ESTADISTICA ESPAÑOLA
Vol. 34, Núm. 129, 1992, págs. 151 a 171

Análisis de reglas de depuración de datos

ILDEFONSO VILLAN CRIADO

Instituto Nacional de Estadística

RESUMEN

Este trabajo revisa distintos análisis a los que se pueden someter las reglas de depuración de datos estadísticos. Se distingue entre los análisis que se pueden aplicar a las reglas de depuración de datos cuantitativos y los que se pueden aplicar a las reglas de depuración de datos cualitativos. Se presta especial atención a las reglas de conflicto, aunque también se hace referencia a las reglas de imputación determinística.

Palabras clave: Edit, regla determinística de imputación, análisis de consistencia.

Clasificación AMS: 62-04.

1. LAS REGLAS DE DEPURACION COMO PARTE DEL MODELO DE LA REALIDAD

En la realización de investigaciones estadísticas se utilizan dos tipos de representaciones de la realidad, una a nivel de datos individuales o microdatos y otra a nivel de datos agregados o macrodatos. En ambas representaciones los estadísticos utilizan, implícita o explícitamente, un modelo mediante el cual simplifican, para hacerla más manejable, la compleja realidad con la que se enfrentan. Para la construcción del modelo se apoyan en diferentes instrumentos que dependen de la metodología utilizada. Para los microdatos es estándar la utilización del modelo entidad-relación. En el nivel de los datos agregados no existe una herramienta estándar de modelización, si bien es de gran interés el Modelo Conceptual Estadístico de Battista y Batini (1988) que forma parte de una metodología que integra la modelización de los microdatos con los datos estadísticos agregados.

A pesar de que las metodologías actualmente existentes no las recogen explícitamente, las reglas de depuración forman parte del modelo, al que aportan un fuerte contenido semántico. En el modelo de microdatos, las reglas de depuración plasman ciertas condiciones que los expertos piensan que no se deben presentar en la población (conocidas como edits de rechazo) o condiciones que piensan que cumplen todos los objetos de la población (edits de aceptación). Con las reglas de depuración los estadísticos tratan de eliminar aquellas observaciones que, debido a errores producidos en cualquier momento, caen fuera del modelo que ellos tienen de la realidad. También se utilizan reglas de depuración en el modelo de macrodatos, p.e. el establecimiento de límites de aceptación para los valores de una serie temporal. Tradicionalmente ambos tipos de reglas de depuración se han utilizado de forma no integrada; en la actualidad, los procedimientos de macrodepuración vienen a tender un puente entre ambos.

Como otras partes del modelo, las reglas de depuración se pueden analizar para garantizar su coherencia y adaptación a la realidad. El análisis de reglas tiene dos funciones:

— Comprobar la consistencia lógica de las reglas. Este análisis se limita a comprobar si las reglas determinan una región de aceptación demasiado reducida, o incluso vacía. No se tienen en cuenta los datos actuales, únicamente el modelo que se tiene de la realidad.

— Comprobar si las reglas se adaptan a la realidad observada. Esta comprobación se realiza contrastando las reglas con los datos reales a depurar. Si una regla no se falla nunca, puede ser eliminada, si una regla se falla un número excesivo de veces puede ser señal, o bien de un error sistemático, o bien de que la regla no se adapta a la realidad. Si una regla ha detectado como erróneas situaciones que se comprueba que son correctas la regla no se adapta a la realidad.

En este documento nos vamos a limitar a considerar el análisis lógico de las reglas de depuración a nivel de microdatos y entre ellas unos tipos simplificados, que son los que las metodologías actuales emplean.

En el apartado 2 se presentan los tipos de reglas de depuración considerados, en el 3 las principales estrategias de análisis de reglas, centrándonos en los apartados 4, 5 y 6 en el análisis de los edits numéricos, categóricos y categóricos numéricos respectivamente. Se utiliza un enfoque muy informal presentando únicamente los posibles conflictos que se pueden encontrar. En los apartados 7 y 8 se da una breve referencia a una posible forma de implementación de los análisis de edits numéricos y categóricos, el apartado 9 está dedicado al análisis de las reglas de imputación determinística y el 10 al análisis conjunto de reglas de imputación determinística y edit. Finalmente el apartado 11 recoge las conclusiones fundamentales.

2. TIPOS DE REGLAS DE DEPURACION

Vamos a considerar una clasificación de las reglas de depuración que es relevante a efectos de su tratamiento por el análisis de reglas.

En una primera aproximación distinguimos dos tipos de reglas de depuración:

1. Los edits, que definen condiciones inaceptables (edits de rechazo), o condiciones que deben ser satisfechas por los datos para ser aceptados como válidos (edits de aceptación). Los edits no contienen ninguna acción correctiva, dejando éstas para tratamientos manuales o para módulos de localización de variables a imputar e imputación.

2. Las reglas de depuración determinística, generalmente de la forma IF (condición de error) THEN (acción correctiva). En ellas no sólo se determina la condición inaceptable, también se incorpora una «solución» para la misma. Tanto unas como otras se pueden aplicar a datos de una misma unidad o a datos de varias unidades. Nosotros nos limitaremos a considerar el caso «intra-registro» que es el único tratado por las metodologías actuales de análisis de reglas.

Los edits, por su parte, se pueden clasificar en tres tipos (1): i) numéricos, ii) categóricos, iii) condicionales numéricos.

i) edits numéricos

Los edits numéricos se pueden representar por medio de una igualdad o desigualdad lineal entre funciones de los valores de las variables. Los sistemas generales actuales (GEIS, SPEER) sólo consideran el caso de funciones lineales, obtenidas en ocasiones mediante sencillas transformaciones de reglas no lineales.

Ejemplos:

$x_1 \leq x_2 * x_3$ tomando logaritmos en ambos miembros de la expresión se transforma en $y_1 \leq y_2 + y_3$.

$x_1/x_2 \leq 3$ se transforma en $x_1 \leq 3 * x_2$

(1) GILES (1989) considera un cuarto tipo: los edits condicionales categóricos, pero es obvio que estos edits se pueden reducir a edits categóricos.

Sin embargo, transformaciones tan sencillas como las anteriores pueden dar lugar en ocasiones a problemas poco manejables. Consideremos, por ejemplo, el caso de coexistir las dos reglas siguientes:

$$x_1 \leq x_2 * x_3$$

y

$$x_1 + x_2 \leq x_3$$

en este caso, mediante la transformación logarítmica no podríamos sustituir las x_i por las y_i, deberían coexistir ambas y manejar su relación de alguna forma.

Los edits numéricos suelen expresar condiciones de aceptación. Para manipular edits numéricos lineales se suelen utilizar técnicas de programación lineal que introducen la restricción adicional de usar variables no negativas. Esta restricción se resuelve con la transformación, usual en programación lineal, de sustituir cada variable x por la diferencia de dos variables no negativas: $x = x_1 - x_2$, donde $x_1 = max\ \{x, 0\}$ y $x_2 = max\ \{-x, 0\}$.

El siguiente párrafo extraído de Giles, nos proporciona la notación necesaria para tratar los edits numéricos lineales a efectos de su análisis.

«Los edits numéricos se pueden expresar usando notación algebraica:

Sean

n	número de variables
m_1	número de edits con desigualdad
m_2	número de edits con igualdad
$m = m_1+m_2$	número total de edits
x_j, $j = 1, ..., n$	valor de la variable j
a_{ij}, $i = 1, ..., m, j = 1, ..., n$	coeficiente de xj en el edit i
b_i, $i = 1, ..., m$	constante del edit i
A	matriz de coeficientes a_{ij}
B	vector con constantes b_i
x	vector con los valores de los datos x_j

Con esta notación los edits se pueden expresar en forma matricial como:

$$A_1\ x \leq B_1$$
$$A_2\ x = B_2$$
$$x \geq 0$$

donde las matrices A y B se han particionado adecuadamente.»

Los edits establecen la región de casos aceptables que es un poliedro convexo.

ii) edits categóricos

Los edits categóricos son expresiones lógicas que relacionan mediante operadores AND y/o OR conjuntos de posibles respuestas para las variables del cuestionario. Aplicando repetidamente la propiedad distributiva, los edits categóricos se pueden transformar en expresiones en las que el único operador utilizado para relacionar conjuntos de valores de diferentes variables es el AND. Estos son los llamados edits en forma normal en la metodología de Fellegi y Holt (1976). Otras formulaciones se pueden reducir también a la forma normal.

Ejemplos:

[EDAD(<15) AND (E_CIVIL (casado alguna vez) OR RELA (CABEZA_FAMILIA)]

es equivalente a:

EDAD(<15) AND RELA (CABEZA_FAMILIA)
EDAD(<15) AND E_CIVIL (casado alguna vez)

EDAD1 < EDAD 2 equivale a:

EDAD1(0) EDAD2(>0)
EDAD1(1) EDAD2(>1)
...
EDAD1(99) EDAD2(>99)

iii) edits condicionales numéricos

Son de la forma:

IF «expresión lógica» THEN edit numérico

donde la expresión lógica determina el conjunto poblacional al que es aplicable el edit numérico.

Ejemplo:

IF sector = 301 THEN (Salario-total/Número-empleados) $\leq$ 500

3. PRINCIPALES ESTRATEGIAS PARA EL ANALISIS DE LAS REGLAS

Hay dos tipos principales de estrategias para el análisis de reglas:

— El análisis *lógico,* en el que utilizando algoritmos se comprueba la consistencia lógica del conjunto de reglas especificadas. Este tipo de análisis tiene el inconveniente de exigir complejos programas de ordenador para su realización, siendo generalmente costoso en tiempo de ejecución. Suele estar asociado a un tipo de reglas específico. Su principal ventaja es la facilidad que tiene para los expertos estadísticos, al realizarse de manera automática sin exigir complejas especificaciones o revisiones.

— El *exhaustivo,* en el cual los expertos estadísticos especifican un conjunto de datos de prueba en el que recogen todas las situaciones que consideran aceptables (o inaceptables). Este juego de prueba se enfrenta con el conjunto de reglas para comprobar si el comportamiento de éstas es aceptable. Este es un procedimiento de comprobación que en muchos casos puede ser suficientemente satisfactorio, pues no requiere programación adicional, aunque puede ser muy enojoso para los expertos estadísticos. Cuando el número de combinaciones posibles es muy alto, es muy difícil realizar y analizar pruebas exhaustivas.

En este documento nos vamos a centrar en el análisis lógico de reglas.

4. ANALISIS DE EDITS NUMERICOS

El análisis de edits numéricos más completo implementado en la actualidad es el del Sistema GEIS, desarrollado por Statistics Canada, descrito por Giles (1989) y basado en Sande (1976).

Utilizaremos la notación introducida en el apartado 2, recordando que en este caso los edits son reglas de aceptación. En GEIS se proponen los siguientes chequeos de los edits:

4.1. Análisis de consistencia

Tiene como objetivo el determinar si la región de aceptación establecida por los edits es vacía, es decir si no hay ningún registro que satisface todos los edits. Si la región de aceptación es vacía los expertos tendrán que determinar cuál o cuáles son los edits responsables.

4.2. Eliminación de redundancias

Un edit es redundante si no puede ser fallado a menos que otro edit sea fallado, es decir, no aporta nada a la delimitación de la región de aceptación.

Ejemplo:

El edit $x_1 \leq 4$ es redundante respecto a

$x_1 \geq 0,\ x_2 \geq 0,\ x_1 + x_2 \leq 3$

4.3. Análisis de determinancias

Los edits pueden hacer que la región aceptable se reduzca a un único punto, o que para alguna variable haya un único valor aceptable. Son situaciones que, si bien no son necesariamente señal de que los edits sean inconsistentes, sí son sospechosas y requieren una revisión por los expertos.

4.4. Límites inferior y superior de cada variable

Los edits determinan un límite inferior y un límite superior para cada variable (que en algún caso puede ser infinito). Estos límites deben ser comprobados por los expertos para verificar si los intervalos de aceptación son demasiado pequeños o grandes. Esto puede dar lugar a añadir, eliminar, o modificar edits.

4.5. Igualdades escondidas

Puede ocurrir que un conjunto de edits, especificados como desigualdades, den conjuntamente lugar a una restricción de tipo igualdad. Es un caso en el que la región de aceptación se reduce en una dimensión, siendo señal de restricciones demasiado fuertes.

Ejemplo:

Los edits $2\ x_1 + 3\ x_2 \leq 10$ y $2\ x_1 + 3\ x_2 \geq 10$ equivalen al edit: $2\ x_1 + 3\ x_2 = 10$

4.6. Edits implicados

A partir de los edits especificados explícitamente por los expertos, se pueden obtener otros edits, conocidos como edits implícitos, que son condiciones que se pueden deducir lógicamente a partir de los edits especificados por los expertos.

Los edits implícitos fueron introducidos por Fellegi y Holt (1976) con el objeto de permitir determinar las variables a ser imputadas y los posibles valores a imputar, para ellos su utilización en el análisis de reglas es secundario. Si un edit implicado es inaceptable es señal de que al menos uno de los edits originales es inaceptable.

Los edits implicados se obtienen formando las combinaciones líneales positivas de los edits explícitos en las que se elimina el coeficiente de alguna variable.

Ejemplo:

de $3\ x_1 - 2\ x_2 \leq 4$ y
$x_1 + 5\ x_2 \leq 1$ se obtiene
$17\ x_1 \leq 22$

4.7. Puntos extremos

Los edits originales determinan una región de aceptación que es un poliedro convexo. Los puntos extremos de dicho poliedro son muy interesantes a efectos de análisis, puesto que cualquier punto interior, y por lo tanto aceptable, se puede obtener por medio de una combinación líneal convexa de los puntos extremos. Los edits implicados se presentan a los expertos para que puedan estudiarlos y decidir si son aceptables o no.

En el anexo 1 se presenta un ejemplo de los distintos análisis para un caso de dos variables, lo que permite representar gráficamente en dos dimensiones los edits para facilitar la comprensión.

5. ANALISIS DE EDITS CATEGORICOS

El análisis de los edits categóricos aquí descrito está basado en la metodología de Fellegi y Holt (1976) y en su implementación en el Sistema de detección e imputación automática DIA, desarrollado en el Instituto Nacional de Estadística, ver García Rubio y Villan (1988 y 1990). Revisaremos de forma rápida y resumida las definiciones y resultados de la metodología de Fellegi y Holt relevantes para el análisis.

Introduciremos la siguiente notación:

Sean las variables $V_1, \ldots, V_n$

Sea $E_i = 1, \ldots, n_i$ $(i = 1, \ldots, n)$ el conjunto de valores posibles de V_i

Un edit, en forma normal, se puede representar como:

$$e_i: E_{i1} \cap E_{i2} \cap \ldots \cap E_{in}$$

donde cada E_{ij} es un subconjunto, propio o no, de Ei. Estos edits indican condiciones de error.

Si E_{ij} es un subconjunto propio de E_i, decimos que la variable V_i esta activa en el edit e_i.

El conjunto de edits en forma normal especificados por los expertos lo llamamos *edits explícitos.*

Lema:

«Sean e_i $i \in S$ un conjunto de edits,

la expresión:

$$e_G : E_{G1} \cap \ldots \cap E_{Gn} \text{ donde:}$$

$$E_{Gj} = \bigcap_{i \in S} E_{ij} \qquad j \neq k$$

$$E_{Gk} = \bigcup_{i \in S} E_{ik} = E_k$$

y ninguno de los E_{Gj} $j \neq k$ es vacío, es un edit implicado.»

Además se puede probar que todos los edits implicados se pueden obtener con este procedimiento de generación.

Si todos los E_{ik} son subconjuntos propios de E_k, entonces decimos que el edit implicado e_G es un *edit esencialmente nuevo,* el campo k se llama *campo generador.*

Ejemplo:

De los edits e_1: EDAD(<15) ECIVIL (≠ SOLTERO) y

e_2: ECIVIL (NO CASADO ACTUALMENTE) RELACION_CABEZA (ESPOSA)

se deduce

$e_{1\text{-}2}$: EDAD(<15) RELACION_CABEZA(ESPOSA)

(e_1 equivale a EDAD(<15) IMPLICA ECIVIL (SOLTERO), e_2 equivale a ECIVIL (NO CASADO ACTUALMENTE) IMPLICA RELACION_CABEZA (≠ESPOSA).

De ambos se deduce:

EDAD (<15) IMPLICA RELACION_CABEZA (≠ESPOSA)

que equivale a $e_{1\text{-}2}$).

El conjunto de los edits explícitos, junto con los edits esencialmente nuevos forman el *conjunto completo de reglas.*

5.1. Análisis de consistencia

Decimos que un conjunto de edits es inconsistente si conjuntamente implican que haya valores permisibles de un campo que causen automáticamente, independientemente de los valores de los demás campos, el fallo de algún edit. Esto se traduce en que un edit implicado tendrá sólo un campo activo.

Para la detección de inconsistencias, en el sentido de Fellegi y Holt, es necesario generar el conjunto completo de edits, y verificar si alguna de las reglas generales está únicamente activa en un campo.

5.2. Eliminación de redundancias

Un edit e_1 es redundante respecto a otro e_2, si todas las situaciones detectadas por e_1 lo son también por e_2.

La eliminación de redundancias es conveniente a efectos de reducir el número de edits a considerar y por tanto el tiempo de proceso.

Ejemplo:

e_1 : EDAD(<15) ECIVIL(CASADO) es redundante respecto a:
e_2 : EDAD(<15) ECIVIL (≠SOLTERO)

Otra forma de redundancia es cuando dos edits son iguales en todos los campos menos en uno, en este caso se pueden fusionar en un único edit.

Ejemplo:

e_1 : EDAD(<15) ECIVIL(CASADO) y
e_2 : EDAD(<15) ECIVIL(VIUDO)

se pueden fusionar en

e_{12} : EDAD(<15) ECIVIL(CASADO O VIUDO)

Un edit resultado de una fusión puede dar lugar a nuevas fusiones o dominar a algún edit. Por lo tanto, si se realizan fusiones la eliminación de redundancias es un proceso iterativo.

5.3. Análisis de los edits implicados

En la metodología de Fellegi y Holt los edits implicados tienen una doble función, por una parte son necesarios para determinar para un registro detectado como erróneo qué variables hay que imputarle y qué posibles valores hay para cada variable a imputar. Por otra parte sirven para analizar los edits. Según vimos en 5.1, si durante su proceso de generación se produce un edit con un único campo activo se detecta una inconsistencia. Una vez que se dispone del conjunto de todos los edits implicados, los expertos los deben revisar, pues si observan un edit implicado demasiado restrictivo es señal de que alguno de los edits originales también lo es.

6.1. Análisis de edits condicionales numéricos

El análisis de los edits condicionales numéricos se puede realizar utilizando los procedimientos descritos para el caso de los edits numéricos. Para ello se realiza una estratificación de los edits, agrupando todos los edits numéricos que se correspondan con una misma condición. El análisis se realiza por separado para cada estrato. Por ejemplo, si se han especificado conjuntos de edits numéricos con una condición de aplicación que indica el sector de actividad al que son aplicables, realizaríamos el análisis para cada conjunto de edits aplicable a cada sector.

Este tipo de análisis tiene dos debilidades, la primera de índole práctica, pues su realización puede ser muy engorrosa para los expertos, sobre todo si el número de estratos es muy grande. La segunda es conceptual, pues descansa en la hipótesis implícita de que los valores de las variables que intervienen en la condición son correctos.

7. IMPLEMENTACION PRACTICA DEL ANALISIS DE EDITS NUMERICOS

La implementación práctica de los edits numéricos, en el caso de utilizar edits de tipo igualdades y/o desigualdades lineales:

$$\begin{aligned} A_1\, x &\leq B_1 \\ A_2\, x &= B_2 \\ x &\geq 0 \end{aligned} \qquad (1)$$

se puede realizar utilizando técnicas tradicionales de programación lineal.

7.1. Análisis de consistencia

En este caso basta buscar una solución, con cualquier función objetivo $c'x$ del problema de minimizar $c'x$ sujeto a las restricciones (1).

7.2. Eliminación de redundancias y detección de determinaciones para los edits de igualdad $A_2\, x = B_2$

Los edits de igualdad forman un sistema lineal de m_2 ecuaciones con n incógnitas, que puede ser resuelto para comprobar si tiene solución única (determinancia) o no. También se puede comprobar que edits son redundantes respecto a los demás (el número máximo de edits no redundantes es el mínimo entre m_2 y n).

7.3. Eliminación de redundancias para los edits de desigualdad $A_1\, x \leq B_1,\, x \geq 0$

Para cada edit $A_{1i}\, x$ (una fila de $A_1\, x \leq B_1$) se maximiza $A_{1i}\, x - b_i$ sujeto a $A_1\, x \leq B_1,\, x \geq 0$

Si el máximo es positivo estamos ante un edit redundante, pues los edits que limitan el conjunto aceptable alcanzan en la frontera el valor 0.

7.4. Obtención de límites para las variables y detección de determinaciones

Los límites inferior y superior que los edits determinan para cada variable x_j se obtienen resolviendo los problemas:

min x_j sujeto al conjunto de restricciones (1), lo que nos da el límite inferior x_{j1}, y

max x_j sujeto a (1), lo que nos da el extremo superior x_{j2}.

Si x_{j1} es igual a x_{j2} se produce una determinancia para la variable x_j.

7.5. Obtención de los edits implicados

Dados dos edits

$$a_{i1}\, x_1 + \ldots\ldots + a_{in}\, x_n \leq b_i$$

$$a_{k1}\, x_1 + \ldots\ldots + a_{kn}\, x_n \leq b_k$$

Si $a_{kj} > 0$ y $a_{ij} < 0$, multiplicando el primer edit por a_{kj} y el segundo por $-a_{ij}$, y sumando los resultados se obtiene un edit en el que la variable x_j no aparece explícitamente: un edit implicado por e_i y e_k generado en el campo j. Con el siguiente algoritmo se podrían obtener todos los edits implicados:

```
E = el conjunto inicial de edits.
DO mientras haya parejas no chequeadas en E
  Para la pareja no chequeada e_i y e_k
    DO l = 1 to n
    comprobar si e_i y e_k generan en el campo l
    Si generan un edit e_ik añadirlo a E
    END
END
```

Un algoritmo conceptualmente equivalente a éste está implementado en SPEER, Greenberg (1982). GEIS utiliza uno completamente distinto basado en el algoritmo de Chernikova, Schiopu-Katrina y Kovar (1989).

7.6. Obtención de los puntos extremos

Los puntos extremos del poliedro convexo definido por (1) se obtienen resolviendo todos los sistemas de n ecuaciones que se pueden formar con las $m+n$ reglas. Esto puede dar lugar a un número formidable de combinaciones a verificar $(m+n)\ !/n!m!$. En la práctica (GEIS) se utilizan algoritmos más eficientes, Schiopu-Kratina y Kovar (1989).

8. IMPLEMENTACION PRACTICA DEL ANALISIS DE EDITS CATEGORICOS

La implementación de la metodología de Fellegi y Holt se suele hacer representando los edits como tiras de bits. A cada código posible de cada variable se le asigna un bit, que en un edit dado tendrá un 1 si el código es relevante para el fallo del edit o un 0 en caso contrario.

Ejemplo:

Supongamos tres variables *A*, *B* y *C*, con códigos válidos 1, 2 y 3 cada una de ellas.

El edit

e: *A*(1) *C*(2, 3) que expresa una incompatibilidad existente entre valores de las variables *A* y *C*, independientemente del valor de la variable *B*, se representaría:

A	*B*	*C*
100	111	011

8 1. Análisis de redundancias

Con esta representación la eliminación de redundancias se puede hacer de una forma sencilla. Para comprobar si dos edits son redundantes se calcula el producto escalar de sus vectores de bits. Si el producto coincide con la cardinalidad (número de 1*s*) de uno de los vectores de los edit, dicho edit es redundante.

Ejemplo:

e_1: *A*(1,2) *B*(1)

cuyo vector es: 110 100 111 de cardinalidad 6

e_2: *A*(2) *B*(1)

con vector: 010 100 111 de cardinalidad 5

El producto escalar da 5, por lo tanto e_2 es redundante.

8.2. Análisis de consistencia

El proceso de generación en un campo se puede realizar haciendo uniones lógicas en el campo generador e intersecciones lógicas en todos los demás.

Ejemplo:

e_1: *A*(1,2) *B*(1,2)

e_2: *B*(2,3) *C*(3)

con vectores:

e_1: 110 110 111 y

e_2: 111 011 001

Haciendo unión lógica en el segundo campo e intersección en los otros dos obtenemos:

e_{12}: 110 111 001

que equivale a: e_{12}: $A(1,2)C(3)$

Mediante el siguiente algoritmo se pueden obtener todos los edits implicados:

E = conjunto de edits originales
DO mientras quede alguna combinación de edits no analizada
  DO *I* = 1 to *n*
  I es el campo generador
  formar todas las combinaciones, no chequeadas hasta ahora, de edits de *E* activos en *I*
  DO para cada combinación
  aplicarle el procedimiento de generación usando *I* como campo generador.
  Si genera un edit esencialmente nuevo añadirlo a *E*
  END
  END
END

Con este algoritmo se pueden obtener todos los edits implicados. Para cada edit implicado se comprueba si está activo en una única variable (inconsistencia) o no.

Para conseguir que este algoritmo sea eficiente, es necesario modificarlo profundamente mediante la introducción de varios filtros y estrategias. En particular un teorema de truncación, desarrollado por la Oficina Central de Estadística de Hungría, permite acortar el tiempo de proceso al no requerirse la obtención del conjunto completo de reglas para la detección de inconsistencias.

9. ANALISIS DE REGLAS DE IMPUTACION DETERMINISTICA

Considerando las reglas de imputación determinística como reglas del tipo:

r : IF (condición) THEN (imputación).

se pueden realizar los siguientes análisis:

9.1. Eliminación de redundancias

Dadas dos reglas r_1 y r_2, tales que dan lugar a una misma imputación, entonces

a) Si la condición de r_1 domina a la de r_2 entonces r_2 es redundante y se puede eliminar.

Ejemplo:

r_1: IF (SALARIO < 50.000 & SECTOR = 027) THEN (SALARIO = = 60.000)

r_2: IF (SALARIO < 45.000 & SECTOR = 027) THEN (SALARIO = = 60.000)

r_2 se puede eliminar.

b) Si la condición de r_1 es igual a la de r_2 en todas las variables menos en una, entonces r_1 y r_2 se pueden fusionar en una única regla.

Ejemplo:

r_1: IF (SALARIO < 50.000 & SECTOR = 027) THEN (SALARIO = = 60.000)

r_2: IF (SALARIO < 50.000 & SECTOR = 032) THEN (SALARIO = = 60.000)

se pueden fusionar en la regla:

r_{12}: IF [SALARIO < 50.000 & (SECTOR = 027 | SECTOR = 032)] THEN (SALARIO = 60.000)

Si en el análisis de redundancias se realizan fusiones de reglas es necesario realizar una nueva iteración del proceso búsqueda y eliminación de redundancias. Una regla resultante de una fusión puede dominar o fusionarse con otras reglas.

9.2. Eliminación de inconsistencias

Entre las posibles inconsistencias que se pueden presentar en las reglas de imputación determinística están:

a) Dos reglas ofrecen distintas situaciones a una misma situación conflictiva.

Ejemplo:

r_1: IF (SALARIO < 35.000 & SECTOR = 011) THEN (SALARIO = = 40.000)

r_2: IF (SALARIO < 25.000 & SECTOR ≠ 022) THEN (SALARIO = = 55.000)

en este caso ambas reglas están proponiendo para los registros con SALARIO (<25.000) y SECTOR (011) imputaciones distintas.

b) La imputación de un registro por una regla de imputación determinística provoca el fallo de otra regla de imputación determinística que antes no se fallaba.

Ejemplo:

Dadas las reglas:

r_1: IF (SEXO = mujer & SECTOR = 012) THEN (SEXO = varón)
r_2: IF (SEXO = varón & SALARIO <25.000) THEN (SALARIO = 35.000)

un registro con SEXO = mujer, SECTOR = 012 y SALARIO = 20.000 falla r_1 y no r_2, pero tras ser imputado por r_1 pasa a fallar r_2.

Este tipo de situaciones, no son necesariamente conflictivas, pero sí sospechosas.

10. ANALISIS CONJUNTO DE EDITS Y REGLAS DE IMPUTACION DETERMINISTICA

No es frecuente la realización de un análisis conjunto de reglas de imputación determinística y edits. Las razones son, por una parte la dificultad de tal análisis, y por otra el manejo de hipótesis, en muchos casos implícitas, acerca del procedimiento de depuración (la hipótesis más general es que las reglas de imputación determinística son muy «seguras», apoyándose en casos perfectamente determinados, y con soluciones claras. Domina por tanto la imputación determinística sobre la depuración apoyada en los edits).

Un análisis de este tipo depende mucho de la estrategia global de depuración, y de si se va a realizar una imputación apoyándose en los edits (siguiendo la metodología de Fellegi y Holt por ejemplo).

Por este motivo únicamente se apuntarán, apoyándonos en ejemplos, algunas de las posibles inconsistencias que se pueden producir (2).

---

(2) El Sistema DIA tiene un analizador de reglas que incluye un analizador edit-reglas de imputación determinística bastante completo, no limitándose únicamente a detectar situaciones conflictivas, sino que ofrece para aquellas menos graves una solución estándar. Estas soluciones están integradas en la propia metodologia del sistema y son sólo aplicables si se utiliza para realizar imputación determinística e imputación probabilística basada en la metodología de FELLEGI y HOLT.

10.1. Una regla de imputación determinística produce el fallo de un edit

Ejemplo:

r_1: IF (SALARIO < 50.000 & SECTOR = 012) THEN (SALARIO = = 60.000)

e_1: (SALARIO > 55.000) & NIVEL-EDUCATIVO (analfabeto)

en este caso, para un registro con SALARIO (<50.000), SECTOR (012) y NIVEL-EDUCATIVO (analfabeto)

la regla de imputación determinística forzaría el fallo del edit.

10.2. Dos reglas de imputación determinística producen el fallo de un edit

Ejemplo:

Ilustraremos la situación con un ejemplo poco realista.

Sean las reglas determinísticas:

r_1: IF (SALARIO > 100.000 & SECTOR = 012) THEN (SALARIO = = 60.000)

r_2: IF (SEXO = mujer & SECTOR = 012) THEN (SEXO = varón)

y el edit

e: SEXO (varón) SALARIO (<100.000)

Un registro con

SALARIO (120.000) SECTOR (012) SEXO (mujer)

que no fallaba el edit e, tras ser imputado por ambas reglas determinísticas pasa a fallarlo.

10.3. Una regla de imputación determinística se apoya en una variable que es sospechosa de acuerdo con un edit

Ejemplo:

r_1: IF (SALARIO < 50.000 & SECTOR = 012) THEN (SALARIO = = 60.000)

en esta regla, se considera implícitamente que el sector es «correcto», modificándose en su función el SALARIO, pero SECTOR puede ser sospechoso si existe un edit fallado como:

e_1: SECTOR (012) & SEXO (MUJER)

11. CONCLUSIONES

El análisis de reglas descrito en los apartados anteriores tiene la limitación de restringirse a aspectos parciales de la depuración (al no considerar la depuración interregistros), y ser únicamente aplicable a unos tipos limitados de reglas; que por otra parte son los más frecuentes. Además, no considera el tratamiento conjunto de reglas de distintos tipos.

Otra limitación viene dada por el tiempo de proceso necesario para su realización, y por la necesidad de disponer de software adecuado, software que por otra parte existe únicamente como parte de paquetes generales.

Su principal ventaja está en permitir detectar problemas, antes de enfrentarnos con los datos a depurar, sin necesidad de tener que prever complejos juegos de pruebas que cubran todas las posibilidades. Permite detectar en etapas tempranas problemas que más tarde sería difícil diagnosticar y corregir. La eliminación de redundancias permite obtener mejores tiempos de proceso, al hacerse este con menos reglas.

El análisis de reglas permite mejorar la calidad de las especificaciones de los expertos al analizarlas automáticamente.

En cualquier caso, se utilice un analizador lógico de reglas o no, los expertos no están libres de la labor de comprobación de la adaptación de sus especificaciones a la realidad. Los expertos deben comprobar que ninguna regla detecta como erróneos un número excesivo de registros. Con las reglas de imputación determinística el control ha de ser aún mayor, pues pueden introducir importantes sesgos.

ANEXO 1

Consideremos el siguiente conjunto de edits numéricos:

$$\begin{aligned} x_1 &\geq 0 \\ x_2 &\geq 0 \\ x_1 + x_2 &\leq 4 \\ x_1 &\leq 3 \end{aligned} \qquad (1)$$

que definen la región de aceptación que aparece rayada en la figura 1.

Si se añade el edit:

$$2\,x_1 + x_2 \leq 8$$

se observa que la región de aceptación no varía, dicho edit es redundante.

Si se añade el edit:

$$x_1 \geq 5$$

la región de aceptación se convierte en vacía, el conjunto de edits es inconsistente.

Si se añade el edit:

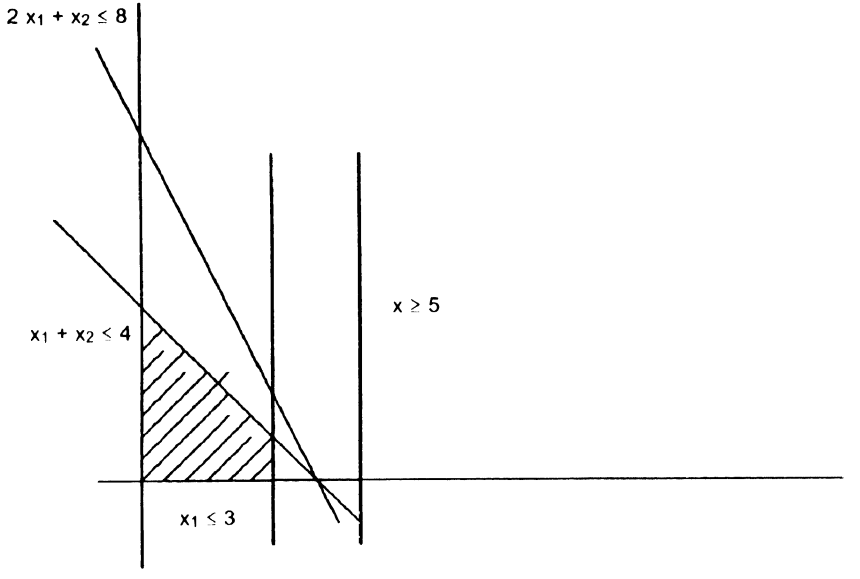
$$x_1 + x_1 \geq 4$$

la región de aceptación resultante se reduce al segmento de la recta $x_1 + x_2 = 4$, comprendido entre los puntos [(0,4) y (3,1)]. Existe una igualdad escondida entre los edits.

Los edits del conjunto (1) establecen los siguientes límites para las dos variables:

$$0 \leq x_1 \leq 3$$
$$0 \leq x_2 \leq 4$$

Figura 1.



REFERENCIAS

Data Editing Joint Group (1989). Data Editing System Guidelines for Concepts and Specification. *Work Session on Statistical Data Editing. Ginebra.*

FELLEGI, I. P. y HOLT, D. (1976). A systematic approach to automatic editing and imputation. *J. Amer. Statist. Assoc.*, 71, 17-35.

GARCÍA-RUBIO, E. y VILLAN, I. (1988). Sistema DIA: Descripción del sistema. *Instituto Nacional de Estadística de España.*

GARCÍA-RUBIO, E. y VILLAN, I. (1990). DIA SYSTEM: software for the automatic editing and imputation of qualitative data. *U.S. 6th Annual Research Conference Proceedings.*

GILES, P. (1989). Analysis of edits in a Generalized edit and imputation system. *Statistics Canada Working Paper SSDM-89-004-E.*

GREENBERG, B. (1982). Using an edit system to develop editing specifications. *Proceedings of the Section on Survey Research Methods, ASA.*

SANDE, G. (1976). Diagnostic capabilities for a numerical edit specifications analyzer. *Statistics Canada Technical report, BSMD.*

SCHIOPU-KRATINA, I. y KOVAR, J. G. (1989). Use of Chernikova's algorithm in the Generalized Edit and Imputation System. *Statistics Canada Working Paper BSMD-89-001E.*

SANTA, J. (1991). On the Fellegi and Holt rule analysis. *UNDP/SCP2/WP.81.*

SANTA, J. (1991). Rule Analyzer. *UNDP/SCP2/WP.83.*

SUMMARY

ANALYSIS OF STATISTICAL DATA EDITING RULES

In this paper different kind of statistical data editing rules analysis are revised. We distinguist between the analysis which could be applied to editing rules for numerical data and the analysis that could be applied to editing rules for categorical data. Our main focus are the conflict rules, edits, but deterministic imputation rules are also considered.

Key words: Edit, deterministic imputation rule, consistency analysis.

AMS Classification: 62-04.