MÉTODOS AUTOMÁTICOS DE FUSIÓN DE REGISTROS Y SU UTILIZACIÓN EN EUSTAT

EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADISTICA

Donostia-San Sebastián, 1
01010 VITORIA-GASTEIZ
Tel.: 945 01 75 00
Fax.: 945 01 75 01
E-mail: eustat@eustat.es
www.eustat.es

Presentación

En esta publicación se presenta el trabajo desarrollado en los últimos años en relación a la aplicación de nuevas metodologías a la producción estadística, en materia de fusión de registros.

El trabajo realizado se enmarca en una de las operaciones del Plan Vasco de Estadística 2005-2008, relativa a I+D+i de métodos estadísticos, y cuya finalidad se dirige a investigar y aplicar nuevas metodologías estadístico-matemáticas en las operaciones estadísticas. El espíritu que anima estos estudios se encuadran en una gestión de la excelencia, que está definido por un proceso de mejora continua.

La creciente demanda de información estadística, los recursos limitados de las oficinas estadísticas y el propósito de evitar una excesiva carga de respuesta a los encuestados, hace necesario el uso eficiente de la información proveniente de los censos y encuestas, así como de la utilización de información administrativa.

La aplicación de métodos automáticos de fusión de registros comenzó en EUSTAT en los años 90. Primeramente, se aplicaron métodos deterministas, y desde el año 2002 se han venido desarrollado los métodos probabilísticos, los cuales son el objeto principal de esta publicación.

Espero que esta publicación sea de utilidad para todos los interesados en este ámbito de la estadística.

Vitoria-Gasteiz, marzo 2007
Josu Iradi Arrieta
Director General.

MÉTODOS AUTOMÁTICOS DE FUSIÓN DE REGISTROS Y SU UTILIZACIÓN EN EUSTAT.

RESUMEN

La fusión de registros se refiere a la tarea de identificar registros que corresponden a la misma entidad en dos o más archivos distintos cuando no se dispone de identificadores únicos. En EUSTAT[1] se ha trabajado con procedimientos de fusión de forma continuada, pero hace algún tiempo se comenzaron a estudiar nuevas técnicas, como la fusión de registros probabilística que se describe en este cuaderno técnico. Lo primero que se hace en este documento es una breve introducción al concepto de fusión de registros y se establece la situación de EUSTAT cuando empezó el estudio de esta nueva técnica. A continuación, se describe la metodología utilizada, que fundamentalmente se basa en el artículo "*A theory for Record Linkage*" de Ivan P. Fellegi y Alan B. Sunter. Posteriormente se detalla el programa informático que se ha desarrollado en EUSTAT para llevar a cabo la fusión basada en el método probabilístico de forma automática, siguiendo las directrices marcadas por la metodología anteriormente mencionada. Una vez visto el programa se examinan algunas de las aplicaciones de fusión que se han realizado en el Instituto. Por último, se ofrecen las conclusiones extraídas tras el estudio.

[1] EUSTAT quiere agradecer el excelente trabajo de investigación desarrollado inicialmente por Leire Legarreta y posteriormente por Laura Otero, en el marco de las becas de investigación y metodología estadístico-matemática que promueve EUSTAT

Índice

Capítulo 1

Introducción

Introducción y objetivos

La fusión de registros se define como el uso de técnicas algorítmicas con el objetivo de identificar pares de registros procedentes de dos archivos distintos cuando no se dispone de identificadores únicos.

Históricamente, esta tarea se encargaba a personal administrativo, quien tenía que revisar manualmente las listas de registros, obtener información adicional cuando ésta estaba ausente o era contradictoria, y tomar las decisiones de fusión según unas reglas previamente establecidas. Generalmente, estas listas se ordenaban por nombre o por alguna de las características de la dirección para facilitar el proceso. Si algún nombre contenía una variación tipográfica inusual, podía no encontrarse su correspondencia. Si los ficheros eran muy extensos, las listas estaban divididas en varias hojas, así que a veces algunos pares se perdían. A pesar de un entrenamiento exhaustivo, las decisiones que se tomaban no eran siempre consistentes. Todo ello motivó el estudio de técnicas computacionales para llevar a cabo esta tarea de fusión de forma automática.

Hay varias técnicas para realizar este trabajo. Principalmente se pueden clasificar en dos grupos: métodos deterministas y métodos probabilísticos.

La **fusión de registros determinista** busca acuerdos y desacuerdos exactos sobre una o más variables de fusión de dos o más ficheros. Por ejemplo, se puede simplemente usar el campo DNI que sea común a dos ficheros. Sin embargo, los errores de codificación producidos al registrar esta variable en cualquiera de los ficheros hacen que se pierdan algunos de los *matches* verdaderos (*match*: un par de comparación de dos registros en diferentes ficheros de la misma persona).

La **fusión de registros probabilística** usa la información de un número mayor de variables, y tiene en cuenta toda la información proporcionada por cualquiera de los acuerdos o desacuerdos sobre las variables de fusión. A pesar de tener en cuenta todas las variables, cada una de ellas tendrá un poder decisivo diferente. Por ejemplo, el acuerdo sobre el número de la seguridad social es más indicativo de ser un *match* que un acuerdo sobre el sexo. También, acuerdos sobre valores raros de una variable de fusión dada (por ejemplo, el apellido Galzarsoro) son más indicativos que un acuerdo sobre valores comunes (por ejemplo, García).

El objetivo principal de esta publicación es presentar el trabajo realizado en Eustat sobre las técnicas de fusión de registros basadas en métodos probabilísticos. En el Instituto se decidió estudiar nuevos métodos de fusión de registros para tratar de mejorar algunos de los resultados obtenidos mediante los métodos deterministas que se estaban utilizando hasta ese momento. Partiendo del trabajo desarrollado por Fellegi y Sunter se ha diseñado un programa de fusión programado en lenguaje SAS que permite la fusión de dos ficheros utilizando técnicas probabilísticas.

En este cuaderno se describe en primer lugar la metodología que se ha seguido para el posterior desarrollo del programa de fusión. También se detalla dicho programa, haciendo un repaso por cada una de las macros que lo forman y se explicará cuáles son las funciones de cada una de ellas y qué posición ocupan dentro del programa de fusión. Posteriormente se enumeran algunas de las aplicaciones de dicho programa que se han llevado a cabo en el Instituto y se reflejan los resultados obtenidos. Por último, se expondrán las conclusiones a las que ha llegado EUSTAT sobre este tipo de procedimientos probabilísticos.

Descripción del Proyecto

Para conseguir estos objetivos se comenzó haciendo un estudio exhaustivo de la documentación disponible acerca de distintos métodos de fusión de registros, principalmente, métodos probabilísticos. Una vez que se tuvo suficiente información al respecto, se decidió llevarla a la práctica y para ello se desarrollaron una serie de macros en lenguaje SAS que ejecutasen de manera automática la fusión entre dos ficheros con registros correspondientes a individuos.

Una vez desarrollado el programa, se ejecutó con diversos ficheros para estudiar su fiabilidad y poder comparar los resultados obtenidos mediante este método con aquellos resultados que se obtenían utilizando procedimientos basados en técnicas deterministas que fijaban a las variables unos pesos concretos.

Antecedentes

Antes de comenzar el estudio de métodos probabilísticos, en EUSTAT ya se llevaban a cabo procedimientos de fusión basados en técnicas deterministas que distinguían la entidad de las variables asignándoles a cada una un peso distinto.

Una de las aplicaciones de este procedimiento de fusión se llevó a cabo con el objetivo de asociar a los registros del Padrón Municipal de Habitantes las claves de personas (Unidad de Población Básica) que están en el Registro Estadístico de Población y las claves de territorio (Unidad Territorial de Emplazamiento) de la base de datos del Territorio.

Para los ficheros de Movimientos Padronales de Personas sólo se asocian las claves de personas. Este proceso se realiza en varios pasos y módulos, dependiendo del fichero que se trate.

Se describen los distintos módulos:

1) **Fusión más estricta.**

   Para el caso del Padrón Municipal de Habitantes, se intenta buscar las claves de personas y las claves de territorio en el Registro Estadístico de Población y Territorio, respectivamente. La clave de Unidad Territorial de Emplazamiento se intenta recuperar primero por medio de la clave de Unidad Poblacional Básica.

Se asociarán los registros del Padrón Municipal de Habitantes con los registros de las tablas del Registro Estadístico de Población que contengan los datos de fecha de nacimiento, lugar de nacimiento, sexo, dirección postal, nombre, apellidos y DNI, mediante los siguientes criterios:

- Igualdad en DNI. Se utiliza el DNI completo y sin letra. Se excluyen los de frecuencia mayor para evitar que aparezcan muchos duplicados, ya que éstos se supone que son DNI incorrectos.

- Igualdad en nombre y apellidos. Previamente se normalizan el nombre y los apellidos y se les aplica una función de generación de alfaclaves. Estas alfaclaves son literales equivalentes que facilitan el tratamiento de dichas variables. Para la asociación se utilizan las tres alfaclaves simultáneamente.

- Igualdad en dirección postal, una vez que se ha comprobado la igualdad en la fecha de nacimiento y en el sexo.

Cada registro de la tabla de Movimientos Padronales de Personas hay que relacionarlo con el Registro Estadístico de Población mediante la clave de Unidad Poblacional Básica. El proceso a realizar es el mismo que el utilizado para el Padrón Municipal de Habitantes, sin tener en cuenta las comparaciones con la dirección postal ya que ésta no viene en el fichero de Movimientos Padronales de Personas.

Posteriormente se calculan los pesos, que mediante comparaciones sobre las distintas variables va dando puntuaciones a las coincidencias para quedarse luego con las de mayor valor.

2) Fusión menos estricta.

Dado que algunas relaciones válidas podrían escapar por ser los primeros criterios de fusión demasiado estrictos, se decidió que se flexibilizaran algunas de estas condiciones.

El primer ajuste fue que, para los registros no fusionados con el primer proceso, se buscaran igualdades por combinaciones del nombre y de los dos apellidos: nombre y primer apellido, nombre y segundo apellido, primer y segundo apellido. Los pesos que se utilizan en este proceso son los mismos que en el módulo anterior.

A los que quedan sin fusionar después de ese ajuste, se les aplica otro criterio mediante el cual se buscaban igualdades por dos de los tres elementos en su fecha de nacimiento: día y mes, día y año, mes y año.

Con estas fusiones menos estrictas, ocurría que a veces se tomaban por válidas relaciones que tenían una puntuación total superior al límite, pero debido a la coincidencia de elementos aislados. Por tanto, se podían dar dos situaciones erróneas: por un lado, que sea la misma persona, pero que tenga distinta clave, y por otra, que sean distintas personas con la misma clave. Se está trabajando en la mejora de estos casos por la vía de la normalización de nombres y apellidos.

Capítulo 2

Metodología

El modelo teórico en el que se basa el programa diseñado en Eustat para la fusión probabilística de ficheros es el presentado por Fellegi y Sunter en el artículo "A *theory for Record Linkage*" [1] en 1969. Las bases de este modelo son las siguientes.

Modelo teórico

Los métodos probabilísticos de fusión de registros llevan a cabo la comparación de registros de dos archivos distintos. Se denotan por A y B los archivos a ser comparados y por a y b sus elementos, respectivamente.

Se supone que ambos archivos tienen elementos comunes, y por tanto, el objetivo de la fusión es reconocer, de entre todos los pares de AxB que podrían formarse, aquellos que se refieran a la misma persona, objeto o entidad. Es decir, el objetivo es dividir el conjunto

$$AxB = \{(a, b) \mid a \in A, b \in B\}$$

en la unión de los conjuntos disjuntos

$$M = \{(a, b) \mid a = b, a \in A, b \in B\} \quad (1)$$

y

$$U = \{(a, b) \mid a \neq b, a \in A, b \in B\} \quad (2)$$

a los que se llaman conjunto de *matches* y *no-matches* respectivamente.

Cada unidad de la población tiene unas características asociadas, tales como, nombre, apellidos, edad o dirección. Se han de identificar aquellos registros que se refieren a una misma persona, objeto o entidad. Sin embargo, el proceso de creación de los ficheros podría introducir errores o imprecisiones (errores de codificación, trascripción y tecleo, variaciones tipográficas o fonéticas, pérdida de datos, etc.) en los registros generados. Como resultado de estos errores, dos miembros de A y B que no se refieren al mismo individuo podrían generan registros idénticos y, más frecuentemente, dos miembros idénticos de A y B podrían producir registros distintos. Se denotan los registros correspondientes a los miembros de A y B por $\alpha(a)$ y $\beta(b)$ respectivamente.

También se supone que los individuos que han sido registrados provienen de muestras aleatorias seleccionadas de A y de B que se denotan por As y Bs. No se excluye la posibilidad de que As = A y Bs = B. De ahora en adelante, para simplificar la notación, se elimina el subíndice s.

El primer paso al intentar emparejar registros de dos archivos es compararlos. El resultado de la comparación es un conjunto de códigos, que son codificados en afirmaciones del tipo: “el nombre coincide en ambos registros”, “el nombre coincide y es Pedro”, “el nombre no coincide”, “el nombre está en blanco en uno de los registros” o “existe acuerdo en la zona de la ciudad de la dirección pero no en la calle”. Formalmente, se define el vector comparación como un vector función de los registros α(a) y β(b) en la forma:

$$\gamma[\alpha(a), \beta(b)] = \{\gamma^1[\alpha(a), \beta(b)], \ldots, \gamma^k[\alpha(a), \beta(b)]\} \quad (3)$$

Se ve que γ es una función definida sobre AxB. Se puede escribir γ(a, b), γ(α, β) o simplemente γ. El conjunto de todas las posibles realizaciones de γ se llama *espacio de comparación* y se denota por Γ.

Durante el proceso de la operación de fusión, se observa γ(a, b) y se tiene que decidir si (a, b) es un emparejamiento, (a, b) Є M (se llama a esta decisión *link* y se denota por A1) o si es un no-emparejamiento, (a, b) Є U (se llama a esta decisión *no-link* y se denota por A3). Sin embargo, pueden existir situaciones para las cuales sea imposible tomar una de estas dos decisiones para niveles específicos de error, por lo que se permite una tercera decisión, denotada por A2, a la que se llama *posible link*.

En estas condiciones, se define una *regla de fusión* L como una aplicación del espacio de comparación Γ sobre el conjunto de funciones de decisión aleatorias D = {d(γ)} donde

$$d(\gamma) = \{P(A1|\gamma), P(A2|\gamma), P(A3|\gamma)\};\ \gamma \in \Gamma \quad (4)$$

y

$$\sum_{i=1}^{3} P(Ai \mid \gamma) = 1 \quad (5)$$

En otras palabras, para cada valor observado de γ, la regla de fusión asigna las probabilidades de tomar cada una de las tres posibles decisiones.

Hay que considerar los niveles de error asociados con cada regla de fusión. Se asume que un par de registros [α(a), β(b)] es seleccionado aleatoriamente para ser comparado de acuerdo con algún proceso probabilístico. El vector de comparación resultante γ[α(a), β(b)] es por tanto una variable aleatoria. Se denota la probabilidad condicional de que se dé γ dado que (a, b) Є M como m(γ), y será:

$$m(\gamma) = P\{\gamma[\alpha(a), \beta(b)] \mid (a,b) \in M\} = \sum_{(a,b)\in M} P\{\gamma[\alpha(a), \beta(b)]\}.P[(a,b) \mid M] \quad (6)$$

De manera similar, se denota la probabilidad condicional de γ dado que (a, b) Є U por u(γ). Por tanto,

$$u(\gamma) = P\{\gamma[\alpha(a), \beta(b)] \mid (a,b) \in U\} = \sum_{(a,b)\in U} P\{\gamma[\alpha(a), \beta(b)]\}.P[(a,b) \mid U] \quad (7)$$

Hay dos tipos de errores asociados a esta regla de fusión. El primero se da cuando al comparar pares de registros que no se corresponden con *matches* se decide asignarlos como *links* y tiene como probabilidad:

$$P(A1|U)=\sum_{\gamma\in\Gamma}u(\gamma).P(A1|\gamma) \quad (8)$$

El segundo tipo de error sucede cuando un *match* es comparado y es considerado como *no-link*, y tiene como probabilidad:

$$P(A3|M)=\sum_{\gamma\in\Gamma}m(\gamma).P(A3|\gamma) \quad (9)$$

Una regla de fusión en el espacio Γ se dice que es una regla de fusión a los niveles de error μ, λ (0<μ<1, 0<λ<1) y se denota por L(μ, λ, Γ) si

$$P(A1 \mid U) = \mu \quad (10)$$

Y

$$P(A3 \mid M) = \lambda \quad (11)$$

Se dice que la regla de fusión L(μ, λ, Γ) es óptima si la relación

$$P(A2 \mid L) \leq P(A2 \mid L') \quad (12)$$

se mantiene para cualquier L'(μ, λ, Γ ') entre todas las reglas de fusión que verifican las relaciones anteriores.

Se observa que, según esta definición, una regla de decisión óptima es aquella que maximiza las probabilidades de adoptar disposiciones de comparación positivas (es decir, decisiones A1 o A3) sujeta a unos niveles fijos de error. Esto parece una decisión razonable, dado que el adoptar una decisión A2 requerirá de costosas operaciones manuales de fusión. Además, por otro lado, parece que si la probabilidad de A2 no es pequeña, el proceso de fusión será de dudosa utilidad.

No es difícil darse cuenta de que para ciertas combinaciones de μ y λ, el conjunto de reglas de fusión que satisfacen (10) y (11) es vacío. Por tanto, se consideran tan sólo aquellas combinaciones de μ y λ para las cuales es posible satisfacer las ecuaciones (10) y (11) simultáneamente para algún conjunto D de funciones de decisión como el definido por (4) y (5). Cuando esto se cumple, se dice que (μ, λ) es un *par admisible de niveles de error*.

Los autores proponen una regla de fusión óptima. Para ello, empiezan por definir un orden único en el conjunto finito de todas las posibles realizaciones de γ de la siguiente manera: si algún valor de γ es tal que tanto m(γ) como u(γ) son iguales a cero, entonces la probabilidad de que se dé este γ es igual a cero, y por tanto no es necesario incluirlo en Γ. Siguiendo esto, se asigna un orden arbitrario para cualquier γ para el cual m(γ) > 0, pero u(γ) = 0.

El resto de los γ se ordena de manera que la correspondiente secuencia de m(γ) / u(γ) sea monótonamente decreciente. Cuando el valor de m(γ) / u(γ) sea el mismo para más de un γ, entonces el orden que se asigne será arbitrario.

Se indica el conjunto ordenado {γ} por el subíndice i; (i = 1, 2,…, N_Γ), donde N_Γ es el número de puntos de Γ y se escribe ui = u(γi); mi = m(γi).

Sean (μ, λ) un par admisible de niveles de error de manera que no sean ambos demasiado grandes. Sean n, n' enteros tales que

$$\mu = \sum_{i=1}^{n} u_i \quad \lambda = \sum_{i=n'}^{N_\Gamma} m_i \quad 0 < n \leq n' < N_\Gamma$$

Se define

$$T_\mu = \frac{m(\gamma_n)}{u(\gamma_n)}$$
$$T_\lambda = \frac{m(\gamma_n')}{u(\gamma_n')} \qquad (13)$$

Entonces Fellegi y Sunter demostraron que la mejor regla de fusión a los niveles de error (μ, λ) venía dada por la forma:

$$d(\gamma) = \begin{cases} (1,0,0) & si\, T_\mu \leq m(\gamma)/u(\gamma) \\ (0,1,0) & si\, T_\lambda < m(\gamma)/u(\gamma) < T_\mu \\ (0,0,1) & si\, m(\gamma)/u(\gamma) \leq T_\lambda \end{cases}$$

En muchas aplicaciones, podría ser posible tolerar niveles de error suficientemente altos para eliminar la posibilidad de la acción A2. En este caso, se considera n y n' o bien T_μ y T_λ de manera que el conjunto medio de γ en la forma anterior sea vacío. En otras palabras, cada par (a, b) es localizado bien en M o bien en U. De hecho, esta es la decisión que nosotros hemos adoptado en nuestro programa de fusión, de tal modo que establecemos un límite único $T_\mu = T_\lambda$.

Algunos supuestos de simplificación

En la práctica, el conjunto de los distintos valores que puede tomar γ puede ser tan largo que la estimación de las correspondientes probabilidades m(γ) y u(γ) se puede convertir en algo impracticable. Por tanto, se necesita hacer algunas suposiciones que simplifiquen la distribución de γ.

Se asume que las componentes de γ pueden re-ordenarse y agruparse de tal modo que $\gamma = \left(\gamma^1, \gamma^2, \cdots, \gamma^k\right)$ y las componentes vectoriales sean estadísticamente independientes todas con respecto a cada una de las distribuciones condicionales. Entonces,

$$m(\gamma) = m_1(\gamma^1) \cdot m_2(\gamma^2) \cdot \cdots \cdot m_k(\gamma^k) \qquad (14)$$

$$u(\gamma) = u_1(\gamma^1) \cdot u_2(\gamma^2) \cdot \cdots \cdot u_k(\gamma^k) \qquad (15)$$

donde m(γ) y u(γ) están definidas por (6) y (7) respectivamente y

$$m_i(\gamma^i) = P(\gamma^i \,|\, (a,b) \in M)$$
$$u_i(\gamma^i) = P(\gamma^i \,|\, (a,b) \in U)$$

Por simplicidad de notación se escribe $m(\gamma^i)$ y $u(\gamma^i)$ en lugar de las expresiones técnicamente más precisas $m_i(\gamma^i)$ y $u_i(\gamma^i)$. Como ejemplo, en una comparación de registros que correspondan a personas, γ^1 podría incluir toda las componentes de comparación que se refieran a los apellidos, γ^2 todas las componentes de comparación que se refieran a direcciones. Las propias componentes γ^1 y γ^2 son vectores en sí mismos; las subcomponentes de γ^2 pueden representar, por ejemplo, los resultados codificados de comparar las diferentes componentes de las direcciones (ciudad, calle, número, etc.). Si dos registros son un *match* (es decir, cuando en realidad representan la misma persona), podría ocurrir un desacuerdo en la configuración debido a errores. Nuestra asunción dice que los errores en el nombre, por ejemplo, son independientes de los errores en la dirección. Si dos registros son un *no-match* (es decir, cuando en realidad representan individuos diferentes) entonces la asunción es que un acuerdo accidental en el nombre, por ejemplo, es independiente de un acuerdo accidental en la dirección.

En otras palabras, se asume que γ^1, γ^2,…, γ^k están distribuidas de forma condicionalmente independiente. Se remarca que no se está asumiendo nada acerca de la distribución no condicional de γ.

Está claro que cualquier función monótona creciente de $m(\gamma)/u(\gamma)$ podría servir igualmente bien como test estadístico para el propósito de las reglas de fusión.

En particular será ventajoso usar el logaritmo de este ratio y definir

$$w^k(\gamma^k) = \log m(\gamma^k) - \log u(\gamma^k) \qquad (16)$$

Entonces se puede escribir

$$w(\gamma) = w^1 + w^2 + \cdots + w^k \qquad (17)$$

y usar $w(\gamma)$ como nuestro test estadístico entendiendo que si $u(\gamma)=0$ o $m(\gamma)=0$ entonces $w(\gamma) = +\infty$ (o $w(\gamma) = -\infty$) en el sentido en que $w(\gamma)$ es mayor (o menor) que cualquier número finito dado.

Se supone que γ^k puede tomar n_k configuraciones distintas $\gamma_1^k, \gamma_2^k, \cdots, \gamma_{n_k}^k$. Se define

$$w_j^k = \log m(\gamma_j^k) - \log u(\gamma_j^k) \qquad (18)$$

Es efectivo para la interpretación intuitiva del proceso de fusión que los pesos estén definidos como positivos para aquellas configuraciones para las cuales $m(\gamma_j^k) > u(\gamma_j^k)$, y como negativos para aquellas configuraciones para las cuales $m(\gamma_j^k) < u(\gamma_j^k)$, y esta propiedad está preservada por los pesos asociados a la configuración total γ.

El número total de configuraciones (es decir, el número de puntos γЄΓ) es obviamente $n_1 \cdot n_2 \cdot \cdots \cdot n_k$. Sin embargo, debido a la propiedad aditiva de los pesos definidos para las componentes, será suficiente con determinar $n_1 + n_2 + \cdots + n_k$ pesos. Se puede entonces determinar el peso asociado con cada γ empleando esta aditividad, reduciendo de este modo el cardinal de Γ de forma notable.

Cálculo de pesos

Uno de los puntos clave en el desarrollo del procedimiento de fusión es el cálculo de los pesos m(γ) y u(γ) para cada uno de los posibles γ que se pueden presentar. Existen varios métodos propuestos; de hecho, el propio trabajo de Fellegi y Sunter presenta dos maneras distintas de enfrentarse a esta cuestión. Para este trabajo se ha optado por un método basado en las frecuencias con las que cada una de las configuraciones de los datos registrados aparece en el conjunto de los archivos a fusionar.

Supóngase que una de las componentes de los registros de ambos archivos es el campo *apellido*. La comparación de apellidos de ambos registros dará como resultado una componente del vector comparación. Esta componente puede ser una simple comparación del tipo "*el apellido coincide*" o "*el apellido no coincide*" o "*el apellido no aparece en uno o ambos archivos*".

En cualquiera de los dos archivos el apellido puede ser registrado con algún error. Se asume que es posible confeccionar una lista con todas las realizaciones de los apellidos de ambos archivos libres de error y también con el número de individuos de los respectivos archivos que tienen cada uno de estos apellidos.

Sean las respectivas frecuencias en A y B las siguientes:

$$f_{A1}, f_{A2}, \cdots, f_{Am}; \quad \sum_{j=1}^{m} f_{Aj} = N_A$$

y

$$f_{B1}, f_{B2}, \cdots, f_{Bm}; \quad \sum_{j=1}^{m} f_{Bj} = N_B$$

Sean las correspondientes frecuencias en A∩B:

$$f_1, f_2, \cdots, f_m; \quad \sum_{j=1}^{m} f_j = N_{AB}$$

Para concretar,

- f_{A1} es la frecuencia con la cual un valor concreto de una de las variables aparece en el archivo A
- f_{B1} es la frecuencia con la cual dicho valor aparece en el archivo B
- f_1 es la frecuencia con la cual dicho valor aparece tanto en A como en B

Es necesaria la siguiente notación:

e_A o e_B	las probabilidades respectivas de que el valor de una variable se haya registrado erróneamente en cada uno de los dos archivos. (Se asume que la probabilidad de ser mal registrado es independiente de cada uno de los valores particulares).
e_{A0} o e_{B0}	las probabilidades respectivas de que el valor de una variable no haya sido registrado en cada uno de los dos archivos. (Se asume que la probabilidad de no ser registrado es independiente de cada uno de los valores particulares).
e_T	la probabilidad de que el valor de una variable aparezca de forma distinta en ambos archivos a pesar de estar bien registrado en ambos. (Esto puede ocurrir, por ejemplo, si ambos archivos fueron creados en momentos distintos y el individuo cambió de nombre).

Finalmente se asume que e_A y e_B son suficientemente pequeñas como para que la probabilidad de una coincidencia entre dos entradas idénticas, aunque erróneas, sea insignificante y que las probabilidades de estar mal registrado, no registrado y haber cambiado sean independientes unas de otras.

Se dan unas reglas para el cálculo de m y u correspondientes a las siguientes configuraciones de γ: la variable coincide y es la j-ésima listada, la variable no coincide, la variable está ausente en alguno de los archivos.

$$m(\text{la variable coincide y es la j-ésima listada}) =$$
$$\frac{f_j}{N_{AB}}(1-e_A)(1-e_B)(1-e_T)(1-e_{A0})(1-e_{B0}) = \tag{19}$$
$$\frac{f_j}{N_{AB}}(1-e_A-e_B-e_T-e_{A0}-e_{B0})$$

$$m(\text{la variable no coincide}) =$$
$$\left[1-(1-e_A)(1-e_B)(1-e_T)\right](1-e_{A0})(1-e_{B0}) = \tag{20}$$
$$e_A+e_B+e_T$$

$$m(\text{la variable está ausente en alguno de los archivos}) =$$
$$1-(1-e_{A0})(1-e_{B0}) = \tag{21}$$
$$e_{A0}+e_{B0}$$

$$u(\text{la variable coincide y es la j - ésima listada}) =$$

$$\frac{f_{Aj}}{N_A}\frac{f_{Bj}}{N_B}(1-e_A)(1-e_B)(1-e_T)(1-e_{A0})(1-e_{B0}) = \qquad (22)$$

$$\frac{f_{Aj}}{N_A}\frac{f_{Bj}}{N_B}(1-e_A-e_B-e_T-e_{A0}-e_{B0})$$

$$u(\text{la variable no coincide}) =$$

$$\left[1-(1-e_A)(1-e_B)(1-e_T)\sum_j \frac{f_{Aj}}{N_A}\frac{f_{Bj}}{N_B}\right](1-e_{A0})(1-e_{B0}) = \qquad (23)$$

$$\left[1-(1-e_A-e_B-e_T)\sum_j \frac{f_{Aj}}{N_A}\frac{f_{Bj}}{N_B}\right](1-e_{A0}-e_{B0})$$

$$u(\text{la variable está ausente en alguno de los archivos}) =$$

$$1-(1-e_{A0})(1-e_{B0}) = \qquad (24)$$

$$e_{A0}+e_{B0}$$

Las proporciones f_{Aj} / N_A, f_{Bj} / N_B, f_j / N pueden tomarse, en muchas aplicaciones, como si fueran la misma. Este sería el caso, por ejemplo, si los dos archivos se supone que han sido extraídos de la misma población. Estas frecuencias pueden estimarse directamente de los propios archivos.

Se recuerda que de acuerdo a (18) la contribución de la componente de una variable al peso total es log(m/u) y que las comparaciones con un peso mayor que un número concreto serán consideradas como *links*, mientras que aquellas cuyo peso está por debajo del número concreto serán consideradas como *no-links*. Está claro de (19-24) que un acuerdo sobre una variable producirá un peso positivo y que cuanto más raro sea el valor de la variable mayor será el peso. De forma análoga, un desacuerdo sobre una variable producirá un peso negativo que decrece con los errores e_A, e_B, e_T. Si el valor de la variable está ausente en alguno de los registros, el peso será cero.

Creación de listas estandarizadas

Para utilizar el método de cálculo de pesos descrito anteriormente es necesario disponer de un listado con todas las configuraciones libres de error de cada una de las variables que intervienen en el proceso. Esto no supone ningún problema para, por ejemplo, variables del tipo sexo, ya que se sabe que existen dos únicas configuraciones correctas y que cualquier valor que no corresponda con estos dos valores es erróneo.

En el caso de variables tipo nombre o apellido esta tarea no es tan simple. El número de errores tipográficos que presentan estas variables suele ser bastante elevado por lo cual antes de extraer un listado de las configuraciones se tienen que corregir aquéllas que presentan errores.

A continuación se describe un método que se ha desarrollado en EUSTAT y que a partir de todas las configuraciones de una de estas variables tipo nombre o tipo apellido intenta relacionar aquellas configuraciones que provienen de la misma configuración original pero que son distintas porque dicha configuración puede tener más de una posible forma correcta o porque se ha producido un error en alguna de ellas.

La idea es la siguiente: Se parte de un listado original C con todas las configuraciones de una variable que aparecen en ambos ficheros a fusionar. El objetivo es generar un listado de códigos D y una aplicación a la que se denomina *est*, de C en D, de manera que dos elementos de C que equivalgan a un mismo valor tengan la misma imagen por *est*. Como se ha comentado, dos elementos de C pueden tener la misma imagen, bien debido a que un mismo valor se puede representar mediante diferentes configuraciones o bien debido a que al registrar alguna de ellas, se haya producido un error.

Entonces este proceso consiste, por un lado, en relacionar aquellas configuraciones que, aún representando ambas el mismo valor de una variable, son diferentes aunque correctas. Ejemplos de esta situación son nombres como ENEKOITZ-ENEKOIZ, LEZURI-LEXURI o JAVIER-XABIER. Por tanto, se debe garantizar que la imagen por *est* de estas configuraciones sea la misma.

Para ello, se han seguido ciertos criterios de estandarización, que incluyen los siguientes:

- Se corrigen errores producidos por la ocurrencia del símbolo ¥ en lugar del carácter Ñ.

- Se eliminan guiones, puntos y comas.

- Se sustituyen por un missing aquellos valores del tipo NO TIENE, NO HAY, NO CONSTA, NO SE SABE, …

- Se elimina el acento o la diéresis de aquellas vocales que aparezcan con cualquier tipo de acento o diéresis.

- Se eliminan los números. (1, 2, 3, 4, 5, 6, 7, 8, 9)

- Se eliminan ciertos caracteres. (‘, ´, `, ¨, ^, *, “, º, ª, #, Ç, ¥)

- Se sustituyen los ceros por la letra ‘O’.

- Se eliminan partículas. (D, DE, DEL, DA, DI, DO, L, LA, LAS, EL, LOS, Y)

- Se traducen ciertos diminutivos.

- Se identifican caracteres o grupos de caracteres con una similitud fonética o gráfica.

- Se estudian pares de nombres y apellidos que coincidan en todas las letras excepto una, y se analiza si pueden provenir de la misma configuración. Estos se estudian aparte porque existen grafías que, a pesar de no tener una similitud fonética o gráfica evidente en algunos casos, pueden ser susceptibles de representar el mismo carácter.

Por otro lado, se debe detectar aquellas configuraciones para las cuales se ha producido algún tipo de error. Una vez estén identificadas, su imagen por la aplicación *est* será la misma que la de la configuración correcta de la cual procedan. Así, configuraciones como FERANNDO, FERNNDO y FERANANDO se corresponden bajo la aplicación *est* con el mismo código de FERNANDO.

El primer obstáculo que se encuentra al plantearse esta tarea es cómo detectar aquellas configuraciones en las que se ha producido algún error. Para llegar a deducciones lógicas se parte de las frecuencias con las que las configuraciones aparecen en los archivos a fusionar. Para cada c∈C, $freq_0(c)$ representará la frecuencia con la cual c aparece en ambos archivos.

Sea c una cadena de caracteres de longitud n que aparece en alguno de los ficheros a fusionar y ha sido, supuestamente, correctamente registrada.

Se denota por

e = Prob(carácter haya sido registrado correctamente)

1-e = Prob(carácter haya sido registrado erróneamente)

Dado que la cadena de caracteres c tiene n caracteres, la probabilidad de que dicha cadena de caracteres haya sido registrada correctamente (como se ha supuesto) será de:

P(0 errores | longitud(c) = n) = e^n.

Si el número de individuos de la población original cuya variable toma el valor c es freq(c), dado que $freq_0(c)$ representa el número de casos en los que se da la configuración c en los archivos que se están considerando, se tiene que:

freq(c) * P(0 errores | longitud(c) = n) = $freq_0(c)$

Por lo tanto, se puede estimar freq(c) como:

freq(c) = $freq_0(c)/e^n$

Análogamente, para cada configuración correcta de C se puede estimar el número total de aquellas configuraciones para las cuales se ha producido un error en su trascripción, dos errores, etc.

$$P(1\,\text{error}\,|\,\text{longitud(c)} = \text{n}) = \binom{n}{1} e^{n-1}(1-e) = n \cdot e^{n-1} \cdot (1-e)$$

Y de esta manera se puede estimar el número total de configuraciones que se han podido generar a partir de c con un error como:

$$freq_1(c) = P(1\,\text{error}\,|\,\text{longitud(c)} = \text{n}) \cdot\ freq(c) = \frac{freq(c) \cdot n \cdot (1-e) \cdot e^n}{e}$$

Del mismo modo se estima el número total de configuraciones que se han podido producir a partir de c con dos errores, como sigue:

$$P(2\ \text{errores}\,|\,\text{longitud(c)} = \text{n}) = \binom{n}{2} e^{n-2}(1-e)^2 = \frac{n\cdot(n-1)\cdot e^{n-2}\cdot(1-e)^2}{2}$$

$$freq_2(c) = P(2\ \text{errores}\,|\,\text{longitud(c)} = \text{n})\cdot freq(c) = \frac{freq(c)\cdot n\cdot(n-1)\cdot(1-e)^2\cdot e^n}{2\cdot e^2}$$

En esta situación el objetivo es relacionar aquellas configuraciones erróneas de una variable con las correctas, conocida la estimación que se ha realizado del número total de configuraciones erróneas que pueden aparecer. Aquellos casos con un número de errores mayor que 2 no son considerados, por suponer que la probabilidad de que se produzcan es muy pequeña. De todos modos, si se desea podrían tenerse en cuenta siguiendo el razonamiento hecho para los casos con 1 o 2 errores.

La idea de la que se parte es la de comparar las configuraciones que es posible que sean erróneas con aquellas que se pueden suponer correctas, y generar un listado en el que unas y otras aparezcan relacionadas en aquellos casos en los cuales el número de posibles errores que separan unas configuraciones de otras es menor o igual a 2.

Una idea intuitiva es suponer que aquellas configuraciones que se dan con altas frecuencias en los archivos son correctas, dado que es poco probable que se cometa el mismo error en gran cantidad de ocasiones. Por tanto, se fija un límite en función del cual se divide el conjunto C en los subconjuntos:

$C_1 = \{c \in C \mid freq(c) > lim\}$

$C_2 = \{c \in C \mid freq(c) \leq lim\}$

Se comparan los valores en C_2 con los de C_1 con el objeto de relacionarlos en aquellos casos en los cuales el número de errores que los separa es bajo. De este modo se obtiene un listado en el que cada individuo de C_2 puede aparecer relacionado con ninguno, uno o varios individuos de C_1 y a la inversa.

A continuación, de todas esta relaciones se seleccionan aquellas que finalmente se van a unir entre sí. Para ello, partiendo de los supuestos detallados anteriormente, el número total de configuraciones relacionadas con una dada nunca podrá superar el valor previamente estimado.

De este modo, se ha relacionado cada elemento de C_2 con uno o varios elementos de C_1 y a la inversa. De entre todo este conjunto de relaciones se ha de seleccionar un subconjunto final en el cual cada elemento de C_2 se encuentre relacionado a lo sumo con uno de C_1, y el número total de configuraciones relacionadas con una dada c de C_1 no sea mayor que $freq_1 + freq_2$.

$$freq_1 + freq_2 = \frac{freq(c)\cdot n\cdot(1-e)\cdot e^n}{e} + \frac{freq(c)\cdot n\cdot(n-1)\cdot(1-e)^2\cdot e^n}{2\cdot e^2}$$

Para realizar esta selección se fija el siguiente criterio: del listado de partida, se selecciona un subconjunto en el cual cada elemento de C_2 tan sólo aparezca relacionado con uno de C_1, dado que una configuración posiblemente errónea tan sólo puede proceder de una configuración correcta original. En caso de que una configuración del subconjunto C_2 aparezca relacionada con más de una de C_1, se selecciona aquella que aparezca con mayor frecuencia en los archivos, o bien aquella de la que diste menos número de errores.

Una vez se ha generado un listado de estas características, es probable que cada elemento de C_1 se encuentre relacionado con más de uno de C_2. Por tanto, de entre todos ellos, se seleccionan aquellos que son más probable que provengan del elemento de C_1, tras haberse cometido algún tipo de error, teniendo en cuenta que el número total de configuraciones relacionadas con una de éstas no puede ser superior a la estimación que se ha realizado.

Se repite el proceso hasta que todos los elementos de C_2 hayan sido relacionados con alguno de C_1 o bien todos los elementos de C_2 hayan sido relacionados con un número de configuraciones igual a la estimación realizada.

A parte de este método que se utiliza para tratar de corregir errores en las variables tipo nombre o tipo apellido, más adelante en el programa se hace otro tipo de consideraciones sobre estas variables para aquellos pares que no son fusionados en los primeros intentos y sobre los cuales se hace un estudio un poco más exhaustivo.

En primer lugar se comprueba si ocurre que uno de los nombres es diminutivo del otro considerando algunos casos concretos conocidos.

También se verifica, tanto para nombres como para apellidos, si en un archivo aparece un valor compuesto mientras que en el otro archivo aparece sólo una de las componentes.

Otro modo de establecer relación entre dos nombres o dos apellidos es mediante el comparador de cadenas de caracteres de Jaro. Éste calcula el número de caracteres comunes a las dos cadenas, la longitud de ambas y el número de trasposiciones para calcular una medida de similitud que va entre 0.0 y 1.0.

Las trasposiciones vienen determinadas por pares de caracteres comunes que están fuera de lugar. Entonces, el valor del comparador de cadenas de caracteres de Jaro para las cadenas α y β viene dado por la siguiente fórmula:

$$S_J = \frac{1}{3}\left(\frac{comunes}{longitud(\alpha)} + \frac{comunes}{longitud(\beta)} + \frac{comunes - trasposiciones}{comunes}\right)$$

Se comprueba que si las cadenas son idénticas (comunes = longitud(α) = longitud(β) y trasposiciones = 0), entonces $S_J = 1$.

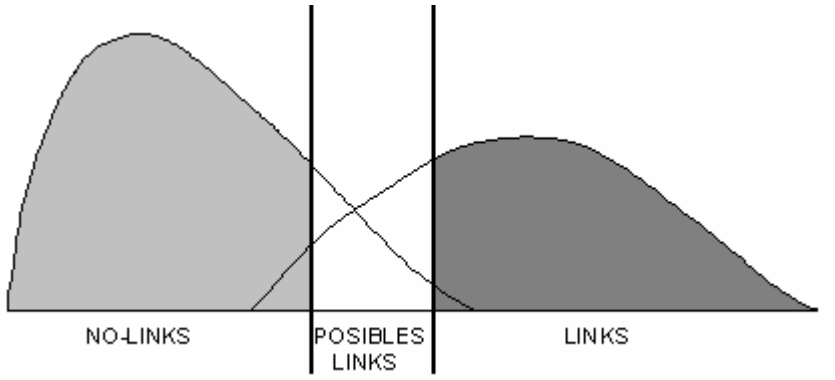
Otra consideración que se hace cuando se tiene como variables de fusión los dos apellidos del individuo es comprobar si éstos han sido intercambiados. Como se explica más adelante, al observar los resultados obtenidos tras una ejecución del programa de fusión, había ciertos pares de registros que correspondían al mismo individuo y que no se fusionaban puesto que los apellidos habían sido cambiados de orden.

El último contraste tiene que ver con una peculiaridad de nuestro ámbito. En la Comunidad Autónoma Vasca conviven dos lenguas oficiales, el castellano y el euskera. Por tanto, existen ciertas equivalencias onomásticas que pueden producir configuraciones distintas en dos ficheros para una misma ocurrencia. Lo que se hace es crear una base de datos con varias equivalencias de nombres castellano-euskera y para cada par de nombres se comprueba si están incluidos en esta lista. Si es así, se establece que hay relación entre ellos.

Después de todas estas apreciaciones, se recalculan los pesos y se seleccionan aquéllos que ahora tengan un peso superior al límite establecido por el usuario anteriormente.

Método para establecer el límite

Habiendo especificado todas las configuraciones relevantes γ_j^k y determinado sus pesos asociados $w_j^k; k = 1,2,\ldots,K;\ j = 1,2,\ldots,n_k$ falta establecer los valores límite T_μ y T_λ correspondientes a los μ y λ dados. Estos dos límites T_μ y T_λ dividen los pares de registros en tres zonas.



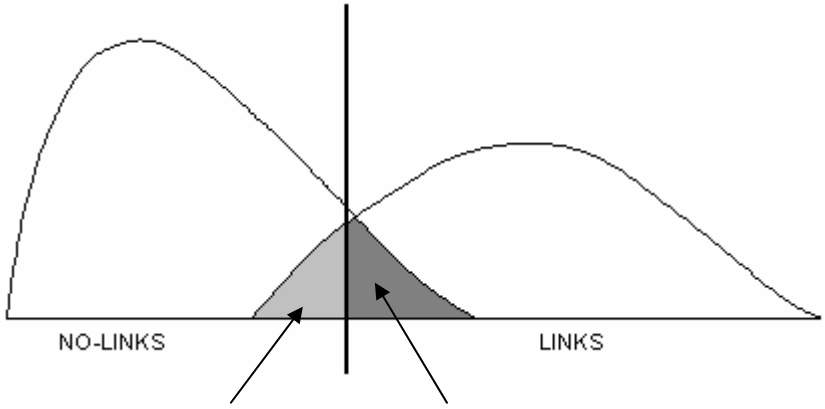
La zona a la izquierda del límite T_μ corresponde a los pares de registros que van a ser clasificados como *no-links*. En la zona a la derecha de T_λ están los pares de registros que se archivarán como *links*. Y la zona intermedia serán los pares de registros que se llaman *posibles links* y que necesitan una revisión manual para decidir en qué grupo clasificarlos.

Al tomar estas decisiones se está cometiendo dos tipos de errores. Uno, se comete al establecer como link dos registros que no pertenecen al mismo individuo. Y otro, al establecer como *no-link* un par de registros que en realidad corresponden al mismo individuo. Estos errores pueden cuantificarse de la forma:

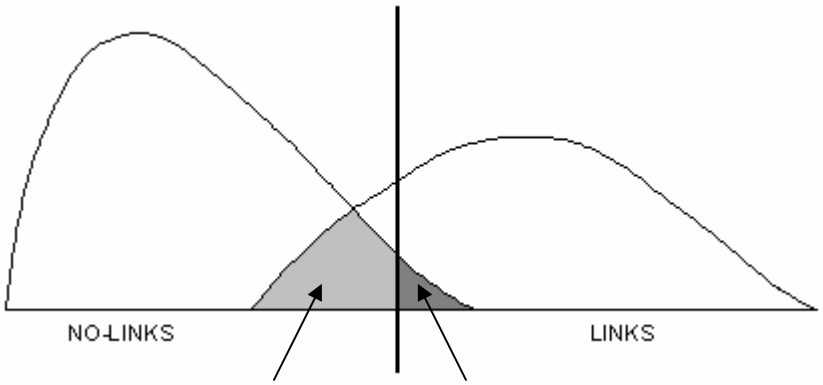
$\mu = P(A1 \mid U) = \sum_{\gamma \in A1} u(\gamma)$ Proporción de pares *link* en U.

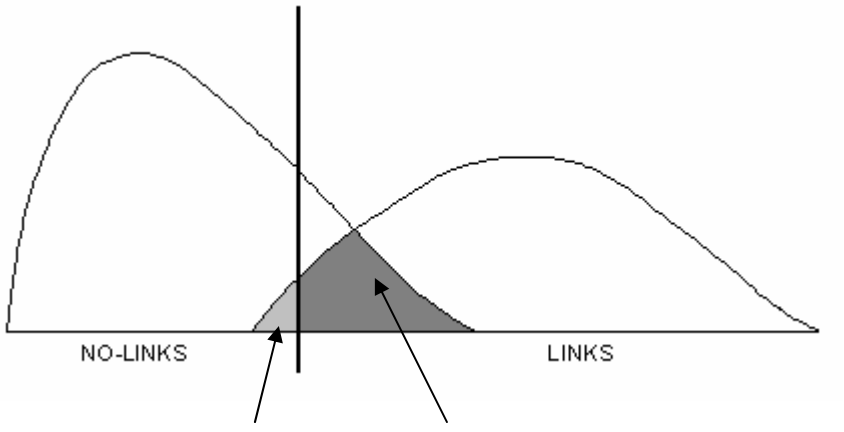
$\lambda = P(A3 \mid M) = \sum_{\gamma \in A3} m(\gamma)$ Proporción de pares *no-link* en M.

Para evitar el procesamiento manual de los pares de registros que se encuentran entre ambos límites, en EUSTAT se decidió tomar un solo límite $T = T_\mu = T_\lambda$.



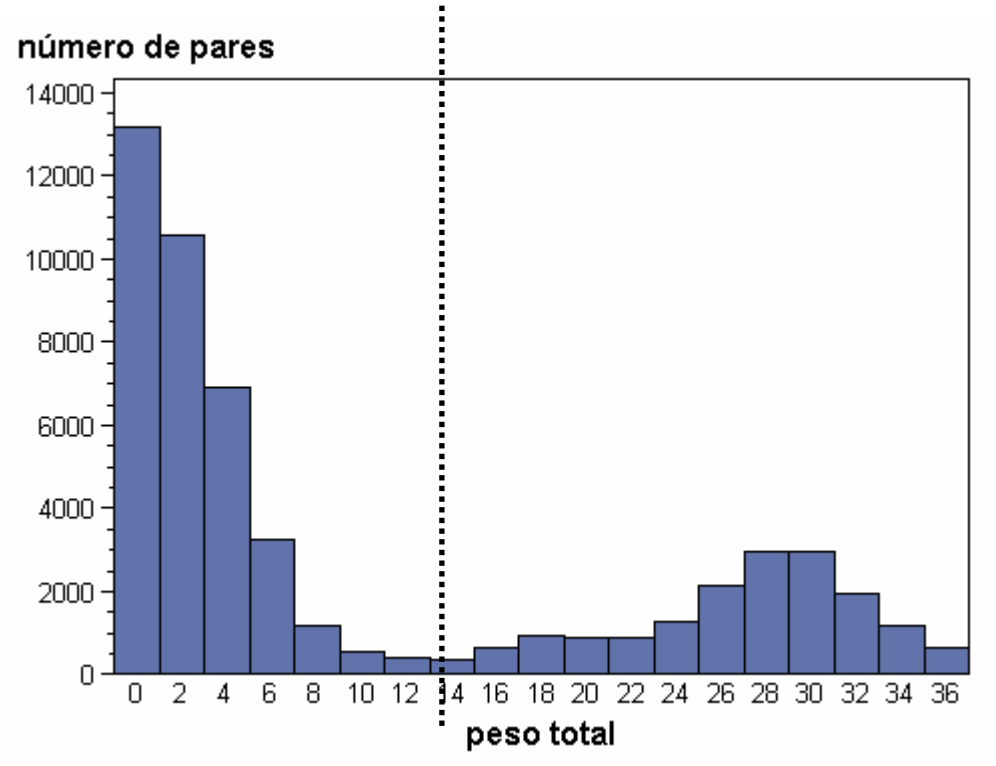
Igualando ambos límites, a medida que aumente el valor del límite único también aumentará el número de falsos negativos pero decrecerá el número de falsos positivos. De forma análoga, disminuyendo el valor del límite único, disminuirá el número de falsos negativos a costa de aumentar el número de falsos positivos.





Un método simple para establecer adecuadamente este límite se basa en la observación del histograma de frecuencias de todos los pesos totales calculados. El histograma proporciona valiosa información para ayudar a tomar la decisión, basándose en la forma que éste toma para los distintos valores de los pesos. Esto es, los pares correspondientes al mismo individuo deberían tener un peso alto mientras que los pares de registros que no correspondan al mismo individuo deberían producir un peso bajo (es más, negativo).

En la práctica esto no es exactamente así debido a errores en las variables y a coincidencias casuales en las variables. De todas formas, el histograma de frecuencias tendrá una forma similar a la siguiente:



Así se decide el valor del límite y todos aquellos pares de registros con un peso superior a dicho valor se establecen como *links* mientras que aquellos pares con pesos inferiores se señalan como *no-links*.

Blocking

Lo ideal a la hora de fusionar dos ficheros A y B sería comparar cada uno de los registros de A con todos los registros de B, pero esto, para dos ficheros de tamaño medio, supondría un total de |AxB| comparaciones, que es un número demasiado grande.

Por lo tanto, lo que se hace es utilizar un esquema que extraiga únicamente pares de registros que sean razonablemente susceptibles de corresponder con un *match*. Este proceso es lo que se llama *blocking*.

De esta manera, lo que se pretende es reducir el número de comparaciones que debe llevar a cabo el programa repartiendo los ficheros en bloques mutuamente exclusivos y exhaustivos diseñados para aumentar la proporción de pares *matches* observados mientras se disminuye el número de pares de registros a comparar.

El *blocking* generalmente se implementa mediante clasificaciones de los dos ficheros sobre una o varias variables. Por ejemplo, si ambos ficheros se clasifican por el código postal, los pares que se considerarían serían aquellos donde el código postal coincida. Los pares de registros que no tengan el mismo código postal registrado no se analizarían y por tanto serían automáticamente considerados *no-matches*.

Obviamente, si por cada registro se buscase su correspondencia en el otro fichero completo, la probabilidad de encontrar un verdadero *match* sería mayor dado que no se excluye ningún registro en la consideración.

Sin embargo, el coste de tal proceso sería excesivo. Así, el *blocking* es una compensación entre el coste computacional (examinar demasiados pares de registros) y las proporciones de falsos *no-matches* (clasificar los pares de registros como *no-matches* si los registros no son miembros del mismo bloque). Para suavizar este inconveniente es aconsejable utilizar más de una variable de *blocking* y que sean independientes entre ellas.

También es muy importante la variable que se escoja como criterio de *blocking*. Una componente de *blocking* ideal debería contener un número considerable de valores que estén bastante uniformemente distribuidos. Es decir, ninguno de los bloques en que se divida el archivo debe ser excesivamente grande ni el número de bloques debe ser excesivamente pequeño para que el ahorro de tiempo de ejecución sea real. También deberían ser variables que tengan una probabilidad de registrar errores baja (es decir, un peso alto) puesto que si se ha cometido un error en esa variable los registros afectados serán colocados en distintos bloques, por tanto, no serán comparados y en consecuencia no se podrán fusionar.

Ahora se describen cuáles son los tipos de *blocking* implementados en nuestro programa de fusión. Cabe mencionar que estos criterios han sido programados de forma modular de modo que se facilite la futura inclusión de otros criterios distintos.

Criterio fecha de nacimiento.

Este criterio establece bloques según la variable fecha de nacimiento del registro. Es decir, sólo estudiará aquellos pares de registros cuya fecha de nacimiento coincida.

No se considerarán los registros cuyo valor para esta variable esté ausente ya que no existe ningún motivo que indique pensar que el hecho de que en un fichero esta variable no aparezca, implique que tampoco vaya a aparecer en el segundo fichero.

Este es uno de los motivos principales por lo que se aconseja no utilizar únicamente este criterio de *blocking*, especialmente si la probabilidad de que la variable fecha de nacimiento esté vacía es alta.

Criterio clave.

Este criterio utiliza para la clasificación de los archivos una variable de tipo clave como puede ser la variable *identificador de vivienda*. Dicha variable no es identificadora de cada registro, ya que puede que haya varios individuos con el mismo *identificador de vivienda* pues todos residan en la misma vivienda.

Criterio código de texto 1.

Este criterio se basa en un cierto código de caracteres que se forma a partir de los apellidos del individuo. Para cada fichero se crean dos variables nuevas que almacenan la combinación de la primera letra y las dos siguientes consonantes de cada apellido. La variable de *blocking* final se construye concatenando las dos variables anteriores de forma que se crea una clave alfabética a partir de ambos apellidos. Se utiliza esta serie de caracteres en lugar de los apellidos completos para evitar los errores que hayan podido ocurrir al registrar los valores de esta variable.

Criterio código de texto 2.

Este criterio se basa en un cierto código de caracteres que se forma a partir de los apellidos del individuo. Para cada fichero se crean dos variables nuevas que almacenan las tres primeras letras de cada apellido. La variable de *blocking* final se construye concatenando las dos variables anteriores de forma que se crea una variable alfabética a partir de ambos apellidos. Como en el criterio anterior, se utiliza esta serie de caracteres en lugar de los apellidos completos para evitar los errores que hayan podido ocurrir al registrar los valores de esta variable.

Capítulo
3

Programación SAS

A continuación y de manera general se describe el programa de fusión desarrollado en EUSTAT detallando las macros SAS que lo conforman.

Variables Input

Antes de ejecutar el programa el usuario debe introducir ciertos datos necesarios para la correcta actuación del mismo. Los datos que se deben definir son los siguientes.

Localización de carpetas.

El usuario debe definir tres variables con la localización de las siguientes carpetas.

1 – Carpeta en la que se encuentran los dos ficheros que van a ser fusionados. En esta carpeta también se guardarán los ficheros resultados del programa. Serán tres: uno con los pares de registros fusionados y otros dos con los registros que no se han conseguido fusionar de cada uno de los dos ficheros.

2 – Carpeta en la que se almacenaran los ficheros de datos que se van creando durante la ejecución del programa.

3 – Carpeta en la que se almacenan los ficheros de datos auxiliares que se van creando y eliminando durante la ejecución del programa.

Nombre de los ficheros.

El usuario también debe indicar los nombres de los dos ficheros a fusionar.

Nombres y tipo de las variables de fusión.

Para cada una de las variables de fusión que quieran utilizarse debe introducirse en el programa el nombre que dicha variable tiene en cada uno de los ficheros y el tipo de variable que es, siendo posibles los siguientes tipos:

0 – Clave identificadora de registro.

1 – Variable tipo Nombre.

2 – Variable tipo Apellido.

3 – Variable tipo Sexo.

4 – Variable tipo Fecha de nacimiento.

5 – Variable tipo Año de nacimiento.

6 – Variable tipo Clave formada sólo por números.

7 – Variable tipo Clave formada por números y letras.

Es obligatorio que al menos se definan una variable tipo clave identificadora para distinguir los registros, otra tipo nombre, otra tipo apellido y otra tipo sexo.

Indicadores de sexo.

También debe comunicársele al programa cuáles son los valores que indican el sexo masculino y el sexo femenino.

Tipos de criterio de *blocking*.

Debe decidirse con qué criterios de *blocking* se quiere ejecutar el programa.

Los tipos posibles son:

1 – Fecha de nacimiento.

2 – Clave.

3 – Código de texto 1.

4 – Código de texto 2.

Cada uno de estos criterios ha sido detallado en el apartado correspondiente del capítulo anterior.

Parámetros.

Para cada variable de fusión deben establecerse el valor de ciertos parámetros necesarios para el cálculo de los pesos. Son los siguientes:

- Probabilidad de que un valor de dicha variable haya sido mal registrado en el primer archivo.

- Probabilidad de que un valor de dicha variable haya sido mal registrado en el segundo archivo.

- Probabilidad de que un valor de dicha variable sea distinto en ambos ficheros a pesar de estar correctamente redactado en ambos.

Además también debe asignar un valor a las siguientes variables generales:

- Probabilidad de que un carácter haya sido registrado correctamente.

- Valor que se utiliza para establecer cuando un nombre o apellido es frecuente o no lo es.

Límite.

También será necesario establecer un valor límite que discrimine qué pares de registros hay que fusionar. En un momento dado la ejecución del programa se detiene, se muestra un histograma y a partir de él, se establece y se introduce el valor del límite.

Macros SAS

Nombre macro	Descripción
ejecución	Es la macro principal del programa. Se hacen dos llamadas a esta macro. En la primera llamada, se estudian los ficheros y las variables de fusión, se realiza la estandarización y homogeneización de las variables, se efectúa el procedimiento de asignación de valores según las listas estandarizadas descritas anteriormente, se calculan los pesos, se dividen los bloques y se estudia la distribución de los pesos de los pares de registros que pertenecen a algún bloque. Esta distribución se muestra en forma de histograma. El programa devuelve el histograma al usuario a partir del cual establece el valor del límite que utilizará para distinguir entre *links* y *no-links* y llamará por segunda vez a esta macro para que realice el último paso que consiste en estudiar los pares de registros cuyo peso está por debajo del límite pero es muy próximo a él.
paso1_analisis	Se crean las librerías en las que se va a trabajar. Se distingue el archivo1 del archivo2 (el archivo1 será el menor, el que tenga menos registros, ya que se supone que se quiere fusionar completamente este fichero). Se analizan las variables de que se dispone y se comprueba que no haya incongruencias en los datos introducidos por el usuario.
recuento	Se calcula el número de registros de un fichero de datos.
paso2_confeccion_listas	Esta macro llama a otras macros para realizar un proceso de homogeneización y estandarización de las variables de tipo nombre y apellido, dado que este tipo de variables son susceptibles de contener errores y además una misma configuración puede aparecer representada de distintas formas.

enies	Corrige aquellas configuraciones en las que aparece el símbolo ¥ en sustitución de la letra Ñ.
estandarizar	Estandariza las variables tipo nombre o apellido según unos ciertos criterios como la eliminación de signos ortográficos, de puntuación o la identificación de caracteres con similitud fonética o gráfica.
emparejar	Algunos caracteres a pesar de no tener una similitud fonética o gráfica relevante pueden ser utilizados a veces con un mismo significado. Por tanto, la macro estudia las configuraciones que sólo se diferencian en un carácter y evalúa si proceden del mismo origen.
comunes_y_no_comunes	Estudia la frecuencia con la cual cada una de las configuraciones aparece en el conjunto de los dos archivos a fusionar. Tras este estudio se distinguen los valores entre comunes y no comunes según dicha frecuencia. Notar que se considerarán como la misma configuración aquellos valores que, tras los pasos de estandarización anteriores, hayan resultado en el mismo código.
buscar_errores	Se supone que los valores con frecuencias bajas (no comunes) son configuraciones erróneas procedentes de alguno de los valores frecuentes (comunes). Entonces, se establece una medida de distancia y se seleccionan pares de valores formados por un valor común y un valor no común que tengan una distancia suficientemente pequeña como para considerar que puedan provenir de la misma configuración.
confeccion_listas	De entre todos los pares extraídos de la macro anterior, se seleccionan aquéllos que según unos criterios de frecuencias y probabilidades de error, se tienen suficientes indicios como para considerar que corresponden a pares de nombres o apellidos provenientes de la misma configuración original.
indicador_sexo	Una vez realizada la estandarización de la variable tipo nombre esta macro identifica si cada valor pertenece a un hombre o a una mujer.

paso3_asignar_estandarizados	Estudia aquellas configuraciones de nombres que aparezcan tanto en el listado de hombres como en el de mujeres y toma una decisión sobre si incluirlo en una tabla u en otra según las frecuencias de dicha configuración en ambos listados. Una vez realizada esta tarea reinicia los códigos asignados a cada configuración de las variables nombre de hombres, nombre de mujeres y apellidos. Habrá dos series de códigos, una para apellidos y otra, tanto para nombre de hombres como de mujeres pero en este orden, es decir, que existirá un número de código en concreto que se almacenará en una variable y para el cual, todo código inferior a él corresponderá a un nombre masculino y todo código superior a él, corresponderá a un nombre femenino.
est_final	Macro auxiliar que se llama desde la macro paso3_asignar_estandarizados y que reinicia los códigos para que éstos sean una serie consecutiva de números naturales.
paso4_calculo_pesos	Se calculan los pesos m y u descritos en la metodología.
paso5_blocking	Se ejecutan uno por uno todos los criterios de *blocking* que haya establecido el usuario. Se reúnen todos los pesos recuperados de estos procedimientos de *blocking* y se dibuja un histograma (con aquellos pesos mayores que 5, para que dicho histograma sea visiblemente más claro) a partir del cual se decidirá el valor de la variable límite.
blocking_fecha_nacimiento	Efectúa un *blocking* de los registros para coincidencia en la variable fecha de nacimiento.
blocking_clave	Efectúa un *blocking* de los registros para coincidencia en una variable tipo clave.
blocking_codigo_texto1	Efectúa un *blocking* de los registros para coincidencia de unas cadenas de caracteres compuestas por la primera letra y las dos siguientes consonantes de cada uno de los apellidos.
blocking_codigo_texto2	Efectúa un *blocking* de los registros para coincidencia de unas cadenas de caracteres compuestas por las tres primeras letras de cada uno de los apellidos.

fusion	Macro auxiliar que se llama desde cada una de las macros correspondientes a un criterio de *blocking* y que efectúa el cálculo de los pesos m y u para cada todas las configuraciones de cada variable.
paso6_posibles_links	Se seleccionan los pares de registros que se consideran *matches*. Primero se toman aquellos cuyo peso total esté por encima del límite establecido y después se hace un estudio un poco más exhaustivo donde se seleccionan algunos de los pares cuyos pesos rozan el límite sin superarlo.
asignación_lineal	Dado un grupo de pares de registros susceptibles de provenir del mismo origen, toma asignaciones biunívocas, escogiendo cuando haya varias posibilidades aquella con un peso mayor.
iml	Lleva a cabo un procedimiento de asignación lineal (lsap) para la asignación 1-1.
comparador_strings	Calcula el comparador de cadenas de caracteres de Jaro para medir la distancia entre dos configuraciones.

.

Capítulo 4

Aplicaciones

A continuación se describen algunas de las aplicaciones de fusión[2] que se han realizado en EUSTAT y que utilizan métodos de fusión probabilísticos. Las primeras son casos particulares en los que se introdujo algún detalle del método probabilístico. A raíz de sus resultados y de la literatura existente al respecto, se desarrolló el programa de fusión que se ha descrito en el capítulo anterior. El resto de aplicaciones son ya consecuencia de los resultados obtenidos al ejecutar dicho programa.

Fichero de Matrimonios y Registro Estadístico de Población

Las primeras pruebas en la implantación de metodologías de fusión de registros mediante técnicas probabilísticas en el Instituto Vasco de Estadística fueron llevadas a cabo con el objeto de fusionar el fichero que contiene la información de los datos personales de los titulares de los matrimonios y el que contiene los datos de cada Unidad Poblacional Básica, es decir, el Registro Estadístico de Población.

Las fusiones que habían sido realizadas hasta ese momento conseguían un porcentaje bajo de fusión, dado que en estos ficheros no se dispone de ningún identificador único como el DNI o algún otro tipo de clave identificadora.

En primer lugar, se toman de la base de datos Oracle los dos ficheros de datos, uno con los registros de los matrimonios con fecha posterior a enero de 1996 y otro con los datos de todos los registros correspondientes a cada Unidad Poblacional Básica.

Se necesita determinar qué variables de fusión van a utilizarse. Para ello, se observan los formatos en los que aparecen las variables. Se puede comprobar fácilmente que algunas variables no son en absoluto útiles para el proceso de emparejamiento, dado que muchas de ellas aparecen poco informadas, o son registradas de manera diferente en cada fichero. En otros casos, como ocurre en la variable sexo, se necesita una recodificación previa para utilizar dicha variable en el proceso puesto que en un fichero viene registrada como 'H' y 'V' y en el otro como '6' y '1'.

Las variables que en este caso se van a utilizar como variables de fusión van a ser:

Nombre
Primer apellido
Segundo apellido
Sexo
Fecha de nacimiento

[2] Todas las aplicaciones de fusión mencionadas en este texto se realizaron dentro del ámbito de EUSTAT y bajo el amparo del secreto estadístico (regulado en la ley 4/1986 de Estadística de la Comunidad Autónoma de Euskadi) que afecta a todo el personal implicado en su desarrollo.

En total se tenía un conjunto de 64.384 registros de matrimonios de los cuales se conseguía fusionar mediante los procedimientos anteriores 30.802, es decir, un 47,84% de los registros.

En primer lugar se fusionaron mediante un procedimiento de SAS los registros que coincidiesen directamente en todas las variables de fusión para aplicar el procedimiento de fusión probabilística a aquellos registros cuya coincidencia fuera más complicada de justificar. Los registros que no se fusionaron en este intento, se pasan por el programa de fusión probabilística utilizando dos criterios de *blocking* distintos. El primer criterio fue por fecha de nacimiento y el segundo por las iniciales del nombre, primer apellido y segundo apellido. En total, se consiguieron fusionar 60.055 registros que suponen un 93,28% del fichero original.

Registro Mercantil y Directorio de Actividades Económicas

Otro de los primeros casos en el que se aplicaron procedimientos de fusión probabilística fue la adaptación del programa general de fusión de dos ficheros para unir empresas del Registro Mercantil con el Directorio de Actividades Económicas de Eustat. El objetivo de esta fusión era poder utilizar la información económica que aporta el Registro Mercantil como fuente de actualización del Directorio de Actividades Económicas, y de esta manera, aumentar la cobertura de la operación, al ser el Directorio de Actividades Económicas el marco posterior de elevación de dicha información.

El primer paso para llevar a cabo esta tarea fue obtener un fichero único del Registro Mercantil. Éste es la unión de 6 ficheros con igual estructura, 2 por territorio histórico. En cada Territorio uno de los ficheros tiene origen digital y el otro se obtiene tras un proceso de escaneo y posterior reconocimiento de caracteres (OCR). Del fichero resultante se eliminan los registros duplicados (conservando el más reciente).

El siguiente proceso consiste en cruzarlo con el Directorio de Actividades Económicas por el campo CIF/DNI. Como resultado del mismo se obtiene un fichero con los registros que no han coincidido de forma directa con el directorio. Es a este fichero con los registros que no han coincidido de forma directa al que se le aplica el proceso de fusión, previa estandarización del CIF/DNI y del nombre. De todos los analizados se recuperan los nuevos registros coincidentes con el directorio que cumplen ciertos criterios.

Las variables que se consideran de fusión son:

CIF/DNI
Nombre
Código postal
Teléfono.

Se realizan distintas fusiones según las siguientes variables:

- CIF/DNI, Nombre, Código postal y Teléfono.
- CIF/DNI y Nombre.
- Nombre.
- CIF
- Teléfono
- Código Postal

En los casos de las fusiones por CIF y por teléfono, se exigía también que existiera cierta similitud en los nombres, además de coincidir los valores del CIF estandarizado y el teléfono. Esta similitud se calculaba mediante la ejecución de una macro SAS diseñada para ello y que consideraba cuatro supuestos de relación:

1. Los nombres comprimidos coinciden.

2. El comparador de cadenas de caracteres de Jaro es superior a 0.85.

3. Uno de los nombres de empresa es una abreviatura o siglas del otro nombre.

4. El número de letras de las palabras comunes en ambos nombres es superior que un cierto valor.

La macro no sólo seleccionaba los pares de registros cuyos nombres cumplían alguno de los criterios anteriores sino que además marcaba cuál de los cuatro criterios era el que se cumplía para cada par de registros de establecimientos.

Hasta el momento nada de lo que se ha descrito para esta explotación se refiere a métodos probabilísticos. Éstos se utilizan en la última fusión, la del código postal. Evidentemente, fusionar dos empresas porque coincida el código postal no tiene ninguna lógica ya que bajo un mismo código postal habrá registradas muchas empresas. En este caso, lo que se hace es, además de comprobar una cierta similitud entre los nombres de las empresas como se hace en las fusiones anteriores, se asigna un determinado peso a cada par de registros según la frecuencia del nombre.

Es decir, se seleccionan aquellos registros para los cuáles coincida el código postal y sus nombres estén relacionados mediante alguno de los criterios enumerados anteriormente. A cada uno de estos pares se le asigna un peso. Este peso está directamente relacionado con la frecuencia con que las palabras comunes a ambos nombres aparecen en los dos ficheros.

Por último, se fusionan los pares de registros que cumplan:

(1) Coincide el CIF (estandarizado).

(2) Coincide el nombre (estandarizado).

(3) Coincide el teléfono y existe relación entre los nombres.

(4) La variable peso es mayor que un cierto valor dado.

(5) Si todas las palabras que aparecen en uno de los nombres están en el otro nombre.

Los resultados fueron que, de los 30.000 registros de empresas iniciales, se consiguieron fusionar 22.500 mediante el procedimiento exacto. Los 7.500 registros que quedaron sin fusionar, pasaron a ser estudiados mediante el procedimiento descrito y se consiguió relacionar aproximadamente 900 de ellos. Es decir, en total, se aumentó el porcentaje de fusión de un 75% a un 78% gracias a este procedimiento.

Cabe mencionar que muchos de los registros que no se consiguieron fusionar en ninguna de las dos etapas corresponden al sector agrario y sus condiciones particulares impiden que puedan ser fusionados por ningún método.

Estadística de la Renta Personal y Familiar

Otra de las experiencias de EUSTAT en la que se utilizan algunos procesos relacionados con la fusión probabilística fue la llevada a cabo sobre la Estadística de Renta Personal y Familiar elaborada para el ejercicio de 2001.

La fusión de registros en la Estadística de Renta Personal y Familiar se realiza entre los ficheros recibidos de las Haciendas Forales y el Registro Estadístico de Población.

Esta fusión se concibe como un proceso de comparación de campos de fuentes distintas, detección de valores iguales y asignación de un peso de fusión por cada coincidencia para posteriormente calcular un peso total, en función del cual se asignará la clave de Unidad Poblacional Básica correspondiente (del Registro Estadístico de Población). Es decir, se le aplica el proceso de fusión exacta con unos pesos determinados.

La novedad en esta aplicación es que previo a la ejecución del procedimiento de fusión se realiza un tratamiento de homogeneización de la variable *Nombre y apellidos* basado en los métodos probabilísticos desarrollados.

A continuación se describen los tratamientos de homogeneización de la variable *Nombre y apellidos*, imprescindible en el procedimiento de fusión.

Se reciben tres ficheros de las respectivas haciendas forales con diferentes diseños de registros y es necesario un tratamiento de homogeneización de las variables. El campo *Nombre y apellidos* se debe ajustar al esquema de tres subcampos: nombre, primer apellido y segundo apellido, separados por asteriscos.

Se explica a continuación cuáles son exactamente los problemas que se encuentran al tratar esta variable.

- En algunos registros que contienen alguno de los apellidos compuestos, el campo *Nombre y apellidos* es erróneo. Si el primer apellido es compuesto, la última partícula de ese apellido se convierte en el segundo apellido, y éste va como nombre sin separar con asterisco del verdadero nombre. Si el segundo apellido es el compuesto, la última partícula del mismo pasa al campo de nombre sin separar con asterisco del nombre verdadero.

- Otros registros se ajustan al esquema de nombre, primer apellido y segundo apellido separados por espacios en blanco.

Como base para la homogeneización se utilizan unas bases de datos generadas a partir de los nombres y apellidos de los individuos registrados en el Censo de Población y Vivienda 2001 y en la Encuesta de Población en Relación con la Actividad.

Se han tomado los datos de nombres y apellidos de estos ficheros, se ha calculado la frecuencia con la que cada uno de ellos aparece en los dos ficheros y en función de estas frecuencias se han creado las siguientes bases de datos:

Nombre base de datos	Descripción
listado_final variables: *variable* *resultado*	Se genera a partir de los datos de los nombres y apellidos de los individuos del Censo de Población y Vivienda 2001 y la Encuesta de Población en Relación con la Actividad. Se han tomado cada una de las componentes de los campos nombre y apellidos y según la frecuencia con la que aparecen se les ha asignado un valor en el campo resultado que será uno de estos posibles. 1 = Nombre 2 = Nombre menos frecuente. 3 = Apellido 4 = Apellido menos frecuente. 5 = Nombre y Apellido
datos_2 variables: *variable* *var1* *var2* *resultado* *n_datos*	También está generada a partir de los nombres y apellidos de los individuos del Censo de Población y Vivienda 2001 y la Encuesta de Población en Relación con la Actividad. Se ha generado para intentar detectar los datos correspondientes a nombres y/o apellidos compuestos por 2 componentes. Los campos var1 y var2 serán las correspondientes componentes del nombre y/o apellidos compuesto, n_datos valdrá 2 para indicar que se tratan de dos componentes y el campo resultado toma uno de los valores: 1 = Nombre compuesto 3 = Apellido compuesto
datos_3 variables: *variable* *var1* *var2* *var3* *resultado* *n_datos*	Análogo a la base de datos anterior pero con los datos correspondientes a los nombres y apellidos compuestos por 3 componentes. Los campos var1, var2 y var3 serán dichas componentes, n_datos vale 3 y el campo resultado toma uno de los valores: 1 = Nombre compuesto 3 = Apellido compuesto
datos_4 variables: *variable* *var1* *var2* *var3* *var4* *resultado* *n_datos*	Análogo a la base de datos anterior pero con los datos correspondientes a los nombres y apellidos compuestos por 4 componentes. Los campos var1, var2, var3 y var4 serán dichas componentes, n_datos vale 4 y el campo resultado toma uno de los valores: 1 = Nombre compuesto 3 = Apellido compuesto

El fichero de entrada es el fichero de Hacienda con los campos *antes*, *variable* y *después*. El tratamiento se aplica sólo al campo *variable* que corresponde a los nombres y apellidos.

En primer lugar se hacen unas pequeñas modificaciones para la estandarización de los nombres y apellidos (se eliminan acentos, diéresis, símbolos, números, letras consecutivas repetidas, excepto RR y LL, y se traducen algunas contracciones de nombres y apellidos conocidos como M por MARIA o PZ por PEREZ).

Estas modificaciones se realizan para facilitar la identificación de los nombres y apellidos de los ficheros de Hacienda con los obtenidos del Censo de Población y Vivienda 2001 y la Encuesta de Población en Relación con la Actividad, ya que esas mismas modificaciones se hicieron a éstos últimos anteriormente.

Ahora se separan por palabras el campo *Nombre y apellidos* de los ficheros de Hacienda. Se eliminan aquellas palabras que sirvan de unión para algunos nombres y apellidos compuestos (como DE, LA, LOS, SAN,…). Se toman grupos de palabras y se comparan con aquellos nombres y apellidos compuestos que están almacenados en las bases de datos listado_final, datos_2, datos_3 y datos_4 que han sido descritas antes. Si se encuentran coincidencias, se almacena el número correspondiente de la variable *resultado* y de la variable *n_datos*.

Hay que tener en cuenta que puede que se encuentre más de una combinación posible a la hora de considerar los compuestos (por ejemplo, PEDRO JOSE RUIZ OTALORA se puede descomponer como PEDRO*JOSE*RUIZ-OTALORA o como PEDRO-JOSE*RUIZ*OTALORA). Por tanto, lo que se hace es resaltar esos registros mediante una variable que actúe como un marcador y que indique que ese registro es un caso especial a estudiar con más detenimiento.

Se asignan los códigos correspondientes según el resultado obtenido al buscar los nombres y apellidos de los registros de Hacienda con las bases de datos que se tiene de referencia. Las uniones que se utilizan a veces en compuestos y que antes se habían eliminado también se codifican pero mediante símbolos en lugar de números.

Sean las siguientes asignaciones:

Partícula	Código
DE, DEL, Y	&
Cualquier palabra de longitud 1 excepto Y	+
DA, DI, DO, SAN, DOS, SA, SANTA, SANTO, SAO, BON	*
LA, LAS, EL, LOS	$

Por último, se revisan aquellos registros que mediante una variable marcador han sido consideradas especiales y que se han separado en una base de datos para estudiar su especificidades más a fondo. En función de los valores que tengan se toman una serie de decisiones sobre cuál de ellas mantener.

El objetivo final es obtener los ficheros de salida *_asignados* y *_no_asignados* con los valores finales del campo *Nombre y apellidos*.

Una vez realizado el tratamiento de homogeneización de nombres y apellidos se procede a la fusión de ficheros en sí misma, en este caso, una fusión determinista.

Intervienen las siguientes variables (con repercusión en el tratamiento del RPF) para la asignación de pesos:

- NOMBRE (alfaclave de nombre, alfaclave de apellido 1 y alfaclave de apellido 2).
- RESIDENCIA (Territorio, Municipio, Distrito, Sección, Entidad, Calle, Número, Piso y Mano).
- DNI (a 8, 7, 6, 5 y 4 dígitos).
- FECHA DE NACIMIENTO (Día, Mes y Año de Nacimiento).

La asignación de la clave de Unidad Poblacional Básica se realizaría una vez computados todos los pesos, de manera que la función eliminaría, por peso asignado según el conjunto de los dominios, los registros de menor peso, y posteriormente los duplicados. Así se seleccionaría la clave de Unidad Poblacional Básica que tuviera mayor peso y, a su vez, no estuviera de baja. Tras el proceso de homogeneización descrito y el procedimiento de fusión determinista se consiguen fusionar más del 98% de los registros.

Censo de Población y Encuesta de Población en Relación con la Actividad

Estas técnicas de fusión también han sido utilizadas como aplicación a la validación del Censo de Población y Vivienda de la CAE del 2001, a partir de los datos obtenidos de la Encuesta de Población en Relación con la Actividad. Esta operación se realiza con periodicidad trimestral y tiene como objetivo plantear información estadística continua sobre el volumen y las características de los principales colectivos en que se puede clasificar a la población del País Vasco según su participación en las distintas Actividades económicas. Entonces, para aplicar el proceso de fusión se parte de dos ficheros. Uno, el correspondiente al Censo de Población y Vivienda de la CAE del 2001 que cuenta con 2.082.587 registros. Y otro, el correspondiente a la Encuesta de Población en Relación con la Actividad relativo al trimestre más próximo a la fecha de referencia del Censo que tiene 10.831 registros.

Los campos comunes a ambos ficheros y que se utilizan como variables de fusión son:

Nombre
Primer apellido
Segundo apellido
Sexo
Identificador de vivienda
Fecha de nacimiento

Como criterios de *blocking* se utilizan:

Fecha de nacimiento
Identificador de vivienda
Código de texto formado por la primera letra y las dos siguientes consonantes de ambos apellidos
Otro código formado por las tres primeras letras de los dos apellidos.

Tras aplicar el proceso de fusión los resultados fueron que se consiguieron encontrar en el fichero del Censo de Población y Vivienda, 10.535 registros de los 10.831 de la Encuesta de Población en Relación con la Actividad. Esto supone un porcentaje de fusión del 97,27%.

Padrón Municipal de Habitantes de Vitoria-Gasteiz y Registro Estadístico de Población

También se ha realizado la fusión entre los ficheros del Registro Estadístico de Población y el Padrón Municipal de habitantes de Vitoria-Gasteiz. El número de registros que tiene cada uno de estos ficheros es:

Padrón Municipal de Habitantes de Vitoria-Gasteiz	233.670
Registro Estadístico de Población	2.456.831

El objetivo es localizar todos los individuos registrados en el Padrón Municipal de Habitantes de Vitoria-Gasteiz en el Registro Estadístico de Población.

Las variables de fusión que se han utilizado han sido:

Primer apellido
Segundo apellido
Nombre
Fecha de nacimiento
Sexo

Y como criterios de *blocking* se han utilizado:

Fecha de nacimiento
Código de texto1
Código de texto2

Los resultados que se obtuvieron en una primera aproximación fueron:

Nº de pares de registros fusionados	231.646
Nº de registros del Padrón Municipal de Habitantes de Vitoria-Gasteiz sin fusionar	2.357
Nº de registros del Registro Estadístico de Población sin fusionar	2.225.188

Por sentido común, si se sumase el número de registros fusionados con el número de registros que quedan sin fusionar en cada uno de los ficheros, debería dar como resultado el número total de registros de los ficheros originales. Pero al fijarse en los números, se advierte que no ocurre así.

Esto se debe a que hay registros del Padrón Municipal de Habitantes de Vitoria-Gasteiz que se han fusionado con más de un registro del Registro Estadístico de Población puesto que éste tiene duplicados.

Lo mismo ocurre, pero en menor medida, al contrario. Varios registros del Padrón Municipal de Habitantes de Vitoria-Gasteiz han sido fusionados con el mismo registro del Registro Estadístico de Población.

En total, hay 231.313 registros distintos del Padrón Municipal de Habitantes de Vitoria-Gasteiz que han sido fusionados con algún registro del Registro Estadístico de Población. Esto supone un porcentaje de fusión del 98,99%.

Como se indicó en el apartado "Antecedentes" ya se había efectuado anteriormente un procedimiento de fusión a estos ficheros mediante técnicas no probabilísticas. Como se disponía de los resultados de esta otra fusión lo que se decidió fue comparar los resultados obtenidos por ambos métodos.

Hay 635 registros del Padrón Municipal de Habitantes de Vitoria-Gasteiz que han sido fusionados con un registro del Registro Estadístico de Población distinto por cada método de fusión. Estos registros han sido revisados manualmente y se han comparado los dos registros con los que fusionan. En general se ha observado que aproximadamente en 4 de cada 5 registros es más ajustado el registro que ha sido fusionado mediante el método probabilístico. En cambio, parece que funciona mejor el método determinista tan sólo en 1 de cada 25 registros. En el resto de registros, o bien no se encontró ningún criterio para determinar cuál de los dos era mejor o bien ambos parecían haber fusionado con el mismo individuo sólo que con distintas claves.

Por otro lado, hay 535 registros que han sido fusionados mediante el programa de fusión probabilística pero que no habían sido fusionados mediante el procedimiento anterior, y de forma análoga, existen 1.126 registros que sí fueron fusionados mediante el procedimiento anterior pero no lo han sido durante la ejecución del programa que se describe en este documento.

Algunas de las observaciones que se obtuvieron revisando estos últimos registros son:

- El método determinista causa problemas a la hora de fusionar nombres de individuos extranjeros ya que en algunos casos varios registros con el mismo nombre son fusionados entre sí a pesar de tener apellidos diferentes. Esto ocurre en general porque el segundo apellido de muchos individuos extranjeros no aparece registrado y el método determinista establece que la comparación entre dos campos vacíos es positiva y junto a la coincidencia del nombre produce un resultado de *match* a pesar de no haber coincidencia en el primer apellido. Este problema se resuelve con el método probabilístico pues evalúa en conjunto la coincidencia o similitud de todos los campos.

- En ambos métodos se ha detectado que existen errores a la hora de distinguir hermanos, especialmente en los casos de gemelos y mellizos.

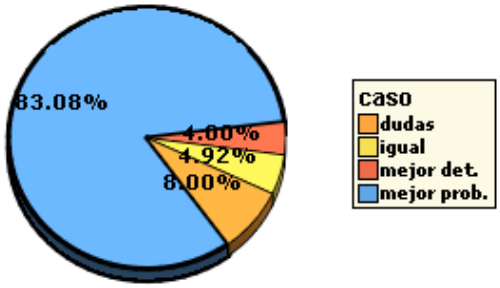
- En algunos casos los apellidos del individuo aparecen traspuestos. Es decir, en un fichero lo que aparece como primer apellido corresponde con lo que en el otro fichero es el segundo apellido y viceversa. Este intercambio de apellidos hace que el método probabilista no los identifique como un *match* a pesar de serlo (el método determinista sí los identifica correctamente).

Teniendo en cuenta esta última observación se planteó incluir código en el programa de fusión que tratase con esta situación. Así se hizo y se volvieron a calcular los resultados. En esta segunda aproximación éstos fueron:

Nº de pares de registros fusionados	231.818
Nº de registros del Padrón Municipal de Habitantes de Vitoria-Gasteiz sin fusionar	2.185
Nº de registros del Registro Estadístico de Población sin fusionar	2.225.016

Como ocurría antes hay registros del Padrón Municipal de Habitantes de Vitoria-Gasteiz que fusionan con más de un registro del Registro Estadístico de Población y por tanto producían más de un par de fusión. En total, hay 231.485 registros distintos del Padrón Municipal de Habitantes de Vitoria-Gasteiz fusionados, lo que supone, un 99.06% del total. Con esta modificación, se han conseguido fusionar 172 registros más.

Tras esta segunda ejecución del programa, se volvieron a buscar los registros que habían sido fusionados de manera distinta según el método utilizado. En esta ocasión se recogió que en el 83% de los casos funciona mejor el método probabilístico, en el 4% es mejor el resultado del método determinista, en el 5% ambos métodos fusionan igual y por último queda un residuo del 8% de casos en los cuales no se ha sabido tomar una decisión.



Encuesta de Población en Relación con la Actividad y Registro Estadístico de Población

Otro fichero que ha sido fusionado con el Registro Estadístico de Población ha sido la Encuesta de Población en Relación con la Actividad. Pero en este caso no se han tratado de localizar todos los registros de la Encuesta de Población en Relación con la Actividad sino sólo aquellos que no tenían clave de Unidad Poblacional Básica.

La Encuesta de Población en Relación con la Actividad se basa en una muestra probabilística continua, es decir, un panel de viviendas que se va renovando continuamente. La muestra de viviendas se extrae aleatoriamente del Directorio de Viviendas de modo estratificado a nivel de Territorio Histórico.

Una vez obtenida la muestra de las viviendas, se asocian los datos de las personas de las viviendas seleccionadas a partir del Registro Estadístico de Población. Por este motivo, al sacar la muestra de todos los individuos de cada vivienda, todos poseen clave de Unidad Poblacional Básica.

Sin embargo, sucede a veces que cuando se acude a encuestar una vivienda las personas que residen en esa vivienda no son los que aparecen en la muestra (por ejemplo, si los que residían antes se han mudado).

En este caso se encuesta a las personas que actualmente residen en la vivienda. Como consecuencia, estas nuevas personas no tendrán rellenado el campo de clave de Unidad Poblacional Básica, y van a ser estos individuos los que se intenten localizar posteriormente en el Registro Estadístico de Población mediante el programa de fusión.

En el fichero de la Encuesta de Población en Relación con la Actividad había 41.568 individuos distintos, de los cuales, poco más del 10% no tienen clave de Unidad Poblacional Básica. Entonces, se fusionan los siguientes ficheros con los siguientes números de registros:

Encuesta de Población en Relación con la Actividad sin clave de Unidad Poblacional Básica	4.117
Registro Estadístico de Población	2.456.831

Las variables de fusión que se han utilizado han sido:

Primer apellido
Segundo apellido
Nombre
Fecha de nacimiento
Sexo

Y como criterios de *blocking* se han utilizado:

Fecha de nacimiento
Código de texto1
Código de texto2

Los resultados obtenidos fueron:

Nº de pares de registros fusionados	3.277
Nº de registros de la Encuesta de Población en Relación con la Actividad sin clave de Unidad Poblacional Básica sin fusionar	855
Nº de registros del Registro Estadístico de Población sin fusionar	2.453.556

En total son 3.262 registros distintos de la Encuesta de Población en Relación con la Actividad que no tenían clave de Unidad Poblacional Básica y que se han conseguido encontrar mediante la fusión probabilística. Esta cantidad supone un 79.23% del total de los registros.

Atendiendo sólo al porcentaje de fusión puede parecer que el resultado no es tan favorable como el obtenido en la explotación descrita anteriormente del Padrón Municipal de Habitantes. En este caso hay que hacer una pequeña aclaración y es que para algunos de esos individuos entrevistados en la Encuesta de Población en Relación con la Actividad es imposible encontrarlos en el Registro Estadístico de Población puesto que pueden no estar en él. Es el caso de los individuos que se localizan en una vivienda y que no aparecen en la muestra original (y por tanto carecen de clave de Unidad Poblacional Básica) y que corresponden a individuos que proceden de fuera de la Comunidad Autónoma Vasca y que, por tanto, no están registrados en el Registro Estadístico de Población.

Verificar cuántos de los registros de la Encuesta de Población en Relación con la Actividad que no tienen clave de Unidad Poblacional Básica no han sido fusionados debido a que no aparecen en el Registro Estadístico de Población es una tarea muy difícil. Lo que sí se ha podido comprobar es cuántos de esos registros no han sido fusionados puesto que son individuos que han nacido posteriormente a la última actualización del Registro Estadístico de Población. En este caso, hay 187 registros en la Encuesta de Población en Relación con la Actividad cuya fecha de nacimiento es posterior a la última registrada en el Registro Estadístico de Población.

Capítulo 5

Conclusiones

Después de todos los resultados obtenidos hasta el momento, la conclusión general es que el método de fusión probabilístico es de gran utilidad a la hora de fusionar ficheros que no tienen como variable común una clave identificadora de registro.

La principal ventaja del método probabilístico respecto al determinista reside en una mayor eficacia en los casos más difíciles, frente a la simplicidad y mayor rapidez del segundo. Ello implica la eliminación del tratamiento manual en algunos casos y por tanto una reducción de los costes.

Una exigencia del método probabilista es una alta capacidad computacional cosa que hoy en día cada vez es menos importante.

EUSTAT se plantea no sólo seguir trabajando con este método para la fusión de individuos sino también ampliarlo para poder considerar la fusión de ficheros de empresas.

Para ficheros de empresas, se observa que la principal diferencia entre fusionar individuos y fusionar empresas es el tipo de variables de las que se dispone en cada caso. En los registros de empresas, la variable *Nombre de la empresa*, no puede ser tratada de la misma forma que se hace cuando se trata del nombre de un individuo, dado que la casuística es totalmente distinta. Existe más diversidad en los nombres de empresa y además, la probabilidad de que ocurran errores es mayor, debido a que la complejidad es también mayor. Otra diferencia está en que una variable que aparece con gran asiduidad en los ficheros de empresas y que no se ha tratado en el caso de individuos, es la variable dirección. Esta variable puede aparecer registrada de muchas maneras distintas y su estandarización exigirá un trabajo adicional.

Por otra parte, el hecho de que se haya observado una gran utilidad en los métodos probabilísticos no implica que se dejen de utilizar en el Instituto los métodos deterministas. Debido a que los resultados de fusión dependen en gran parte de la calidad de los ficheros, habrá casos en los cuales sea mejor utilizar otro tipo de métodos en lugar de los probabilísticos. Es más, puede que para algunos ficheros en concreto la opción ideal sea una combinación de varios tipos de métodos.

Actualmente, EUSTAT está trabajando en la construcción de un módulo de fusión que optimice y facilite su utilización computacional de todos estos métodos para aplicar en cada caso el método más adecuado.

Capítulo 6

Bibliografía

[1] FELLEGI, IVAN P. AND SUNTER, ALAN B.

A theory of Record Linkage. Journal of the American Statistical Association, December vol. 64, nº 328, pp. 1183-1210 (1969).

[2] JARO, M.A.

Record Linkage research and the calibration of record linkage algorithms. U.S. Bureau of the Census (1984).

[3] JARO, M.A.

Advances in record linkage methodology as applied to matching the 1985 Census of Tampa, Florida. Journal of the American Statistical Association.

[4] BLAKELY, T. AND SALMOND, C.

Probabilistic record linkage and a method to calculate the positive predictive value. Internation Journal of Epidemiology (2002).

[5] AYESTARÁN, MARINA AND LEGARRETA, LEIRE.

Applying methods of record linkage for census validation in the Basque Statistics Office. Instituto Vasco de Estadística (2004).

[6] WINKLER, WILLIAM E.

Matching and Record Linkage. Bureau of the Census (1993).

[7] CHRISTEN, PETER AND CHURCHES, TIM.

Febrl – Freely extensible biomedical record linkage. Australian National University (2003).

[8] YANCEY, WILLIAM E.

An Adaptive String Comparator for Record Linkage. U.S. Bureau of the Census, Statistical Research Division (2004).