTOS

Un Edit Mixto consta de dos clases de componentes: "cualitativo = clase C" y "aritmético = clase A". Para que se produzca el fallo de un "edit mixto" es necesario el fallo de sus componentes activos (" no vacios ").

A su vez un componente (cualquiera que sea su clase) puede ser de tipo:

0 : El componente considerado es un conjunto vacio, es decir, en él no está activa ninguna variable.

1 : El componente consiste en una condición de fallo en la cual está activa solamente una variable.

2 : El componente consiste en un edit y, en consecuencia, en él habrá al menos dos variables activas.

Así pues, se tiene los 9 tipos de "edit en sentido amplio", según se describe en el cuadro siguiente:

TIPO DE EDIT SEGÚN TIPO DE SUS COMPONENTES

TIPO DEL COMPONENTE CUALITATIVO	TIPO DEL COMPONENTE ARITMÉTICO		
	0	1	2
0	1	2	3
1	4	5	6
2	7	8	9

El tipo 1 de edit carece de sentido puesto que no existe ninguna condición de fallo.

Los tipos 2 y 4 no son estríctamente edits de consistencia, puesto que en ellos solamente está activa una variable, de tipo A y C, respectivamente.

Los tipos 3 y 7 son edits de consistencia, en sentido estricto, de tipo A y C, respectivamente.

Los tipos 5, 6, 8 y 9 son realmente los únicos edits mixtos.

No obstante, por razones de brevedad, en lo que sigue usaremos la denominación genérica de edits mixtos, cualquiera que sea su tipo.

Para ilustrar la forma de operar con edits mixtos, se supone que:

Las variables aritméticas son de tipo continuo y positivas (en general, se admitirá también valores enteros y valores no positivos).

El usuario ha determinado seis edits mixtos.

Han de tener un valor positivo (en general, se admitirá también valores negativos).

El registro a imputar tiene los siguientes valores:

Rº	Aº=2	Bº=2	Cº=2	Xº=5	Yº=10	Zº =3	CTE=1

En base a los datos anteriores y a los edits mixtos explícitos, se obtiene la siguiente matriz de edits mixtos MX.

MATRIZ DE EDITS (MX)

EDIT	VARIABLES CUALITATIVAS											VARIABLES ARITMETICAS				TIPO DE EDIT COMPLETO.			F=FALLADO N=NO FALLADO I=INACTIVO			VARIABLES ACTIVAS EN EL EDIT
	A			B				C				X	Y	Z	CTE							
	1	2	3	1	2	3	4	1	2	3	4	MATRIZ DE COEFICIENTES				C	A	E	C	A	E	
1	1	1	1	1	1	1	1	1	1	1	1	2	0	- 1	0	0	2	3	I	F	F	X Z
2	0	1	1	1	1	1	1	0	1	1	0	0	0	0	0	2	0	7	F	I	F	A C
3	1	1	1	0	1	1	0	0		1	1	0	0	4	- 10	2	1	8	N	F	N	B C Z
4	1	0	1	1	1	1	0	1	1	1	1	- 1	2	4	0	2	2	9	N	F	N	A B X Y Z
5	1	1	0	1	1	1	0	0	1	1	0	0	3	0	- 5	2	1	8	F	F	F	A B C Y
6	1	1	1	0	0	1	1	1	1	1	0	5	0	- 2	2	2	2	9	N	F	N	B C X Z
Rº		*1*			*1*				*1*			*5*	*10*	*3*	1	----------						

DETERMINATION DEL CONJUNTO MINIMO (CM) DE VARIABLES IMPUTADAS

En el caso de edits mixtos, la determinación del CM de campos a imputar es un proceso análogo al realizado en el caso estricto de edits cualitativos y edits aritméticos, es decir, se basa en la determinación:

De los edits fallados (F).

Del IS correspondiente, en cada reiteración, a cada variable sospechosa que sea candidata a ser incluida en el CM.

Así pues, en base a los datos actuales:

F = (1 2 5).

variable A : F/A = 2/3.

variable B : F/A = 1/4.

variable C : F/A = 2/4.

variable X : F/A = 1/3.

variable Y : F/A = 1/2.

variable Z : F/A = 1/4.

En la 1ª reiteración el IS máximo corresponde a la variable A, que cubre los edits fallados 2 y 5. En la 2ª reiteración, para cubrir el edit fallado 1, se selecciona la variable X por ser su F/A(1/3) mayor que la F/A(1/4) correspondiente a la variable Z.

Así pues, CM = (A X).

IMPUTACION DE LA VARIABLE A

Se tiene el siguiente conjunto de edits:

TE = (1 2 3 4 5 6).

NPF = (3 6) : su fallo es evitado por los valores fijos B^{o}= 2 y C^{o}=2.

PF = (1 2 4 5).

El edit **e** pertenece al conjunto EFE (A) si verifica las siguientes condiciones (AND):

- e ∈ PF = (1 2 4 5).
- e ∈ AV(A) = (2 4 5).
- e ∈ NAV(X) = (2 3 5).

De aquí, se tiene EFE (A) = (2 5). La UNION en A de los edits 2 y 5 es el vector de bits:

2	0	1	1
5	1	1	0
∪	1	1	1

Dado que la UNION es un vector unitario, la variable A es vacía. Así pues, es necesario generar un nuevo edit mixto, de la forma indicada a continuación:

FORMA DE GENERAR EDITS MIXTOS

CASO 1: La variable vacía es de tipo C (Cualitativo).

1.1: El componente cualitativo(CC) se genera del modo siguiente :

Variable generadora: la variable vacía (en este caso, la variable A).

Edits contribuyentes: los del correspondiente EFE: en este caso EFE(A) = (2 5).

Sistema generador: el empleado al operar solamente con edits cualitativos (UNION para la variable generadora; INTERSECCION para las demás variables).

1.2: El componente aritmético(CA) se genera del modo siguiente :

Variable generadora: no existe variable generadora de tipo A (es de tipo C).

Edits contribuyentes: los del correspondiente EFE: en este caso EFE(A) = (2 5).

Sistema generador: CA = $\sum$ A(i), siendo A(i) el componente aritmético correspondiente al i-ésimo edit mixto perteneciente al EFE.

En este caso, CA= A(2) + A(5).

CASO 2: La variable vacía es de tipo A (Aritmético)

2.1 : El componente cualitativo se genera del modo siguiente :

Variable generadora: no existe variable generadora de tipo C (es de tipo A).

Edits contribuyentes: el par de edits determinado por el MGEI (Módulo Generador de Edits Implicitos).

Sistema generador: INTERSECCION de los edits contribuyentes, aplicada a todas las variables.

2.2 : El componente aritmético se genera del modo siguiente :

Variable generadora: la variable vacía.

Edits contribuyentes: el par de edits determinado por el MGEI (Módulo Generador de Edits Implícitos)

Sistema generador: el empleado al operar solamente con edits aritméticos.

Así pues, el componente cualitativo generado en este caso (∪ en A ; ∩ en B y en C), sería el siguiente :

EDIT	VARIABLES CUALITATIVAS										
	A			B				C			
	1	2	3	1	2	3	4	1	2	3	4
2	0	1	1	1	1	1	1	0	1	1	0
5	1	1	0	1	1	1	0	0	1	1	0
7	1	1	1	1	1	1	0	0	1	1	0

Por otra parte, el componente aritmético generado en este caso caso [A(2) + A(5)], sería el siguiente :

EDIT	VARIABLES ARITMETICAS			
	X	Y	Z	CTE
	MATRIZ DE COEFICIENTES			
2	0	0	0	0
5	0	3	0	-5
7	0	3	0	-5

Los últimos valores obtenidos para la matriz MX son los siguientes:

MATRIZ DE EDITS (MX)

EDIT	VARIABLES CUALITATIVAS											VARIABLES ARITMETICAS				TIPO DE COMPONENTE			F= FALLADO N= NO FALLADO I= INACTIVO			VARIABLES ACTIVAS EN EL EDIT
	A			B				C				X	Y	Z	CTE							
	1	2	3	1	2	3	4	1	2	3	4	MATRIZ DE COEFICIENTES				Q	A	E	Q	A	E	
1	1	1	1	1	1	1	1	1	1	1	1	2	0	- 1	0	0	2	3	I	F	F	X Z
2	0	1	1	1	1	1	1	0	1	1	0	0	0	0	0	2	0	7	F	I	F	A C
3	1	1	1	0	1	1	0	0	0	1	1	0	0	4	- 10	2	1	8	N	F	N	B C Z
4	1	0	1	1	1	1	0	1	1	1	1	- 1	2	4	0	2	2	9	N	F	N	A B X Y Z
5	1	1	0	1	1	1	0	0	1	1	0	0	3	0	- 5	2	1	8	F	F	F	A B C Y
6	1	1	1	0	0	1	1	1	1	1	0	5	0	- 2	2	2	2	9	N	F	N	B C X Z
7	1	1	1	1	1	1	0	0	1	1	0	0	3	0	- 5	2	1	8	F	F	F	B C Y
Rº		*1*			*1*				*1*			*5*	*10*	*3*	1	----------						

Al variar el conjunto de edits puede cambiar el CM:

DETERMINACION DEL CONJUNTO MÍNIMO (**CM**) DE CAMPOS A IMPUTAR

Se tienen los siguientes datos:

F = (1 2 5 7).

variable A : F/A = 2/3.

variable B : F/A = 2/5.

variable C : F/A = 3/5.

variable X : F/A = 1/3.

variable Y : F/A = 2/3.

variable Z : F/A = 1/4.

En base al IS se selecciona primero la variable C, que cubre los edits fallados 2,5 y 7. Para cubrir el edit fallado pendiente de cubrir (1), el sistema selecciona la variable X por ser su F/A (1/3) mayor que la F/A (1/4) de la variable Z.

Así pues, las variables seleccionadas por el IS son: CM = (C X).

IMPUTACION DE LA VARIABLE C

Se tienen los siguientes conjuntos de edits:

TE = (1 2 3 4 5 6 7).

NPF = (4 6) : su fallo es evitado por los valores fijos Aº = 2 y Bº= 2.

PF = (1 2 3 5 7).

El edit **e** pertenece al conjunto EFE (C) si se cumplen las siguientes condiciones (AND):

e $\in$ PF = (1 2 3 5 7)

e $\in$ AV(C) = (2 3 5 6 7)

e $\in$ NAV(X) = (2 3 5 7)

De aquí, se tiene EFE (C) = (2 3 5 7). La UNION en C de los edits 2, 3, 5 y 7 es el vector de bits:

2	0	1	1	0
3	0	0	1	1
5	0	1	1	0
7	0	0	1	1
$\cup$	0	1	1	1

El módulo de imputación decide C*= 1, valor que evita el fallo de los edits 2, 3, 5 y 7, y se pasa a imputar la variable X.

IMPUTACION DE LA VARIABLE X

Se tienen los siguientes conjuntos de edits:

PF = (1).

AV(X) = (1 4 6).

Por lo tanto, se tiene EFE (X) = (1).

X se puede imputar si se verifican las siguientes condiciones:

C1: $2X - Z \leq 0 \rightarrow 2X - 3 \leq 0 \rightarrow X \leq 3/2$.

Supuesto que el valor imputado fuese X*=1, se tiene finalmente un registro R* libre de errores.

R*	Aº=2	Bº=2	**C*=1**	**X*=1**	Yº=10	Zº =3

Referencias bibliográficas

(1) I. P. FELLEGI and D. HOLT, "A Systematic Approach to Automatic Edit and Imputation", Journal of the American Statistical Association. March 1976, Volume 71, Number 353, 17-35.

(2) DIA (versión 2): Descripción del Sistema, Instituto Nacional de Estadística (INE) Madrid, abril de 1994 (S.G.A. Metodología Informática).

PERFIL PROFESIONAL DE JOSE MANUEL GÓMEZ ALONSO:

- Estadístico del INE.
- Profesor Asociado de la Universidad Nacional de Educación a Distancia.

KEYWORDS

- Conjunto de Edits Explícitos (CEE).
- Conjunto de Edits Implícitos (CEI).
- Conjunto de Edits Implícitos, a nivel de registro (CEI-R°).
- Conjunto Completo de Edits, a nivel de registro (CCE -R°)
- Conjunto Mínimo de edits a imputar (CM).
- Variable vacía.
- Errores aleatorios.
- Errores sistemáticos.
- Reglas de imputación determinística (RID)
- Edit de Imputaciòn Determinística (EID)
- Edits Fallados Pendientes de Cubrir (FPC)
- Índice de Sospecha (IS)
- Edits en los cuales está Activa una Variable X (AV(X))
- Edits cuyo Fallo solamente puede ser Evitado por el Valor imputado a la Variable X EFE(X)
- Edits caracterizados por No ser Activos en ninguna de las variables pendientes de imputación NAV(α)
- Módulo Generador de Edits Implícitos , necesarios para corregir el registro Rº (MGEI)
- Sistema Generador de Edits Aritméticos.
- Cotas e Intervalos de Valores Admisibles.