Estadística Española • Volumen 57, número 188 / 2015, pp. 261-285

TEIDE2: una nueva herramienta para la depuración de encuestas*

Mª Salomé Hernández García
Facultad de Matemáticas. Universidad de La Laguna. Tenerife

Juan José Salazar González
Facultad de Matemáticas. Universidad de La Laguna. Tenerife

Resumen

Los institutos de estadística reciben gran cantidad de datos a través de encuestas. De la calidad de éstas dependen las conclusiones que se extraen. Por ello es fundamental depurar las encuestas, eliminando los errores inherentes. Los institutos de estadística dedican gran cantidad de recursos económicos y humanos a este importante objetivo. TEIDE2 es una herramienta informática que intenta reducir estos esfuerzos automatizando al máximo la detección de incoherencia y la imputación de valores perdidos, a la vez que se mantienen las propiedades estadísticas de la encuesta original. TEIDE2 se ha desarrollado y evaluado ampliamente durante los últimos diez años, y recientemente se ha colocado en internet como software abierto y gratuito al público. Este artículo ilustra su uso sobre algunas encuestas reales con el objetivo de despertar el interés de nuevos usuarios, quienes a su vez están invitados a mejorarlo con la única condición de mantener futuras versiones también abiertas y gratuitas al público interesado.

Palabras claves: depuración automática, encuestas, imputación, algoritmos.

Clasificación AMS: 90-08.

TEIDE 2 (Techniques for Editing and Imputation of Statistical Data)

Abstract

Statistical agencies collect lot of data through surveys. The quality of the information released by the agencies depends on the correctness of these surveys. Therefore it is essential to debug surveys, eliminating the inherent errors in records. For this purpose, the agencies devote a large amount of economic and

* TEIDE2 es el resultado del trabajo conjunto de los autores de este artículo con técnicos de diversos institutos de estadística. En especial queremos agradecer el entusiasmo, dedicación y financiación aportados por el Instituto Canario de Estadística (ISTAC). Este trabajo ha sido parcialmente financiado por el proyecto nacional MTM2012-36163-C06-01 y por el proyecto europeo "Data without Boundaries" (Grant agreement 262608).

human resources to editing and imputation. TEIDE2 is a computer tool which attempts to reduce this resource consumption by automating the inconsistency detection and imputation of missing values, while maintaining the statistical properties of the original survey. We illustrate the use of TEIDE2 on some realistic surveys.

Keywords: automatic editing, surveys, imputation, algorithms.

AMS classification: 90-08.

1. Introducción

La depuración de las encuestas es una tarea fundamental que todo instituto de estadística debe acometer permanentemente. Intenta detectar incoherencias en las respuestas a cuestionarios, reemplazando los potenciales errores por valores coherentes con el resto de los datos. Su aplicación intenta garantizar que los institutos de estadística procesen datos sin errores, y en consecuencia incrementa la calidad de los análisis que se realicen. Cada respuesta en un cuestionario se identifica con una *variable* y para la validación de un cuestionario se asume disponer de un conjunto de *reglas de depuración*. Ejemplos de variables son la "edad" y el "estado civil", y un ejemplo de regla de depuración es "si la edad es menor de 16 años entonces el estado civil no puede ser divorciado".

La importancia de esta tarea es tal que periódicamente se organizan congresos internacionales enfocados exclusivamente a la depuración de encuestas (lo que en inglés se denomina "data editing and imputation"). Citamos a título de ejemplo la serie de encuentros "UNECE worksession on statistical data editing", que se han celebrado en Cardiff (18-20 octubre 2000), Helsinki (27-29 mayo 2002), Madrid (20-22 octubre 2003), Ottawa (16-18 mayo 2005), Bonn (25-27 septiembre 2006), Viena (21-23 abril 2008), Neuchàtel (5-7 octubre 2009), Slovenia (9-11 mayo 2011), Oslo (24-26 septiembre 2012) y Paris (28-30 abril 2014). Los trabajos presentados en esta serie de encuentros pueden obtenerse en la página web http://www.unece.org/statistics/meetings-and-events.html. También la importancia de la depuración de encuestas queda patente del proyecto europeo EUREDIT financiado por la Comisión Europea a través de Eurostat con algo más de dos millones de euros en el Quinto Programa Marco. El objetivo de este proyecto realizado entre el 2000 y el 2003 fue el desarrollo y la evaluación de nuevos métodos de depuración de encuestas. Pueden verse los resultados de este proyecto en la página web http://www.cs.york.ac.uk/euredit/. Individualmente algunos institutos nacionales de estadística también dedican esfuerzos para afrontar esta tarea, tanto en investigación de nuevas metodologías (véase por ejemplo Arbués, González y Revilla (2009,2010)) como en el desarrollo de herramientas informáticas. En esta última línea destacan Statcan de Canadá con las herramientas CANEDIT y GEIS, U.S. Census de Estados Unidos con la herramienta SPEER, Istat de Italia con la herramienta DIESIS, CBS de Holanda con las herramientas SLICE and WAID, y el INE de España con la herramienta DIA (véase por ejemplo Gómez Alonso (1980) y Villán Criado (1992)).

Lamentablemente estas herramientas informáticas no son abiertas y públicamente disponibles, lo que dificulta su integración y uso por parte de otros institutos de estadística diferentes de sus creadores. En el mejor de los casos, algunas son comercializadas a través de complejos procesos, y otras están disponibles sólo en el organismo donde fueron creados y pueden usarse sólo enviando los datos y esperando un tiempo. Estas políticas no contribuyen al correcto desarrollo de las tareas de depuración de datos en, por ejemplo, institutos autonómicos de estadística. Nuestro artículo describe TEIDE2, una herramienta que se ha creado para dar transparencia a los procesos de depuración e imputación mediante software abierto y gratuito. TEIDE2 no sustituye al experto en el instituto de estadística, sino que le permite centrarse más en la parte más desafiante y técnica de la depuración al realizar TEIDE2 la parte menos relevante y tediosa de forma automática.

El artículo está estructurado como sigue. En la sección 2 damos una breve descripción de cómo nace y se desarrolla nuestro proyecto. La sección 3 introduce los elementos fundamentales del software resultante del proyecto, TEIDE2. La sección 4 resume su forma de uso. La sección 5 muestra resultados computacionales sobre encuestas reales. Por último, la sección 6 termina con varias conclusiones.

2. ¿Cómo nace TEIDE?

Las siglas TEIDE proceden de "Técnicas de Edición e Imputación de Datos Estadísticos", que fue el título del proyecto nacional TIC2002-00896 financiado por el Ministerio de Ciencia y Tecnología entre los años 2003 y 2005. Gracias al apoyo de este proyecto desarrollamos nuevos modelos y algoritmos para el tratamiento de encuestas con datos numéricos. Concretamente en el artículo de Riera y Salazar (2007a) se presenta un algoritmo exacto de tipo "ramificación y corte" para la resolución del problema con hasta 100 variables cualitativas y 40 reglas de depuración lineales en tiempos de cálculos próximos al minuto sobre un ordenador personal. El artículo de Riera y Salazar (2007b) describe un algoritmo heurístico experimentado con encuestas con hasta 1000 variables cualitativas y 400 reglas de depuración. También a raíz del proyecto nacional TEIDE se desarrolló un prototipo de software para afrontar encuestas con variables cuantitativas y cualitativas, de utilidad práctica en institutos de estadísticas. Este prototipo despertó el interés del Instituto Canarias de Estadística (ISTAC), y como fruto de una colaboración conjunta se desarrolló la herramienta informática TEIDE1. Esta herramienta estuvo desarrollada en Borland C++, trabajaba con bases de datos Microsoft Access, y sirvió para depurar encuestas del ISTAC como la Encuesta de Salud 2005 y la Encuesta de la Cesta de la Compra. Algunos de sus aspectos aparecen en el artículo de Delgado y Salazar (2008). La experiencia adquirida al usar TEIDE1 sobre encuestas reales y el interés mostrado por diversos institutos de estadística, autonómicos y nacionales, favoreció la implementación de una nueva herramienta informática que llamamos TEIDE2. Esta nueva herramienta ha sido desarrollada bajo la metodología del "software libre y abierto" y se ha beneficiado de la experiencia acumulada durante 10 años trabajando sobre encuestas reales en estrecha colaboración con técnicos de institutos de estadística.

TEIDE2 se creó para cubrir la carencia de una herramienta de software libre y gratuito que organismos públicos pudiesen integrar en sus sistemas de gestión de datos sin necesidad de codificar herramientas básicas de depuración de encuestas. La metodología implementada en TEIDE2, descrita en la siguiente sección, está ampliamente aceptada por la comunidad de expertos en depuración. Consiste en una primera etapa conocida como "localización de errores" y en una segunda etapa basa en la "imputación por donación". Sigue la metodología pionera de Fellegi y Holt (1976) según la cual los registros incorrectos deben modificarse cuando no cumplen un conjunto de reglas de coherencia. TEIDE2 ha implementado procedimientos dando servicio a diversos institutos de estadística autonómicos en España. Las encuestas que motivaron nuestra investigación tuvieron sus variables de tipo categórico, con escasas variables continuas. Por este motivo TEIDE2 está dotado de una técnica de imputación por donación sofisticada (descrito en la sección 3.5), con algunas técnicas de imputación por regresión muy simples para variables continuas. El hecho de que TEIDE2 es software libre y gratuito facilita que otras reglas de imputación puedan incorporarse en el futuro si son deseables o necesarias. Véase por ejemplo de Waal et al (2011) para conocer más detalles sobre estos procedimientos.

3. Metodología

Para afrontar la depuración de una encuesta se necesita disponer de tres elementos fundamentes: variables, datos, y reglas de depuración. En esta sección describimos cada uno de ellos, desde nuestro punto de vista, y exponemos la forma de proceder de TEIDE2.

3.1 Variables

Como se ha adelantado en la introducción, las variables representan respuestas en un cuestionario. Pueden ser categóricas (como por ejemplo *estado civil*) o continúas (como por ejemplo, *ingresos*). Las primeras asumen un número discreto de valores mientras que las segundas se mueven dentro de un intervalo. Tanto los valores posibles de una variable categórica como los extremos del intervalo que definen la variable continua se determinan al definir el rango de la variable. En la versión actual TEIDE2 no contempla variables que combinen valores de categóricas y de continuas a la vez, en algunas ocasiones llamadas *semicontinuas*. El motivo de no considerar estas otras variables es que en los años de uso de TEIDE2 nunca necesitamos su tratamiento, pero es posible ampliar TEIDE2 para que también las considere aprovechando que el código de TEIDE2 es libre y abierto. Un tercer tipo de variables asumen texto libre (como *apellidos*) y aquí supondremos que no se desean modificar.

3.2 Datos

Cada persona o entidad que responde al cuestionario genera un registro en una base de datos, y el conjunto de todos los registros se llama *muestra*.

3.3 Reglas de depuración

El primer paso del proceso de depuración consiste en clasificar cada registro como válido o inválido, y para ello se necesita disponer de un conjunto de reglas de depuración. Éstas son condiciones lógicas y darán como resultado "coherente" o "incoherente" sobre cada registro de la muestra. Las reglas de depuración deben ser generadas por algún experto del instituto de estadística atendiendo a lo que se espera que respondan los encuestados. Esta tarea entender bien el cuestionario (preferiblemente haber participado en su diseño), y por ello es una buena práctica que las reglas de depuración se escriban a la vez que se diseña la encuesta (es decir, antes de iniciar la recogida de datos). TEIDE2 no construye reglas de depuración, aunque puede usarse para ver si una regla concreta está sintáctica y léxicamente bien escrita, y para evaluar si cada registro cumple o no con una posible regla. El análisis sintáctico busca, entre otras cosas, que no le sobren o falten paréntesis, por ejemplo, ni nexos entre clausulas diferentes que compongan la regla. El análisis léxico consiste en asegurar que las variables usadas en las clausulas estén entre las variables inicialmente declaradas. En el caso de que se desee incorporar nuevas reglas de depuración durante la recogida de datos (no recomendable), TEIDE2 también facilita la tarea porque puede usarse como un simulador: distintas reglas pueden lugar a distintas incoherencias. No obstante, reiteramos que, tal y como se cita en de Waal et al (2011), lo recomendable es que las reglas de depuración se diseñen antes de comenzar la recogida de datos.

Las reglas en TEIDE2 se agrupan en tres tipos que describimos en las siguientes secciones.

3.3.1 Reglas de Rango.

Una regla de rango es la especificación del conjunto de valores que puede asumir una variable. Determina si una variable es cualitativa o cuantitativa, y en cada caso qué valores admite. Concretamente las variables cuantitativas (o categóricas) vienen definidas por una lista de valores enteros (positivos o negativos), cada uno con una etiqueta que interpreta el valor. Las variables cualitativas (o continuas) vienen definidas por los extremos del intervalo donde asumen valores, y estos extremos pueden ser valores muy grandes para representar variables no acotadas (en uno o ambos extremos del intervalo). También el rango de una variable muestra si admite valores especiales (a los que llamamos *missings*) como "no sabe", "no contesta", "no procede", etc.

3.3.2 Reglas de Filtro.

Es muy común la existencia de condiciones en los cuestionarios para definir cuándo se debe formular una pregunta al encuestado y/o cuándo no. Por ejemplo, si la respuesta a la pregunta "¿trabaja?" es "no" entonces no corresponde esperar un "sueldo" en las respuestas (o en otras palabras, se espera que la variable "sueldo" contenga el valor "no procede"). La regla que refleja esta forma correcta de proceder se llama *regla de filtro* de la variable "sueldo". Estas reglas también se conocen como reglas "if-then" en otros enfoques (véase de Waal et al (2011)). Cuando la muestra de datos se recoge mediante dispositivos electrónicos, estos dispositivos obligan el cumplimiento de las reglas de flujo y por tanto no es necesaria su posterior verificación. Sin embargo, los datos de

muchas encuestas (especialmente en institutos de estadística autonómicos) se recogen sobre papel y por tanto estas reglas de filtro necesitan ser consideradas como cualquier otra regla de depuración.

Cada variable X de la encuesta puede tener asociada una condición lógica B para reflejar su regla de filtro. Caben tres formas diferentes de entender esta condición para esta variable:

- Forma A: Si "no B" entonces "X = no procede"
- Forma B: Si "B" entonces "X ≠ no procede"
- Forma C: "B" si y sólo si "X ≠ no procede"

En TEIDE2 se da la posibilidad de asociar una regla de filtro a cada variable introduciendo la condición y la forma en la que debe entenderse.

3.3.3 Reglas de depuración explícitas.

Llamamos *reglas de depuración explícitas* a todas las demás reglas que no caigan en las dos categorías expuestas en las secciones 3.3.1 y 3.3.2. No se refieren a las reglas implícitas que se citan en Fellegi y Holt (1976), que TEIDE2 no construye, sino a las reglas que introduce directamente un técnico en la herramienta TEIDE2 y que no son ni las de rango ni las de filtro. Pueden ser estructuras lógicas, expresiones matemáticas con relaciones lineales, o combinaciones de ambas. Lo único que no se admite en la versión actual de TEIDE2 es el uso de funciones no-lineales como logaritmo, raíz cuadrada, trigonométricas, exponentes, etc. El motivo de excluir estas funciones es porque añaden una complejidad grande en la imputación y no se ha visto aún su necesidad sobre las encuestas en las que se ha usado TEIDE2 hasta el momento. No obstante, si en algún momento se necesitase trabajar con reglas no-lineales, es posible extender el código de TEIDE2 para que las procese.

3.4 El proceso de edición.

Tradicionalmente el proceso de clasificar cada registro como coherente o incoherente se llama *edición*, y TEIDE2 lo realiza en dos fases: evaluación de reglas de rango y de filtro, y evaluación de las reglas explícitas. En cada fase se examina cada regla sobre cada registro, y se clasifica éste como coherente o como incoherente según cumpla o no todas las reglas, respectivamente.

3.5 El proceso de imputación.

Una vez que cada registro ha sido clasificado como coherente o incoherente comienza la fase llamada *imputación*. Cuando un registro incoherente incumple un porcentaje elevado de reglas de depuración puede excluirse de esta fase en TEIDE2, pero por simplicidad en la exposición asumiremos que el instituto desea aplicar esta fase a todos los registros incoherentes. La imputación consiste en determinar las variables a modificar y sus nuevos valores en cada registro incoherente. El primer problema (llamado "localización de errores") se resuelve tradicionalmente mediante técnicas de optimización que combinan la enumeración, la programación matemática y métodos aproximados como el tradicional

"método del registro donante". Véase Delgado-Quintero y Salazar-González (2008) como referencia a la metodología que sigue TEIDE2.

El método del registro donante selecciona un subconjunto de registros coherentes de gran parecido a cada registro incorrecto. Los registros se llaman "donantes potenciales" y el parecido se establece en función de una distancia que se construye comparando los distintos valores de cada variable en cada registro. Antes de detallar cómo se construye esta distancia necesitamos introducir el concepto de grafo de reglas.

3.5.1 Grafo de reglas.

El proceso de imputación hace uso de un grafo. Este grafo viene dado por G = (N,A), donde el conjunto de nodos N representa las variables en la encuesta y el conjunto de arcos A une dos variables cuando ambas están presentes en una misma regla de depuración.

3.5.2 Variables básicas y extendidas a imputar.

Cada registro incoherente tiene asociado un conjunto de variables que podrían necesitar ser modificadas para convertir este registro en coherente. Llamamos *variables básicas* a las variables implicadas directamente en reglas de depuración que el registro incumple. Es evidente que al menos alguna de estas variables necesita ser modificada. Pero puede suceder que modificando alguna variable, alguna regla de depuración que antes sí se verificaba, ahora pase a no cumplirse. Llamamos *variables extendidas directas* a aquéllas que están en reglas de depuración donde también hay variables básicas, no importa que tales reglas las verifiquen o incumpla el registro incoherente. Finalmente llamamos *variables extendidas* a las variables que están en reglas de depuración donde también hay variables básicas o variables extendidas. En otras palabras, las variables extendidas directas son las variables adyacentes a variables básicas mediante arcos A, y las variables extendidas se corresponden con los nodos en las componentes conexas que contienen una variable básica en el grafo G.

Dado un registro m_i denotaremos a sus variables básicas por $v_{bas}(i)$ y a sus variables extendidas por $v_{ext}(i)$.

3.5.3 Distancia entre dos registros.

Los conceptos anteriores nos permiten definir ahora la distancia que se usa para comparar dos registros. Se basa en funciones que miden la similitud de valores en cada variable. Concretamente, cuando una variable es cualitativa, la función vale 1 cuando los valores de esta variable en los dos registros son diferentes, y 0 en otro caso. Cuando la variable es cuantitativa la función es la diferencia en valor absoluto entre los valores, dividido por la diferencia entre el mayor y el menor valor que puede asumir dicha variable. Conviene notar que el denominador puede ser un valor muy grande, lo cual tiene un efecto directo en el cálculo de la distancia.

La distancia entre dos registros suma la evaluación de la función indicada sobre todas las variables, ponderando las variables básicas con escalares diferentes, las variables extendidas directas, las variables extendidas, y las restantes variables. Basados en nuestras

experiencias computacionales, TEIDE2 coloca por defecto las ponderaciones a valores 0.01, 0.10, 0.01, 0.05, respectivamente. Un usuario puede cambiar estos valores en función de su conocimiento de la encuesta, e incluso analizar los resultados tras valoras diferentes alternativas. Dependiendo de la encuesta, el resultado de TEIDE2 depende fuertemente de estos valores, por lo que no se trata de un procedimiento robusto frente a estos parámetros. Al contrario, los resultados de TEIDE2 son sensibles a los valores de estos y otros parámetros, y por ello se recomienda al técnico que haga un análisis de sensibilidad de sus resultados antes de terminar la depuración de una encuesta.

3.5.4 Algoritmo de imputación

Para cada registro incorrecto se seleccionan los k registros correctos (potenciales donantes) ordenados usando la distancia antes definida. Por defecto usamos k=500, basado en nuestras experiencias computacionales, pero este valor lo puede cambiar el usuario. Luego se va probando con cada uno de los registros correctos hasta encontrar el mejor donante, y éste será el que corrija el registro incorrecto. TEIDE2 también permite la imputación de una variable mediante valores de estadísticos descriptivos, como la media, moda, mediana, e incluso expresiones introducidas por un usuario para aplicar imputación por regresión (ver sección 4).

Un registro se corregirá de manera satisfactoria cuando el número de variables modificadas sea menor que el número de variables básicas, o bien de manera insatisfactoria (que TEIDE2 muestra como *warning*) cuando el número de variables modificadas sea mayor que el número de variables básicas. Nótese que este segundo tipo de corrección, aunque no deseada, puede ser inevitable, y es por ello por lo que TEIDE2 atrae la atención del usuario sobre ellos (en algunos casos revelan errores importantes en la construcción de las reglas de depuración o del cuestionario).

Determinado el registro que donará los valores al registro incorrecto, el problema que se plantea es: ¿cuántos valores debe donar el registro donante para hacer que el registro incorrecto cumpla todas las reglas de depuración? Para responder esta cuestión iniciamos la corrección de las variables con valor erróneo de rango copiando los valores del donante en estas variables. Luego se continúa con la corrección de las variables que incumplen las reglas. La donación se puede dividir en dos fases. En una primera fase, se intentará arreglar el registro incorrecto probando con todas las combinaciones de donación de variables básicas a imputar. Si para el registro incorrecto m_i tenemos un número de variables básicas a imputar $t = |v_{bas}(i)|$, probaremos a donar 2^t combinaciones, lo que puede ser excesivo en la práctica. Para limitar el tiempo de cálculo en esta exploración, sólo se llevarán a cabo si el número de variables básicas a imputar está por debajo de un umbral definido (por defecto, no más de 10). Si en alguna de estas combinaciones el registro pasa a cumplir todas las reglas de depuración, salimos de la imputación con éxito. Si no es así, pasamos a la segunda fase del algoritmo.

En esta segunda fase se donan una a una el resto de variables extendidas a imputar en el registro incorrecto m_i. El orden en el que se van considerando estas variables conviene que no sea aleatorio, sino que conviene explorar primero las de mayor aparición en las reglas de depuración. Es decir, si $ed(v_k)$ devuelve el número de reglas en las que aparece

la variable v_k, se intenta donar en cada iteración la variable extendida con mayor $ed(v_k)$. En el peor de los casos, este proceso puede terminar donando todo el conjunto de variables básicas a imputar y todo el conjunto de variables extendidas a imputar, con lo que se asegura que el registro incorrecto tendrá todas sus variables conflictivas modificadas. El registro será correcto y TEIDE2 los visualizará como *warning* para que el usuario le preste especial atención. También este proceso puede terminar clasificando el registro como *no-corregido*, llamando la atención del usuario sobre él. Recordemos que, aunque TEIDE2 intenta automatizar completamente la depuración, no hay garantía de que siempre lo consiga. Referimos al lector a Delgado-Quintero y Salazar-González (2008) para la metodología que inspiró la implementación en TEIDE2.

4. ¿Cómo funciona TEIDE2?

Nuestra aplicación informática toma como datos de entrada un fichero llamado metafile que contiene, entre otra información, el nombre de la base de datos Oracle, Microsoft Office Access (.mdb), Microsoft Office Excel (.xls) o en formato abierto XML. Esta base de dato es otro fichero con las tablas necesarias:

Variables: son los campos del cuestionario, cada uno con su descripción:

ID: identificador de la variable (es decir, código).

NOMBRE: nombre de la variable.

INFO_VARIABLE: información sobre el significado de la variable.

TIPO: indica si es discreto en rango, discreto en lista, continuo, o texto.

RANGO: valores que puede tomar (de una lista o de un rango de valores).

FILTRO: condición que define su regla de filtro, descrita en la sección 3.3.2 (asociada al valor NO_PROCEDE en la variable).

INFO_FILTRO: descripción del significado de la regla de filtro.

SENTIDO_FILTRO: puede tomar los valores a, b o c, con el siguiente significado:

a: if (nofiltro) then (valor = No Procede)

b: if (filtro) then (valor != No Procede)

c: a y b

Toma el valor a por defecto.

IMPUTABLE: indica si la variable debe o no imputarse

Missings: Enumera los distintos valores especiales que puede tomar la variable, y que deben existir en otra tabla dentro de la base de datos. Hablamos de la "tabla Missing" que puede declarar valores como por ejemplo NO_PROCEDE, NO_SABE, NO_CONTESTA, NS_NC.

PESO: indica que ponderación tiene esta variable con respecto a las demás. Se utiliza en la definición de distancia introducida en la sección 3.5.3.

MAPPING: indica el nombre de la "tabla mapping" que contiene el significado de los valores que toma esa variable.

IMP_NUM: indica el tipo de imputación que se desea aplicar en caso de necesitar modificar su valor. Por defecto es por medio del registro donante que se explica en la sección 3.5.4, y se realiza cuando este campo se deja vacío. Otros valores que puede tomar son:

- MEDIANA: Es el valor de la variable que separa en dos grupos los valores de las variables, ordenadas de menor a mayor. También se puede indicar el número de donantes a tener en cuenta, en el caso de no especificar se tomará el número de donantes como 10.
- MEDIA_R: Se obtiene calculando la media de los valores de la variable salvo un porcentaje de los más grandes y el mismo % de los más pequeños. También se puede indicar el número de donantes a tener en cuenta y el número de valores recortados, en el caso de no especificar se tomará el número de donantes como 10 y el número de valores recortados como 2.
- MEDIA: Es la suma de todos los valores de la variable dividida entre el número total de elementos. También se puede indicar el número de donantes a tener en cuenta, en el caso de no especificar se tomará el número de donantes como 10.
- MODA: Es el valor de la variable que más veces se repite, es decir, el valor que tenga mayor frecuencia absoluta. También se puede indicar el número de donantes a tener en cuenta, en el caso de no especificar se tomará el número de donantes como 10.
- Ecuación de regresión: Se puede poner una función de regresión constituida por números, variables, sumas, restas y productos.
- Campo vacío: En el caso de que el campo esté vacío la aplicación realizará la imputación mediante el registro donante.

Microdatos: contiene todos los valores que toman las variables en cada registro.

Reglas: son las reglas de depuración que tienen que cumplir los registros para ser considerados correctos. La tabla contiene

ID: identificador de la regla;

CONDICION: la regla de depuración;

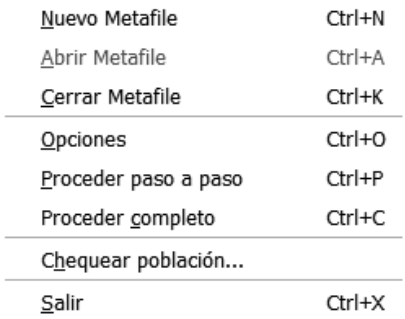
DESCRIPCION: información sobre el significado de la regla.

Missing: contiene los valores especiales que pueden asumir las variables. Por ejemplo:

No_Procede	-1
No_Contesta	-2
No_Sabe	-7
Ns_Nc	-9

Mapping: contiene el significado de cada valor "missing" que puede tomar esa variable.

Al ejecutar TEIDE2 el usuario tiene que elegir una opción del menú que se muestra a continuación:



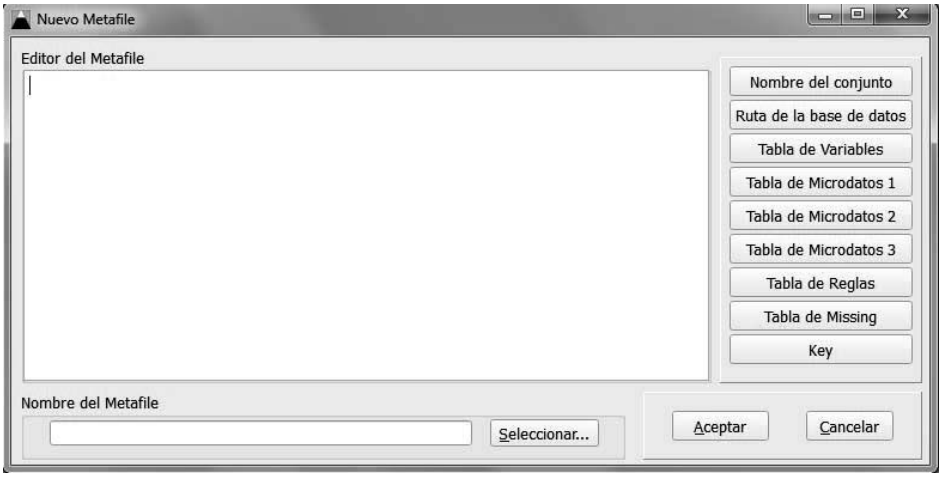
Si se desea crear un metafile, el programa le ayuda con la opción "Nuevo Metafile" (Figura 1). Por defecto, los metafiles tienen extensión .vme y se almacenan en la carpeta metafiles. Un ejemplo es:

```
<NOM> "SAMPLE"
<RBD> "data\SAMPLE.mdb"
<VAR> "VARIABLES"
<MD1> "DATOS"
<EDT> "EDITS"
<MIS> "INFO_MISSING"
<KEY> "TH01"
```

donde NOM se asocia al nombre de la encuesta, RBD indica cómo se llama la base de datos, MDn es la tabla con los microdatos (tantas como necesitemos), EDT describe la tabla con las reglas, MIS tabla missing y KEY variable clave.

Figura 1

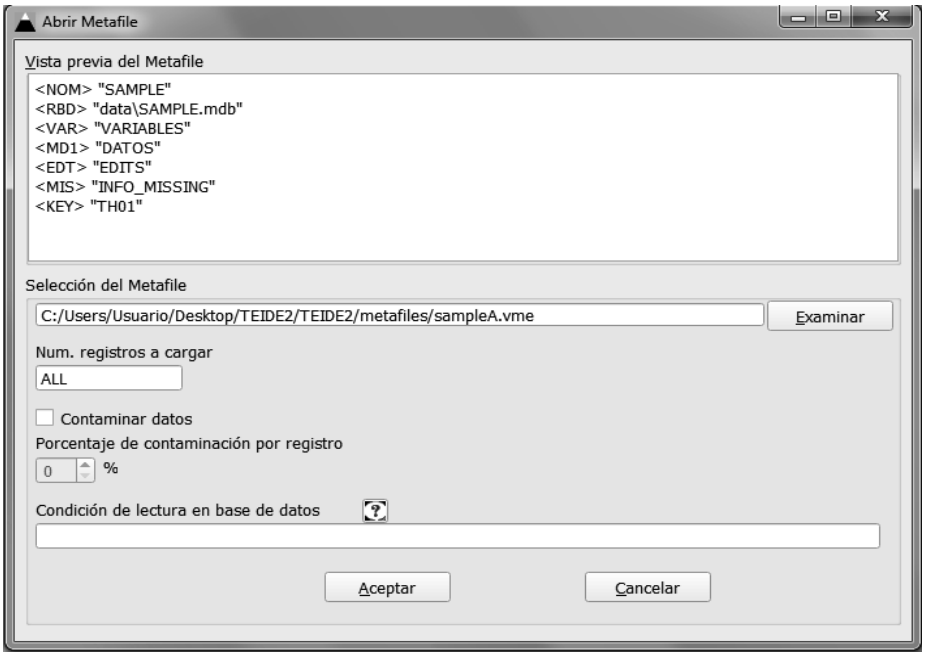
Crear un metafile nuevo



El usuario tiene que abrir un metafile creado para poder incorporar los datos (Figura 2).

Figura 2

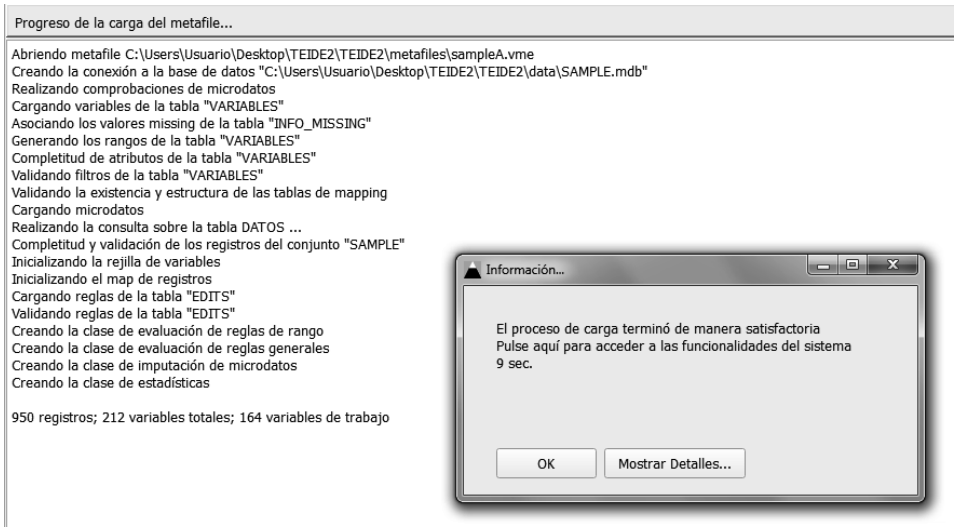
Abrir un metafile existente



Cuando se selecciona el metafile, comienza el proceso de carga en el que el programa leerá todas las tablas y mostrará la información en las pestañas: Variables, Microdatos y Reglas. Véase a Figura 3.

Figura 3

Progreso de la carga.



En la pestaña de variables (Figura 4) se muestra la información de cada variable como puede ser su nombre, rango, filtro (en el caso de que lo tenga), si es imputable, etc. Se puede acceder a una determinada variable o atributo de forma rápida mediante el panel de desplazamiento situado en la parte inferior de la pantalla, con sólo indicar el número o nombre de la variable y también al atributo indicando su nombre.

Figura 4

Pestaña Variables

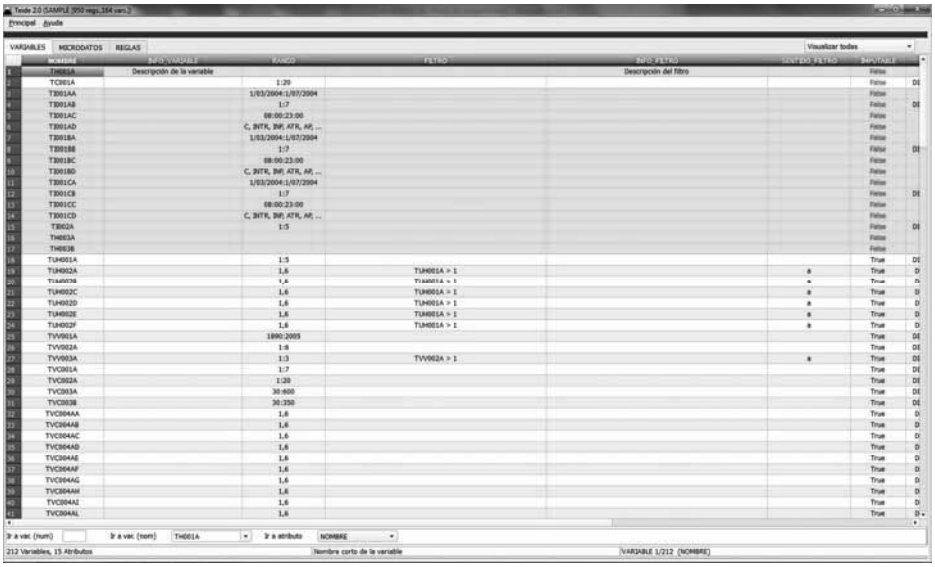
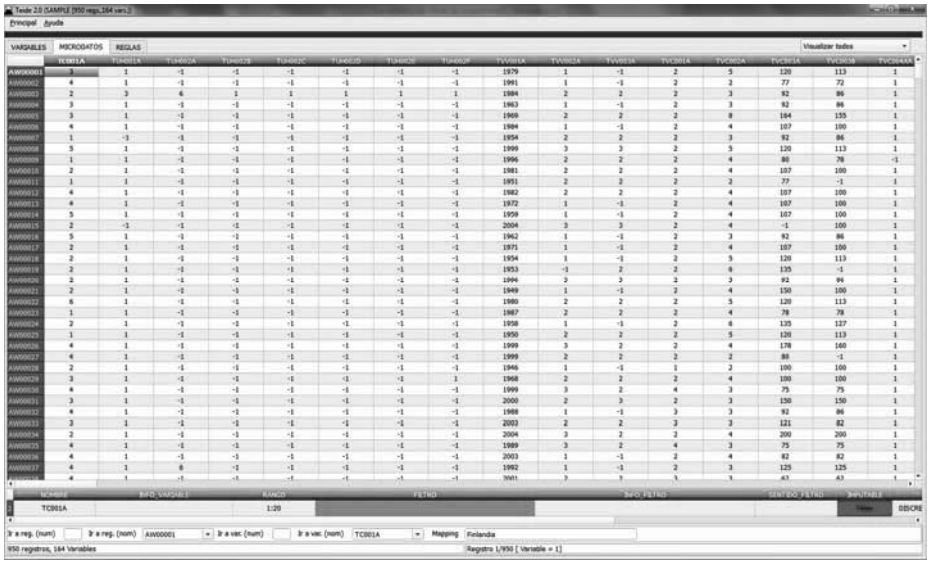


Figura 5

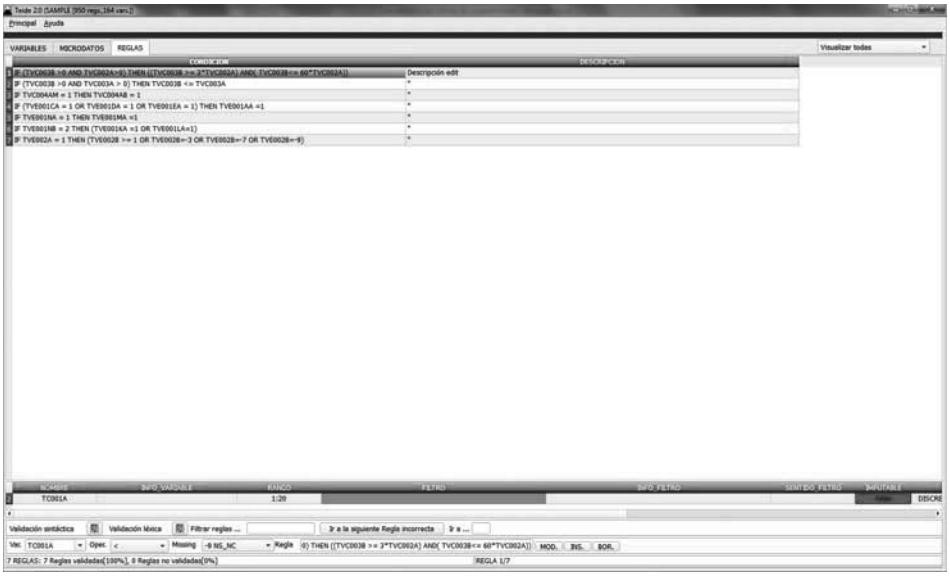
Pestaña Microdatos



La pestaña de microdatos (Figura 5) muestra todos los registros de los que consta la encuesta y los valores que toman las variables en cada uno de ellos. También muestra información sobre la variable seleccionada. Permite acceder de forma cómoda a los registros o variables con sólo indicar a cuál se desea visualizar.

Figura 6

Pestaña Reglas



La pestaña de reglas (Figura 6) muestra todas las reglas de depuración explicitas asociadas a la encuesta. Mediante colores muestra si son correctas o no desde un punto de vista gramático y sintáctico. Permite modificar, borrar o insertar una nueva regla de depuración de forma fácil.

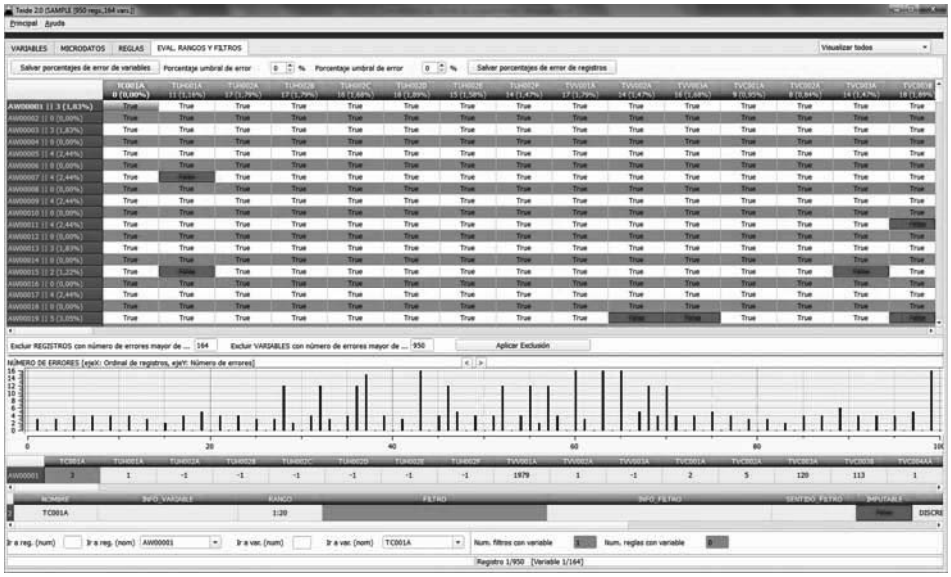
Una vez mostrada la información del proceso de carga, la aplicación TEIDE2 pasa a la fase de edición e imputación. Se puede proceder de dos maneras: o bien paso a paso, o bien de forma ininterrumpida. Si la opción elegida es paso a paso, el usuario podrá ir siguiendo el desarrollo del algoritmo descrito en la sección 3.5.4, controlando si en algún momento lo quiere detener o si se desea introducir alguna modificación. En cambio si la opción es de forma ininterrumpida el usuario no podrá detener ni el proceso de edición ni el de imputación, y la aplicación le mostrará todas las pestañas una vez ha terminado todo el proceso. Cuando se depura una encuesta, para las primeras ejecuciones se recomienda la ejecución paso a paso, mientras que en las últimas ejecuciones (donde quizás se desea experimentar con distintos parámetros) se puede preferir una ejecución ininterrumpida.

Ante una ejecución paso a paso, la siguiente ventana que se visualiza corresponde a la evaluación de rangos y filtros (Figura 7), en la cual se muestran todos los registros con

sus variables, indicando para cada variable si se verifica correctamente su rango y/o su filtro. También muestra información sobre los valores de un determinado registro o una determinada variable. Dispone de una barra de desplazamiento para poder acceder de forma cómoda a un registro o variable.

Figura 7

Pestaña Evaluación de Rangos y Filtros

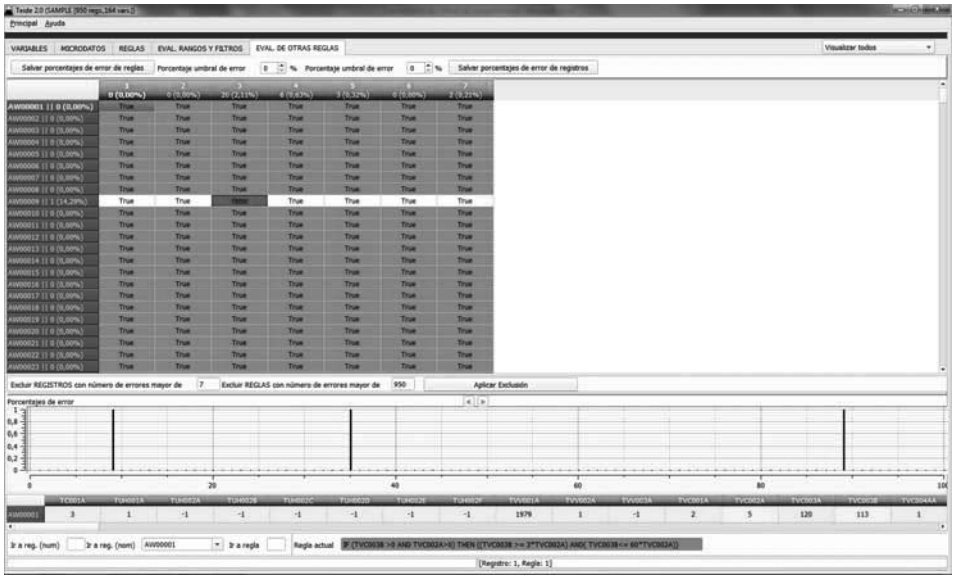


.

Cuando se han estudiado los rangos y filtros, el siguiente paso es el estudio de las reglas de depuración explícitas (Figura 8). La pestaña de evaluación de otras reglas muestra qué registros cumplen todas las reglas propuestas. Para los registros incorrectos muestra los valores que toman las variables incluidas en las reglas que incumple (variables básicas, que muestra en rojo). Permite acceder de forma instantánea a un determinado registro o a una determinada regla para visualizar su información y ayudar al usuario en la comprensión del problema.

Figura 8

Pestaña Evaluación de otras Reglas



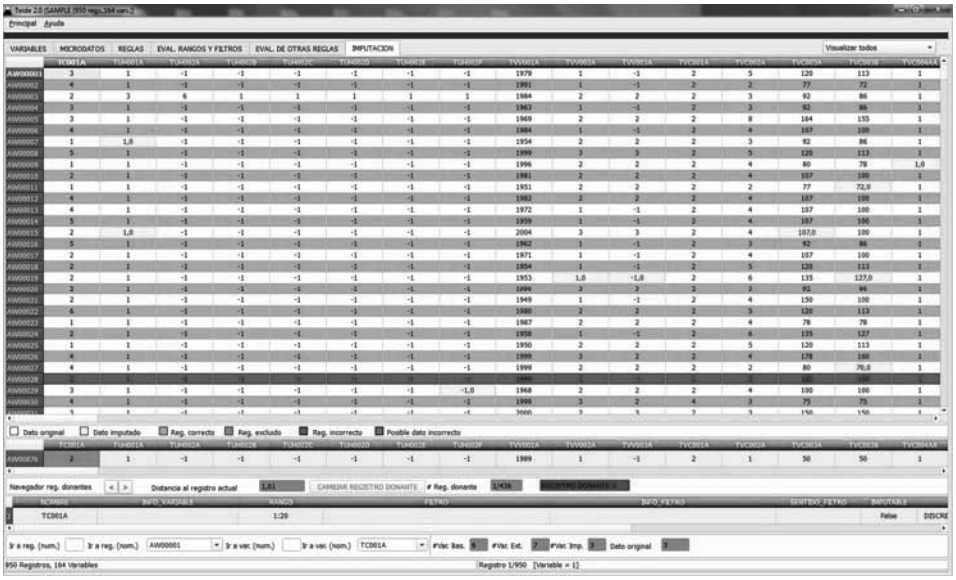
Una vez terminados los procesos de evaluación, el siguiente paso es la imputación (Figura 9). La pestaña de imputación muestra, para todos los registros y sus variables, la información siguiente:

- Registros correctos y corregidos, en color blanco y azul alternativamente.
- Registros incorrectos y que no se han podido corregir, en color rojo.
- Valores de variables en un registro incorrecto que TEIDE2 no ha podido corregir pero que puede que sea responsable de que ese registro sea incorrecto (variables extendidas) en color magenta.
- Registros excluidos (quizás porque incumplen demasiadas reglas) en color naranja.
- Valores de un registro incorrecto que TEIDE2 ha modificado, en color amarillo.

Mediante el botón derecho del ratón se pueden guardar los microdatos imputados en un fichero de texto, o en una base de datos con el mismo formato que tenían los microdatos originales.

Figura 9

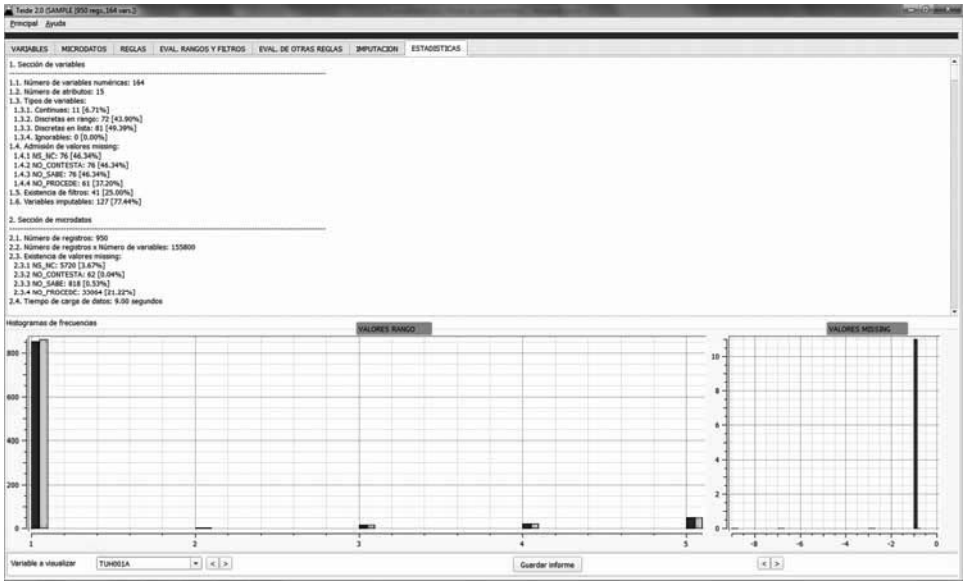
Pestaña Imputación



Cuando ha terminado la edición e imputación se muestra la pestaña de estadísticas (Figura 10). Se trata de un informe con toda la información relativa a los resultados obtenidos. Así como dos gráficas en las que se pueden ver para cada variable los valores que tenían antes y los que tiene después de la depuración. Una gráfica muestra la distribución de valores de rango y la otra gráfica muestra la distribución de los valores especiales (missing) por variable.

Figura 10

Pestaña Estadísticas



La información que muestra la pestaña de estadísticas se puede almacenar en un fichero de texto y es la siguiente:

1. **SECCIÓN DE VARIABLES**: Información referente a las variables leídas en la aplicación.
2. **SECCIÓN DE MICRODATOS**: Información referente a los microdatos cargados en la aplicación.
3. **SECCIÓN DE REGLAS**: Información referente a las reglas de depuración cargadas en la aplicación.
4. **SECCIÓN DE RANGOS Y FILTROS**: Información referente al proceso de evaluación de rangos.
5. **SECCIÓN DE TEST**: Información referente al proceso de evaluación de reglas explícitas.
6. **SECCIÓN DE IMPUTACIÓN**: Información referente al proceso de imputación.
7. **LISTADO DE REGLAS DE RANGO**: Especificación formal de los rangos en forma de reglas.
8. **LISTADO DE REGLAS DE FILTRO**: Especificación formal de los filtros en forma de reglas.
9. **LISTADO DE REGLAS GENERALES**: Especificación formal de las reglas explícitas.

10. **INCUMPLIMIENTO DE REGLAS**: Información sobre las variables imputadas:

    VBI (variables básicas a imputar): son aquellas variables (imputables o no imputables) que pertenecen a las reglas que incumplen un determinado registro.

    vbi (variables básicas a imputar imputables): son aquellas variables (imputables) que pertenecen a las reglas que incumplen un determinado registro.

    VEI(C) (variables extendidas cortas a imputar): variables obtenidas a partir de las básicas y que tienen reglas en común.

    VEI(L) (variables extendidas largas a imputar): variables obtenidas a partir de las básicas y de las extendidas cortas, y que tienen reglas en común.

    VI: número de variables imputadas.

11.1 **VARIABLES IMPUTADAS**: Información sobre las variables imputadas en cada registro, así como su registro donante.

11.2 **REGISTROS DONANTES**: Listado de los registros donantes y del número de donaciones (a registros incorrectos) realizadas por cada uno.

12. **PERTENENCIA DE VARIABLES A REGLAS**: Información de las variables correspondientes a cada regla.

13. **EXCLUSIÓN**: Información referente al proceso de exclusión (si lo hubo).

14. **COMPARATIVA EN RESUMEN ESTADÍSTICO DE VARIABLES**: Comparativa pre/post imputación de las distribuciones de frecuencia de las variables.

15. **LISTADO SOBRE REGISTROS**: Listado sobre registros excluidos, incorrectos y warning.

16. **LISTADO SOBRE VARIABLES/REGLAS**. Listado de los errores de rango, filtro o reglas para cada variable o regla.

17. **IMPUTACIÓN SOBRE VARIABLES**: Información sobre la imputación realizada sobre cada variable.

18. **CORRESPONDENCIA ENTRE NOMBRE E ÍNDICES DE VARIABLES**: Correspondencia entre nombre e índices de variables.

En esta etapa se analiza la realización de las tareas de ejecución y los datos obtenidos con vistas a decidir si los resultados tienen un nivel de calidad aceptable. Se analizan y evalúan el nivel de no respuesta, los errores producidos durante la grabación, el nivel y distribución de los errores detectados y las imputaciones realizadas.

5. Experiencia Computacional

La herramienta TEIDE2 ha sido desarrollada usando Qt 5.3 (http://qt-project.org/) como librería de funciones gráficas y GCC 4.9 como compilador para generar

ejecutables a partir del código fuente, que ha sido escrito en C++. De esta forma, a partir del código fuente es posible generar ejecutables para diversas plataformas (32 o 64 bits, Windows o Linux, etc). En esta sección mostramos los resultados computacionales de usar TEIDE2 para depurar algunas encuestas procedentes de datos reales cedidos por algunos institutos de estadística autonómicos. El ordenador usado para obtener estas experiencias es un ordenador personal con procesador Intel Core Duo 2 3,34 Ghz y 4 GB de memoria RAM dotado de tres sistemas operativos: Linux Ubuntu 13.10 a 32 bits, Linux Ubuntu 13.10 a 64 bits, y Microsoft Windows 7 a 64 bits.

Para ilustrar el comportamiento de TEIDE2 sobre un caso práctico real hemos usado una encuesta anonimizada amablemente cedida por el Instituto Canario de Estadística (ISTAC). Se trata de la Encuesta sobre Gasto Turístico (EGT), que tiene como objetivo conocer el gasto realizado por los turistas que visitan Canarias, su perfil sociodemográfico, así como las características generales de su viaje, permitiendo la comparación de la situación turística entre los principales municipios de las islas. Consiste en 235 variables, de las que 29 son de tipo texto, siendo imputables las otras 206 variables. Cada una de estas 206 variables tiene su rango, y 150 tienen una condición que define un filtro. Adicionalmente hay 57 reglas de depuración explícitas. La encuesta es trimestral, pero para poder demostrar mejor la capacidad de TEIDE2 hemos unido los datos por anualidad, aumentando con ello el tamaño de la muestra de datos recogidos. Disponiendo de los datos en los años 2011 y 2012, hemos creado dos escenarios numéricos sobre los que evaluar el comportamiento de TEIDE2. El primer escenario (EGT 2011) reúne 36556 registros, y el segundo escenario (EGT 2012) reúne 35488 registros. Ambos escenarios están almacenados en bases de datos en formato XML.

Tabla 1

Tiempos de cálculo

	TCD		*TSRF*		*TST*		*TSI*	
	EGT 2011	*EGT 2012*	*EGT 2011*	*EGT 2012*	*EGT 2011*	*EGT 2012*	*EGT 2011*	*EGT 2012*
Windows 7 MinGW 32 bits	155 seg.	147 seg.	31 seg.	31 seg.	64 seg.	61 seg.	20 seg.	7 seg.
Windows 7 MinGW 64 bits	151 seg.	141 seg.	27 seg.	26 seg.	55 seg.	53 seg.	17 sg.	4 seg.
Windows 7 VS2012 32 bits	135 seg.	129 seg.	33 seg.	32 seg.	61 seg.	53 seg.	16 seg.	6 seg.
Windows 7 VS2012 64 bits	120 seg.	115 seg.	30 seg.	28 seg.	50 seg	49 seg.	14 seg.	4 seg.
Ubuntu 13.10 GCC 32 bits	128 seg.	123seg.	259seg.	253seg.	390seg.	183seg.	17 seg.	8 seg.
Ubuntu 13.10 GCC 64 bits	117 seg.	112 seg.	168 seg.	168seg.	241 seg.	265seg.	17 sg.	7 seg.

La tabla 1 muestra tiempos de cálculo (en segundos) de diversas fases de TEIDE2 según el compilador usado para su generación y el sistema operativo del ordenador. Las fases son cuatro y se disponen por columnas en la tabla, mientras que las combinaciones de compilador y sistema operativo son seis y se disponen por filas. Las cuatro fases son:

* Carga de datos. TCD es el tiempo que se ha utilizado en la lectura de los datos y volcado de la información en tablas.

* Proceso de rangos y filtros. TSRF es el tiempo empleado en la comprobación de posibles errores de rango o filtro.

* Proceso en reglas explícitas. TST es el tiempo consumido en la comprobación de errores en las reglas de depuración que no son ni de rango ni de filtro.

* Imputación. TSI es el tiempo empleado en la corrección de los registros erróneos.

Las seis combinaciones se han generado al considerar las opciones 32 bits y 64 bits, con dos compiladores bajo el sistema Windows (el comercial Microsoft Visual Studio (VS) 2012, y el gratuito MinGW) y con el gratuito que viene en Ubuntu (GNU GCC). Cada celda de la tabla muestra los tiempos sobre las dos encuestas (EGT 2011 y EGT 2012). Según la tabla 1, TEIDE2 se comporta mejor cuando es generado con Visual Studio y ejecutado sobre Windows 64 bits. No obstante la ejecución bajo Ubuntu implica tiempos computacionales también razonables teniendo en cuenta las dimensiones de nuestras encuestas.

Tabla 2

Dimensión de las dos encuestas analizadas

	Registros correctos	*Registros incorrectos*	*Registros corregidos*	*Registros no corregidos*	*Registros warning*
EGT 2011	36.527	29	29	0	2
EGT 2012	35.464	24	24	0	0

La tabla 2 resume los resultados generados por TEIDE2 sobre los dos escenarios. De esta tabla se observa que TEIDE2 ha corregido todos los registros erróneos, destacando sólo dos en la encuesta EGT 2011 que clasificó como "warning" porque para corregirlos tuvo que modificar bastantes variables. En la práctica ambos registros deberían ser examinados por un experto del instituto de estadística, por si cree oportuno usarlos así, o modificarlos con algún criterio, o contactar con el proveedor de esos registros para volver a realizarle la encuesta.

Tabla 3

Descripción de los resultados sobre las dos encuestas consideradas

	EGT 2011	*EGT 2012*
Promedio de variables imputadas por registro sobre el total de imputados	2,55	4,00
Promedio de variables imputadas por registro sin considerar los registros warning	1,00	4,00
Promedio errores en rango por registro	0,03	0,00
Promedio de variables involucradas en reglas explícitas no cumplidas por registro incorrecto	2,55	5,75
Promedio de variables involucradas en reglas no cumplidas por registro incorrecto	2,59	5,75
Promedio de variables involucradas en componentes conexas por registro incorrecto	116,41	113,25
Promedio de distancias a registros donantes	0,00	0,00

TEIDE2 genera adicionalmente un informe final para destacar características sobre la calidad de la imputación. La tabla 3 muestra algunas de estas características. Desde esta tabla observamos que el número medio de variables con un valor erróneo de rango o filtro es 2,55 para "EGT 2011" y de 5,75 para "EGT 2012". Después de la fase de imputación, el número medio de modificaciones en un registro erróneo fue de 2,55 para "EGT 2011" y de 5.75 para "EGT 2012". Dado que las variables con valor erróneo necesitan ser modificadas, estas cifras indican que TEIDE2 no tuvo que cambiar variables adicionales para obtener registros válidos, incluso teniendo en cuenta las otras reglas de depuración. Además también es relevante observar que las componentes conexas en ambas encuestas contiene una media bastante alta, lo que significa que la imputación en ambas encuestas es complicada.

Cerramos la sesión observando que estos experimentos se han realizado sobre microdatos que previamente han sido depurados parcialmente por TEIDE2 en varias ocasiones. De hecho, la forma natural de usar una herramienta de depuración no es aplicarla únicamente cuando estén disponibles todos los registros, sino repetidamente sobre microdatos parciales. De esta forma se detectan errores en la interpretación del cuestionario, en la fase de grabación o en la redacción de la reglas de depuración que se pueden ir corrigiendo durante el proceso de recogida de datos. Es habitual que los primeros registros recogidos tengan más errores que los últimos gracias a los errores que se detectan con herramientas como TEIDE2. Al ir evaluado trozos de la muestra de datos recogidos según llegan al instituto, aún sin imputar ningún dato en sus registros incorrectos, la muestra final resulta de mucha mayor calidad. De hecho, en la práctica, la ejecución paso a paso disponible en TEIDE2 es altamente útil ya que permite entender lo que sucede en la fase de depuración sobre los primeros registros que se reciben. Las encuestas usadas en la experiencia descrita en esta sección representan bien

otras encuestas reales sobre las que hemos trabajado, y que en su fase final tienen comportamientos análogos.

6. Conclusión

Este artículo describe una herramienta para la edición e imputación de datos estadísticos, llamada TEIDE2, desarrollada bajo la metodología de software abierto y gratuito, y el código fuente está disponible en el repositorio público https://github.com/Teide2/teide2 donde además puede encontrarse documentación, modo de instalación, blogs de uso y debate, etc.

Organismos interesados en depurar datos pueden integrar TEIDE2 en sus sistemas de gestión de datos, y mejorarlo con nuevas funcionalidades. Es multiplataforma, lo que favorece que se pueda compilar y ejecutar prácticamente en cualquier dispositivo.

Realiza una edición e imputación de datos automática a través de una interfaz gráfica y amigable. Todos los procesos de cómputo permiten cambiar parámetros que viene pre-establecidos. De esta forma se puede experimentar con diversas opciones y optar por la imputación más razonable entre un conjunto de opciones. Por ejemplo, a modo de simulador, es posible analizar el impacto en la depuración de aumentar o disminuir el peso de unas variables. De hecho, es altamente aconsejable que el técnico experimente con diferentes valores para los parámetros que ofrece TEIDE2 porque sus resultados son muy sensibles a ellos. Adicionalmente, al disponer del código fuente, es posible inserir nuevas funcionalidades para experimentar empíricamente con ellas, y ofrecer nuevas metodologías de imputación a la comunidad de usuarios. Este artículo no pretende ser el manual exhaustivo de TEIDE2, sino un resumen de sus características principales.

TEIDE2 se ha utilizado y utiliza para depurar encuestas como la "Encuesta de Ingresos y Condiciones de Vida de los Hogares Canarios", la "Encuesta sobre la Implantación de Tecnologías de la Información y la Comunicación en los Hogares Canarios", la "Encuesta sobre la Implantación y Uso de Tecnologías de la Información y la Comunicación en las Empresas Canarias", y la "Encuesta sobre Gasto Turístico". Este artículo muestra el comportamiento de TEIDE2 sobre esta última encuesta con datos del 2011 y 2012.

Referencias

Arbués, I., González, M., Revilla, P., (2009), «Selective editing as a stochastic optimization problem», *Boletín de Estadística e Investigación Operativa*, Vol. 25, 32-41.

Arbués, I., González, M., Revilla, P., (2010), «A Class of stochastic optimization problems with application to selective data editing», INE documento de trabajo 02/2010.

De Waal, T., Pannekoek,J., Scholtus, S. (2011). «Handbook of Statistical Data Editing and Imputation». Wiley, Hoboken.

Delgado-Quintero, S., Salazar-González, J.J., (2008), «A new approach for Data Editing and Imputation», *Mathematical Methods of Operations Research*, Vol. 68, 407-428.

Fellegi, I.P., Holt D., (1976), «A Systematic Approach to Automatic Edit and Imputation», *Journal of the American Statistical Association*, Vol. 71, 17-35.

Gómez Alonso, J. M., (1980), «Un enfoque sistemático de la edición e imputación automáticas», *Estadística Española*, Vol. 88, 33-80.

Riera-Ledesma, J., Salazar-González, J.J., (2007a), «A branch-and-cut algorithm for the Error Location Problem in Data Cleaning", *Computers & Operations Research*, Vol. 34, 2790- 2804.

Riera-Ledesma, J., Salazar-González, J.J., (2007b), «A Heuristic Approach for the Continuous Error Localization Problem in Data Cleaning», *Computers & Operations Research*, Vol. 34, 2370-2383.

Villán Criado, I. (1992), «Análisis de reglas de depuración de datos», *Estadística Española*, Vol. 34, 151-171.