FUSIÓN DE REGISTROS ADMINISTRATIVOS DE UNIDADES ECONÓMICAS MEDIANTE TÉCNICAS PROBABILÍSTICAS

Laura Otero Franco

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

Un registro administrativo es un documento que contiene información relacionada con una unidad, que bien puede ser una persona, un establecimiento u otra entidad, que un servicio administrativo recaba para sus propios fines. En los últimos años, el potencial de la estadística basada en la explotación de registros administrativos ha avanzado a un ritmo vertiginoso.

Los registros administrativos se han convertido en un potente instrumento para la medida y el análisis de la realidad en diversos ámbitos. Uno de los usos estadísticos más frecuentes de los registros administrativos es la construcción y mantenimiento de los directorios, debido a su eficiencia y racionalidad de costes.

No obstante, la información de origen administrativo presenta diversas dificultades a la hora de ser utilizada en el entorno estadístico. Por ejemplo, es frecuente que los registros de origen administrativo hayan sido puestos en marcha en periodos anteriores o por unidades muy diferentes a la que va a iniciar la operación estadística, y principalmente (como característica intrínseca a todo registro administrativo) no han sido concebidos ni diseñados con fines estadísticos. De ahí que la información proporcionada por estos registros necesite un tratamiento previo a su utilización en el ámbito estadístico.

Uno de estos tratamientos previos es la fusión o enlace de registros. Estas técnicas permiten enlazar la información referente a una misma unidad contenida en distintos registros administrativos.

Vitoria-Gasteiz, Noviembre 2010

JAVIER FORCADA SAINZ

Director General de EUSTAT

ÍNDICE

Capítulo 1

INTRODUCCIÓN

Introducción y objetivos

La fusión o enlace de registros se define como el procedimiento de encontrar pares de elementos en dos registros administrativos distintos (un elemento por registro) que representen a la misma unidad. En caso de que ambos registros administrativos fuesen el mismo se estaría hablando de localizar elementos duplicados.

Hay dos técnicas generales dentro de la fusión de registros:

- **Fusión determinista**: Los elementos son relacionados cuando concuerdan exactamente en todos los campos o en un número predeterminado de campos.

- **Fusión probabilística**: Se asigna un determinado peso probabilístico a cada par de elementos y se consideran fusionados aquellos pares cuyo peso sea suficientemente alto.

Este cuaderno técnico tiene por objeto presentar el estudio de la fusión de registros administrativos que contienen información relativa a unidades económicas mediante métodos probabilísticos. Más concretamente, el estudio se focaliza en la fusión de los siguientes ficheros: el Directorio de Actividades Económicas de EUSTAT (DIRAE) y el fichero de la Seguridad Social.

Descripción del Proyecto

El proyecto se divide en tres fases:

I. Estudio y adaptación de la metodología de fusión probabilística propuesta por Fellegi y Sunter para el tratamiento de los registros administrativos de unidades económicas.

II. Programación en SAS de una aplicación que efectúe de manera automática la fusión de dos registros administrativos de unidades económicas.

III. Análisis de los resultados obtenidos.

Para presentar las distintas fases del proyecto se ha dividido este cuaderno técnico en los siguientes capítulos:

En el capítulo 1 se introducen los objetivos del cuaderno técnico, las fases del proyecto y los antecedentes de EUSTAT en fusión de registros.

En el capítulo 2 se introduce el concepto de registro administrativo y se presentan sus características, ventajas y desventajas y se particulariza en el caso de los registros administrativos con información de unidades económicas.

En el capítulo 3 se describe la metodología de fusión de registros mediante la técnica probabilística presentada por Fellegi y Sunter en 1969 en el artículo "A theory for Record Linkage".

En el capítulo 4 se presenta la estructura del programa de fusión de registros administrativos de unidades económicas mediante técnicas probabilísticas desarrollado por EUSTAT.

En el capítulo 5 se analizan los resultados obtenidos al ejecutar dicho programa de fusión con los ficheros que promovieron el estudio, esto es, el Directorio de Actividades Económicas de EUSTAT (DIRAE) y el fichero de la Seguridad Social.

Por último, en el capítulo 6 se muestran las conclusiones a las que ha llevado el estudio de estas técnicas y la aplicación del programa de fusión desarrollado.

Antecedentes

Cuando en 1996 tuvo lugar la última renovación del Padrón Municipal de Habitantes, EUSTAT decidió comenzar a trabajar con registros administrativos con el fin de conformar un Registro Estadístico de Población. Para ello, se utilizó la información contenida en el Padrón Municipal completo a 31 de diciembre de cada año y las Estadísticas del Movimiento Natural de la Población (Nacimientos, Defunciones y Matrimonios).

Para tratar dichos ficheros se comenzaron a implantar técnicas de fusión determinista utilizando variables identificativas comunes (nombre, apellidos, DNI/NIE, fecha de nacimiento, dirección postal, etc.). Posteriormente, dentro del marco de las becas de investigación y metodología estadístico-matemática que promueve EUSTAT, se comenzó el estudio de técnicas probabilísticas basadas en el modelo teórico presentado por Fellegi y Sunter en 1969 en el artículo "*A theory for Record Linkage*".

Dada la necesidad de generalizar estos procedimientos, tanto los deterministas como los probabilísticos, se llevó a cabo la programación de una aplicación de fusión independiente, llamada Módulo de Fusión (MDF).

Movidos por el éxito obtenido con la fusión de registros de individuos, se decidió estudiar la aplicación de los mismos métodos de fusión probabilística para la fusión de ficheros de unidades económicas. En un principio se pensó en adaptar el programa existente para la fusión de individuos pero se observó que no era algo trivial dada las peculiaridades de los registros de unidades económicas.

Capítulo 2

REGISTROS ADMINISTRATIVOS

Los registros administrativos son una fuente de datos estadísticos, como lo son las encuestas o los censos. La principal diferencia con las otras dos fuentes se fundamenta en que el registro administrativo se debe a criterios normativos que establece el servicio administrativo dueño del registro, en lugar de deberse a criterios metodológicos que toma el estadístico.

Existe un amplísimo número de registros de uso administrativo en manos de las diferentes administraciones públicas susceptibles de uso estadístico, los cuales recogen una gran variedad de datos con información relativa a personas, empresas, instituciones y otras poblaciones, datos que a priori no sería necesario recabar de nuevo a los informantes, disminuyendo así el coste económico de la recogida.

Con mayor frecuencia en los últimos años, las instituciones estadísticas han elaborado listas de registros susceptibles de aprovechamiento estadístico y se han dado grandes pasos en el suministro de información por parte de los grandes detentadores de datos, así la administración tributaria, educativa, sanitaria y de la seguridad social. Otros registros tales como el Registro Civil han venido usándose desde tiempo atrás para uso estadístico.

Características

A continuación se enumeran algunas de las características de los registros administrativos:

1) Son documentos normativos que registran un acontecimiento administrativo, es decir, no son concebidos para fines estadísticos.

2) No siguen una lógica de pasos estadísticos ya que el propósito con el que se realizan es de planificación, registro, seguimiento, coordinación y/o control administrativo.

3) Poseen instrumentos característicos de captación de información (plantillas, formularios, tarjetas, fichas, cuadernos o libros de actas, etc.) que obedecen a cubrir las necesidades del seguimiento y control administrativo.

4) Su periodicidad puede estar establecida o no de acuerdo a la legislación correspondiente. Por ejemplo, puede registrarse una sola vez (como los títulos universitarios), puede tener periodicidad (como los registros de tránsito y licencias) o puede ser aleatoria (como los registros de salud).

5) Son de amplia cobertura según la tenga el organismo dueño del registro administrativo.

6) Definen diversas unidades a ser registradas, personas, edificios o establecimientos, hechos (sucesos o actividades), servicios, recursos, transacciones comerciales, etc. y por tanto, múltiples variables.

7) Las unidades registradas y las variables medidas pueden ser objeto de conversión a variables estadísticas mediante metodologías relativamente sencillas.

8) Son almacenados en archivos de diferente formato que van desde papel, ficheros o expedientes hasta soporte digital.

Ventajas y desventajas

Se distinguen las siguientes ventajas de los registros administrativos frente a otras fuentes:

1) Bajo coste en la producción de datos. Generalmente los registros administrativos responden a un procedimiento normativo de alguna institución que necesita controlar una acción administrativa, por lo tanto, la información se toma generalmente en las oficinas de la institución y así no existen costes operativos de campo.

2) Menor carga en la cumplimentación de los formularios para los informantes. En general, las encuestas nacionales suelen ser largas, aunque hay que tener en cuenta que también existen registros administrativos engorrosos como pueden ser las declaraciones de impuestos.

3) Permite la circulación de la información entre los órganos de gobierno y con ello evita la duplicación de esfuerzos en la administración pública. Esta ventaja teórica obliga a sincronizar a los organismos de la administración pública en términos de requerimientos de datos. Este proceso es muy complejo debido a que las plantillas de cada organismo son independientes de los de otro o pueden variar las poblaciones sobre las que se aplican los instrumentos.

4) Logra una cobertura completa de la población objetivo. En muchos casos los registros administrativos logran cobertura completa como por ejemplo en áreas de salud, justicia o educación. Sin embargo, existen casos como victimización donde los registros administrativos no logran cobertura puesto que pueden existir pocas denuncias de criminalidad.

5) Los errores de no respuesta son menores que en otras fuentes, no hay errores muestrales. Es cierto que no existen errores muestrales pues no se toma una muestra estrictamente, pero este error se sustituye por el error de cobertura.

6) Es posible la desagregación en subpoblaciones. Esta ventaja es muy importante debido a que el registro administrativo puede contener una variedad de datos interesante con los cuales poder obtener subpoblaciones. Por esto es importante el análisis del registro administrativo para evaluar su pertinencia.

7) Fortalece a los sistemas de información en todos los ámbitos territoriales de un país. Es evidente cuando el Estado tiene una política de fortalecimiento de sus sistemas de información lo que hace de esta ventaja más una oportunidad.

8) La calidad de la información aumenta al tener la posibilidad de construir formularios con los detalles que requiere el tema de interés. Sin embargo, modificar formularios es complejo, aunque la oportunidad existe.

9) Constituye una base cierta para la construcción de series de datos. Los registros administrativos logran llevar la historia del proceso administrativo, lo que lo potencia la construcción de la serie temporal.

Por otro lado, los registros administrativos presentan ciertas desventajas como las siguientes:

1) La falta de correspondencia entre las unidades administrativas y las estadísticas. Evidentemente las unidades de análisis pueden no coincidir debido a que el registro administrativo puede referirse a una persona individual y no necesariamente al hogar o a un establecimiento comercial.

2) Diferencias en las definiciones de las variables. Generalmente las plantillas y formatos no contienen metodologías que definan las variables de modo operacional o pueden ser variables de identificación o descripción sin profundizar tal y como lo requeriría un estudio descriptivo.

3) Falta de conversión entre los códigos administrativos y estadísticos.

4) Datos y períodos de referencia no coincidentes con la finalidad estadística. Esto puede ocurrir pues de hecho los registros administrativos no tienen finalidad estadística, pero existen procesos de conversión de registros administrativos en datos estadísticos.

5) Efectos de los cambios políticos en la continuidad de los registros administrativos. Ciertamente las plantillas y procedimientos administrativos pueden variar con el vaivén político. Este aspecto es muy importante y declara una debilidad del registro administrativo.

6) Falta de un identificador común en los registros para realizar la conciliación de los datos y falta de personal estable para una tarea

7) Falta de una visión a largo plazo que conduzca al desarrollo del sistema estadístico y que privilegie la coyuntura.

8) Carencia de una política de cooperación entre los órganos que suministran los registros, la inexistencia de un acuerdo entre todos los actores e instituciones partícipes y la ausencia de leyes estadísticas para todas las desagregaciones territoriales.

Registros de empresas y otras unidades jurídicas de interés para la estadística económica

Algunos de los registros más importantes para la detección y actualización de las unidades estadísticas de producción económica son el Impuesto de Actividades Económicas (IAE), las unidades de cotización de la Seguridad Social (SS) y los Registros Mercantiles (RM).

El IAE es un impuesto municipal que grava anualmente el ejercicio de actividades empresariales, profesionales y artísticas. No obstante, su gestión centralizada se lleva a cabo en nuestro caso por las Haciendas Forales y en territorio común por la Agencia Estatal de Administración Tributaria (AEAT). Cada actividad en un lugar precisa de una licencia, de modo que un sujeto pasivo puede presentar varias licencias en uno o más domicilios de situación. Entre otros datos de identificación del sujeto y del domicilio, la licencia conlleva fechas de alta y de baja, un código o epígrafe de la actividad (basado en la antigua CNAE-74) y un literal descriptivo de la misma.

El IAE abarca casi la totalidad de la actividad mercantil llevada a cabo por las empresas, personas físicas y jurídicas, los profesionales y los artistas (en nuestro caso, en el ámbito de los territorios de la CAE). Todos los sujetos del Impuesto de Valor Añadido (IVA) están obligados a su declaración en la matrícula del IAE.

No obstante, la normativa actual establece la exención del pago del impuesto a todas las personas físicas y a las personas jurídicas por debajo de un cierto nivel de facturación. Esto ha tenido la inmediata consecuencia de que con frecuencia estos sujetos no se dan de baja al cesar en la actividad, dando una imagen abultada de la realidad económica.

Por lo demás, el IAE no ofrece información sobre agricultores, ganaderos y otras actividades del sector primario, ni tampoco claro está sobre las actividades no mercantiles de las administraciones públicas y demás instituciones sin fin de lucro. Con frecuencia, los ficheros que facilitan las haciendas forales tampoco incluyen los datos de grandes empresas con licencia nacional para el ejercicio de la actividad, con las cuales mantienen acuerdos especiales para el prorrateo de la parte del impuesto correspondiente a los municipios vascos.

Los registros de la SS de interés para los registros de empresas y unidades jurídicas están constituidos por los Códigos de Cuenta de Cotización (CCC), por los registros del Régimen Especial de Trabajadores Autónomos (RETA) y por los del Régimen Especial Agrario (REA, en extinción tras su incorporación al RETA) y del Régimen Especial del Mar (REA). Actualmente, se dispone de ficheros trimestrales con las unidades de alta a una fecha determinada, en nuestro caso siempre delimitado al ámbito geográfico de nuestra comunidad autónoma.

Los CCC contienen datos elementales del Empresario persona física o jurídica, solo identificación fiscal y nombre, y datos de cada cuenta de cotización relativa a las personas asalariadas contratadas por el empresario. Cada empresario puede disponer de uno o más CCC en un lugar determinado si tiene grupos de personas contratadas bajo diferentes modalidades de cotización a la Seguridad Social. Cada CCC contiene datos de la dirección de la unidad de cotización, código CNAE de la actividad, fechas de alta y de situación, tipo de relación laboral y número de trabajadores bajo ese CCC.

Los registros del RETA contienen datos descriptivos de las personas sujetas a ese régimen, así como datos de la dirección de la actividad y, escasas veces, código CNAE. No distingue entre empresarios propiamente dichos y otros trabajadores, socios principales, cooperativistas, autónomos dependientes y otros, para los cuales la Seguridad Social obliga a su inclusión en este régimen en lugar del régimen general.

De lo anterior se deduce que no disponemos de dato alguno procedente de SS sobre la gran cantidad de empresas societarias, particularmente sociedades limitadas, cuyos trabajadores (no más de dos o tres, las más de las veces) son también los propietarios o accionistas principales y que cotizan en RETA en lugar de en un CCC del régimen general.

Los datos de la SS suelen estar mejor actualizados, ya que los empresarios han de soportar multas simplemente por el retraso en las declaraciones y pagos de las cotizaciones o incluso por no dar de baja sus unidades de cotización al cesar la actividad. Las unidades de cotización abiertas pero obsoletas se refieren casi exclusivamente a sociedades dolosamente abandonadas por sus dueños, hasta tanto se den de baja de oficio por las autoridades del registro.

Los Registros Mercantiles constituyen la principal herramienta para proporcionar seguridad jurídica a las empresas y otros interesados en el tráfico mercantil. Tienen obligación de registrarse y declarar ciertos sucesos todas las sociedades mercantiles y, voluntariamente, las personas físicas y otras no jurídicas que ejerzan actividades mercantiles. Además, las sociedades tienen obligación de depositar sus cuentas en los registros para conocimiento público. La falta del depósito anual puede suponer una multa para la sociedad.

Los ficheros disponibles de RM nos permiten conocer datos elementales de las sociedades mercantiles cuyo domicilio social esté en alguna de las tres provincias de nuestra comunidad autónoma: Números de identificación registral y también fiscal, razón social, dirección de la sede social, objeto social y otros datos de menor interés. Paralelamente, los depósitos de cuentas tienen gran interés para encuestas y estadísticas económicas.

Al igual que con la SS, en RM podemos disponer de sociedades “abandonadas” no liquidadas por sus propietarios o que, aun en activo, llevan varios años sin depositar las cuentas.

En todos los casos de las tres fuentes administrativas mencionadas, IAE, SS y RM, los contenidos de las variables respectivas son totalmente dispares, adoleciendo cada una de su propia casuística, omisiones, errores, etc., consecuencia de servicios administrativos muy diferentes entre si y muy dispersos por toda la geografía.

Capítulo 3

FUSIÓN DE REGISTROS

Como se ha visto, los registros administrativos no son una fuente estadística propiamente dicha (de origen), en el sentido de los censos y las encuestas, ya que, por una parte, su finalidad es administrativa y por otra, sirve para control normativo, es decir, registra un evento o acto individual referido a una entidad y le afecta directamente.

Además, una operación estadística requiere definiciones y clasificaciones acordes con los objetivos de la investigación, mientras que los registros administrativos no necesariamente coinciden con estos aspectos metodológicos.

Una de las tareas que requiere el tratamiento de registros administrativos como fuente estadística es la fusión de registros. Mediante esta técnica se consigue utilizar información de diversas fuentes administrativas de manera adecuada para el uso estadístico. En este capítulo se describe la metodología que se ha seguido en EUSTAT para el desarrollo de técnicas de fusión probabilística.

Metodología

Para la elaboración del programa de fusión automática se ha seguido el modelo teórico presentado por Fellegi y Sunter en el artículo “A theory for record linkage” en 1969. Las bases de este modelo son las siguientes:

Modelo teórico

Se denota por A y B los registros administrativos a ser fusionados y sean a y b los miembros genéricos de los registros administrativos, respectivamente.

Se supone que ambos registros tienen miembros 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, en nuestro caso, a la misma unidad económica. 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

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

y

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

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

Cada unidad de la población en estudio tiene unas características asociadas, tales como, denominación, dirección postal, empleo, etc. Se han de identificar aquellos miembros que se refieren a una misma unidad económica. Sin embargo, el proceso de creación de los registros administrativos podría introducir errores o imprecisiones (errores de codificación, trascripción y tecleo, variaciones tipográficas o fonéticas, pérdida de datos, etc.) en los elementos generados. Como resultado de estos errores, dos miembros de A y B que no se refieren a la misma unidad económica podrían generar elementos idénticos y, más frecuentemente, dos miembros idénticos de A y B podrían producir elementos diferentes. Se denotan los elementos correspondientes a los miembros de A y B por $\alpha(a)$ y $\beta(b)$, respectivamente.

El primer paso al intentar emparejar elementos de dos registros administrativos es compararlos. El resultado de la comparación es un conjunto de códigos, que son codificados en afirmaciones del tipo: "la denominación coincide en ambos elementos", "la denominación coincide y es Almacenes Garrido", "la denominación no coincide", "la denominación está ausente en uno de los elementos" o "existe acuerdo en parte de la denominación pero no en toda". Formalmente, se define el ***vector comparación*** como un vector función de los elementos $\alpha(a)$ y $\beta(b)$ en la forma:

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

Se ve que γ es una función definida sobre AxB. Se puede escribir $\gamma(a,b)$, $\gamma(\alpha,\beta)$ 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 $\gamma(a,b)$ y se tiene que decidir si:

- (a,b) es un emparejamiento, $(a,b) \in M$ (se llama a esta decisión ***link*** y se denota por A_1)
- (a,b) es un no-emparejamiento, $(a,b) \in U$ (se llama a esta decisión ***no-link*** y se denota por A_3)

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 A_2, 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 $D = \{d(\gamma)\}$ donde:

$$d(\gamma) = \{P(A_1 \mid \gamma),\ P(A_2 \mid \gamma),\ P(A_3 \mid \gamma)\};\quad \gamma \in \Gamma$$

y

$$\sum_{i=1}^{3} P(A_i \mid \gamma) = 1$$

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 a cada regla de fusión. Se asume que un par de elementos $[\alpha(a), \beta(b)]$ es seleccionado aleatoriamente para ser comparado de acuerdo a un proceso probabilística. El vector de comparación resultante $\gamma[\alpha(a), \beta(b)]$ es por tanto una variable aleatoria. Se denota la ***probabilidad condicional de*** γ ***dado que*** $(a,b) \in M$ como $m(\gamma)$, y será:

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

De manera similar, se denota la ***probabilidad condicional de*** γ ***dado que*** $(a,b) \in U$ por $u(\gamma)$. Por tanto,

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

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

$$P(A_1 \mid U) = \sum_{\gamma\in\Gamma} u(\gamma).P(A_1 \mid \gamma)$$

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

$$P(A_3 \mid M) = \sum_{\gamma\in\Gamma} m(\gamma).P(A_3 \mid \gamma)$$

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

$$P(A_1 \mid U) = \mu \quad y \quad P(A_3 \mid M) = \lambda$$

Se dice que la regla de fusión $L(\mu, \lambda, \Gamma)$ es óptima si la relación

$$P(A_2 \mid L) \leq P(A_2 \mid L')$$

se mantiene para cualquier $L'(\mu, \lambda, \Gamma')$ 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 A_1 y A_3) sujeta a unos niveles fijos de error. Esto parece una decisión razonable, dado que el adoptar una decisión A_2 requiere de

costosas operaciones manuales de fusión. Además, por otro lado, parece que si la probabilidad de A_2 no es pequeña, el proceso de fusión es de dudosa utilidad.

Los autores proponen una regla de fusión óptima a los niveles de error (μ, λ) que viene dada de la siguiente forma:

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

donde $T_\mu = \dfrac{m(\gamma_n)}{u(\gamma_n)}$, $T_\lambda = \dfrac{m(\gamma_{n'})}{u(\gamma_{n'})}$ y n, n' dos enteros tales que $0 < n \leq n' < N_\Gamma$.

En muchas aplicaciones sería posible tolerar niveles de error suficientemente altos para eliminar la posibilidad de la acción A_2. En este caso, se consideran 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 se ha adoptado en EUSTAT a la hora de desarrollar el programa automático de fusión, de modo que se establece un límite único $T_\mu = T_\lambda$.

Para ver más detalles de la construcción de la regla de fusión óptima, el cálculo de los pesos $m(\gamma)$ y $u(\gamma)$ y otros pormenores del modelo teórico se puede consultar el artículo mencionado de Fellegi y Sunter [1] o el cuaderno técnico "Métodos automáticos de fusión de registros y su utilización en EUSTAT" [2].

Capítulo

4

PROGRAMACIÓN

Siguiendo la metodología expuesta en el capítulo anterior se han elaborado un programa en SAS para la realización de la fusión entre dos ficheros de unidades económicas. El programa está orientado principalmente a la fusión de dos ficheros concretos, estos son, el Directorio de Actividades Económicas de EUSTAT (DIRAE) y el fichero de la Seguridad Social.

Debido a las particularidades de ambos ficheros no se ha podido construir un programa genérico para la fusión de dos ficheros cualesquiera, sino que se ha programado de manera *ad hoc* para estos ficheros en concreto. Sin embargo, el programa de fusión es un programa modular y secuencial, de tal manera que muchas macros a las que llama el programa pueden utilizarse en posteriores fusiones con ficheros diferentes.[1]

En este capítulo se describe la estructura del programa de fusión. Más adelante, en el siguiente capítulo, se hará hincapié en los ficheros objeto de estudio y en los procedimientos auxiliares necesarios para implementar la fusión entre ellos.

Programa general

El programa de fusión consta de los siguientes ficheros SAS:

→ Programa principal donde el usuario define los argumentos necesarios para la ejecución del mismo.

→ Programa que contiene las macros a las que llama el programa principal y que efectúan las distintas etapas de la fusión.

Además de contener las macros propias de la fusión basada en el modelo teórico de Fellegi y Sunter descrito en la sección anterior, también contiene unas macros para ejecutar un procedimiento de blocking. Este procedimiento no es imprescindible en la teoría, pero sí que lo es a nivel práctico. La comparación de todos los elementos de un registro administrativo con todos los elementos de otro registro administrativo no se puede asumir a nivel computacional, y por lo tanto, es necesario utilizar algún criterio de blocking que seleccione un subconjunto de pares de elementos susceptibles de ser fusionados, y así evitar la comparación masiva de todos los elementos.

Además de estos ficheros se ha creado un programa auxiliar que construye dos tablas externas necesarias para la correcta ejecución del programa de fusión. Las tablas que construye son las siguientes:

- Un conjunto de datos que contiene algunas partículas (preposiciones, artículos, etc.) que se eliminan de las variables de fusión alfabéticas en el paso de estandarización y homogeneización.

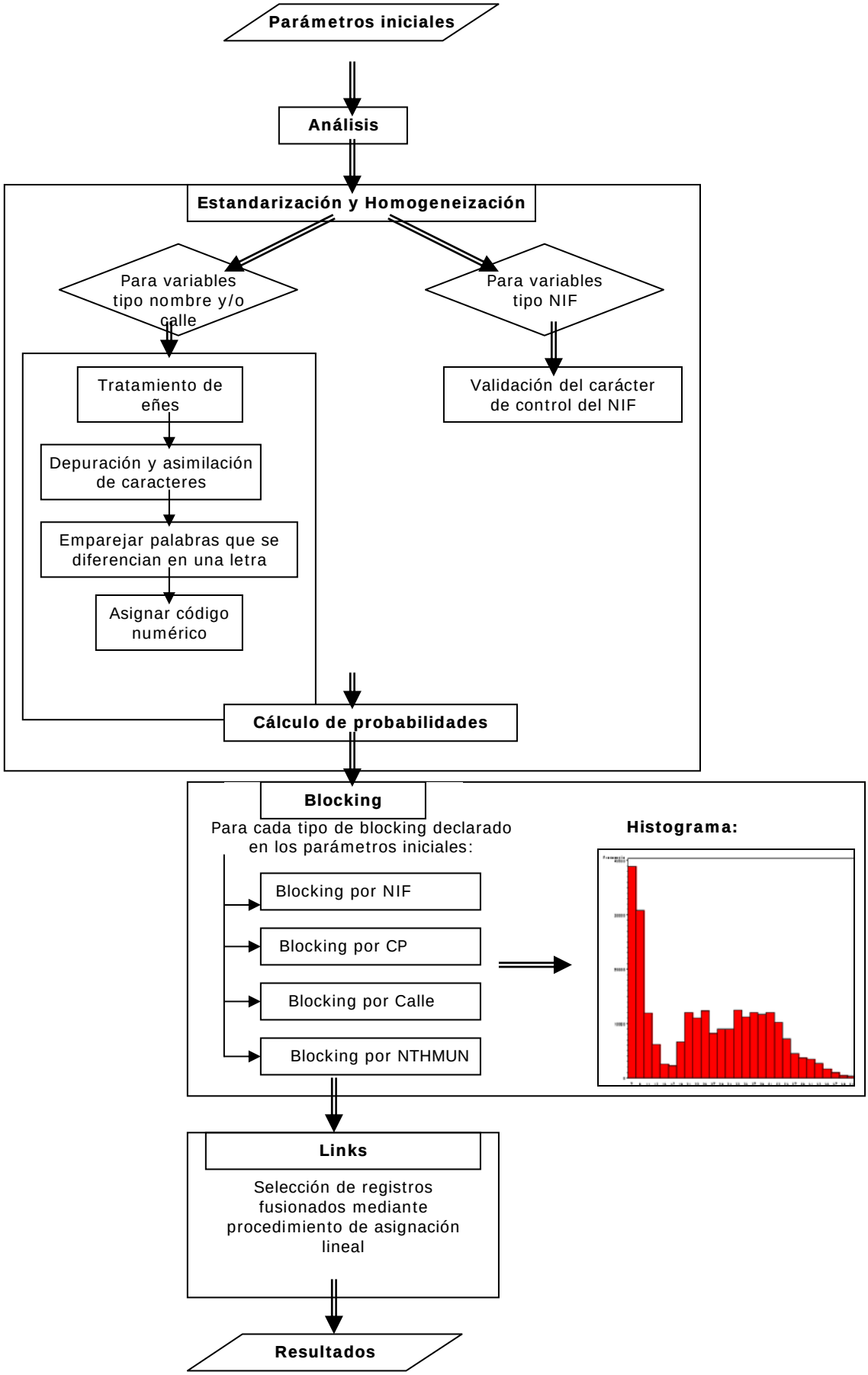
[1] Las macros de SAS desarrolladas en este proyecto estan a disposición los interesados previa petición a EUSTAT

- Un conjunto de datos que contiene algunas siglas de razón social que se eliminan de las variables de fusión alfabéticas en el paso de estandarización y homogeneización.

Este programa auxiliar debe ejecutarse al menos una vez previamente a la ejecución del programa de fusión para que las tablas existan cuando se ejecute el programa.

Estas tablas se han construido con un número determinado de registros, pero es conveniente ir añadiendo registros a estas tablas a medida que se vayan obteniendo resultados, de manera que lleguen a contener el mayor número de partículas y siglas de razón social posible, y así que el paso de estandarización y homogeneización sea más eficiente.

A continuación se muestra un diagrama de flujo que representa gráficamente el algoritmo del programa de fusión, que se describe posteriormente:


Parámetros iniciales
Análisis
Estandarización y Homogeneización
Para variables tipo nombre y/o calle
Para variables tipo NIF
Tratamiento de eñes
Depuración y asimilación de caracteres
Emparejar palabras que se diferencian en una letra
Asignar código numérico
Validación del carácter de control del NIF
Cálculo de probabilidades
Blocking
Para cada tipo de blocking declarado en los parámetros iniciales:
Blocking por NIF
Blocking por CP
Blocking por Calle
Blocking por NTHMUN
Histograma:
Links
Selección de registros fusionados mediante procedimiento de asignación lineal
Resultados

Parámetros iniciales

Para el correcto funcionamiento del programa de fusión se han de inicializar algunos parámetros antes de proceder a ejecutar el programa de fusión. A continuación se presenta brevemente los datos input que debe introducir el usuario:

- Hay que indicar cuales son las variables de fusión, esto es, el nombre que toma la variable en cada uno de los dos ficheros y de qué tipo de variable se trata. Los posibles tipos de variable que pueden utilizarse en la fusión y el código que les corresponde son los siguientes:

CÓDIGO	VARIABLE DE FUSIÓN
TIPO 0	CLAVE IDENTIFICADORA de registro
TIPO 1	NOMBRE
TIPO 2	CALLE
TIPO 3	MUNICIPIO
TIPO 4	PROVINCIA
TIPO 5	TELÉFONO
TIPO 6	CÓDIGO POSTAL
TIPO 7	NIF

Se debe destacar que el TIPO 0 (CLAVE IDENTIFICADORA de registro) no es una variable de fusión sino un código que identifica de manera unívoca cada elemento dentro de su registro administrativo. Sin embargo es una variable imprescindible puesto que los pares de elementos fusionados serán identificados mediante sus correspondientes claves en los registros administrativos.

- Hay que indicar al menos un criterio de blocking. Como se ha comentado anteriormente a pesar de que el blocking no es una técnica implícita de fusión si que es imprescindible en la práctica cuando se está trabajando con ficheros de cierta dimensión. A continuación se muestran los posibles criterios de blocking que pueden utilizarse y el código que les corresponde.

CÓDIGO	TIPO BLOCKING	DESCRIPCIÓN
Criterio 1	Blocking NIF	Coincidencia del NIF
Criterio 2	Blocking CP	Coincidencia del código postal o códigos postales (si hay más de uno)
Criterio 3	Blocking Calle	Coincidencia del código del literal de la calle o calles (si hay más de una)
Criterio 4	Blocking NTHMUN	Coincidencia del código de la primera palabra del nombre de la empresa además del territorio histórico y del municipio de la empresa

- Hay que establecer algunos parámetros iniciales teóricos (ver bibliografía [1]):

e_A , e_B	Probabilidad de que el valor de la variable de fusión se haya registrado erróneamente en cada uno de los ficheros, respectivamente. Se asume que la probabilidad de estar mal registrado es independiente de cada uno de los valores particulares
e_T	Probabilidad de que el valor de la variable de fusión aparezca de forma distinta en ambos ficheros a pesar de estar bien registrado en ambos. Esto puede ocurrir, por ejemplo, si los ficheros fueron creados en momentos distintos y la unidad económica cambió de nombre

Se asume que los valores de los parámetros e_A y e_B son suficientemente pequeños como para que la probabilidad de una coincidencia entre dos entradas idénticas, aunque erróneas, sea insignificante.

- En un momento dado el programa solicita al usuario introducir un valor para el límite que determina a partir de qué peso se consideran fusionados los pares de unidades económicas.

Análisis

El primer paso del programa es el análisis de las variables de fusión y de la variable tipo clave declaradas previamente.

Para ello se realizan ciertas comprobaciones como que se haya establecido al menos una variable de fusión y un criterio de blocking.

En caso de que haya algún error o no se haya introducido correctamente algún dato, el programa finaliza su ejecución y se muestra por pantalla en el *log* un mensaje identificando el error cometido que el usuario debe solventar.

Si por el contrario se ha introducido toda la información necesaria el programa continúa su ejecución.

Estandarización y homogeneización.

Las tareas de estandarización y homogeneización son de gran importancia en los procesos de fusión puesto que un buen tratamiento previo de las variables puede aumentar considerablemente el número de pares fusionados, y lo que es más importante, fusionar pares de mayor calidad, es decir, proporcionan mayor seguridad de que realmente corresponden a la misma unidad.

Hay dos clases de variables de fusión que deben ser estandarizadas: las variables de fusión alfabéticas (nombre de la unidad económica y calle) y el NIF. Cada una de estas clases tiene su propia estandarización. A continuación se describe el proceso de estandarización y homogeneización de cada una de estas clases de variable de fusión.

Variables de fusión alfabéticas.

La estandarización y homogeneización de las variables de fusión alfabéticas consta de diversas etapas. El objetivo final es obtener dos nuevos campos: *est_var** y *cod_var**, suponiendo que la variable de fusión alfabética a estandarizar sea *var**.

El nuevo campo *est_var** contiene el literal estandarizado de la variable de fusión original *var**, mientras que el nuevo campo *cod_var** contiene un código numérico que identifica el valor de la variable de fusión. De esta manera ocurre que un grupo de nombres de unidades económicas o calles que en su valor original son diferentes pero se han estandarizado de la misma manera están representados por el mismo código *cod_var**.

A continuación se describe brevemente los diferentes tratamientos de estandarización y homogeneización que se desarrollan en el programa de fusión:

- ***Tratamiento de eñes***

  Debido a que los ficheros de registros administrativos pueden haberse generado con diferentes codificaciones de caracteres en los editores de texto o programas, es usual que ciertos caracteres como las vocales acentuadas o las Ñ aparezcan mal codificados. Esta macro corrige la aparición de los símbolos /, # y ¥ en lugar de Ñ.

  La macro estudia todas las palabras que contienen los símbolos /, # y ¥ y se analiza si pueden ser errores tipográficos que en realidad corresponden a la letra Ñ. Hay que tener cuidado puesto que se pueden encontrar ocurrencias de los símbolos /, # y ¥ que no deban ser corregidos por una Ñ.

  Este tipo de errores se evitan corrigiendo únicamente aquellas palabras que cumplen que la nueva palabra formada al sustituir el símbolo /, # y ¥ por Ñ en la palabra original existe en cualquier otro elemento.

- ***Depuración y asimilación de caracteres***

  En esta etapa se procura homogeneizar lo máximo posible las variables de fusión de manera que dos literales que corresponden a un mismo valor no parezcan diferentes debido a errores tipográficos.

  Para ello se efectúan las siguientes tareas de depuración:

  - Se eliminan signos de puntuación tales como el punto, la coma o el guión.
  - Si la variable de fusión alfabética es nombre de empresa o establecimiento se eliminan los caracteres de sigla social.
  - Se eliminan los acentos de las vocales.
  - Se eliminan caracteres no alfanuméricos como paréntesis, asteriscos, almohadillas, etc.
  - Se eliminan artículos y preposiciones que no aportan información relevante a la variable de fusión alfabética.
  - Se consideran con una única grafía los siguientes caracteres o grupos de caracteres susceptibles de error debido a similitudes fonéticas o gráficas.

Caracteres o grupos de caracteres que se representan con la misma grafía		
Y	I	
TX, TS, CH	TZ	
K	C	Si preceden a las vocales A, O, U
K	QU	Si preceden a las vocales E, I
N	M	Si preceden a las consonantes B, P
V	B	
Ñ	N	
GU	G	
Z	C	Si preceden a las vocales E, I excepto cuando la Z es precedida por una T

- ***Emparejar palabras que se diferencian en una letra***

En esta fase se estudian pares de palabras que sólo se diferencian en una letra, siendo ésta una letra en concreto. En la siguiente tabla se presentan los pares de letras que se analizan en el programa.

C	K	K	C	K	QU	QU	K
C	TZ	TZ	C	L	LL	LL	L
C	Z	Z	C	M	N	N	M
C	X	X	C	Q	QU	QU	Q
C	Q	Q	C	R	RR	RR	R
C	QU	QU	C	S	X	X	S
G	J	J	G	S	TZ	TZ	S
I	J	J	I	S	Z	Z	S
I	LL	LL	I	TZ	X	X	TZ
J	X	X	J	TZ	Z	Z	TZ
K	Q	Q	K	X	Z	Z	X

Si al cambiar en una palabra una letra por su correspondiente pareja coincide con otra palabra existente en el conjunto de los ficheros se consideran la misma, y por tanto se les asigna el mismo código numérico.

- ***Asignar códigos numéricos***

  Se tiene un listado de todas las palabras de la variable de fusión alfabética que se está tratando junto con su estandarizado y su código numérico. Lo que se hace en esta última fase es asignar a los valores originales de la variable de fusión alfabética en cada uno de los ficheros a fusionar los estandarizados y códigos numéricos correspondientes.

NIF.

Esta fase de estandarización clasifica y valida perfectamente todos los códigos fiscales que se usan en España. Se analiza un campo de 9 caracteres alfanuméricos y devuelve un valor numérico para cada tipo de código analizado donde todos los valores positivos (mayores que cero) indican que el código fiscal es correcto.

Se asume que el código fiscal es correcto cuando el carácter de control es el que le corresponde. El carácter de control es una función de las cifras o letras que componen la identificación fiscal y eventualmente de las posiciones que ocupan. Su inclusión tiene como fin detectar y, en consecuencia, evitar los errores de trascripción y digitación de dicha identificación. Por tanto, lo que se hace es calcular el carácter de control que corresponde al código y comprobar si coincide.

A continuación se muestran los valores que puede devolver la función programada para la validación del carácter de control del NIF:

Tipo	Desconocido	NIF	CIF	NIE	CIF Temporales
Correcto:		1	2	3	4
Incorrecto:	0	-1	-2	-3	

Esta macro cumple con las siguientes especificaciones de las leyes españolas:

Decreto 2423/1975, de 25 de septiembre

Real Decreto 338/1990, de 9 de marzo

Real Decreto 1624/1992, de 29 de diciembre que modifica el 338/1990

Real Decreto 155/1996, de 2 de febrero

Orden de 3 de julio de 1998, por la que se modifica el Anexo del Decreto 2423/1975

Real Decreto 1065/2007, de 27 de julio

Orden EHA/451/2008, de 20 de febrero de 2008

Orden INT/2058/2008, de 14 de julio de 2008

A continuación se describe el procedimiento de validación que se realiza a cada código fiscal:

- ***NIF de personas físicas***

Con carácter general, para las personas físicas españolas se utiliza como número de identificación fiscal el propio número del Documento Nacional de Identidad (DNI), mientras que para las personas físicas extranjeras se utiliza el Número de Identificación de Extranjero (NIE), asignados ambos por el Ministerio de Interior.

Para las personas físicas extranjeras la identificación comienza con una de las letras X, Y o Z.

Además, para los españoles menores de 14 años sin DNI o residentes en el extranjero que no disponen de NIE, Hacienda les asigna una clave de identificación que comienza con las letras K, L o M, según el caso.

El número del DNI consta a lo sumo de ocho cifras, mientras que para los demás casos de personas físicas españolas o extranjeras consta a lo sumo de siete (una vez eliminada la letra inicial K, L, M, X, Y o Z). En todos los casos el carácter de control es siempre una letra.

En la siguiente tabla se muestra una descripción de los distintos NIF de personas físicas:

Tipo	Formato	Comentario
DNI	Ocho números + dígito de control	Españoles con documento nacional de identidad asignado por el Ministerio de Interior
NIF K	K + 7 números + dígito de control	Españoles menores de 14 años
NIF L	L + 7 números + dígito de control	Españoles residentes en el extranjero sin DNI
NIF M	M + 7 números + dígito de control	NIF que otorga la Agencia Tributaria a extranjeros que no tienen NIE
NIF X	X + 7 números + dígito de control	Extranjeros identificados por la Policía con un número de identidad de extranjero (NIE) asignado hasta el 15 de julio de 2008
NIF Y	Y + 7 números + dígito de control	Extranjeros identificados por la Policía con un NIE asignado desde el 16 de julio de 2008 (Orden INT/2058/2008, BOE del 15 de julio)
NIF Z	Z + 7 números + dígito de control	Letra reservada para cuando se agoten los 'Y' para extranjeros identificados por la Policía con un NIE

A continuación se describe cómo se calcula el carácter de control:

DNI

Se toma el resto que se obtiene de dividir el número formado por los 8 números del DNI entre 23 y se le asigna la correspondiente letra según la siguiente tabla:

0	T	8	P	16	Q
1	R	9	D	17	V
2	W	10	X	18	H
3	A	11	B	19	L
4	G	12	N	20	C
5	M	13	J	21	K
6	Y	14	Z	22	E
7	F	15	S	23	T

NIF X, Y, Z

Se sustituye respectivamente la X por un 0, la Y por un 1 y la Z por un 2 y se procede del mismo modo que en el caso de un DNI estándar.

NIF K, L, M

Su carácter de control se calcula como si fuese un NIF de persona jurídica (descrito más adelante).

- ***NIF de personas jurídicas y entidades en general***

El Número de Identificación Fiscal es asignado por Hacienda a las personas jurídicas y a las entidades sin personalidad jurídica –sociedades mercantiles, instituciones, agrupaciones, etc.- y consta de nueve caracteres, siendo el noveno de ellos el carácter de control (un dígito o una letra).

El primer carácter corresponde a la forma jurídica, que puede ser A, B, C, D, E, F, G, H, J, N, U, V y W para las sociedades y entes mercantiles y P, Q, R y S para las congregaciones religiosas y entes y organismos de la administración pública.

En la siguiente tabla se describen los valores que puede tomar la primera letra según su naturaleza jurídica:

Letra	**Naturaleza jurídica**	**Carácter de control**
A	Sociedades anónimas	Numérico
B	Sociedades con responsabilidad limitada	Numérico
C	Sociedades colectivas	Numérico
D	Sociedades comanditarias	Numérico
E	Comunidades de bienes y herencias yacentes	Numérico
F	Sociedades cooperativas	Numérico
G	Asociaciones	Numérico
H	Comunidades de propietarios en régimen de propiedad horizontal	Numérico

J	Sociedades civiles, con o sin personalidad jurídica	Numérico
P	Corporaciones locales	Alfabético
Q	Organismos públicos	Alfabético
R	Congregaciones e instituciones religiosas	Alfabético
S	Órganos de la Administración del Estado y de las Comunidades Autónomas	Alfabético
U	Uniones temporales de empresas	Numérico
V	Otros tipos no definidos en el resto de claves	Numérico
N	Entidades extranjeras	Alfabético
W	Establecimientos permanentes de entidades no residentes en España	Alfabético

Se siguen los siguientes pasos para calcular la letra o dígito de control (recordar que este cálculo sirve tanto para calcular el carácter de control de un NIF de persona jurídica y entidades en general como el de los NIF de persona física que comienzan con las letras K, L y M). Para ello se parte de las siete cifras situadas en las posiciones 2ª a 8ª, en el orden en que están escritas. Sean $a_1, a_2, \ldots, a_7$, entonces

1. Sea $A = a_2 + a_4 + a_6$ la suma de los dígitos con subíndice par

2. Sea $B = b_1 + b_3 + b_5 + b_7$ donde $b_i =$ suma de los dígitos $2xa_i$, $i = 1,3,5,7$

3. Sea $C = A + B$, $E =$ último dígito de C y $D = 10 - E$ (si $E = 0$, entonces $D = 0$)

4. Si el carácter de control es un dígito entonces es D. En cambio si el carácter de control es alfabético se toma la letra correspondiente al valor de D en la siguiente tabla:

Valor de D	1	2	3	4	5	6	7	8	9	0
Carácter de control	A	B	C	D	E	F	G	H	I	J

Cálculo de probabilidades.

Al comparar los valores de una variable de fusión de un par de registros se pueden dar tres tipos distintos de indicadores:

$$\gamma = \begin{cases} \gamma_1 & \equiv \quad los\ valores\ coinciden\ y\ son\ el\ \ j-ésimo \\ \gamma_2 & \equiv \quad los\ valores\ no\ coinciden \\ \gamma_3 & \equiv \quad a\lg ún\ valor\ está\ ausente \end{cases}$$

donde $j = 1, \ldots, m$ indica un valor concreto de los m distintos valores que puede tomar la variable de fusión en cuestión.

En esta fase se calculan las siguientes probabilidades:

- Probabilidad de que los valores de la variable de fusión coincidan y sean igual al j-ésimo valor posible de la variable de fusión, dado que el par de registros representa a la misma unidad económica ($m(\gamma_1)$).
- Probabilidad de que los valores de la variable de fusión coincidan y sean igual al j-ésimo valor posible de la variable de fusión, dado que el par de registros no representa a la misma unidad económica ($u(\gamma_1)$).
- Probabilidad de que los valores de la variable de fusión no coincidan, dado que el par de registros representa a la misma unidad económica ($m(\gamma_2)$).
- Probabilidad de que los valores de la variable de fusión no coincidan, dado que el par de registros no representa a la misma unidad económica ($u(\gamma_2)$).
- Probabilidad de que alguno de los valores de la variable de fusión esté ausente, dado que el par de registros representa a la misma unidad económica ($m(\gamma_3)$).
- Probabilidad de que alguno de los valores de la variable de fusión esté ausente, dado que el par de registros no representa a la misma unidad económica ($u(\gamma_3)$).

Los cálculos detallados de estas probabilidades pueden consultarse en la documentación referida en la bibliografía.

Blocking.

A continuación se efectúan los criterios de blocking declarados por el usuario en el programa. Además, para el conjunto de pares de elementos que selecciona el criterio de blocking se calcula el peso final de cada uno de ellos.

Según los criterios de blocking que haya declarado el usuario se ejecuta la macro correspondiente para obtener finalmente un conjunto de datos con aquellos pares de elementos que cumplen alguno de los criterios de blocking.

- **%blockingNIF**. Selecciona los pares de elementos para los cuales coincide el NIF de la empresa.
- **%blockingCP**. Selecciona los pares de elementos para los cuales coincide el código postal. En este caso puede que haya más de un código postal. Por lo tanto, en el caso de que haya más de una variable de fusión tipo código postal, el blocking se realiza para la coincidencia de todas las variables de ese tipo. Es decir, se relacionan los pares de elementos para los cuales coinciden todas y cada una de las variables de tipo código postal.

- **%blockingCalle**. Selecciona los pares de elementos para los cuales coincide el código asociado al literal de la calle. En este caso puede que haya más de una variable de fusión tipo literal de la calle. Por lo tanto, en el caso de que haya más de una variable de fusión tipo literal de la calle, el blocking se realiza para la coincidencia de todas las variables de ese tipo. Es decir, se relacionan los pares de elementos para los cuales coinciden todas y cada una de las variables tipo literal de la calle.

- **%blockingNTHMUN**. Selecciona los pares de elementos para lo cuales coincide la primera palabra del nombre de la empresa codificada además del territorio histórico y el municipio de la empresa.

Se pueden declarar tantos criterios de blocking como se quiera. Se recomienda declarar más de uno para evitar que un par de elementos que representan a la misma unidad económica no sea fusionado debido a errores en la variable de blocking.

Una vez construido el conjunto de pares de elementos que cumplen algún criterio de blocking se calcula para cada uno de ellos su peso total de la siguiente manera

$$w = \sum_{k=1}^{K} w_k \text{ donde } w_k = \log \frac{m(\cdot)}{u(\cdot)} = \log m(\cdot) - \log u(\cdot), \quad k = 1, \ldots, K$$

siendo K el número de variables de fusión declaradas por el usuario en el programa.

Una vez que se ha calculado el peso total para todos los pares de elementos que cumplen algún criterio de blocking se muestra al usuario un histograma con la frecuencia de dichos pares para que determine el peso límite.

Links

Dado un par de elementos puede ocurrir dos cosas: que los elementos representen a la misma unidad económica o que sean unidades económicas diferentes.

En una situación teórica ideal, si los dos elementos representasen a la misma unidad económica deberían coincidir todas sus variables de fusión y por lo tanto, tener un peso total muy alto. Sin embargo, esto no ocurre en todos los casos puesto que pueden existir diferencias en algunas variables de fusión debido a errores tipográficos o a cambios de estado o debido a que algún valor de alguna variable de fusión está ausente.

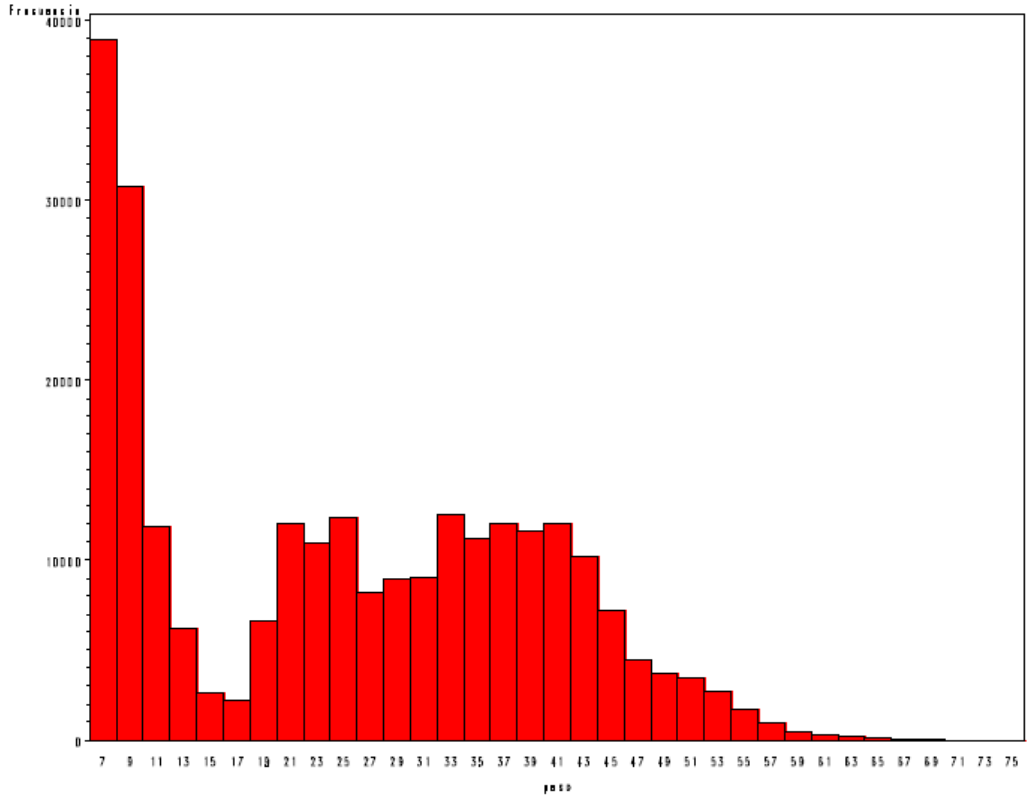
Por el contrario, si los dos elementos representan unidades económicas diferentes no deberían coincidir las variables de fusión y el par de elementos debería tener un peso muy bajo (incluso negativo). Sin embargo, pueden existir coincidencias casuales en algunas variables de fusión que hagan aumentar el peso.

En consecuencia, en la situación teórica ideal el histograma estaría compuesto de dos claras secciones, una a la derecha alrededor de un peso alto que correspondería con los pares de elementos que representan a la misma unidad económica, y otra en el lado izquierdo alrededor de un peso bajo que correspondería a los pares de elementos que representan unidades económicas diferentes.

En la práctica estas secciones no están tan claramente diferenciadas y existen valores en los pesos intermedios formando una curva.

Entonces, el usuario debe observar el histograma y establecer un peso a partir del cual todos los pares de elementos con peso superior o igual al peso establecido como límite serán considerados fusionados. Se recomienda tomar como peso límite el valle del histograma.

A continuación se muestra un ejemplo de histograma de pesos:



Una vez que el usuario ha decidido el valor del peso límite se realiza un procedimiento de asignación lineal con los pares de unidades que han sido seleccionados en la fase de blocking para establecer los elementos que se consideran fusionados.

Capítulo

5

ANÁLISIS DE LOS RESULTADOS

En esta sección se va a describir la aplicación del programa de fusión de registros orientada a ficheros administrativos con información de unidades económicas descrito previamente a los siguientes ficheros: el Directorio de Actividades Económicas (DIRAE) y el fichero de la Seguridad Social.

Descripción de los ficheros

Directorio de Actividades Económicas (DIRAE)

La información del Directorio de Actividades Económicas viene proporcionada por los siguientes tres conjuntos de datos:

- Un fichero que contiene los datos de DIRAE correspondientes a unidades de actividad económica local.

  La unidad de actividad económica local se trata de una subdivisión de la unidad local en base a criterios de actividad, exclusiva de DIRAE.

- Un fichero que contiene los datos del DIRAE correspondientes a unidades jurídicas.

  Las unidades jurídicas son tanto personas jurídicas cuya existencia está reconocida por la ley independientemente de las personas o instituciones que las posean o sean miembros de ellas como las personas físicas que, en calidad de independientes, ejercen una actividad económica.

- Un fichero que contiene los datos del DIRAE correspondientes a unidades locales.

  La unidad local corresponde a una empresa o a una parte de empresa (taller, fábrica, almacén, oficinas, mina, depósito) sita en un lugar delimitado topográficamente. En dicho lugar o a partir de él se realizan actividades económicas a las que –salvo excepciones- dedican su trabajo una o varias personas (llegado el caso, en jornada parcial) por cuenta de una misma empresa.

Seguridad Social

La información del fichero de la Seguridad Social viene dada en dos conjuntos de datos:

- Un fichero que contiene datos de la Seguridad Social correspondientes a empresarios.

- Un fichero que contiene datos de la Seguridad Social correspondientes a unidades de cotización.

El fichero de la Seguridad Social es de actualización trimestral por lo tanto en un año natural se reciben 4 ficheros diferentes. El fichero empleado en la fusión de registros es un fichero que computa todas las variantes de un registro durante un año. Es decir cuando un establecimiento ha variado alguna de sus características este registro se añade al fichero y puede haber establecimiento que se encuentren hasta cuadriplicados.

Análisis de los ficheros

El paso previo a la fusión de los ficheros administrativos es la construcción de los conjuntos de datos que van a fusionarse. Para ello se construye para cada fuente un único conjunto de datos con toda la información necesaria para la fusión, es decir, las claves que identifican cada unidad en el registro y todas las variables en común de ambos registros que sirvan para la fusión. En nuestro caso se construyeron los conjuntos de datos **ula.sas7bdat** con la información de cada unidad local de DIRAE y **uco.sas7bdat** con la información de las unidades de cotización de la Seguridad Social.

A continuación se muestra una tabla con las variables que van a considerarse en la fusión:

DIRAE			**SEG. SOCIAL**		
UNIDAD JURÍDICA	UJA_CIF	CIF	**EMPRESARIO**	EMP_CIFDNI	CIF/DNI
	UJA_NOMBRE	Denominación		EMP_NOMBRE	Denominación
	UJA_PROV	Provincia		EMP_TH	Territorio histórico
	UJA_MUN	Municipio		EMP_MUN	Municipio
	UJA_CP	Código postal		EMP_CP	Código postal
	UJA_CALLE	Literal de la calle		EMP_T_CALLE	Literal de la calle
UNIDAD LOCAL	ULA_TH	Territorio histórico	**UNIDAD DE COTIZACIÓN**	UCO_TH	Territorio histórico
	ULA_MUN	Municipio		UCO_MUN	Municipio
	ULA_CP	Código postal		UCO_CP	Código postal
	ULA_CALLE	Literal de la calle		UCO_T_CALLE	Literal de la calle

Como se puede observar, las variables de fusión disponibles son pocas. La variable *CIF/DNI* es una variable de fusión que identifica de manera unívoca a la empresa. La variable *Denominación* también tiene un gran poder discriminante, aunque es

susceptible de contener muchos errores. Al contrario, el resto de variables de localización (*Provincia, Municipio, Código postal y Literal de la calle*) son poco representativas.

Además del poder discriminante de cada variable de fusión también es importante tener en cuenta que no todos los elementos de los registros administrativos contienen información para todas las variables. En la siguiente tabla se muestra el número de elementos que tienen ausentes cada una de las variables:

DIRAE (ula.sas7bdat)			**SEG. SOCIAL (uco.sas7bdat)**		
200.675 registros			554.177 registros		
Variable	**NMISS**	**PCTMISS**	**Variable**	**NMISS**	**PCTMISS**
UJA_CIF	0	0%	EMP_CIFDNI	265	0.05%
UJA_NOMBRE	0	0%	EMP_NOMBRE	1	0.00%
UJA_PROV	0	0%	EMP_TH	143673	25.93%
UJA_MUN	0	0%	EMP_MUN	143673	25.93%
UJA_CP	18	0.01%	EMP_CP	140996	25.44%
UJA_CALLE	179	0.09%	EMP_T_CALLE	143673	25.93%
ULA_TH	0	0%	UCO_TH	112027	20.22%
ULA_MUN	1437	0.72%	UCO_MUN	112027	20.22%
ULA_CP	5958	2.97%	UCO_CP	101856	18.38%
ULA_CALLE	1653	0.82%	UCO_T_CALLE	104213	18.81%

Tal y como se muestra en la tabla anterior a pesar de que el DIRAE es muy completo, las variables de localización del fichero de la Seguridad Social tienen muchos valores ausentes, lo cual dificulta sobremanera la fusión.

Se han hecho pruebas utilizando distintos criterios de blocking. En todos los casos se ha utilizado el blocking por NIF puesto que es una variable que aporta mucha información y además, está bastante bien registrada.

Esto se sabe gracias a la macro programada para estudiar la validez del carácter de control. A continuación se muestra una tabla con los resultados obtenidos en dicho análisis:

Resultado de ValidarNIF.sas	**DIRAE**	**Seguridad Social**

-3	8	1
-2	38	0
-1	55	6
0	140	265
1	102109	424931
2	92616	112602
3	5706	16372
4	3	0
Total	200675	554177

Se recuerda que los valores negativos indican que el carácter de control es erróneo y el 0 indica que el formato del NIF no es estándar (es decir, no consta de 9 caracteres). Por lo tanto, en la tabla se puede observar que:

- En cuanto a las unidades locales del DIRAE el 0.05% de los establecimientos tienen un NIF erróneo y el 0.07% tiene un formato del NIF incompleto. Además, hay 3 establecimientos con un NIF temporal (código de validación del NIF= 4) que no se puede validar.

- En cuanto a las unidades de cotización de Seguridad Social, el 0.001% tienen un NIF erróneo y el 0.05% un formato de NIF incompleto.

Resultado de la fusión

A continuación se muestra una tabla con el resumen de los resultados obtenidos al ejecutar el programa de fusión con los distintos criterios de blocking.

Criterios de blocking	**Peso límite**	**Tiempo de ejecución**	**Fusionados**		
			Fase 1	**Fase 2**	**Total**
NIF	20	10 horas	14748	111948	126696
NIF + Calle	28	Abortado tras 39horas	--------	---------	---------
NIF + NTHMUN	32	6 horas	14748	80040	94788

En la Fase 1 se fusionan los pares de elementos para los cuales coinciden todas las variables de fusión, es decir, se trata de una fusión directa. En la Fase 2 se fusionan los pares de elementos seleccionados por el algoritmo de asignación lineal entre los que tienen un peso superior al del peso límite, es decir, son los pares de elementos fusionados por el procedimiento probabilístico.

Capítulo

6

CONCLUSIONES

Usar información de origen administrativo presenta muchas dificultades a la hora de ser utilizada en el entorno estadístico. Los registros de origen administrativo normalmente no han sido concebidos ni diseñados con fines estadísticos y usar la información que nos proporciona necesita de un tratamiento previo.

Con este proyecto hemos convertido el registro de la seguridad social en un agil instrumento para uso estadístico en la construcción y mantenimiento del DIRAE. El DIRAE es para nosotros un directorio de referencia fundamental sin el que no se podría llevar a cabo una estadística fiable y de calidad. La información procedente de la Seguridad Social se emplea para actualizar variables tan importantes como el número de empleados.

Los resultados de la aplicación del método de fusión probabilístico junto con los tratamientos previos, presentan unos porcentajes más que aceptables de fusión de registros, teniendo en cuenta la calidad en las variables de fusión. Tal como hemos descrito en el cuaderno, muchas de las variables auxiliares empleadas en la fusión presentan valores missing que dificultan cualquier tipo de tratamiento. La fusión realizada de un modo “directo” no llegaría ni a un 10% de los establecimientos.

La idea general del proyecto se planteo para fusionar dos registros cualesquiera de establecimientos, tal y como se hizo anteriormente con individuos. Tras el estudio de los registros de establecimientos disponibles se observó que las peculiaridades de cada uno impedían un tratamiento conjunto, por tanto se opto por programar una macro secuencial y modular de tal modo que los módulos puedan ser reutilizados y adaptados para cualquier futura fusión de establecimientos.

El estudio se podrá mejorar en la medida que los registros administrativos mejoren también y en la medida que se estudien nuevas vías para realizar un blocking efectivo, que permita dividir el fichero en bloques lo suficientemente pequeños para que el tiempo de ejecución (numero de comparaciones) sea aceptable y lo suficientemente amplio para no perder calidad y posibles parejas de datos.

Para evaluar los resultados obtenidos se realizó un control de calidad seleccionando varias muestras aleatorias de pares de unidades fusionados en la segunda fase de la ejecución del programa de fusión. En la primera fase se fusionaron los pares de unidades económicas para los cuales coincidieron todas las variables de fusión. En la segunda fase se fusionaron los pares de registros que determina la metodología probabilística empleada. Para la mayoría de los registros fusionados se comprobó que la fusión era correcta o por lo menos lógica. Por ejemplo, se consigue fusionar establecimientos únicos de la misma empresa aunque tengan importantes variables ausentes de fusión.

El resultado de este proyecto ha sido por tanto más que satisfactorio y a partir de ahora nos permitirá disponer de una herramienta más, que redundará en la calidad del directorio de actividades económicas de EUSTAT

Capítulo 7

BIBLIOGRAFÍA

[1] EUSTAT

MÉTODOS AUTOMÁTICOS DE FUSIÓN DE REGISTROS Y SU UTILIZACIÓN EN EUSTAT. http://www.eustat.es/document/datos/ct_15_c.pdf

[2] I.P. FELLEGI AND A.B. SUNTER

A theory for record linkage. *Journal of the American Statistical Association, 64(328):1183-1210, 1969*

[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] AYESTARAN, MARINA AND LEGARRETA, LEIRE.

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