tecnim@p Sevilla 2006
30 mayo a 2 junio
060.es
24x7x365

Comunicación

039

TEIDE: UNA NUEVA HERRAMIENTA SEMI-AUTOMÁTICA PARA DEPURAR ENCUESTAS

Fayna Álamo Santana
Jefa de Servicio de Estadísticas Demográficas y Sociales
ISTAC

Sergio Delgado Quintero
Ingeniero Informático
Universidad de La Laguna

José Molina González
Jefe de Proyectos del Servicio de Informática Estadística y Banco de Datos
ISTAC

Juan José Salazar González
Profesor Titular
Universidad de La Laguna

Fayna Álamo Santana
Sergio Delgado Quintero
José Molina González
Juan José Salazar González

Palabras clave

Software, Edición e Imputación.

Resumen de su Comunicación

TEIDE es una herramienta software que permite la depuración semi-automática de encuestas, partiendo de un conjunto de reglas definidas. La aplicación se encarga de resolver el problema de edición e imputación de datos. TEIDE es el acrónimo de "Técnicas de Edición e Imputación de Datos Estadísticos". En las diferentes fases que tiene una encuesta, los datos sufren alteraciones que pueden introducir incoherencias. La detección y corrección de estas incoherencias es fundamental para obtener una alta calidad en la información, y por consiguiente, en las conclusiones inferidas sobre ella. Por ello, los procesos de depuración de información son muy importantes, e históricamente han consumido una gran cantidad de recursos humanos y económicos. TEIDE ocupa este lugar dentro del tratamiento de una encuesta y se convierte en un sistema de ayuda al técnico estadístico, proporcionándole métodos automáticos para la depuración de datos, además de ofrecerle elementos de visualización y gestión de errores. TEIDE es capaz de trabajar con una gran cantidad de datos y con variables de tipo cualitativo y cuantitativo. Además, las reglas de coherencia que se introducen en el sistema pueden ser bastante complejas como para cubrir casi cualquier situación. Esto hace que la herramienta sea muy versátil y general, pudiendo ser usada en cualquier encuesta, siempre y cuando se respeten los formatos en los datos que la aplicación necesita.

El problema de edición e imputación de datos estadísticos es un problema difícil dentro del campo de los problemas de optimización, y TEIDE es capaz de proporcionar soluciones satisfactorias con un consumo de recursos razonable. Sus algoritmos se basan en los fundamentos dados en Fellegi y Holt (ver [1]), y hacen uso de modernas técnicas de Optimización Matemática para determinar qué cambios hacer en los datos para buscar un resultado coherente con el menor impacto. Es una herramienta "semi-automática" porque aunque por defecto realiza correcciones para garantizar datos coherentes, sólo personal con experiencia en encuestas anteriores será capaz de ajustar apropiadamente los parámetros internos de manera para definir apropiadamente los criterios de optimalidad. En otras palabras, TEIDE no puede ni pretende eliminar personal de los institutos de estadística, sino favorecer el que utilicen su tiempo en tareas más creativas e interesantes. TEIDE realizará los cálculos masivos a partir de escenarios descritos por expertos, y en este sentido TEIDE puede ser también utilizado como un "simulador" para estudiar estadísticamente los datos de una encuesta antes, durante y posteriormente a la salida a campo.

TEIDE ya ha sido y está siendo usado para depurar encuestas reales en el ISTAC (Instituto Canario de Estadística) proporcionando resultados satisfactorios tanto en la visualización y gestión de errores como en los procedimientos automáticos de depuración. El objetivo de este trabajo es mostrar algunas de las ventajas que se obtiene con su uso en la esperanza de que también otros institutos manifiesten su interés y puedan beneficiarse de nuestro trabajo. Recíprocamente, también esperamos enriquecer TEIDE con las experiencias y críticas de expertos.

Fayna Álamo Santana
Sergio Delgado Quintero
José Molina González
Juan José Salazar González

TEIDE: UNA NUEVA HERRAMIENTA SEMI-AUTOMÁTICA PARA DEPURAR ENCUESTAS

1. Introducción

Los institutos de estadística se encargan de proporcionar información que ayude a la toma de decisiones de una región. Esta información es inferida a partir de un conjunto de datos recogidos que constituyen una encuesta. El trabajo de campo suele introducir errores en la información tomada, bien por equivocaciones del informante, bien por equivocaciones del propio encuestador o bien por otros motivos. Una vez que esta información está recogida, es necesario pasarla a un formato digital, y en esta etapa pueden también aparecer errores de grabación. Es decir, que desde que "apuntamos" la respuesta del informante hasta que ese dato está disponible para realizar un tratamiento estadístico, es posible que haya sufrido alteraciones con respecto al original. Ahora bien, para realizar inferencias correctas sobre los datos debemos partir de información coherente. El proceso de depuración de datos recibe el nombre de edición e imputación. TEIDE es una herramienta que automatiza estas tareas para servir de apoyo al técnico estadístico en el proceso de corrección de datos.

2. Características de TEIDE

TEIDE es el acrónimo de "Técnicas de Edición e Imputación de Datos Estadísticos". Se trata de un software especialmente orientado a las oficinas de estadística y a los organismos que desarrollan encuestas. Su principal cometido es realizar una depuración automática de los datos, pero además proporciona funcionalidades auxiliares que se verán más adelante.

La aplicación puede trabajar con un gran número de registros y de variables, así como con campos discretos o continuos. Una vez que la información está cargada en el sistema, ésta es susceptible de ser visualizada, evaluada y posteriormente corregida. Estamos hablando de una aplicación de carácter general que puede aplicarse a cualquier encuesta, siempre y cuando los datos de entrada sigan el formato que la herramienta requiere.

Es importante señalar que esta herramienta posee un marcado enfoque visual, con una potente interfaz gráfica que permite al usuario navegar por los datos de manera cómoda y tener una visión global de los mismos a "golpe de ratón".

TEIDE es un software creado para trabajar sobre plataformas Windows donde los datos están una base de datos Access. Está implementado en C++ haciendo uso de librerías auxiliares como flex y bison.

En este trabajo, aprovechando la descripción de las distintas pantallas que tiene la aplicación, se presentan las funcionalidades que proporciona TEIDE, a la vez que se comentan algunos de los métodos matemáticos internos de cómputo que utiliza.

3. Interfaz de Usuario

En este apartado se presenta la interfaz de usuario de TEIDE junto con el modo de uso habitual que se suele seguir para un procesamiento correcto de los datos.

3.1 Arranque del programa

El programa es ofrecido en un único fichero ejecutable (.exe) sin otro tipo de dependencias adicionales. Así pues, para su puesta en funcionamiento sólo debemos abrir este fichero.

Una vez que se abre el programa aparece una ventana splash con el nombre de la aplicación y el nombre de los autores (ver Figura 1).

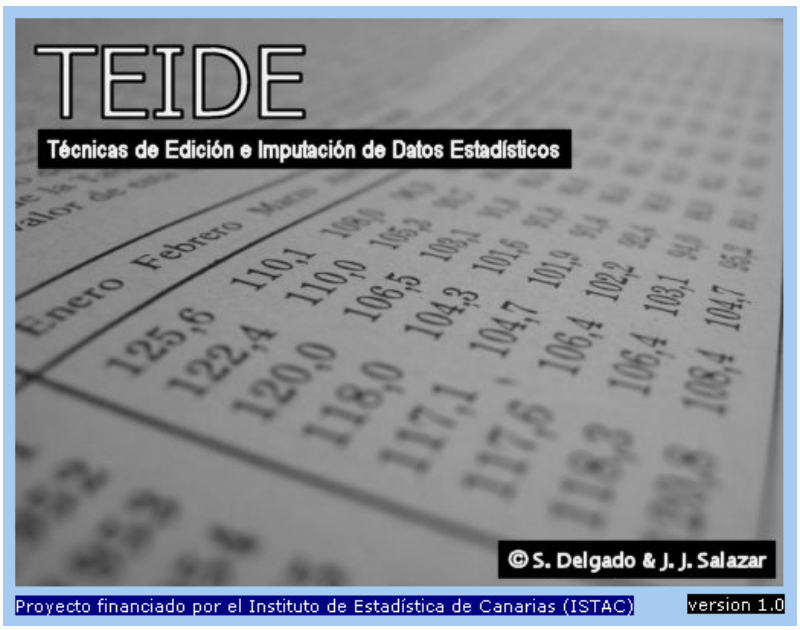

Figura 1

Cuando se cierra esta primera pantalla de bienvenida se abre la pantalla inicial del programa con dos menús principales: Principal y Ayuda. En el primero encontramos las opciones básicas de la aplicación y en el segundo un enlace a la página de ayuda.

3.2. Carga de datos

El siguiente paso que debemos realizar en la aplicación es cargar los datos que queremos depurar. A continuación explicamos qué elementos necesita TEIDE para poder trabajar y cuáles son los formatos asociados a cada elemento.

Como ya se ha comentado en la Sección 2, la aplicación lee, procesa y escribe en una base de datos Microsoft Access. La base de datos debe contener 3 tablas principales, además de otras auxiliares, para que TEIDE sea capaz de funcionar adecuadamente. La primera tabla debe contener los datos en bruto de la encuesta. La segunda tabla debe contener la información asociada a cada una de las variables de la encuesta. Y la tercera tabla debe contener las reglas que determinan la coherencia dentro de los datos (en este contexto las reglas se denominan edits). Además de esto, pueden existir tablas auxiliares que hacen corresponder valores discretos a cadenas de texto, información que aprovecha la herramienta para su posterior visualización.

Fayna Álamo Santana
Sergio Delgado Quintero
José Molina González
Juan José Salazar González

La tabla de datos (o tabla de microdatos) contiene tantas columnas como campos y tantas filas como registros. Un campo generalmente está asociado a una pregunta y un registro generalmente está asociado a un informante.

La tabla de variables contiene tantas columnas como atributos y tantas filas como variables. Los atributos sirven para definir perfectamente las características de una variable. Actualmente TEIDE trabaja con los siguientes atributos:

- Nombre: nombre corto de la variable.
- Descripción: explicación más extensa de la variable.
- Rango: posibles valores de la variable.
- Filtro: condición necesaria para que la variable pueda ser respondida.
- Descripción del Filtro: explicación más extensa del filtro de la variable.
- Imputable: indica si los valores de la variable pueden ser modificados.
- Tipo: las variables pueden ser discretas o continuas.
- Admisión de valores missing: hace referencia a la posibilidad de que los valores de la variable puedan ser "no procede", "no sabe", "no contesta" o "no sabe/no contesta".
- Peso: importancia de la variable dentro de la encuesta.
- Mapping: referencia a la tabla que hace corresponder valores y cadenas.
- Imputación Numérica: para aquellas variables de tipo continuo, se pueden especificar métodos de imputación alternativos.

La tabla de edits contendrá tantas filas como reglas se quieran definir en el sistema. Las reglas utilizadas se suelen escribir con el siguiente formato:

IF (condición1) THEN (condición2),

lo que significa que si se cumple condición1 entonces debe cumplirse también condición2. El típico ejemplo que se suele utilizar para este tipo de reglas es:

IF (estado_civil=casado) THEN (edad>18).

Dado que estas tres tablas pueden tener cualquier nombre y dado que se puede utilizar cualquier base de datos, necesitamos un fichero que determine dónde ir a buscar cada información y cuáles son los nombres asociados. Se trata del llamado metafile, y consiste en un fichero de texto normal formateado con ciertos tags (al estilo html). Cada línea especifica el valor de un campo. Los campos necesarios son los siguientes:

- Nombre del conjunto de datos.
- Ruta a la base de datos Access donde está la información (fichero .mdb).
- Nombre de la tabla de variables.
- Nombre de la tabla de microdatos.
- Nombre de la tabla de edits.
- Campo clave para cada registro de microdatos.

Todos estos elementos (campos de datos, tipos, reglas, etc.) han debido ser generados a partir de la encuesta inicial mediante alguna herramienta informática que los coloque en la base de datos Microsoft-Access que hemos descrito y que TEIDE necesita para iniciar el proceso de carga y luego el de depuración.

Una vez explicados los elementos que contienen la información relevante para TEIDE, continuamos ahora con el proceso de carga de datos. Para ello, debemos seleccionar el "metafile" deseado. La aplicación en este momento se encarga de leer la información existente en la base de datos y de realizar una gran cantidad de validaciones sobre los mismos. Entre estas validaciones podemos citar varias: comprobar que

la base de datos y las tablas existen, comprobar si todas las variables que se usan en los microdatos están definidas en la tabla de variables, comprobar que no existan variables duplicadas, comprobar que el tipo y el rango de las variables sea coherente, comprobar que los edits tienen una estructura sintáctica y léxica correcta, comprobar la validez de las tablas de mapping, etc.

TEIDE aprovecha este momento de la carga de datos para crear los elementos visuales necesarios y para construir las estructuras de memoria que le harán falta. Por ello, el proceso de carga es computacionalmente intensivo, pero una vez acabado disponemos de los datos preparados para su proceso.

De la misma manera que la información de entrada se dividía en tres tablas, la visualización también se divide en tres pestañas, cada una asociada con microdatos, variables y edits. Una vez completado el proceso de carga de datos, el sistema nos muestra estas tres pestañas que describimos en las siguientes secciones.

3.3. Pestaña de variables

Esta pestaña es la que queda activa tras la carga de datos (ver Figura 2). En ella podemos encontrar una rejilla principal con la información referente a todas las variables y un navegador en la parte inferior para poder acceder de forma rápida a cualquier variable tanto por su nombre como por su índice. En la parte inferior de la pestaña se dispone de una barra de estado con información de la situación actual.

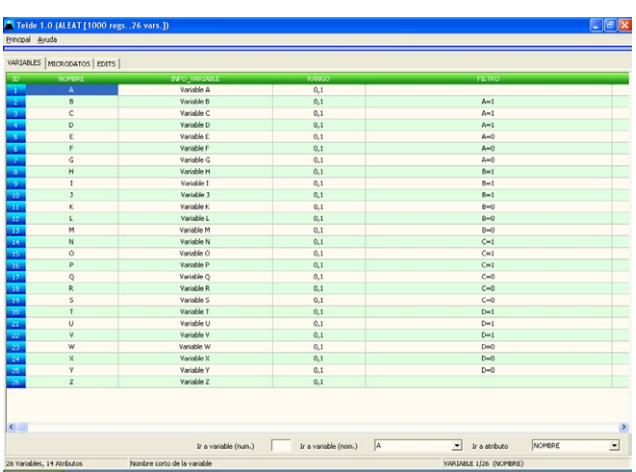

Figura 2

3.4. Pestaña de microdatos

Esta pestaña está situada justo a la derecha de la pestaña de variables (ver Figura 3). En ella podemos encontrar una rejilla principal con todos los microdatos dispuestos por filas y todas las variables dispuestas por columnas. También encontramos una pequeña rejilla auxiliar que nos permite visualizar la información de la variable seleccionada y que se va modificando de forma dinámica mientras nos movemos por ella.

Fayna Álamo Santana
Sergio Delgado Quintero
José Molina González
Juan José Salazar González

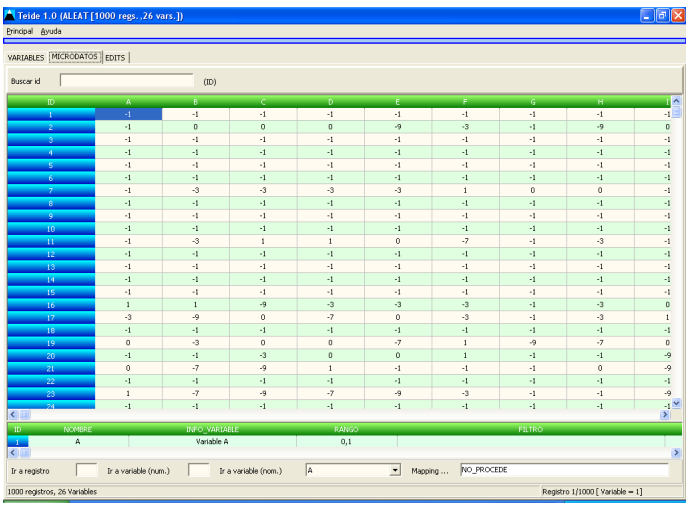

Figura 3

Al igual que antes existen navegadores para movernos a un determinado registro o a una determinada variable. Podemos incluso buscar un valor del identificador principal del registro. Además, en la parte inferior derecha tenemos la posibilidad de visualizar cuál es el texto asociado a cada valor numérico siempre y cuando se haya especificado la tabla de mapping correspondiente a la variable. En la parte inferior de la pestaña se dispone de una barra de estado con información de la situación actual.

3.5. Pestaña de edits

Esta pestaña se encuentra justo a la derecha de la pestaña de microdatos (ver Figura 4). En ella podemos distinguir varios elementos. En primer lugar se presenta una rejilla principal que contiene todos los edits definidos en el sistema, con una columna auxiliar que explica el significado de la regla. Debajo de esta rejilla encontramos un bloque de controles destinados a la gestión de los edits. Con estos controles podremos insertar nuevos edits, borrarlos o modificarlos.

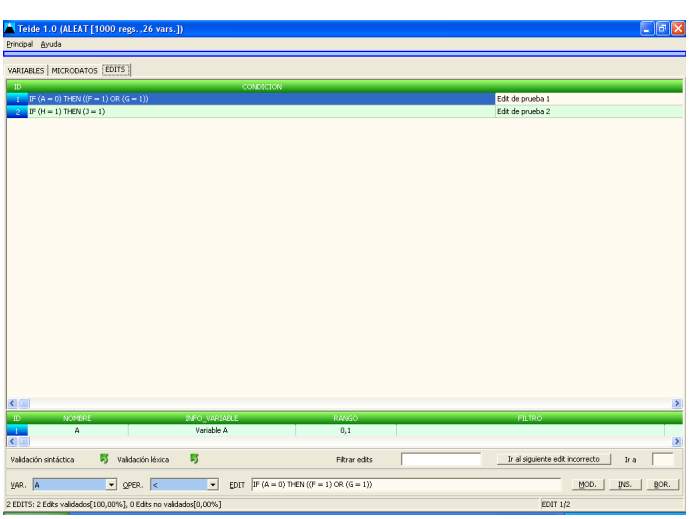

Figura 4

A la hora de insertar un nuevo edit en el sistema, esta pestaña ofrece sistemas de ayuda a la escritura tales como poder seleccionar el nombre de la variable o poder seleccionar el operador. Cuando una variable es seleccionada, todas sus características son mostradas en la rejilla que está situada debajo de la principal. Una vez insertado el nuevo edit, el sistema realiza una evaluación del mismo y es capaz de decir si la validación sintáctica y léxica es correcta o no. Esto nos puede ayudar a ir insertando sólo aquellas reglas que cumplan estos requisitos.

En esta misma pestaña, tenemos la posibilidad tanto de filtrar aquellos edits que contengan una determinada variable, como de ir navegando por aquéllos que son incorrectos, para así poder corregirlos. Al igual que en otras pestañas, se dispone de un navegador para moverse sobre los edits y de una barra de estado que muestra información de la situación actual.

3.6. Pestaña de evaluación de rangos y filtros

Hasta aquí la aplicación únicamente nos ha mostrado la información que teníamos dentro de la base de datos. A partir de este momento la aplicación genera información muy valiosa para el proceso de edición e imputación de datos. Lo primero que debemos hacer es realizar la evaluación de edits de rango y de filtro. Como ya hemos dicho, dentro de la definición de las variables existen dos atributos que nos dicen cuál es el rango y cuál es el filtro para una determinada variable. Al final, tanto el rango como el filtro son reglas, así que podemos referirnos a ellos como edits.

La evaluación de estos edits consiste en comprobar que se cumplan dos condiciones: por un lado ver que el valor que contiene la variable está dentro de los valores permitidos por su rango, y por otro ver si esa variable podía o no ser contestada en función de la evaluación de su filtro.

La pestaña de evaluación de rangos y filtros (ver Figura 5) es la que nos permite visualizar estas evaluaciones de una manera muy gráfica y directa. En esta pestaña existen muchos elementos que presentamos a continuación.

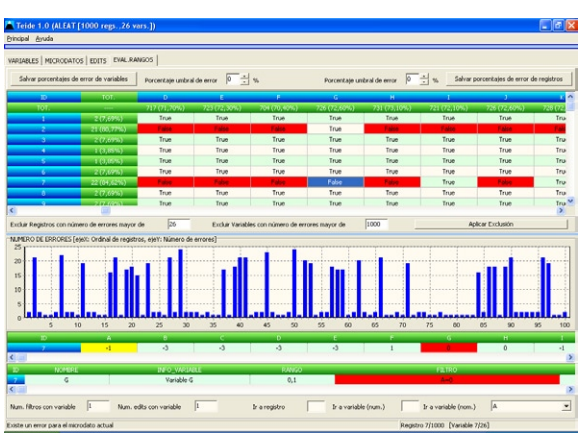

Figura 5

En la mitad superior de la pestaña podemos distinguir la rejilla principal. En esta rejilla tenemos una fila por cada registro de datos y una columna por cada variable. Se puede observar que cada celda de esta rejilla tiene un valor True ó False. Si aparece "True" esto significa que el microdato asociado a esa celda cumple tanto el edit de rango como el edit de filtro. En caso contrario aparece "False". Además los valores falsos están coloreados en rojo así que es muy fácil su identificación.

Para poder tener una idea global del número de errores, justo al lado de cada identificador de registro y nombre de variable, tenemos tanto el número absoluto como el número relativo de errores que existen. Estos datos pueden ser guardados en ficheros de texto con los dos botones que se encuentran en la parte superior, e incluso podemos especificar un umbral de error para que sólo se graben aquellos registros/variables que superen dicho umbral. Debajo de la rejilla principal también tenemos un panel que nos permite excluir del proceso aquellos registros o variables que se pasen de un número de errores determinado. Aquellos casos donde el número de errores resulte excesivamente alto, quizás no merece que sean corregidos y podemos optar por desecharlos de la depuración.

Siguiendo el análisis de arriba hacia abajo, nos encontramos con un histograma. Este histograma representa el número de errores absoluto de cada registro y de cada variable. Está escalado en el eje de las abscisas de 100 en 100, y podemos navegar hacia delante o hacia atrás.

Las dos rejillas que aparecen justo debajo representan la información del registro y de la variable asociados a la celda seleccionada en la rejilla principal. Cuando nos movemos por ella, esta información se actualiza según la celda seleccionada. En la rejilla del registro se representan en amarillo aquellas variables que forman parte del filtro de la variable de la celda seleccionada.

Por último, como siempre, también disponemos de navegadores para registros y variables, junto con una barra de estado para mostrar la información relevante.

3.7. Pestaña de evaluación de edits generales

Una vez que TEIDE ha evaluado los rangos y los filtros, el siguiente paso en este proceso es evaluar los edits generales, aquellos que están definidos en una tabla separada y que tienen la forma "IF( ) THEN ( )".

Para llevar a cabo las evaluaciones, tanto de los rangos, los filtros como de los edits generales, TEIDE hace uso de analizadores léxicos y sintácticos que le permiten construir un árbol en memoria a través del cual ir evaluando las expresiones que se encuentra. Algo así como una calculadora avanzada de expresiones lógicas.

La pestaña resultante de este proceso es la pestaña de evaluación de edits generales (ver Figura 6). Es una pestaña muy similar a la que se ha visto en la Sección 2.5, pero con algunas particularidades que pasaremos a comentar a continuación.

Al igual que en el caso de la evaluación de los edits de rango y de filtro, en esta pestaña también tenemos una rejilla principal, opciones para guardar los totales de errores a fichero, un sistema de exclusión y un histograma para visualizar el número de errores. Ahora lo único que cambia es que en vez de estar hablando de variables estamos hablando de edits. En la rejilla principal, las filas siguen representando los registros pero ahora las columnas representan edits.

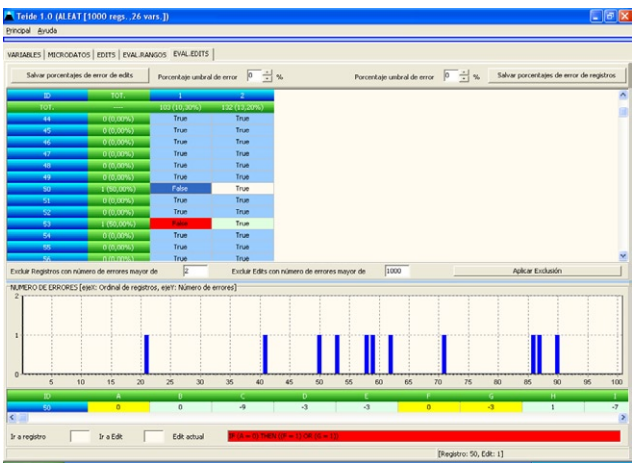

Figura 6

Debajo del histograma seguimos teniendo otra pequeña rejilla que representa los valores del registro seleccionado actualmente, y también tenemos un lugar para el texto del edit que está siendo seleccionado. El código de colores es el mismo que en el caso anterior. También se dispone de navegadores y de barra de estado.

3.8. Pestaña de imputación

En este punto del proceso, ya tenemos evaluados todos los edits definidos en el sistema (rangos, filtros, generales). Por lo tanto podemos empezar con la fase de edición e imputación propiamente dicha.

Para abordar el problema de edición e imputación de datos, TEIDE se basa en los estudios realizados por Fellegi y Holt [1] y en la metodología del registro donante [2]. En ellos se establece que un buen método de edición e imputación debe buscar el menor número de cambios posibles en un registro y debe alterar lo menos posible las distribuciones de frecuencia de las variables.

En esta fase se realizan muchas operaciones entre las que cabe destacar las siguientes. En primer lugar TEIDE construye el conjunto de registros donantes. Un registro donante es aquél que ha satisfecho todos los edits del sistema. Por lo tanto, el resto de los registros son incorrectos y deben ser corregidos. En segundo lugar TEIDE asocia a cada registro su conjunto de registros donantes ordenados en función de una distancia. Es decir, que se tienen los registros que se parecen más en una posición cercana y los registros que difieren más en una posición más lejana. A continuación, el sistema empieza a procesar cada registro incorrecto de manera individual.

Para intentar corregir un registro, debemos modificar los valores de algunas variables. La selección de estas variables es guiada por ciertos heurísticos, y una vez localizadas sus valores son tomados del primer registro donante asociado. Así, cada vez que se modifique un valor necesitamos reevaluar los edits para comprobar que el registro está correcto o, por el contrario, debe seguir en el proceso de depuración. Si el procedimiento no es capaz de hacer que el registro satisfaga todos los edits, se vuelven a repetir estos mismos pasos pero tomando el siguiente registro donante de la lista, y así sucesivamente.

Una vez que se han procesado todos los registros incorrectos, el sistema muestra la pestaña de imputación, análoga a la pestaña de microdatos (ver Figura 2), pero donde los valores están ya modificados y los registros corregidos.

3.9. Pestaña de estadísticas

El proceso automático de edición e imputación ha sido completado, pero debido a la gran cantidad de parámetros, decisiones y procesos que han tenido lugar sobre los datos, se hace necesario un informe final que resuma todos estos detalles

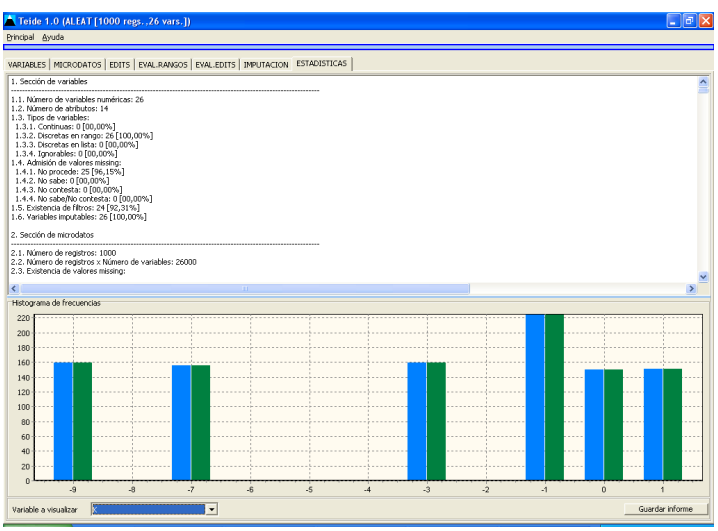

Figura 7

Ese informe se presenta en la pestaña de estadísticas (ver Figura 7) junto con un histograma que permite visualizar las distribuciones de frecuencia de las variables antes y después del proceso de edición e imputación.

El informe detalla minuciosamente todos los procesos y las estadísticas asociados a ellos, que han tenido lugar en TEIDE, además de información sobre los datos de entrada, los datos de salida, las variables y los edits.

Finalmente, el usuario puede guardar los datos imputados en la base de datos y también hacer una copia del informe para tener conocimiento de las operaciones realizadas por la aplicación.

4. Experiencias computacionales

TEIDE no es sólo un prototipo fruto de investigación científica, sino una herramienta práctica diseñada para trabajar sobre encuestas reales, y así lo corrobora su aplicación en la Encuesta de Ingresos y Condiciones de Vida de los Hogares Canarios 2005, y en la Encuesta de Salud Canaria, ambas desarrolladas por el Instituto de Estadística de Canarias (ISTAC). Los resultados obtenidos por TEIDE en estas encuestas han sido satisfactorios.

NOTA: Todas las capturas de pantalla que se presentan en esta comunicación provienen de una ejecución de TEIDE con una encuesta totalmente inventada y con datos aleatorios.

5. Bibliografía

[1] Fellegi, I.P., Holt, D., (1976). A systematic approach to automatic edit and imputation. Journal of the American Statistical Association 71, 17-35.

[2] Ford, B.F., (1983). An overview of hot-deck procedures. Incomplete data in sample surveys: Theory and bibliographies 2, 185-207.