Procedimientos de fusión: la experiencia de Eustat y nuevos desarrollos.

MORAN ALAEZ Enrique

GONZALEZ HERNANDEZ Martín

MARTINEZ ROLLON Pilar

ORTUÑEZ ALONSO Alejandro

OTERO FRANCO Laura

22 - 29 AUG, Lisboa 2007

1.- **INTRODUCCIÓN**

Desde 1996, cuando tuvo lugar la última renovación del Padrón Municipal de Habitantes, Eustat ha venido trabajando con registros administrativos a fin de conformar un Registro Estadístico de Población. Para ello se decidió preparar una regulación legal, basada en las competencias atribuidas al Gobierno Autónomo del País Vasco, mediante la cual se requería a los Ayuntamientos vascos la entrega anual de dos ficheros: el primero con el Padrón Municipal completo a 31 de diciembre de cada año y el segundo con los movimientos padronales ocurridos durante el año, es decir, todo tipo de altas y bajas padronales: nacimientos, defunciones, inmigraciones, emigraciones, cambios de domicilio dentro del municipio, altas por omisión y bajas por duplicados e inclusiones indebidas. De esta manera, se preveía realizar un contraste entre los movimientos registrados a lo largo del año y la situación al final del mismo, teniendo en cuenta que la situación de partida era conocida previamente.

Por otra parte, Eustat ha venido trabajando así mismo con las Estadísticas del Movimiento Natural de la Población, Nacimientos, Defunciones y Matrimonios, cuya fuente original son los asientos de dichos sucesos en los Registros Civiles, que en España están asociados ya sea a los Juzgados, en los municipios mayores, ya a los propios Ayuntamientos, en los más pequeños. Esta información registral se traslada a unos cuestionarios estadísticos pero, en la práctica, las informaciones relativas a las personas afectadas coinciden. La información nominativa que contienen ha ido variando a lo largo del tiempo y, en particular, el DNI o equivalente se ha introducido tardíamente, pues si en un primer momento se consideraba irrelevante a efectos estadísticos, más tarde ha pasado a ser fundamental como una de las vías de actualización de los Registros de Población.

Los registros municipales de población contienen una información sobre los residentes que está fijada legalmente y que comprende el nombre y apellidos, la fecha de nacimiento, el sexo, el número del DNI o equivalente y la dirección completa del domicilio. Aunque cabría la posibilidad de enlazar los registros de una persona mediante el DNI, esta opción está limitada por la falta de información para los menores de 14 años y por las incorrecciones observadas en su cumplimentación. Estas carencias de los identificadores personales numéricos han obligado a enfrentarse al problema de la fusión de ficheros a través de procedimientos automáticos de asociación de personas que contemplen toda la variedad de elementos de identificación.

2.- **SITUACIÓN ACTUAL**

Se dispone de procedimientos de fusión específicos para la asociación de registros procedentes de los ficheros padronales y de los ficheros de movimientos padronales de población. También se dispone de procedimientos de fusión para los nacimientos, las defunciones y los matrimonios, es decir, el conjunto de las estadísticas vitales. Estos procedimientos están insertos en el Registro Estadístico de Población de Eustat, en concreto en una base de datos Oracle denominada “Población”, y no están accesibles desde otras bases de datos diferentes o llamados por aplicaciones externas, lo cual constituye una limitación importante.

Se basan en la equivalencia de los identificadores personales que hemos ido mencionando: nombre y apellidos de las personas, número del DNI, si lo hubiere, o alternativamente NIE –número de identificación de extranjeros, fecha de nacimiento, sexo y dirección postal de la vivienda. Se asignan pesos a cada una de las coincidencias, aceptando inclusive coincidencias en parte de cada uno de los identificadores (por ejemplo, en un solo apellido), pero ponderándolos menos en tal caso. De esta forma, se seleccionan todos los pares de candidatos a la fusión que superan un cierto umbral de coincidencia, eligiéndose al que tiene la mejor puntuación en la comparación.

Naturalmente, antes es preciso diferenciar los casos en que una misma persona puede aparecer repetida en un fichero (por ejemplo, varios movimientos migratorios) frente a aquellos en que esto es imposible (por ejemplo, una defunción).

De forma más precisa, el procedimiento de fusión aplicado a los datos de los Padrones Municipales tiene las siguientes fases: primero, los ficheros de entrada –padrón y movimientos- se dividen en cabecera (vivienda) y detalle (personas) y se separan los nombres y apellidos que contienen, escribiéndolos en una tabla especialmente protegida y encriptada para protegerla; segundo, esta información se vuelca a una tabla de maniobra incorporando una normalización sencilla del DNI y de los nombres y apellidos, los cuales no son utilizados directamente sino mediante una transformación; tercero, se procede a asociar los registros de la tabla de maniobra con los existentes en el Registro Estadístico de Población mediante la igualdad del DNI o identificador equivalente, del nombre y los dos apellidos, de la fecha de nacimiento completa y del sexo y de la dirección postal completa, compuesta por los campos: provincia, municipio, calle, número, piso, mano y puerta; cuarto, se asignan pesos a cada una de las parejas encontradas, ya que las sucesivas búsquedas por cada uno de los tres primeros caminos señalados puede dar lugar a más de un candidato potencial a la asociación, teniendo en cuenta que para elegir a uno de ellos se precisa que se haya superado una cierta puntuación mínima –equivalente a la igualdad de nombre y apellidos, fecha de nacimiento y sexo o DNI, fecha de nacimiento y sexo, por ejemplo, y que se elegirá al que tenga una puntuación más elevada en caso de que hubiera

varios; finalmente, de la pareja seleccionada, si la hubiera, se tomará la clave de identificación del Registro Estadístico de Población.

Esta descripción somera del procedimiento se complica en seguida si tenemos en cuenta que las coincidencias a puntuar son un total de 30, repartidas entre 7 coincidencias parciales y total de nombre y apellidos, un número equivalente de puntuaciones para la fecha de nacimiento, hasta 9 puntuaciones de distintos grados de coincidencia en el DNI o identificador similar, incluida su ausencia –ya que no está definido para los menores de edad, otras 3 puntuaciones sobre la dirección postal de la vivienda y 4 relativas a otros elementos de comparación, tales como el sexo y el lugar de nacimiento, que también forman parte de las informaciones obligatorias de los Padrones Municipales.

Existe la posibilidad de que en el paso anterior no se haya encontrado un candidato adecuado para la fusión e incluso de que no se haya encontrado ningún candidato en absoluto. Esto es aceptable puesto que todo proceso de actualización conlleva altas de personas que, lógicamente, no pueden asociarse a ningún individuo preexistente en el fichero.

De todas formas, por si la existencia de errores en los datos fuera el motivo que hubiera impedido la fusión, se realizan dos nuevas búsquedas, menos exigentes, la primera de las cuales parte de tres variantes simplificadas del nombre y apellidos, perdiendo cada vez uno de los tres elementos, y la segunda se realiza por la vía de parte de la fecha y el lugar de nacimiento.

1. Población por puntuación obtenida en la fusión y tipo (%)

	Total	Tipo de fusión		
		1ª	2ª	3ª
Total	100	88,0	4,0	1,2
Sin candidato	6,8			
Por debajo del umbral	2,3	0,2	1,2	0,9
En el umbral	0,2	0,0	0,2	0,0
Por encima del umbral	63,8	60,9	2,6	0,3
Puntuación máxima	26,9	26,9	0,0	0,0

La tabla adjunta muestra los resultados de una fusión relativa al Padrón Municipal de Habitantes de 2002 en la que se observa la existencia de un porcentaje del 9% de personas no fusionadas, entre las que se incluye un 4% de altas por nacimiento e inmigración y el predominio de la fusión principal. La segunda forma de la fusión y, sobre todo, la tercera añaden pocos emparejamientos e introducen algunos erróneos.

Para concluir este apartado, se constata la existencia de un porcentaje del 4 al 5% de la población que no ha sido adecuadamente fusionada con el procedimiento actual, lo que requiere la introducción de mejoras en el mismo en la medida en que dicho procedimiento se ha ido convirtiendo cada vez más en la forma de entrada de las informaciones externas en el Sistema Integrado de Información de Eustat.

La existencia de registros no fusionados, en general interpretados como altas en el Registro Estadístico de Población, tiende a incrementar de forma artificial el número de personas cuantificado por esta vía, lo hace además de manera diferente en los distintos ámbitos territoriales y para las diversas operaciones estadísticas y deteriora la calidad del sistema. Eustat evalúa periódicamente esta calidad, mediante encuestas y otros procedimientos, y realiza otras tareas de revisión y actualización extaordinarias para asegurar los datos, pero de todas formas se precisa introducir mejoras en el procedimiento de fusión y desarrollar nuevas funcionalidades para mantener la calidad de los datos en un contexto de producción de datos con base en fuentes administrativas.

Por otro lado, en Eustat se lleva varios años estudiando nuevas técnicas, como la fusión de registros probabilística y se ha desarrollado una metodología que fundamentalmente se basa en el artículo “A theory for Record Linkage” de Ivan P. Fellegi y Alan B. Sunter. El desarrollo informático de este método se ha hecho en SAS pero está prevista su reprogramación para adecuarlo al entorno de trabajo en Oracle.

3.- **NUEVOS DESARROLLOS**

Dada la multitud de nuevos ficheros administrativos que figuran entre los recogidos por Eustat para la realización de la Estadística de Población y Viviendas de 2006 y las mejoras que se han ido viendo como necesarias a lo largo de estos años, Eustat está introduciendo una serie de cambios en el procedimiento de fusión que se describen a continuación.

Como ya se mencionó, se hace preciso generalizar los procedimientos de fusión, convirtiéndolo en una herramienta de uso general, disponible para cualquiera que tenga que hacer frente a una fusión de registros de personas procedentes de ficheros distintos. Para ello se está programando una aplicación de fusión independiente, denominada Módulo de Fusión (MDF), con un entorno de seguridad adecuado, a la que se le darán los permisos de acceso a los datos que precise en función del usuario de la misma. Es decir, un técnico estadístico podrá diseñar una ejecución especial de fusión para la operación estadística de la que es responsable dentro de MDF; una vez preparada esta versión, podrá ser llamada desde la aplicación de la operación estadística que recibirá como resultado las claves de persona y de vivienda que precise.

Una vez independiente, MDF debe aceptar diferentes tipos de entradas de datos, que provendrán tanto de fuentes externas al Sistema Integrado de Información de Eustat, en concreto de ficheros de texto, bases de datos Access y libros Excel, como de bases de datos Oracle ya integrados. La descripción de estas entradas corresponde al usuario de MDF y se permite una flexibilidad total en cuanto a la ubicación de los archivos como a su diseño, que se describe en una hoja de Excel, siempre que tenga los permisos necesarios para ello; a estos efectos, el propio MDF genera unos informes de permisos que se precisan, limitados generalmente a lectura de ficheros o tablas de base de datos.

Una de las mejoras imprescindibles se refiere a completar la normalización de los identificadores, tanto por lo que respecta a nombres y apellidos como al DNI o identificador equivalente; se introducen, por ejemplo, una serie de tablas de equivalencias lingüísticas (euskera-español) para el nombre, de abreviaturas usuales de nombres y apellidos y de normalización de grafías –tema este particularmente importante en la actualidad debido a la incorporación de población extranjera de países que no utilizan el alfabeto latino.

Además, como se aceptan ficheros externos de varios tipos se ha previsto también una homogeneización de la fecha de nacimiento y de la codificación de ciertas variables importantes para el contraste y la ponderación (por ejemplo, el sexo); en lo que atañe a la dirección postal de la vivienda, se exige una codificación de los distintos elementos de la misma de acuerdo con los códigos de Eustat y, preferiblemente, la incorporación del identificador de vivienda creado por Eustat en 1996.

Otro elemento novedoso es la adición de variables de identificación indirectas, principalmente relativas al cónyuge, al padre y a la madre, ya que, por un lado, algunos ficheros presentan este tipo de información, como por ejemplo, los nacimientos y los matrimonios, en tanto que, por otra parte, el Registro Estadístico de Población conserva información de los vínculos familiares conocidos, aunque se haya terminado la convivencia en la misma vivienda; de esta manera la asociación de un miembro de la pareja pueda servir de vía auxiliar para encontrar a la otra y la identificación de un padre puede ser un elemento decisivo para encontrar también al hijo.

Se introduce una etapa de fusión probabilística que difiere de la anterior sobre todo en el cálculo de los pesos. Se toman los pares de registros susceptibles de provenir del mismo individuo y se comparan todos sus campos. Para cada uno, se calcula un cociente de probabilidades en función del resultado de la comparación y, en caso de coincidencia, de la frecuencia de dicho valor en ambos ficheros. Se suman los valores de cada campo y se obtiene así una puntuación final para cada par. Se fusionarán aquellos cuya puntuación supere el valor de un cierto límite.

A continuación, y una vez ejecutada la fusión, se obtienen una serie de tablas estadísticas que detallan las cifras y proporciones de personas y viviendas fusionadas de acuerdo de con variables de control que se consideran relevantes: territoriales, edad y sexo, por estratos de pesos obtenidos en la comparación. A ellas se añaden 2 más que nos indicarán al número de candidatos –viviendas o personas- que se han considerado a efectos de la fusión.

Finalmente se realiza una verificación estadística de los aciertos y errores, en base a una muestra aleatoria de parejas admitidas y rechazadas, clasificadas por etapas de fusión, de manera que se acredite siempre la proporción de unidades fusionadas correctamente o no. En relación con los no fusionados por no superar el umbral marcado, la revisión estadística tiene la misión de revisar la adecuación de la decisión tomada y la oportunidad de su revisión.

Las informaciones relativas a estas dos últimas fases se archivan como resultados de la fusión, al igual que se archivan los parámetros de la misma –carga, normalización utilizada, opciones de fusión, etc., para que pueda volver a repetirse en el futuro o utilizarse como base para el diseño de un nuevo proceso de fusión.

Con estas modificaciones, fruto de la experiencia de años dedicados al trabajo con ficheros administrativos, Eustat prevé estar en óptimas condiciones para enfrentarse a los retos del trabajo creciente con fuentes administrativas para la producción de datos estadísticos como a la integración de informaciones recogidas mediante encuestas directas en el Sistema Integrado de información, reforzando la calidad de los datos, permitiendo el contraste de fuentes, validando y analizando las faltas de respuesta y produciendo nuevas estadísticas como resultado de este trabajo.

RESUMEN

Desde 1997, cuando tuvo lugar la última renovación del Padrón Municipal de Habitantes, Eustat ha venido trabajando con registros administrativos a fin de conformar un Registro Estadístico de Población. Dada la inexistencia en España de identificadores personales fiables, esto ha obligado a enfrentarse al problema de la fusión de ficheros a través de procedimientos automáticos de asociación de personas presentes en varios de ellos.

En el momento actual se dispone de procedimientos de fusión específicos para la asociación de registros procedentes de los ficheros padronales y de los ficheros de movimientos padronales de población. También se dispone de procedimientos de fusión para los nacimientos, las defunciones y los matrimonios, es decir, el conjunto de las estadísticas vitales.

Estos procedimientos se basan en la equivalencia de los siguientes identificadores: nombre y apellidos de las personas, número del DNI, si lo hubiere, fecha de nacimiento completa y dirección postal de la vivienda. Se asignan pesos a cada una de las coincidencias, aceptando inclusive coincidencias en parte de cada uno de los identificadores (por ejemplo, en un solo apellido), pero ponderándolos menos en tal caso. De esta forma, se selecciona el candidato a la fusión que, superando un cierto umbral de coincidencia, tiene la mejor puntuación en la comparación. Naturalmente, antes es preciso diferenciar los casos en que una misma persona puede aparecer repetida en un fichero (por ejemplo, varios movimientos migratorios) frente a aquellos en que esto es imposible (por ejemplo, una defunción).

Dada la multitud de nuevos ficheros administrativos que figuran entre los recogidos por Eustat para la realización de la Estadística de Población y Viviendas de 2006, la experiencia y los conocimientos adquiridos en estos últimos años, se hace preciso independizar el procedimiento de fusión, de forma que acepte diferentes tipos de entradas de datos, al mismo tiempo que se completa con una mejora de la normalización de los identificadores, con la adición de variables de identificación indirectas, relativas a los familiares directos, y con una verificación sistemática de los aciertos y errores derivada de una muestra aleatoria de pares de registros asociados.

Instituto Vasco de Estadística (Eustat)

c/Donostia-San Sebastián, 1

E-01011 Vitoria-Gasteiz (España)

Enrique-Moran@eustat.es